

Extracting Adverse Drug Events from Clinical Notes

Darshini (Samantha) Mahendran

Bridget T. Mcinnes, Ph.D

Virginia Commonwealth University, Department of Computer Science

Outline

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



Introduction



What is Adverse drug events (ADE)?

- ADEs are unintended incidents that involve the taking of a medication (unwanted effect caused by the administration of a drug)
- Includes overdoses, allergic reactions, drug interactions, and medication errors
- Often Lead to hospitalization, and account for an estimated 12% of all emergency room visits
- Conditions caused by undiscovered ADEs, increase costs and risks further and impact patient economically and mentally

The challenge

- Quickly identifying ADEs in large, can increase both safety and quality of patient health care
- Require information about not just the drug itself, but attributes describing the drug (e.g. strength, dosage) and why the drug was initially being taken (e.g. reason)
- Processing information manually from scientific publications and clinical narratives is challenging



Data

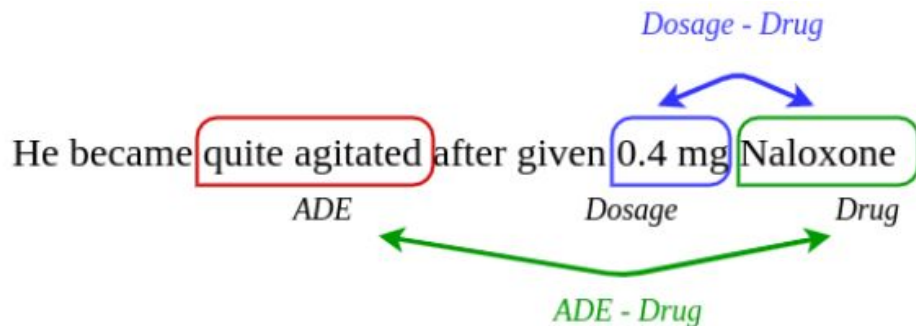


VCU

College of Engineering

Data: n2c2 (2018)

Includes adverse drug events (ADE), drug related attributes and drug related relations from clinical records

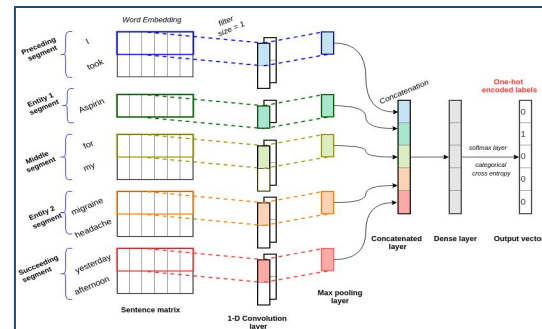
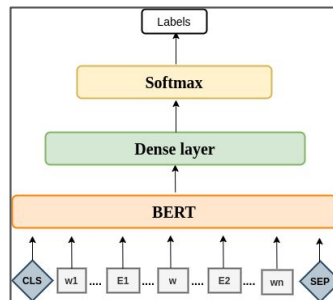
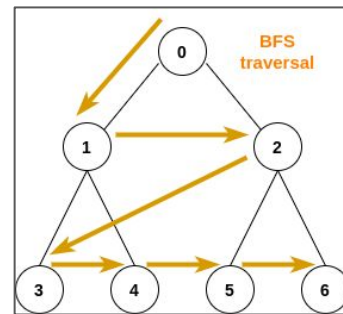


Relation	# train	# test
Drug-Strength	6702	4244
Drug-Duration	643	426
Drug-Route	5538	3546
Drug-Form	6654	4374
Drug-ADE	1107	733
Drug-Dosage	4225	2695
Drug-Reason	5169	3410
Drug-Frequency	6310	4034

Methodology

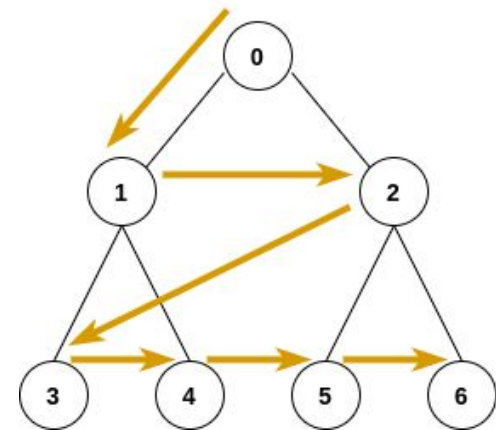
Method

- **RelEx** - a **Relation Extraction** Framework based on Python for RE
- Utilize three approaches for clinical RE:
 - Rule-based approach
 - Left-only traversal
 - Left-Right (bounded & unbounded)
 - Deep learning-based approach
 - Sentence CNN
 - Segment CNN
 - BERT-based approach
 - BERT - cased/uncased
 - Bio-BERT
 - Clinical -BERT



