RelEx: A system for clinical relation extraction via Convolutional Neural Network

Samantha(Darshini) Mahendran

Bridget T. Mcinnes, Ph.D

Virginia Commonwealth University, Department of Computer Science

VCU College of Engineering

VCU NLP Lab

Outline

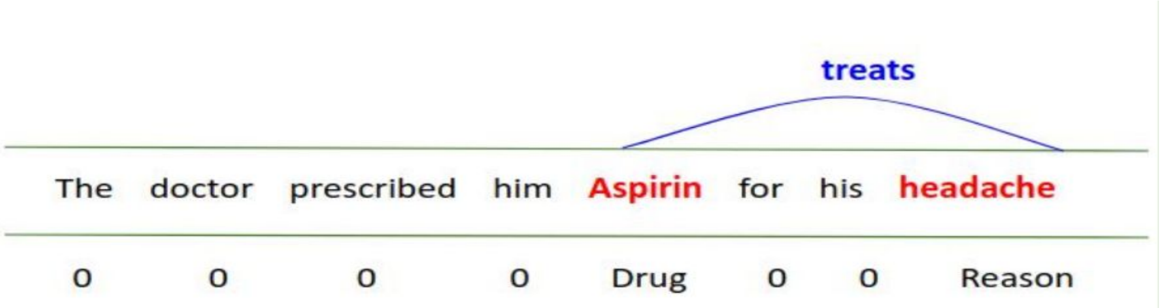
1. Introduction
2. Data
3. Methodology
4. Results and Analysis
5. Conclusion

Introduction

VCU
College of Engineering

What is Relation Extraction?

Task of natural language processing (NLP) to identify and classify the relation between two entities in a text.



She was continued on *midorine* 5mg for a month

Drug ***Dosage*** ***Duration***

Challenge

- Exponential growth of text in recent years
- Manual relation extraction is impossible
- Relation extraction in the clinical domain is more challenging as clinical records can contain multiple pairs of medical entities in the same sentence

VCU
College of Engineering

Data

Data

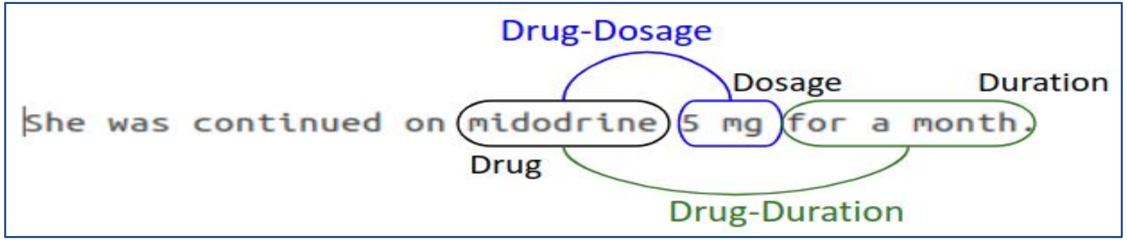
n2c2-2018

1. ***i2b2 (2010)*** dataset includes problem related attributes and relations from patient discharge summaries

2. ***n2c2 (2018)*** dataset contains adverse drug events (ADE), drug related attributes and drug related relations from clinical records

Data: n2c2 (2018)

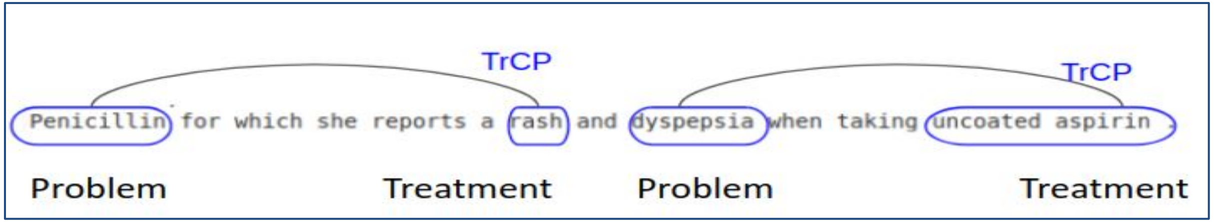
Example:



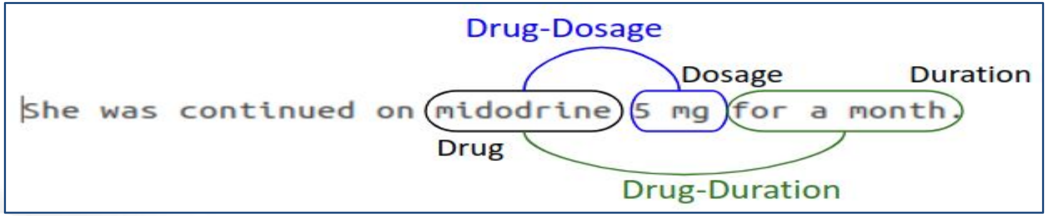
Relation	No of train instances	No of test instances
Drug-Strength	6702	4244
Drug-Duration	643	426
Drug-Route	5538	3546
Drug-Form	6654	4374
Drug-ADE	1107	733
Drug-Dosage	4225	2695
Drug-Reason	5169	3410
Drug-Frequency	6310	4034

