Automatic Document Categorisation by User Profile in Medline

Euripides G.M. Petrakis[1], Angelos Hliaoutakis[2]

Dept. Of Electronic and Comp. Engineering, Technical Univ. of Crete (TUC), Chania, Crete, Greece, [1]euripides@intelligence.tuc.gr , [2]angelos@softnet.tuc.gr

We investigate potential improvements to the problem of term extraction related to document representation and indexing in large document collections such as Medline, the premier bibliographic database of the U.S. National Library of Medicine (NLM). Using term extraction methods such as $AMTE_X$ and MMT_X, document representations are semantically compact and more efficient, being reduced to a limited number of meaningful multi-word terms (phrases), rather than large vectors of single-words, part of which may be void of distinctive content semantics. We show how this information can be used for the automatic categorisation of medical documents by user profile (i.e., novice users and experts). This is achieved by mapping document terms to external lexical resources such as WordNet, and MeSH (the medical thesaurus of NLM). Evaluation results of all methods are presented and discussed.

Keywords

document categorisation, health informatics, term extraction

1. Introduction

Typically, medical information systems such as Medline[1], are designed to serve health care professionals (expert users such as clinical doctors and medical researchers). Expert users are familiar with the type and content of medical resources (such as the NLM dictionaries and databases) and use medical terminology for their searches. However, the spread and availability of medical information over the Web have made this information available also to consumer (i.e., novice) users. Unlike experts, consumers are usually unfamiliar with the content and type of specialized medical text resources and typically use the Web for their searches. They are often uncertain as to the exact type of information they are looking for and they do simple searches using natural language (rather than domain specific) terms.

A medical information system must be capable of providing dedicated, domain specific answers to experts or, simple, easy to comprehend answers to novice users respectively. A similar categorisation of medical information applies in existing systems such as MedScape[2], MedlinePlus[3], Wrapin[4] and MedHunt[5] (maintained by HON, the Health on the Net Foundation, a non profit organisation aiming at providing authoritative and trustworthy information on the Web). All systems referred to above focus on providing medical information to consumer or expert users but, rely solely on the manual categorisation of information, a solution which requires intervention by human experts and, therefore, is slow and does not scale-up for large document collections. PubMed[6] of NLM is of particular interest to us. It provides free access to Medline document abstracts and to articles in selected life sciences journals not included in Medline.

Medline documents are currently indexed by human experts by assigning to each one, a number (typically 10 to 12) of terms, based on a controlled list of indexing terms, deriving from a subset of the

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

[2] http://www.medscape.com

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

[4] http://www.wrapin.org

[5] http://www.hon.ch/HONsearch/Patients/medhunt.html

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

UMLS[7] (Unified Medical Language System) Metathesaurus, the MeSH[8] (Medical Subject Headings) thesaurus. 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 MMT_X[9] (MetaMap Transfer tool).

The limitations of MMT_X in term extraction have been analyzed in detail by Hliaoutakis et al. [1]. The experiments with the MMT_X application on Medline documents have shown that the MMT_X not only fails to extract all domain terms, but it also over-generates terms by producing general terms, which diffuse the document concept leading to inaccurate retrieval of Medline documents. The latter reflects an inherent limitation of MMT_X, which was not designed by default to focus on MeSH terms, whereupon Medline indexing has been based. Additionally, the variant generation process of MMT_X is found to account for the over-generation problem for retrieval purposes. $AMTE_X$ (Automatic Medical Term Extraction) [1] aims at improving the efficiency of automatic term extraction, using a hybrid linguistic/statistical term extraction method, the C/NC-value method [2]. $AMTE_X$ is based on the extraction and mapping of document terms to the MeSH Thesaurus, rather than the full UMLS Meta-thesaurus mapping of MMT_X. It is therefore more selective resulting in more compact document representations than MMT_X.

