Neuroinformatics

ISSN 1539-2791/03/215–238/$25.00

Original Article

Text Mining Neuroscience Journal Articles to Populate Neuroscience Databases

Chiquito J. Crasto,[*,1,2] *Luis N. Marenco,*[1] *Michele Migliore,*[2,5] *Buqing Mao,*[1] *Prakash M. Nadkarni,*[1] *Perry Miller,*[1,3,4] *and Gordon M. Shepherd*[2]

[1]Center for Medical Informatics, [2]Department of Neurobiology, [3]Department of Anesthesiology, [4]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT; [5]Institute of Biophysics, National Research Council, Palermo, Italy.

Abstract

We have developed a program NeuroText to populate the neuroscience databases in SenseLab (http://senselab.med.yale.edu/senselab) by mining the natural language text of neuroscience articles. NeuroText uses a two-step approach to identify relevant articles. The first step (pre-processing), aimed at 100% sensitivity, identifies abstracts containing database keywords. In the second step, potentially relevant abstracts identified in the first step are processed for specificity dictated by database architecture, and neuroscience, lexical and semantic contexts. NeuroText results were presented to the experts for validation using a dynamically generated interface that also allows expert-validated articles to be automatically deposited into the databases. Of the test set of 912 articles, 735 were rejected at the pre-processing step. For the remaining articles, the accuracy of predicting database-relevant articles was 85%. Twenty-two articles were erroneously identified. NeuroText deferred decisions on 29 articles to the expert. A comparison of NeuroText results versus the experts' analyses revealed that the program failed to correctly identify articles' relevance due to concepts that did not yet exist in the knowledgebase or due to vaguely presented information in the abstracts. NeuroText uses two "evolution" techniques (supervised and unsupervised) that play an important role in the continual improvement of the retrieval results. Software that uses the NeuroText approach can facilitate the creation of curated, special-interest, bibliography databases.

Index Entries: Text mining; natural language processing; neuroscience; databases; supervised and unsupervised learning.

*Address to which all correspondence and reprint requests should be sent. E-mail: chiquito.crasto@yale.edu

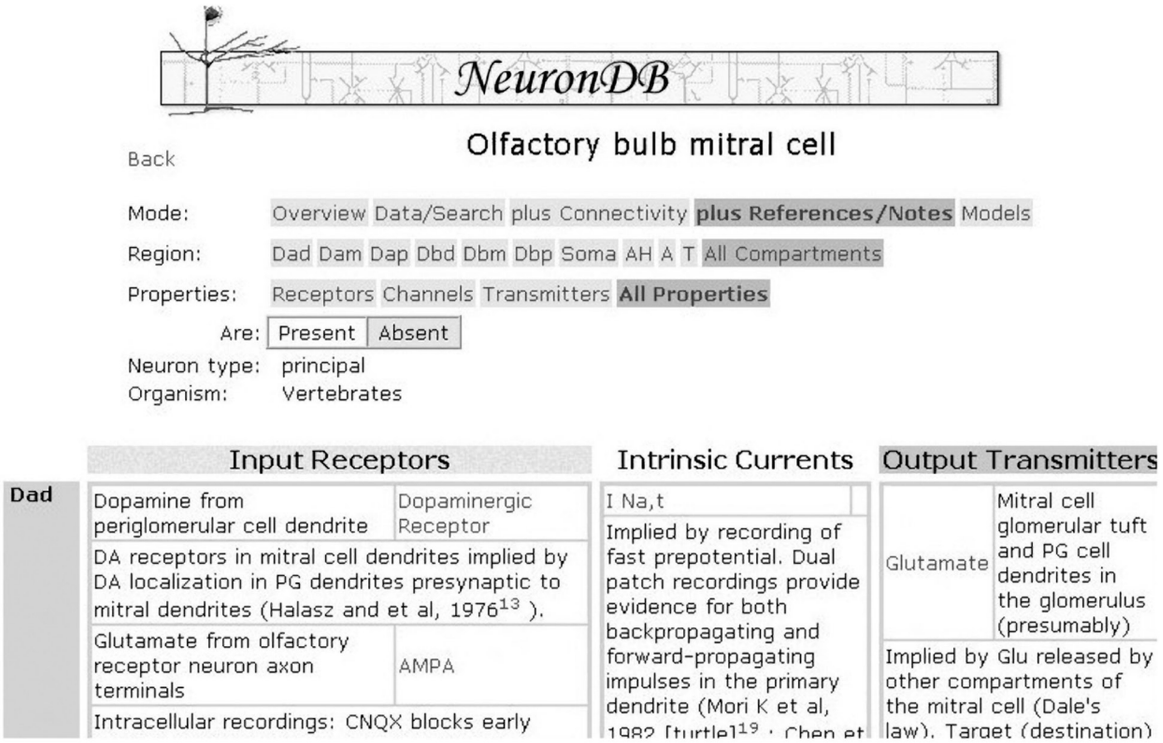

Fig. 1. NeuronDB: Relevant references for properties with expert-supplied annotations related to compartments of the olfactory mitral cells in the olfactory bulb.

Introduction

Neuroscientists studying membrane signaling mechanisms in neurons are faced with rapidly increasing literature. Databases are needed to assist in enabling experimenters to search for and extract data relevant to their particular neuronal system and compare it to those of other systems. To this end, CellProperties Database (CellPropDB) and NeuronDB have been constructed in the SenseLab (http://senselab.med.yale.edu/senselab) databases (Marenco et al., 1999; Shepherd et al., 1998).

Populating these databases has initially been done manually (Migliore et al., 2003), but the rapidly expanding literature requires the development of more automated procedures. This study reports the development of tools that constitute a first step toward this goal.

NeuronDB and CellPropDB provide annotated, bibliographic information related to three essential elements of rapid neuronal signaling: neurotransmitters, neurotransmitter receptors, and intrinsic ion channels in different types of neurons (Fig. 1).

The information in CellPropDB (neuronal properties of a cell as a whole) and NeuronDB (properties that have been experimentally localized to specific neuronal compartments) includes bibliographic citations to articles in the neuroscience literature. Sources for this information include research articles in journals, textbooks, and monographs. The NeuroGuide website (http://neuroguide.com) reports that there are more than 300 online

resources that publish neuroscience-related articles. Each article is a potential CellPropDB and NeuronDB citation.

At present, mining relevant information from such widespread literature sources manually is daunting. An expert can sometimes derive enough information from the title or a few relevant keywords from the abstract to determine whether an article is relevant and should be included in the database, but at other times it requires more extensive reading. The development of automated approaches that offer significant help to the expert is therefore highly desirable. This paper describes such an approach.

Our specific approach to the general problem of mining unstructured text to populate databases involves the following key steps:

- A knowledgebase needs to be established. It should contain specific information that will help identify relevant articles in the correct *hierarchy* (if any). This knowledgebase contains *keywords* and their *synonyms*, and information relevant to the scoring of these keywords.
- The text must also be examined to determine whether the keywords occur in a *context* relevant to the database domain. Key lexical and semantic relationships between keywords to correctly identify *relationships* have to be scanned and identified.
- The lexical scanner should also be able to identify the affirming/negating context of the text. Correctly identifying journal articles that specifically refute the presence of a property is important.
- The program must accommodate the evolution of information in the database domain. The knowledgebase has to be updated (ideally, dynamically) when new keywords, synonyms, and relationships are identified, as well as when existing knowledgebase information becomes irrelevant.
- A successful program should ideally be extensible to other domains without significant algorithmic modifications.
- The program requires close collaboration between the informatics expert and the experimental expert to validate the search results.

We have developed NeuroText to identify potentially relevant articles from neuroscience sources to help populate the SenseLab databases. This paper describes the design of NeuroText and a pilot study of its operation. The *Journal of Neuroscience* (http://www.jneurosci.org), the source for this study, contains full-length articles of cellular and molecular studies, development, plasticity and repair, and behavioral systems, in addition to a small number of short articles—"Rapid Communications." Using the *Journal of Neuroscience* allows our strategy to test different areas of neuroscience while focusing on a single article source.

The development of NeuroText does not address the full natural language understanding problems for general, unstructured text as studied by researchers in the areas of Artificial Intelligence and Computational Linguistics (Baeza-Yates and Ribeiro-Neto, 1999). We process text from neuroscience articles with a focused end-goal in mind: populating a database with specific information about neuronal membrane properties. The nature of this desired information dictates very focused retrieval strategies.

Specific features of NeuroText's approach include the following.

- Keyword counting as an initial basis for potential relevance: NeuroText is premised on the straightforward assumption that researchers presenting information will mention concepts and keywords related to that information more frequently than other non-related or distantly-related concepts. For example, if the research focuses on the **serotonin** receptor expressed in the **Purkinje** cells of the **cerebellum**, elementary counting statistics should be able to help differentiate this from a more tangential mention of some other property

such as "**CA1** pyramidal cells in the **hippocampus**" found in the same journal article.

