Journal of Biomedical Informatics 42 (2009) 912–922

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Using UMLS to construct a generalized hierarchical concept-based dictionary of brain functions for information extraction from the fMRI literature

Mei-Yu Hsiao [a], Chien-Chung Chen [b], Jyh-Horng Chen [a,*]

[a] Interdisciplinary MRI/MRS Laboratory, Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
[b] Department of Psychology, National Taiwan University, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 27 May 2008
Available online 22 April 2009

Keywords:
Neuroinformatics
Term recognition
Term classification
Brain-function association model
Human brain study

ABSTRACT

With a rapid progress in the field, a great many fMRI studies are published every year, to the extent that it is now becoming difficult for researchers to keep up with the literature, since reading papers is extremely time-consuming and labor-intensive. Thus, automatic information extraction has become an important issue. In this study, we used the Unified Medical Language System (UMLS) to construct a hierarchical concept-based dictionary of brain functions. To the best of our knowledge, this is the first generalized dictionary of this kind. We also developed an information extraction system for recognizing, mapping and classifying terms relevant to human brain study. The precision and recall of our system was on a par with that of human experts in term recognition, term mapping and term classification. Our approach presented in this paper presents an alternative to the more laborious, manual entry approach to information extraction.

1. Introduction

Functional magnetic resonance imaging (fMRI), which measures the blood metabolism of neural activities [1], is a non-invasive approach for studying human brain function. Over the past 15 years, fMRI has become extremely popular among neuroscientists. As a result, a profusion of fMRI studies are published every year, in numerous journals. Hence, it is becoming difficult for researchers to keep up with the literature by reading papers, as this activity is extremely time-consuming and labor-intensive. Therefore, automated analysis of unstructured texts using a text mining technique such as that proposed here, may be of great help to scientists.

The main challenge in text mining, especially in the field of neuroscience, is the problem of data entry. Presently, the automatic retrieval and extraction of large amounts of data for analysis is still difficult, and most databases rely on inputting information manually. For instance, the first annotated database for published neuroimaging studies was BrainMap [2,3], the interface of which allowed researchers to search by querying experimental parameters, bibliographical details, or specialized locations in Talairach space [4]. However, the scale of BrainMap was limited; presently, only 61 behavior related terms or function names, and 72 experimental paradigms, are available in the database. Such a limited scale may be inevitable, because BrainMap relies on researchers to input data manually. The lack of automatic processing severely limits the scope of the database and reduces its usefulness.

The goal of the work presented in this paper was to build an automatic information extraction (IE) tool that can extract terms related to brain anatomy, function and experimental tasks from the fMRI literature. Such an approach should allow for the processing of a large amount of text data in a relatively short period and could overcome the shortcomings of the manual entry approach, as discussed above. In addition, to demonstrate the possible applications of this system, we constructed a brain-function co-occurrence association model to assist with identifying terms associated with human brain function.

IE is a method that allows automatic recognition of meaningful words or phrases from unstructured text. A variety of IE methods have been applied to bioinformatics, either in dictionary-based [5] or rule-based approaches [6,7], in applications including detection of disease, protein and gene names [8–11]. IE methods have also been used for identifying relationships between different terms—for example, protein–protein interactions [12–15]. While much progress has been made in applying IE to bioinformatics, unfortunately, it is difficult to apply these achievements directly to a different domain, such as neuroscience, as the textual features subject to IE are highly domain-specific.

Not many studies on IE for the neuroscience literature have been published. Natural language processing (NLP) has not been applied to neuroinformatics until recently [16–20]. One well-known example however is NeuroScholar, which is a knowledge management system for the neuroscience literature that allows

* Corresponding author. Address: Interdisciplinary MRI/MRS Lab, Department of Electrical Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 106, Taiwan. Fax: +886 2 2369 9465.
E-mail address: jhchen@ntu.edu.tw (J.-H. Chen).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.04.003

authors to annotate papers, make notes and share them with others who also use that same platform [21–23]. Presently, NeuroScholar is developing useful NLP tools to make it easier for the user to make annotations efficiently. NeuroScholar and our system may be complementary to each other. Since the NeuroScholar does not provide an association model, whereas our approach does, and it is difficult to extract the relations between different terms in NeuroScholar. Hence, our approach can help authors in NeuroScholar to expedite the annotation of brain functions, experimental tasks and their relations easily throughout the literature.

In a different approach to mining a neuroimaging database, Nielsen et al. [24] report a method that searches for associations between Talairach coordinates and textual descriptions in the abstract using bag-of-words feature vectors. However, this method uses high-frequency unigrams as candidate terms of brain function; thus, it can only identify popular topics and cannot retrieve multi-word names (e.g. "working memory"). An alternative approach to term recognition is machine learning. This approach has only limited use however, since an annotated corpus for supervised learning does not at present exist in neuroscience.

Another issue for automatic information extraction is that of identifying what to extract. One approach is to develop a set of ontologies/terminologies that can work together with the information extraction algorithms. A critical issue is the task of identifying, defining and mapping concept definitions. In an fMRI study, a neuroscientist is most interested in (1) brain responses to a certain experimental task; (2) the areas of the brain in which these responses occur; and (3) the brain functions implied by these responses. That is, the three key concepts in an fMRI paper are brain anatomy, experimental tasks that the participants were asked to perform during the scan and the brain functions that are involved in these tasks. While the anatomy of the brain may be acquired from numerous knowledge sources, such as the Foundational Model of Anatomy [25] and NeuroNames [26,27], no automatic tool yet exists for extracting information regarding brain functions and experimental tasks from the vast literature. Thus, we constructed a generalized brain function dictionary from the Unified Medical Language System (UMLS), which is supported by the National Library of Medicine. UMLS is an important resource for information extraction in the biomedical domain [28,29]. To resolve the problem of synonymous terms, that is, the failure of extracting relevant literature that uses synonyms rather than the exact form of the search word, UMLS used a concept-based Metathesaurus to cluster synonymous terms together. This approach allows users to find the concepts and not just the keywords. Each concept in UMLS Metathesaurus has been assigned one or more semantic types, which are high-level categories in the biomedical domain. However, while UMLS does include concepts for brain functions, each concept involves a number of terms from a number of dictionaries. Hence, the challenge here is to retrieve and merge concepts from various sources.

In this study, we used UMLS for constructing a hierarchical dictionary of brain function terms collected from a variety of vocabulary sources and a hybrid method that combined a dictionary and a rule-based approach for recognizing and classifying concepts related to human brain studies. We present two examples of using the extracted concepts to construct the brain-function co-occurrence association model for studying areas of the brain relevant to memory function, and brain functions relevant to the amygdala. Our approach should provide an improvement to current text mining methods by resolving the data entry problem in the field of neuroinformatics.

