Semantic Search on Digital Document Repositories based on Text Mining Results

Chikashi Nobata,[1,2] Yutaka Sasaki,[1,2] Naoaki Okazaki,[3]
C.J. Rupp,[1,2] Jun'ichi Tsujii,[1,2,3] Sophia Ananiadou,[1,2]

[1] National Centre for Text Mining, Manchester Interdisciplinary Biocentre
131 Princess Street, Manchester, M1 7DN, UK
[2] School of Computer Science, University of Manchester
Oxford Road, Manchester, M13 9PL, UK
[3] Department of Computer Science, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan

Abstract. In this paper, we describe a search tool, Kleio, for digital document repositories in the biomedicine domain. This tool makes use of semantic metadata to index the document with semantic concepts, for genes, proteins and other biomedical terms. The necessary semantic metadata results from text mining processes, which in turn make use of external databases for this domain. The system offers a combination of textual and metadata searches across MEDLINE, the search functionality is enhanced by leveraging terminology management technologies. It covers the entire MEDLINE abstracts. Named entities (NEs) are annotated on these abstracts in advance and category information can also be used as a part of queries to narrow down the target articles. Query expansion is provided at no extra overhead, by the normalization of biomedical terms. Identifiers from external databases are also included in the indexed data. Notably, the system implements knowledge acquisition in automatically generating its acronym dictionary from the corpus of documents.

1 Introduction

A known and persistent problem with document repositories is getting what you want out of them. Conventional search methods focus on the distribution of keywords and most of the information that repositories themselves associate with a document concern the publishing process. While some users may be well served by a search based on bibliographic metadata, a much more, need is for documents that are about a specific topic. It follows that a search tool focusing on what the document is about would be both more convenient and more intuitive. However, the prerequisite for such a tool is the availability of semantic metadata, *i.e.* information about what the document is about.

The requirement for semantic metadata can be met by making systematic use of text mining results, and, in particular, the results of Named Entity Recognition (NER). This is a process by which the expressions in a document that are clearly about something are located and classified. Typically, this will also include determining the canonical name of the object they refer to. The term NER derives from the original application of such processing to news-wire articles [29] where people, places and companies were the terms with the highest priority. More recently considerable effort has been put into NER in the biomedical domain where it is scientific terms that must be identified and their referents are genes, proteins and other chemical substances.

It is no coincidence that some of the most pressing needs for searching for documents in large repositories are apparent in the biomedicine domain. We describe a tool for accessing a document repository in biomedicine that builds on the text mining results to provide a form of semantic search. This system is dubbed Kleio, after the Greek muse of history. We highlight several attractive properties of a semantic search at the level of concepts, rather than keyword matching. The ability to recognise usage as a particular concept is a primary and demonstrable advantage when compared to just the occurrence of a particular word form. The classification of terms allows the organisation of links between documents in an intuitive manner, so as to aid the incremental refinement of a query and potentially lead to genuine knowledge discovery. Conversely, determining the canonical form of a term, and in particular of acronyms and abbreviations, allows for query expansion to cover all the forms of words that may refer to that concept.

We describe in some detail the text mining processes on which this tool is built. This includes the process by which acronyms and their expansions are learnt from the document collection itself. We also describe the tools used for indexing the semantic metadata and crucially the design of the interface that supports several modes of semantic search.

2 Named Entity Recognition (NER)

The purpose of text mining is to make explicit knowledge encoded in text. This produces semantic metadata associated with documents, as whole, and with expressions and passages within those documents. Subsequent processing can build on that metadata, as in our case a semantic search tool. As such, NER is a primary text mining process. In the sense that it is the most common text mining process, but equally the one that requires the least analysis of the language of the text. Perhaps we could characterise this as “open cast” text mining.

What NER does require is informational resources on which the association between terms in context and semantic categories can be built. This can take two forms, and will be used in slightly different ways. An annotated corpus can be used to learn the association between terms and categories. This is most useful where there is a high degree of ambiguity and the context of use provides vital cues to the correct association. However, it takes time and effort to produce annotated corpora covering all the categories that are of interest to biochemists. On the other hand, databases of technical terms created by domain experts are often available in the biochemical domain. These can function as a resource for the more general case of NER where ambiguity is not as prevalent. Nevertheless, it is essential to make a specific link between annotated NEs and the compounds denoted, because various synonyms can be used to denote the same compound. This can be achieved by assigning unique identifiers like SwissProt IDs [42] to the annotated NEs.