Difference from other datasets

- relations in both datasets are fundamentally different
  - i2b2 (2010): multiple relations per entity pair



  - n2c2 (2018): single relation per entity pair



Methodology

College of Engineering

Method

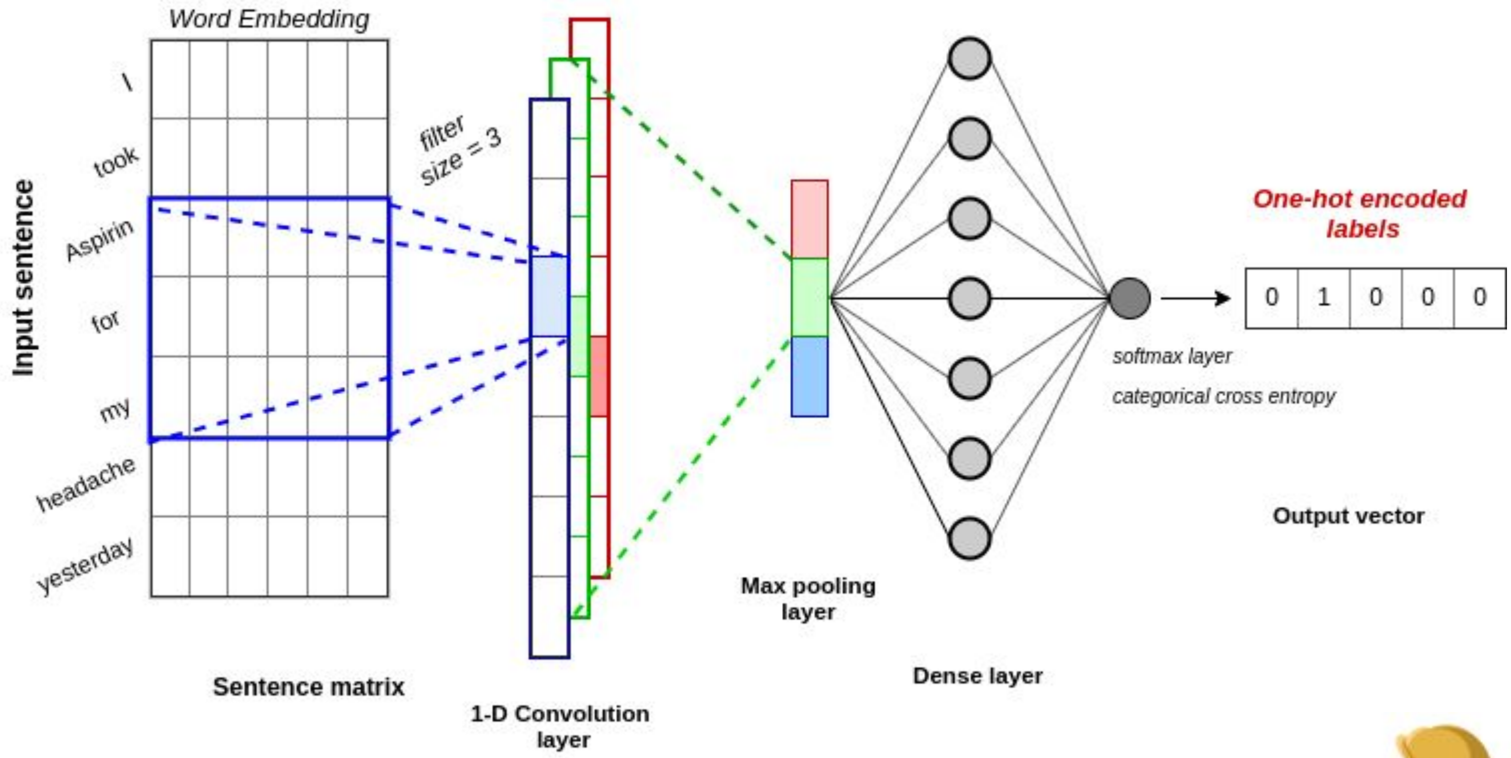
RelEx - Relation extraction system for identifying and classifying relations from clinical text using CNNs

Three approaches:

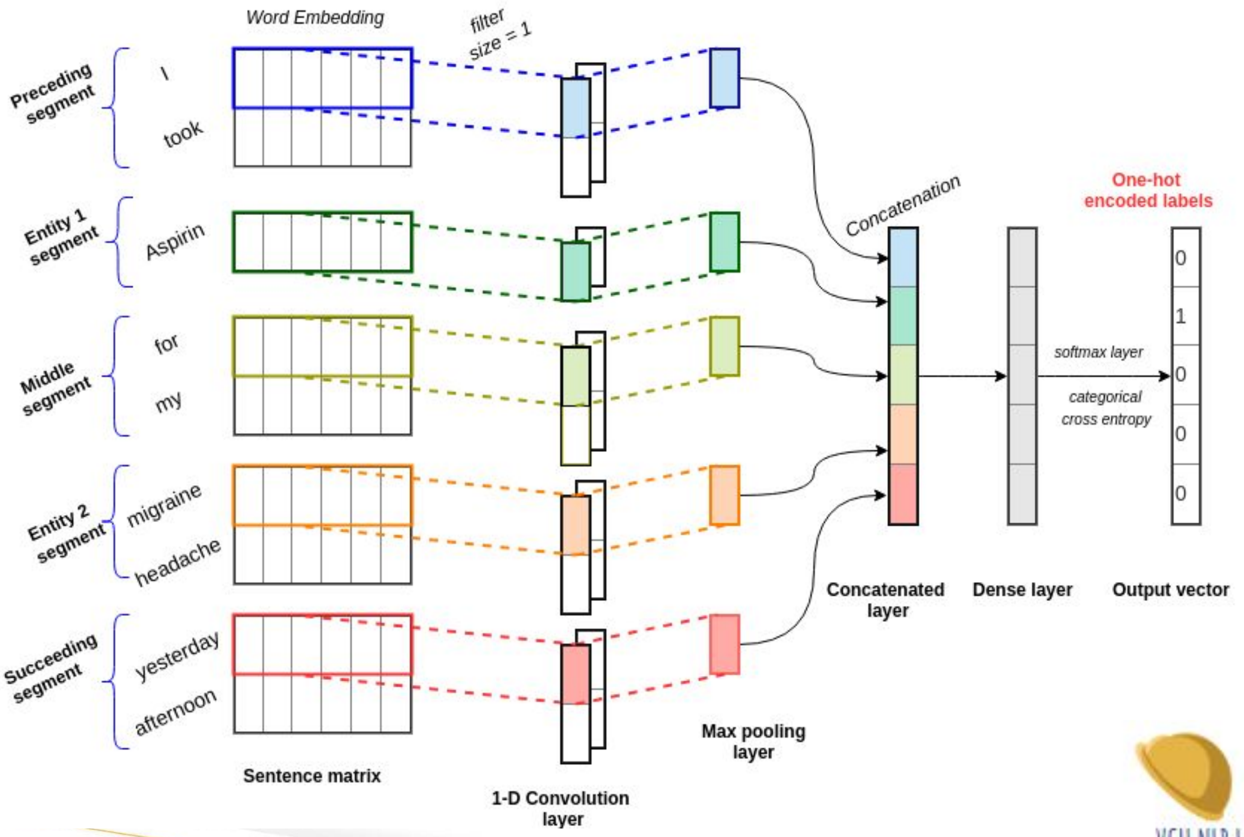
- Single label Sentence-CNN*
- Segment-CNN*
- ***Multi label Sentence-CNN***

** based on Luo et al's paper*

Sentence CNN(Single label)



