The AMTEx Approach in the Medical Document Indexing and Retrieval Application

Angelos Hliaoutakis, Kaliope Zervanou, Euripides G.M. Petrakis[1]

Department of Electronic and Computer Engineering,
Technical University of Crete (TUC), Chania, Greece
angelos@softnet.tuc.gr, kelly@intelligence.tuc.gr, petrakis@intelligence.tuc.gr

Abstract

AMTEx is a medical document indexing method, specifically designed for the automatic indexing of documents in large medical collections, such as MEDLINE, the premier bibliographic database of the U.S. National Library of Medicine (NLM). AMTEx combines MeSH, the terminological thesaurus resource of NLM, with a well-established method for extraction of terminology, the C/NC-value method. The performance evaluation of two AMTEx configurations is measured against the current state-of-the-art, the MetaMap Transfer (MMTx) method in four experiments, using two types of corpora: a subset of MEDLINE (PMC) full document corpus and a subset of MEDLINE (OHSUMED) abstracts, for each of the indexing and retrieval tasks respectively. The experimental results demonstrate that AMTEx performs better in indexing in 20-50% of the processing time compared to MMTx, while for the retrieval task, AMTEx performs better in the full text (PMC) corpus.

Keywords

document indexing, medical document retrieval, term extraction, MMTx, AMTEx

[1] Corresponding Author

1. Introduction

The availability of large medical online collections, such as MEDLINE [1], poses new challenges to information and knowledge management. MEDLINE constitutes the primary medical repository of the U.S. National Library of Medicine, including over 15 million computer-readable records and is expanding rapidly. It is a rich resource of medical, biological and biomedical information, requiring efficient management and retrieval. MEDLINE documents are currently indexed by human experts, based on a controlled list of indexing terms, deriving from a subset of the UMLS Metathesaurus [2], the MeSH thesaurus [3]. The automatic mapping of biomedical documents to UMLS term concepts has been undertaken by the U.S. National Library of Medicine with the development of MMTx [4].

MMTx was originally developed to improve retrieval of bibliographic material, such as MEDLINE citations [5]. Its applications also include semi-automatic and fully automatic indexing, hierarchical indexing and text mining for various medical and biological concept and relation extraction [5]. The limitations of MMTx in term extraction and in the UMLS Metathesaurus mapping have been analysed in detail in in the studies by Pratt and Yetisgen-Yildiz [6] and Divita et al. [7]. Our experiments in a pilot study of MMTx and AMTEx on a small MEDLINE corpus showed that MMTx performance was low in precision and that its output greatly suffers by over-generating terms, which diffuse the document concept leading to inaccurate indexing of MEDLINE documents [8], [9]. This reflects a design choice in MMTx, which attempts to favour recall by not focusing on MeSH, whereupon MEDLINE indexing has been based, and by incorporating a variant generation process which leads to term over-generation.

In this article, we briefly review the MMTx approach and we present our alternative method, the Automatic MeSH Term Extraction method (AMTEx). AMTEx aims at improving the efficiency of automatic term extraction, using a hybrid linguistic/statistical term extraction method, the C/NC value method [10]. Additionally, AMTEx aims at improving efficiency and accuracy in indexing and retrieval of MEDLINE documents, based on the extraction and mapping of document terms to the MeSH Thesaurus, rather than the full UMLS Metathesaurus mapping of MMTx.

The remainder of this paper first presents related work in the field of term extraction and, in particular, approaches to the extraction of medical terminology for indexing purposes. Subsequently, we present the MMTx resources and processes in more detail, and the resources used in the AMTEx approach, namely the MeSH thesaurus and the C/NC value method for term extraction. Then, the AMTEx approach is presented and, finally, our experiments and results evaluation. We conclude with a discussion on our results and future work.

2. Term Extraction

Term Extraction aims at the identification of linguistic expressions denoting specialised concepts, namely domain or scientific terms. Terms are word or multi-word expressions, which, contrary to general language words, are deliberately created within a scientific or technical linguistic community not only for concept naming purposes, but also for specialised concept distinction and classification purposes [11]. The automatic identification of terms is of particular importance in the context of information management applications, because these linguistic expressions are bound to convey the principal informational content of a document. In early approaches,

terms have been sought for indexing purposes, using mostly *tf · idf* counts [12]. Term extraction approaches largely rely on the identification of term formation patterns (e.g. [13], [14], [15]). Statistical techniques may also be applied to measure the degree of unithood or termhood of the candidate multi-word terms (e.g. [16]). Later and current approaches tend to follow a hybrid approach combining both statistical and linguistic techniques (e.g. [10], [17], [18]).

The extraction of terms for the medical, biological and biomedical domain has greatly motivated research for both indexing, as well as knowledge extraction purposes [15], [19], [20], [21]. In the specific context of term extraction for indexing purposes, the main objective of the term extraction process is the identification of discrete content indicators, namely index terms. A traditional technique for automatic indexing has been the *tf · idf* method [12]. In traditional indexing techniques, query and document representations ignore multi-word and compound terms, which may perform quite efficiently split into isolated single word index terms. However, compound and multi-word terms are very common in the biomedical domain [17] and are often used in indexing medical documents. Multi-word terms carry important classificatory content information, since they comprise of modifiers denoting a specialisation of the more general single-word, head term [14]. For example, the compound term *"heart disease"* denotes a specific type of disease. A study by Milios et al. [22] of the extraction of multi-word terms for retrieval purposes shows that multi-word term methods may complement other methods to improve results. Currently machine learning techniques are also applied for indexing, such as the Naïve Bayes learning model implemented in the KEA (Automatic Keyphrase Extraction, [23]). Comparative experiments of *tf · idf*, KEA and the C/NC value term extraction

methods by Zhang et al. [24] show that C/NC value significantly outperforms both *tf · idf* and KEA in a narrative text classification task using the extracted terms.

3. Background

3.1 The MMTx Approach and Resources

The MMTx approach uses the UMLS Metathesaurus® and the UMLS SPECIALIST^{TM} lexicon as its lexicographic resources. In this section we first briefly present the structure of UMLS and the limitations related to its design and content. Then we present an outline of the MMTx approach.

The UMLS Medical Knowledge Resource

The *Unified Medical Language System (UMLS)* is a source of medical knowledge developed and maintained by the U.S. National Library of Medicine. UMLS consists of the Metathesaurus, the Semantic Network and the SPECIALIST lexicon.

The *Metathesaurus*^{TM} is a large, multi-purpose, and multi-lingual vocabulary database. It integrates about 800,000 concepts from 50 families of vocabularies. In the Metathesaurus, equivalent terms are clustered into unique concepts. Each concept is an abstract representation of the term phrases which are considered as synonymous in the medical domain. Thus, each concept is linked to its respective term variants, i.e. graphical and lexical variants, and in some cases translations into other languages. However, the terms integrated in the Metathesaurus do not all share a common structure, i.e. same properties and characteristics; they inherit the organisational principles governing their respective source vocabularies. Moreover, certain types of relationships, including synonymy and hierarchical relationships, are not defined.