2. Methods

Fig. 1 shows the system architecture of our method. This system takes abstracts of papers from a publicly available database, such as Medline, and breaks each abstract into sentences. Then the

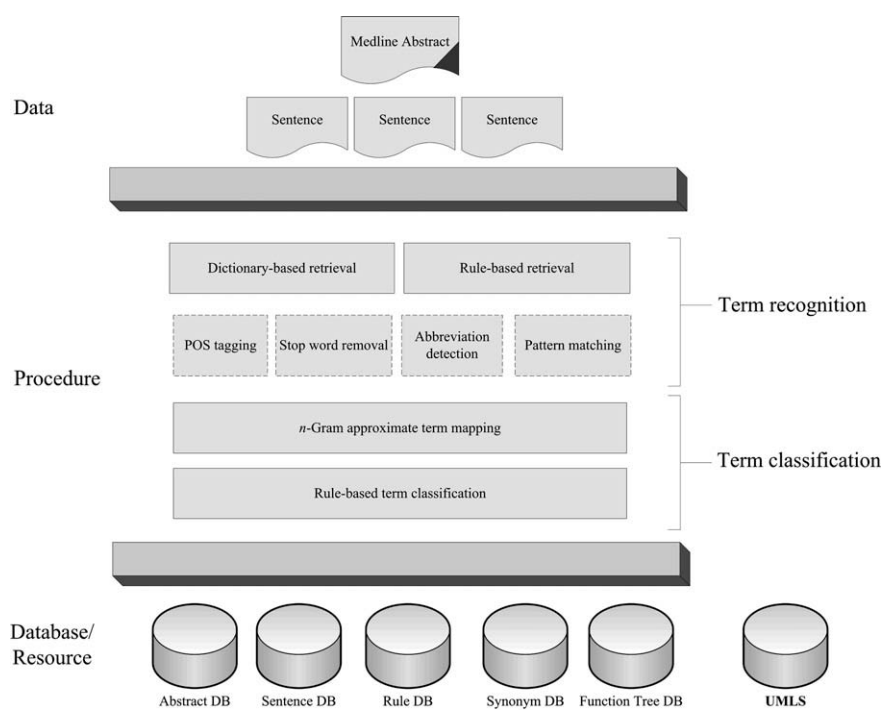

Fig. 1. System architecture of the human brain information extraction system. The data layer splits the Medline abstract into sentence processing units. The procedure layer is the core of the information extraction task. The terms in each sentence were first extracted by term recognition and then classified by term classification stages. In the database layer, synonym and function tree databases were constructed using UMLS and were used to support both term recognition and term classification. See text for detail.

system uses a two-step approach to extract terms. The first step is term recognition. As described in Section 2.2, this combines a dictionary and a rule-based approach to identify words that describe experimental tasks and brain functions. The second step is term classification. As described in Section 2.3, this uses *n*-gram approximate term mapping to identify terms and assign them to categories. Both steps are supported by a dictionary. However, to the best of our knowledge, there is no appropriate, adequately comprehensive resource of brain function vocabulary available. Thus, we constructed a hierarchical concept-based brain function tree to serve as our dictionary, using UMLS. Section 2.1 describes the process of dictionary construction.

2.1. Construction of a hierarchical concept-based dictionary of brain functions

To support term recognition, we first constructed a generalized hierarchical concept-based dictionary of brain functions relevant to the field of neuroscience from UMLS. In this study, we utilized two properties of UMLS to form a hierarchical concept-based brain function dictionary. The first property is that UMLS combines terms with similar meanings into concepts; the second is that terms are arranged with a "parent–child" hierarchy. Thus, our task is to construct the relations among concepts from relations among terms.

We used a command line in UMLS (http://umlsks.nlm.nih.gov/kss/servlet/CommandLine) to map a term to its concept. The command line with parameter "command=ks –meta –cst –def -chd" allowed retrieval of the term's concept (allowing that there could be more than one concept), synonyms, definition, semantic type and child terms included in the different vocabulary sources of UMLS. We traced a complete hierarchical tree starting from the UMLS terms "mental process", "cognitive processes" and "psychological function".

The dictionary construction was a recursive process. It operated as follows: (1) Map a term, say, Term A, to its concept. Initially, Term A would be one of the starting terms mentioned above and would be updated in step 3. (2) If the concept of Term A is already in our database—that is, this concept has been traced before and thus Term A is a synonym of this concept—then add Term A to the synonyms list of its concept and stop tracing Term A. (2′) Otherwise, retrieve relevant synonyms, definitions, semantic and child terms of the concept of Term A. (3) Trace all child terms across various vocabulary sources. This is achieved by updating Term A with its child terms and then repeating steps 1 and 2. (3′) If a child term has been traced before, that is, this child term is also a child of other parent terms and is already in our database, then add this term as a child of Term A and stop tracing this child term.

Notice that the various vocabulary sources in UMLS contain a diversity of terms, since the sources were constructed with different users in mind. Hence, not all the terms emerging from our dictionary construction process were relevant to brain function. For these terms, we used the UMLS semantic type to filter them. Experts in neuroscience were involved in deciding which semantics or concepts were not appropriate for our purpose. After the pruning process, we re-arranged the brain function tree as a hierarchical concept-oriented tree.

2.2. Term recognition

The purpose of term recognition is to extract a target term from the fMRI literature. Our algorithm (Section 2.2.1) utilized the brain function tree (Section 2.1) as the dictionary for term extraction. However, a dictionary constructed from existing sources (UMLS) might not be comprehensive. In order to avoid missing important terms, we also utilized a rule-based approach, which recognized terms surrounded by specific information items, to extract brain functions and experimental tasks (Section 2.2.2). In addition, we discuss the detail of text processing modules shared by both dictionary-based and rule-based algorithms, in Section 2.2.3.

2.2.1. The dictionary-based term recognition system

Our dictionary-based term recognition method used an approximate matching algorithm to improve recognition performance [11]. In addition, to deal with spelling variation, we extended entries of synonyms in the dictionary for a more comprehensive coverage [10,11].

We used the brain function tree described in Section 2.1 as the dictionary source and a part-of-speech (POS) paradigm [30] to detect left and right boundaries of a term in a sentence. The POS-based approximate matching algorithm extracted not only terms that exactly matched a term in the dictionary but also those that partially matched. The advantage of approximate matching was an improvement in efficiency of term mapping, and the collecting of more information about this term [11]. For example, "memory" in the dictionary not only retrieved "memory", but also retrieved "delayed recall memory" from abstracts. In addition, the extracted term "delayed recall memory" was mapped to a concept of "delayed memory", not "memory" (the detailed algorithm of term mapping is described in Section 2.3.1).

For even more comprehensive coverage, we extended entries of synonyms in the dictionary dealing with word order and punctuation. For example, "memory, short-term" was expanded to six entries: "memory, short-term", "memory, short term", "memory short-term", "memory short term", "short-term memory" and "short term memory". It has been shown that this approach can improve term recognition performance over the straightforward approach of looking-up a table [10,11]. In addition, we eliminated some words at the end of entries because they could decrease retrieval precision. For example, "episodic memory, function" was modified to "episodic memory". This allowed for a retrieval of more terms than those using only that exact wording. Some highly ambiguous terms, such as "mr", which could be an abbreviation for either "mental retardation" or "magnetic resonance", were removed manually. This prevented the retrieval of many false positive terms, especially in the fMRI literature (e.g. T2-weighted MR).

