Automatic Document Indexing in Large Medical Collections

Angelos Hliaoutakis† angelos@softnet.tuc.gr

Kalliopi Zervanou† kelly@intelligence.tuc.gr

Euripides G.M. Petrakis† petrakis@intelligence.tuc.gr

Evangelos E. Milios‡ eem@cs.dal.ca

† Dept. of Electronic and Comp. Engineering, Technical University of Crete (TUC), Chania, Crete, Greece
‡ Faculty of Comp. Science, Dalhousie University, Halifax, Nova Scotia, Canada

ABSTRACT

Term extraction relates to extracting the most characteristic or important terms (words or phrases) in a document. This information is commonly used for improving the accuracy of document indexing and retrieval in large text collections. It also allows for faster and better understanding of the contents of a document collection without first browsing through the contents of its documents. This paper presents $AMTE_X$, an automatic term extraction 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). $AMTE_X$ combines MeSH, the terminological thesaurus resource of NLM, with a well-established method for extraction of domain terms, the C/NC-value method. The performance evaluation of various $AMTE_X$ configurations in the indexing task is measured against the current state-of-the-art, the MMTx method. The experimental results on a subset of MEDLINE documents demonstrate that $AMTE_X$ achieves better precision and recall than MMTx.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods, Dictionaries, Indexing methods, Linguistic processing, Thesauruses*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis*

General Terms

Algorithms, Management, Performance, Experimentation

Keywords

term extraction, medical document retrieval, document indexing

1. INTRODUCTION

New technological developments in communications have not only increased our ability to disseminate information in electronic form, but also the amount of the communicated information, which we need to process and assimilate. The availability of large medical collections, such as MEDLINE [1] (Medical Literature Analysis and Retrieval System Online), 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 [2] (Unified Medical Language System) Metathesaurus, the MeSH [3] (Medical Subject Headings) thesaurus. The automatic mapping of biomedical documents to UMLS term concepts has been undertaken by U.S. National Library of Medicine with the development of MMTx [4] (MetaMap Transfer tool).

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 a pilot study by Divita et al. [11]. Our experiments with the MMTx application on MEDLINE documents have shown that the MMTx output suffers, not only in recall (as noted by [11]), failing to extract all domain terms, but also because it over-generates by producing terms that are too general, which diffuse the document concept leading to inaccurate retrieval of MEDLINE documents. The latter reflects an inherent limitation of MMTx, which was not designed to focus on MeSH terms, whereupon MEDLINE indexing has been based. Additionally, the variant generation process of MMTx is found to account for the over-generation problem for retrieval purposes.

In this paper, we briefly review the MMTx approach and we propose an alternative method, the Automatic MeSH Term Extraction method ($AMTE_X$). $AMTE_X$ aims at improving the efficiency of automatic term extraction, using a hybrid linguistic/statistical term extraction method, the C/NC-value method [12]. Additionally, $AMTE_X$ aims at improving indexing and retrieval of MEDLINE documents, based on the extraction and mapping of document terms to the MeSH Thesaurus, rather than the full UMLS

HIKM'06, November 11, 2006, Arlington, Virginia, USA.
Copyright 2006 ACM 1-59593-528-2/06/0011 ...$5.00.

[1] http://www.nlm.nih.gov/databases/databases_medline.html

[2] http://www.nlm.nih.gov/research/umls

[3] http://www.nlm.nih.gov/mesh

[4] http://mmtx.nlm.nih.gov

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 $AMTE_X$ approach, namely the MeSH thesaurus and the C/NC-value method for term extraction. Then, the $AMTE_X$ 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 words 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, but also for specialised concept distinction and classification purposes [1]. 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 informational content load of a document. In early approaches, terms have been sought for indexing purposes, using mostly $tf \cdot idf$ counts [18]. Term extraction approaches largely rely on the identification of term formation patterns (e.g. [2, 10, 14]). Statistical techniques may also be applied to measure the degree of unithood or termhood of the candidate multi-word terms (e.g. [9]). Later and current approaches tend to follow a hybrid approach combining both statistical and linguistic techniques (e.g. [13, 19, 16]).