Thus, the Metathesaurus on its own does not have a hierarchical structure, neither fulfils ontological requirements.

The *Semantic Network* consists of 134 semantic types categorising the Metathesaurus concepts. The purpose of the Semantic Network is to provide a consistent categorisation of all concepts represented in the Metathesaurus and a set of useful relationships among these concepts. Every concept in the Metathesaurus™ is assigned to at least one semantic type in the Semantic Network. Two high semantic level hierarchies are defined, one for entities related to pathology, and one for events (treatment for diseases). The Semantic Network may be viewed as an upper level ontology of the biomedical domain. In this perspective, the Metathesaurus entities constitute the properties of the semantic network concepts (i.e. they can be inherited by concepts related by an IS-A relationship). Thus, the Semantic Network of UMLS provides a basis for an ontology of the biomedical domain. Nevertheless, the Semantic Network was not originally designed as an ontology. Problems inherent in the design of the Semantic Network include, among others, circular hierarchical relationships, inconsistencies in the categorisation of concepts and discrepancies between the semantic structure of the Metathesaurus and the Semantic Network. Moreover, the lack of relationships between concepts in the Metathesaurus and the Semantic Network has been also observed.

Finally, the *SPECIALIST lexicon* is intended to be a general English lexicon which includes many medical and biomedical terms. The lexicon entry for each word or term records the syntactic, morphological and orthographic information of the respective lemma.

The MMTx Approach

MMTx uses the Metathesaurus™ and SPECIALIST lexicon knowledge resources during the term extraction process. This process maps arbitrary text to Metathesaurus term concepts and works in the following steps [5]:

Parsing: The document text is parsed, using the Xerox part-of-speech tagger and the SPECIALIST minimal commitment parser to perform a shallow syntactic analysis of the text. A simple linguistic filter of the form *(Adj | Noun)+ Noun* isolates noun phrases [25]. The SPECIALIST parser provides information on the internal syntactic structure of the noun phrase, identifying the head and modifier components of the phrase. For example, the term *"ocular complications"* is analysed as:

```
[mod(ocular), head(complications)]
```

where *complications* is the head, namely the term that is being modified/specialised and *ocular* is the modifier, namely the concept specialising the term *complications*.

Variant Generation: Variant generation is performed in iterative manner. First, the multi-word term phrase is split into generators. A variant generator is considered any meaningful subsequence of words in the phrase. That is either a single word or a term existing in the SPECIALIST lexicon [26]. For example, the term *"liquid crystal thermography"* would be split into the generators: *"liquid crystal thermography"*, *"liquid crystal"*, *"liquid"*, *"crystal"* and *"thermography"* [25]. In the second phase, for each of the generators, all possible semantic (synonyms, acronyms and abbreviations) and derivational variants are identified using the SPECIALIST lexicon and a supplementary database of synonyms. At this stage, please note that, although we have started the process of variant generation of a noun phrase, we may have

derivational and semantic variants belonging to other parts-of-speech, such as verbs. All these variants are in turn used as generators and their respective variants are recomputed. Finally, inflectional and spelling variants are generated based on all word-forms found in the previous processes.

Candidate Retrieval: At this stage, the candidate set of all Metathesaurus term mappings is retrieved. The main criterion of the retrieval is that the Metathesaurus term string should contain at least one of the variants found during the variant generation process [27]. The mapping process may vary [5]. We may have:

- *simple match*, where, for example, *"intensive care unit"* maps to *"Intensive Care Units"*;
- *complex match*, where *"intensive care medicine"* maps to *"Intensive Care and Medicine"*;
- *partial match-gapped*, where *"ambulatory monitoring"* maps to *"Ambulatory Cardiac Monitoring"*;
- *normal and overmatch*, where *"application"* maps to *"Job Application"*, *"Heat/Cold Application"* and *"Medical Informatics Application"*.

The normal partial match is assumed as a good matching for correctness, where at least one word of either the noun phrase or the Metathesaurus string (or both) does not participate in the matching (e.g. *"liquid crystal thermography"* maps to *"Thermography"*, where the mapping does not involve *"liquid crystal"*.

Candidate Evaluation: The candidate set of Metathesaurus mappings is evaluated. The evaluation process computes the mapping strength between the candidate

Metathesaurus string and the text string. The mapping strength weight is calculated by a linguistically principled function consisting of a weighted average of four criteria [28]:

i) *Centrality* indicates whether the Metathesaurus string involves the head of the text phrase and its value is 1 (yes) or 0 (no);

ii) *Variation* is the distance score between the phrase and its variants (this is computed during variant generation);

iii) *Coverage* denotes the length of the text phrase and the Metathesaurus candidate string participating in the match.

iv) *Cohesiveness* is similar to coverage and denotes the non-intermittent words of the text phrase and the Metathesaurus term participating in the match.

The weight for the last two criteria, coverage and cohesiveness, is doubled in the scoring function and their measures are normalised to a value between 0 and 1,000.

3.2 The AMTEx Method Resources

The C/NC-Value Method for Term Extraction

The C/NC value method [10] is a hybrid method for term extraction. C/NC value is domain-independent and combines statistical and linguistic information for the extraction of multi-word and nested terms. In this method, the text is first tokenised and tagged by a part-of-speech tagger. Subsequently, a set of rules and linguistic filters is used to identify in text candidate term phrases. The three filters available are:

N+ N

(Adj | N)+ N

((Adj | N)+|((Adj | N)(N P)?)(Adj | N)*) N*

where *N* is a noun, *Adj* is an adjective and *P* stands for a preposition. Obviously, the linguistic filters used have an impact on the precision and recall of the system. Using a rather closed filter, such as the first one, will result in increased precision and decreased recall, whereas an open filter, such as the last one will increase recall and decrease precision [10]. The current implementation of C/NC value in our approach uses all three linguistic filters. The generated list of candidate noun phrases is then filtered through a stoplist.