2.2.2. The rule-based term recognition system

The method of dictionary-based term recognition was used to extract brain function terms from the brain function dictionary. There is, however, no dictionary available for experimental tasks, another important component in an fMRI study. Furthermore, some names for an experimental task are compound words created or modified by authors. One example is the *n*-back task, where the number '*n*' is modified for a particular experimental design, such as a 2-back task. For this reason, a rule-based term extractor was an essential supplement for retrieval of experimental task names.

A rule-based system uses rules to extract named entities surrounded by specific information items [6]. In our approach, a rule consisted of four components: the trigger word, the mapping direction, boundary conditions and maximum word length. A trigger word started the extraction when a parsing algorithm found a match between a term in a sentence and a trigger. The mapping direction would then specify whether an extraction should proceed to the left or to the right from the trigger word in a sentence. The boundary conditions specified the stop criteria of the parsing algorithm. The condition was a compound of symbols, as shown in Fig. 2. The maximum word length specified the maximum length of an extracted term. If the extracted term length was greater than the length limit, the extracted result would be discarded. Each rule

[A, B]	Option, but at least fit one of them. Ex: fit A or B.
{A, B}	Order. Ex: fit A then fit B.
!	Not
*	POS tags (Part of speech)

Fig. 2. Condition pattern symbols. Parameters A and B can be a POS tag or a specific string.

was given a class label for later term classification (Section 2.3). An example of term extraction is shown in Fig. 3.

For the rule-based approach to work appropriately, the key task is to choose proper trigger words. To address this issue, we used mining collocates to find trigger words [9], and extracted new terms from the collected abstracts. We first decomposed sentences into tokens, that is, individual words. We then removed stop words, or words with high-frequency but irrelevant to the IE, such as "the", "that", or "what", and anatomical names. A stemming process was then applied to the remaining tokens. We then calculated the frequency of each token and chose terms from the 10 most frequently used tokens as trigger words. Fig. 4 shows the flowchart of how these trigger words were acquired. Table 1 shows the 10 stem candidate trigger words determined by this method. For instance, "role" was one of the trigger words. In most abstracts, "role" describes a function that a particular brain structure performs. For example, "The corpus callosum plays a "role" in mediating interhemisphere communication". Thus, using "role" as a trigger word allowed for identification of anatomical brain structure function.

Sometimes, a recognition rule may miss important information. For example, suppose that we have two sentences: (1) "*Functional imaging has consistently shown that attention-related areas of medial frontal and posterior parietal cortices are active during the attentional conflict induced by color naming in the presence of distracting words (Stroop task)*" and (2) "*Performance of the conventional Stroop specifically activated the anterior cingulate, insula, premotor and inferior frontal regions*". In Sentence (1), we can retrieve "Stroop" using "task" as a trigger word. However, the rule-based algorithm would fail to extract the same word in Sentence (2) because there were no trigger words in the sentence to start the algorithm. In order to avoid such losses, we implemented a feedback learning module to extract them. In our feedback learning module, once an extracted phrase such as "Stroop" is extracted by the rule-based algorithm and is classified as an experimental task (see Section 2.3 for the classification method), we treat it as an entry in our dictionary and locate the same term in other sentences with the dictionary-based extraction algorithm (Section 2.2.1). In addition, since most cognitive tests take an initial capital letter, to reduce false positives, we choose only terms with a capital letter in the feedback learning module.

2.2.3. The text processing modules

Term recognition in our system included four reusable text processing modules: POS tagging, stop word removal, abbreviation detection and pattern matching. For each input sentence, "POS tagging" was first employed to code the part-of-speech tag of each word and to detect term boundaries. Then, "stop word removal" was used for eliminating unwanted words. The "abbreviation detection" module was applied when a term was followed by parentheses. Note that only simple abbreviations consisting of initial characters were detected. The pattern matching module matched a user-defined pattern which was either a POS tag or a specific string.

Pattern:

class: task, trigger word: task, mapping direction:Left

boundary condition: [*verb, *prep], maximum word length: 6

Sentence:

A novel event-related potential (ERP) elicited by a visuospatial recognition memory task was recorded in 20 patients with temporal lobe epilepsy using depth electrodes sited in the temporal lobes.

The extraction result:

visuospatial recognition memory

Fig. 3. Sample pattern, sentence and extracted result.

Fig. 4. The pipeline process for acquiring trigger words.

Table 1
Trigger word candidates: The top 10 stem tokens.

	Stem word	Original term
1	Active	Activate activated activates activation activations active actively activity
2	Associ	Associate associated association associations associative
3	Studi	Studied studies study
4	Role	Role
5	Process	Process processes processing
6	Involv	Involve involved involvement involves involving
7	Function	Function functionality functions
8	Task	Task tasks
9	Suggest	Suggest suggesting suggestive suggests
10	Find	Find finding findings

After pattern matching, we used two filtering strategies to remove unwanted phrases and to increase the efficiency of term recognition. The first one was the POS filter. Since nouns are a major part of the named entities, we used the POS filter to discard extracted phrases that did not contain a noun. To avoid filtering out the truly functional terms that used only an adjective tag, we used POS to tag all entries in the brain function dictionary and put those with only an adjective tag in a list. Subsequently we retained and extracted terms that were in that list. With this strategy, we effectively increased the precision of term recognition. We also applied the same POS method to deal with the issue of conjunctions. If an extracted phrase had a POS pattern like "adjective_1 and adjective_2 noun", such as "spatial and verbal working memory" we could recover phrases like "adjective_1 noun" and "adjective_2 noun", such as "spatial working memory" and "verbal working memory", rather than "adjective_1" and "adjective_2 noun", such as "spatial" and "verbal working memory. This strategy thus can enhance recall by extracting more correct terms.

A second strategy was to filter out the functional designation of a brain area. Brain regions are generally named according to structural or functional significance. For example, the area surrounding the calcarine sulcus is generally called the primary visual cortex. However, a search for the dictionary term "visual" would produce a retrieval error. Most functional designations are compound words, with a function term followed by a noun, such as "area", "cortex", "region", etc. In our study, we discarded terms that were related to functional designation of an anatomical entity. This increased precision in the searches.

2.3. Term classification

Once the text was extracted, the algorithm of *n*-gram approximate term mapping was used for term mapping and term classification. Mapping of extracted phrases to a concept space improved precision of query results. For example, when searching "short-term memory", the system returned data not only in "short-term memory" but also in "immediate memory" and "short memory", because these are terms for the same concept in UMLS. Term classification was a task that assigned terms to categories. In this study, we focused on two categories: the experimental tasks given by the experimenters and the brain functions that are supposed to be revealed by these experimental tasks. The rules of category assignment are listed below:

(1) If a term can be extracted by the dictionary-based method with the brain function dictionary, it is a function.