The extraction of terms for the medical, biological and biomedical domain has greatly motivated research for both indexing, as well as knowledge extraction purposes [14, 27, 26, 28]. 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 $tf \cdot idf$ method [18]. Although terms (domain terms [5]) may be discovered in such a process, neither all terms are useful index terms, nor all index terms are terms. For example, a valid term appearing very frequently in a document collection is useless for the retrieval of a specific document. Moreover, query and document representations traditionally 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 [16] and are often used in indexing medical documents. Multi-word terms carry important classificatory content information, since they comprise modifiers denoting a specialization of the more general single-word, head term [10]. For example, the compound term *"heart disease"* denotes a specific type of *"disease"*. A recent study by Milios et al. [20] of the extraction of multi-word terms for retrieval purposes suggests that multi-word term methods may complement other methods to improve results. Currently machine learning techniques are also applied for indexing, such as the Naive Bayes learning model implemented in the KEA (Automatic Key-phrase Extraction, [25]). Comparative experiments of $tf \cdot idf$, KEA and the C/NC-value term extraction methods by Zhang et al. [29] show that C/NC-value significantly outperforms both $tf \cdot idf$ and KEA in a narrative text classification task using the extracted terms.

Since term extraction is primarily based on surface term form patterns, it inherently suffers from two problems: ambiguity and variation. Ambiguity relates to the semantic interpretation of a given term form and it arises when this form can be interpreted in more than one way. Variation is generally defined as the alteration of the surface term form of a terminological concept. According to Jacquemin [16], variation is more specifically defined as a transformation of a controlled multi-word term and can be of three types: morphological, syntactic or semantic. Many approaches, such as [19], [16] and [14], including MMTx and our $AMTE_X$, attempt to resolve the problems of ambiguity and variation in terminological concepts by combining simple text normalisation techniques, statistics, or more elaborate rule-based, linguistic techniques, with existing thesaurus and lexicon information. In a previous work, we implemented MedSearch [15], a retrieval system that discovers semantically similar terms in documents and queries based on the computation of semantic similar terms in different taxonomies using our SSRM statistical method [24].

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.

3.1.1 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*® 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 linguistic utterance which is 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, and it does not fulfills 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

[5] At this point we wish to make a distinction between (a) the notion of *term* which, depending on the scientific community, may refer to the terminologically acceptable notion of *domain or scientific term*, as defined in the beginning of this section; and (b) the notion of *index term*, namely a key concept, word or phrase, which semantically labels and conceptually categorises the content of a document for information management purposes, such as retrieval. In the rest of this paper, the notion of term refers mainly to index terms, though in the C/NC-value approach used in our method, the design objective is *domain term* extraction, rather than indexing.

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.

3.1.2 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 performs the following steps [5]:

1. **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 $(\text{Adj}|\text{Noun})^{+}\text{Noun}$ isolates noun phrases [4]. 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*.

2. **Variant Generation:** Variant generation is performed in an 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 [8]. For example, the term *"liquid crystal thermography"* would be split into the generators: *"liquid crystal thermography"*, *"liquid crystal"*, *"liquid"*, *"crystal"* and *"thermography"* [4]. 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.

3. **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 [6]. The mapping process may vary [4]. 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*).

4. **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 [7]:

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

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

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

   Cohesiveness is similar to coverage and denotes the continuous 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 $AMTE_X$ Method Resources

3.2.1 The C/NC-value Method for Term Extraction

The C/NC-value method [13] 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:

$$N^{+}N$$

$$(A|N)^{+}N$$

$$((A|N)^{+}|((A|N)^{*}(N\ \ P)?)(A|N)^{*})N$$

where N is a noun, A 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 [12]. 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

$$\textit{C-value} = \begin{cases} log_2|a|f(a), \\ log_2|a|(f(a) - \frac{1}{P(T_a)}\sum_{b\in T_a} f(b)). \end{cases} \quad (1)$$

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 words that tend to appear with terms and is computed as

$$weight(w) = \frac{t(w)}{n}, \quad (2)$$

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

$$\textit{NC-value(a)} = 0.8 \cdot \textit{C-value(a)} + 0.2 \cdot CF(a). \quad (3)$$

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

$$CF(a) = \sum_{w\in C_a} f_a(w) \cdot weight(w), \quad (4)$$

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 [3]), eye pathology medical records [12], biomedical business newswire texts [28] and computer science papers [20].