The statistical part defining the termhood of the candidate phrases aims to get more accurate terms than those obtained by the pure frequency of occurrence method, especially terms that may appear as nested within longer terms, such as the term *"enzyme inhibitors"* nested in *"Angiotensin-converting enzyme inhibitors"*. The measurement used for this estimation is C-value. C-value is defined as the relation of the cumulative frequency of occurrence of a word sequence in the text, with the frequency of occurrence of this sequence as part of larger proposed terms in the same text. Depending on whether the term is nested or not C-value is defined as:

$$C-value = \begin{cases} \log_2|a| \times f(a), & a\ is\ a\ non-nested\ term, \\ \log_2|a| \times (f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b)), & a\ is\ a\ nested\ term. \end{cases}$$

Equ. 1

In the above, the first C-value measurement is for non-nested terms and the second for nested terms, where *a* denotes the word sequence that is proposed as a term, $|a|$ is the

length of this term in words, *f(a)* is the frequency of occurrence of this term in the corpus (both as an independent term and as a nested term within larger terms), T_a denotes the set of extracted terms that contain a and $P(T_a)$ is the number of these terms. The C-value algorithm produces a list of proposed terms ranked with decreasing term likelihood. The NC-value takes into account the context of each term and assigns weights to specific verbs, adjectives and nouns that appear in candidate term context. The weight factor of a context word *w* is higher for the respective words that tend to appear with terms and is computed as

$$weight(w) = \frac{t(w)}{n},$$

Equ. 2

where *t(w)* is the number of terms the word *w* appears with and *n* is the number of all terms. Finally, the NC-value is defined by

$$NC - value(a) = 0.8 \times C - value(a) + 0.2 \times CF(a).$$

Equ. 3

Here, *a* is the proposed term, *C-value(a)* is calculated as shown in (Equ. 1), and *CF(a)* is computed as

$$CF(a) = \sum_{w \in C_a} f_a\,(w) \times weight(w),$$

Equ. 4

where C_a is the set of context words of term *a*, *w* is a context word in C_a, *weight(w)* is the weight of *w* and $f_a(w)$ is its frequency as context word of *a*.

C/NC-value has been successfully tested in various domains, such as molecular biology (nuclear receptors [29]), eye pathology medical records [10], biomedical business newswire texts [21] and computer science papers [22].

The MeSH Thesaurus

The MeSH Thesaurus (Medical Subject Headings) is a taxonomy of medical and biological terms and concepts suggested by the U.S National Library of Medicine. The MeSH terms are organized in IS-A hierarchies, where more general terms, such as *"chemicals and drugs"*, appear in higher levels than more specific terms, such as *"aspirin"*. MeSH is organised in 15 taxonomies, including more than 22,000 terms. A term may appear in more than one taxonomy. Each MeSH term is described by several properties, the most important being:

1. *MeSH Heading (MH)*: the term name or identifier;

2. *Scope Note (SN)*: a text description of the term;

3. *Entry Terms (ET)*: mostly synonym terms to the MH.

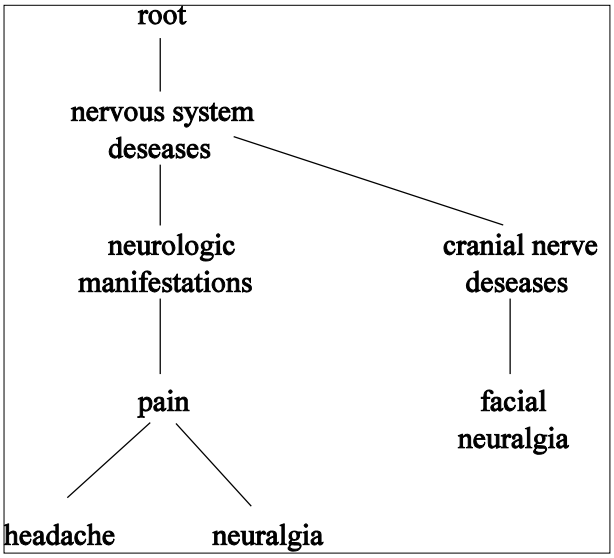

Figure 1: A fragment of the MeSH IS-A hierarchy

Entry Terms (ET) also include stemmed MH terms and are sometimes referred to as quasi-synonyms (they are not always exact synonyms). In our AMTEx approach, all *ET* terms are treated as synonyms. Each MeSH term is also characterised by its MeSH tree number (or code name), indicating the exact position of the term in the MeSH tree taxonomy, for example `"D01,029"` is the code name of term *"Chemical and drugs"*. A fragment of the MeSH IS-A hierarchy is illustrated in Figure 1.

4. The AMTEx method

Based on the study of the MMTx algorithm and resources, we observe the following:

- During the variant generation stage, the iterative expansion of the initial text phrase to all possible variants is quite exhaustive. MMTx extracts term variants, not only based on the terms found in the original text phrase, but also from their variant terms. This is due to an obvious attempt to increase recall of Metathesaurus mappings, a known limitation of MMTx as discussed in [7]. However, this process also results in term over-generation and increased term ambiguity, which diffuse the original term concept, leading to inaccurate indexing.

- MMTx extracts general Metathesaurus terms, not MeSH terms. Although MMTx was originally developed to improve retrieval of bibliographic material, such as MEDLINE citations [5], MMTx mappings were not based on the MeSH Thesaurus, which contains the controlled list of MEDLINE indexing terms. This design option broadens the application domain of MMTx, but it also affects its accuracy in the MEDLINE indexing task, as shown in our experiments in section 5.

- Term selection is based on a scoring function, for evaluating the importance of all candidate terms, using the SPECIALIST lexicon as an external lexical resource. Moreover, the scoring function, though partly based on valid linguistic principles, such as the centrality criterion, it is arbitrarily and empirically defined, making it possible for unrelated terms to be included in the list of extracted terms. The C/NC-value scoring functions are especially tuned to multi-word terms, taking into consideration nested terms and term context words. Additionally, C/NC-value has been proven to extract up to 98% of correct terms [29], [10], [21], [22] in various application domains. Finally, WordNet and MeSH can be used as additional lexical resources, if needed, for both general and medical terms.

Based on the above observations we propose two basic changes towards the development of an improved term extraction method that could substitute MMTx:

i) Term extraction based on a well-established method, the C/NC-value method;

ii) Use of MeSH Thesaurus as lexical resource, both for (limited) term variant retrieval, and candidate term mapping.

Input: Document *d*, MeSH Thesaurus.
Output: MeSH terms *t*.
1. *Multi-word Term Extraction:* C/NC-value method
2. *Term Ranking:* NC-value ranking (Equ. 3)
3. *Term Mapping:* Only MeSH terms are retained.
4. *Single-word Term Extraction:* Single-word MeSH terms are added.
5. *Term Variants:* Stemmed terms are added.
6. *Term expansion:* Semantically similar terms from MeSH

Table 1: AMTEx Algorithm

4.1 The AMTEx Algorithm

An outline of the AMTEx procedure is illustrated in Table 1. In particular, the AMTEx method has the following processing stages:

1. *Multi-word Term Extraction:* The C/NC-value method is used for term extraction. This method is domain independent, does not require any lexical resources and has been proven to be particularly effective in multi-word and nested term extraction both in medical and general document collections. During term extraction in AMTEx the document text is annotated for part-of-speech, using the C/NC-value implementation part-of-speech tagger and linguistic filters.

2. *Term Ranking:* Extracted candidate terms are ordered, first by C-value and subsequently by NC-value score. The final candidate term list is ranked by decreasing term likelihood (Equ. 3). Top ranked terms are more important than terms ranked lower in the list and are more likely to be included in the final list of extracted terms.

3. *Term Mapping:* Candidate terms are mapped to terms of the MeSH Thesaurus (by applying simple string matching). The list of terms now contains only MeSH terms.

