A proposal for determining the evidence types of biomedical documents using a drug-drug interaction ontology and machine learning

Linh Hoang1, Richard D. Boyce2, Mathias Brochhausen3, Joseph Utech3, Jodi Schneider1

1 University of Illinois at Urbana-Champaign 2 University of Pittsburgh 3 University of Arkansas for the Medical Sciences

Background

• Knowledge of potential drug-drug interactions (PDDIs) is important for clinicians to make safe treatment decisions.
• It is challenging for clinicians to keep track of new knowledge about PDDIs due to a large amount of new research about PDDIs from a variety of resources including journal articles and drug labels e.g., 4,71 new papers about DDIs in PubMed published in 2017 [1].
• PDDI knowledge tends to exist in silos requiring search and synthesis by drug experts. However, experts disagree about how to search and evaluate PDDI evidence.
• We propose to combine machine learning with a formal representation of the DDRI knowledge in order to assist humans in the process of searching, assessing, and summarizing PDDI evidence for clinical use.

Approach

We are building a hierarchical classifier, which is a combination of multiple sub-classifiers that automatically predict the specific type of DIDE ontology provided in scientific documents, based on a set of 44 evidence types formally defined in the DIDEO ontology.

Error Analysis

Examples of the most informative unigrams for each evidence type

<table>
<thead>
<tr>
<th>Evidence type</th>
<th>Non randomized (non parallel) DDI Clinical Trial</th>
<th>Non randomized parallel DDI Clinical Trial</th>
<th>Randomized DDI Clinical Trial</th>
<th>Non-polymorphic enzyme/Transport PK Trial</th>
<th>Genotyped PK Trial</th>
<th>Phenotype PK Trial</th>
</tr>
</thead>
<tbody>
<tr>
<td>confidence</td>
<td>outreach</td>
<td>crossover</td>
<td>accumulation</td>
<td>carriers</td>
<td>undetectable</td>
<td></td>
</tr>
<tr>
<td>concentration</td>
<td>output</td>
<td>randomized</td>
<td>fitted</td>
<td>hydroxyclarithromycin</td>
<td>primer</td>
<td>nesiritide</td>
</tr>
<tr>
<td>furosemide</td>
<td>diuretic</td>
<td>placebo</td>
<td>ephetol</td>
<td>polymorphisms</td>
<td>interpreted</td>
<td></td>
</tr>
<tr>
<td>significant</td>
<td>coinciding</td>
<td>Double</td>
<td>elderly</td>
<td>genetic</td>
<td>oxotremorine</td>
<td></td>
</tr>
</tbody>
</table>

DIDE Ontology (viewed in Protégé)

The DIDEO Ontology is a formal representation of DDRI knowledge.

Implementation

Stage 1: Data Preparation

• Collect and annotate data: An existing dataset which contains 189 PDDI evidence items that were partially annotated with the evidence types in a previous study. We had an expert further annotate papers, resulting in a manual gold standard of labels.
• Preprocess data: Automatically collected the papers’ metadata (title, abstract, publication type) through the PubMed API. Manually collected full-text PDFs and programmatically converted them to plain text.

Stage 2: Classifier Development

• Features: Stemmed TF-IDF of unigrams taken from the titles, from abstracts and from the Methods sections; drug entities from the titles and abstracts as detected by MetaMap.
• Machine learning model: All sub-classifiers are trained and tested using Support vector machine (SVM), cross validation (5 folds), class weighting mechanism. All papers are used to train and test the top-level sub-classifier. A subset of the dataset from the top-level classifier are used to train and test the next level sub-classifiers.
• Evaluation metrics: ROC AUC, precision, recall and F1-score.

Classification Results

Sub-classifiers’ prediction performance:

<table>
<thead>
<tr>
<th>Classifier</th>
<th>ROC AUC</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1 – Clinical vs. PK trials</td>
<td>0.74</td>
<td>0.88</td>
<td>0.87</td>
<td>0.85</td>
</tr>
<tr>
<td>Level 2 – Randomized vs. non-randomized</td>
<td>0.96</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Level 2 – Polymorphic vs. non-polymorphic enzyme/transport</td>
<td>0.95</td>
<td>0.96</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>Level 3 – Parallel vs. non-parallel</td>
<td>0.95</td>
<td>0.87</td>
<td>0.93</td>
<td>0.90</td>
</tr>
<tr>
<td>Level 3 – Genotyped vs Phenotype</td>
<td>0.95</td>
<td>0.81</td>
<td>0.82</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Future Work

• Expand the existing classifiers to other evidence types in the DIDEO ontology.
• Run the classifiers on new articles (e.g. from PubMed Central) to get predictions of PDDI evidence types as well as to identify potentially new evidence types.
• Run the model with different features: MedShar terms, drug entity recognizer.

References


Partially supported by NLM R01LM010817 and NIH R01LM011838.