NER has been extensively studied in the biomedical domain, and NER systems for proteins and genes have also been developed by many research groups [39, 17, 8, 13, 15, 20]. There is also research on the creation knowledge resources for NER on chemical names. JNLPBA2004 [9] evaluated NER systems with the GENIA corpus [18]. The GENIA corpus has expanded the target NE categories and defined 47 categories as a hierarchy. Corbett et al. [10] annotated five NE types of chemical names (*i.e.* chemical compound, chemical reaction, chemical adjective, enzyme, chemical prefix) in 42 full-text chemistry papers. They achieved the inter-annotator agreement F-score of 93%. Kulick et al. [23] created corpora in two domains: gene oncology domain and CYP inhibition domain. In the latter domain, CYP450 enzymes, other substances, and quantitative measurements are the target NE categories. Kolárik et al. [19] examined available chemical name dictionaries and also annotated MEDLINE abstracts with six chemical classes, which contain IUPAC(-like) names (names derived from the chemical structure, *e.g.* 1-hexoxy-4-methyl-hexane), partial IUPAC class names (*e.g.* 17beta-), trivial names (commonly used names, *e.g.* aspirin, estrogen), abbreviations and acronyms, chemical family names, and formula/atoms/molecules. Spasić et al. [37] describe a methodology for rapid development of controlled vocabularies. Their approach is to utilise an Information Retrieval (IR) system, automatic term recognition (ATR), and a thesaurus to acquire terms automatically as a practical alternative to both manual term collection and tailor-made NER methods.

When annotated corpora are available, we can obtain context information from them and improve the performance of NER. However, as the target types of relationships or events are broad, different categories will need to be recognised as well as existing NE categories, such as names of disease or experimental methods. It will therefore be helpful if we could perform NER even when a training corpus is not readily available for the target NE categories.

A dictionary-based approach uses existing terminological resources in order to locate term occurrences in text [4]. The approach is to find a term sequence in text that matches an entry in a given dictionary. Spelling variations and term ambiguities of the target entities are major causes of degradation in the performance of the dictionary-based NE. Dictionary-based NER systems which handle these problems combine different methods to improve the performance. Krauthammer et al. [21] use BLAST [2, 3] to identify gene and protein names. Names or English sentences are converted into nucleotide sequences (*e.g.* zgap1 → AGATAAGCAAACACCCAGCG), and approximate string matching is performed on the converted sequences by BLAST. Tsuruoka et al. [41] employ different methods to handle the problem, such as filtering annotation using machine learning with a training corpus, edit-distance operations to allow an approximate term matching, and dictionary expansion with a variant generator.

2.1 NER Method Used for Kleio

Our method of recognizing NEs is based on the system described in [34], which consists of two components. The first part, dictionary-based tagging, finds candidates for entities using a dictionary. The dictionary maps strings to parts-of-speech (POS), whose tag set is a slight extension of the Penn Treebank POS tag set [33]. The NE dictionary is introduced as a subset of noun dictionary, and the NE tag names are given a new POS tag such as NN-PROTEIN.

We use an open-source morphological analyzer Mecab [27] for NER with the additional term lists for this part. In practice, the dictionary-based NER acts as a part-of-speech (POS) tagger. The POS tagger that we used is trained with Conditional Random Fields (CRFs) [24] using the POS information in the GENIA corpus[4]. The Viterbi algorithm is used to find the most probable path of tags for the input sequence. The NER system converts a sentence into all possible sequences of words that are registered in given word dictionaries, and selects the most plausible sequence based on the estimated cost [22]. Therefore, words that match entries in the dictionary are annotated as NEs when they are in the selected sequence. This method makes it possible to annotate NEs with multiple possible sequences of POS taggings, because the system can efficiently handle n-best POS sequences for NER and also avoid errors inherent in a single-best POS sequence when a POS tagger is used as a separate pre-processor.

The second part, statistical sequential labelling, is a supervised method with JNLPBA-2004 training data [9]. The module uses results of dictionary-based NER as well as word, orthographic and POS information as features to predict the NE labels. Word features are the surface form of the word and the postfixes (the last two and four letters of the word). Orthographic features represent the first letter and last four letters of the word, in a normalized form. Upper case letters are mapped to “A”, lower case to “a” and digits to“0”, so that AA-0 may represent the suffix pattern of IL-2.

The NE labels adopts IOB2 format [40], *i.e.* the first token of the target sequence is labeled with “B” of “Beginning” (*e.g.* B-protein), the intermediate and the last tokens in the target sequence are labeled with “I” of “Intermediate” (*e.g.* I-protein) and other tokens are labeled just as “O” of “Others”. For instance, the sequence “dendritic cell-specific transmembrane protein” is annotated as “B-protein I-protein I-protein I-protein”. CRF models are used to predict the IOB2 labels with the above features. For gene and protein names, both first and second part are performed for NER. For metabolites and medical terms, only the first part is performed because currently there is no available training corpus for these terms.

[4] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi

2.2 Dictionary Data

We are currently working on extracting relationships between proteins/metabolites and biomedical terms, therefore additional annotations need to be added to the documents. To annotate biomedical terms other than proteins and genes, we extracted entries from external databases. For metabolites, we used HMDB (The Human Metabolome Database) [43], and for drug names we used DrugBank [44]. For medical terms we extracted several categories from UMLS (Unified Medical Language System) Metathesaurus [6].