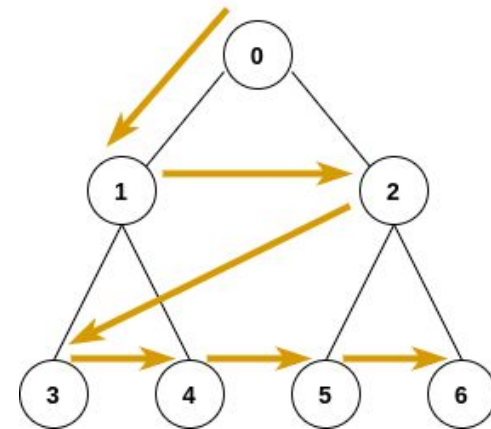
Rule-based approach

- Utilizes co-location information to determine whether a relation exists between two entities
- Graph-based algorithm is used for traversal
- Different traversal techniques are applied and best traversal technique for each relation type is determined
 - traverse left side only
 - traverse right side only
 - traverse left first then right
 - traverse right first then left



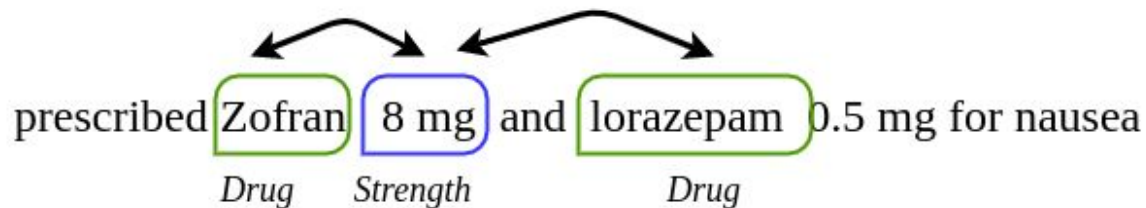
Rule-based approach

- Utilizes co-location information to determine whether a relation exists between two entities
- Graph-based algorithm is used for traversal
- Different traversal techniques are applied and best traversal technique for each relation type is determined
 - traverse left side only
 - traverse right side only
 - traverse left first then right
 - traverse right first then left



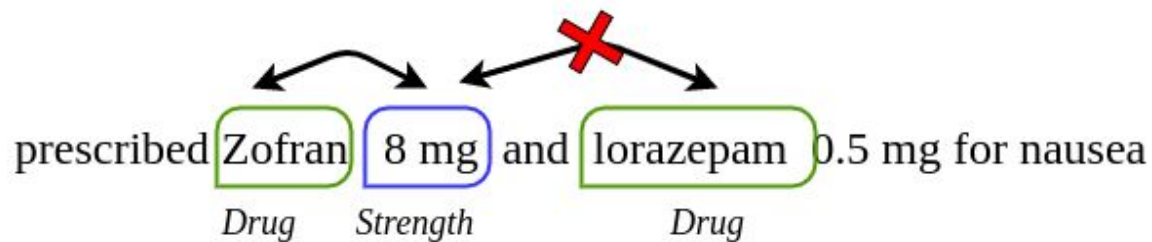
Rule-based approach

- Different traversal techniques are applied and best traversal technique for each relation type is determined
 - traverse left side only
 - traverse right side only
 - traverse left first then right
 - traverse right first then left



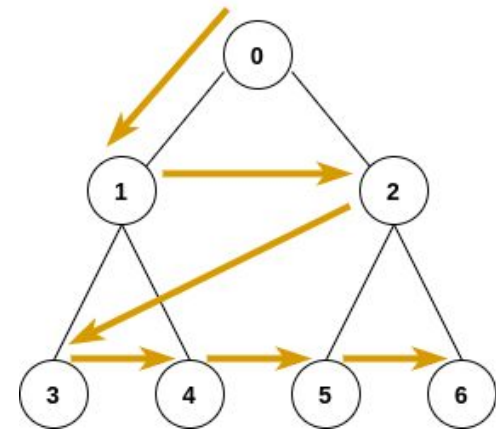
Rule-based approach

- Different traversal techniques are applied and best traversal technique for each relation type is determined
 - traverse left side only
 - traverse right side only
 - traverse left first then right
 - traverse right first then left