In this work, the performance of $AMTE_X$ is compared against the current state-of-the-art, the MetaMap Transfer (MMT_X) method using two types of corpora: a subset of Medline (PMC) full document corpus and a subset of Medline (OHSUMED) abstracts. Subsequently, the experimental results demonstrate that $AMTE_X$ performs better in indexing in 50% of the processing time compared to MMT_X. We show how term extraction methods (such as MMT_X and $AMTE_X$) can be used for filtering medical information for targeted audiences such as experts and novice users. An obvious application of this filtering operation will be retrieval on medical information by user profile. This approach is automatic and relies on the categorisation of medical terms to terms comprehendible by novice users and to more involved terms typically used by experts (e.g., medical doctors, practitioners etc.). This is made possible with the aid of WordNet[10], a thesaurus for natural language terms of the English language. It is based on the observation that up to 30% of the terms participating in MeSH vocabulary are general terms (terms that can be found in Wordnet as well) while, the remainder 70% are domain specific UMLS terms that do not belong to Wordnet. The performance of the method is assessed using the OHSUMED subset of Medline based on relevance assessments provided by naive users and experts.

An almost orthogonal issue is speed of search. Indexing might not only increase the speed of access to the huge amounts of medical information, but also make this information usable and easily accessible by subject (topic of interest). Without an indexing process, the search engine of a medical information system would scan every document in the corpus, which would require considerable time and computing power. This work, tackles all these issues.

Data and algorithmic resources such as the text extraction methods considered in this work (C-Value/NC-Value, $AMTE_X$ and MMT_X, UMLS and Medline) are presented in Section 2. Document categorisation is discussed in Section 3. Evaluation results are presented in Section 4 followed by conclusions and issues for further research in Section 5.

2. Background and Resources

Automatic indexing and categorisation of medical documents relies mainly on term extraction for the identification of discrete content indicators, namely index terms. Traditional indexing techniques ignore multi-word and compound terms, which are split into isolated single word index terms. However, compound and multi-word terms are very common in the biomedical domain [3] and are often used in indexing medical documents. Multi-word terms carry important classificatory content information, since

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

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

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

[10] http://wordnet.princeton.edu

they comprise of modifiers denoting a specialisation of the more general single-word, head term. For example, the compound term “*heart disease*” denotes a specific type of disease.

In this work, we focus our attention on multi-word terms and AMTE$_X$ [1] for extracting multi-word terms from medical documents. *AMTE$_X$* and *MMT$_X$* have been shown to be more suitable than single-word term extraction methods not only for document indexing and retrieval, but also, for general concept description and ontology construction tasks [4]. *AMTE$_{X,}$* in particular, has been shown to be more selective than the MetaMap Transfer (*MMT$_X$*) method of NLM which maps arbitrary text to concepts in the UMLS Metathesaurus (equivalently, it discovers Metathesaurus concepts in text).

2.1 Unified Medical Language System (UMLS)

The Unified Medical Language System (UMLS) is a source of medical knowledge developed by the U.S. NLM. UMLS consists of the Metathesaurus, the Semantic Network and the SPECIALIST lexicon. Metathesaurus is a large, multi-purpose, and multi-lingual vocabulary database. It integrates about 800,000 concepts from 50 families of vocabularies. In Metathesaurus, equivalent terms are clustered into unique concepts. Each concept is an abstract representation of 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 to translations into other languages. However, the terms integrated in Metathesaurus do not all share a common structure (i.e., same properties and characteristics) and they inherit the organisational principles governing their respective source vocabularies. Moreover, certain types of relationships, including synonymy and hierarchical relationships are not defined. Thus, Metathesaurus on its own neither have a hierarchical structure, nor fulfils ontological requirements.

The Semantic Network (SN) consists of 134 semantic types categorising the Metathesaurus concepts. The purpose of the SN is to provide a consistent categorisation of all concepts represented in Metathesaurus and a set of useful relationships among these concepts. Every concept in Metathesaurus is assigned to at least one semantic type in the Semantic Network. High semantic level hierarchies are defined, such as, one for entities relating to pathology, and one for events (treatment for diseases). The SN 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 SN 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 Metathesaurus and Semantic Network. Moreover, the lack of relationships between concepts in Metathesaurus and Semantic Network has been also observed [5].