(2) If a term is extracted by the rule-based method with a "task" class rule, and it can be mapped to the brain function dictionary, it is also a function. Otherwise, it is a task. The reason for this rule is that in the fMRI literature, authors often use the word "task" to describe an experiment that studies a particular brain function but not what the participants were performing. For example, "working memory task" is a task to study working memory function while in contrast, "1-back task" describes what the participants needed to perform in an experiment. Hence, "face perception" can be extracted from the sentence "*Autistic people fail to activate the 'fusiform face area' during face perception tasks*" with a rule of "task" class and can be mapped to a brain function concept and therefore it is a function. On the other hand, "Stroop" from "*Functional imaging has consistently shown that attention-related areas of medial frontal and posterior parietal cortices are active during the attentional conflict induced by color naming in the presence of distracting words* (*Stroop task*)", cannot be mapped to a function in our brain function dictionary and therefore it is a task.

(3) However, if a term extracted by the rule-based method is capitalized and can be mapped to a brain function concept, then this term is a proper noun and belongs to both the task and function categories (e.g. "Continuous Visual Memory Test" is a task but also maps to the function "visual memory").

2.3.1. n-Gram approximate term mapping

It is reported that the *n*-gram approach can successfully detect medical compound words [31]. We generated *n*-gram candidate terms and used rules to determine the best concept. The *n*-gram algorithm takes a small group of words and produces compounds from combinations of these words. Then, these compounds, rather than words in their original order, are used for term mapping. This algorithm has the advantage both of reducing the noisy string problem and of increasing the string matching rate. Here, the term "*n*-gram" refers to "*n*" number of words in a phrase. We assumed that the longest matching *n*-gram (maximum *n*-gram) covered the best definition of the extracted phrase. We set the upper limit for "*n*" to six for efficiency. The algorithm of *n*-gram approximate matching is described below:

(1) If the extracted phrase exactly matches a concept, retrieve this concept and stop. Otherwise, go to step 2.
(2) Perform *n*-gram decomposition to produce candidate terms t_1–t_n.
(3) Match all decomposed terms to concepts c_1–c_n in the brain function dictionary.
(4) For each concept c_1–c_n, check whether c_j is a child of c_k, where $j = 1, 2, \ldots, n$; $k = 1, 2, \ldots, n$, in the brain function hierarchy tree. If it is, remove the parent concept c_k from the candidate list. Repeat this step recursively until there is no relation between candidates.
(5) For each remaining concept, calculate the *n*-gram degree of the concept as the number of words of the original decomposed terms in step 2.
(6) If there are concepts with the same degree, go to step 7. Otherwise return the concept with the largest degree as the best concept.
(7) If there are concepts with the same *n*-gram degree, return the one mapped to the term located closest to the end of the extracted phrase.

In our generalized hierarchical concept-based dictionary of brain functions, a child term usually referred to as a narrower concept than the parent term; for example, visuospatial memory (child term) is a specific kind of memory (parent term). We used this heuristic information to map terms to narrower concepts and therefore chose the child term, rather than the parent term, as the result of term mapping in step 4.

Fig. 5 shows an example, with the *n*-gram decomposition of the extracted phrase "visuospatial recognition memory". Only "visuospatial memory" ($n = 2$), "recognition" ($n = 1$) and "memory" ($n = 1$) were mapped exactly in the brain function dictionary. After step 4, we removed the broader concept "memory" because "memory" is a parent of "visuospatial memory" in the hierarchical tree of brain function. In step 6, we mapped the phrase "visuospatial recognition memory" as the concept "visuospatial memory".

Phrase:
visuospatial recognition memory
n=2:
visuospatial recognition, visuospatial memory, recognition memory
n=1:
visuospatial, recognition, memory

Fig. 5. Sample phrase and the *n*-gram decomposition result.

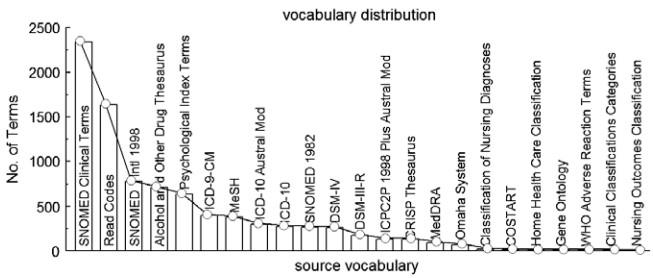

Fig. 6. Distribution of source vocabulary in the brain function dictionary.

In order to improve the efficiency of term mapping, three more steps were used. First, if a term was a single word and capitalized, we checked the abbreviation list retrieved by the abbreviation extraction module described in Section 2.2.3. We retrieved and replaced the abbreviated term with its long form. For instance, WM was replaced by *working memory*, TOL was replaced by *Tower of London*. Second, we established rules for transferring terms to proper nouns or restoring the singular form (e.g. a term with the suffix -ies was replaced with -y). Third, to decrease ambiguity of term mapping to two concepts, we chose the semantic that belonged to a "mental process" in the first instance. Alternatively, we chose the concept with the same word. If these conditions were unavailable, we made a random choice.

Table 2
Semantic types in the brain function dictionary.

Semantic	No. of concepts
Mental or behavioral dysfunction	2039
Mental process	672
Language	642
Sign or symptom	208
Individual behavior	112
Pathologic function	39
Organ or tissue function	21
Temporal concept	11
Physiologic function	11
Research activity	6
Spatial concept	3
Neoplastic process	3
Fungus	1

3. Results and evaluation

3.1. Representation of the brain function tree (brain function dictionary)

The brain function dictionary was constructed using UMLS. We began with 7429 terms in 5566 concepts. After pruning, we ended up with 3720 concepts marked as functions in 13 semantic types from 23 vocabulary sources (Fig. 6). In the synonym list, 9546 out of 9702 synonyms were each mapped to only one concept. This reduced ambiguity while achieving optimal concept selection. As shown in Table 2, most concepts belonged to two semantics: "mental or behavior dysfunction" or "mental process". It is also possible to split the brain function tree into two categories: brain function (positive relation) and dysfunction (negative relation). Fig. 7 shows a subset of brain function terms combined with the multiple source view and the concept view.

3.2. Evaluation

Evaluating our system performance was not an easy task because, to the best of our knowledge, there is no well-established test for fMRI IE. Thus, we conducted two experiments to evaluate the performance of our algorithm against manual entry approach. We used "(fMRI + human) not animal" as our keywords to retrieve abstracts from PubMed. The basic processing unit for term recognition was a sentence. Then, five cognitive psychologists annotated the answers in each experiment. The majority-voting scheme was used to decide the correct answer.