4. *Single-word Term Extraction:* For the multi-word terms which do not fully match MeSH, their single word constituents are used for matching. If mapped to a single word MeSH term, this is also added to the candidate term list, retaining its original C/NC ranking value.

5. *Term Variants:* Term variants are included in the candidate term list. The C/NC-value implementation in *AMTEx* includes inflectional variants of the extracted terms. Also, MeSH itself can be used for locating variant terms, based on the MeSH term, Entry Terms property. However, only the stemmed term-forms are used in *AMTEx* since the full list of Entry Terms may contain terms, which often are not synonymous.

6. *Term Expansion:* The list of terms is augmented with semantically (conceptually) similar terms from MeSH. Figure 2 illustrates this process: A term is represented by its MeSH tree hierarchy. The neighbourhood of the term is examined and all terms with similarity greater than threshold $T_{Expansion}$ are also included in the query vector. This expansion may include terms more than one level higher or lower than the original term depending on the value of $T_{Expansion}$.

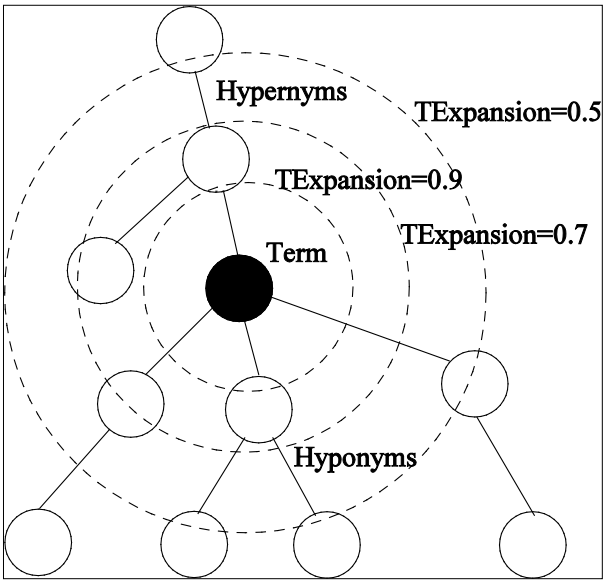

Figure 2: Term expansion thresholds using MeSH

AMTEx in its current state does not include a syntactic parser, such as the SPECIALIST minimal commitment parser used in MMTx. This is due to the fact that AMTEx uses an alternative, well established method for term extraction, the C/NC-value, which relies on linguistic filtering rules and where the head/modifier information is indirectly inferred through the statistical measures, namely the nested term estimations. In AMTEx v2 presented here, the estimated head of a multi-word term is successfully used for the refinement of the Single-word Term Extraction process.

Our approach to *Term Variant* generation is more limited than MMTx. This constrains our term recall to terms that are closer to the original term in text. As we observe in the results of our experiments in section 5, we manage to achieve better precision in a fraction of the processing time taken for MMTx. This is partly due to the fact that our term extraction method outperforms MMTx in suggesting candidate terms. It is also due to the fact the AMTEx approach to variant generation is limited to MeSH and does not operate iteratively, generating variants out of already found variants, thus avoiding the diffusion of the original concept to unrelated concepts.

In *Term Expansion*, the method used in AMTEx for discovering semantically similar terms, is based on the semantic similarity method by Li et al. [30]. The evaluation of the semantic similarity methods indicated that this method is particularly effective, achieving up to 73% correlation with results obtained by humans [34]. An important observation and a desirable property of this method is that it tends to assign higher similarity to terms which are close together (in terms of path length) and lower in the hierarchy (more specific terms), than to terms which are equally close together but higher in the hierarchy (more general terms). Therefore, expanding with threshold

$T_{Expansion}$ will introduce new terms depending also on the position of the terms in the taxonomy: More specific terms (lower in the taxonomy) are more likely to expand than more general terms (higher in the taxonomy). Figure 2 illustrates this process for various values of the threshold $T_{Expansion.}$

Because no synonymy relation is defined in MeSH, we did not apply expansion to the Entry Terms of terms. Word sense disambiguation [31] can also be applied for detecting the correct sense to expand (here, expansion is applied to the most common sense of each term).

4.2 Refining the AMTEx Method

In order to determine the optimal set of indexing terms, namely one increasing recall and precision, there exist three thresholds in the AMTEx process that could be refined:

i) *C-Value threshold* (T_{Cvalue}) for the term extraction, which in our initial experiments presented in [8] was set to its recommended value ($T_{Cvalue} = 1.5$) to limit output to the most valid terms;

ii) *Term expansion threshold* ($T_{Expansion}$), whereupon we have experimented in our pilot small scale experiments with AMTEx [8] ;

iii) *Final list threshold* ($T_{FinalList}$), which determines the minimum value a mapped to MeSH candidate index term must have to be included in the final index term list. In our experiments presented in [8], all candidate terms were retained.

The optimal value for each of these thresholds is not easy to determine, as each of these affects term recall at different stages of the AMTEx process [8]. A simple approach to this optimisation problem would be to consider only the threshold applied at the end of the process, the $T_{FinalList}$. Moreover, precision or recall alone should not determine an optimal threshold, since an increase in precision for example, simultaneously affects recall. A balanced measure such as an F-measure, where recall and precision are equally weighted (shown on Equ. 5 below), would provide us a better indicator for our final threshold.

$$F = \frac{2 \times precision \times recall}{precision + recall}.$$

Equ. 5

Thus, in our AMTEx v2, we have chosen to be exhaustive with both T_{Cvalue} (i.e. $T_{Cvalue}=0$) and $T_{Expansion}$ (i.e. $T_{Expansion}=0.5$) thresholds and use the maximum F-measure to determine the $T_{FinalList}$. Moreover, in the Term Expansion step, the semantic similar terms ($T_{Expansion}=0.5$) added to the candidate list are assigned a weight, as shown on Equ. 6 below:

$$weight(w) = sim \times weight(s),$$

Equ. 6

where a term *w*, semantically similar to term *s*, has ranking weight, *weight(w)*, combining its semantically similar term weight, *weight(s)*, and the similarity value, *sim*, by which *w* is similar to *s*. In this way, in AMTEx v2 the final candidate list ranks accordingly terms which are added to it by the Term Expansion process. In AMTEx v1, these terms were merely assigned the *weight(s*) of Equ. 2.

In our pilot experiments with AMTEx v1 [8], in the *Single-word Term Extraction* step, we were attempting to find partial matches in MeSH, for all word constituents of an unmatched multi-word term. We have observed that single term insertion in our candidate list through that process produced worse results. In our AMTEx v2, we have chosen to conceptually limit our search for single-word mappings using only the head word of the multi-word term. The experiments presented in section 5 of this paper show that this type of *Single-word Term Extraction* slightly improves both recall and precision. Regarding ranking weight for these terms, we consider it equal to its source, i.e. the original multi-word term weight.