2.2 The MeSH Thesaurus

The MeSH Thesaurus (Medical Subject Headings) is a taxonomy of medical and biological terms and concepts suggested by the U.S NLM. 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 (2006) is organised in 15 taxonomies, including 23,884 terms. A term may appear in more than one taxonomy. Each MeSH term is described by several properties, the most important being i) MeSH Heading (MH): the term name or identifier; ii) Scope Note (SN): a text description of the term; iii) Entry Terms (ET): 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 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*”.

2.3 WordNet

WordNet is an on-line lexical reference system developed at Princeton University. WordNet attempts to model the lexical knowledge of a native speaker of English. WordNet v.2.0 (2006) contains around

127,361 terms, organized into taxonomic hierarchies. Nouns, verbs, adjectives and adverbs are grouped into synonym sets (synsets). The synsets are also organized into senses (i.e., corresponding to different meanings of the same term or concept). The synsets (or concepts) are related to other synsets higher or lower in the hierarchy defined by different types of relationships. The most common relationships are the Hyponym/Hypernym (i.e., Is-A relationship), and the Meronym/Holonym (i.e., Part-Of relationship). There are nine noun and several verb Is-A hierarchies (adjectives and adverbs are not organized into Is-A hierarchies). WordNet sometimes is referred to as an ontology of natural language terms although (similarly to the UMLS Metathesaurus) does not fulfil ontological requirements.

2.4 The MMT$_X$ Approach

MMT$_X$ 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:

- *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.
- *Variant Generation*: First, the multi-word term phrase is split into generators. A variant generator is considered to be any meaningful subsequence of words in the phrase. That is either a single word or a term existing in the SPECIALIST lexicon. For example, the term "*liquid crystal thermography*" would be split into the generators: "*liquid crystal thermography*", "*liquid crystal*", "*liquid", "crystal*" and "*thermography*". In this 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. 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*: 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. 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: 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.

2.5 The AMTE$_X$ Approach

AMTE$_X$ implements the C/NC-value [2], domain-independent method for the extraction of multi-word and nested terms. In this approach, noun phrases are initially selected by linguistic filtering. The subsequent statistical component defines the candidate noun phrase termhood by two measures: C-value and NC-value. The first measure, the C-value, is based on the hypothesis that multi-word terms tend to consist of other terms (nested in the compound term). For example, the terms "*coronary artery*" and "*artery disease*" are nested within the term "*coronary artery disease*". Thus, 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. In that respect, C-Value clearly favours longer terms. The second measure, the NC-value, is based on the hypothesis that terms tend to appear in specific context and often co-occur with other terms. Thus, NC-value refines C-value by assigning additional weights to candidate terms which tend to co-occur with specific context words.

The $AMTE_X$ method has the following processing stages:

- *Multi-word Term Extraction*: The C/NC-value method is used for term extraction.
- *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. Top ranked terms are more important than terms ranked lower in the list.
- *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.
- *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.
- *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.
- *Term Expansion*: The list of terms is augmented with semantically (conceptually) similar terms from MeSH by applying the algorithm by Li et al. [6].

2.6 Medline

The Medline database is a collection of biomedical articles. It consists of abstracts of medical publications together with metadata that is information on the organisation of the data, the various data domains, and the relations between them. Publications in the Medline database are manually indexed by NLM using MeSH terms, with typically 10-12 descriptors assigned to each publication by human experts. Hence, the MeSH annotation defines for each publication a highly descriptive set of features. Medline (version 2008) contains over than 16 million publications.

