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

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

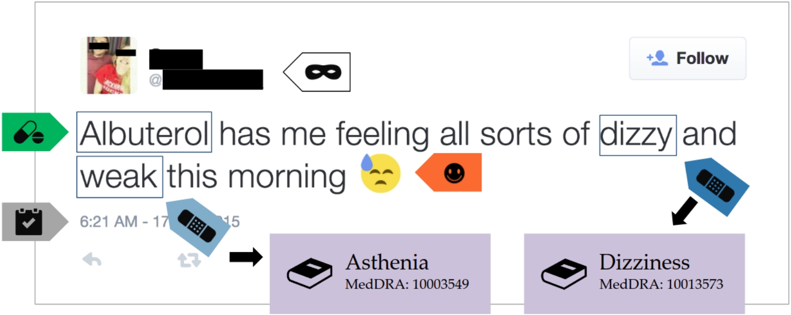
Trazadone – Originally trialed as antidepressant unsuccessfully, now used as sleep aid

Drug repurposing
using Text Mining

Solution stages:

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

Drug repurposing



[redacted]

@[redacted]

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

6:21 AM - 17 … 15

Asthenia
MedDRA: 10003549

Dizziness
MedDRA: 10013573

Dataset

CADEC corpus

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

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

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

https://github.com/dartrevan/ChemTextMining

Dictionaries

- UMLS dictionary
- Manually validated terms from UMLS
- ADR lexicon
- Multi-word expressions dictionary
- Drug names dictionary

Results

Method	Exact maching			Partial maching		
	P	R	F1	P	R	F1
Dictionary-based	.503	.502	.494	.836	.546	.625
3-layer LSTM	.718	.629	.670	.801	.872	**.812**
3-layer GRU	**.735**	.629	.678	.793	**.876**	.811
CRF	.702	**.680**	**.691**	**.852**	.790	.794

Retrospective analysis

proof of concept

Vyvance (Lisdexamfetamine Dimesylate)

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

Extracted from social media: decrease in appetite (2007). appetite decreased , appetite suppression, no appetite (2008).

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