- Using contextual constraints to refine potential relevance: The *context* of the occurrence of relevant keywords, however, precludes reliance on the mere counting of database keywords. Context is defined by the constraints of the specific database domain (specifically) and of neuroscience (as a whole). CellPropDB and NeuronDB present information (structured in the region-neuron-property hierarchy) for a subset of normal, healthy neurons from in vivo or in vitro studies. An efficient program must therefore eliminate articles related to neurons not currently in the database, work related to diseased cells or neurological disorders (e.g., Alzheimer's and Parkinson's diseases), and work pertaining specifically to in vitro cell cultures. The program also must determine whether the negated concepts are contributory to the publication. As a specific example of a highly domain-specific contextual constraint, the experts (Michele Migliore and Gordon M. Shepherd, hereafter MM and GMS) have determined that bibliographic citations to publications about trophic factors (related to growth or metabolism) such as brain-derived neurotrophic factor (BDNF), whose levels are a determinant in the cause of depression in humans) (Barde et al., 1982) should not be covered in the current database. The program has to recognize, however, when BDNF is merely mentioned as a related study without being central to the article being analyzed.
- Identifying important relationships between keywords: Identifying relationships (in the text) between regions, neurons, and the properties they express can be difficult, especially if more than one property has been identified with more than one neuron (or neuronal compartment). Therefore, information about semantic relationships (neuroscience and lexical) must be incorporated in the program to define such relationships between keywords as clearly as possible.
- Easy automated updating of the Knowledgebase: In the range of articles that NeuroText might be asked to analyze, the context and content are quite unrestricted and can essentially encompass all of neuroscience. Concepts hitherto unidentified (e.g., a neurological disorder found in neuroscience text that should serve as a contextual constraint) need to be incorporated into the knowledgebase to guide the system's operation. The domain expert should be able to add this new knowledge easily as he uses NeuroText to analyze articles.
- The domain expert needs to make the final decision: A survey of the efficacy of knowledge engineering studies reveals that successful programs identify target articles approx 70% of the time—where the parameters determined as **precision** and **recall** relate to **specificity** and **sensitivity** of the retrieval strategy, respectively (Korfhage, 1997; Raghavan et al., 1989; Tague-Sutcliffe, 1992). Since the eventual aim of our study is to populate databases with authoritative information, the expert/curator is charged with the responsibility of depositing articles with 100% accuracy.
- Presenting the results of NeuroText's analysis to the expert for validation and automatic deposition: The final necessary step is to present the results of text mining in an interface to the expert. The interface should highlight keywords, concepts of interest, and affirming and negating sentences if these are potentially decisive in determining whether an article is citable. An outline of the program's analysis should be readily understandable to the expert. The interface should also provide the expert with the tools to dynamically override erroneous results of the program. The interface should allow the expert to store validation results for continuous monitoring of the efficacy of the algorithms.

Background

Natural language understanding involves the development of computational systems to process and assimilate unstructured, non-annotated written and spoken language in a fashion

that mimics human understanding. Humans are capable of understanding the nuances of spoken or written text as a matter of course—our comprehension enhanced by knowledge of the world and by a constant process of learning. Whether speaking standard English, nongrammatical English, other languages or even sign language, communication has an instinctual component (Pinker, 1994).

Research in pursuing natural language techniques is of growing importance because of the increased availability of online text, according to Forrester Research (http://www.forrester.com) a business technology company. Human intervention in cataloguing millions of bytes of such data is impractical. Natural language processing (NLP), an information science endeavor for well over 20 years, includes projects in many areas, for example: word-indexing and retrieval of relevant articles, syntactic parsing of sentences, separating relevant keywords from random noise in a sentence, restructuring retrieved information to and from databases, interfacing programs with audio media, and translation of documents between languages.

NLP in the Information Science Domain

A wide range of NLP tools have been created in projects focusing on information science generally. IBM (International Business Machines) has devised various NLP tools. For example, Intelligent Miner is based on clustering algorithms that seek to identify and categorize vocabulary by concepts using lexical relatedness (Justeson and Katz, 1995). Intelligent Miner also uses a Multilanguage interface "LINGUINI" (Prager, 1999).

Several other organizations have also created practical software to analyze natural language, e.g., Verity Inc.'s Knowledge Organizer (Verity.com, 2000), TextWise's CINDOR (Multilanguage) (CindorSearch.com, 2002) and TextWise's Content Repurposing Suite (Textwise.com, 2002). Such software enables businesses to process documents, emails, and other online text for categorization and retrieval. WEBSOM is an internet tool for clustering of documents online newsgroups, created at the Helsinki University of Technology in Finland (Lagus, 2000).

NLP in the Life Science Domain

NLP techniques have also been used in the clinical and biological field. Krauthammer and coworkers reported a method of identifying protein and gene names from the literature by using the techniques of BLAST searching (Krauthammer et al., 2000). Several groups have created NLP systems for molecular biology: GENIES (Friedman et al., 2001), EcoCyc (Karp et al., 1999), TEXTQUEST (Iliopoulos et al., 2001) and RIBOWEB (Chen et al., 1997) are a few examples.

Friedman and coworkers have also utilized NLP methods in the clinical domains: MEDLEE seeks to identify radiology concepts from reports and medical discharge summaries using NLP (Friedman et al., 1994). Hersh et al. have created SAPHIRE, which maps clinical keywords to their Unified Medical Language System (UMLS) identifiers (Hersh et al., 2002). UMLS concepts have also been used by Aronson and coworkers who seek to map clinical keywords for query, search, and retrieval strategies (Aronson, 2001; Kim et al., 2001; Weeber et al., 2001).

In NEGFINDER, Mutalik and coworkers mapped medical concepts to their UMLS unique identifiers while identifying negated concepts in discharge summaries and medical reports (Mutalik et al., 1999). Two systems, BRENDA, which is based at the University of Cologne and presents information related to enzymes extracted from 46,000 articles (Schomburg et al., 2002), and NTDB, a database for thermodynamic properties of nucleic acids (Chiu et al., 2001), similarly rely on published scientific literature as their sources.

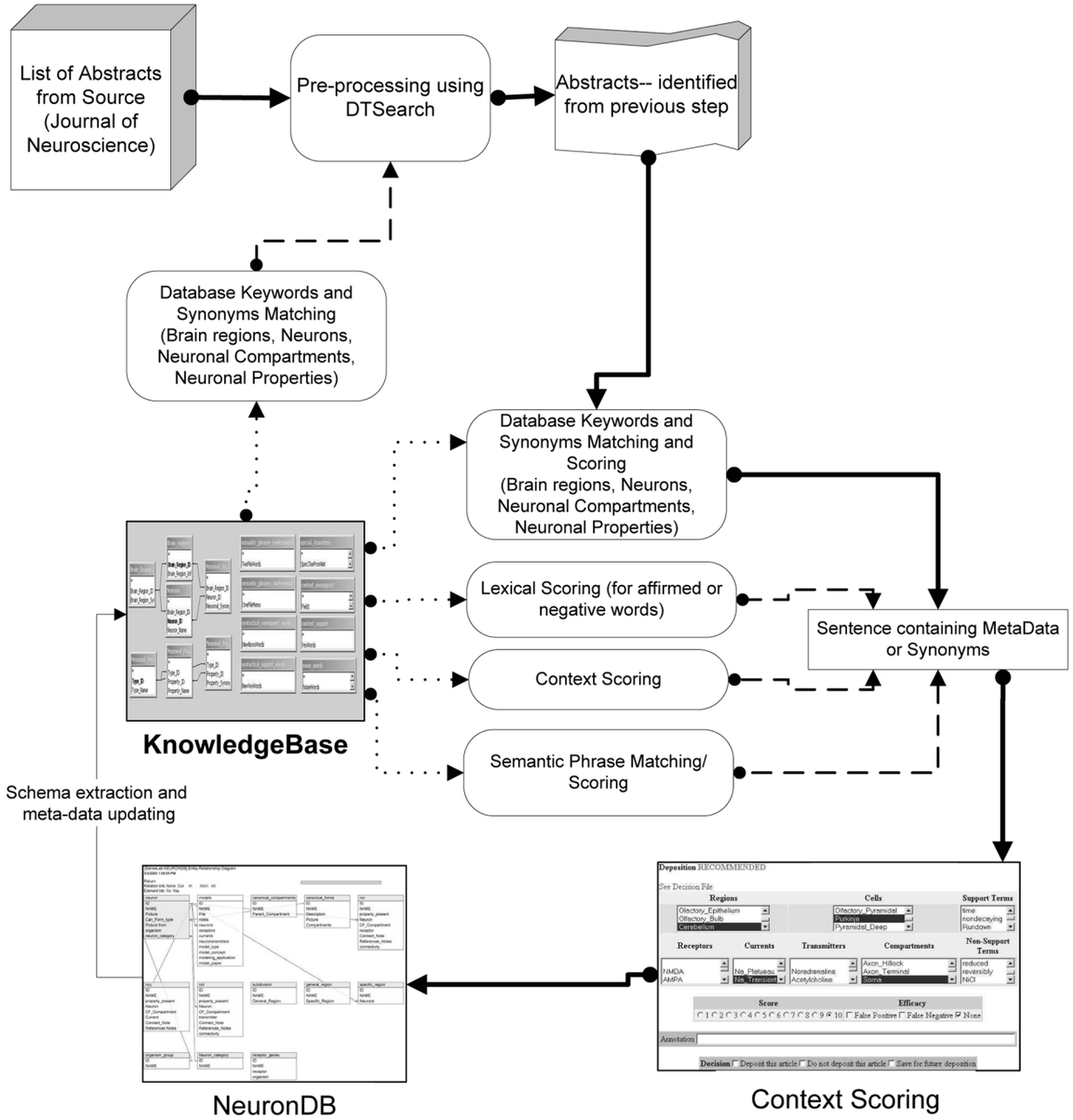

Fig. 2. The Process Flow Diagram shows the entire NeuroText process, which is a single PERL Script. The bold lines show the main process, dashed lines show the intermediate read and write processes, and dotted lines show access to the information stored in the knowledgebase. The abstracts are first processed using DTSearch, which uses database keywords and their synonyms from the knowledgebase. The articles that meet the search constraints are post-processed to identify and score abstracts that contain database keywords. The sentences that contain the keywords are further scored based on context matching, semantic phrase matching, and lexical (affirmed and negated word) matching. The abstract thus processed is presented to the expert in a dynamic interface. The expert then deposits the relevant abstract information into the databases.

