Identifying Drug Related Events from Social Media

Jeongho Noh
Jisu You
Yoonju Lee
Woo Jin Kye
Sungho Kim

CS4624 Multimedia, Hypertext, and Information Access

Professor: Edward A. Fox
Client: Weiguo Fan, Long Xia

May 2, 2017

Virginia Tech, Blacksburg, VA 24061

Innovative information system and processing steps

(Crawl social network reviews on drugs that are used to treat diabetes - by client)

1. Label the crawled data manually
2. Generate side effect dictionary to recognize side effect entities.
3. Visualize the resulting information for doctors and patients
4. Create confusion matrix to see result

1. Data Labeling

- Manual labeling is necessary to build a problem specific dictionary.
- Labeled about 235,000 words for named entity recognition.

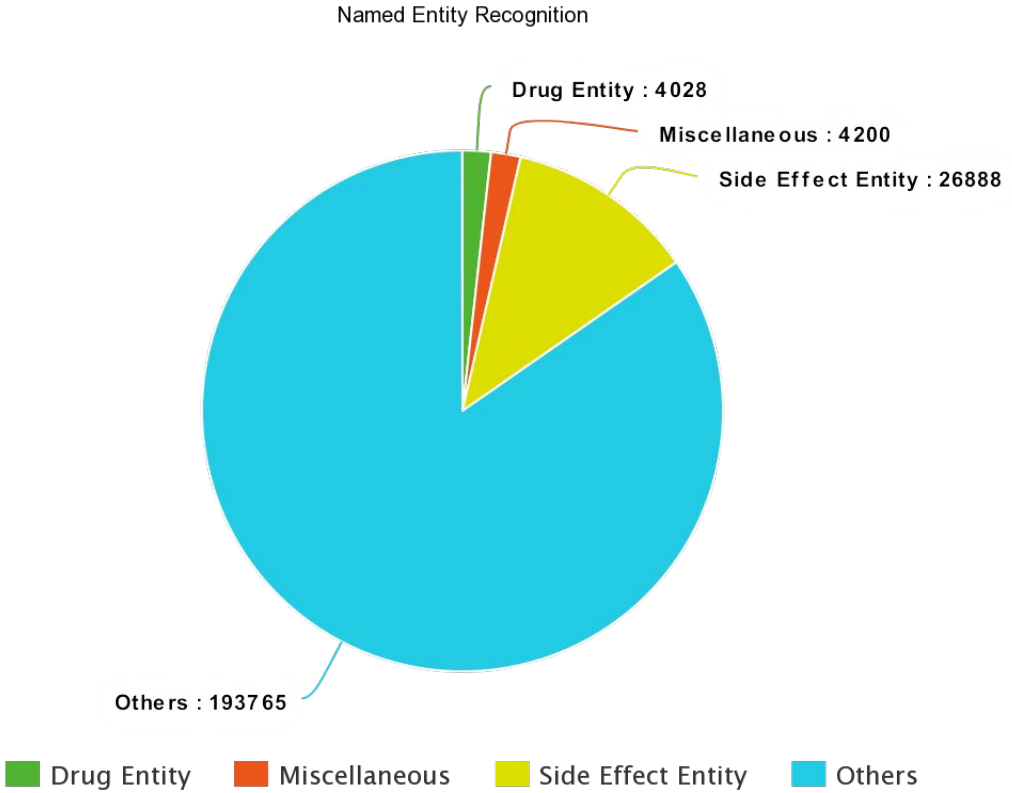
Words	Words
My	I
doctor	had
tried	swelling
switching	,
me	and
from	I
lantus	could
to	barely
toujeo	feel
because	my
my	toes
blood	due
sugar	to
was	poor
always	circulation
high	
in	
the	
morning	

1. Data Labeling - Named Entity Recognition

- Four different labels for different entities:
  - D – drug entity
  - S – side effect entity
  - M – miscellaneous medical terms that are not a drug entity or a side effect entity
  - O – others

Words	Label	Words	Label
My	O	I	O
doctor	O	had	O
tried	O	swelling	S
switching	O	,	O
me	O	and	O
from	O	I	O
lantus	D	could	O
to	O	barely	S
toujeo	D	feel	S
because	O	my	S
my	O	toes	S
blood	M	due	O
sugar	M	to	O
was	O	poor	M
always	O	circulation	M
high	O		
in	O		
the	O		
morning	O		

1. Data Labeling - Named Entity Recognition Cont.



- There are a total of 2242 unique side effect entities and 412 unique drug entities out of 228881 named entities.

