BIOINFORMATICS

Vol. 19 no. 3 2003, pages 402–407
DOI: 10.1093/bioinformatics/btg010

Identification of key concepts in biomedical literature using a modified Markov heuristic

W. H. Majoros[1,*], G. M. Subramanian[2] and M. D. Yandell[3]

[1]The Institute for Genomic Research, Rockville, MD, 20850, USA, [2]Human Genome Sciences, Rockville, MD, 20850, USA and [3]Howard Hughes Medical Institute, Berkeley, CA, 94720, USA

Received on June 27, 2002; revised on September 26, 2002; accepted on September 29, 2002

ABSTRACT

Motivation: The recent explosion of interest in mining the biomedical literature for associations between defined entities such as genes, diseases and drugs has made apparent the need for robust methods of identifying occurrences of these entities in biomedical text. Such concept-based indexing is strongly dependent on the availability of a comprehensive ontology or lexicon of biomedical terms. However, such ontologies are very difficult and expensive to construct, and often require extensive manual curation to render them suitable for use by automatic indexing programs. Furthermore, the use of statistically salient noun phrases as surrogates for curated terminology is not without difficulties, due to the lack of high-quality part-of-speech taggers specific to medical nomenclature.

Results: We describe a method of improving the quality of automatically extracted noun phrases by employing prior knowledge during the HMM training procedure for the tagger. This enhancement, when combined with appropriate training data, can greatly improve the quality and relevance of the extracted phrases, thereby enabling greater accuracy in downstream literature mining tasks.

Contact: bmajoros@tigr.org

INTRODUCTION

The last few years have seen a remarkable growth of interest in mining the biomedical literature for various types of knowledge, including protein–protein interactions (e.g. Ono *et al.*, 2001; Jenssen *et al.*, 2001; Marcotte *et al.*, 2001; Stapley and Benoit, 2000; Ng and Wong, 1999; Blaschke *et al.*, 1999; Thomas *et al.*, 2000), novel hypotheses about disease (e.g. Swanson and Smalheiser, 1999), relations between drugs, genes and cells (e.g. Friedman *et al.*, 2001; Rindflesch *et al.*, 2000; Tanabe *et al.*, 1999), protein structure (e.g. Demetriou *et al.*, 2000), and protein function (e.g. Andrade and Valencia, 1998). Although these works collectively employ a diverse array of techniques in the pursuit of an equally diverse set of goals, an obvious feature of most of them is a strong dependence on their ability to reliably identify named entities of interest in the text in order to produce accurate results. For example, attempts to infer statistical associations between genes or other entities based on their co-occurrence within documents or within sentences can be quite sensitive to both false-positives and false-negatives in term identification (Jenssen *et al.*, 2001; Yandell and Majoros, 2002). For this reason, techniques which rely on a controlled vocabulary or curated lexicon of terms can be severely limited by the quality of that lexicon. A lexicon which is poorly structured, contains ambiguous terms, or is in some other way inadequate can render an otherwise promising algorithm much less useful. These ideas are depicted in Figure 1, which illustrates the dependence of various mining tasks on the *concept indexing* phase of these systems.

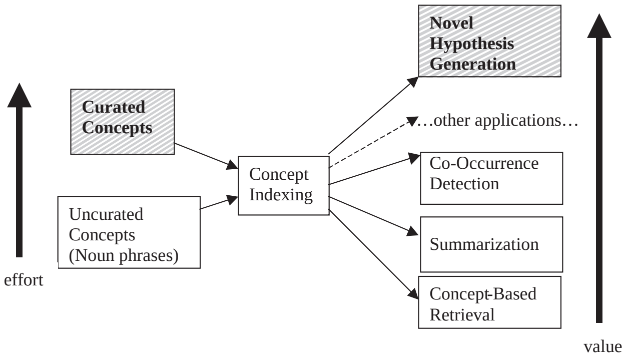

Fig. 1. The central role of concept indexing in biomedical text mining. More difficult (and valuable) tasks and resources are shown higher on the page, with those most elusive goals shown with shading.