3.3 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:

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

Scope Note: a text description of the term;

Entry Terms: mostly synonym terms to the MH.

Entry terms also include stemmed MH terms and are sometimes referred to as quasi-synonyms (they are not always exact synonyms). In our *AMTE*$_X$ approach, all entry 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 Fig. 1.

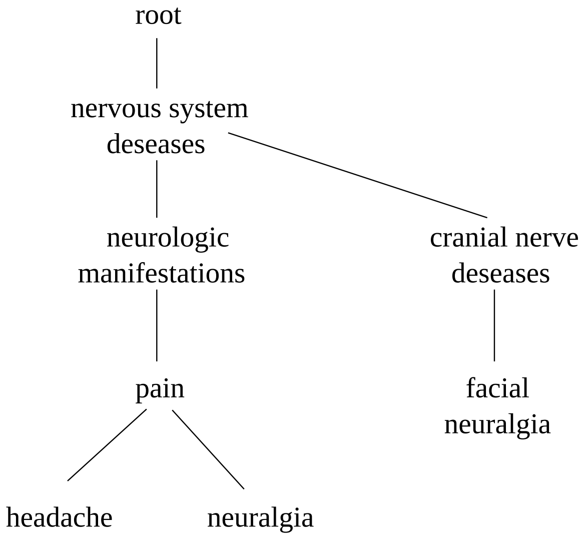

Figure 1: A fragment of the MeSH IS-A hierarchy.

4. THE *AMTE*$_X$ METHOD

Based on the study of the MMTx algorithm and resources, discussed in section 3.1, 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 [11]. However, this process also results in term over-generation and increased term ambiguity, which diffuse the original term concept, leading to inaccurate retrieval.

- 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 retrieval 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 [3, 12, 28, 20] 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:

1. Term extraction based on a well-established method, the C/NC-value method;
2. Use of MeSH Thesaurus as lexical resource, both for (limited) term variant retrieval, and candidate term mapping.

Input: Document d, MeSH Ontology.

Output: MeSH terms t.

1. Multi-word Term Extraction: The C/NC-value method is applied.

2. Term Ranking: Extracted terms are ranked by NC-value (Eq. 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 are added.

Figure 2: *AMTE$_X$* Algorithm.

An outline of the *AMTE$_X$* procedure is illustrated in Fig. 2. In particular, the *AMTE$_X$* 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 *AMTE$_X$* the document text is parsed, using the C/NC-value part-of-speech tagger and linguistic filters.
2. **Term Ranking:** Extracted candidate terms are evaluated, first by C-value and subsequently by NC-value score. The final candidate term list is ranked by decreasing term likelihood (Eq. 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. In this work we kept all 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:** The C/NC-value tends to produce compound (multi-word) terms (it does not produce single-word terms). Often such terms include shorter terms (mostly single-word terms) which are also MeSH terms. Single-word terms are also extracted and are added to the candidate term list (the text is scanned and each word is checked against the MeSH vacabulary).
5. **Term Variants:** Term variants are included in the candidate term list. The C/NC-value implementation in *AMTE$_X$* 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 *AMTE$_X$* since the full list of Entry Terms may contain terms, that often are not synonymous.
6. **Term Expansion:** The list of terms is augmented with semantically (conceptually) similar terms from MeSH. Fig. 3 illustrates this process: A term is represented by its MeSH tree hierarchy. The neighborhood of the term is examined and all terms with similarity greater than threshold T 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.