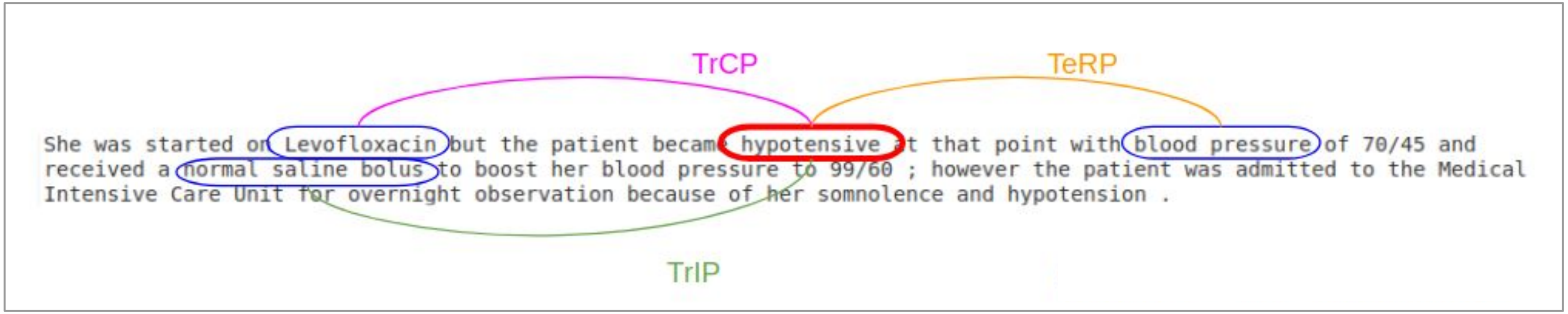
Segment CNN



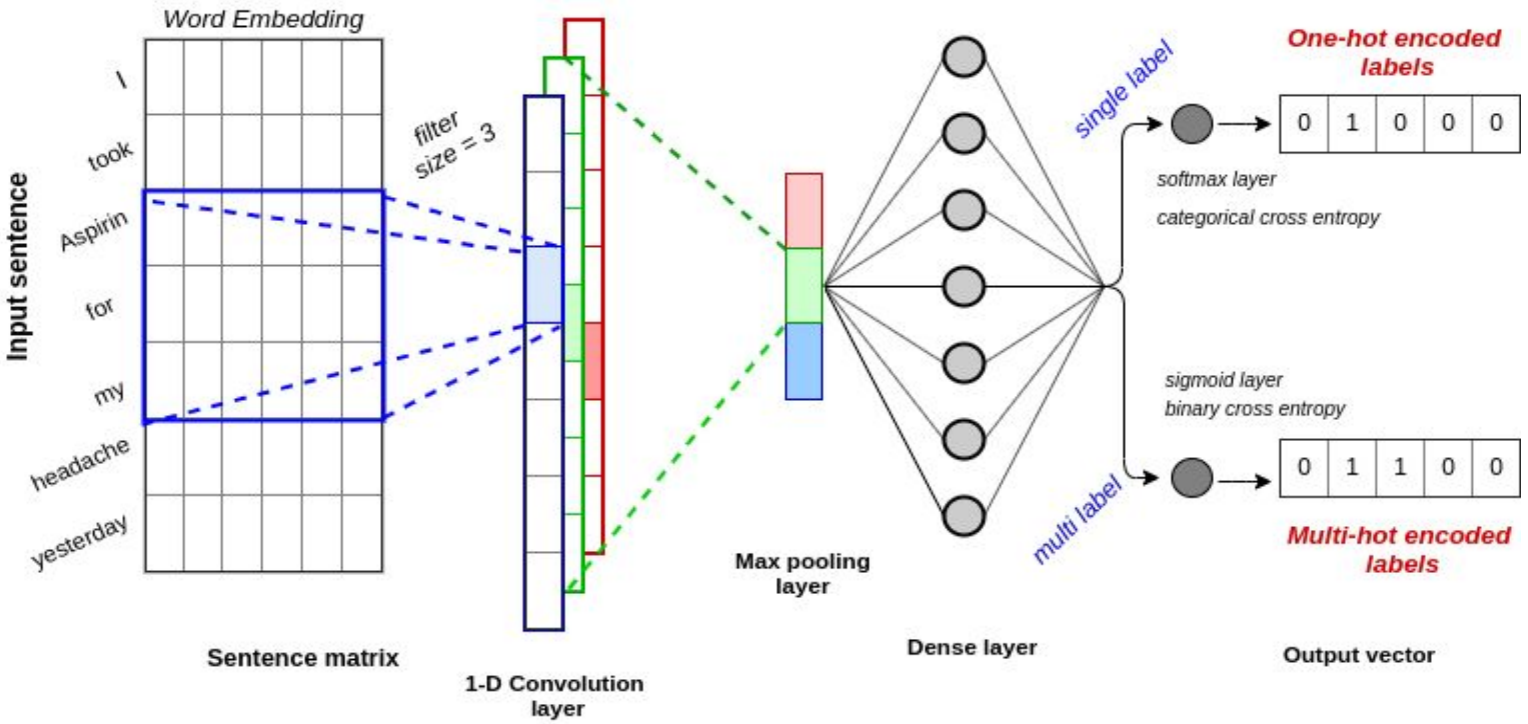
Word Embedding
filter
size = 1
Preceding segment
I
took
Entity 1 segment
Aspirin
Middle segment
for
my
Entity 2 segment
migraine
headache
Succeeding segment
yesterday
afternoon
Concatenation
One-hot encoded labels
0
1
0
0
0
softmax layer
categorical cross entropy
Concatenated layer
Dense layer
Output vector
Max pooling layer
Sentence matrix
1-D Convolution layer

