Mining Officially Unrecognized Side effects of drugs by combining Web Search and Machine learning

Carlo Carino, Yuanyuan Jia, Bruce Lambert, Patricia West and Clement Yu

University of Illinois at Chicago

Motivation

- Drugs have side effects
- Food and Drug Administration (FDA) requires drug companies to do extensive clinical trials before a drug enters the market place
- Not all side effects of a given drug are officially recognized by the FDA

  Recent case: Vioxx

Objective

- Find unrecognized side effects of drugs

Approach

- Submit a query (drug name, side effect) to a Web search engine
- Extract from the retrieved pages all side-effects which are found in those pages

Problems associated with this approach

(1) May not retrieve enough relevant documents

(2) May retrieve a lot of irrelevant documents

(3) May not want to have entire documents

Want to have more relevant documents

Modify the query to become

< drug name OR active ingredients>

Active ingredients = the chemical compounds forming the drug

Want to reduce the number of irrelevant documents

Web retrieval

|

|

Classification

(machine learning algorithm)

Machine Learning algorithm

Neural network

Many features: words such as

actual side-effects; “side-effects”.

“adverse effects”; “safe”, “not” etc

Hundreds of such features

Training phase, then test phase

Train on 7 drugs: identify the relevant and irrelevant pages retrieved for these 7 drugs;

Test on 20 other drugs

Results

Each drug retrieves 100 pages from Google;

Using the classification algorithm only 16.4 pages/drug are retained by our system;

Average precision-accept: 90.3%

Average precision -reject: 87.5%

Average precision of top 17 pages for Google: 61.2%

Reduce the amount of manual efforts for collecting training data

Generate “positive examples”

and “negative examples” automatically with high probabilities

Automatic generation of training data

Find a drug, d, such that the diseases it treats, T, are disjoint from its known side-effects, S.

A retrieved page in response to

<d, t in T>, if does not contain any known side effect S is “negative”;

A retrieved page in response to <d, s in S>, if does not contain “safe” or “not” in the vicinity of s is “positive”.

Accuracy of generating training examples automatically

Accuracy of generating 50 positive examples: 98%

Accuracy of generating 50 negative examples: 96%

Validation of unrecognized side-effects

Validated by licensed pharmacist and drug information specialist

Prilosec: pneumonia

Accutane: Watery eye

Uroxatral: Yellowing of skin or eyes

Retrieve passages instead of pages

For each passage of certain number of words, compute the degree of “relevance”.

Output the passage with highest relevance, if it exceeds a threshold.

Also output frequencies of side-effects

Summary

Proposed system can retrieve unrecognized side-effects of drugs

Improve accuracy of retrieve;

Need a good medical dictionary to recognize highly related side effects

for example, nausea and vomiting

Extensions

Our approach, retrieval followed by classification, can be applied to other retrieval problems:

Examples:

(1) Complications of medical procedures/ operations;

(2) Processing of queries of specialized types