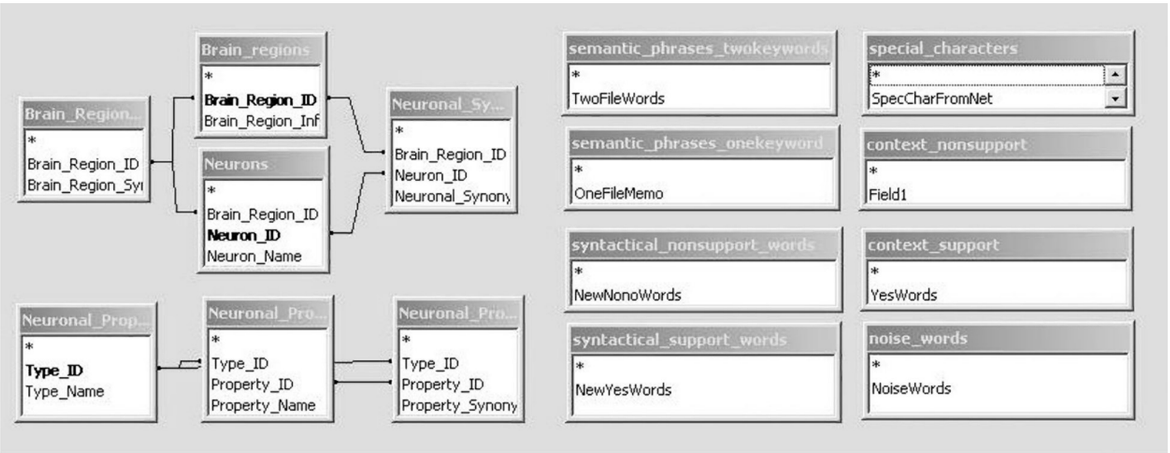

Fig. 3. Entity-Relationship (ER) diagram showing the knowledgebase that each abstract scanned in NeuroText is accessed by. The brain regions *are related* to neurons present in that region, in compliance with the database hierarchy. The neuronal properties are independent as they appear in the text of the abstract, as are the syntactic, semantic, and context word (phrase lists). Each metadata (regions, neurons, and properties) table is related to its synonyms table. Every synonym identified would map back to the metadata.

Methods

This section gives an overview of NeuroText's operation and describes aspects of its design. Figure 2 provides a process flow diagram for NeuroText.

A *single* PERL server-side script runs the *entire* NeuroText Program—including the pre- and post-processing steps. The only entry that the expert/curator NeuroText requires is the volume number from the *Journal of Neuroscience*. The principal steps in NeuroText's operation are:

- Automatic downloading of abstracts from the *Journal of Neuroscience*. (Alternately, this step can be performed manually.)
- Dynamic creation of a set of DTSearch text-searching macros, including keywords and synonyms, from the database that contains NeuroText's knowledgebase. (DTSearch is a commercial text-indexing and retrieval program [DTSearch, 1999]).
- Pre-processing the downloaded abstracts using DTSearch, searching for relevant abstracts that meet defined search criteria.
- Post-processing abstracts filtered through the previous step. This post processing involves identifying and scoring keywords and synonyms, lexical scanning, context matching, and semantic phrase matching.
- Creating a dynamic, web-based interface that allows the expert to assess the search results and deposit relevant abstracts into a SenseLab database. This dynamic interface, generated by the PERL NeuroText program, is also a server-side PERL CGI-script as opposed to text-based HTML pages. Such a script is necessary because the interface contains embedded forms in which commands for automatic deposition into the neuroscience databases are encoded.
- Each NeuroText result after pre- and post-processing is tested by the experts (MM and GMS) to validate NeuroText's decision to "Deposit" or "Not to Deposit." If the expert decision can-

not be made after scanning the abstract, NeuroText provides a link to the full text article. While recourse to the full text is not an objective test of NeuroText's validity since only abstracts are scanned, it is important in achieving the ultimate goal: populating the databases with 100% accuracy.

NeuroText's Knowledgebase

The NeuroText knowledgebase is stored in a Microsoft Acess database. The knowledgebase design is represented in Fig. 3.

Each table in the knowledgebase contains information accessed during the various search steps, e.g., information regarding database keywords and their synonyms, and words and phrases that serve as contextual and lexical determinants. The entity-relation (ER) diagram in Fig. 3 shows that only the brain region (and synonyms) and the neurons (and synonyms) are "joined" (because there exists a specific neuroscience relationship, namely, location). An SQL statement queries the relationship (specific location) between regions and neurons. The tables containing neuronal properties, and lexical, semantic, context word/phrase lists are independent because NeuroText attempts to identify concepts for their occurrence in the text. Regions, neurons and neuronal properties, and their synonym tables are also related primarily because every synonym maps back to its keyword that is housed in the CellPropDB and NeuronDB databases. Each word list against which the abstract text is scanned is in an independent table. They are not related to any other keywords and no such relationships need be established.

Keywords and Synonyms

The first set of tables is associated with keywords (related to brain regions, neuron names, neuronal properties, neuronal compartments and neuronal connectivities) and their synonyms. An example from the neurons table for the thalamic reticular neuron is:

SenseLab_Object_ID: 14
InternalKeyWord_Name:
Thalamic_Reticular
Visual Keyword_Name: thalamic reticular
Synonym1: nucleus reticularis thalami
Synonym2: NRT
Synonym3: perigeniculate

This entry defines the search string "thalamic reticular" and a set of synonyms and closely related terms. When any of these strings are found, the system will initiate a count for the concept "Thalamic_Reticular." (The "SenseLab object identifier" ensures that any keyword or synonym in the text maps to the database internal keyword and its unique identifier.)

Neuroscience Context Word Lists

The second set of tables contains lists of words that help determine the neuroscience context in which particular keywords occur. The context tables were created after extensive consultation with the experts. There are two sub-types of such lists: supporting and non-supporting concepts (as defined by experts). The words: *potentiation*, *polarization*, and *spatiotemporal* are examples of supporting concepts; *seizures*, *dementia*, and *epilepsy* are examples of non-supporting concepts. (They imply the context of disease rather than normal function.)

Word Lists for Affirmation and Negation

Additional tables, similar to the context tables, store words and phrases which imply affirmation (e.g., *significant*, *marked*, and *certain*) or negation (e.g., *nullify*, *refute*, and *uncertain*) words. The word lists, consisting of 828 (349-affirmed and 479-negated) words, were created using Roget's Thesaurus (http://humanities.uchicago.edu/orgs/ARTFL/forms