3. Document Categorisation by User Profile

Our method for categorizing Medline documents by user profile relies on our observation that MeSH terms are distinguished into i) general medical terms expressing known concepts (e.g., "*pain*", "*headache*") which are easily conceived by all users, ii) domain specific terms which are used mainly by experts, iii) general - non medical terms. The more expert terms a document contains, the higher its probability to be a document for expert (i.e., one that consumer users cannot comprehend) and the reverse. Figure 1 illustrates the respective categorisation of Medline documents and MeSH terms.

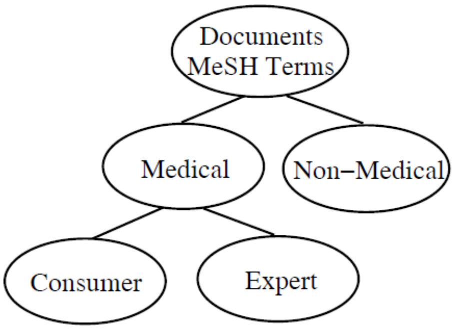

Figure 1 Categorisation of Medline documents and MeSH terms.

By combining information from WordNet and MeSH the following three term vocabularies are constructed:

- *Vocabulary of General Terms (VGT)*: these are terms that belong to WordNet but not to MeSH:
  $$\boldsymbol{VGT} = (\boldsymbol{WordNet}) - (\boldsymbol{MeSH})$$
  It follows that VGT contains 105.675 general (WordNet) terms.
- *Vocabulary of Consumer Terms (VCT)*: these are terms that belong to both, WordNet and MeSH:
  $$\boldsymbol{VCT} = (\boldsymbol{WordNet} \cap (\boldsymbol{MeSH})$$
  It follows that VCT contains 7,165 consumer (MeSH) terms.
- *Vocabulary of Expert Terms (VET)*: these are MeSH terms that do not belong to WordNet:
  $$\boldsymbol{VET} = (\boldsymbol{MeSH}) - (\boldsymbol{WordNet})$$
  It follows that VET contains 16,719 consumer (MeSH) terms.

Documents are represented by term vectors extracted by *AMTE*$_X$ or *MMT*$_X$ respectively. Each term in such a vector is represented by its weight. The term frequency-inverse document frequency model is used for computing the weight of each multi-word term: the weight $\boldsymbol{d_i}$ of a term $\boldsymbol{i}$ in a document is computed as $\boldsymbol{d_i} = \boldsymbol{tf_i} \cdot \boldsymbol{idf_i}$, where $\boldsymbol{tf_i}$ is the frequency of term $\boldsymbol{i}$ in a document and $\boldsymbol{idf_i}$ is the inverse document frequency of $\boldsymbol{i}$ in the whole document collection [7].

In this work, document categorisation by user profile is realized by computing the percentage of expert (VET) and consumer (VCT) terms in a document term vector. For example, a document with VET% = 0.62 has 62% probability of being a document suitable for experts.

3.1 Information Retrieval by User Profile in Medline

In the following, we design an information retrieval method capable of both i) ranking documents by similarity with a query, and ii) bringing documents matching a given user profile higher in the ranked list of similar documents.

As it is typical in information retrieval (IR), the similarity between a query $\boldsymbol{q}$ and a document $\boldsymbol{d}$ is computed by matching their term vectors according to VSM (Vector Space Model) [7]:

$$\boldsymbol{Document - Similarity} = \frac{\sum_i \boldsymbol{q_i} . \boldsymbol{d_i}}{\sqrt{\sum_i \boldsymbol{q_i^2}}\sqrt{\sum_i \boldsymbol{d_i^2}}}$$

where $\boldsymbol{i}$ denote terms in the query and the document and $\boldsymbol{q_i}$ and $\boldsymbol{d_i}$ are their $\boldsymbol{tf \cdot idf}$ weights in their respective vector representations. More specifically, the query is matched against all Medline documents and the returned list of documents is ranked by decreasing similarity. For ranking query results by user profile we distinguish between the following two cases:

- *Known user profile*: the user identifies her/himself as an expert (or consumer) prior to issuing a query. The similarity score by VSM is multiplied by its percentage of VET (or VCT) terms that is, its probability of being a document for experts (or consumer users respectively).
- *Unknown user profile*: the system determines her/his profile from the query. If the query contains at least one expert term, the user is considered to be an expert (a consumer otherwise). Retrievals are then processed similar to the previous case.