The evaluation of the algorithm was based on three standard information retrieval metrics—precision, recall and *F*-score, as defined below:

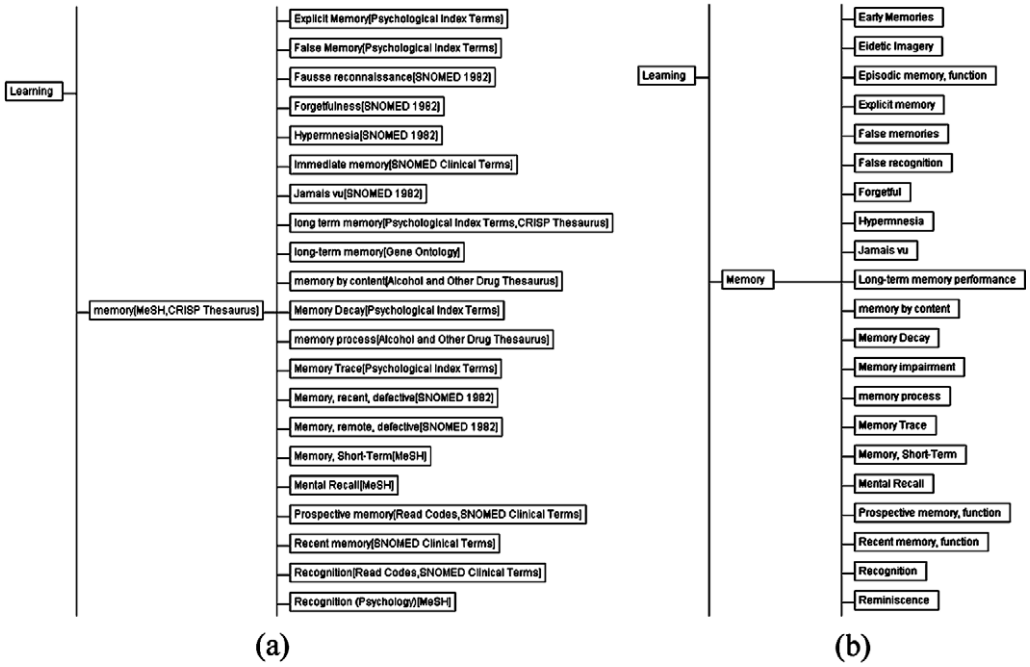

Fig. 7. A subset of the brain function dictionary. (a) Terms combined with the multiple source view. (b) Terms combined with the concept view.

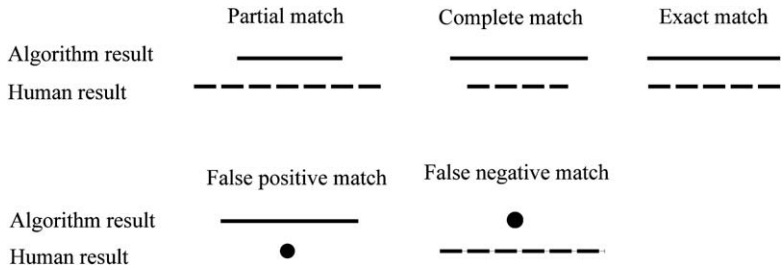

Fig. 8. Five possible types of term recognition for evaluating system and human recognition ability.

Precision: No. of correctly extracted text/no. of extracted text = TP/(TP + FP),

Recall: No. of correctly extracted text/no. of correct text = TP/(TP + FN), and

F-score: 2 ∗ (precision ∗ recall)/(precision + recall),

where TP, FP and FN are the numbers of true positives, false positives and false negatives, respectively.

3.2.1. Evaluation of term recognition

To evaluate our extraction method, we randomly selected 100 sentences from our database. Eighty of these sentences contained the named entities extracted by our algorithm (to test hit rate and false negative) while 20 were from sentences without the extracted named entities (to test false positive). Five cognitive psychologists (two post docs and three 5th or 6th year graduate students with majors in cognitive psychology from the Department of Psychology, National Taiwan University) manually annotated the 100 sentences for terms relating to experimental tasks or function/dysfunction. Correct answers were established by a majority vote. For each term, there were five types of recognition (Fig. 8):

(1) *Partial match:* only part of a text's constitutive words were matched, regardless of how many constitutive words were matched in that text.
(2) *Complete match:* all constitutive words of a text were matched.
(3) *Exact match:* the text from system extraction and the correct answer were exactly matched.
(4) *False positive match:* text extracted by the system, but not by humans.
(5) *False negative match:* system misses an extraction.

Table 3 summarizes the results. For an exact match, there was 58% precision, 59% recall, and an *F*-score of 58%. The 10 partial match terms had the same meaning as human results after term mapping. Including partial and complete matches, we obtained an approximate match of precision, recall and *F*-score up to 72%, 73%, and 72%, respectively. The false positives were mostly from terms with high-frequency and multiple meanings in the fMRI literature (e.g. "orientation" and "complication") or names referring to a dysfunction that had a wide range of symptoms (e.g. "schizophrenia"). The false negatives were mostly from terms that were either not in the dictionary or context sensitive (e.g. "abstraction" and "categorization").

Table 3
System term recognition results.

	Match type	No. of terms
(a)	Partial match	10
(b)	Complete match	5
(c)	Exact match	61
(d)	False positive match	30
(e)	False negative match	28

Retrieved terms: (a) + (b) + (c) + (d).
Correct answer: (a) + (b) + (c) + (e).
Approximate match: (a) + (b) + (c).

Table 4
Comparison of the five human results and the algorithm performance.

	Precision (%)	Recall (%)	*F*-Score (%)
1	89	91	90
2	67	93	78
3	90	93	91
4	91	88	89
5	94	72	82
Human	86	87	86
Algorithm	72	73	72

Annotation results from the five experts were also compared with the correct answers (Table 4). By bootstrapping analysis [32], the average precision was 86, with confidence intervals of [83, 89] at the 95% confidence level. Bootstrapping involved a permutation of responses across the five experts for each item and then recalculation of precision for each individual. This procedure was repeated 10,000 times to derive a distribution of human performance. The level of human performance could be considered as a reference for evaluating the performance of algorithms. Note that even experts in the field did not reach 100% precision.

Table 5 shows results from term classification of 76 approximate extractions. We achieved 93% in category assignment in the task class, suggesting that term classification performs well in task category assignment (row 1). It should be noted that 13 terms classified as functions by our system were classified as tasks by experts. Upon closer examination, nine (69%) of these terms were annotated to both task and function by two of the five experts. This indicated that these terms might belong to multiple categories for experts.

3.2.2. Evaluation of term mapping

In order to evaluate term mapping, we randomly selected 100 sentences containing named entities extracted by our algorithm that were relevant to memory concepts. Five cognitive psychologists (one post doc and two graduate students from the previous experiment and two 5th year graduate students from a biopsychology of learning and memory laboratory) who are experts in

Table 5
Term classification result.

System result	Expert answer		
	Task/test	Function/dysfunction	Both
Task/test	28	0	2
Function/dysfunction	13	25	6
Both	2	0	0

Table 6
Examples of term mapping error.