Although many specialized lexica are available, we are not aware of a lexicon which is both comprehensive

[*]To whom correspondence should be addressed.

and ideally suited for concept indexing in biomedicine. For example, the UMLS Metathesaurus (Lindberg *et al.*, 1993), one of the best known sources of controlled vocabulary for medicine, is a seemingly comprehensive terminology resource in some technical areas, though shown to be inadequate due to redundancy, omissions, homonyms, acronyms, abbreviations, elisions, proper names, and spelling errors (Nadkarni *et al.*, 2001). For this and other reasons, various groups have resorted to utilizing *uncurated terminology*, or phrases extracted dynamically from literature, sometimes in combination with existing curated terminology (Yoshida *et al.*, 2000; Thomas *et al.*, 2000; Rindflesch *et al.*, 2000; Ng and Wong, 1999; Sekimizu *et al.*, 1998; Fukuda *et al.*, 1998; Lindberg *et al.*, 1993). We believe that the most useful types of dynamically extracted phrases are noun phrases, which (not coincidentally) appear to account for the vast majority of specialized terminology in biomedical text.

To gain an appreciation of the richness of specialized terminology represented in the noun phrases in MEDLINE, it is instructive to note that while UMLS contains $\sim$1.7 million concept phrases, it is possible to identify $\sim$7.4 million distinct core noun phrases in MEDLINE. Upon inspection, many of these appear to represent useful biomedical concepts, worthy of use in various mining efforts. In fact, we have found them to be very useful for various forms of co-occurrence analysis, hypothesis generation, and summarization tasks (unpublished data). Uncurated concepts are thus of demonstrable value, notwithstanding the obvious drawbacks stemming from their lack of curation and placement within a standard ontology.

Another motivation for the use of noun phrases as surrogates for curated concepts derives from the continual emergence of new gene and protein names in published literature (Friedman *et al.*, 2001; Ng and Wong, 1999). Simply tagging them as nouns should allow most of these new gene and protein names to be reliably extracted. Indeed, in the absence of strong suffix cues, modern part-of-speech taggers tend to tag unknown words as nouns (Manning and Schütze, 2000; Brill, 1992; Brill and Marcus, 1992; Cutting *et al.*, 1992), which will generally favor the accurate identification of noun phrases, since specialized words often belong in noun phrases.

Reliable extraction of useful noun phrases from biomedical text is not yet a perfected art, however. The main difficulty stems from the lack of a highly accurate part-of-speech tagger for biomedical text. Existing taggers are generally pre-trained on relatively generic text, such as the Brown Corpus (Francis and Kučera, 1982) or the SUSANNE Corpus (Sampson, 1994), neither of which are primarily medical in content. Hence, such a tagger would be confronted with a large number of unfamiliar words when applied to a corpus such as MEDLINE, resulting in a reduction of tagging accuracy (Manning and Schütze, 2000). Although many taggers provide retraining features, such retraining invariably requires a large sample of manually (or semi-manually) tagged training sentences (Brill, 1992; Cutting *et al.*, 1992).

Some taggers attempt to make an ‘educated’ guess at the part-of-speech for an unfamiliar word based on the word ending (Manning and Schütze, 2000; Charniak *et al.*, 1993; Brill, 1992; Brill and Marcus, 1992; Cutting *et al.*, 1992). For example, an -ed or -ing ending is often taken as an indication that the word is a verb. However, many such verbs encountered in biomedical text are actually behaving as participles. These can be effectively treated as adjectives for the purpose of noun phrase extraction (e.g. *striated muscle*, or *nictitating membrane*). However, not all -ed and -ing verbs act as adjectives (even those identified by a tagger as participles). Thus, guesses based on word morphology are often incorrect.

Finally, phrases extracted from MEDLINE using the simple heuristic of finding continuous runs of adjectives and nouns are often only a subphrase of what most people would agree is the more desirable, complete phrase (for example, *immunodeficiency syndrome* versus *acquired immunodeficiency syndrome*; *cell* versus *nucleated cell*), and thus extracted noun phrases often lack the specificity of a genuinely informative biomedical concept. Although recent work on ‘chunking’ methods to find noun phrases has produced encouraging results, virtually all of that work has been directed toward nonmedical corpora such as the Wall Street Journal (Ramshaw and Marcus, 1998; Pla *et al.*, 2000; Zhou and Su, 2000; Sang *et al.*, 2000; Veenstra and Buchholz, 1998; Zhou *et al.*, 2000), and many of these systems assume that part-of-speech tags have already been correctly assigned before chunking is carried out.