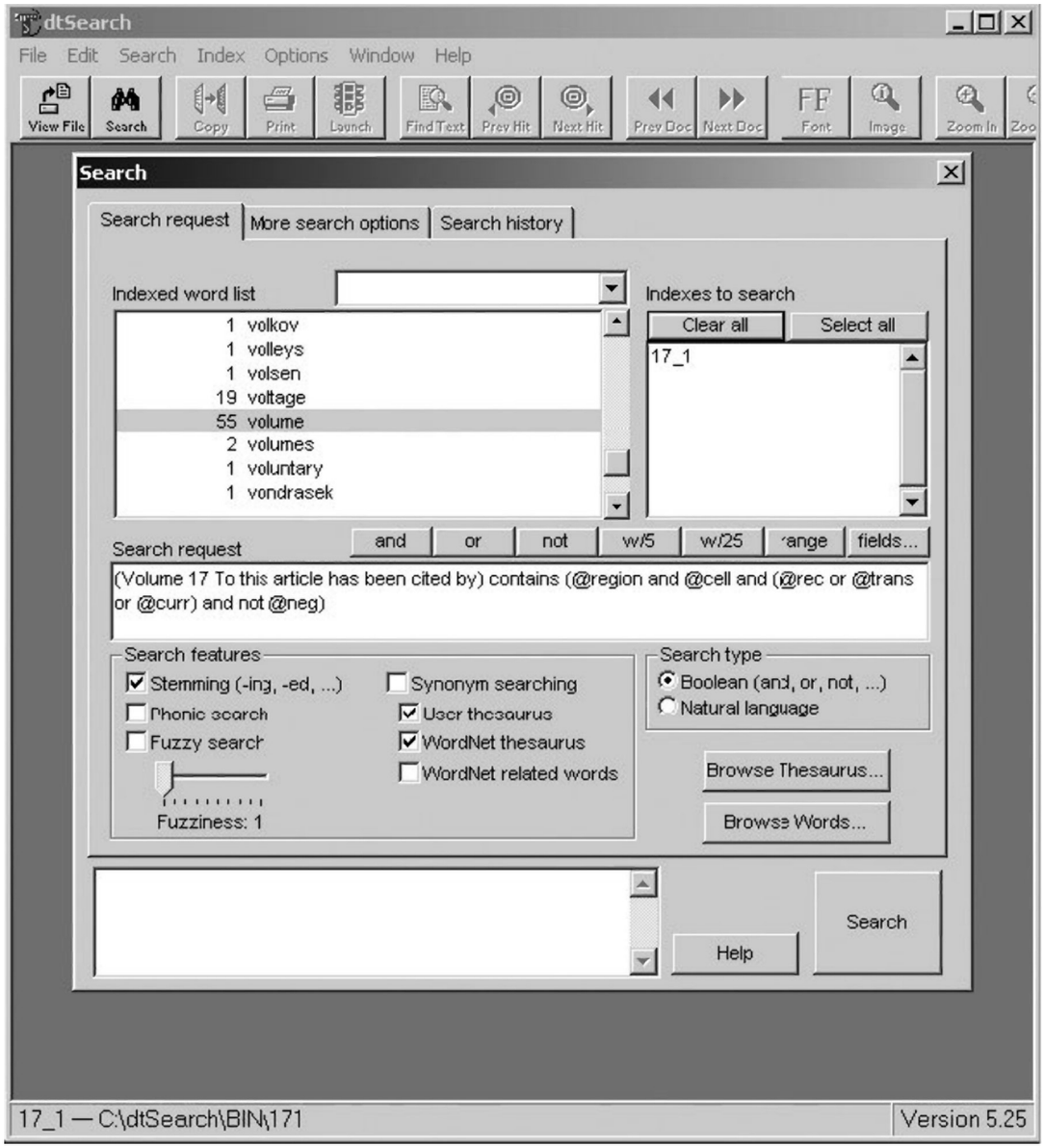

Fig. 4. Searching indexes of neuroscience articles using DTSearch. The search command illustrates the use of macros where the database keywords and their synonyms are embedded.

_unrest/ROGET.html) in addition to affirming or negating words found in common English usage.

The word lists were generated using the seed words "certain" and "uncertain" for the affirmed and negated word lists, respectively.

Semantic Phrases That Can Be Augmented by "Unsupervised Learning"

Semantic phrase tables were created *a priori* from a training set of 50 full-length neuroscience articles. Additional semantic relationships were derived from information already stored in the target databases. These stored relationships generally define the brain-region-neuron-neuronal property hierarchy. The training set articles were scanned to identify, primarily, semantic phrases that related brain regions to neurons, and neurons (or neuron compartments) to neuronal properties typically used by researchers in journal articles and neuroscience parlance. The semantic phrase tables store phrases that can identify relationships that would enhance the probability of identifying properties key to the search as opposed to properties occurring randomly in the text. Examples include "expressed in" and "mediated predominantly." As described later in the paper, NeuroText is designed to recognize new semantic phrases and add them to its knowledgebase automatically.

Archival Tables to Help Maintain the Knowledgebase

The final component of NeuroText's knowledgebase consists of archival tables. Archival tables are used to store words, concepts, and phrases that have been removed from the database. This information is not deleted; it can be restored to active use (if desired) when the knowledgebase is updated.

The NeuroText Program

A single script is used 1) to process all the abstracts downloaded for one volume of the *Journal of Neuroscience*, and 2) to create a web-based interface that allows the expert to validate NeuroText results. NeuroText identifies abstracts relevant for deposition in CellPropDB and NeuronDB in two steps: a sensitivity search followed by post-processing for specificity. Abstracts were chosen for analysis in favor of full-length text because 1) abstracts generally captured the main themes of the articles avoiding irrelevant keywords; 2) Scanning abstracts also sped up computation time.

Sensitivity Search

The sensitivity search makes use of a commercial indexing program DTSearch© (DTSearch, 1999). Figure 4 illustrates the user-interface to DTSearch when used manually.

(Since NeuroText used DTSearch in automated mode, the NeuroText user does not see this interface.) Database keywords and synonyms from the NeuroText knowledgebase are dynamically incorporated into the DTSearch control files. A batch script dynamically generated from the ACCESS database indexes the abstracts into word lists for each issue in each volume and creates a search query patterned after the hierarchy described previously: 1) brain region, 2) neuron, and 3) property.

The search is designed to identify, if possible, at least one region, one neuron, and one property in each abstract. The search first scans the abstract 1) for all brain regions to find one or more, then 2) for all the neurons to find one or more, and finally 3) for all the properties to find one or more. The top-level DTSearch command that coordinates this scanning process is:

"(@JournalName **TO** @AbstractLimiter) **CONTAINS** (@regions **AND** @cells **AND** (@receptors **AND** @currents **AND** @transmitters))" (Expression 1)

The first part of this expression ("@Journal Name **TO** @AbstractLimiter") restricts the scanning of the abstract to avoid erroneously scanning titles of articles cited in the abstracts' reference lists. For our pilot study, @JournalName = "Journal of Neuroscience."

In DTSearch, each "@" represents a macro. A macro can be one word or several words connected by simple (AND or OR) or complicated Boolean operators. For example, "Ip,q" is a calcium-ion current. In the literature, this current may be referred to as Ip,q or P,Q currents and P,Q-type or Purkinje currents (Sun and Dale, 1998). To identify these variants, NeuroText uses the following macro:

```
(I p,q OR Purkinje OR ((P OR Q) w/3
   (type OR channel)))
```

"W/3" means that the connected words (or phrases) should occur within *three* words of each other. Specific word ranges for different database keywords were determined after a careful survey of articles in the training set. Each macro may contain one or several layers of macros embedded in it.

The aim of this first step (the sensitivity search) is to identify as many articles as possible that contain keywords or concepts associated with database keywords in the correct hierarchy. During this step, NeuroText does not seek to relate region, neuron, and property. Every abstract-search that meets the sensitivity search criterion in Expression 1 is combined into an initial search report that becomes the starting point for further analysis.

Post-Processing for Specificity

Post-processing is designed to help ensure the specificity of the neuroscience abstract for deposition into CellPropDB or NeuronDB: that the neuron mentioned in the article does indeed belong to a specific region in the brain, and the property was indeed found (or was not found) in that neuron and/or its compartments. Post-processing helps ensure that the contextual and lexical constraints of information being suggested for deposit into NeuronDB and CellPropDB are adhered to. The individual post-processing steps are described in the following paragraphs.

Scanning the Abstract Text

Every word of every sentence in the abstract is scanned using the NeuroText knowledgebase to identify keywords and their synonyms. Each time a keyword is identified, a count for this keyword is initiated. At the same time, every word in the entire sentence where a keyword occurs is scanned against the context table to identify contextual patterns that might enhance or reduce the score that the keyword receives. (If no discernable context is identified, then the initial score for each keyword is retained.)

The entire abstract is included in the search report. This enables NeuroText to search for additional keywords related to neuron compartments (axonal, somatic, and dendritic) that were not part of the earlier search. In this second search, for example, certain fiber pathways and interneurons, which provide key connectivity information as to their originating and terminating neurons, can also be identified. An example of a pathway is **mossy fibers**. The term "mossy fibers" applies to axons that arise from dentate granule cells and terminate on CA3 cells in the hippocampus (Claiborne et al., 1986), and axons that arise from inferior olive cells and terminate on Purkinje cells in the cerebellum. Occurrence of the word "mossy fibers" in the text of an article therefore counts as a score increment for dentate granule axons and the presynaptic axon input to CA3 pyramidal neurons. NeuroText can distinguish mossy fibers associated with Purkinje cells as irrelevant to the database—primarily because they are not currently in the databases.

Each sentence of the abstract's text that contains keywords is also scanned for affirming or negating contexts. Sentences in the text where either concept occurs are flagged. Keywords in these sentences are scored in a variety of ways. For example, only sentences that contain keywords for properties (receptors, currents and transmitters) and neuronal compartments are scored if affirmed or negat-

A

160.pdf-Link to Full-Text of Article 160.xml XML-Link

Volume 17, Number 1, Issue of January 1, 1997 pp. 160-170

Low-Threshold Ca 2+ Currents in Dendritic Recordings from **Purkinje** Cells in Rat **Cerebellar Slice** ~~Cultures~~

Received July 12, 1996; revised Oct. 15, 1996; accepted Oct. 22, 1996

Brain Research Institute, University of Zurich, CH-8029 Zurich, Switzerland

Voltage-dependent Ca 2+ conductances were investigated in **Purkinje** cells in rat **cerebellar slice** ~~cultures~~ using the whole-cell and cell-attached configurations of the patch-clamp technique. In the presence of 0.5 m M Ca 2+ in the

a slow (304 ± 46 msec time constant), and a nondecaying component. Rundown of the slow and sustained components of the current, or application of antagonists for the P/Q-type Ca 2+ channels, allowed isolation of the **fast-inactivating** Ca 2+

current, which had a threshold for **activation** of - 60 mV and reached a maximal amplitude of 0.7 nA at a membrane potential of - 33 mV. Both **activation** and steady-state inactivation of this **fast-inactivating** Ca 2+ current were described

with Boltzmann equations, with half-activation and inactivation at - 51 mV and - 86 mV, respectively. This Ca 2+ current was nifedipine-insensitive, *but* its amplitude was reduced reversibly by bath-application of NiCl 2 and amiloride, thus

allowing its identification as a T-type Ca 2+ current. Channels with a conductance of 7 pS giving rise to a fast T-type ensemble current (insensitive to omega -Aga-IVA) were localized with a **high** density on the dendritic membrane. Channel

in **somatic** membrane patches.