Our focus: Sentence CNN

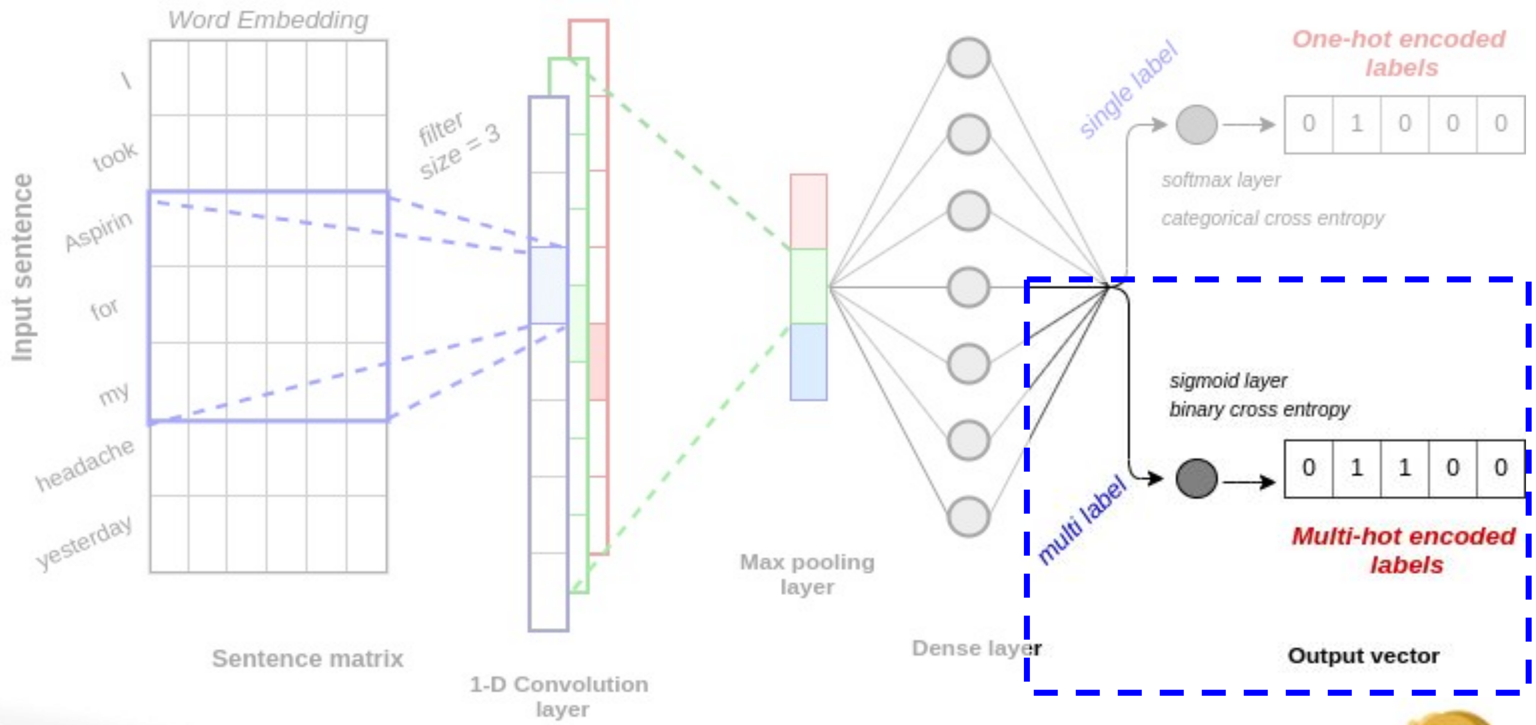
A sentence can contain more than one distinct mentions of relation (pair of entities) with its own context



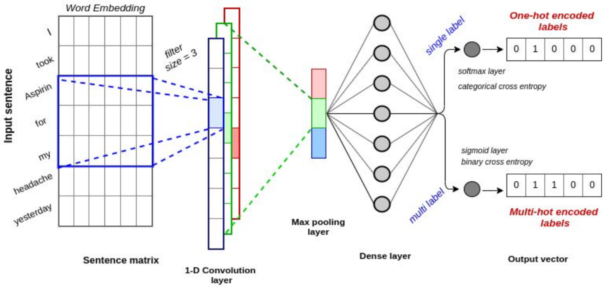
Sentence CNN(Multi label)



Sentence CNN(Multi label)



Sentence CNN (Multi-label)



1. Loss function

   binary cross entropy function is used as the problem is considered as binary classification of each label

2. Choice of output layer - sigmoid activation function

   models the probability of a class as bernoulli distribution and calculates the conditional probabilities of each target class independent from the other class probabilities
   Output falls in the range of 0 to 1

3. Multi-hot-encoding of labels

   threshold of 0.5 which is the inflection point of sigmoid function to determine the class label

Feature Representation

Word2Vec

- Trained over MIMIC - III (Medical Information Mart for Intensive Care)
  - Experimented: 200d, 300d, 400d
- Performed well with *Segment - CNN*

GloVe

- Trained over Wikipedia (2014) and Gigaword 5
  - Experimented: 100d, 200d, 300d
- Performed well with *Sentence - CNN*

Results & Analysis

VCU College of Engineering

i2b2 (2010) dataset - Results

*statistically significant

	Sentence CNN (Single label)			Sentence CNN (Multi label)			Segment CNN		
Relation	***Precision***	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***
Problem - Treatment(TrP) (6)	0.68	0.69	0.69	0.71	0.62	0.66	0.7	0.71	**0.71**
Problem - Test(TeP) (3)	0.68	0.68	0.68	0.75	0.7	0.72*	0.78	0.79	**0.79**
* Problem - Problem(PP) (2)	0.87	0.88	0.87	0.93	0.89	**0.92**	0.87	0.86	0.87
Average	*0.75*	*0.75*	*0.75*	*0.8*	*0.74*	*0.77*	*0.78*	*0.79*	***0.79***

