Information Extraction

University Policy Change

- Due to covid:
    - Required to use your midterm exam grade as your final grade


APRIL
FOOLS

Hit record

Information Extraction

extract structured information from unstructured text

Information Extraction

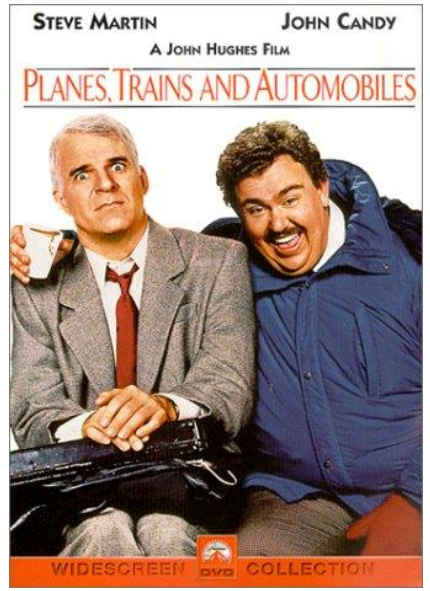
extract structured information from unstructured text

- Named Entity Recognition
- Relationship Extraction

Named Entity Recognition

- traditionally anything that can be referred to with a proper name

• People (PER):	Individuals, fiction characters, …
• Organization (ORG):	Companies, Agencies, …
• Location (LOC):	Physical extents, mountain ranges, seas, ….
• Geo-political entity (GPE):	Countries, states, …
• Facility (FAC):	Bridges, airports, buildings
• Vehicles (VEH):	Planes, trains and automobiles



Citing high fuel prices, [ORG United Airlines] said [TIME Friday] it has increased fares by [MONEY $6] per round trip on flights to some cities also served by lower cost carriers. [ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PER Tim Wagner] said, a unit of [ORG UAL Corp.], said the increase took effect [TIME Thursday] and applies to most routes where it competes against discount carriers, such as [LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco].

Text contains

- 13 mentions
  - 5 organizations (ORG)
  - 4 locations (LOC)
  - 2 times (TIME)
  - 1 person (PER)
  - 1 money (MONEY)

Citing high fuel prices, [ORG United Airlines] said [TIME Friday] it has increased fares by [MONEY $6] per round trip on flights to some cities also served by lower cost carriers. [ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PER Tim Wagner] said, a unit of [ORG UAL Corp.], said the increase took effect [TIME Thursday] and applies to most routes where it competes against discount carriers, such as [LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco].

Traditionally NER

- focuses on the: person, location and organization

- specialized applications:
  - weapons
  - nano-particle characteristics
  - works of art
  - proteins/genes
  - drugs/diseases

Extend NER

- The notion of NER is commonly extended to include things that are not entities per se
  - Temporal expressions
    - dates, times, named events
  - Numerical expressions
    - Measurements, prices, counts

Difficulty with NER: Ambiguity

- Two types of ambiguity
  - the mention can refer to different entities of the same type
    - JFK could refer to the former president or his son
  - the mention can refer to more than one entity
    - JFK could be a person (PER) or an airport (LOC)

NER as sequence labeling

- Standard approach to NER:
  - word-by-word labeling task
  - classifiers are trained to label the tokens in a text with tags that indicate the presence of a particular kind of name entity

[ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PER Tim K Wagner] said.

Words	Label
American	Borg
Airlines	Eorg
,	O
a	O
unit	O
of	O
AMR	Borg
Corp.	Eorg
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	Bper
K	Iper
Wagner	Eper
said	O
.	O

NER as sequence labeling

- Standard approach to NER:
  - word-by-word labeling task
  - classifiers are trained to label the tokens in a text with tags that indicate the presence of a particular kind of name entity

[ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PER Tim K Wagner] said.

Words	Label
American	org
Airlines	org
,	O
a	O
unit	O
of	O
AMR	org
Corp.	org
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	per
K	per
Wagner	per
said	O
.	O

Algorithms

- Conditional Random Fields
- Recurrent Neural Networks
  - bi-Long Short Term Memory (bi-LSTM)
    - often with a CRF output layer

Conditional Random Fields

- Sequence modeling algorithm
- CRF take context into account when predicting the label of a token
  - linear chain CRF

$$P(\boldsymbol{Y}|\boldsymbol{X}) = P(\boldsymbol{l}|\boldsymbol{s})$$

where

Y *= the sequence of labels =* ***l***

X *= the words in the sentence =* ***s***

$$P(\boldsymbol{Y}|\boldsymbol{X}) = P(\boldsymbol{l}|\boldsymbol{s})$$

where

Y *= the sequence of labels =* ***l***

X *= the words in the sentence =* ***s***

Dose
DoseUnit
Frequency
Route of administration (RA)
Drug
Reason
Adverse Reaction
Form