4. Experiments and Evaluation

We conducted two groups of experiments. The first set of experiments is designed to demonstrate the relative effectiveness of *AMTE*$_X$ and *MMT*$_X$ methods in indexing medical documents. The second group of experiments is designed to demonstrate the categorisation effectiveness of both *AMTE*$_X$ and *MMT*$_X$ in retrieving medical documents according to user profile.

The main data sources used in the experiments are:

- *PMC*: a corpus of 5,819 full documents from PubMed Central[11] indexed in Medline and selected out of 60 Journals. The documents were selected on the basis of having an identification (UID) number, which was used to retrieve their respective Medline index sets. This index set for each document is manually assigned by Medline experts.
- *OHSUMED*: it is a standard TREC[12] collection of 348,566 medical document abstracts from Medline, published between 1988 -1991. OHSUMED is commonly used in benchmark evaluations of IR applications. OHSUMED provides 64 queries and the relevant answer set (documents) for each query. The correct answers were compiled by the editors of OHSUMED and are also available from TREC. For the evaluations, we applied all 64 queries available.

Both, data store and access mechanisms are implemented using Lucene[13]. A document is indexed by the two-layered indexing structure of Figure 2 by mapping the terms of its document vector to their semantic categories of the Semantic Network at the top level and then, to their MeSH topics at the lower level.

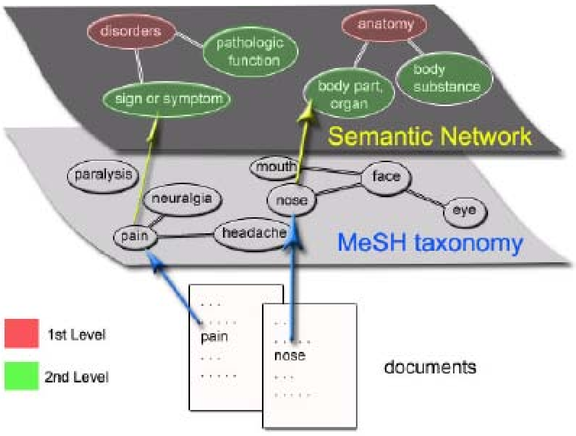

Figure 2 A two-layered indexing structure for Medline documents.

4.1 Term extraction experiment

In the following, $AMTE_X$ and MMT_X are evaluated in terms of precision and recall against the Medline provided MeSH index terms which constitute the ground truth. Because MMT_X is slow, a subset of the OHSUMED TREC collection is selected for this experiment consisting of 10% of OHSUMED (i.e., 34,000 documents). Because it is not possible for a statistically-based term extraction method such as $AMTE_X$ to work for abstracts, we treat the totality of the corpus as a single document during the extraction step. Subsequently, the extracted terms are associated with their respective source abstract. We also run the same experiment using the PMC full document corpus. Table 1 below summarize these results.

Table 1 Average precision and recall of $AMTE_X$ and MMT_X.

Data Set	Method	Number of Terms	Precision	Recall	Time (hours)
OHSUMED	$AMTE_X$	8	0.125	0.101	7.383
	MMT_X	40	0.089	0.336	14.516
PMC	$AMTE_X$	25	0.034	0.062	1.387
	MMT_X	72	0.033	0.162	2.727

[11] http://www.ncbi.nlm.nih.gov/pmc
[12] http://trec.nist.gov/data/t9_filtering.html
[13] http://lucene.apache.org

In both cases, *AMTE$_X$* performs much faster than *MMT$_X$*. This is due to the algorithmic simplicity of *AMTE$_X$* compared to *MMTx*, especially in regards to the variant generation phase of *MMT$_X$*. For OHSUMED, *AMTE$_X$* demonstrates improved precision and a reasonable recall compared to *MMT$_X$* by merely a fifth of the average term output of *MMT$_X$*. For PMC, the results are similar. Notice that, *MMT$_X$* is tuned towards higher recall (by revealing more indexing terms through an exhaustive variant generation phase).