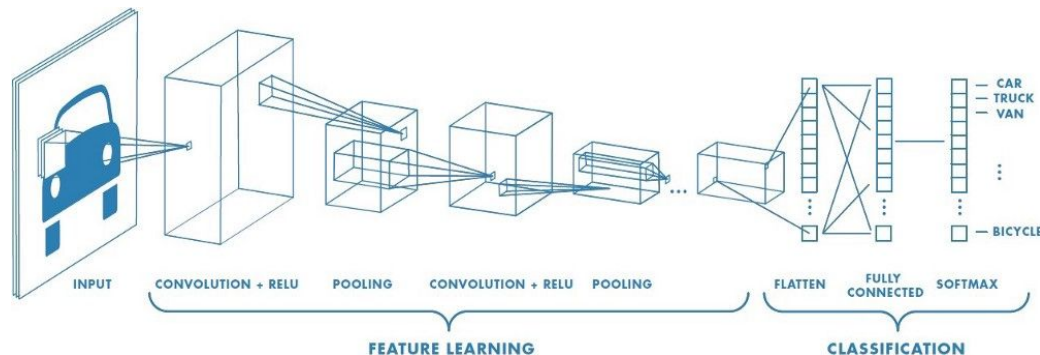
Rule-based approach

- Different traversal techniques are applied and best traversal technique for each relation type is determined
 - traverse left side only
 - traverse left first then right
- Conduct traversals in two modes:
 - bounded - limiting traversal to only a single relation per relation class
 - unbounded - allows a entity to be linked to multiple other entity classes
with same relation

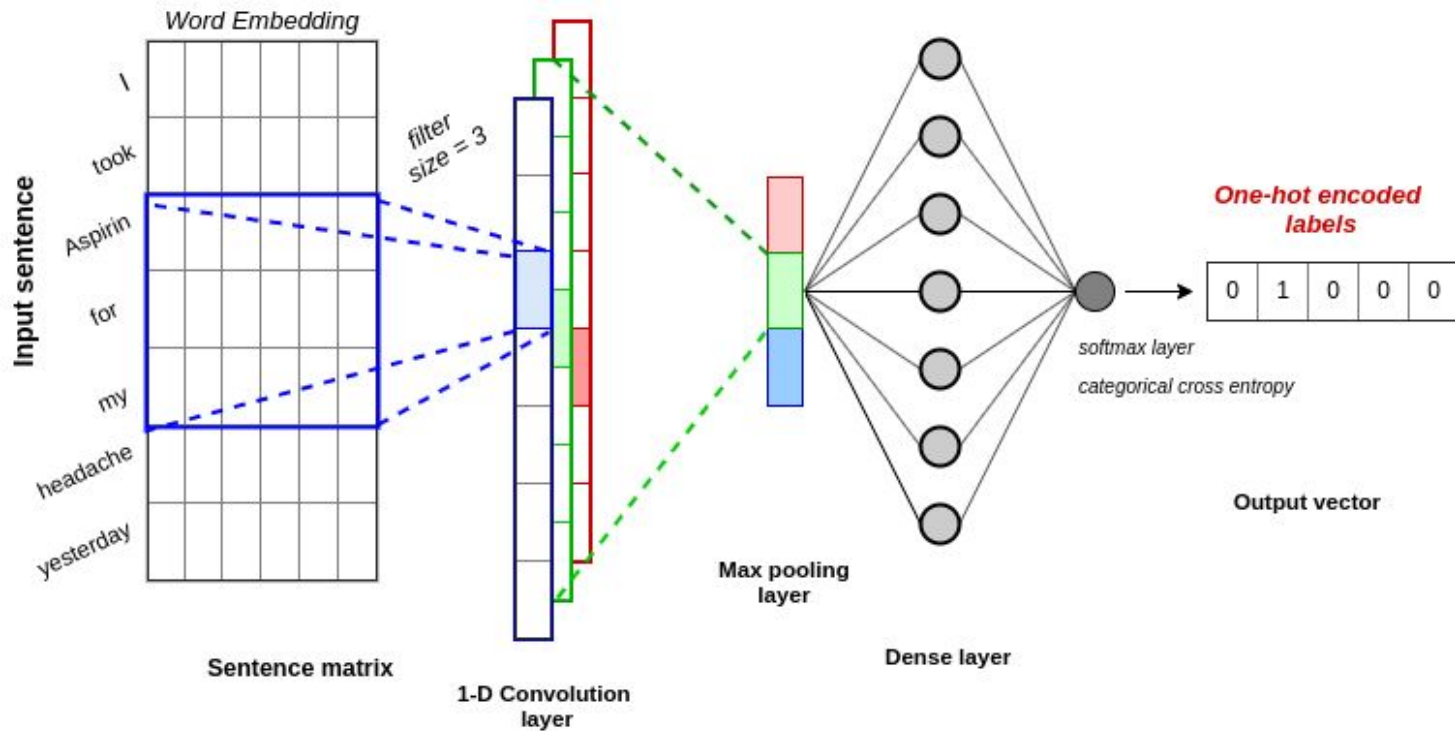


Deep learning-based approach

- Use Convolutional Neural Networks (CNN) in our approaches
- CNN - class of deep neural networks (NN), works well with data that consists of hidden patterns or complex relations among entities
- Our deep learning-based approach includes two CNN architectures:
 - Sentence-CNN
 - Segment-CNN