Cerebellum; Purkinje□; ; Ca; Na_Transient; Soma;

B

1848.pdf-Link to Full-Text of Article 1848.xml XML-Link

~~Estradiol~~ Increases the Sensitivity of **Hippocampal** CA1 Pyramidal Cells to NMDA Receptor-Mediated Synaptic Input. **Correlation** with Dendritic Spine Density

Received Sept. 12, 1996; revised Dec. 12, 1996; accepted Dec. 19, 1996

Previous studies have shown that ~~estradiol~~ induces new dendritic spines and synapses on **hippocampal** CA1 pyramidal cells. We have assessed the consequences of estradiol-induced dendritic spines on CA1 pyramidal cell intrinsic and synaptic

electrophysiological properties. **Hippocampal** slices were prepared from ovariectomized rats treated with either ~~estradiol~~ or oil vehicle. CA1 pyramidal cells were recorded and injected with biocytin to visualize spines. The association of

dendritic spine density and electrophysiological parameters for each cell was then tested using linear regression analysis. We found a *negative* relationship between spine density and **input** resistance; however, no other intrinsic property

measured was **significantly** associated with dendritic spine density. Glutamate receptor autoradiography demonstrated an estradiol-induced **increase** in binding to NMDA, *but not* AMPA, receptors. We then used input/output (I/O) curves (EPSP slope

vs stimulus intensity) to **determine** whether the sensitivity of CA1 pyramidal cells to synaptic **input** is correlated with dendritic spine density. Consistent with the lack of an ~~estradiol~~ effect on AMPA receptor binding, we observed no

relationship between the slope of an I/O curve generated *under* standard recording conditions, in which the AMPA receptor dominates the EPSP, and spine density. However, recording the pharmacologically isolated NMDA receptor-mediated

component of the EPSP revealed a **significant** correlation between I/O slope and spine density. These results indicate that, in parallel with estradiol-induced increases in spine/synapse density and NMDA receptor binding, ~~estradiol~~ treatment

increases sensitivity of CA1 pyramidal cells to NMDA receptor-mediated synaptic input; further, sensitivity to NMDA receptor-mediated synaptic **input** is well correlated with dendritic spine density.

Adams, M. M., Fink, S. E., Shah, R. A., Janssen, W. G. M., Hayashi, S., Milner, T. A., McEwen, B. S., Morrison, J. H. (2002). ~~Estrogen~~ and Aging Affect the Subcellular **Distribution** of ~~Estrogen~~ Receptor-alpha in the **Hippocampus** of Female

NMDA; AMPA;

ed; affirmed or negated scoring is not carried out if a sentence contains neurons and brain regions keywords.

It is relatively easy to associate cells with regions in the brain. For example, the olfactory mitral cell is the principal neuron in the olfactory bulb (Mori et al., 1981). In making these associations, NeuroText also needs to be able to discriminate ambiguities (e.g., whether granule cells arise from the dentate gyrus versus the cerebellum). Differentiating properties found in neurons, especially if keywords for more than one are found in the same abstracts, can pose difficulties. The situation is further complicated if properties are identified with more than one compartment in a neuron. Isolating a property to a neuronal compartment is difficult to extract from natural text (unless explicitly stated, e.g., "occurs in" or "expressed in [by]"), especially if the information is contained in keywords scattered throughout the text. In such cases, very often, the expert relies on *a priori* knowledge to extract relevant information. As a result, to handle such complex cases, our presentation interface (discussed later) provides a link to the full-length article to help the expert make the final decision to cite an article.

Semantic Phrases and Unsupervised Learning

During post-processing, when a "neuronal property" keyword occurs in a sentence by itself or with another keyword, the sentence is scanned against the "semantic phrase" component of NeuroText's knowledgebase. If a relevant phrase is identified, the keyword score is enhanced. It is also flagged as related to the neuron (or compartment) or region. The sentence is also scanned for an affirming or negating tone. If a negating word or concept is identified with the property, the score for that property is not enhanced. Each property (or neuronal compartment) keyword is thus scored differently from a region or a neuron.

If potential "relation" phrases are not identified, the sentence containing the keyword matches is stripped of database keywords and extraneous noise words, and then appended to the end of the semantic relationship table in the knowledgebase. For example, the phrase "rapidly activated" relates an ion channel to a neuron. This phrase was stored in the semantic phrase table. If a sentence in the abstract contains two keywords, and the sentence contains the phrase "rapidly activated" then NeuroText assigns a score increment if the keyword relates to a property or a neuronal compartment. If the phrase does not find a match, it is appended to the table of such semantic phrases. Any subsequent abstract or full-length article that NeuroText scans will avail of this new phrase.

Fig. 5. (left) **(A)** Results of post-processing in NeuroText as presented to the expert. In this example, NeuroText determined that the article should be deposited. The lexically negated sentence has been highlighted along with database keywords and support and non-support terms. Here calcium and sodium currents have been identified in the Cerebellar Purkinje cells. The decision file for this abstract is at http://senselab.med.yale.edu/textmine/1604.html. **(B)** In this example, NeuroText determined that the article should not be deposited. The lexically negated sentence has been highlighted along with database keywords and support and non-support terms. The figure shows that while the receptors NMDA and AMPA are identified, no regions in the brain or neurons (currently present in the databases) were identified. The decision tree for this file can be found at http://senselab.med.yale.edu/textmine/1848.html.

Scoring for Relevance

After the abstract text has been scanned, the scores for each neuroscience keyword are examined. The maximum score for keywords of each type is determined. Regions are identified with neurons. Regions and neurons that do not match are discarded. If more than one region-cell pair or property or neuronal compartment exists in the text, keywords whose count is less than one-fourth of the count for the maximum of that class of keywords are discarded as likely "random" occurrences. (The presentation interface discussed in the following allows the expert to dynamically change the results if a keyword is mistakenly identified or discarded.) In this way, the scores are based on database and neuroscience context, on affirming or negating sentence tone (in case of properties and compartments), and on semantic relation-phrase matching.

Organizing the Citation Information

NeuroText also identifies and separates citation data including first and last page numbers, volume number, month and year of publication, authors, and title. This information is used when populating the NeuronDB and CellPropDB databases.

Interface for Expert Evaluation and Deposition

Interface Design

Every abstract processed in the previous step is presented to the expert (Fig. 5A,B).

This interface, which is generated as the post-processing step proceeds, is a dynamically generated PERL (CGI) script that performs information presentation and data deposition. An expert can access this interface remotely on the internet, make decisions, and deposit relevant data. The file presents the abstract with relevant words and sentences highlighted. Database keywords are enlarged; enhancing concepts are bolded; non-support concepts have "strikethroughs" in them. Negating words are italicized; affirming words have a different font. Sentences with negating words are white text on black background; sentences that point to lexically affirming tones have a gray background. Sentences with conflicting negating and affirming tones or no discernible tones are not highlighted. The abstract is followed by NeuroText's assessment regarding the relevance of the article, as well as a link to an HTML file that contains a step-wise explanation leading to the NeuroText decision.

NeuroText's Assessment Decision

After analyzing the abstract, NeuroText presents one of three decisions in a dynamically generated web page to the expert:

- A "Deposition Recommended" decision is made if a region, a neuron *in* that region, and a property identified *with* the neuron or its compartment is clearly identified.
- "Deposition Not Recommended" is the NeuroText decision if the score for keywords in an abstract during the sensitivity search is nullified, if the sentences are identified by the context that would negate the scores, or if identified regions are not associated with identified neurons.
- A "Deposition Under Advisement" decision is made if a region-cell pair is properly identified in NeuroText but either no specific property (from the database) is identified, or if the property scores are negated but the region-neuron pairs are not. NeuroText assumes a possible relation between a neuron (and its compartments) and its properties, and directs the expert to take a closer look before making a decision.