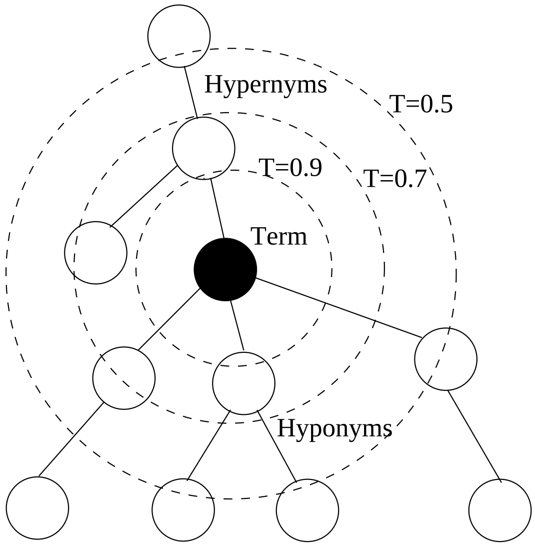

Figure 3: Term expansion using MeSH.

AMTE$_X$ 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 *AMTE$_X$* 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. Nevertheless, information about the *head* of a term phrase could be easily included in subsequent implementations to refine single-word term extraction, based on a single-word term *head*.

Regarding our approach to variant generation, it is more limited than MMTx. This could constrain our term recall to terms that are more related to the original term in text but are not included in MeSH. However, as we observe in the results of our experiments in section 5, we manage to achieve not only better recall, but also better precision. We believe that 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 $AMTE_X$ 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 the original concept to unrelated concepts.

Regarding *Term Expansion*, the method suggested for $AMTE_X$ for discovering semantically similar terms, is based on the semantic similarity method by Li et al. [17]. The evaluation of the semantic similarity methods indicated that this method is particularly effective, achieving up to 73% correlation with results obtained by humans [23]. 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 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).

Because no synonymy relation is defined in MeSH, we did not apply expansion to the Entry Terms of terms. Word sense disambiguation [21] can also be applied for detecting the correct sense to expand (here, expansion is applied to the most common sense of each term). The specification of T requires further investigation (e.g., appropriate threshold values can be learned by training).

A final observation is that term ranking in $AMTE_X$ is currently based on NC-value scores. However, at this stage we could also incorporate, in future versions of the system, the semantic similarity scoring, estimated during the detection of term variants.

5. EVALUATION

In order to assess the performance of our $AMTE_X$ method, we have experimented in the retrieval task of MEDLINE documents, using various $AMTE_X$ configurations and comparing it to MMTx, which is considered the benchmark method for this task. In this section, we first describe the design of our experiments and then we discuss our results and observations, compared to MMTx.

Our testing corpus consisted of a set of 61 full MEDLINE documents. Notice that MEDLINE stores mainly publication abstracts with citation information and index (MeSH) terms. We chose to work on full documents, rather than abstracts or mere document titles, because the original manual indexing was based on the full article content. Moreover, the terms found in abstracts or titles may be neither enough, nor adequate for our purposes. The documents were downloaded from PMC database of NCBI Pubmed [6] giving as input the term "*pain*". Out of the first 100 results, 61 documents were finally selected for our evaluation experiments. The rest were judged inadequate, either because they were too small in size, or because they constituted a document collection, rather than a single document, appearing as a single pdf file. The same corpus was used as input in all our experiments, both for $AMTE_X$ and MMTx.

Fig. 4 illustrates an an example of how MMTx and AMTEx work in practice on an example document selected at random from the test collection. $AMTE_X$ produced a more compact and coherent set of index terms than MMTx. Almost all terms are related to the document. MMTx (although restricted to the generation of MeSH terms in this example) still results in term over-generation which diffuses the original document concept (leading to topic drift).

The performance of all methods was measured in terms of precision and recall. Gold standard for evaluation was considered the set of MeSH terms appearing in each MEDLINE document index (provided by experts). Thus, in our evaluation, *precision* is the total number of correctly extracted terms, compared to the MeSH terms appearing in the respective document index. Similarly, *recall*, in our evaluation, is the total number of correctly retrieved terms, compared to the total number of terms in the MeSH index gold standard.