Thus, an important and as yet unsolved problem is how to readily obtain a part-of-speech tagger specifically geared toward biomedical text, so that high-quality noun phrases can be extracted for use in concept indexing and other downstream mining tasks.

We investigated whether the retraining of a tagger for MEDLINE could be automated by incorporating existing sets of curated phrases into the training process in a well-defined and principled way. Our hypothesis was that beginning with a corpus of text tagged by a naïve tagger and constraining the training process to respect known phrases, a less naïve tagger would be obtained which would be able to identify not just the curated phrases provided during training, but also other phrases having similar phrase structure. Such an approach would be more feasible than traditional retraining practices, because it eliminates the need to manually tag large sets of training sentences, benefiting instead from existing lists of curated terminology.

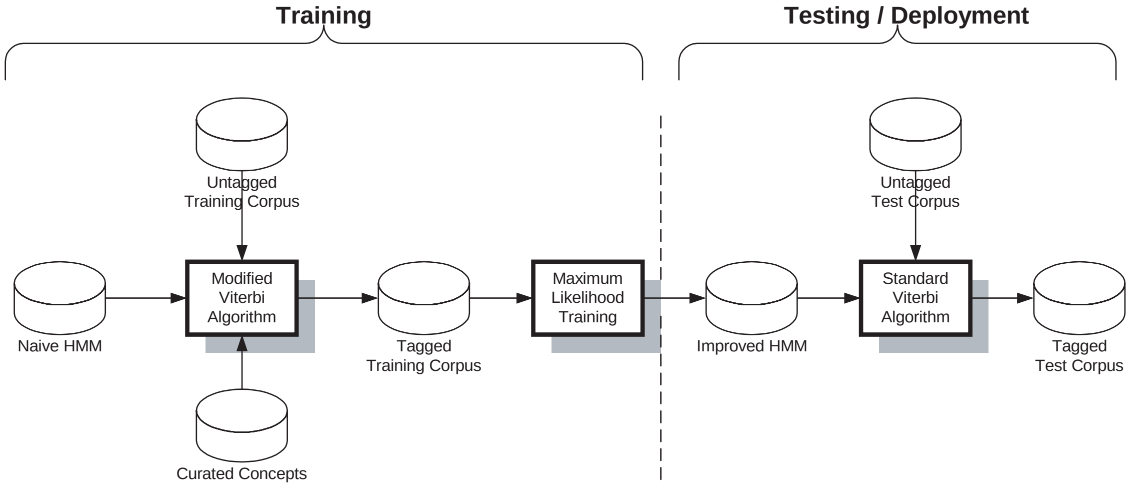

Fig. 2. The process used to retrain an HMM tagger by incorporating prior knowledge in the form of curated concepts.

METHODS

The use of Markov models for part-of-speech tagging is common practice (Manning and Schütze, 2000; Charniak *et al.*, 1993; Cutting *et al.*, 1992), and is perhaps the simplest of the proven methods for tagging. Our approach included modifying a basic Markov model tagger with states corresponding to part-of-speech tags and an alphabet of symbols corresponding to individual words. This process is depicted schematically in Figure 2.

We used a curated lexicon of phrases extracted from UMLS (including genes, diseases, post-translational modifications, molecular functions, biological processes, and anatomical terms) to modify the HMM probabilities so that words comprising those phrases are more likely to be tagged as adjectives and nouns, not only when they occur in those phrases, but also when they occur in novel phrases not included in the lexicon.

However, we wished to avoid simplistic tuning of the parameters in an unprincipled way. In particular, our goal was to modify the emission and transmission probabilities in tandem, and in a way which allowed the surrounding context of the curated phrases to influence the precise tagging of the individual words in the phrase. We posited that this would eliminate unintended side-effects that could reduce overall tagging accuracy.