	Test sentence	System	Expert
(a)	We recently reported that patients who had received unilateral temporal lobectomy, including the amygdala and hippocampus, show impaired acquisition in a fear conditioning task, indicating a deficit in [emotional memory]	Memory	Memory by content
(b)	Most literature on the human hippocampus stresses its [non-spatial memory] functions, but older work in rodents and some other species emphasized the role of the hippocampus in spatial learning and memory as well	Spatial memory	Memory by content
(c)	We hypothesize that activity within these multiple frontal regions provides a functional influence (input) to medical temporal regions that bind the information together into a [lasting episodic memory trace]	Memory trace	Episodic memory, function

The test phrase was marked with square bracket in the test sentence.

memory research, manually mapped every term to 99 concepts (including terms related to function and dysfunction). Although it was possible to assign multiple concepts to each phrase, the experts were forced to choose the best one. On average, each expert took 5 h to complete this experiment.

Applying our *n*-gram approximate term mapping algorithm, we could reach up to 61% consistency with the human experts while the base-line method of simple mapping could achieve only 8% consistency. The 39 mapping inconsistencies between the algorithm and the human experts were of the following categories:

(1) Eighteen inconsistencies occurred because there were no relevant synonyms in the list. In a forced choice situation the expert simply chose what they considered as the best match from the available concepts in the list. (e.g. (1) in Table 6).
(2) Seven were due to an outdated dictionary. For example, "working memory" and "short-term memory" were used interchangeably in the earlier literature but have come to be considered as two distinct concepts in recent years (as reflected in recent cognitive psychology textbooks [33]), but were not considered as separate terms in UMLS. Thus, experts were unable to find the corresponding concept in the list.
(3) Four were context dependent (e.g. (2) in Table 6).
(4) Three (not including the problems mentioned in (2)) were due to the fact that no suitable concept exists in the brain dictionary and thus were misclassified.
(5) Three were considered to be assessments of memory, not a term describing brain function.
(6) Only three were actual algorithm mapping errors (e.g. (3) in Table 6).
(7) One was a word order problem or a missed synonym problem (e.g. "small memory delay" should be mapped to "delayed memory" not "memory").

Six of the 39 terms, while they were mapped inconsistently by the experts and by the algorithm, should be considered an acceptable alternative. This indicates that some terms might have been mapped to multiple concepts. If these terms, and the seven terms which were mapped incorrectly due to the dictionary update problem, are considered, the system's accuracy actually reached 74%.

4. Examples of information usage

This section provides examples of how the extracted information may be used. The most important information to be gained from fMRI experiments is an understanding of brain activities. For this purpose, we gave researchers brain-function co-occurrence association models from various publications. In the brain-function association models, if brain anatomy and function are mentioned together in a large number of MEDLINE sentences, it implies that there is an underlying biological relationship between the two.

We used "(fMRI + human) not animal" as our key words to retrieve abstracts from PubMed MEDLINE. We have 124,450 abstracts (3156 journals) from 1985 to the current period in our database, which updates automatically monthly. Function and task terms were extracted using our method. Brain anatomy was extracted using NeuroNames [26,27], a neuro-anatomical thesaurus in UMLS, and filtered with the Talairach database [4]. Below, we present a model of memory and a model of amygdala.

Fig. 9 presents the results of the brain-function co-occurrence association model for memory. The results showed that a clear and strong relationship existed between memory function and the temporal lobe, the hippocampus and the frontal lobe. This is consistent with previous empirical reports. For example, Kolb and Whishaw [34] showed that these brain regions all play an import role in memory. For each brain region, users could expand the search further to identify more memory related brain functions in these areas. As shown in Fig. 10, the temporal lobe and the hippocampus contribute not only to memory in general but also to short-term memory in particular, and the temporal lobe is also important in many content related memory functions, such as visual memory, spatial memory and semantic memory. The brain-function association model for the amygdala is shown in Fig. 11. Using our hierarchical brain function dictionary to classify function, we were able to show that the amygdala is involved in emotion and learning. This result is consistent with neurobiological studies.

It is anticipated that scientists will be able to better understand and compose a meaningful representation of human brain activities through the integration of data related to brain functions and structural areas. Furthermore, it is hoped that this will help scientists reduce human reading time and increase efficiency significantly, by allowing them to digest specific knowledge before designing experiments and comparing their results with the present literature.

5. Discussion and conclusion

As described in Section 2, we developed a system to extract terms related to experimental tasks and brain functions. We also found that results using POS tagging to retrieve multi-word named entities were better than from use of the unigram method. To the best of our knowledge, this is the first study to extract brain functions and experimental tasks automatically. The generalized hierarchical concept-based dictionary we have developed can be helpful for further studies in text mining, as can algorithms for automatic retrieval of brain functions and their hierarchical relationships for cross-referencing.

Conventional approaches of constructing a dictionary by having experts input information manually have the advantage of achieving a more accurate dictionary but are limited in the number of terms, synonyms, concepts and the relationships between each concept. This task requires a number of experts with knowledge of neuroscience working together for a long time to accomplish. In our dictionary, we merged 23 vocabulary sources to yield a

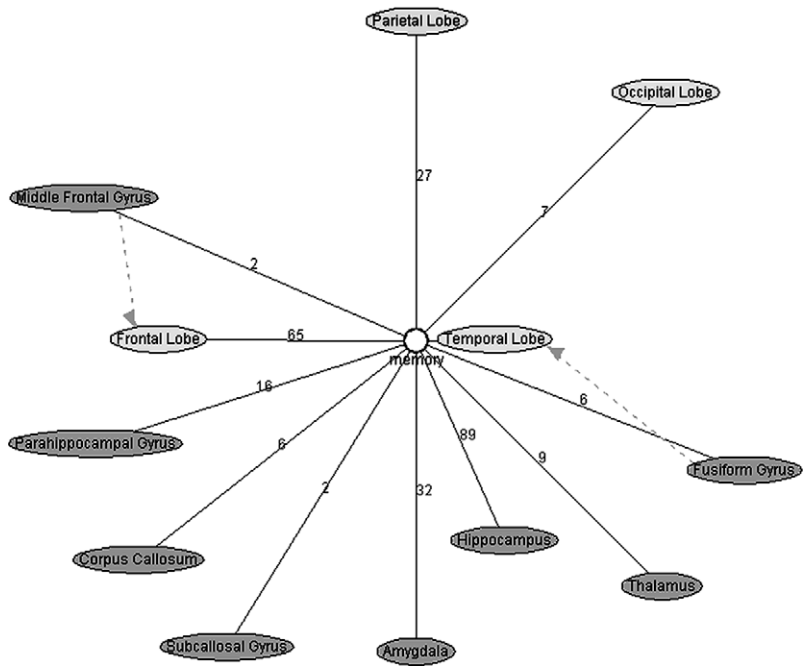

Fig. 9. The brain-function association model of memory with co-occurrence coefficients ⩾2. Co-occurrence coefficients shown on the lines indicate the number of sentences in which brain function and brain area co-occur in the same sentence. The gray arrow lines between the two anatomical nodes indicate the "a-part-of" relation and the nodes to which the arrows point are the parent nodes.