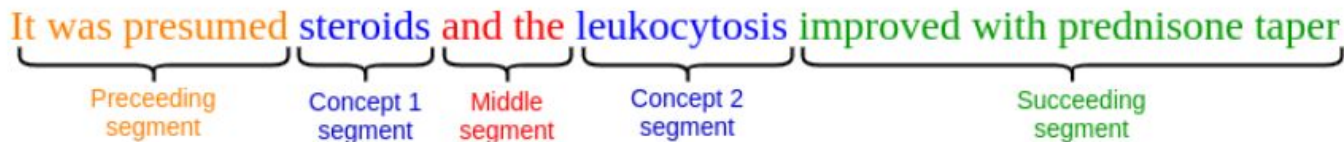


Sentence CNN

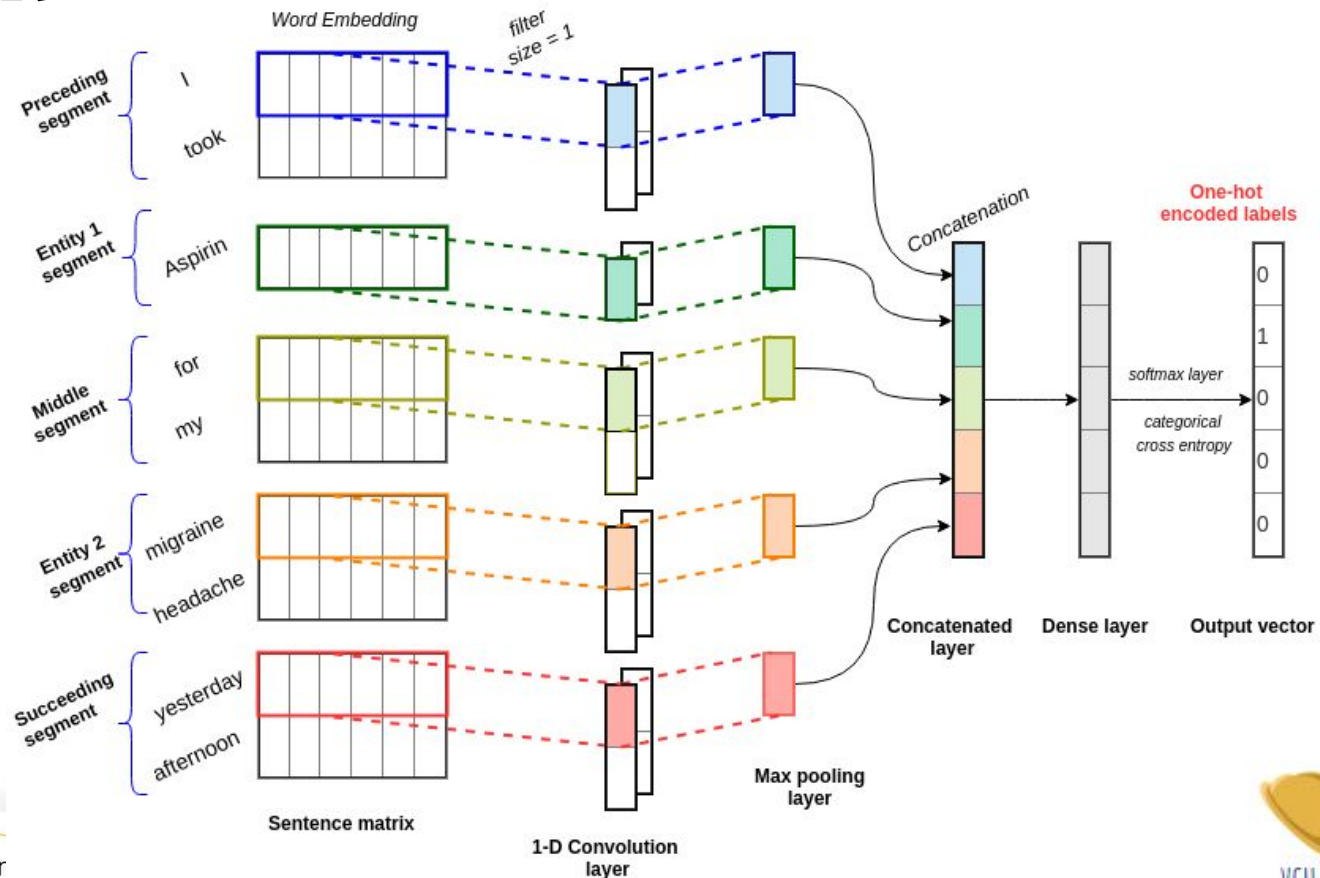


Segment CNN

- Different segments play different role in determining the relation class
- Divide the sentence 5 into segments based on the position of the entities in the sentence
 - preceding - tokenized words before the first entity
 - entity 1 - tokenized words in the first entity
 - middle - tokenized words between the two entities
 - entity 2 - tokenized words in the second entity
 - succeeding - tokenized words after the second entity



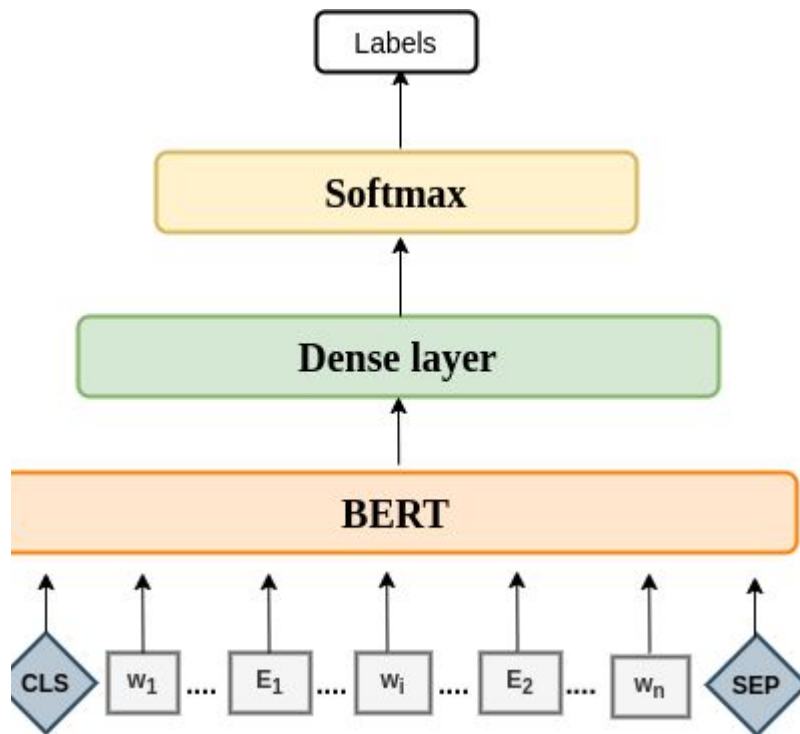
Segment CNN



Bidirectional Encoder Representations from Transformers (BERT)

- Introduced by Google in 2018
- BERT embeddings - Context-based representation of a token is generated based on the surrounding words in the text.
- Pre-trained BERT models we used:
 - BERT (uncased) } *trained on English data: BookCorpus (800M words) &*
 - BERT (cased) } *Wikipedia (2500M words)*
 - BioBERT - *general BERT, trained over research articles from PubMed abstracts*
 - Clinical BERT - *BioBERT, further fine-tuned over MIMIC-III*

BERT-based approach