5. Experiments and evaluation

5.1 Developing AMTEx v2

Defining $T_{FinalList}$ threshold for AMTEX v2

In order to determine the $T_{FinalList}$, we have experimented with a corpus of 5,819 full PMC documents selected out of 60 Journals. The documents were selected on the basis of having an UID number, which we used to retrieve their respective MEDLINE index sets. This index set for each document is manually assigned by MEDLINE experts and is used in our experiment as our ground truth. In our evaluation, precision is the percentage of correctly retrieved terms compared to the total number of retrieved terms, and recall is the percentage of correctly extracted terms compared to the MeSH terms appearing in the respective MEDLINE document index. In this experiment F-measure of equally weighted precision and recall is used, as shown on Equ. 5, illustrated above.

In AMTEx v2, as discussed in section 4.2, we attempted to modify the *Single-word Term Extraction* process, using only the *head* term constituent for MeSH mapping. Nevertheless, we needed to ascertain that the single-word term extraction step significantly contributes to AMTEx performance, rather than unnecessarily complicating the AMTEx algorithm. Thus, we conducted a second experiment on the same dataset, where the single-word term extraction step was not included in the process. The comparative results in Figure 3 show clearly that *Single-word Term Extraction* improves AMTEx performance.

The peak of a curve in Figure 3 indicates the optimal F-measure performance for our corpus. We observe that the optimal F-measure performance is reached before the 20th point of a curve. Thus, in AMTEx v2, the $T_{FinalList}$ is set to the 20 top terms in the list.

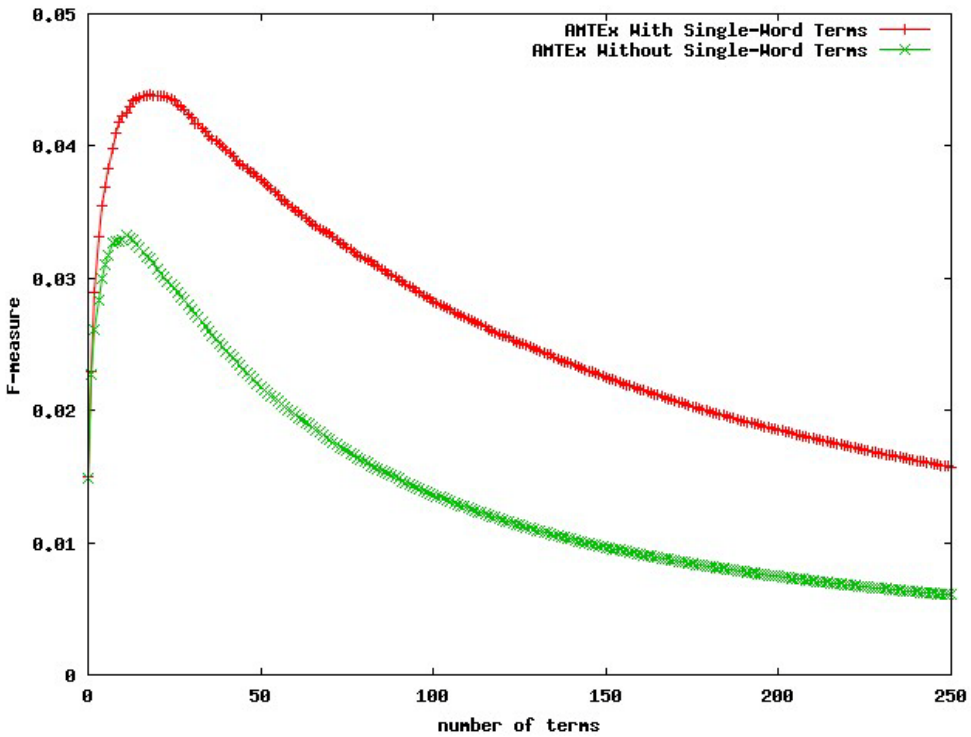

Figure 3: AMTEx with/without single-word extraction and $T_{FinalList}$ threshold in PMC dataset

5.2 MMTx vs AMTEx Method

In our pilot experiments presented in [8], we have compared our first AMTEx version performance to MMTx, which is considered the benchmark method, using a small set of 61 full documents. In this paper, we present a series of comparative experiments we conducted to test our approach in:

- a significantly larger corpus of full documents,
- a corpus of document abstracts,
- using both versions of AMTEx, v1 and v2,
- for indexing and retrieval tasks,
- against MMTx, v24B.

For this reason, we conducted four experiments, comparing AMTEx v1 and v2 to MMTx v2.4B: the first two experiments assess the performance of AMTEx vs. MMTx in the indexing task on a corpus of document abstracts (*Abstract Indexing experiment*) and on a large dataset of full documents (*Full Document Indexing experiment*). The other two experiments compare the performance of AMTEx vs. MMTx v2.4B in the retrieval task using again the respective document abstract (*Abstract-based Retrieval experiment*) and full document (*Full Document-based Retrieval experiment*) datasets.

We should note that in MMTx term ranking is less rigorous than AMTEx. In MMTx valid term output has mostly a weight value of 1000, whereas in AMTEx each term is ranked based on its individual weight. Thus, the evaluation score value of the 10th or 100th best answer of MMTx is not particularly adequate, since all its results may be

equally weighted. This fact makes hard any controlling processes for the over-generated extracted MMTx terms.

Please also note that for our indexing experiments, we thought it to be fair for MMTx to restrict its term mapping process to MeSH, rather the full UMLS, similarly to our AMTEx, since our ground truth consists of the MEDLINE provided index sets, which are based on MeSH.

Abstract Indexing experiment

This first experiment was conducted to test the performance of the three systems in the indexing task in a document abstracts corpus. The problems related to processing document abstracts were first identified in our pilot experiments with AMTEx [8]. These relate to the abstract size, which is quite limited to be used as input to a method using statistics, such as AMTEx. Moreover, the content of the abstract has not been found to contain all necessary textual information for accurately indexing the full document. We have concluded at the time that we needed to consolidate our AMTEx approach before embarking on such an experiment.

For the *Abstract Indexing experiment* presented here, we selected a corpus subset of the OHSUMED standard TREC collection corpus [32]. OHSUMED is a collection of MEDLINE document abstracts used for benchmarking information retrieval systems evaluation. Our selected subset consisted of 10% of OHSUMED, i.e. 30,000 document abstracts (because MMTx is slow, processing of the entire OHSUMED was not feasible). These were again evaluated in terms of precision and recall against the MEDLINE provided MeSH index term sets.

For processing of document abstracts, AMTEx algorithm was slightly modified to respond to the problems of document limited size and content that we have identified. Thus, both AMTEx versions first treat the totality of the corpus as a single document input during the term extraction step. Subsequently the extracted terms are associated to their respective source document by string matching. This modification of the AMTEx process has been thought necessary, since AMTEx term extraction is not only linguistic but also statistically based.