VCU College of Engineering

** binary classification*
() no of classes

i2b2(2010) dataset - Analysis

	Sentence CNN (Single label)			Sentence CNN (Multi label)			Segment CNN		
Problem - Treatment (TrP)	***Precision***	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***
NTrP (1702)	0.78	0.79	0.78	0.64	0.57	0.60	0.76	0.86	**0.81**
TrAp (885)	0.57	0.67	0.61	0.76	0.82	**0.79**	0.59	0.62	0.6
TrCP (184)	0.75	0.23	**0.34**	0.73	021	0.33	0.91	0.14	0.22
TrNAP (62)	0.84	0.24	**0.36**	1.00	0.09	0.17	0.7	0.01	0.17
TrIP (51)	0.4	0.04	0.07	0.5	0.03	0.05	0.2	0.04	**0.07**
TrWP (24)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
System	***0.68***	***0.69***	***0.69***	***0.71***	***0.62***	***0.66***	***0.7***	***0.705***	***0.705***

i2b2(2010) dataset - Analysis

	Sentence CNN (Single label)			Sentence CNN (Multi label)			Segment CNN		
Problem - Treatment (TrP)	***Precision***	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***
NTrP (1702)	0.78	0.79	**0.78**	0.64	0.57	0.60	0.76	0.86	0.81
TrAp (885)	0.57	0.67	0.61	0.76	0.82	**0.79**	0.59	0.62	0.6
TrCP (184)	0.75	0.23	**0.34**	0.73	0.21	0.33	0.91	0.14	0.22
TrNAP (62)	0.84	0.24	**0.36**	1.00	0.09	0.17	0.7	0.01	0.17
TrIP (51)	0.4	0.04	0.07	0.5	0.03	0.05	0.2	0.04	**0.07**
TrWP (24)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
System	***0.68***	***0.69***	***0.69***	***0.71***	***0.62***	***0.66***	***0.7***	***0.71***	***0.71***

i2b2(2010) dataset - Analysis

	Sentence CNN(Single label)			**Sentence CNN (Multi label)**			**Segment CNN**		
Problem - Test (TeP)	***Precision***	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***
NTeP (993)	0.73	0.75	0.73	0.68	0.62	0.65	0.74	0.86	**0.77**
TeRP (993)	0.66	0.71	0.68	0.78	0.84	**0.81**	0.81	0.71	0.75
TeCP (166)	0.51	0.08	0.13	0.81	0.32	**0.46**	0.73	0.16	0.25
System	**0.687**	**0.687**	**0.687**	**0.75**	**0.7**	**0.72**	**0.78**	**0.79**	**0.79**

i2b2(2010) dataset - Analysis

	Sentence CNN (Multi label)										
	Single labels only				Multiple labels only				All labels		
Relation	***# instances***	***Precision***	***Recall***	***F-measure***	***# instances***	***Precision***	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***
Problem - Treatment (TrP)	644	0.65	0.64	0.65	240	0.96	0.59	**0.73**	0.71	0.62	0.66
Problem - Test(TeP)	738	0.61	0.82	0.70	209	0.88	1.00	**0.93**	0.75	0.7	0.72
* Problem - Problem (PP)	1039	0.93	0.68	0.78	469	1.00	0.78	0.88	0.93	0.89	**0.92**
System	*2421*	*0.73*	*0.71*	*0.71*	*918*	*0.95*	*0.79*	***0.85***	*0.8*	*0.74*	*0.77*

* *binary classification*

n2c2(2018) dataset - Results

*statistically significant

	Sentence CNN (Single label)			Sentence CNN (Multi label)			Segment CNN		
Relation	***Precision***	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***
Drug-Strength	0.83	0.32	0.46	0.77	0.61	0.66*	0.96	0.92	**0.94**
Drug-Duration	0.66	0.21	0.32	0.83	0.73	0.78*	0.91	0.86	**0.88**
Drug-Route	0.30	0.57	0.39	0.91	0.9	0.9*	0.95	0.97	**0.96**
Drug-Form	0.47	0.62	0.53	0.91	0.9	0.91*	0.97	0.97	**0.97**
Drug-ADE	0.84	0.04	0.07	0.72	0.61	0.66*	0.79	0.64	**0.69**
Drug-Dosage	0.60	0.21	0.30	0.88	0.83	0.86*	0.91	0.95	**0.93**
Drug-Reason	0.59	0.78	0.67	0.85	0.84	0.84*	0.90	0.94	**0.92**
Drug-Frequency	0.39	0.38	0.38	0.92	0.94	0.93*	0.96	0.96	**0.96**
Average	*0.59*	*0.46*	*0.46*	*0.87*	*0.87*	*0.87**	*0.94*	*0.93*	***0.94***