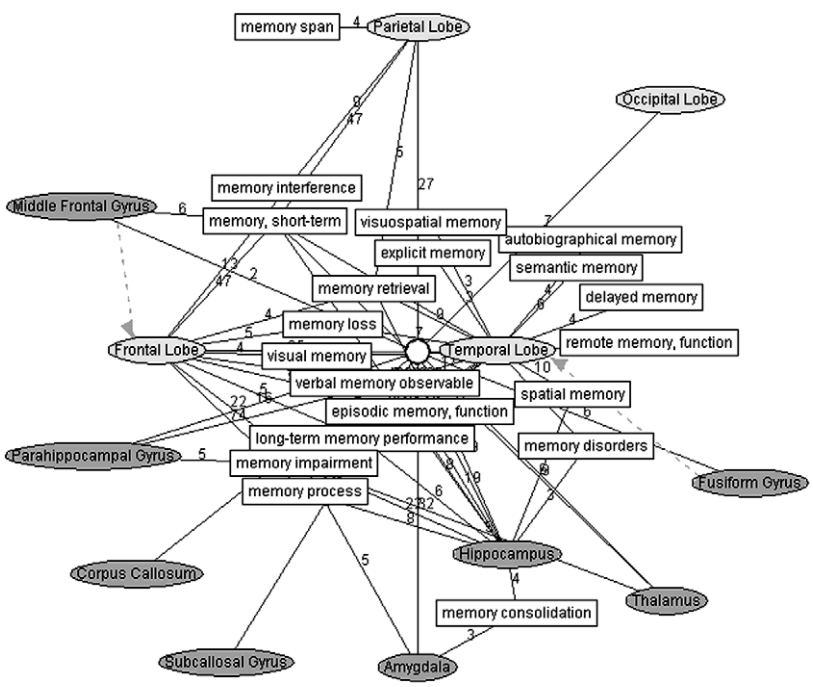

Fig. 10. The expanded memory brain-function association model. The ellipse shape indicates the brain area node and the rectangle shape indicates the extended brain function node.

broader coverage. This allows experts to just insert, delete and reorganize the relationships of brain functions based on large and authentic dictionary sources, such as MeSH, Psychological Index Terms and the like. It is easier and quicker for experts to construct a comprehensive brain function dictionary in this way.

In our term mapping evaluation experiment, each expert took about 5 h to complete the assignment. This illustrates how time-consuming and tedious term mapping is for human experts. Thus, even in a manually constructed knowledge source, our system can help experts to annotate a large number of terms (such as brain functions, brain regions and experimental tasks) and map terms to their concepts in a very short period, and thus overcome the shortcomings of manual entry approaches.

Our system, however, does come with limitations. In the term recognition task, our method was limited by the terms available in the dictionary. Consequently, our system is liable to miss terms

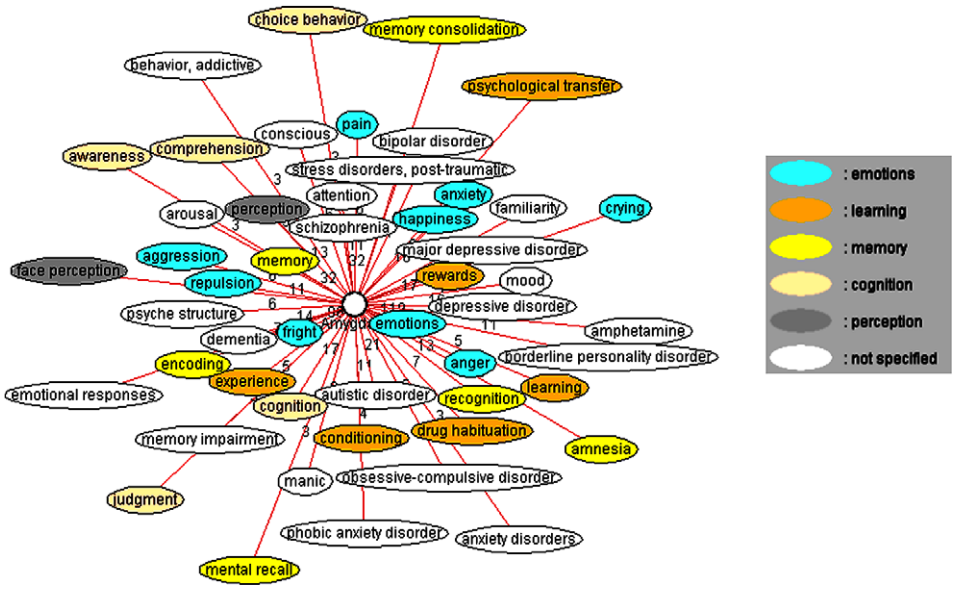

Fig. 11. The brain-function association model of amygdala with co-occurrence coefficients $\geqslant 3$. Legend shows the color of high order concepts in brain function order by the sum of the child nodes co-occurrence number.

that are not in the dictionary. The rule-based algorithm may be able to find terms that are not in the dictionary. However, it is next to impossible to implement all possible rules used by an expert while keeping the efficiency of the system. Hence, rule-based algorithms, even with the help of the feedback-learning algorithm, are not foolproof with respect to any text. In term classification, a function term extracted by the rule-based approach may be classified as an experimental task if this term is not in the brain function dictionary. That is, a function term correctly extracted in the term recognition stage may be misclassified if the system cannot find it in the brain function dictionary.

These limitations showed that the main constraint of our system is the brain function dictionary, which in turn is restricted by the incomplete UMLS. To the best of our knowledge, there is no comprehensive dictionary specific for brain functions. In addition, as new terms for brain function continue to emerge, neuroscientists by no means have consensus on the nomenclature of brain functions. There are databases containing relevant terms. Some of them contain more brain function terms than others. For instance, the SNOMED Clinical Terms contributes 2339 terms while MeSH contributes only 382 terms (Fig. 6). In our dictionary, we merged terms from 23 vocabulary sources to yield a broader coverage than any single source. With such a broad, though incomplete, coverage, our system can achieve 72% precision and 73% recall for term recognition, and this is on a par with human expert performance. Hence, we expect that performance can improve when there are better vocabulary sources available.

The system could be improved by making the dictionary more extensive. The brain function dictionary constructed for this study could also provide a platform from which experts are able to expand, validate and maintain the entries in the dictionary. Furthermore, the phrases extracted by our system could be used as an extension of the available concepts in our brain function dictionary. For instance, "visual working memory" extracted here can be taken as a branch of "working memory" in our brain function dictionary. Such extensions may help us discover more knowledge in the future.

Presently, there is no corpus of annotated research papers available in neuroscience for a large-scale evaluation of the information extraction algorithms. At this stage, a user online feedback system would be an important provision for constructing an annotated corpus by marking errors (such as errors of phrase boundary detection and term mapping) and providing correct answers. Only annotated errors can accelerate the construction of such a corpus. This system could help our approach by adjusting errors and the annotated corpus would be helpful for further study in evaluating the performance of information extraction.