This web page also serves as a script which will be used to deposit information into the databases. The NeuroText "decision" is accompanied by a link to a decision tree which enables the experts and curators to view a stepwise breakdown of NeuroText's scoring. This is followed by a deposition form (Fig. 6) containing

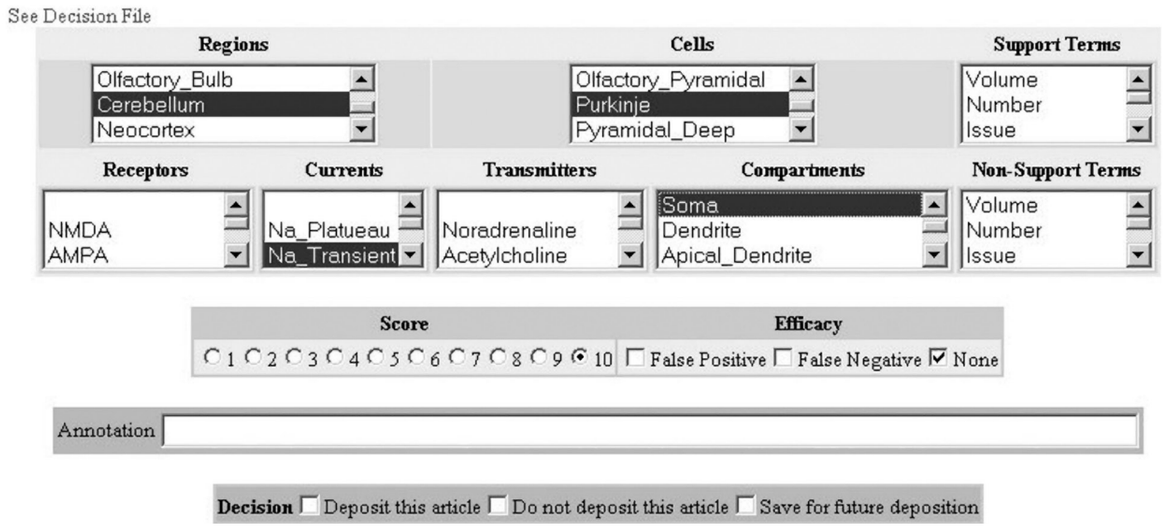

Fig. 6. The decision form in the deposition interface for an abstract. Each of the keywords identified and scored are automatically highlighted. Complete lists are included to enable the experts to override erroneous findings in NeuroText. The knowledgebase can be augmented by clicking relevant words and phrases in the lists marked "Support Terms" and "Non-Support Terms."

tabulated, scrolling lists in which keywords for regions, neurons, receptors, ionic currents, neurotransmitters and neuronal compartments are identified.

The expert can also ascertain if the decision as a whole or in part contains false-positives or false-negatives, and record these while scoring the search efficacy. The expert can override NeuroText based on their assessment of an abstract by making changes in this information before it is deposited into the SenseLab database (by clicking on the correct information in the scrolling lists).

Supervised Learning

Supervised learning was included in NeuroText to allow the system to update and modify the knowledgebase in a continually evolving fashion. To allow such *supervised* learning, two additional scrolling lists, besides those associated with keywords, are presented to the expert in the interface. These word-concept lists are identical. The lists are created from an array of words from the abstract's text after stripping away every word already present in the knowledgebase. If a concept not present in the knowledgebase is deemed necessary to enhance the score or diminish the score of keywords, the expert can click on that word or concept in the appropriate list. When the information is submitted for deposition, the negating concepts are added to the file containing non-support concepts in the knowledgebase.

Any phrases or sentences containing these words are also removed from the phrase knowledgebase and placed in an archival table. Similarly, new support concepts identified by the experts, are placed in the file containing support terms. The program also scans the archival tables to retrieve any phrases associated with this affirmed concept that might have been previously placed in the archive.

Deposition in the SenseLab Databases

During the post-processing, as each abstract is scanned, an XML file is generated. Embedded in the XML tags are citation data that will be deposited into the databases (Crasto et al., 2002). At the end of each abstract, the expert is prompted to decide whether the abstract information should or should not be deposited. When the information is submitted, the XML files for articles to be deposited are collated into a submission file.

Validation

In order to determine the efficacy of NeuroText results, the entire corpora of abstracts (*Journal of Neuroscience*, vol. 17) were presented to both experts (MM and GMS) at both steps. At the pre-processing step, identifying errors (specifically false-negatives, since false positives at this step will be further evaluated at post-processing) is important because they point to any deficiencies in the search constraints used by DTSearch. After post-processing, all the abstracts are presented to the experts. The experts verified the result for each abstract with the NeuroText scoring scheme. If the expert could not ascertain an article's suitability for deposition even from an independent perspective, the full text of the article was accessed.

SenseLab Database: Presenting Neuronal Property Information to the User

This subheading describes how SenseLab itself presents data to the user once it is deposited into a SenseLab database. Researchers can navigate the web pages of NeuronDB and CellPropDB and access information related to each neuron. For example, clicking on the "Plus/Reference" notes for the olfactory mitral cell in the olfactory bulb page in NeuronDB (http://senselab.med.yale.edu/senselab/NeuronDB/ndbEavSum.asp?id=267&mo=4&) reveals citations related to articles for different properties of neuronal compartments: "Intracellular recordings: CNQX blocks early component of EPSP response to olfactory nerve volley (Chen WR and Shepherd GM, 1997 [rat][16]" (Chen and Shepherd, 1997) is one example (Fig. 1). A user can click on the citation superscript "16" to obtain the reference. The experts (MM and GMS) provide the annotations. The above information relates to the AMPA receptor in the distal apical dendrite of the olfactory mitral neuron in the olfactory bulb. Each datum of information—title, annotation, author names, volume number, journal name, publication year, and page numbers—is then stored in the database whose architecture is based on the Entity Attribute Values with Classes and Relationship (EAV/CR) schema—a flexible schema devised to store and retrieve heterogeneous data (Nadkarni et al., 1999).

Results of the Pilot Study

To perform an initial pilot test of NeuroText in operation, 912 abstracts from volume 17 of the *Journal of Neuroscience* (1997) were downloaded (http://www.jneurosci.org). Figures 5A and B present the results from NeuroText for a sample abstract taken from this set. At the same time, these articles were scanned independently by the experts. The results of NeuroText post-processing (for every volume studied) are available at the SenseLab website at http://senselab.med.yale.edu/textmine/NeuroText.pl. This web page allows users to view NeuroText results (Internet Explorer [version 5 and above] Netscape [version 7 and above] work best for this page).

The page dynamically created for deposition into the databases is only available to the experts. Table 1 summarizes the results of NeuroText's analysis compared to that of the experts.

Table 1 —Summary of Results of NeuroText Versus Expert Analyses of 148 Articles from the *Journal of Neuroscience* Volume 17.*

	Articles for Deposition by Expert	Articles not for Deposition by Expert
Articles for Deposition by NeuroText	28 (True Positives [TP])	13 (False Positives [FP])
Articles Not for Deposition by NeuroText	9 (False Positives [FP])	98 (True negatives [TN])

*Not included are 29 articles that NeuroText deemed as "Deposition Under Advisement." The tabulated entries will be used to calculate Specificity and Sensitivity presented in the "Results" subheading.

$$Specificity = \frac{FN \times 100}{FN + FP} \; ; \; Sensitivity = \frac{TP \times 100}{TP + FN}$$

Of the 177 article abstracts identified for post-processing, 1) 29 were deferred by NeuroText for final decision to the experts, 2) 126 were correctly identified, 3) 13 were incorrectly identified (false positives), and 4) 9 were identified by the experts that NeuroText judged "Deposition Not Recommended" (false negatives).

Using the values from Table 1, of 148 articles for which NeuroText did not defer decision, NeuroText identified 126 articles correctly (in agreement with the experts) and 22 incorrectly, for an accuracy of 85%. Similarly, the proportions (Cicchetti and Feinstein, 1990; Spitzer and Fleiss, 1982) for identifying true positives was 72% and for true negatives 90%. Alternatively, using the *odds-ratio test* (Agresti, 1990), the odds ratio of a correct identification of an article (as citable or not-citable) by NeuroText is approx 26:1.

Identifying true positives correctly is important for accuracy of deposition. Correctly identified articles for deposition describe the sensitivity (recall) and specificity (precision) (Table 1). NeuroText identified 28 true positives, 13 false positives for a specificity of 90%; and 9 false negatives for a sensitivity of 76%. Ninety-eight articles were correctly identified as true negatives.

Subsequent analysis of the results revealed that almost all articles that NeuroText deferred (deemed as "Under Advisement") required that the experts consult the full text (available as a link in NeuroText results) before deciding whether or not to cite the article. Most NeuroText false positives were deemed as possible weak citations—that the abstracts did not contain novel information. Most of the false negatives were due to inadequacies of the knowledgebase. With a subsequently enhanced knowledgebase, the number of falsely identified articles decreased significantly. The experts did not call into question the algorithmic details nor the search and scoring strategies for any of the articles analyzed. Every volume of the *Journal of Neuroscience* contains approx 1000 articles. NeuroText's time for processing 1000 abstracts in a volume is less than two hours.

Discussion

In this subheading, we discuss examples of NeuroText's classification of abstracts, including cases where its results are in disagreement with the experts' judgment. The experts (MM and GMS) who are the main decision makers as to an abstract's citability also served as impartial evaluators of NeuroText results.

