

Recognition and Interactions of Drugs

Advanced Human Languages Technologies

Pau Rodriguez Asaf Badouh

Universitat Politècnica de Catalunya
MIRI - Data Science

May 2018



Outline

- 1 Introduction
- 2 Approach
- 3 Architecture
- 4 Features
- 5 Experiments
- 6 Conclusion



Drug name recognition (DNR) is a critical step for drug information extraction. DNR is the first step to identifying unknown drug interactions (DDI). DDI is broadly described as a change in the effects of one drug by the presence of another drug. Because of the lack of labeled corpora, early studies on DNR are mainly based on ontologies and dictionaries. To promote the research on drug information extraction, MAVIR research network and University Carlos III of Madrid in Spain organized two challenges successively: **DDIExtraction 2011** and **DDIExtraction 2013**. Both of the two challenges provide labeled corpora that can be used for machine learning-based DNR.

We are presenting several approaches to solve **SemEval-2013 Task 9.1 and Task 9.2**.



Overview:

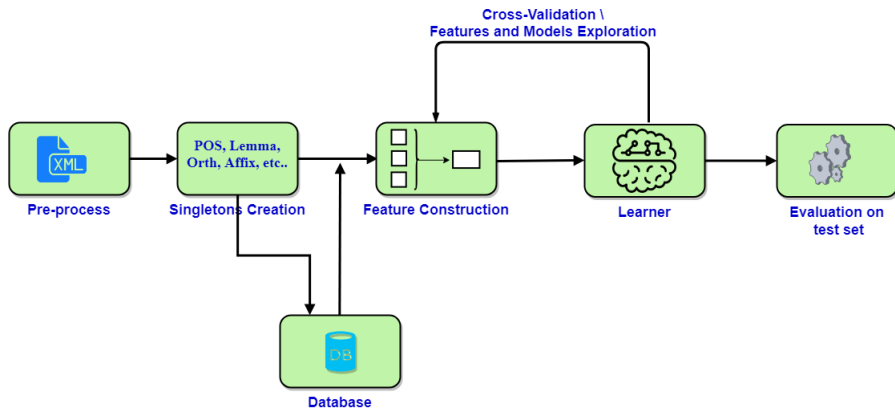
Transform and parse the input XML data into new user friendly data base (Json based).

- **Drug Name Recognition:** For each word - create Features vector with information we extract.
- **Drug-Drug Interaction:** For each Drug pair in sentence - Create feature vectors.

Train and examine different learners (SVM and CRF).

- **Sources of inspiration:**
 - Reference papers.^[1, 2, 3, 4]

Solution's Architecture



Learners:

Support Vector Machine (linear kernel)

Conditional Random Fields

Software:

Scikit-Learn, sklearn-crfsuite

Freeling, NLTK, DrugBank^[5], PubMed^[6, 7]

PyStatParser, Maltparser



Features

DNR features

f_1	Word feature	
f_2	Part-Of-Speech	
f_3	Lemmatization	
f_4	Orthographic - basic features	5 classes: all-capitalized, is-titlecase, all-digits, alphanumeric, hyphen or not
f_5	Orthographic - affix features	Prefixes and suffixes of the length 3,4,5
f_6	Orthographic - Word shape features	Generalized word class: Xxxxx00xxOxx Brief word class: Xx0xOx
f_7	Dictionary features	Lookup on Drugbank
f_8	Chunk feature	NLTK noun phrase chunking tag
f_9	Word Embeddings features	Word2Vec and classification

Drug-Drug Interaction Features

f_1	All DNR SVM features	for each entity of the pair
f_2	All DNR SVM features in a window	for each word in a window 0 to 5
f_3	Appearance of most frequent 3-gram sequences	from dependency tree shortest path
f_4	Appearance of most frequent Word/Lemma	
f_5	Appearance of most frequent POS	from CFG parse tree shortest path
f_6	Word count in the shortest path	from dependency tree shortest path
f_7	Counts of POS tags in sentence	specific tags: VB, CC, MD, DT

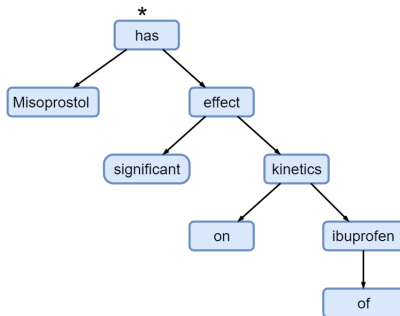
Word embedding features based on Word2Vec model, We used pre-processed database from Pubmed^[7].