DrugBank is:
"a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information. The database contains nearly 4800 drug entries including more than 1,480 FDA-approved small molecule drugs, 128 FDA-approved biotech (protein/peptide) drugs, 71 nutraceuticals and more than 3,200 experimental drugs."[5]

HMDB is:
"a freely available electronic database containing detailed information about small molecule metabolites found in the human body." "The database is designed to contain or link three kinds of data: 1) chemical data, 2) clinical data, and 3) molecular biology/biochemistry data. The database currently contains nearly 2500 metabolite entries including both water-soluble and lipid soluble metabolites as well as metabolites that would be regarded as either abundant (> 1 uM) or relatively rare (< 1 nM)."[6]

UMLS Metathesaurus is:
"a very large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships." "the Metathesaurus is built from the electronic versions of various thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing and cataloging biomedical literature, and/or basic, clinical, and health services research."[7]

The total number of entries in the Metathesaurus we used (2007AC) is about 7.4 million, from which we extracted the following seven categories: Disease, Symptom, Organ, Diagnostic/therapeutic procedure (*e.g.* "MRI", "cerebral blood flow"), Indicator/reagent/diagnostic aid (*e.g.* "hydrogen peroxide", "sulfhydryl reagent"), Phenomenon/process (*e.g.* "UV radiation", "automobile accident"), and Pathologic functions (*e.g.* "hyperventilation", "anaphylactic shock").

Table 1 shows the specifications of terms obtained from databases. The rightmost column shows the average number of synonyms per identifier in each category. The number of synonyms per identifier is large in drug and metabolite names compared to the medical terms.

Figure 1 more clearly shows the differences in the number of synonyms between resources. The x-axis is the number of synonyms and the y-axis is the

[5] http://www.drugbank.ca/

[6] http://www.hmdb.ca/

[7] http://www.nlm.nih.gov/research/umls/about_umls.html

Table 1. Numbers of Entries Extracted from Databases

Category	# Terms	# Identifiers	Ave.
Drugs	26,272	1,200	21.9
Metabolites	48,179	2,966	16.2
Medical Terms	764,529	262,900	2.9
Diseases	301,029	86,624	3.5
Indicators	27,735	12,249	2.3
Organs	143,509	65,794	2.2
Pathologic func.	33,193	10,067	3.3
Phenomena	6,329	2,230	2.8
Procedures	229,934	79,773	2.9
Symptoms	22,800	6,163	3.7

number of identifiers. The distinct number of identifiers in the dictionary is 262,900. For the entries from UMLS, the 35.2% (92,622 / 262,900) of identifiers have only one entry. On the other hand, the number of identifiers that have only one entity is 0 in metabolites, and 20 in drugs. Some of the names of metabolites and drugs have a large number of synonyms, because the synonyms also include chemical names. For example, synonyms of 'Acetaminophen' includes not only its variations (*e.g.* Acetaminofen, Paracetamol), but also its chemical IUPAC name (*i.e.* N-(4-hydroxyphenyl)acetamide) and its chemical formula (*i.e.* C8H9NO2).

2.3 Statistics of the NER Results

We applied NER to all 17 million MEDLINE [28] abstracts. The frequency of the entries are shown in Figure 2. The total number of annotated metabolite/medical terms is more than 72.3 million. The results show that most of the entries found in existing databases do not appear in the MEDLINE abstracts. 713,083 of 838,980 entries (85.0%) in our NE dictionary are not found in the abstracts. On the other hand, the top 80 terms appear more than 100,000 times. The tendency is the same when the frequency is accumulated for each identifier. 214,588 of 267,066 identifiers (80.3%) in our NE dictionary are not found in MEDLINE abstracts, and terms which have one of the top 135 identifiers appear more than 100,000 times.

Most of the entries that are not found in the abstracts contain negligible variations, (*e.g.* word orders are reversed for searching head words), but some of them are systematic chemical names such as IUPAC names or chemical formulae, and also brand names of drugs. Though such names rarely appear in the abstracts, these entries would be beneficial when we apply our NE methods to a large set of full-text papers, which is one of our directions of future work.

In contrast, terms that account for many of the MEDLINE entries are acronyms (such as DNA, ATP, CT) and names that can also be used as general nouns (like 'alcohol', 'water'). Acronyms are often ambiguous, which leads us to introduce a method that uses an acronym dictionary to handle them. These general