Interface of NeuroText Results

Figure 5A and B illustrates two examples where NeuroText and the experts are in agreement. In A, calcium and sodium currents are identified in the Purkinje cells of the cerebellum. The database keywords are highlighted. The scores for the keywords are enhanced from concepts entered into the knowledgebase. The sentence with a black background reflects the identification of a negated tone from the word "but." A closer view of the sentence shows that the word "but" does not have a bearing on the context of the article. It directs the expert to a sentence that might potentially negate the finding of the keyword in that sentence, and which may have consequences on the overall decision.

Enhancement of the Knowledgebase by the Expert

Figure 5B illustrates an abstract that NeuroText and the experts deemed "Not for Deposition." The word "estradiol" in the non-support file of the knowledgebase decreases the score of the region (hippocampus) and neuron (CA1 pyramidal cells) keyword matches. When these keywords-match scores are tallied, NeuroText deems them insufficient in the final decision. This method of counting illustrates the importance of a knowledgebase. If "estradiol" was not in the knowledgebase, NeuroText would have positively scored keywords and probably erroneously flagged the abstract as "fit for deposition."

As a result of this pilot test, certain keywords were added to lists of enhancing and negating keywords as determined by the experts. Such contextual information was not previously available to NeuroText, which failed to account for these while scoring the occurrence of a keyword in the databases. The nine non-support keywords and acronyms which were added to the knowledgebase following analysis of the results included among them: ischemia, parvalbumin, calcineurin, and estrogen.

Analyses of the results also allowed the entry of a synonym for thalamic reticular neuron in the thalamus not present in the knowledgebase—"perigeniculate."

Potential NeuroText Failures

Mismatches or Incomplete Matches

As mentioned earlier, it is impossible to create an all-encompassing knowledgebase. We anticipate that NeuroText's knowledgebase will continually evolve and that the domain expert will make the final decisions about deposition into the SenseLab databases.

- For example, CA1 and CA3 are regions in the hippocampus. Several of the false-positives arose from misidentification of CA1 and CA3 neurons as CA1 and CA3 pyramidal neurons.
- Dopaminergic cells were also misidentified as dopaminergic receptors. Specific peptide and enzyme information (not in the databases) was also not flagged as negated because of opioid receptor peptides being present in the database.

The knowledgebase when minimally modified to take into account these discrepancies properly identified these articles, e.g., replacing "CA1" with "CA1 pyramidal" for a keyword match. Researchers sometimes use CA1 and CA1 pyramidal neurons interchangeably; in which case, NeuroText would report a false hit. NeuroText's use of partial matches resulted in false-positives:

- NeuroText falsely mapped the sub-thalamic region (not in the database) as the thalamus. The entorhinal cortex was identified and scored as a hit for neocortex, which some might find problematic. In order to avoid such errors and problems, the knowledgebase was modified to an extent that only exact matches would be allowed. Such a step might prove detrimental in the future as useful information in the form of partial matches might be ignored in NeuroText.
- Specific properties such as long-term potentiation (LTP) and long-term depression (LTD) were initially considered of secondary relevance to the databases because of the overwhelming number of these studies. Articles related to these terms may contain relevant information that might merit deposition in the databases. In the future, these articles will appear under the decision "Under Advisement."

The advantage of presenting the interface to the experts with the tools to allow the dynamic modification of NeuroText decisions while making modifications to enhance the knowledgebase ensures that the information deposited is accurate. By changing the knowledgebase, most NeuroText failures can be remedied such that subsequent articles scanned would benefit from these changes. This would result in better agreements between expert and the computer program.

There are some instances where NeuroText in its current form would most likely fail, irrevocably. One such example is articles pertaining to diseases:

- NeuronDB and CellPropDB allow only articles describing research related to normal brains; thus, terms like Parkinson, Huntington, and Alzheimer's diseases and several other neurobiological disorders are contextually negated. NeuroText however, cannot differentiate between the articles (where these keywords occur) that deal with the biology of cells and those that describe clinical work. According to the expert, the latter may be deposited, the former not. The knowledgebase would have to undergo considerable enhancement and domain-extension to include clinical articles—beyond the scope of NeuroText in its current form.
- In the analysis of the results, the expert often termed an article as too general or too specific to be deposited. This qualitative determination can be neither borne out by NeuroText results nor by a careful perusal of the decision tree. The telling detail in these NeuroText failures is that they are not consistently false-positives or false-negatives. The failures encompass both in equal measure—indicative of information that might be inferred from the abstracts in the absence of keywords or a concept that could enhance or negate the scores for the keywords, if present.

With an aim to providing relevant information for the knowledgebase, the experts often had to access the full text of the article seeking information that could be condensed and added to the knowledgebase—most times such attempts met with failures. Four specific examples follow.

NeuroText False-Negatives

- The decision tree for the abstract for article: "Inhibition of Synaptic Transmission by Neuropeptide Y in Rat Hippocampal Area CA1: Modulation of Presynaptic Ca^{2+} Entry" (Qian et al., 1997) is at http://senselab.med.yale.edu/textmine/8169.html. The decision tree shows that no CA1-CA3 hippocampal pyramidal neurons were identified even though the abstract clearly shows that they express a calcium ion channel. NeuroText judged this article as not to be deposited since no neuron was clearly identified. The experts however, determined that this article merited deposition into CellPropDB. They opined that the CA3-CA1 synaptic pairs are unique. They are made by Schaeffer collaterals of CA3 pyramidal neurons on the middle region of the apical dendrites of the CA1 pyramidal neu-

rons. The information even without the associated keywords would map onto relevant pyramidal neurons, hence meriting deposition.

- In the abstract of the article: "Dopaminergic Modulation of Sodium Current in Hippocampal Neurons via cAMP-Dependent Phosphorylation of Specific Sites in the Sodium Channel α Subunit" (Cantrell et al., 1997), one region was identified—the hippocampus—without specific mention of CA1 or CA3 pyramidal neurons—key to the databases. Nigral dopaminergic neurons were also identified without specific mention of their associated region—Substantia Nigra. The NeuroText decision tree rejected this abstract (http://senselab.med.yale.edu/textmine/7330.html) for lack of specific neurons associated with identified regions. The experts however, found that by tracking the dopaminergic input into the hippocampus, the actions of specific dopamine receptors that modulate specific presynaptic terminal properties would be interesting enough to deposit. This is an example where the experts determine that the article despite the non-specific nature of the study is interesting and novel enough to present to database users.

NeuroText False-Positives

- The first of two abstracts termed "weak" (NeuroText false-positives) by the experts was: "Ca^{2+} or Sr^{2+} Partially Rescues Synaptic Transmission in Hippocampal Cultures Treated with Botulinum Toxin A and C, But Not Tetanus Toxin" (Capogna et al., 1997). The NeuroText decision tree (http://senselab.med.yale.edu/textmine/7190.html) indicates that the Hippocampus and CA3 pyramidal neurons were identified, along with a calcium ion channel—strontium not being part of the databases. The knowledgebase has been modified to specifically identify CA3 pyramidal neurons in the hippocampus, as opposed to unspecified CA3 neurons. Identification of CA3 in the abstract comes from the words "CA3 pairs." The expert's opinion was that these words might or might not mean CA3 pyramidal neurons. In this particular case, NeuroText failures arose from an uncertainty in identifying keywords whose names may not be in standard neuroscience usage.
- The abstract "Instantaneous Perturbation of Dentate Interneuronal Networks by a Pressure Wave-Transient Delivered to the Neocortex" (Toth et al., 1997) was also determined as not meriting deposition by the experts: NeuroText identified the AMPA and glutamate receptors in the Soma of Dentate Granule cells in the Dentate Gyrus, therefore deciding that this abstract merited deposition. Keyword matches were also found for Neocortex; these hits were discarded as random occurrences as no matching neurons were identified (http://senselab.med.yale.edu/textmine/8106.html). In the experts' opinion however, the AMPA and Glutamate receptors were identified on interneurons and were not related to granule cells. The receptor property identified as related to dentate granule cells was incorrect—and the interneurons are not part of the database. This difficulty in identifying specific properties being expressed in specific cells from "unhelpful" text was alluded to in the "Methods" subheading.

Scalability and Interoperability

Scalability is an important consideration. Obvious questions arise as to the effort it takes to build a knowledgebase for a domain being studied. Our recommendation is to create a training set depending on availability and accessibility of articles that, according to the expert(s), contains key information. Once an initial knowledgebase is established, the evolution tools of NeuroText, (i.e., the word-lists of support and non-support terms available to experts and curators in the presentation and deposition interface for each abstract) will enable knowledgebase enhancement.

The validation-deposition interface is linked to the SenseLab databases that use a specific architecture. Naturally, such links would not

be useful to populate databases with different articles on different platforms. To address this issue, every NeuroText result also contains a dynamically created XML file in whose nested fields the mined information is embedded. The XML files are created with a view to interoperability. Researchers who wish to use the NeuroText tool would have to simply create an XML parser (XML parsers are available for different platforms and programming languages) to extract relevant data and link (or post) it to a database or storage medium of their choice.

