Recommending medical documents by user profile

Kleanthi Lakiotaki, Angelos Hliaoutakis, Serafim Koutsos and Euripides G.M. Petrakis

Department of Electronic and Computer Engineering
Technical University of Crete

13th IEEE International Conference on BioInformatics and BioEngineering, Chania, 2013
1. Introductory Remarks
   - The problem
   - Motivation

2. Methods and tools
   - Language Processing
     - The Unified Medical Language (UMLS) System
     - The Medical Subject Headings (MeSH) method
     - The Automatic MeSH Term Extraction (AMTEx) method
     - Readability formulas
   - Classification
     - The UTilites Additives (UTA) method
     - Decision Trees

3. Proposed work
   - Medical Document Recommendation by User Profile

4. Concluding Remarks
Medical Document Classification

The Problem

- **Experts**: medical professionals, i.e. physicians, pharmacists, clinicians, nurses, e.t.c.
- **Consumers**: Non experts seeking medical information, i.e. patients, relatives of patients, e.t.c.
Medical Document Classification

Why?

- 72% of internet users say they looked on-line for health information of one kind or another. *National survey by the Pew Research Centres Internet & American Life Project in 2013*
Medical Document Classification

Why?

- 72% of internet users say they looked on-line for health information of one kind or another. *National survey by the Pew Research Centres Internet & American Life Project in 2013*

- Medical information systems such as MEDLINE are designed to serve health care professionals.
Medical Document Classification

Why?

- 72% of internet users say they looked on-line for health information of one kind or another. *National survey by the Pew Research Centres Internet & American Life Project in 2013*

- Medical information systems such as MEDLINE are designed to serve health care professionals.

- There is a gap between consumer and medical experts terminology, i.e. stomach ache vs. gastralgia.
Medical Document Classification

Why?

- 72% of internet users say they looked on-line for health information of one kind or another. *National survey by the Pew Research Centres Internet & American Life Project in 2013*

- Medical information systems such as MEDLINE are designed to serve health care professionals.

- There is a gap between consumer and medical experts terminology, i.e. stomach ache vs. gastralgia.

- An automatic system able to classify medical documents as "consumer specific", or "expert specific" is valuable to both groups, by significantly reducing their effort and time on information seeking task.
Unified Medical Language System
What is the UMLS?

- The **UMLS**, brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems.
- The **UMLS** has three tools:
  1. **Metathesaurus**: a large, multilingual thesaurus concerning biomedical and health related concepts from over a hundred source vocabularies.
  2. **Semantic Network**: Broad categories (semantic types) and their relationships (semantic relations).
  3. **SPECIALIST Lexicon and Lexical Tools**: Natural language processing tools
The UMLS-Semantic Network

What is the UMLS-Semantic Network?

- The **UMLS-Semantic Network** consists of:
  - a set of broad subject categories, or **semantic categories**, that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus,
  - a set of useful and **important relationships**, or semantic relations, that exist between semantic types.

- There are 133 semantic types (i.e. organism, biologic function, chemical, e.t.c.) and 54 semantic relations (i.e. isa, part of, location of, e.t.c.).
**Medical Subject Headings (MeSH)**

**What is the MeSH**

- **MeSH** is the U.S. National Library of Medicine’s controlled vocabulary thesaurus.
- It consists of sets of terms naming **descriptors** in a hierarchical structure that permits searching at various levels of specificity. There are 26,853 descriptors in 2013 **MeSH**.
- National Library of Medicine (NLM) indexers use **MeSH** to describe the subject content of journal articles for MEDLINE.
The AMTEx method [DKE 2009]

Automatic MeSH Term Extraction

- A term extraction approach for automatic indexing of medical documents
- **Automatic MeSH Term Extraction**

Main idea:

1. Initial term extraction based on a hybrid linguistic/statistical approach, the C/NC-value
2. Extracted terms are validated against MeSH
3. Extracts general single and multi-word terms
**Readability Formulas**

- A **readability formula** can be simply considered as a measure of the ease with which a document can be read.
- We applied the **Flesh Reading Ease Score**:
  \[
  FRES = 206.835 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right)
  \]
- and the **Flesch-Kincaid Grade Level**:
  \[
  FKG = 0.39 \left( \frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \left( \frac{\text{total syllables}}{\text{total words}} \right) - 15.59
  \]
- Documents with low readability score, are considered difficult to read.
The UTA method

How it works?

- The UTA method considers as input a weak-order ranking structure on a set of alternatives, together with the performances of the alternatives on a set of criteria.
- Returns as output a set of criteria weights (trade-offs), corresponding to the significance by which each criterion participates in the initial weak-order ranking structure.
- This is accomplished by means of special linear programming techniques.
EXPERIMENTAL SET-UP
DATA REPRESENTATION AND MODELING

UMLS Semantic Network Categories

- Organism
- Chemical Viewed Structurally
- Event
- Virus
- Disease or Syndrome
- Body Substance
- Embryonic Structure
- Antibiotic
- ...

D_1
- gastralgia
- lactase
- esophageal

D_2
- calcitonin
- retrovirus
- sickle

D_n
- receptor
- escherichia
- fibrosis

W_1
- W_2
- W_3
- ...
- W_n

LAKIOTAKI et. al

RECOMMENDING MEDICAL DOCUMENTS BY USER PROFILE

BIBE 2013 13/18
Our approach

\[
\begin{bmatrix}
    w_1 \\
    w_2 \\
    \vdots \\
    w_n
\end{bmatrix}
\cdot
\begin{bmatrix}
    a_1 \\
    a_2 \\
    \vdots \\
    a_n
\end{bmatrix}
\]

\(d_i \geq \text{thres}\)
EVALUATION RESULTS

Mean values of Readability formulas

<table>
<thead>
<tr>
<th></th>
<th>OHSUMED</th>
<th></th>
<th>PubMed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Consumer</td>
<td>Expert</td>
<td>Consumer</td>
<td>Expert</td>
</tr>
<tr>
<td>Flesch Reading Ease</td>
<td>34.31</td>
<td>29.45</td>
<td>40.17</td>
<td>37.03</td>
</tr>
<tr>
<td>Flesch Kincaid Grade</td>
<td>13.22</td>
<td>14.45</td>
<td>13.8</td>
<td>12.54</td>
</tr>
</tbody>
</table>

Classification accuracy measures

<table>
<thead>
<tr>
<th></th>
<th>OHSUMED</th>
<th></th>
<th>PubMed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UTA</td>
<td>Decision Trees</td>
<td>UTA</td>
<td>Decision Trees</td>
</tr>
<tr>
<td>Accuracy</td>
<td>95.80</td>
<td>83.75</td>
<td>70.73</td>
<td>63.00</td>
</tr>
</tbody>
</table>
Concluding Remarks

Summary and Conclusions

- We investigated the problem of automatic classification of medical information on two common types of users (consumers and experts) and showed that this problem cannot be solved by simply measuring readability easiness of the documents.
- On the contrary, we proved that:
  - by representing documents as vectors of AMTEx terms we achieve high classification accuracy,
  - the UMLS Semantic Network category terms can act as criteria for the categorization of a medical documents,
  - the UTA, successfully identifies the significance of Semantic Network category terms in their classification ability.
Concluding Remarks

Future work

- Investigate the role of other language tools in classifying medical documents, such as n-grams.
- Classify medical documents according to thematic categories (i.e. pneumonia, cancer, e.t.c.).
Thank you for your attention

CONTACT INFORMATION

http://www.intelligence.tuc.gr/