Table 2 demonstrates the comparative performance of AMTEx v1 and v2 against MMTx v2.4B in terms of average document precision and recall. We observe that AMTEx shows improved precision compared to MMTx, and a reasonable recall by merely a fifth of the average term output compared to MMTx.

OHSUMED Dataset	AMTEx v1.0	AMTEx v2.0	MMTx 2.4B
Average number of terms in abstracts	8	8	40
Precision	0.124	0.125	0.089
Recall	0.101	0.101	0.336

Table 2: AMTEx vs. MMTx performance on the OHSUMED data set

Full Document Indexing experiment

In our second experiment we have assessed the performances of our two versions of AMTEx against the MMTx v.2.4B in the indexing task using a full document dataset, our 5,819 PMC full document corpus. The results were evaluated for precision and recall, against our ground truth, i.e. the MEDLINE document index set (assigned manually by the experts). All methods process single document input during the term extraction step.

The results in Table 3 show average term output, precision and recall for each document, for all three systems. We observe that AMTEx v1, shows a precision result that is higher than MMTx, whereas the average extracted terms are much less. AMTEx v2 demonstrates the best recall of the two AMTEx systems, for a fraction of the average MMTx term output.

PMC Dataset	AMTEx v1.0	AMTEx v2.0	MMTx 2.4B
Average number of terms in documents	16	25	72
Precision	0.052	0.034	0.033
Recall	0.054	0.062	0.162

Table 3: AMTEx vs. MMTx performance on the PMC data set

Finally, Table 4 illustrates the comparative results of all systems, in both full document PMC and OHSUMED document abstracts indexing experiments in terms of time efficiency. We observe that the time taken for OHSUMED processing was longer in all systems. Nevertheless, both AMTEx systems are shown to perform much faster than MMTx. We believe that this is due to the algorithmic simplicity of AMTEx compared to MMTx especially with regards to variant generation and term expansion processes (even though MMTx was tested using MeSH rather than the full UMLS).

Time Intervals	AMTEx v1.0	AMTEx v2.0	MMTx 2.4B
PMC Dataset	1721.4	4994.6	9819.5
OHSUMED Dataset	9161.9	26582.5	52261.8

Table 4: Time intervals (in seconds) of AMTEx and MMTx for PMC & OHSUMED data set

Abstract- based Retrieval experiment

In our third experiment we attempted to test AMTEx performance in the medical document retrieval task based on the document abstracts dataset. Documents are represented by term vectors produced by AMTEx (v2.0) and MMTx respectively. Document matching is performed by Vector Space Model (VSM, [34]). Both methods (i.e., retrieval by AMTEx and MMTx vectors) are compared against retrieval using vectors of MEDLINE provided index term sets, i.e. the terms used as ground truth in our indexing experiments. We have used the OHSUMED standard TREC collection corpus subset used in the indexing experiment. However, for this task the results were evaluated against 64 TREC provided queries and answers [32]. These constituted our ground truth for all systems performance.

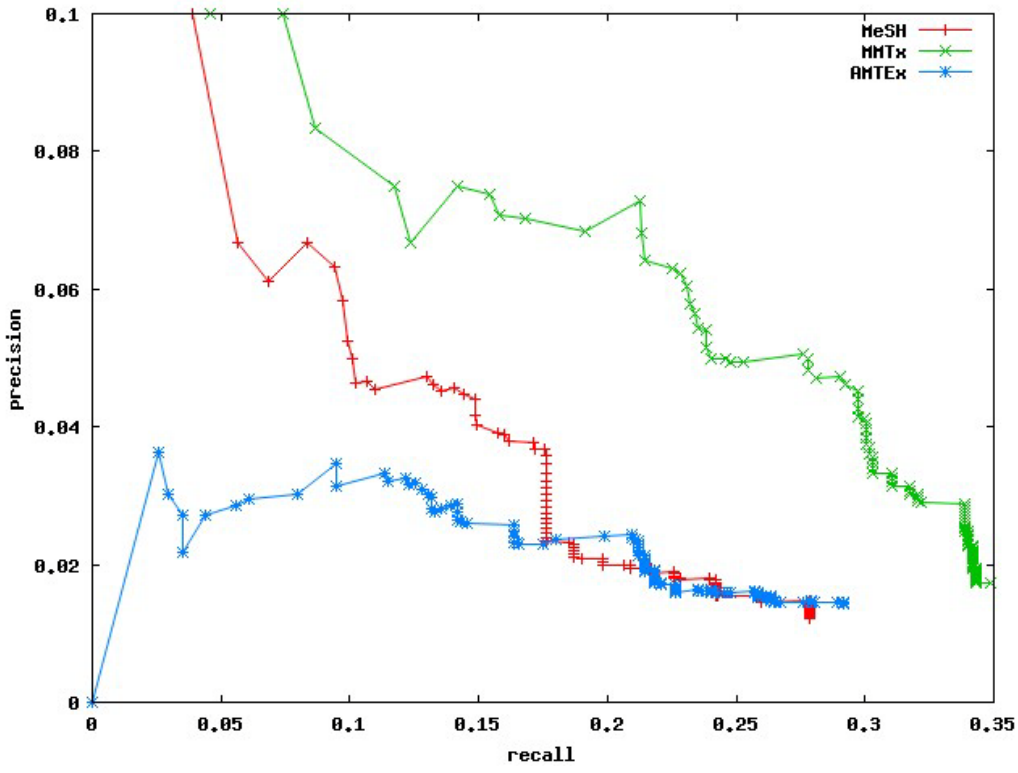

Figure 4: Precision/Recall of AMTEx vs. MMTx on OHSUMED dataset retrieval task

Figure 4 illustrates the performance of AMTEx v2.0 compared to MMTx and the MEDLINE provided index term sets. Each method is represented by a precision/recall curve. For each query, the best 100 answers were retrieved. Precision and recall values are computed after each answer (from 1 to 100) and therefore, each curve contains exactly 100 points. Each point in a curve is the average precision/recall over 64 queries. The top-left point of a curve corresponds to the average precision/recall values for the best answer or best match (which has rank 1), while the right-most point corresponds to the average precision/recall values for the entire answer set.

We observe that for this retrieval task based on the OHSUMED document abstracts dataset AMTEx approaches the performance of the manually assigned MeSH terms as we gradually reach the entire answer set, while the increased recall of MMTx results in significantly better precision than both the manually assigned MeSH terms and AMTEx.

The poor performance of *AMTEx* is due on the combined effect of two reasons. First, given the nature of the corpus, namely the OHSUMED document abstracts, our method AMTEx method due to its statistical part for term extraction, was slightly modified to treat the whole OHSUMED collection as a single document, rather than processing the very small individual document abstracts. The term results of this process were subsequently mapped to individual documents. At this stage the MMTx has the advantage of extracting few terms, even for small document size, which can be subsequently expanded, thus increasing MMTx term recall.