n2c2(2018) dataset - Results

*statistically significant

	Sentence CNN (Single label)			Sentence CNN (Multi label)			Segment CNN		
Relation	***Precision***	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***
Drug-Strength	0.83	0.32	0.46	0.77	0.61	0.66	0.96	0.92	**0.94**
Drug-Duration	0.66	0.21	0.32	0.83	0.73	0.78	0.91	0.86	**0.88**
Drug-Route	0.30	0.57	0.39	0.91	0.9	0.9	0.95	0.97	**0.96**
Drug-Form	0.47	0.62	0.53	0.91	0.9	0.91	0.97	0.97	**0.97**
Drug-ADE	0.84	0.04	0.07	0.72	0.61	0.66	0.79	0.64	**0.69**
Drug-Dosage	0.60	0.21	0.30	0.88	0.83	0.86	0.91	0.95	**0.93**
Drug-Reason	0.59	0.78	0.67	0.85	0.84	0.84	0.90	0.94	**0.92**
Drug-Frequency	0.39	0.38	0.38	0.92	0.94	0.93	0.96	0.96	**0.96**
Average	*0.59*	*0.46*	*0.46*	*0.87*	*0.87*	*0.87**	*0.94*	*0.93*	***0.94***

NLP LAB

n2c2(2018) dataset - Results

	Single labels only			**Multi label only**			**All labels**		
Relation	***Precision***	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***	***Precision***	***Recall***	***F-measure***
Drug-Strength	0.61	0.91	0.73	0.87	0.79	**0.83**	0.77	0.61	0.66
Drug-Duration	0.67	0.71	0.69	0.91	0.74	**0.82**	0.83	0.73	0.78
Drug-Route	0.54	0.8	0.64	0.96	0.91	**0.94**	0.91	0.9	0.9
Drug-Form	0.77	0.93	0.84	0.95	0.9	**0.92**	0.91	0.9	0.91
Drug-ADE	0.72	0.66	**0.69**	0.76	0.41	0.54	0.72	0.61	0.66
Drug-Dosage	0.7	0.83	0.76	0.93	0.84	**0.88**	0.88	0.83	0.86
Drug-Reason	0.82	0.87	**0.85**	0.89	0.81	**0.85**	0.85	0.84	0.84
Drug-Frequency	0.62	0.86	0.72	0.98	0.94	**0.96**	0.92	0.94	0.93
System	*0.71*	*0.85*	*0.77*	*0.94*	*0.87*	***0.9***	*0.87*	*0.87*	*0.87*

Conclusion & Future work

College of Engineering

Conclusions

	Sentence - CNN (Single label)	Sentence - CNN (Multi-label)	Segment - CNN
Pros	• Good for multi class classification • Not computationally expensive	• Good for multi label classification • Not computationally expensive	• Explicitly distinguish segments • Solves to multi label classification problem
Cons	• Not suitable for multi label classification • Do not consider the positional information of entities	• Do not consider the positional information of entities	• Computationally expensive

Future Work

- Explore additional segment-CNN architectures
  - incorporate CRF layer while concatenating segments
  - incorporate biLSTM
  - incorporate transformer with attention mechanism

- Explore different feature representations :
  - *Feature-based representation*
    - incorporate semantic similarity, relatedness and association
  - *Featureless representation*
    - Character embeddings
    - Combine word and character embeddings
    - Contextual representation (e.g. BERT, ELMO)


Thank you!

hyper parameter tuning

dataset	relation types	Sentence CNN (Single label)	Sentence CNN (Multi label)	Segment CNN
i2b2 - 2010	Pr-Tr	Glove 200d	Glove 300d	MIMIC 200d
	Pr-Te	Glove 200d	Glove 300d	MIMIC 200d
	Pr-Pr	MIMIC 200d	Glove 300d	MIMIC 300d
n2c2 - 2018	All	Glove 200d	Glove 200d	MIMIC 200d

* *binary classification*
() *no of classes*

t-test & p values

dataset	relation types	t-test	p value	Statistically significant
i2b2 - 2010	Pr-Tr	1.57	0.15	no
	Pr-Te	-2.97	0.02	yes
n2c2 - 2018	All	-95.22	1.65 e-13	yes

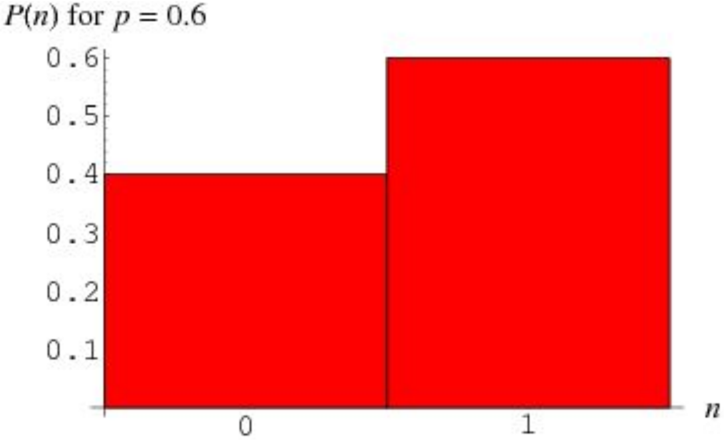
* *binary classification*
() no of classes