In order to understand brain activities in fMRI study further, it is desirable to expand our algorithm to more categories beyond the three used in this study: brain anatomy, brain functions and experiment tasks, such an expansion may help meta-analysis of published material and in turn benefit experts in this field. Besides, integration of the brain-function association models into fMRI analysis tools may be helpful in comparing experimental results from different studies.

In conclusion, we have constructed a hierarchical concept-based dictionary of brain functions. To the best of our knowledge, this is the first generalized brain function dictionary. We have also presented a two-step approach for term recognition and classification to extract useful information from large-scale fMRI Medline abstracts. Evaluations were performed by comparing the performance of the algorithm to that of human experts. Our system produced promising results. To demonstrate the possible applications of this system, we constructed models for memory and for amygdala with the brain-function co-occurrence association model. While this study has its limitations, it can serve as a basis for further text mining studies in neuroinformatics.

Acknowledgments

We would like to thank Dr. Chung-Ming Chen, Dr. Der-Yow Chen, Dr. Hsin-Hsi Chen and Dr. Keng-Chen Liang for their helpful discussions in this work. The authors thank anonymous reviewers for their constructive comments. This research was supported in part by National Science Council, Taiwan under contracts NSC95-2752-H-002-008-PAE, NSC96-2752-H-002-009-PAE, and NSC97-2321-B-002-044.

References

[1] Ogawa S, Tank DW, Menon R, Ellermann JM, Kim SG, Merkle H, et al. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping

with magnetic resonance imaging. Proc Natl Acad Sci USA 1992;89(13):5951–65.
[2] BrainMap. Available from: http://brainmap.org/.
[3] Laird AR, Lancaster JL, Fox PT. BrainMap: the social evolution of a human brain mapping database. Neuroinformatics 2005;3(1):65–78.
[4] Lancaster JL, Woldorff MG, Parsons LM, Liotti M, Freitas CS, Rainey L, et al. Automated Talairach Atlas labels for functional brain mapping. Hum Brain Mapp 2000;10(3):120–31.
[5] Zhou L, Tao Y, Cimino JJ, Chen ES, Liu H, Lussier YA, et al. Terminology model discovery using natural language processing and visualization techniques. J Biomed Inform 2006;39(6):626–36.
[6] Califf ME, Mooney RJ. Relational learning of pattern-match rules for information extraction. Working notes of AAAI spring symposium on applying machine learning to discourse processing 1998:6–11.
[7] Hu ZZ, Narayanaswamy M, Ravikumar KE, Vijay-Shanker K, Wu CH. Literature mining and database annotation of protein phosphorylation using a rule-based system. Bioinformatics 2005;21(11):2759–65.
[8] Jimeno A, Jimenez-Ruiz E, Lee V, Gaudan S, Berlanga R, Rebholz-Schuhmann D. Assessment of disease named entity recognition on a corpus of annotated sentences. BMC Bioinformatics 2008;9(Suppl. 3):S3.
[9] Hou WJ, Chen HH. Enhancing performance of protein and gene name recognizers with filtering and integration strategies. J Biomed Inform 2004;37(6):448–60.
[10] Schuemie MJ, Mons B, Weeber M, Kors JA. Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification. J Biomed Inform 2007;40(3):316–24.
[11] Tsuruoka Y, Tsujii J. Improving the performance of dictionary-based approaches in protein name recognition. J Biomed Inform 2004;37(6):461–70.
[12] Saric J, Jensen LJ, Ouzounova R, Rojas I, Bork P. Extraction of regulatory gene/protein networks from Medline. Bioinformatics 2006;22(6):645.
[13] Domedel-Puig N, Wernisch L. Applying GIFT, a gene interactions finder in text, to fly literature. Bioinformatics 2005;21(17):3582–3.
[14] Liu Y, Navathe SB, Civera J, Dasigi V, Ram A, Ciliax BJ, et al. Text mining biomedical literature for discovering gene-to-gene relationships: a comparative study of algorithms. IEEE/ACM Trans Comput Biol Bioinform 2005;2(1):62–76.
[15] Rinaldi F, Schneider G, Kaljurand K, Hess M, Andronis C, Konstandi O, et al. Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. Artif Intell Med 2007;39(2):127–36.
[16] Crasto CJ, Marenco LN, Migliore M, Mao B, Nadkarni PM, Miller P, et al. Text mining neuroscience journal articles to populate neuroscience databases. Neuroinformatics 2003;1(3):215–37.
[17] Burns GA, Khan AM, Ghandeharizadeh S, O'Neill MA, Chen YS. Tools and approaches for the construction of knowledge models from the neuroscientific literature. Neuroinformatics 2003;1(1):81–109.
[18] Kotter R. Online retrieval, processing, and visualization of primate connectivity data from the CoCoMac database. Neuroinformatics 2004;2(2):127–44.
[19] Srinivas PR, Gusfield D, Mason O, Gertz M, Hogarth M, Stone J, et al. Neuroanatomical term generation and comparison between two terminologies. Neuroinformatics 2003;1(2):177–92.
[20] Talos IF, Rubin DL, Halle M, Musen M, Kikinis R. A prototype symbolic model of canonical functional neuroanatomy of the motor system. J Biomed Inform 2008;41(2):251–63.
[21] NeuroScholar. Available from: http://www.neuroscholar.org/.
[22] Burns G, Hovy E. The NeuroScholar system as a natural language processing platform for the neuroscience literature. Washington, DC: The Society for Neuroscience Annual Meeting; 2006.
[23] Burns G, Feng D, Ingulfsen T, Hovy E. Infrastructure for annotation-driven information extraction from the primary scientific literature: principles and practice. Services on 2007 IEEE Congress; 2007. p. 122–9.
[24] Nielsen FA, Hansen LK, Balslev D. Mining for associations between text and brain activation in a functional neuroimaging database. Neuroinformatics 2004;2(4):369–80.
[25] Foundational Model of Anatomy. Available from: http://fma.biostr.washington.edu.
[26] NeuroNames. Available from: http://braininfo.rprc.washington.edu/indexabout.html.
[27] Bowden DM, Martin RF. NeuroNames brain hierarchy. Neuroimage 1995;2(1):63–83.
[28] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Database issue):D267–70.
[29] Zhang L, Perl Y, Halper M, Geller J, Hripcsak G. A lexical metaschema for the UMLS semantic network. Artif Intell Med 2005;33(1):41–59.
[30] Brill E. A simple rule-based part of speech tagger. In: Proceedings of the third conference on applied natural language processing. Trento, Italy; 1992.
[31] Heja G, Surjan G. Using *n*-gram method in the decomposition of compound medical diagnoses. Int J Med Inform 2003;70(2–3):229–36.
[32] Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat Sci 1986;1(1):54–77.
[33] Solso RL, MacLin OH, MacLin MK. Cognitive psychology. 8th ed. Boston: Allyn and Bacon; 2007.
[34] Kolb B, Whishaw IQ. Fundamentals of human neuropsychology. 5th ed. New York: Worth Publishers; 2003.