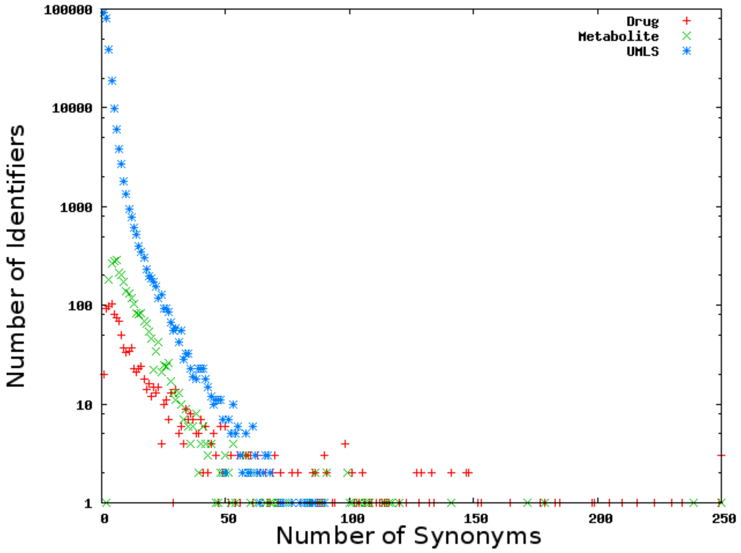


Fig. 1. Number of Synonyms at Each Identifier

nouns are also ambiguous with regard to their usage. They can be related to biochemical pathways, but mostly their usage is more general, *e.g.* as solvents. The disambiguation between these usages is also planned future work. In the following section we describe the acronym handling.

3 Acronym Handling

3.1 Acronym Recognition and Disambiguation

Several studies have been carried out that recognize acronyms and the corresponding long forms (definitions) automatically. These systems use either predefined heuristics/algorithms [1, 5, 35, 38, 45, 46], machine-learning methods [7, 32, 30], or statistics in the source documents [16, 25].

The set of pairs of acronyms and the long forms is created using statistics over the entire collection of MEDLINE abstracts [31]. We utilized a method for recognizing acronym definitions in the abstracts to build an acronym dictionary. The algorithm assumes that parenthetical expressions introduce acronym definitions in the following format:

$$\textit{expanded form} \text{ '(' } \textit{acronym} \text{ ')'} \qquad (1)$$

We regard a parenthetical expression as an acronym if the expression inside the parentheses satisfies these conditions: it consists of, at most, two words; it is between two and ten characters long; it contains at least one alphabetic letter; and the first character is alphanumeric. For each parenthetical expression, the algorithm enumerates candidates for the expanded forms that begin with

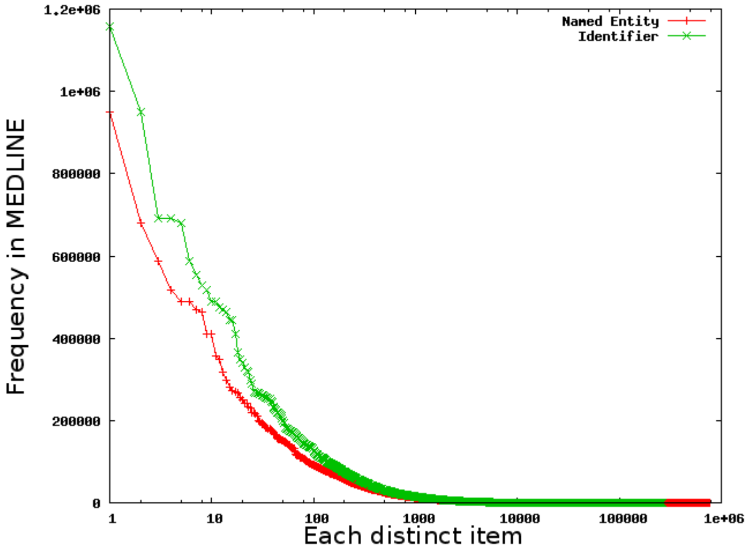


Fig. 2. Frequency of Tagged Entries in MEDLINE

any non-function word and end with the word just before the parenthetical expression. In order to choose correct expanded forms for each acronym a, the algorithm computes a score $\mathrm{LH}_a(c)$ for a candidate of expanded form c:

$$\mathrm{LH}_a(c) = \mathrm{freq}(a,c) - \sum_{t \in T_c} \mathrm{freq}(a,t) \times \frac{\mathrm{freq}(a,t)}{\sum_{t \in T_c} \mathrm{freq}(a,t)}. \tag{2}$$

In this formula, a is an acronym; c is a candidate of expanded form for the acronym a; $\mathrm{freq}(a,c)$ denotes the co-occurrence frequency of the candidate c with the acronym a; and T_c is a set of nested candidates, each of which consists of a preceding word followed by the candidate c. We compile a list of candidates of expanded forms sorted in the descending order of their scores for each acronym. The algorithm takes candidates out of the sorted list one by one. An expanded form is considered valid if: it has a score greater than 2.0; the words in the expanded form can be rearranged so that all alphanumeric letters in the acronym appear in the same order; and it is not nested or an expansion of the previously chosen expanded forms. This method has extracted 886,755 acronym candidates and recognized 300,954 expanded forms, and achieved 99% precision and 82–95% recall on the evaluation corpus.

Acronym recognition is also useful for disambiguating global acronyms (*i.e.* acronyms without their definitions stated explicitly in abstracts) because it provides sense inventories (lists of acronyms definitions), training data (context information of full forms), and local definitions for acronyms. Therefore, we built classifiers for predicting definitions of acronyms by using the context information as training data. Applying the classifiers, we disambiguated the definition of every acronym in the whole of the MEDLINE abstracts.