Evaluation criteria

- Precision (P) - Ratio between correctly predicted mentions over total set of predicted mentions for a specific entity
- Recall (R) - Ratio of correctly predicted mentions over actual number of mentions
- F-1 score (F) - Harmonic mean between precision and recall
- System performance is reported by,
 - Micro average - calculates metrics globally by counting total true positives, false negatives, and false positives

Results & Analysis

Results: Rule-based

	Left-only			Left-Right (unbounded)			Left-Right (bounded)		
	P	R	F	P	R	F	P	R	F
Strength-Drug	0.96	0.95	0.95	0.46	0.90	0.61	0.94	0.94	0.94
Duration-Drug	0.78	0.69	0.73	0.58	0.74	0.65	0.46	0.41	0.43
Route-Drug	0.90	0.89	0.89	0.45	0.64	0.53	0.37	0.36	0.37
Form-Drug	0.98	0.98	0.98	0.62	0.63	0.63	0.67	0.66	0.67
ADE-Drug	0.46	0.39	0.43	0.55	0.75	0.64	0.60	0.51	0.55
Dosage-Drug	0.89	0.89	0.89	0.61	0.57	0.59	0.89	0.88	0.89
Reason-Drug	0.48	0.35	0.41	0.61	0.57	0.59	0.39	0.28	0.33
Frequency-Drug	0.98	0.98	0.98	0.39	0.62	0.48	0.10	0.10	0.10
System (Micro)	0.88	0.83	0.86	0.50	0.67	0.57	0.56	0.53	0.55
System (Macro)	0.85	0.80	0.83	0.61	0.70	0.63	0.58	0.53	0.55