Secondly, this effect is further supported in retrieval, due to the nature of the Vector Space Model (VSM) [33] in document matching. In particular, document matching relies on comparison of term vectors and in VSM partial matching is supported, i.e.

for two documents to be similar the terms of one vector may be a subset of the terms of another vector. Thus, VSM clearly favours representation with many terms, without any regard to excessive terms, while AMTEx output incorporates semantic similarity of terms from the 6th step, not suitable for strictly string matching retrieval results.

As we shall see in our fourth and last experiment, using the PMC full document dataset the combined effect of these two factors is overcome and AMTEx performs clearly better when a full document rather than a document abstract is provided.

Full Document-based Retrieval experiment

In the fourth and last full document retrieval experiment we used the 5,819 PMC full document corpus also used for the indexing task. In this experiment our AMTEx method (v2.0) is again compared to MMTx, which was considered the benchmarking method for this task and to the retrieval results of the manually assigned MeSH terms. However, for this task the results were evaluated against 15 TREC provided queries (for PMC, there are no relevance judgments available by TREC or elsewhere). Relevance judgements on the first 25 answers retrieved by all the three competitive methods (AMTEx, MMTx and manually assigned MeSH terms) for all the 15 queries were provided by a domain expert (it was impossible to evaluate answers for the entire set of the 64 TREC queries as in the previous experiment as this would require 64x20x3 = 3,840 relevance judgments by our domain expert). The queries used for this experiment are presented on Table 5 below.

1. Menopausal woman without hormone replacement therapy	8. Carcinoid tumors of the liver
2. Woman with advanced metastatic breast cancer	9. Female with urinary retention
3. Woman with back pain	10. Stroke and systolic hypertension
4. Patient with hypothermia	11. Female with lactase deficiency
5. Male with pericardial effusion	12. Female some months pregnant
6. Patient with fever or lymphadenopathy	13. Man with sickle cell disease
7. Man with cystic fibrosis	14. Adult respiratory distress syndrome
	15. Young man diabetic

Table 5: PMC Full document Retrieval Experiment Queries

Figure 5 shows AMTEX (v2.0) clearly outperforming MMTx and nearing the performance of the manually assigned MeSH index terms.

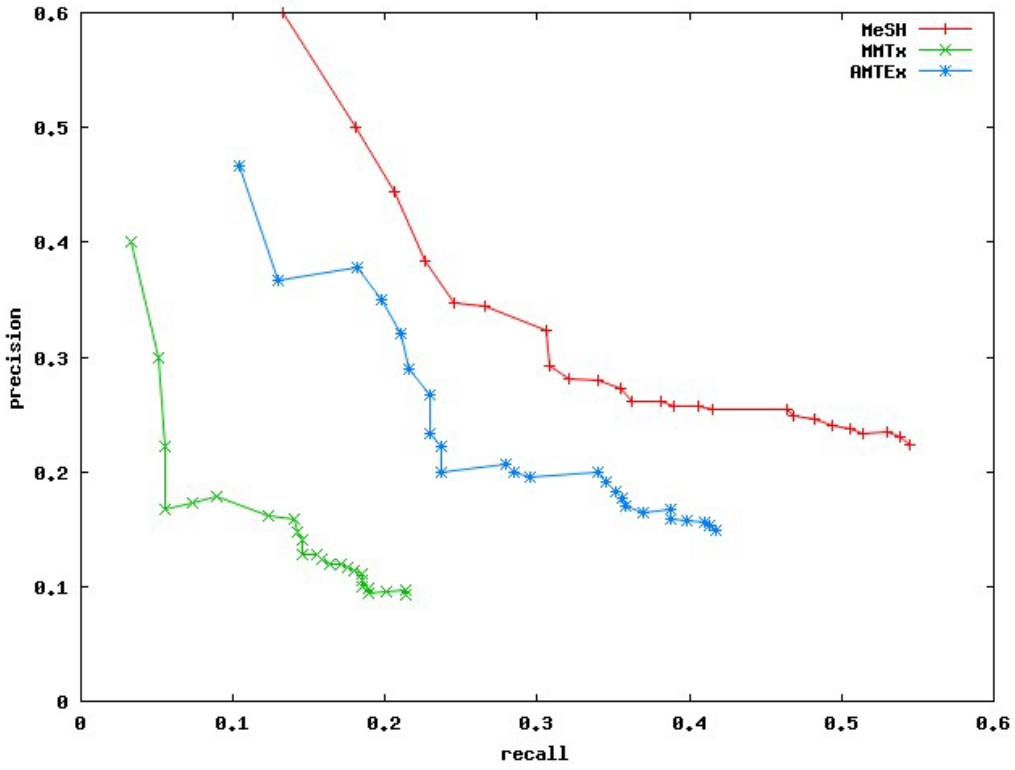

Figure 5: Precision/Recall of AMTEx vs. MMTx on PMC dataset retrieval task

Notice that although MMTx is tuned towards higher recall (by revealing more indexing terms) this does not always lead to improved retrieval performance. A possible explanation could lie in the fact that the document indexing is based on

manual assignment of MeSH terms based on the document conceptual classification done by human experts, whereas retrieval is based on string matching (on terms or term parts found in the document) and the evaluation of these results.

Based on all four experiments we conclude that the AMTEx selective term output method is very well suited for both indexing and retrieval, performing faster and providing a better and concise term output, whereas MMTx increased recall can be well suited in some retrieval cases, where the small document size is prohibitive for the optimal application of AMTEx statistical term extraction process.

6. Conclusions

This article discusses the automatic mapping of documents to the correct MeSH index terms. We consider the term extraction problem for the automatic indexing of documents in large medical collections, such as the MEDLINE collection and we briefly present related approaches to this problem, focusing on the MMTx method, which attempts to map terms in medical documents to UMLS Metathesaurus concepts.

We propose and present the development of an alternative method, the AMTEx method, aiming at providing more accurate and concise terms while being more efficient in terms of processing speed. AMTEx is specifically designed for the automatic indexing of MEDLINE documents, using the MeSH Thesaurus resource and a well-established method for extraction of domain terms, the C/NC-value method.

In this article, we present the experiments we conducted for the refinement of the AMTEx method. We show how we refined our term output by term weighting and by

applying a final cutoff threshold. The AMTEx algorithm is consolidated to include a refined process in Single-word term extraction stage, which is shown to improve its results.

AMTEx is also compared to MMTx in the indexing and the retrieval tasks. The results show that AMTEx performs very well in both tasks, with its average term output being 20 to 50% less than MMTx and its processing speed 3 to 5 times faster than MMTx. MMTx's increased recall may present better results in the small size document retrieval task, where the small document size is prohibitive for the optimal application of AMTEx statistical term extraction process.

Acknowledgements

This work was supported by project TOWL (FP6-STREP, Project No. 026896) of the European Union (EU). We would like to thank Dimitris Makreas, MD of Greek National Health Care System, for his valuable contribution in this work. Dr Makreas proposed a methodology for the intellectual evaluation of the PMC answer sets. He has furthermore carried out the relevance judgments for our experiments, a work of crucial importance for any well-founded evaluation of retrieval techniques.