Thus, we have devised a process which employs two versions of the Viterbi algorithm—a standard, unmodified version for final deployment, and a modified version which is used only during training to force curated concepts to be tagged as noun phrases, thereby modifying the tag frequencies in the tagged training corpus. This tagged training corpus is then used to train the final, improved HMM for deployment. The improved HMM already incorporates to some degree implicit knowledge of the curated concepts and their constituent words in its transition and emission probabilities, so the final tagger can utilize the standard Viterbi algorithm, and need not refer directly to the list of curated concepts.

The principal modification to the Viterbi algorithm is the inclusion of a term $\delta(i, \pi_i)$ which 'zeros-out' the probability of any complete-sentence tag assignment which would assign an undesirable tag to any word participating in a known concept within the sentence:

$$\pi^* = \arg\max_{\pi} \left[\left(\prod_{i=1}^{L} P(x_i|\pi_i)\delta(i,\pi_i) \right) \times P_{\text{start}}(\pi_1) P_{\text{stop}}(\pi_L) \prod_{i=1}^{L-1} P(\pi_{i+1}|\pi_i) \right]$$

$$\delta(i,\pi_i) = \begin{cases} 0 & \text{if } i \in \text{ phrase } \wedge \pi_i \notin \{\text{NOUN, ADJ}\} \\ 1 & \text{otherwise} \end{cases}$$

where $P(x_i \mid \pi_i)$ is the probability of state π_i emitting word x_i, $P(\pi_{i+1} \mid \pi_i)$ is the transition probability from state π_i to state π_{i+1}, and $P_{\text{start}}(\pi_i)$ and $P_{\text{stop}}(\pi_i)$ are the probabilities of starting and stopping in state π_i, respectively. The final tag assignment is given by $\pi^* = (\pi_1, \pi_2, \ldots, \pi_L)$.

In practice, this modification corresponds to 'masking' in the dynamic programming matrix any cell which denotes the assignment of any tag other than NOUN or ADJ (adjective) to a column occupied by a curated concept, as illustrated in Figure 3.

Once masking has been performed, we proceed according to the normal operation of the Viterbi algorithm, choosing the most likely path through the dynamic programming matrix and assigning tags accordingly. The

			phrase m							phrase n		
	w_0	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	.
NOUN												
ADJ												
ADVERB			0	0	0					0	0	
PREP			0	0	0					0	0	
ARTICLE			0	0	0					0	0	
VERB			0	0	0					0	0	
CONJ			0	0	0					0	0	

Fig. 3. Masking in the dynamic programming matrix. Cells which correspond to assignment of a non-noun/non-adjective tag to a word occurring within a curated concept are set to zero probability.

effect of masking is to preclude any path which passes through a masked cell from being chosen, and hence from any word in a concept from being assigned an undesirable tag.

A full specification of the dynamic programming algorithm is given at http://www.tigr.org/imhotep.

Unfortunately, masking can lead to situations where all paths have zero probability, as would happen, for example, if a word in a concept occurrence is only known by the naïve tagger to have one of the undesired tags. In order to ensure that at least one nonzero-probability path passes through a given column, we adjust the local model parameters as necessary to produce nonzero entries in the nonmasked region of the column. We do this as follows.

Each cell is normally defined as the product of three quantities, $v_{k,i-1}$, the value of a cell in the previous column, $P(q_j|q_k)$, the transition probability from that cell to this cell, and $P(x_i|q_j)$, the emission probability for this cell. A zero can result if any of these three terms are zero. We prefer to attribute the lack of nonzero cells in this column to erroneous knowledge about the probability distribution over tags for this word, rather than to incorrect knowledge about transition probabilities in medical text. Therefore, if any products $v_{k,i-1}P(q_j|q_k)$ are nonzero, we obtain a nonzero product by simply assuming a uniform distribution over tags for a given word; i.e. $P(q_j|x_i) = 1/|Q|$ for $|Q|$ the number of states in the HMM. From this estimate of $P(q_j|x_i)$ we obtain its Bayesian inverse, $P(x_i|q_j) = P(x_i)/(P(q_j) \quad |Q|)$, for the final term in the product defining this cell. However, if all products $v_{k,i-1}P(q_j|q_k)$ are zero, we resort to using a uniform distribution for transitions; i.e. $P(q_j|q_k) = 1/|Q|$. By induction, we can assume that at least one $v_{k,i-1}$ is nonzero, so we need only attend to the transition and emission probabilities, as we have done.

