IDENTIFYING DISEASE-RELATED EXPRESSIONS IN REVIEWS USING CONDITIONAL RANDOM FIELDS

Zulfat Miftahutdinov
Elena Tutubalina
Kazan Federal University

Alexander Tropsha
University of North Carolina

2 june 2017
Introduction

- The explosive growth of social media
- Valuable information can be found in social media
- Including for drug repurposing
Drug repurposing is the application of known drugs and compounds to treat new indications (i.e., new diseases).

Examples:

Yaz – first approved for pregnancy prevention, now also used for moderate acne vulgaris

Trazadone – Originally trialed as antidepressant unsuccessfully, now used as sleep aid
Solution stages:

- Extracting disease-related expressions
- Normalization to the medical concepts
- Sentiment analysis
- Relationship extraction
- Set up a repurposing hypothesis
Drug repurposing

Albuterol has me feeling all sorts of dizzy and weak this morning 😞 😞

6:21 AM - 17 July 2015

- Asthenia
  - MedDRA: 10003549

- Dizziness
  - MedDRA: 10013573
Dataset

CADEC corpus

- contains 1250 posts
- 1799 Drug entities
- 6752 Disease entities
Methods

- Dictionary based
- Conditional Random Fields
- Bidirectional Gated Recurrent Unit
- Bidirectional Long Short Term Memory
Hand-crafted features in Conditional Random Fields

- $w$ – Word
- $pos$ – Part-of-speech tag
- $sp$ – Suffix and Prefix
- $context$ – Context
- $wtype$ – Word Type
- $dict$ – Dictionary Look-up
- $b$ – Cluster-based representation
- $emb$ – Word embeddings
Word embeddings

- PubMed word2vec embeddings
- word2vec trained on domain specific reviews

<table>
<thead>
<tr>
<th>Data Source</th>
<th>reviews count</th>
<th>tokens count</th>
</tr>
</thead>
<tbody>
<tr>
<td>webmd.com</td>
<td>284 055</td>
<td>20 794 273</td>
</tr>
<tr>
<td>askapatient.com</td>
<td>113 836</td>
<td>13 649 150</td>
</tr>
<tr>
<td>patient.info</td>
<td>1 472 273</td>
<td>160 750 980</td>
</tr>
<tr>
<td>dailystrength.org</td>
<td>214 489</td>
<td>13 880 025</td>
</tr>
<tr>
<td>drugs.com</td>
<td>93 845</td>
<td>9 191 434</td>
</tr>
<tr>
<td>amazon</td>
<td>428 777</td>
<td>36 499 681</td>
</tr>
</tbody>
</table>

https://github.com/dartrevan/ChemTextMining
Dictionaries

- UMLS dictionary
- Manually validated terms from UMLS
- ADR lexicon
- Multi-word expressions dictionary
- Drug names dictionary
<table>
<thead>
<tr>
<th>Method</th>
<th>Exact matching</th>
<th></th>
<th>Partial matching</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
</tr>
<tr>
<td>Dictionary-based</td>
<td>.503</td>
<td>.502</td>
<td>.494</td>
<td>.836</td>
</tr>
<tr>
<td>3-layer LSTM</td>
<td>.718</td>
<td>.629</td>
<td>.670</td>
<td>.801</td>
</tr>
<tr>
<td>3-layer GRU</td>
<td>.735</td>
<td>.629</td>
<td>.678</td>
<td>.793</td>
</tr>
<tr>
<td>CRF</td>
<td>.702</td>
<td>.680</td>
<td>.691</td>
<td>.852</td>
</tr>
</tbody>
</table>
Vyvance (Lisdexamfetamine Dimesylate)

Approval history: 2007 – Attention-Deficit/Hyperactivity Disorder. 2015 – Moderate to Severe Binge Eating Disorder (BED)


In science: The first clinical study of Lisdexamfetamine in Binge Eating Disorder was started in January 2010.
Conclusion

- The idea is applicable
- CRF is better in exact matching
- Embeddings and Dictionaries can be found at https://github.com/dartrevan/ChemTextMining