Table 2. Re-annotation Results on MEDLINE Abstracts

	# terms	# NEs w/o AF	# NEs w/ AF
ALL	1,877,661,325	72,340,088	72,007,172
Acronym	63,941,442	8,457,864	8,124,948

3.2 Re-annotation of Acronyms

The results of acronym disambiguation are used for checking NE results for acronyms. First, acronyms in the NE dictionary are annotated as NEs. Then, once pairs of a long and short form have been obtained from the acronym pair list, the corresponding long form is searched in the dictionary. If the long form is found in the abstract but not in the dictionary, the acronym's annotation is cancelled. Whereas, when an acronym is not annotated but the long form is found in the dictionary, the acronym is annotated as an NE.

For instance, "DEA" is registered in the dictionary as an acronym of "Diaethanolamin", but "DEA" is also used as an acronym of "Data Envelopment Analysis". With the information in the acronym pair list, the system can check whether "DEA" in a MEDLINE abstract is used as "Diaethanolamin" or "Data Envelopment Analysis", and correctly annotate "DEA" as an NE in the former case. The reverse is also true, *i.e.* we can link long forms in the dictionary with acronyms that are only found in the acronym pair list. For instance, "cardiolipin" is often abbreviated as "CL" in the MEDLINE abstracts, but "CL" as "cardiolipin" is not registered as synonyms in the dictionary. In this case CL is additionally annotated using the acronym data.

3.3 Re-annotation Results

We applied our our re-labelling method throughout the MEDLINE abstracts. In this experiment, we used only local acronyms recognition pairs. Table 2 shows the experimental results with or without acronym filtering (AF). The total number of words in the articles are 1,877,661,325 and about 72,340,088 of them (3.85%) are annotated initially, and reduced to 72,007,172 (3.83%) after re-annotation. On the other hand, the number of acronyms are 63,941,442, and about 8,457,864 of them (13.23%) are annotated initially, and reduced to 8,124,948 (12.71%) after re-annotation. The total number of re-annotations is 7,236,864, and 3,784,890 of them are the ones whose annotations are cancelled, and the remaining 3,451,974 entries are newly annotated acronyms. This indicates that even when only local acronyms are used, we can improve 44.75% (3,784,890 / 8,457,864) of annotated acronyms without supervised learning. Our classifier can also produce disambiguation results for global acronyms. Although disambiguation results of global acronyms have more noisy results compared to local acronyms, it will improve the recall of re-labelling of acronyms, and useful for filtering out spurious annotations to frequently used acronyms the definitions, such as CI ("confidence interval"), CT("computed tomography"), SD("standard deviation"). It is one of

Table 3. Examples of Re-annotation Results

Freq.	Acronym	Original tag	Corresponding synonym	Correct long form
113357	NO	-	-	nitric oxide
56067	CT	Metabolite	3beta,5alpha,6beta-cholestanetriol	computed tomography
52510	AD	Drug	actinomycin D	Alzheimer's disease
48338	CSF	Drug	colony-stimulating factor	cerebrospinal fluid
46874	PKC	Disease	paroxysmal kinesigenic choreoathetosis	protein kinase C
35799	NE	-	-	norepinephrine
34796	MR	Disease	mitral regurgitation	magnetic resonance
30328	CI	Pathologic_func	chemically induced	confidence interval
29106	HPV	-	-	human papillomavirus
28986	ROS	Drug	acrosoxacin	reactive oxygen species
23885	MI	Metabolite	myoinositol	myocardial infarction
20773	HCC	Disease	hypomyelination and congenital cataract	hepatocellular carcinoma
18651	RA	-	-	retinoic acid
18458	IOP	Drug	epinephrine	intraocular pressure
17935	BM	-	-	bone marrow
17161	NA	Metabolite	-	noradrenaline
15660	CBF	-	-	cerebral blood flow
12715	AR	Disease	aortic regurgitation	androgen receptor
12220	MAP	Drug	medroxyprogesteron	mean arterial pressure
11509	HD	Disease	Huntington's disease	hemodialysis

our future tasks to introduce global acronym recognition for re-labelling acronym NEs.

Table 3 shows examples of 20 re-annotation results, in order of (descending) frequency. The frequency of the re-labelling for each acronym is listed in the "Freq." column. "Original tag" shows the semantic categories assigned by our dictionary-based NER, and "Corresponding synonym" shows one of the synonyms to show the definition of the acronym in the dictionary. When acronyms are not originally annotated because they are not in the dictionary, these items are left blank.

We can see that acronym information can be used for modifying annotation in an uniform way. For example, the term "AD" is annotated as a drug name because "AD" is registered as an acronym of "Actinomycin D", but the term is actually often used as an acronym for "Alzheimer's disease". Using the fact that the disease name is mentioned in the article, the system correctly re-annotate the acronym. The acronym information also enables it to find additional annotations. For example, "NO" is not included as an acronym of "nitric oxide" in the dictionary, but with the acronym information "NO" is annotated as an NE (metabolite) after the post-processing. In the same way, "NE" as "norepinephrine" (metabolite) and "RA" as "retinoic acid" (drug) are also newly annotated as NEs.