For a training set we used ∼30 000 MEDLINE documents published since the late 1990s. This corpus contained a total of ∼12 million words drawn from a vocabulary of ∼120 000. Our curated lexicon consisted of ∼246 000 concepts selected from UMLS, and included such entities as genes, diseases, post-translational modifications, protein domains, drugs, tissues, immunology terms, organisms, biological processes, and molecular functions.

Tagged sentences were scanned for noun phrases by finding multiword sequences consisting of at least one noun and an arbitrary number of additional nouns and adjectives. Phrases found by this simple heuristic tend to be good approximations for core, nonrecursive noun phrases. We chose not to include prepositions and conjunctions because doing so often results in ‘run-on’ phrases that are less useful for concept indexing. Nevertheless, we acknowledge that many useful phrases (e.g. ‘cirrhosis of the liver’) are missed in this way, and a that a more sophisticated approach is called for (perhaps by observing the pointwise mutual information between the sub-phrases surrounding the preposition).

In order to estimate the magnitude of the improvement achieved through retraining, we took the first 10 000 noun phrases (approximately 2200 sentences in 270 documents) that were extracted and manually scored the differences between the old and new taggers by examining the questionable phrases in context. Extracted phrases were scored by penalizing a tagger for omitting words that actually belonged in a phrase and for including words which did not belong. Differences between phrase predictions of the two taggers were then classified as either beneficial or detrimental, based on the difference in phrase score.

RESULTS

Of the 108 nontrivial differences found between the old and new taggers, 91 (84%) were judged to be beneficial changes, and 17 (16%) were judged to be detrimental. Although the differences appear to affect only a small minority of all noun phrases (108 out of 10 000 or roughly 1%), it is important to realize that many noun phrases encountered in text are not highly informative terms, or are not specific to biomedicine, and these are of no interest to us. Rather, our method specifically targets multi-word biomedical terms, and it is clear that our changes to the tagger did in fact enhance its ability to extract these. Of the new terms found by the improved tagger, 96% were judged to be highly relevant to biomedicine, and of these, 86% were not already present in our lexica.

Examples of phrases which were improved by the new tagger are shown in Table 1. The improved phrases were often more specific than those produced by the naïve tagger, and therefore more suitable for advanced data

Table 1. Example phrases produced by the original tagger (before retraining) and the new tagger (after retraining). Additional examples can be found at http://www.tigr.org/imhotep

Before retraining	After retraining
Immunodeficiency syndrome	Acquired immunodeficiency syndrome
Tubulin	Tubulin folding intermediates
Erythrocytes	Fetal nucleated erythrocytes
Collagen	Collagen folding diseases
Receptor gene	Human androgen receptor gene
Axons	Mature myelinated axons
Transport	Electrolyte transport
Withdrawal	Androgen withdrawal
Differentiation	Adipocyte differentiation
Cells	Positive multinucleated cells

mining activities. Furthermore, the erroneous differences were very often due to the attachment of an article or preposition to the beginning or end of an otherwise valid phrase, and it is tempting to speculate that these types of simple errors might be rectified by very minor modifications to our algorithm, though we have not investigated this possibility.

To see that the tagger has actually generalized the knowledge provided in the curated concepts, note that a novel discovered phrase, ‘dentate granule cells,’ does not occur in our curated lexicon, although ‘dentate cerebellar ataxia,’ ‘specific granule deficiency,’ and ‘tumor cells’ do. Indeed, even ‘granule cells’ does not occur in the lexicon. Thus, the tagger is finding additional noun phrases which were not explicitly provided during training.

DISCUSSION

The majority of improvements to noun phrases made by our heuristic appear to involve the retagging of certain verbs as adjectives when in the vicinity of the other parts of a noun phrase. However, our method tends to restrict retagging to only those verbs which have been observed acting as adjectives in the curated lexicon. Furthermore, our simple modification to the Viterbi algorithm attends not only to participles, but also to other parts of speech that only occasionally participate in core noun phrases, such as the article ‘A’ in ‘blood group A.’ For this reason, we feel that our method is more elegant than the various ad-hoc approaches that attempt to ‘patch up’ the noun phrase after part-of-speech tagging.