Results: Deep Learning-based

	Segment-CNN			Sentence-CNN		
	P	R	F	P	R	F
Strength-Drug	0.91	0.88	0.90	0.90	0.91	0.90
Duration-Drug	0.39	0.90	0.55	0.41	0.90	0.57
Route-Drug	0.77	0.89	0.83	0.76	0.91	0.83
Form-Drug	0.85	0.95	0.90	0.85	0.96	0.90
ADE-Drug	0.32	0.85	0.46	0.32	0.85	0.46
Dosage-Drug	0.83	0.92	0.87	0.82	0.93	0.87
Reason-Drug	0.27	0.88	0.42	0.27	0.88	0.41
Frequency-Drug	0.56	0.88	0.69	0.56	0.88	0.69
System (Micro)	0.69	0.90	0.78	0.68	0.92	0.78
System (Macro)	0.68	0.90	0.77	0.67	0.91	0.77

Results: BERT-based

	BERT (uncased)			BERT (cased)			BioBERT			Clinical BERT		
	P	R	F	P	R	F	P	R	F	P	R	F
Strength-Drug	0.86	0.88	0.87	0.86	0.99	0.92	0.86	0.90	0.88	0.87	0.82	0.84
Duration-Drug	0.95	0.93	0.94	0.96	0.93	0.94	0.96	0.93	0.95	0.96	0.92	0.94
Route-Drug	0.92	0.99	0.95	0.92	0.97	0.97	0.92	0.97	0.94	0.92	0.95	0.93
Form-Drug	0.96	0.97	0.97	0.96	0.95	0.96	0.96	0.97	0.96	0.96	0.97	0.97
ADE-Drug	0.95	0.99	0.97	0.95	0.99	0.97	0.95	0.99	0.97	0.95	0.99	0.97
Dosage-Drug	0.93	0.96	0.94	0.93	0.96	0.95	0.93	0.96	0.94	0.93	0.89	0.91
Reason-Drug	0.96	0.98	0.97	0.96	0.98	0.97	0.96	0.99	0.97	0.96	0.99	0.97
Frequency-Drug	0.93	0.96	0.94	0.93	0.92	0.93	0.93	0.95	0.94	0.93	0.95	0.94
System (Micro)	0.93	0.96	0.94	0.93	0.96	0.94	0.93	0.95	0.94	0.93	0.96	0.94
System (Macro)	0.92	0.95	0.93	0.92	0.96	0.93	0.92	0.95	0.93	0.92	0.95	0.93

Results: Comparison across our approaches

	Train	Test	Rule-based			Segment-CNN			BioBERT		
	#	#	P	R	F	P	R	F	P	R	F
Strength-Drug	6702	4244	0.96	0.95	0.95	0.91	0.88	0.90	0.86	0.90	0.88
Duration-Drug	643	426	0.78	0.69	0.73	0.39	0.90	0.55	0.96	0.93	0.95
Route-Drug	5538	3546	0.90	0.89	0.89	0.77	0.89	0.83	0.92	0.97	0.94
Form-Drug	6654	4373	0.98	0.98	0.98	0.85	0.95	0.90	0.96	0.97	0.96
ADE-Drug	1107	733	0.46	0.39	0.43	0.32	0.85	0.46	0.95	0.99	0.97
Dosage-Drug	4255	2695	0.89	0.89	0.89	0.83	0.92	0.87	0.93	0.96	0.94
Reason-Drug	5169	3410	0.48	0.35	0.41	0.27	0.88	0.42	0.96	0.99	0.97
Frequency-Drug	6310	4034	0.98	0.98	0.98	0.56	0.88	0.69	0.93	0.95	0.94
System (Micro)			0.88	0.83	0.86	0.69	0.90	0.78	0.93	0.95	0.94
System (Macro)			0.85	0.80	0.83	0.68	0.90	0.77	0.92	0.95	0.93