4.2 Indexing by User Profile

In the following experiment we evaluate our document categorisation method of Section 3. We run a retrieval experiment on the full OHSUMED dataset using VSM [7]. Retrieval using vectors of *AMTE$_X$* and *MMT$_X$* terms is compared with retrieval using vectors of Medline provided MeSH terms (i.e., these terms are used as ground truth). For this experiment, the results were evaluated against all 64 TREC provided queries and answers; 15 out of the 64 queries contain no expert terms and are suitable for consumer users. The remaining queries are suitable for experts.

The objective is to measure the ability of a method in retrieving information for consumer and expert users respectively. We run this experiment twice, once for experts and once for consumer users. A method is deemed successful if it retrieves documents suitable for the particular type of users under consideration. Each method retrieves the best 20 answers for each TREC query, so that each plot below contains exactly 20 points (each method is represented by a curve). The top-left point of a curve corresponds to the average (over all queries) 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.

Figure 3 illustrates the relative performance of the three retrieval methods examined for the consumer retrieval task. Retrievals with the manually assigned MeSH terms perform better than any other method. This result reveals a tendency of the human indexers to assign simpler terms for the indexed documents achieving precision close to 65% for small answer sets with up to 10 answers. Both *AMTE$_X$* and *MMT$_X$* perform similarly (the *AMTE$_X$* method performs better than *MMT$_X$* for small answer sets).

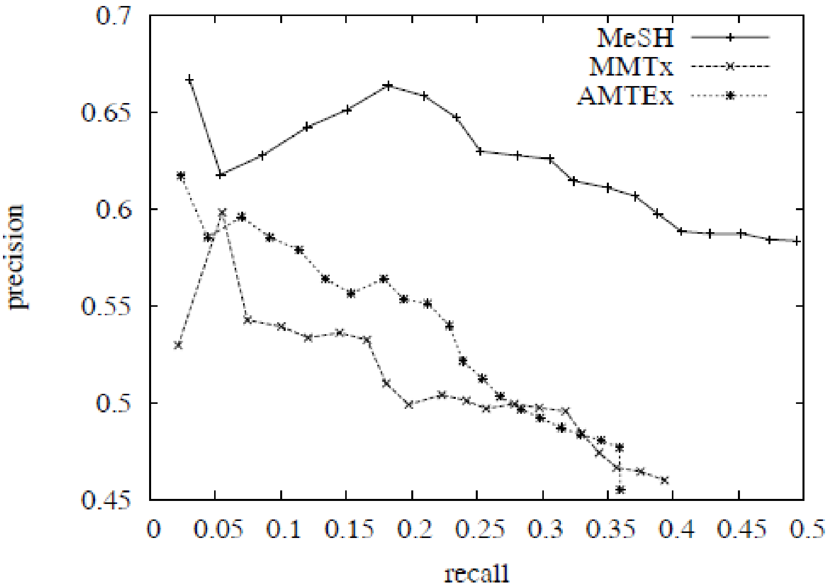

Figure 3 Average precision/recall of *AMTE$_X$* and *MMT$_X$* for the consumer users retrieval task.

Figure 4 illustrates results for the retrieval experiment for expert users. *AMTE$_X$* outperforms all other methods achieving precision up to 75% for small answer sets (i.e., with up to 3 answers). This experiment demonstrates the selective ability of *AMTE$_X$* towards extracting complex medical terms which can be found in the majority of Medline documents. It also reveals a weakness of manually assigning MeSH terms to documents as the human indexers may be not familiar with the content and complexity of domain specific medical publications in Medline.