A somewhat surprising feature of our method is that it has the ability not only to include in noun phrases words which had previously been excluded, but to sometimes exclude those that the naïve tagger would have (erroneously) included. An example is ‘Bacillus thuringiensis exhibits’ which was truncated by the new method to ‘Bacillus thuringiensis.’ Whereas ‘exhibits’ is often used as a verb in scientific text, in less specialized discourse it is often used as a noun, as in ‘The museum has many fine exhibits.’ Naïvely, one might reason that a word not occurring in the curated lexicon would have an unchanged tag distribution and therefore behave as it previously had, but in fact the emission probabilities of the HMM states for such words are affected by the addition of new words to the word class, due to the rescaling of probabilities (which must still sum to 1), and this can apparently reduce the emission probability of other words enough (under the right circumstances) to change their assigned tag. Thus, the example phrases might be said to provide not only ‘positive reinforcement,’ but also ‘negative reinforcement’ through the absence of certain words in those phrases.

Although our method has significantly improved the quality of the extracted phrases, additional improvement is necessary. For example, our tagger still has difficulty with the word ‘in’ in phrases such as ‘in situ’ and ‘in vivo.’ We expect many of these remaining errors to disappear with the use of better and more varied training lexica. Additional work is also necessary to develop methods for accurately extracting larger, recursive noun phrases (i.e. including prepositions and conjunctions).

There are other types of prior knowledge that may be incorporated into the tagger, such as would result by ‘subtracting’ a medical corpus from a non-medical corpus and then down-weighting the probability of the resulting terms from forming useful phrases.

An important direction in which our research needs to be extended is in the handling of unfamiliar words, which we precluded by using identical training and test sets. The class of complete-word Markov models that we employed is not readily applicable to the problem of guessing the tag distributions for novel words, though it is conceivable that similar techniques might be applied to portions of words in order to employ morphological cues specific to biomedical terminology in guessing the appropriate tag for a novel word. For the purposes of the current work, we considered the problem of handling unfamiliar words to be an entirely separate line of inquiry, albeit an interesting one, and worthy of attention. Currently, our tagger simply assumes that all novel words are nouns.

In conclusion, we encourage additional efforts to improve the state of biomedical part-of-speech tagging, and hope that a standard, publicly available tagger will soon emerge, so that future work can concentrate on the more important tasks of extracting real medical knowledge from literature.

REFERENCES

Andrade,M. and Valencia,A. (1998) Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, **14**, 600–607.

Blaschke,C., Andrade,M.A., Ouzounis,C. and Valencia,A. (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. In *Proceedings of the AAAI Conference on Intelligent Systems for Molecular Biology*. pp. 60–67.

Brill,E. (1992) A simple rule-based part of speech tagger. In *Proceedings of Third Conference on Applied Natural Language Processing*. pp. 152–155.

Brill,E. and Marcus,M. (1992) Tagging an unfamiliar text with minimal human supervision. In *Probabilistic Approaches to Natural Language*, American Association for Artificial Intelligence, AAAI Press.

Charniak,E., Hendrickson,C., Jacobson,N. and Perkowitz,M. (1993) Equations for part-of-speech tagging. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*. Washington, DC, pp. 784–789.

Cutting,D., Kupiec,J., Pedersen,J. and Sibun,P. (1992) A practical part-of-speech tagger. In *Proceedings of the 3rd Conference on Applied NLP*. Italy, pp. 133–140.

Demetriou,G., Humphreys,K. and Gaizauskas,R. (2000) Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. In *Proceedings of the Pacific Symposium on Biocomputing*. Hawaii.

Francis,W. and Kučera,H. (1982) *Frequency analysis of English usage. Lexicon and grammar*. Houghton Mifflin, Boston, MA.

Friedman,C., Kra,P., Yu,H., Krauthammer,M. and Rzhetsky,A. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17**, S74–S82.

Fukuda,K., Tsunoda,T., Tamura,A. and Takagi,T. (1998) Toward information extraction: identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing '98*.