4 The Kleio Interface

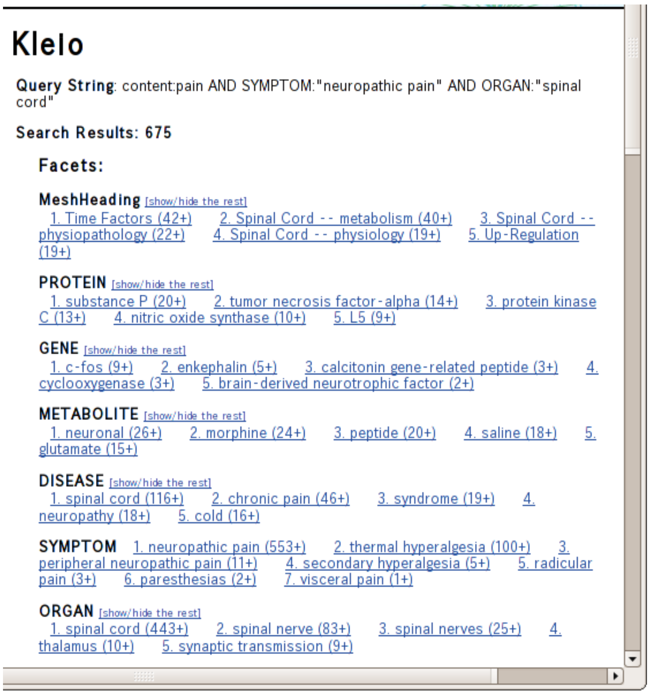
The results of NER based on external dictionaries form the semantic metadata that Kleio uses to provide a range of semantic search functions. First, we employ a standard indexing tool, Lucene [26], to generate an index over the terms for proteins, genes, metabolites and medical terms that have been recognised. This is an index of the concepts that are referred to in the text, rather than individual, or canonical word forms. This means that we can retrieve documents that refer to a specific concept, although the surface form used may differ in each case, as in the use of orthographic variant or acronyms instead of their expansions. We can also base document retrieval on the unique identifier for a concept, providing a link back to the original databases from which the dictionary was generated.

The primary form of a user query can be similar to a search based on word forms, except that the terms in the query are interpreted as a set of concepts. However, the classification of terms into semantic categories allows the user to specify a specific concept, by associating a semantic category with a query term. This can radically reduce the search space. This is one easily demonstrable benefit from basing a search method on text mining results, effectively arising from an initial query. For example, more than 60,000 documents are returned when the word “cat” is given as a query. With semantic annotations, however, the retrieval system can provide a more focused query, because only documents with that semantic category are returned. In fact, the query “PROTEIN:cat” returns about 200 documents with “cat” annotated as a protein.

There are other benefits in what happens next. The list of documents returned by the initial query can be organised according to the set of semantic categories. Effectively, we see these as a set of links between the concepts in the initial query and those occurring in the same immediate context in each document, so for each category we can list the most frequently linked concepts. This forms a faceted interface to the search results, similar to the kind of structured interfaces required for products in an e-commerce application [14]. The structure imposed here reflects the fact that the semantic categories impose on the concept space. Current knowledge resources provide a single level of partitioning into distinct categories. A more complex concept system could be employed, provided that the dictionaries to support it could be made available. A screenshot of the faceted search with Kleio is shown in Figure 3. The faceted search is implemented using Solr [36].

The user may refine the initial query by combining concepts from the faceted interface or may pursue the links to the document representations. The documents themselves are presented with concept markup on all the recognised terms. This markup includes the concept identifiers which provide a direct link to entries in the databases that the dictionaries are derived from. This is a highly desirable functionality for most scientists.

Concept identifiers can also be used as search terms once they have been determined. While query refinement, by combining additional concepts, is an explicit process , query expansion is implicit and falls out naturally from semantic search at the concept level.



Kleio

Query String: content:pain AND SYMPTOM:"neuropathic pain" AND ORGAN:"spinal cord"

Search Results: 675

Facets:

MeshHeading [show/hide the rest]
1. Time Factors (42+) 2. Spinal Cord -- metabolism (40+) 3. Spinal Cord -- physiopathology (22+) 4. Spinal Cord -- physiology (19+) 5. Up-Regulation (19+)

PROTEIN [show/hide the rest]
1. substance P (20+) 2. tumor necrosis factor-alpha (14+) 3. protein kinase C (13+) 4. nitric oxide synthase (10+) 5. L5 (9+)

GENE [show/hide the rest]
1. c-fos (9+) 2. enkephalin (5+) 3. calcitonin gene-related peptide (3+) 4. cyclooxygenase (3+) 5. brain-derived neurotrophic factor (2+)

METABOLITE [show/hide the rest]
1. neuronal (26+) 2. morphine (24+) 3. peptide (20+) 4. saline (18+) 5. glutamate (15+)

