

Lexically-based distinction of readability levels of health documents

Antoine BORST, Arnaud GAUDINAT, Natalia GRABAR, Célia BOYER
Health on the Net Foundation, SIM/HUG, 24 rue Micheli-du-Crest, Geneva, Switzerland

1. Presentation

Problems:

Increasing number of Internet users look for online health information. However, easy to read health documents and very complex documents both co-exist on the web. Users may be unable to understand information they found, even if it's of good quality.

Objectives:

- Guide users to medical content that is suitable to their education level
- Integrate readability information in our search engines to facilitate access to appropriate health information
- Help medical content editors reach their target audience

Approach:

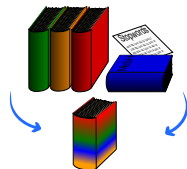
Documents' readability is calculated upon words complexity, sentence length and, in some aspects, lexical density while classical approach are based on words length or syllable number, in addition to sentence length.



2. Material and method

Material used:

- English **generic wordlists** of different sizes (from 12,000 to 264,000 words, including inflections)
- 22,990 medical terms from the **MeSH** (Medical Subject Headings from the National Library of Medicine, Bethesda, USA)
- a list of 27 stopwords
- 1,000 documents to evaluate and 400 documents to calibrate our method



Lexicon construction:

- A large lexicon that associates **words with an estimated complexity** is needed
- Postulate is that **complexity is correlated with rarity**
- Estimating rarity is possible using wordlists of different **exhaustiveness rates**: hypothesis is that a word appearing in a smaller list is very likely to be more common than another showing up only in larger lists (eg. *fever*). MeSH terms appearing in no other lists represent the top complexity (eg. *trichostrongyloidiasis*)

Document scoring:

- Documents complexity is a **weighted mean of words complexity** (consisting of values from 1 to 100)
- Weighting of words decreases upon new similar occurrences: it takes into account the **lexical density**
- Finally the **lexical mean** is combined with a **sentence length** scoring

Categorization:

- Last step is establishing threshold to categorize documents upon their scores
- Two levels were retained: easy and difficult
- The threshold is identified from learning corpora of these two classes

3. Results and perspectives

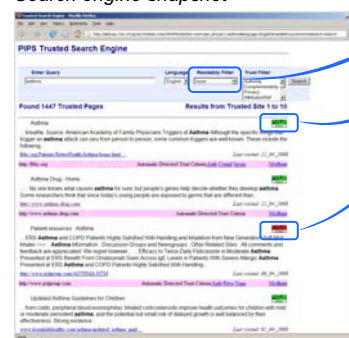
Results:

- Evaluation corpus consists of 500 medical documents especially designed to target non-expert citizens and 500 documents with a high technicality level targeting medical professionals
- Average **accuracy reaches 92%** on the evaluation corpus
- Accuracy is slightly better for easy documents

Perspectives:

- Extend the tool to other languages only by using a large corpus of words
- Add other linguistic parameters and optimize the scoring function
- Try to exploit stemming ability to extend recognition of same terms

Search engine snapshot



Readability filter

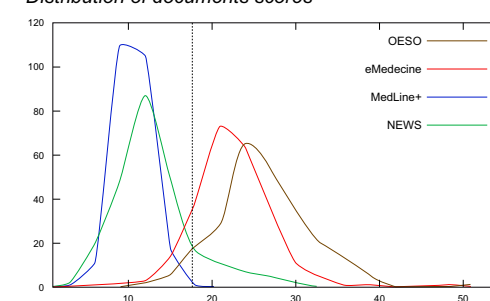
Easy document

Difficult document

Example of integration in one of our health search tool



Distribution of documents scores



Difficult documents: OESO (www.oeso.org) is a foundation specialized in oesophagus diseases, their documents are mostly very scientific; eMedicine is an online resource from WebMD (health portal: www.webmd.com) designed for medical experts.

Easy Documents: Documents from NEWS are coming from the medical news website "HealthDay" (www.healthday.com) which targets laypersons; MedLine+ is an online encyclopedia of consumer health information brought by the National Library of Medicine.

On the vertical axis are number of documents; on the horizontal axis is documents' score.

Vertical dotted line is the threshold between easy (left side) and difficult (right side) documents