Humphreys,B.L. and Lindberg,D.A. (1989) Building the unified medical language system. In *Proceedings of the Thirteenth Annual Symposium Computer Applications in MedicalCare*. pp. 475–480.

Jenssen,T.-K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.*, **28**, 21–28.

Lindberg,D.A.B., Humphreys,B.L. and McCray,A.T. (1993) The unified medical language system. *Methods of information in Medicine*, **32**, 281–291.

Manning,D. and Schütze,H. (2000) *Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marcotte,E., Xenarios,I. and Eisenberg,D. (2001) Mining literature for protein–protein interactions. *Bioinformatics*, **17**, 359–363.

Nadkarni,P., Chen,R. and Brandt,C. (2001) UMLS concept indexing for production databases: a feasibility study. *Journal of the American Medical Informatics Association*, **8**, 80–91.

Ng,S.-K. and Wong,M. (1999) Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Informatics*, **10**, 104–112.

Ono,T., Hishigaki,H., Tanigami,A. and Takagi,T. (2001) Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics*, **17**, 155–161.

Pla,F., Molina,A. and Prieto,N. (2000) Tagging and chunking with bigrams. In *Proceedings of COLING-2000*.

Ramshaw,L.A. and Marcus,M.P. (1998) Text chunking using transformation-based learning. In *Natural Language Proceedings on using Very Large Corpora*. Kluwer, Drodrecht.

Rindflesch,T.C., Tanabe,L., Weinstein,J.N. and Hunter,L. (2000) Edgar: Extraction of drugs, genes and relations from the biomedical literature. In *Proceedings of the Pacific Symposium on Biocomputing*. Hawaii.

Sampson,G. (1994) The SUSANNE Corpus. *ICAME J.*, **17**, 125–127.

Sang,E.F.T.K. and Veenstra,J. (1999) Representing text chunks. In *Proceedings of EACL'99*. pp. 173–179.

Sang,E.F.T.K., Daelemans,W., Dejean,H., Koeling,R., Krymolowski,Y., Punyakanok,V. and Roth,D. (2000) Applying system combination to base noun phrase identification. In *Proceedings of COLING-2000*. pp. 857–863.

Sekimizu,T., Park,H.S. and Tsujii,J. (1998) Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. *Genome Inform. Ser. Workshop Genomic Inform*, **9**, 62–71.

Stapley,B.J. and Benoit,G. (2000) Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. In *Proceedings of the Pacific Symposium on Biocomputing, 2000*. pp. 526–537.

Swanson,D.R. and Smalheiser,N.R. (1999) Link analysis of MEDLINE titles as an aid to scientific discovery. *Library Trends*, **48**, 48–59.

Tanabe,L., Scherf,U., Smith,L.H., Lee,J.K., Hunter,L. and Weinstein,J.N. (1999) MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques*, **27**, 1210–1217.

Thomas,J., Milward,D., Ouzounis,C., Pulman,S. and Carroll,M. (2000) Automatic extraction of protein interactions from scientific abstracts. *Proceedings of the Pacific Symposium on Biocomputing*. Hawaii, pp. 541–551.

Veenstra,J. and Buchholz,S. (1998) Fast NP chunking using memory-based learning techniques. In *Proceedings of BENELEARN'98*. pp. 71–78.

Yakushiji,A., Tateisi,Y., Miyao,Y. and Tsujii,J. (2001) Event extraction from biomedical papers using a full parser. *Proc. Pacif. Symp. Biocomput.*, **6**, 408–419.

Yandell,M.D. and Majoros,W.H. (2002) Genomics and natural language processing. *Nat. Rev. Genet.*, **3**, 601–610.

Yoshida,M., Fukuda,K. and Takagi,T. (2000) PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary. *Bioinformatics*, **16**, 169–175.

Zhou,G.-D. and Su,J. (2000) Error-driven HMM-based chunk tagger with context-dependent lexicon. In *Proceedings of EMNLP/VLC-2000*.

Zhou,G.-D., Su,J. and Tey,T.-G. (2000) Hybrid Text Chunking. In *Proceedings of CoNLL'2000*.