Structured Information Extraction from Pharmaceutical Records

Michaela Bamburová, Zuzana Neveřilová

Faculty of Informatics, Masaryk University

December 6, 2019

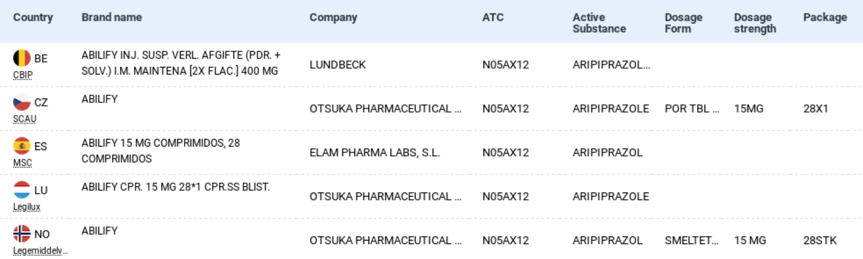
Table of contents

- Introduction
- Data characteristics
- Used methods
- Experiments
- Error analysis
- Results and Future work

Introduction

- *Goal*: Split data with entities such as drug name, dosage strength, dosage form, and package size into the appropriate columns.
- *Issue*: The data is provided by many suppliers are very diverse in terms of structure.
- *Approach*: Rule-based and machine learning methods for parsing the data
    - Iteratively extend the training data set using regular expressions and conditional random fields together with manual corrections.

Structured vs. unstructured data



Country	Brand name	Company	ATC	Active Substance	Dosage Form	Dosage strength	Package
BE CBIP	ABILIFY INJ. SUSP. VERL. AFGIFTE (PDR. + SOLV.) I.M. MAINTENA [2X FLAC.] 400 MG	LUNDBECK	N05AX12	ARIPIPRAZOL...			
CZ SCAU	ABILIFY	OTSUKA PHARMACEUTICAL ...	N05AX12	ARIPIPRAZOLE	POR TBL ...	15MG	28X1
ES MSC	ABILIFY 15 MG COMPRIMIDOS, 28 COMPRIMIDOS	ELAM PHARMA LABS, S.L.	N05AX12	ARIPIPRAZOL			
LU Legilux	ABILIFY CPR. 15 MG 28*1 CPR.SS BLIST.	OTSUKA PHARMACEUTICAL ...	N05AX12	ARIPIPRAZOLE			
NO Legemiddelv...	ABILIFY	OTSUKA PHARMACEUTICAL ...	N05AX12	ARIPIPRAZOL	SMELTET...	15 MG	28STK

Data characteristics

3 categories:

- YELLOW – structured data (103 thousand records)
- BLUE – semi-structured data (1.1 million records)
- GREY – unstructured data (> 2 million records)

Data characteristics

- data mostly in English, but also in Spanish, Finnish, Czech, etc.
- various abbreviated words (e.g., *ml*, *mg*, *inj sol*, *tabl*, *tbl*, *filmtabl*)

Used methods

- Regular expressions
- Conditional Random Fields

Experiments

- first training set only from the data from the yellow category
    - weak predictions on data sets in languages not covered in the training set
    - good predictions of BRAND NAME and DOSAGE STRENGTH
    - overfitting
- tunning regularization parameters
- tunning CRF features
    - features for 2 words before and after the current one
    - features such as is_unit() and is_punctuation() for words

Dosage form feature weights after the first and the last experiment

y='dosage form'	top features
Weight	Feature
+10.693	word.lower():tablet
+7.471	-1:word.lower():surepal
+7.469	word.lower():tabletės
+7.270	word.lower():gelis
+6.921	-1:word.lower():trockensub
+6.780	+1:word.lower():peritonealdialysvätska
+6.513	+1:word.lower():infusionsvätska
+6.176	word[-3:]:eet
+6.061	-1:word.lower():stk
+5.872	word[-3:]:tfl
+5.808	word.lower():tabletė
+5.797	+1:word.lower():injektionsvätska
+5.555	word.lower():por
+5.555	word[-3:]:por
+5.501	word[-3:]:tti
+5.407	word.lower():capsule
+5.243	word.lower():krem
+5.221	word[-3:]:kum
	…
-9.010	word.isdigit()

y='dosage form'	top features
Weight	Feature
+6.280	word.lower():doz
+6.032	word.lower():ampul
+5.625	word[:-3]:table
+5.498	+2:word.lower():er
+4.883	word.lower():cozeltisi
+4.879	word[-2:]:dr
+4.620	-2:word.lower():dos
+4.609	word[-3:]:gas
+4.464	word.lower():setli
+3.898	word[-3:]:tfl
+3.830	word[:-3]:kap
+3.730	word.lower():cozelti
+3.730	word[:-3]:coze
+3.549	word[-2:]:yr
+3.542	word.lower():krem
+3.345	word.lower():flakon
+3.311	word.lower():hard
	…
-3.701	word.isdigit()
-3.873	-1:word.lower():doz

Error analysis

- incorrect PACKAGE SIZE prediction

drug name	dosage strength	package size	dosage form
viron	200 mg	70	kapsul
viron	200 mg 168		kapsul

Error analysis

- order of values to be predicted

Table: Example of input data

	package size
medroxyprogesterone acetate 150 mg/ml inj,susp	1ml

Table: Incorrectly predicted data

drug name	dosage form	other	package size
medroxyprogesterone acetate 150 /	inj, sus	mg ml	1 ml

Error analysis

- unknown words
- languages not covered in the training set

Results

- data from all yellow, 4 blue, and 2 gray data sets
- 5-fold cross-validated results
- F1 score: 95%

	precision	recall	f1-score	support
dosage form	0.95	0.94	0.94	388,489
drug name	0.94	0.92	0.93	257,298
other	0.93	0.91	0.92	98,360
package size	0.94	0.94	0.94	391,810
dosage strength	0.96	0.98	0.97	551,230
accuracy			0.95	1,687,187
macro avg	0.94	0.94	0.94	1,687,187
weighted avg	0.95	0.95	0.95	1,687,187

Future work

- improvements of CRF model
- experiments with recurrent neural networks (also together with CRF)

Thank You for Your Attention!

MUNI
FACULTY
OF INFORMATICS