Experimental details

- Keras 2.3
- Spacy 2.1.3
- Hyper parameters that are tuned:
  - word embeddings (MIMIC III, GloVe)
  - embedding dimensions(100d, 200d, 300d, 400d)
  - sliding window (2, 3, 5)
  - optimizers (Adam, RMSProp)
  - loss (categorical cross entropy, binary cross entropy)

Bernoulli distribution

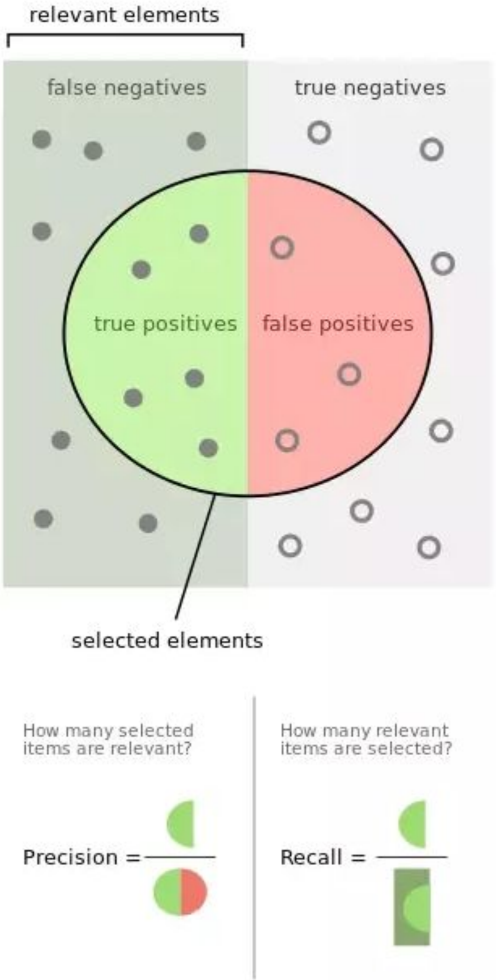
The Bernoulli distribution is a discrete distribution having two possible outcomes labelled by and in which ("success") occurs with probability and ("failure") occurs with probability , where . It therefore has probability density function. (1)



Precision and Recall

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$



Softmax

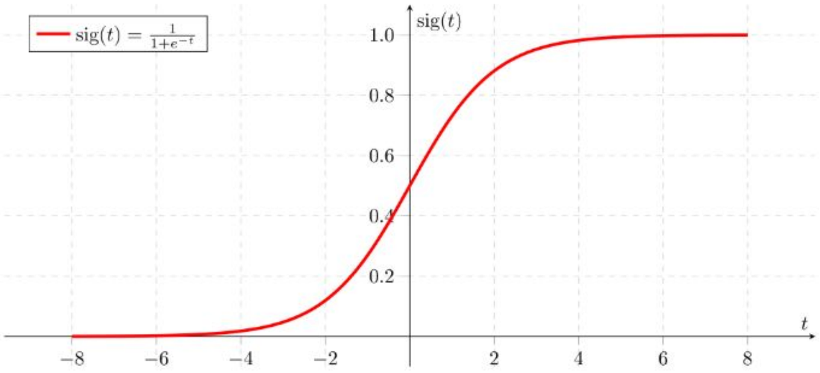
- Softmax calculates the probabilities distribution of the event over ‘n’ different events. (will calculate the probabilities of each target class over all possible target classes).

- Output probabilities range will be 0 to 1, and the sum of all the probabilities will be equal to one.

- If the softmax function used for multi-classification model it returns the probabilities of each class and the target class will have the high probability.

Sigmoid

- Sigmoid function take any range real number and returns the output value which falls in the range of 0 to 1
- When we’re building a classifier for a problem with more than one right answer, we apply a sigmoid function to each element of the raw output independently
- Unlike softmax which gives a probability distribution around n classes, sigmoid functions allow for independent probabilities.



Binary Cross-Entropy Loss

It is a Sigmoid activation plus a Cross-Entropy loss.

Unlike Softmax loss it is independent for each vector component (class), i.e. the loss computed for every CNN output vector component is not affected by other component values.

That’s why it is used for multi-label classification

$$CE = -t_1 log(f(s_1)) - (1 - t_1) log(1 - f(s_1))$$

$$f(s_i) = \frac{1}{1 + e^{-s_i}}$$

Categorical Cross-Entropy Loss

- It is a Softmax activation plus a Cross-Entropy loss.
- If we use this loss, we will train a CNN to output a probability over the n classes for each image.
- It is used for multi-class classification.