Results: Comparison with state-of-art

	Our models				Wei, et al.				Alimova, et al.		
	Cased	Uncased	Bio	Clinical	Cased	Uncased	Bio	Clinical	Uncased	Bio	Clinical
Strength-Drug	0.87	0.87	0.88	0.84	0.98	0.99	0.98	0.99	0.58	0.68	0.68
Duration-Drug	0.94	0.94	0.94	0.94	0.88	0.89	0.88	0.89	0.41	0.66	0.65
Route-Drug	0.95	0.95	0.94	0.93	0.97	0.97	0.97	0.97	0.63	0.74	0.74
Form-Drug	0.97	0.97	0.96	0.97	0.97	0.98	0.98	0.98	0.62	0.81	0.81
ADE-Drug	0.97	0.97	0.97	0.97	0.80	0.80	0.81	0.81	0.10	0.62	0.62
Dosage-Drug	0.94	0.94	0.94	0.91	0.97	0.97	0.97	0.97	0.67	0.82	0.82
Reason-Drug	0.97	0.97	0.97	0.97	0.76	0.76	0.76	0.77	0.22	0.73	0.73
Frequency-Drug	0.94	0.94	0.94	0.94	0.96	0.96	0.96	0.96	0.53	0.79	0.78

Results: Comparison with state-of-art

	Our models				Wei, et al.				Alimova, et al.		
	Cased	Uncased	Bio	Clinical	Cased	Uncased	Bio	Clinical	Uncased	Bio	Clinical
Strength-Drug	0.87	0.87	0.88	0.84	0.98	0.99	0.98	0.99	0.58	0.68	0.68
Duration-Drug	0.94	0.94	0.94	0.94	0.88	0.89	0.88	0.89	0.41	0.66	0.65
Route-Drug	0.95	0.95	0.94	0.93	0.97	0.97	0.97	0.97	0.63	0.74	0.74
Form-Drug	0.97	0.97	0.96	0.97	0.97	0.98	0.98	0.98	0.62	0.81	0.81
ADE-Drug	0.97	0.97	0.97	0.97	0.80	0.80	0.81	0.81	0.10	0.62	0.62
Dosage-Drug	0.94	0.94	0.94	0.91	0.97	0.97	0.97	0.97	0.67	0.82	0.82
Reason-Drug	0.97	0.97	0.97	0.97	0.76	0.76	0.76	0.77	0.22	0.73	0.73
Frequency-Drug	0.94	0.94	0.94	0.94	0.96	0.96	0.96	0.96	0.53	0.79	0.78



Results: Comparison with state-of-art

	Our models				Wei, et al.				Alimova, et al.		
	Cased	Uncased	Bio	Clinical	Cased	Uncased	Bio	Clinical	Uncased	Bio	Clinical
Strength-Drug	0.87	0.87	0.88	0.84	0.98	0.99	0.98	0.99	0.58	0.68	0.68
Duration-Drug	0.94	0.94	0.94	0.94	0.88	0.89	0.88	0.89	0.41	0.66	0.65
Route-Drug	0.95	0.95	0.94	0.93	0.97	0.97	0.97	0.97	0.63	0.74	0.74
Form-Drug	0.97	0.97	0.96	0.97	0.97	0.98	0.98	0.98	0.62	0.81	0.81
ADE-Drug	0.97	0.97	0.97	0.97	0.80	0.80	0.81	0.81	0.10	0.62	0.62
Dosage-Drug	0.94	0.94	0.94	0.91	0.97	0.97	0.97	0.97	0.67	0.82	0.82
Reason-Drug	0.97	0.97	0.97	0.97	0.76	0.76	0.76	0.77	0.22	0.73	0.73
Frequency-Drug	0.94	0.94	0.94	0.94	0.96	0.96	0.96	0.96	0.53	0.79	0.78

Conclusions

1. Explored a rule-based, deep learning-based, and contextualized language model-based approaches for ADE extraction.
2. BERT-based approach outperformed other models overall and obtained state-of-the-art performance
3. However, co-location information is sufficient to identify many relations -
 - a. Rule-based approach obtained a higher Precision and Recall for certain relations, for e.g. Strength-Drug, Form-Drug, Frequency-Drug (order of entities play a vital role)



Thank you!





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

- Ambiguity between the terms ADE and Reason reduces the overall performance
- Performance is low for Drug-ADE (mostly) and Drug - Reason is the ambiguity between the terms
- Most ADE relations are categorized as Drug - Reason relations
- Experiment - Convert ADE and Reason labels to a common term (Symptom) to increase the overall performance