- Vector feature:
 - Word mapped to 200 dimension vector: (0.010, ..., 0.023)
 - Failed to train the model due to hardware limitations.
- Word cluster feature:
 - Use Kmean to cluster words based on their vectors.
 - Reduce the feature from 200 dimension to 1.
 - Failed to do the clustering due to hardware limitation.

Features

Task9.2 DDI - Dependency Tree

Misoprostol has significant effect on kinetics of ibuprofen



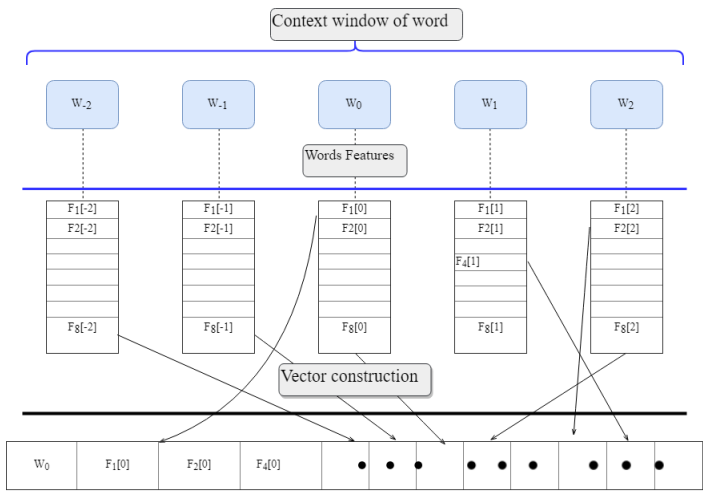
ShortestPath(Misoprostol, ibuprofen) = [has, effect, kinetics]

LenthOfShortestPath = 3

3-grams: [Misoprostol, has, effect], [has, effect, kinetics], [effect, kinetics, ibuprofen]

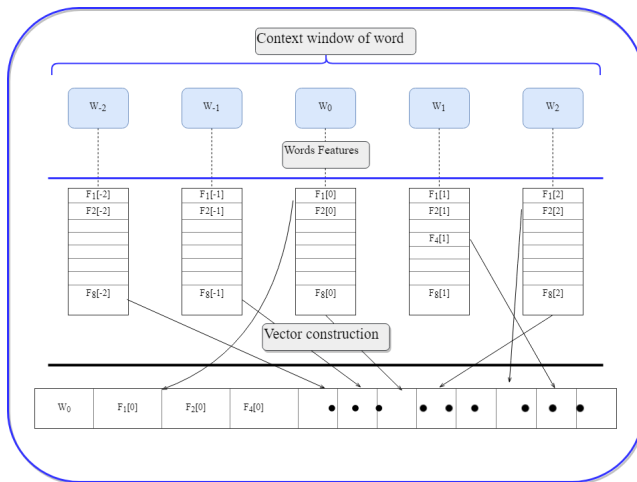
Task9.1 DNR - SVM

Based on the singleton features, we explore different combinations of vectors to feed the learners:



Features

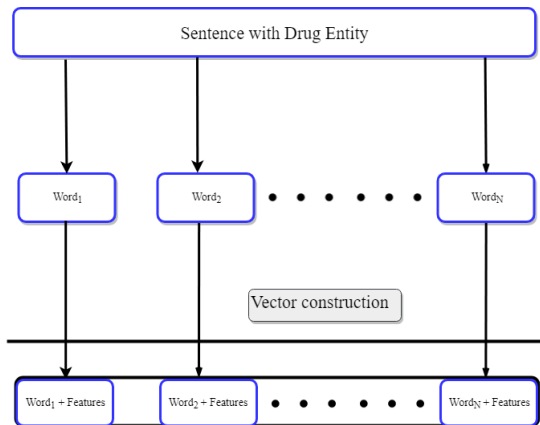
Task9.1 DNR - CRF



Word Entity

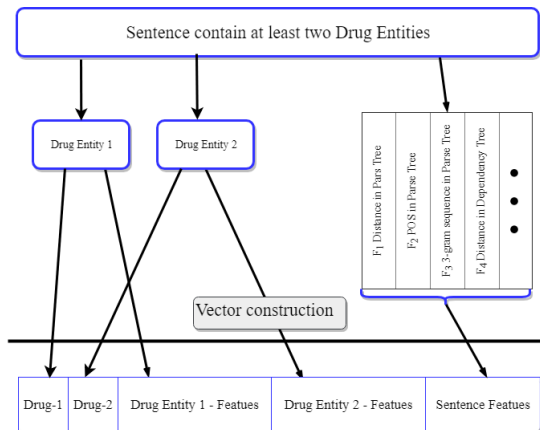
Features

Task9.1 DNR - CRF



Features

Task9.2 DDI - SVM



Experiments & Results

Name	Type	Algo.	Features	# Ftrs	Window	Prec	Rec	F1	M-Prec	M-Rec	M-F1
d24tMDmNERaCRFw1m10	NER	CRF	pos,ort+pos,ort	57	-1:+1	0.9	0.85	0.87	0.92	0.65	0.71
d22btDBmNERaCRFw3m5	NER	CRF	pos,ort,lo+pos,ort	21	-3:+3	0.9	0.82	0.86	0.92	0.63	0.7
d23tDBmNERaLSVMw3m0	NER	SVM	lm,pos,lo,chg,ort + pos,lo,chg,ort	21	-2:+2	0.83	0.79	0.81	0.64	0.54	0.58
SemEval'13 WBI(DrugBank)	NER					0.921	0.914	0.917	0.653	0.659	0.656
SemEval'13 UTurku(MedLine)	NER					0.809	0.521	0.634	0.649	0.528	0.582
ddi006m65(DrugBank)	DDI	SVM	ort+ort	80	-3:+3	0.6513	0.448	0.5308	0.5778	0.3415	0.4293
ddi006m53(DrugBank)	DDI	SVM	lm,ort+lm,ort	96	-3:+3	0.5748	0.4695	0.5168	0.4292	0.3024	0.3548
ddi007m22(MedLine)	DDI	SVM	pos,ort+sent + tw,tri,tl,tp	113	-3:+3	0.3492	0.2316	0.2785	0.0509	0.0253	0.0383
SemEval'13 FBK-irst(DrugBank)	DDI					0.816	0.838	0.827	0.708	0.639	0.672
SemEval'13 FBK-irst(MedLine)	DDI					0.558	0.505	0.53	0.384	0.514	0.44



Proposed improvements:

- Word embedding - use a more suitable ready-made word2vec db or finish clustering.
- feature selection by frequency of appearance in the dataset, or entropy computation
- model selection by cross-validation
- use neural network as a learner



Shengyu Liu, Buzhou Tang, Qingcai Chen, Xiaolong Wang, and Xiaoming Fan.

Feature engineering for drug name recognition in biomedical texts: Feature conjunction and feature selection.

Computational and Mathematical Methods in Medicine, 2015, 2015.



Sun Kim, Haibin Liu, Lana Yeganova, and W. John Wilbur.

Extracting drugdrug interactions from literature using a rich feature-based linear kernel approach.

Journal of Biomedical Informatics, 55:2330, 2015.






Lev Ratinov and Dan Roth.

Design challenges and misconceptions in named entity recognition.

Proceedings of the Thirteenth Conference on Computational Natural Language Learning - CoNLL 09, 2009.



References II

-  Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. Drug name recognition: Approaches and resources. *Information*, 6(4):790810, 2015.
-  The drugbank database is a online database containing information on drugs and drug targets.
<https://www.drugbank.ca/>, 2018 (accessed May 1, 2018).
-  Biomedical natural language processing.
Word2vec database based on pubmed.
<http://bio.nlplab.org/>, 2013 (accessed May 5, 2018).
-  US National Library of Medicine.
Pubmed comprises more than 28 million citations for biomedical literature from medline, life science journals, and online books.
<https://www.ncbi.nlm.nih.gov/pubmed/>.