Conclusions

One of the key features of NeuroText is that it is designed to be extensible to different domains. All the domain-specific information resides in a knowledgebase separate from the program code. The knowledgebase tables contain terms and concepts specific to the SenseLab databases, the neuroscience domain, and to affirming or negating tone. This information could be replaced for use in another similar bioscience domain. Another important feature of NeuroText is its dynamically generated interface, which presents results to the expert and allows the user to override the erroneous results of the automated method while automatically adding to the knowledgebase. NeuroText evolved with the need to populate the SenseLab neuronal databases rapidly and accurately. A side-by-side comparison of the time it takes for an expert to scour a year's worth of articles from the *Journal of Neuroscience*, naturally, is not possible. The time for processing approx one thousand articles (24 issues) by NeuroText is approx one hour and forty minutes. NeuroText also offers continuity and consistency and eases the workload on experts and database curators.

The "evolution" steps in NeuroText ensure that the knowledgebase is constantly enhanced and modified depending on the type of information that the experts and database administrators want to disseminate. This helps ensure that the program increases its accuracy every time it scans a neuroscience article. The knowledgebase is very simple to create, format, modify, and update.

Future Directions

To extend NeuroText abilities beyond the *Journal of Neuroscience*, NeuroText was also tested on 100 articles downloaded from PUBMED using a search for keywords "cerebellum" and "Purkinje." The results are available at http://chutney.med.yale.edu/textmine/Cerebellum_Purkinje.pl. When fully operational, we expect approx 90% of the databases to be populated using the NeuroText tool. We anticipate that approx 10% of the information will be supplied by users of the databases in terms of interesting information not directly available from online sources. We are in the process of extending NeuroText to include all neuroscience publications (monographs, edited volumes and brain atlases), thus helping ensure a comprehensive automatic retrieval and deposition into SenseLab databases.

Acknowledgments

This work was supported in part by NIH grant P01 DC04732 and by NIH grants P20 LM07253, T15 LM07056, and G08 LM05583 from the National Library of Medicine; NIH grant P01 DC04732-03, and Office Of Naval Research/Multidisciplinary Research Program of the University Research Initiative (ONR/MURI) grant DAAG55-98-1-0266

References

Agresti A. (1990) Categorical Data Analysis, Wiley, New York, pp. 59–66.

Aronson A. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc. Am. Med. Inform. Assn. Symp. Washington DC, pp. 17–21.

Baeza-Yates R. and Ribeiro-Neto B. (1999) Modern Information Retrieval, Addison-Wesley, New York, pp. 99–114; 191–224.

Barde Y. A., Edgar D. and Thoenen H. (1982) Purification of a new neurotrophic factor from mammalian brain. EMBO. 1, 549–553.

Cantrell A. R., Smith R. D., Goldin A. L., Scheuer T., and Catterall W. A. (1997) Dopaminergic Modulation of Sodium Current in Hippocampal Neurons via cAMP-Dependent Phosphorylation of Specific Sites in the Sodium Channel a Subunit. J. Neurosci. 17, 7330–7338.

Capogna M., McKinney R. A., O'Connor V., Gähwiler B. H., and Thompson S. M. (1997) Ca^{2+} or Sr^{2+} Partially Rescues Synaptic Transmission in Hippocampal Cultures Treated with Botulinum Toxin A and C, But Not Tetanus Toxin. J. Neurosci. 17, 7190–7202.

Chen W. R. and Shepherd G. M. (1997) Membrane and synaptic properties of mitral cells in slices of rat olfactory bulb. Brain Res. 745, 189–196.

Chiu W. L. A. K., Sze C. N., Ip L. N., Chan S. K. and Au-Yeung S. C. F. (2001) NTDB: Thermodynamic Database for Nucleic Acids. Nucl. Acids Res. 29, 230–233.

Cicchetti D. V. and Feinstein A. R. (1990) High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. 43, 551–558.

Claiborne B. J., Amaral D. G., and Cowan W. M. (1986) A light and electron microscopy study analysis of the mossy fibers of the rat dentate gyrus. J. Comp. Neurol. 246, 435–458.

Crasto C. J., Marenco L., Miller P. L., and Shepherd G. M. (2002) Olfactory receptor database: a metadata driven automated population from sources of gene and protein sequences. Nucl. Acids Res. 30, 354–360.

Friedman C., Alderson P. O., Austin J. H., Cimino J. J., and Johnson S. B. (1994) A general natural language text processor for clinical radiology. J Am Med. Inform. Assn. 1, 161–174.

Friedman C., Jra P., Yu H., Krauthammer M., and Rzhetsky A. (2001) GENIES: a natural-language processing system for extraction of molecular pathways from journal articles. Bioinformatics. 17, S74–S84.

Hersh W. R., Crabtree M. K., Hickman D. H., et al. (2002) Factors Associated with Success in Searching MEDLINE and Applying Evidence to Answer Clinical Questions. J Am Med Inform Assn. 9, 283–293.

Iliopoulos I., Enright A. J., and Ouzounis C. (2001) TextQuest: Document Clustering of MEDLINE Abstracts for Concept Discovery in Molecular Biology, Pacif. Symp. Biocomp. 6, 374–383.

Justeson J. S. and Katz S. (1995) Technical terminology: some linguistic properties and an algorithm for identification in text. Nat. Lang. Eng. 1, 9–27.

Karp P. D., Riley M., Paley S. M., Pellegrini-Toole A., and Krumenacker M. (1999) EcoCyc: Encyclopedia of Escherichia coli genes and metabolism. Nucl. Acids Res. 27, 55–58.

Kim W., Aronson A. R., and Wilbur W. J. (2001) Automatic MeSH term assignment and quality assessment Proc. Am. Med. Inform. Assn. Symb., Washington DC., pp. 310–323.

Korfhage R. R. (1997) Information Storage and Retrieval, John Wiley and Sons, New York, pp. 105–139, 191–215, 219–231.

Krauthammer M., Rzhetsky A., Morozov P., and Friedman C. (2000) Using BLAST for identifying gene and protein names in journal articles. Gene. 259, 245–252.

Lagus K. (2000) Text mining with the WEBSOM. Acta. Polytech. Scand. Math Comput. 110, 1–54.

Marenco L., Nadkarni P. M., Skoufos E., Shepherd G. M., and Miller P. L. (1999) Neuronal database integration: the SenseLab EAV data model. Proc. Am. Med. Inform. Assn. Symp. Washington DC, 102–106.

Migliore M., Morse T. M., Davison A. P., Marenco L., Shepherd G. M., and Hines M. L. (2003) ModelDB: Making Models Publicly Accessible to Support Computational Neuroscience. Neuroinformatics. 1, 135–140.

Mori K., Nowycky M. C., and Shepherd G. M. (1981) Electrophysiological analysis of mitral cells in the isolated turtle olfactory bulb. J. Physiol. (Lond.). 314, 281–294.

Mutalik P. G., Deshpande A., and Nadkarni P. (1999) Use of General-purpose Negation Detection to Augment Concept Indexing of Medical Documents. J. Am. Med. Inform. Assoc. 8, 598–609.

Nadkarni P. M., Marenco L., Chen R., Skoufos E., Shepherd G. M., and Miller P. L. (1999) Organization of Heterogeneous Scientific Data Using the EAV/CR Representation. J. Am. Med. Inform. Assn. 6, 478–493.

Pinker S. (1994) The Language Instinct, Harper-Collins, London, pp. 177–178.

Prager J. M. (1999) Linguini: Language Indentification for Multilingual Documents. Proc. 32nd Hawaii Int. Sys. 1–11.

Qian J., Colmers W. F., and Saggau P. (1997) Inhibition of Synaptic Transmission by Neuropeptide Y in Rat Hippocampal Area CA1: Modulation of Presynaptic Ca2+ Entry. J Neurosci. 17, 8169-8177.

Raghavan V. V., Jung G. S., and Bolling P. (1989) A critical investigation of recall and precision as measures of retrieval system performance. ACM. Tr. Inform. Sys. 7, 205–229.

Schomburg I., Chang A., and Schomburg D. (2002) BRENDA, enzyme data and metabolic information. Nucleic Acids Res. 30, 47–49.

Shepherd G. M., Mirsky J. S., Healy M. D., et al. (1998) The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data. Trends Neurosci. 21, 460–468.

Spitzer R. and Fleiss J. (1982) A design-independent method for measuring the reliability of psychiatric diagnosis. J. Psychiat. Res. 17, 335–342.

Sun Q.-Q. and Dale N. (1998) Differential inhibition of N and P/Q $Ca2^{+}$ currents by 5HT1A and 5HT1D receptors in spinal neurons of Xenopus larvae. J. Physiol. 510, 103–120.

Tague-Sutcliffe J. (1992) Measuring the informativeness of a retrieval process. Proc. 15th Ann. Intern. ACM SIGIR Conf. Res. Dev. Inform. Retrieval. Denmark. pp. 23–36.

Toth Z., Hollrigel G. S., Gorcs T., and Soltesz, I. (1997) Instantaneous Perturbation of Dentate Interneuronal Networks by a Pressure Wave-Transient Delivered to the Neocortex. J. Neurosci. 17, 8106–8117.

Weeber M., Mork J. and Aronson A. R. (2001) Developing a test collection for biomedical word sense disambiguation. Proc. Am. Med. Inform. Assn. Symp. Washington DC, 746–750.