DISEASE [show/hide the rest]
1. spinal cord (116+) 2. chronic pain (46+) 3. syndrome (19+) 4. neuropathy (18+) 5. cold (16+)

SYMPTOM 1. neuropathic pain (553+) 2. thermal hyperalgesia (100+) 3. peripheral neuropathic pain (11+) 4. secondary hyperalgesia (5+) 5. radicular pain (3+) 6. paresthesias (2+) 7. visceral pain (1+)

ORGAN [show/hide the rest]
1. spinal cord (443+) 2. spinal nerve (83+) 3. spinal nerves (25+) 4. thalamus (10+) 5. synaptic transmission (9+)

Fig. 3. Screenshot of the Faceted Search with Kleio

5 Concluding Remarks and Future Work

We have presented a semantic search tool based on text mining results that use NER technology to locate and classify scientific terms in the biomedicine domain. The NER processes are dependent on two kinds of knowledge resources: dictionaries of scientific terms and annotated corpora. The dictionaries provide the broad sweep of term recognition and are based on available databases for the domain. Corpora are used for training machine learning algorithms to determine usage in a specific context. They represent a greater investment of human effort, even requiring specialised humans. The limited availability of such corpora, or the resources to build new ones, means that NER based on machine learning is strategically employed in areas where term ambiguity is a major problem. Fortunately, this is effective.

Similarly, the quality of the term classification based on dictionaries depends on the availability of knowledge sources and the information they contain. The Kleio interface imposes a relatively simple structure on the document results, because that is the structure provided by the knowledge resources. While the technology employed could support more complex structures of concepts, the emphasis is on large scale knowledge resources to get an adequate coverage. The more immediate extension of the concept space will be the addition of new NE types, *e.g.* by employing the OSCAR3 [11] term recognition for chemistry to classify chemical terms according to the categories described above. Generating hierarchical categorizations from descriptive keywords as described in [12] will also be applicable to annotated NEs for enriching the system interface.

We are also in the early stages of applying the technology underlying Kleio to the PMC corpus, under the auspices of the UKPMC project. This would extend the NER techniques from abstracts to full papers. This presents a number of challenges, including the fact that full papers will contain considerably more potential links between concepts, so that the notion of a local context will have to be refined. Full papers are also more structured so that some passages, such as results and conclusions, characterise the content of the document far better than others. We also expect the full text of papers to provide a greater resource for the acquisition of acronym expansions, as there will be more space to define acronyms, particularly more global acronyms that may be left as read in an abstract.

Overall, we have presented an alternative search tool that makes use of semantic metadata, information about what the documents are about. We believe that for many users this will be both a convenient and intuitive tool. However, we are accessing large and important document repositories. There will be various sorts of users with various needs. One size will probably not fit all in this context. As Kleio covers the whole of MEDLINE it can address the whole user community, so that its contribution can mainly be determined by user preferences.

6 Acknowledgements

This work has been partially carried out within the UK PubMed Central project, funded by: Arthritis Research Campaign, BBSRC, British Heart Foundation, Cancer Research UK, Chief Scientist Office, Department of Health - National Institute of Health Research, Medical Research Council, Wellcome Trust. The National Centre for Text Mining is funded by the Joint Information Systems Committee. The authors are grateful to John McNaught for his comments.

References

1. E. Adar. SaRAD: A simple and robust abbreviation dictionary. *Bioinformatics*, 20(4):527–533, 2005.
2. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J. Mol. Biol*, 215:403–410, 1990.
3. S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.
4. S. Ananiadou and J. McNaught, editors. *Text Mining for Biology and Biomedicine.* Artech House, Inc., 2006.
5. H. Ao and T. Takagi. ALICE: An algotirhm to extract abbreviations from MEDLINE. *American Medical Informatics Association*, 12(5):576–586, 2005.