Input: Full text article [22]

MEDLINE index terms: "Aged", "Data Collection", "Humans", "Knee", "Middle Aged", "Osteoarthritis, Knee/complications", "Osteoarthritis, Knee/diagnosis", "Pain/classification", "Pain/etiology", "Prospective Studies", "Research Support, Non-U.S. Gov't"

MMTx terms: "osteoarthritis knee", "retention", "peat", "rheumatology", "acetylcholine", "lysine acetate", "potassium acetate", "questionnaires", "target population", "population", "selection bias", "creativeness", "reproduction", "cohort studies", "europe", "couples", "naloxone", "sample size", "arthritis", "data collection", "mail" 'health status", "respondents", "ontario", "universities", "dna", "baseline survey", "medical records", "informatics", "general practitioners", "gender", "beliefs", "logistic regression", "female", "marital status", "employment status", "comprehension", "surveys", "age distribution", "manual", "occupations", "manuals", "persons", "females", "minor", "minority groups", "incentives", "business", "ability", "comparative study", "odds ratio", "biomedical research", "pubmed", "copyright", "coding", "longitudinal studies", "immunoelectrophoresis", "skin diseases", "government", "norepinephrine", "social sciences", "survey methods", "tyrosine", "new zealand", "azauridine", "gold", "nonrespondents", "cycloheximide", "rheum", "jordan", "cadmium", "radiopharmaceuticals", "community", "disease progression", "history"

$AMTE_X$ terms: "health surveys", "pain", "review publication type", "data collection", "osteoarthritis knee", "knee", "science", "health services needs and demand", "population", "research", "questionnaires", "informatics", "health"

Figure 4: Example illustrating MeSH terms indexing an article in MEDLINE and terms computed by MMTx and $AMTE_X$.

The results of our evaluation are shown in Table 1. In our initial $AMTE_X$ configuration experiments, we observed that 18% (on the average) of the candidate terms extracted from the 61 documents by a similarity measure [17] $T > 0.5$, is also included in the list of MeSH index terms (our gold standard). For this reason, we have decided to experiment with various configurations of the *Term Expansion* threshold, exceeding $T > 0.5$.

The results of MMTx, and of various $AMTE_X$ configurations corresponding to various semantic similarity thresholds T and compound terms (single-word terms are left out) are shown in Table 1. Term expansion with lower values of T (e.g., $T = 0.5$) demonstrated an increase in recall, revealing more correct terms. However, at the same time, precision is decreased, since the expansion step also introduced some unrelated terms. We observe that expansion with low threshold values T (e.g., $T = 0.5$) is likely to

[6] http://www.ncbi.nlm.nih.gov/entrez

Method	Precision	Recall
MMTx	0,013481	0,015109
AMTE$_X$ ($T = 0.5$)	0,186025	0,108085
AMTE$_X$ ($T = 0.55$)	0,205087	0,097531
AMTE$_X$ ($T = 0.6$)	0,21827	0,090039
AMTE$_X$ ($T = 0.65$)	0,227428	0,083379
AMTE$_X$ ($T = 0.7$)	0,235518	0,072318
AMTE$_X$ ($T = 0.8$)	0,235592	0,072243
AMTE$_X$ ($T = 0.9$)	0,23615	0,070267

Table 1: Precision and Recall of MMTx and *AMTE*$_X$ for various expansion thresholds T.

introduce many new terms and diffuse the topic of the query (topic drift).

In the final *AMTE*$_X$ configuration, a semantic similarity threshold $T = 0.9$ was selected as the optimal configuration, which allows term expansion only for very similar terms.

The importance of single-word terms in term extraction is illustrated in Table 2. The results demonstrate that it is possible to improve the recall of the method by including single-word terms in the list of multi-word terms above. This increased recall by 15% but at the same time precision decreased by 12% (some single-word terms, were not identified by the experts). Overall, our *AMTE*$_X$ method outperforms MMTx, reaching up to 21% better precision (*AMTE*$_X$ without single-word terms) and up to 21% better recall (*AMTE*$_X$ with single-word terms).