2. Generating Side Effect Dictionary - Smoke List

Words	Label	Side Effect Prevalence Score	Words	Label	Side Effect Prevalence Score
My	O	398274	I	O	0
doctor	O	0	had	O	915934
tried	O	0	swelling	S	2104240
switching	O	0	,	O	N/A
me	O	0	and	O	1297179
from	O	0	I	O	0
lantus	D	0	could	O	259731
to	O	0	barely	S	420947
toujeo	D	0	feel	S	885157
because	O	0	my	S	398274
my	O	398274	toes	S	657309
blood	M	0	due	O	0
sugar	M	0	to	O	0
was	O	0	poor	M	94930
always	O	52076	circulation	M	123553
high	O	28234			
in	O	893558			
the	O	0			
morning	O	16780			

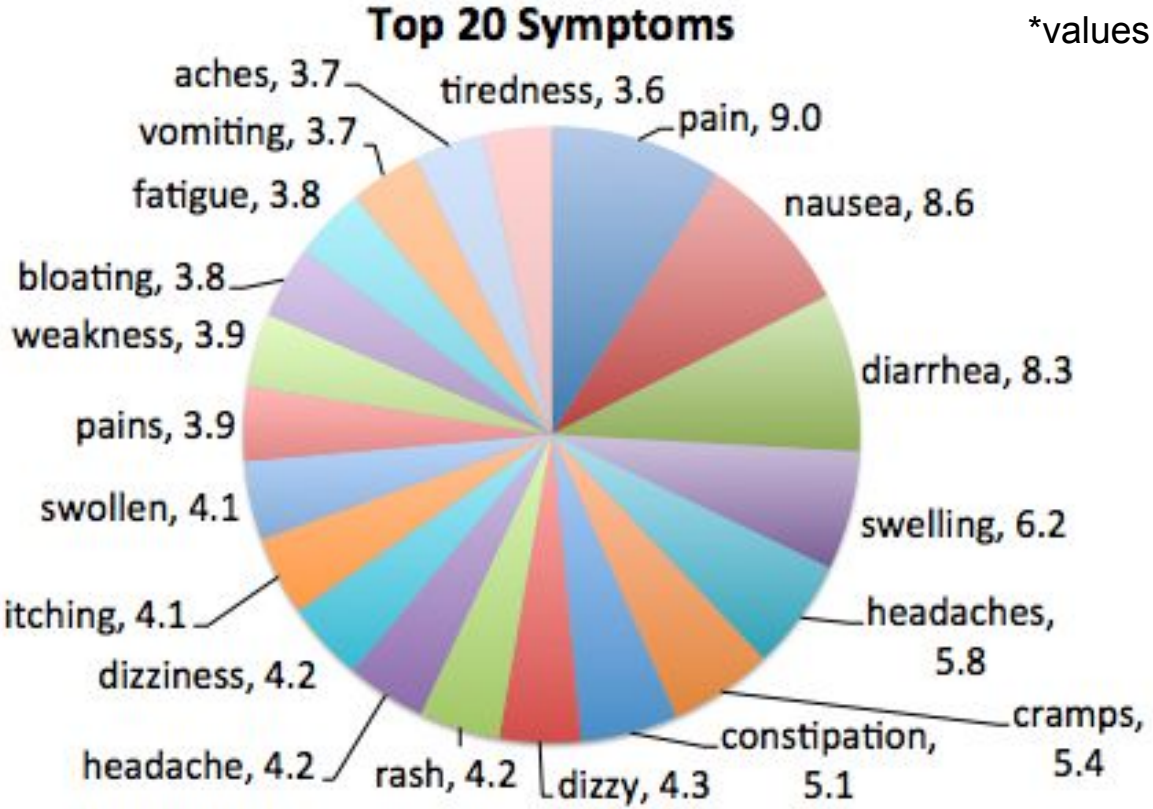
2. Generating Side Effect Dictionary - Filtering

Words	Label	Side Effect Prevalence Score	Actual Side Effect?	Words	Label	Side Effect Prevalence Score	Actual Side Effect?
My	O	398274	n	I	O	0	
doctor	O	0		had	O	915934	
tried	O	0		swelling	S	2104240	y
switching	O	0		,	O	N/A	
me	O	0		and	O	1297179	
from	O	0		I	O	0	
lantus	D	0		could	O	259731	
to	O	0		barely	S	420947	y
toujeo	D	0		feel	S	885157	y
because	O	0		my	S	398274	n
my	O	398274	n	toes	S	657309	y
blood	M	0		due	O	0	
sugar	M	0		to	O	0	
was	O	0		poor	M	94930	y
always	O	52076	y	circulation	M	123553	y
high	O	28234	y				
in	O	893558	y				
the	O	0					
morning	O	16780	y				

2. Generating Side Effect Dictionary - Result

ID	Word	Side Effect Score from NER Training
1	stomach	3420739
2	severe	3067356
3	pain	3040816
4	nausea	2884258
5	diarrhea	2802207
	...	
2072	vile	4138
2073	simple	4138
2074	steadily	4138
2075	appears	4138
2076	hormone	4138

3. Visualization



*values are in percentage

4. Validation

- Confusion Matrix
  - summary of the result.
- Select two hundred reviews from the list of 5585 reviews labeled by PamTAT
  - 100 from the top and 100 from the bottom of the list.
- Hypothesis: All the reviews from the top of the list will contain a mention of side effects, and the reviews from the bottom of the list will not.

	PamTAT Label	
Row Numbers	**No Side Effect**	**Side Effect**
Bottom 100	98	2
Top 100	6	94
Grand Total	**104**	**96**

Demo

https://www.youtube.com/watch?v=QYgBhBQtGsQ