6. O. Bodenreider. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32:267–270, 2004.
7. J. T. Chang and H. Schütze. Abbreviations in biomedical text. In S. Ananiadou and J. McNaught, editors, *Text Mining for Biology and Biomedicine*, pages 99–119. Artech House, Inc., 2006.
8. A. Cohen. Unsupervised gene/protein named entity normalization using automatically extracted dictionaries. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, pages 17–24. Association for Computational Linguistics, 2005.
9. N. Collier, P. Ruch, and A. Nazarenko, editors. *JNLPBA-2004*, Geneva, Switzerland, 2004. http://www.genesis.ch/ natlang/JNLPBAO4.
10. P. Corbett, C. Batchelor, and S. Teufel. Annotation of chemical named entities. In *BioNLP2007*, pages 57–64, 2007.
11. P. Corbett and P. Murray-Rust. High-throughput identification of chemistry in life science texts. pages 107–118. 2006.
12. J. Diederich and W. Balke. Automatically created concept graphs using descriptive keywords in the medical domain. *Methods in Information in Medicine*, 47(3):241–250, 2008.
13. J. Finkel, S. Dingare, C. Manning, M. Nissim, B. Alex, and C. Grover. Exploring the boundaries: gene and protein identification in biomedical text. *BMC Bioinformatics, 2005*, 6(Suppl 1), 2005.
14. M. Hearst. Design recommendations for hierarchical faceted search interfaces. *ACM SIGIR Workshop on Faceted Search*, 2006.
15. L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. Overview of BioCreAtIvE: critical assement of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1:S1), 2005.
16. T. Hisamitsu and Y. Niwa. Extracting useful terms from parenthetical expression by combining simple rules and statistical measures: A comparative evaluation of bigram statistics. In Didier Bourigault, Christian Jacquemin, and Marie-C L'Homme, editors, *Recent Advances in Computational Terminology*, pages 209–224. John Benjamins, 2001.
17. J. Kazama, T. Makino, Y. Ohta, and J. Tsujii. Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 Workshop on Natular Language Processing in the Biomedical Domain*, pages 1–8, 2002.
18. J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. GENIA corpus — a semantically annotated corpus for biotextmining. *Bioinformatics*, 19 (Suppl. 1):180–182, 2003.
19. C. Kolárik, R. Klinger, C. M. Friedrich, M. Hoffman-Apitius, and J. Fluck. Chemical names: terminological resources and corpus annotation. In *Proceeding of Building and Evaluation Resources for Biomedical Text Mining*, pages 51–58, 2008.
20. M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirshman, and A. Valencia. Overview of BioCreAtIvE: critical assement of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1:S1), 2006.
21. M. Krauthammer, A. Rzhetsky, P. Morozov, and C. Friedman. Using BLAST for identifying gene and protein names in journal articles. *Gene*, 259(1–2):245–252, 2000.
22. T. Kudo, K. Yamamoto, and Y. Matsumoto. Applying conditional random fields to japanese morphological analysis. In *EMNLP 2004*, pages 230–237, 2004.
23. S. Kulick, A. Bies, M. Liberman, M. Mandel, R. McDonald, M. Palmar, A. Schein, and L. Ungar. Integrated annotation for biomedical information extraction. In *BioLINK2004*, pages 61–68, 2004.

24. J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML2001*, pages 282–289, 2001.
25. H. Liu and C. Friedman. Mining terminological knowledge in large biomedical corpora. In *PSB 2003*, pages 415–426, 2003.
26. Lucene. http://lucene.apache.org/java/docs/, 2006.
27. MeCab. http://mecab.sourceforge.net/, 2008.
28. MEDLINE. http://www.pubmed.gov/, 2007.
29. *Proceedings of the Sixth Message Understanding Conference*, Columbia, MD, USA, 1995. Morgan Kaufmann.
30. D. Nadeau and P. D. Turney. A supervised learning approach to acronym identification. In *8th Canadian Conference on Artificial Intelligence (AI'2005) (LNAI 3501)*, page 10 pages, 2005.
31. N. Okazaki and S. Ananiadou. Building and abbreviation dictionary using a term recognition approach. *Bioinformatics*, 22(24):3089–3095, 2006.
32. S. Pakhomov. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *ACL2002*, pages 160–167, 2002.
33. B. Santrini. Part-of-speech tagging guidelines for the penn treebank project, June 1990. ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz.
34. Y. Sasaki, Y. Tsuruoka, J. McNaught, and S. Ananiadou. How to make the most of NE dictionaries in statistical NER. *BMC Bioinformatics*, 9(Suppl 11:S5), 2008.
35. A. S. Schwartz and M. A. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. In *PSB2003*, pages 451–62, 2003.
36. Solr. http://lucene.apache.org/solr/, 2007.
37. I. Spasić, D. Schober, S.A. Sansone, D. R. Schuhmann, D. B. Kell, and N. W. Paton. Facilitating the development of controlled vocabularies for metabolomics technologie with text mining. *BMC Bioinformatics, 2008*, 9(Suppl 5), 2008.
38. K. Taghva and J. Gilbreth. Recognizing acronyms and their definitions. *International Journal on Document Analysis and Recognition*, 1(4):191–198, 1999.
39. K. Takeuchi and N. Collier. Use of support vector machines in extended named entity recognition. In *CoNLL-2002*, pages 119–125, 2002.
40. E.F. Tjong Kim Sang and J. Veenstra. Representing text chunks. In *EACL-99*, pages 173–179, Bergen, June 1999.
41. Y. Tsuruoka and J. Tsujii. Improving the performance of dictionary-based approaches in protein name recognition. *Journal of Biomedical Informatics*, 37:461–470, 2004.
42. UniProt. http://www.uniprot.org/, 2002–2008.
43. D.S. Wishart and et al. HMDB: The Human Metabolome Database. *Nucleic Acids Research*, 35(Database issue):D521–D526, Jan 2007.
44. D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, 36:D901–906, 2008.
45. J. D. Wren and H. R. Garner. Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods of Information in Medicine*, 41(5):426–434, 2002.
46. H. Yu, G. Hripcsak, and C. Friedman. Mapping abbreviations to full forms in biomedical articles. *Journal of the American Medical Informatics Association*, 9(3):262–272, 2002.