7. References

[1] MEDLINE: Medical Literature Analysis and Retrieval System Online. [cited 2008 June]; http://www.nlm.nih.gov/databases/databases_medline.html.

[2] UMLS Metathesaurus: The Unified Medical Language System. [cited 2008 June]; http://www.nlm.nih.gov/research/umls.

[3] MeSH: The Medical Subject Headings (MeSH) thesaurus. [cited 2008 June]; http://www.nlm.nih.gov/mesh.

[4] MMTx: MetaMap Transfer tool. [cited 2008 June]; http://mmtx.nlm.nih.gov.

[5] Aronson A R. Effective Mapping of Biomedical Text to the UMLS® Metathesaurus®: The MetaMap Program. In: Proc American Medical Informatics Association Symposium; 2001. p. 17-21.

[6] Pratt W, Yetisgen-Yildiz M. A Study of Biomedical Concept Identification: MetaMap vs. People. In: Proc American Medical Informatics Association Symposium; 2003 Nov; Washington DC, USA. p. 529-33.

[7] Divita G, Tse T, Roth L. Failure Analysis of MetaMap Transfer (MMTx). In: Fieschi M, Coiera E, Li Y C J, editors. MEDINFO 04; 2004 Aug; Amsterdam: IOS Press; 2004. p. 763-7.

[8] Hliaoutakis A, Zervanou K, Petrakis E G M, Milios E. Automatic Document Indexing in Large Medical Collections. In Proc ACM International Workshop on Health Information and Knowledge Management (HIKM); 2006 Nov 11; Arlington, VA, USA. p.1-8.

[9] Hliaoutakis A, Zervanou K, Petrakis E G M Medical Document Indexing and Retrieval: AMTEx vs. NLM MMTx. In Proc 12th International Symposium for Health Information Management Research (ISHIMR), 2007 Jul 18-20, Sheffield, UK.

[10] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: The C-Value/NC-value Method. Int J Digital Libraries 2000; 3(2):117-32.

[11] ISO 704. Principles and Methods of Terminology. Technical report, International Organization for Standardization, Geneva, Switzerland, 2000.

[12] Manning C, Schüzte H. Foundations of Statistical Natural Language Processing. Cambridge: MIT Press; 1999.

[13] Ananiadou S. A Methodology for Automatic Term Recognition. In: COLING-94. Proc; 1994 Aug 5-9; Kyoto, Japan. p. 1034-8.

[14] Bourigault D, Gonzalez-Mullier I, Gros C. LEXTER, a Natural Language Tool for Terminology Extraction. In: Gellerstam M et al., editors. EURALEX'96 Int Congress on Lexicography; 1996 Aug 13-18; Göteborg, Sweden. p. 771-9.

[15] Gaizauskas R, Demetriou G, Humphreys K. Term Recognition in Biological Science Journal Articles. In: Ananiadou S, Maynard D, editors. Proc NLP 2000 Workshop on Computational Terminology for Medical and Biological Applications; June 2000; Patras, Greece. p. 37-44.

[16] Daille B, Gaussier E, Lange J. Towards Automatic Extraction of Monolingual and Bilingual Terminology. In: COLING-94 Proc; 1994 Aug 5-9; Kyoto, Japan. p. 515-21.

[17] Maynard D, Ananiadou S. TRUCKS: A Model for Automatic Multi-Word Term Recognition. J Natural Language Processing 2000; 8(1):101-5.

[18] Jacquemin C. Spotting and Discovering Terms through Natural Language Processing. Cambridge: MIT Press; 2001.

[19] Yu H, Hatzivassiloglou V, Rzhetsky A, Wilbur W J. Automatically Identifying Gene/Protein Terms in MEDLINE Abstracts. J Biomed Inform 2002; 35: 322-30.

[20] Yakushiji A, Tateisi Y, Miyao Y, Tsujii J. Event Extraction from Biomedical Papers using a Full Parser. In: PSB2001 Proc; 2001; Hawaii, USA. p. 408-19.

[21] Zervanou K, McNaught J. A Domain-Independent Approach to IE Rule Development. In: LREC2004 Proc; 2004 May 26-28; Lisbon, Portugal. p. 745-8.

[22] Milios E, Zhang Y, He B, Dong L. Automatic Term Extraction and Document Similarity in Special Text Corpora. In: Proc 6th Conf Pacific Association for Computational Linguistics; Aug 2003; Halifax, Canada. p. 22-5.

[23] Witten I, Paynter G, Frank E, Gutwin C, Nevill-Manning C. KEA: Practical Automatic Keyphrase Extraction. In: Proc 4th ACM Conf on Digital Libraries; Aug 1999; Berkeley, USA. p. 254-5.

[24] Zhang Y, Milios E, Zincir-Heywood N. Narrative Text Classification and Automatic Key Phrase Extraction in Web Document Corpora. In: WIDM2005 Proc 7th ACM Int Workshop on Web Information and Data Management; 2005 Nov 5, Bremen, Germany. p.51-8.

[25] Aronson A R. MetaMap: Mapping Text to the UMLS® Metathesaurus®. 1996 March [cited 2007 March]; http://skr.nlm.nih.gov/papers.

[26] Aronson A R. MetaMap Variant Generation. 2001 May [cited 2007 March]; http://skr.nlm.nih.gov/papers.

[27] Aronson A R. MetaMap Candidate Retrieval. 2001 July [cited 2007 March]; http://skr.nlm.nih.gov/papers.

[28] Aronson A R. MetaMap Evaluation. 2001 May [cited 2007 March]; http://skr.nlm.nih.gov/papers.

[29] Ananiadou S, Albert S, Schuhmann D. Evaluation of Automatic Term Recognition of Nuclear Receptors from Medline. Genome Inform Ser Workshop Genome Inform 2000 Dec 11;11:450-1.

[30] Li Y,Bandar Z A, McLean D. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE Trans Knowl Data Eng 2003 Jul/Aug; 15(4):871-82.

[31] Patwardhan S, Banerjee S, Petersen T. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In Proc Int Conf on Intelligent Text Processing and Computational Linguistics; 2003; Mexico City. p. 17-21.

[32] TREC:Text REtrieval Conference TREC-9 Filtering Track Collections: OHSUMED. [cited 2007 March]; http://trec.nist.gov/data/t9_filtering.html.

[33] Salton G. Automatic text processing: the transformation analysis and retrieval of information by computer. Reading, MA: Addison-Wesley; 1989.

[34] Hliaoutakis A, Varelas G, Voutsakis E, Petrakis E, Milios E, Information Retrieval by Semantic Similarity, International Journal on Semantic Web and Information Systems (IJSWIS), Vol. 3, No. 3, July/September, 2006, pp. 55-73