Method	Precision	Recall
MMTx	0,013481	0,015109
AMTE$_X$	0,23615	0,070267
AMTE$_X$+single-word MeSH terms	0,119629	0,228322

Table 2: Comparative Precision and Recall of MMTx, *AMTE*$_X$ and *AMTE*$_X$ including single-word term extraction.

6. CONCLUSIONS

The paper is about mapping documents to the correct MeSH index terms automatically. We discussed the term extraction problem for the automatic indexing of documents in large medical collections, such as the MEDLINE collection. We have briefly presented related approaches to this problem, focusing on the MMTx method, which attempts to map terms in medical documents to UMLS Metathesaurus concepts. We have developed an alternative method, the *AMTE*$_X$ method, which is specifically designed for indexing and retrieval of MEDLINE documents, using the MeSH Thesaurus resource and a well-established method for extraction of domain terms, the C/NC-value method. Our experimental results show that our *AMTE*$_X$ method outperforms the current benchmark method of MMTx, reaching significantly better precision and better recall.

Future developments on *AMTE*$_X$ method include:

- Incorporating a shallow syntactic parser in our approach and experimenting with single-word term extraction, based on *head* single-word term;

- Experimenting with smaller documents and/or document abstracts, titles;

- Including semantic similarity score in our term ranking function (currently C/NC-value);

- The specification of T threshold could be further investigated, to study, for example whether appropriate threshold values can be learned by training;

- Word sense disambiguation (WSD [21]) could also be applied in detecting the correct sense to expand, rather than expanding the most common sense of each term.

7. ACKNOWLEDGEMENT

This work was supported by project MedSearch/BIOPATTERN (FP6-NoE, Project No. 508803) and by project TOWL (FP6-Strep, Project No. 26896) of the European Union (EU).

8. REFERENCES

[1] I. 704. Principles and Methods of Terminology. Technical report, Intern. Organization for Standardization, Geneva, Switzerland, 1986.

[2] S. Ananiadou. A Methodology for Automatic Term Recognition. In *Proc. of COLING-94*, pages 1034–1038, Kyoto, 1994.

[3] S. Ananiadou, S. Albert, and D. Schuhmann. Evaluation of Automatic Term Recognition of Nuclear Receptors from Medline. *Genome Informatics Series*, 11, 2000.

[4] A. R. Aronson. MetaMap: Mapping Text to the UMLS$^{®}$ Metathesaurus$^{®}$, March 1996. http://skr.nlm.nih.gov/papers.

[5] A. R. Aronson. Effective Mapping of Biomedical Text to the UMLS$^{®}$ Metathesaurus$^{®}$: The MetaMap Program. In *Proceedings of AMIA 2001*, pages 17–21, 2001.

[6] A. R. Aronson. MetaMap Candidate Retrieval, July 2001. http://skr.nlm.nih.gov/papers.

[7] A. R. Aronson. MetaMap Evaluation, May 2001. http://skr.nlm.nih.gov/papers.

[8] A. R. Aronson. MetaMap Variant Generation, May 2001. http://skr.nlm.nih.gov/papers.

[9] D. B, E. Gaussier, and J. Lange. Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *Proc. of COLING-94*, pages 515–521, Kyoto, 1994.

[10] D. Bourigault, I. Gonzalez-Mullier, and C.Gros. LEXTER, a Natural Language Tool for Terminology Extraction. In *EURALEX '96: Proc. I-II, Part II – Papers submitted to the Seventh EURALEX International Congress on Lexicography in Göteborg*, pages 771–779, Göteborg University, Göteborg, Sweden, 1996.

[11] G. Divita, T. Tse, and L. Roth. Failure Analysis of MetaMap Transfer (MMTx). *Medinfo*, pages 763–767, 2004.

[12] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms: The C-Value/NC-value Method. *International Journal of Digital Libraries*, 3(2):117–132, 2000.

[13] K. Franzi and S. Ananiadou. The C/NC Value Domain Independent Method for Multi-Word Term Extraction. *Journal of Natural Language Processing*, 6(3):145–180, 1999.