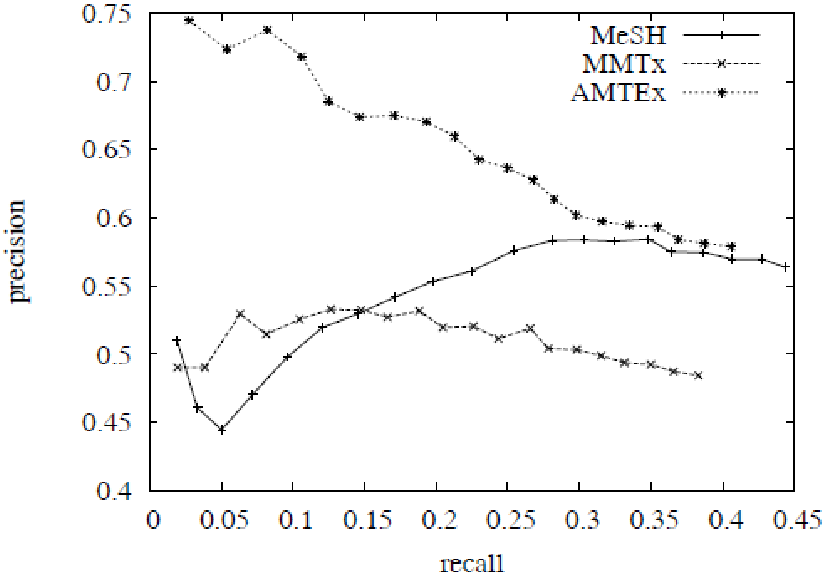

Figure 4 Average precision/recall of *AMTE*$_X$ and MMT for the expert users retrieval task.

4. Conclusions

We investigate potential improvements to the problem of indexing medical documents using *AMTE*$_X$ and we compare its performance against *MMT*$_X$ (the state-of-the-art method of the U.S. NLM). We also introduce to the research community the problem of automatic categorisation of medical information by user profile by investigating two common types of users of medical information (i.e., consumers and experts). Based on our experiments, we conclude that *AMTE*$_X$'s selective term output is very well suited for both problems, performing faster than *MMT*$_X$. However, *MMT*$_X$'s increased recall can be well suited in some retrieval cases, where the small document size is prohibitive for the optimal application of our *AMTE*$_X$ statistical term extraction process [1]. More elaborate experimentation is needed for confirming the performance of *AMTE*$_X$ for general medical collections, such as the Web. For the categorisation of medical documents by user profile problem, future work involves investigation of more elaborate classification methods such as machine learning, fuzzy clustering and document classification.

Acknowledgements

This work was supported by project RT3S (FP7-STREP, Project No 248801) of the European Union.

References

[1] Hliaoutakis, A., Zervanou, K., Petrakis, E. The *AMTE*$_X$ Approach in the Medical Document Indexing and Retrieval Application. Data and Knowledge Engineering 68 (3), 2009, 380–392

[2] Frantzi, K., Ananiadou, S., Mima, H. Automatic Recognition of Multi-Word Terms: The C-Value/NC-value Method. International Journal of Digital Libraries 3(2), 2000, 117–132

[3] Maynard, D., Ananiadou, S.: TRUCKS. A Model for Automatic Multi-Word Term Recognition. Journal of Natural Language Processing 8(1), 2000, 101–125

[4] Drymonas, E., Zervanou, K., Petrakis, E. Unsupervised Ontology Acquisition from Plain Texts: the OntoGain System. In: 15th Intern.l Conf. on Applications of Natural Language to Information Systems (NLDB'2010), Cardiff, Wales, UK, Springer, LNCS 6117, 2010, 277–287

[5] Bodenreider, O., McCray, A. Exploring Semantic Groups through Visual Approaches. Journal of Biomedical Informatics 36(6), 2003, 414–432

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

[7] Baeza-Yates, R., Ribeiro-Neto, B. Modern Information Retrieval. Addison Wesley Longman, 1999