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

Overall Conclusions

- Rule-based approach is applicable for relations with consistent positional information
- Deep learning-based approaches are applicable for labeled data with many training instances
- BERT-based approaches utilize contextualized word embeddings and they perform better than approaches that use non-contextualized word embeddings

t-test & p values

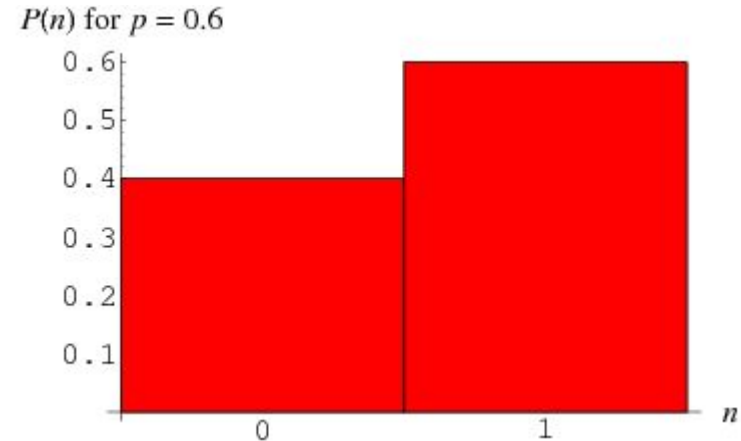
dataset	relation types	t-test	p value	Statistically significant
i2b2 - 2010	Pr-Tr	1.57	0.15	no
	Pr-Te	-2.97	0.02	yes
n2c2 - 2018	All	-95.22	1.65 e-13	yes

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

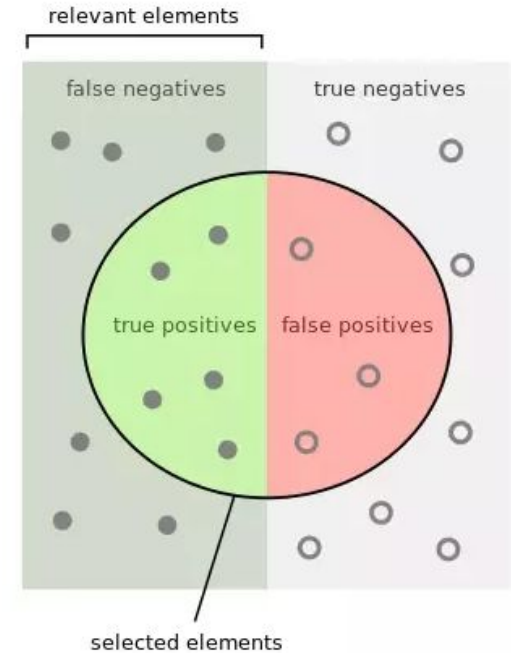
The Bernoulli distribution is a discrete distribution having two possible outcomes labelled by and in which ("success") occurs with probability and ("failure") occurs with probability , where . It therefore has probability density function. (1)



Precision and Recall

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

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



How many selected items are relevant?

$$\text{Precision} = \frac{\text{Green Circle}}{\text{Green Circle} + \text{Red Circle}}$$

How many relevant items are selected?

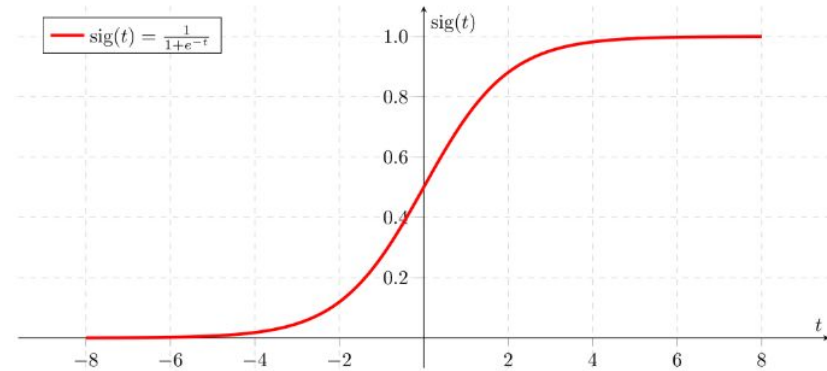
$$\text{Recall} = \frac{\text{Green Circle}}{\text{Green Circle} + \text{Green Rectangle}}$$

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.

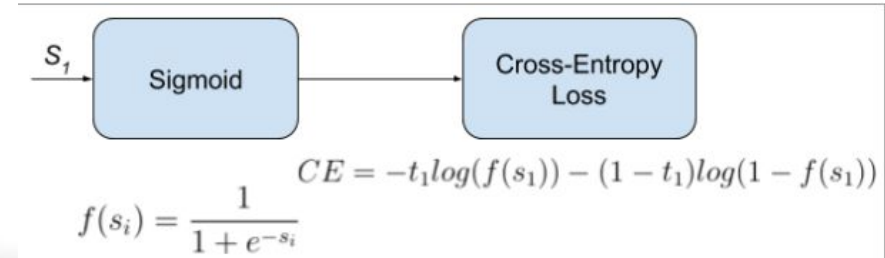


Binary Cross-Entropy Loss

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



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.