[14] R. Gaizauskas, G. Demetriou, and K. Humphreys. Term Recognition in Biological Science Journal Articles. In *Workshop on Computational Terminology for Medical and Biological Applications, (NLP 2000)*, pages 37–44, Patras, 2000.

[15] A. Hliaoutakis, G. Varelas, E. G. Petrakis, and E. Milios. MedSearch: A Retrieval System for Medical Information Based on Semantic Similarity. In *Proc. of the* 10^{th} *ECDL European Conference on Research and Advanced Technology for Digital Libraries (ECDL'2006)*, pages 512–515, Alicante, Spain, September 17-22 2006.

[16] C. Jacquemin. *Spotting and Discovering Terms through Natural Language Processing*. MIT Press, Cambridge, MA, USA, 2001.

[17] Y. Li, Z. A. Bandar, and D. McLean. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Trans. on Knowledge and Data Engineering*, 15(4):871–882, July/Aug. 2003.

[18] C. Manning and H. Schüzte. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, June 18 1999.

[19] D. Maynard and S. Ananiadou. TRUCKS: A Model for Automatic Multi-Word Term Recognition. *Journal of Natural Language Processing*, 8(1):101–105, 2000.

[20] E. Milios, Y. Zhang, B. He, and L. Dong. Automatic Term Extraction and Document Similarity in Special Text Corpora. In *Proc. of the* 6^{th} *Conf. of the Pacific Association for Computational Linguistics*, pages 22–25, Halifax, Aug 2003.

[21] S. Patwardhan, S. Banerjee, and T. Petersen. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In *Intern. Conf. on Intelligent Text Processing and Comutational Linguistics*, pages 17–21, Mexico City, 2003.

[22] G. Peat et al. The Knee Clinical Assessment Study CAS(K). A Prospective Study of Knee Pain and Knee Osteoarthritis in the General Population: Baseline Recruitment and Retention at 18 months, March 2006. http://www.biomedcentral.com/content/pdf/1471-2474-7-30.pdf.

[23] E. G. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou. Design and Evaluation of Semantic Similarity Measures for Concepts Stemming from the Same or Different Ontologies. In 4^{th} *Workshop on Multimedia Semantics (WMS'06)*, pages 44–52, Chania, Crete, Greece, 1998.

[24] G. Varelas, E. Voutsakis, P. Raftopoulou, E. Petrakis, E., and Milios. Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web. In *Proc. of the* 7^{th} *ACM Intern. Workshop on Web Information and Data Management(WIDM 2005)*, pages 10–16, Bremen, Germany, 2005.

[25] I. Witten, G. Paynter, E. Frank, C. Gutwin, and C. Nevill-Manning. KEA: Practical Automatic Keyphrase Extraction. In *Proc. of the* 4^{th} *ACM Conference on Digital Libraries*, pages 254–255, Berkeley, CA, USA, Aug. 1999.

[26] A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii. Event Extraction from Biomedical Papers using a Full Parser. In *Proceedings of the sixth Pacific Symposium on Biocomputing (PSB 2001)*, pages 408–419, Hawaii, U.S.A., 2001.

[27] H. Yu, V. Hatzivassiloglou, A. Rzhetsky, and W. Wilbur. Automatically Identifying Gene/Protein Yerms in MEDLINE Abstracts. *Journal of Biomedical Informatics*, 35:322–330, 2002.

[28] K. Zervanou and J. McNaught. A Domain-Independent Approach to IE Rule Development. In *Proc. of the 4th Intern. Conf. on Language Resources and Evaluation (LREC 2004)*, pages 745–748, Lisbon, Portugal, May 2004.

[29] Y. Zhang, E. Milios, and N. Zincir-Heywood. Narrative Text Classification and Automatic Key Phrase Extraction in Web Document Corpora. In *7th ACM Intern. Workshop on Web Information and Data Management (WIDM 2005)*, pages 51–58, Bremen, German, Nov. 5 2005.