The doctor prescribed him 325 mg Aspirin p.o. 4x/day for 2 weeks

The	**doctor**	**prescribed**	**him**	**325**	**mg**	**aspirin**	**p.o.**	**4x/day**	**for**	**2**	**weeks**
O	O	O	O	Dose	DoseUnit	Drug	RA	Freq	Freq	Freq	Freq

$$P(\boldsymbol{Y}|\boldsymbol{X}) = P(\boldsymbol{l}|\boldsymbol{s})$$

where

Y *= the sequence of labels =* ***l***

X *= the words in the sentence =* ***s***

Dose
DoseUnit
Frequency
Route of administration (RA)
Drug
Reason
Adverse Reaction
Form

The doctor prescribed him 325 mg Aspirin p.o. 4x/day for 2 weeks

The	**doctor**	**prescribed**	**him**	**325**	**mg**	**aspirin**	**p.o.**	**4x/day**	**for**	**2**	**weeks**
O	O	O	O	Dose	DoseUnit	Drug	RA	Freq	Freq	Freq	Freq

$$score(l|s) = \sum_{j=}^{m} \sum_{i=1}^{n} (\lambda_j f_j(s, i, l_i, l_{i-1})$$

First sum: runs over each feature function *j*

Second sum: runs over each position *i* in the sentence

$$P(\boldsymbol{Y}|\boldsymbol{X}) = P(\boldsymbol{l}|\boldsymbol{s})$$

where

Y *= the sequence of labels =* ***l***

X *= the words in the sentence =* ***s***

Dose
DoseUnit
Frequency
Route of administration (RA)
Drug
Reason
Adverse Reaction
Form

The doctor prescribed him 325 mg Aspirin p.o. 4x/day for 2 weeks

The	**doctor**	**prescribed**	**him**	**325**	**mg**	**aspirin**	**p.o.**	**4x/day**	**for**	**2**	**weeks**
O	O	O	O	Dose	DoseUnit	Drug	RA	Freq	Freq	Freq	Freq

$$p(l|s) = \frac{exp[score(l|s)]}{\sum_{l'} exp[score(l'|s)]} = \frac{exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i,l_{i-1})]}{\sum_{l'} exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l'_i,l'_{i-1})]}$$

Transform the score to probabilities by exponentiating and normalizing the scores

$$P(\boldsymbol{Y}|\boldsymbol{X}) = P(\boldsymbol{l}|\boldsymbol{s})$$

where

Y *= the sequence of labels =* ***l***

X *= the words in the sentence =* ***s***

Dose
DoseUnit
Frequency
Route of administration (RA)
Drug
Reason
Adverse Reaction
Form

The doctor prescribed him 325 mg Aspirin p.o. 4x/day for 2 weeks

The	**doctor**	**prescribed**	**him**	**325**	**mg**	**aspirin**	**p.o.**	**4x/day**	**for**	**2**	**weeks**
O	O	O	O	Dose	DoseUnit	Drug	RA	Freq	Freq	Freq	Freq

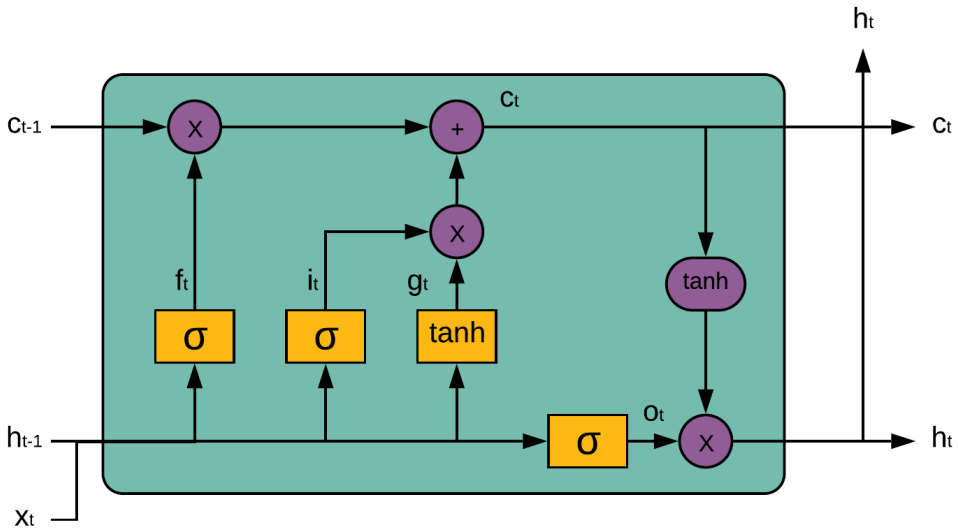
The take away -- Assigning the label RA -> p.o. includes in the probabilities assigning:

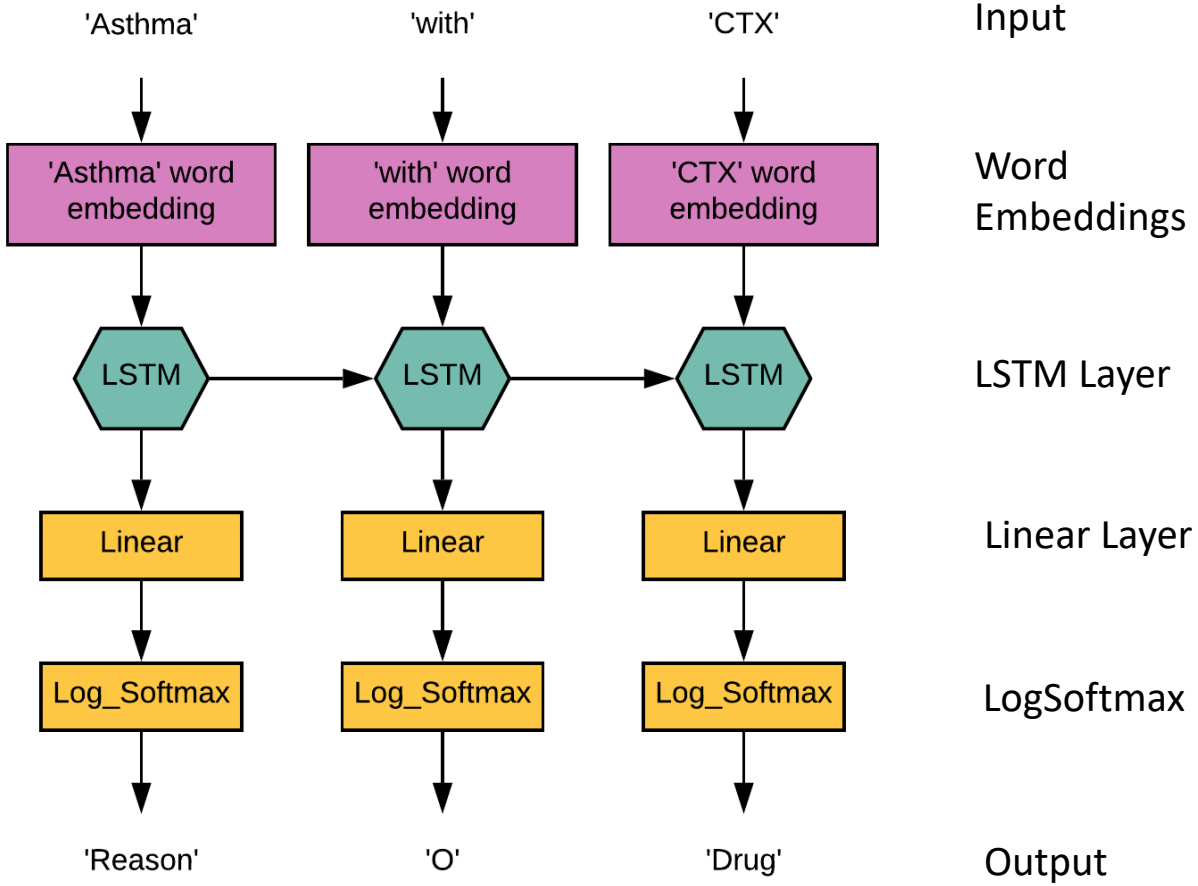
Dose -> 325
DoseUnit -> mg
Frequency-> 4x/day for 2 weeks

bi-LSTM+CRF

- Sequence modeling algorithm
- Takes arbitrarily long contexts into account when predicting the label of a token
- Information flow can run from left-to-right as well as right-to-left
- CRF output layer handles the final predicition

Long Short Term Memory Units





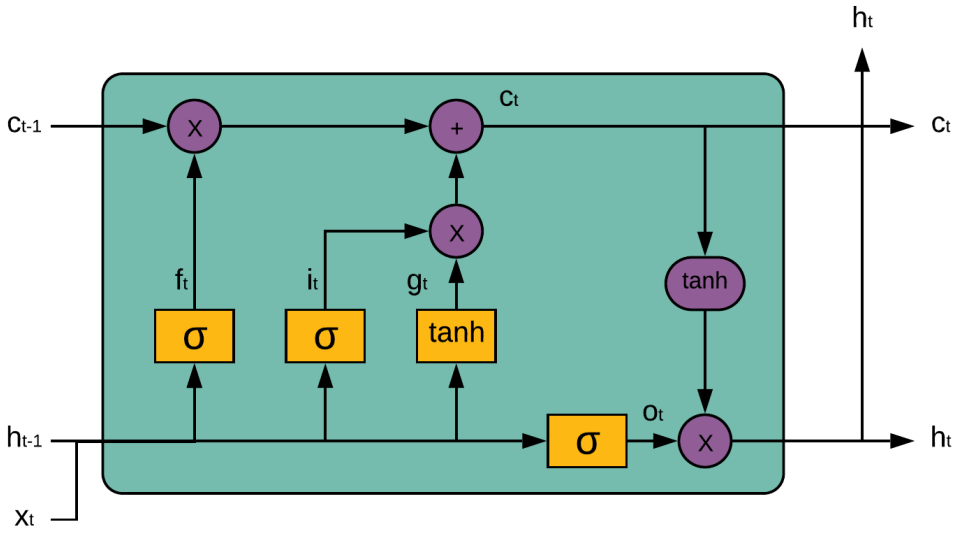
LSTM Cell

x = input
h = hidden state
c = cell state
t = current cell (timestep)

W_{yz} = weights for y at gate z
b_{yz} = bias for y at gate z

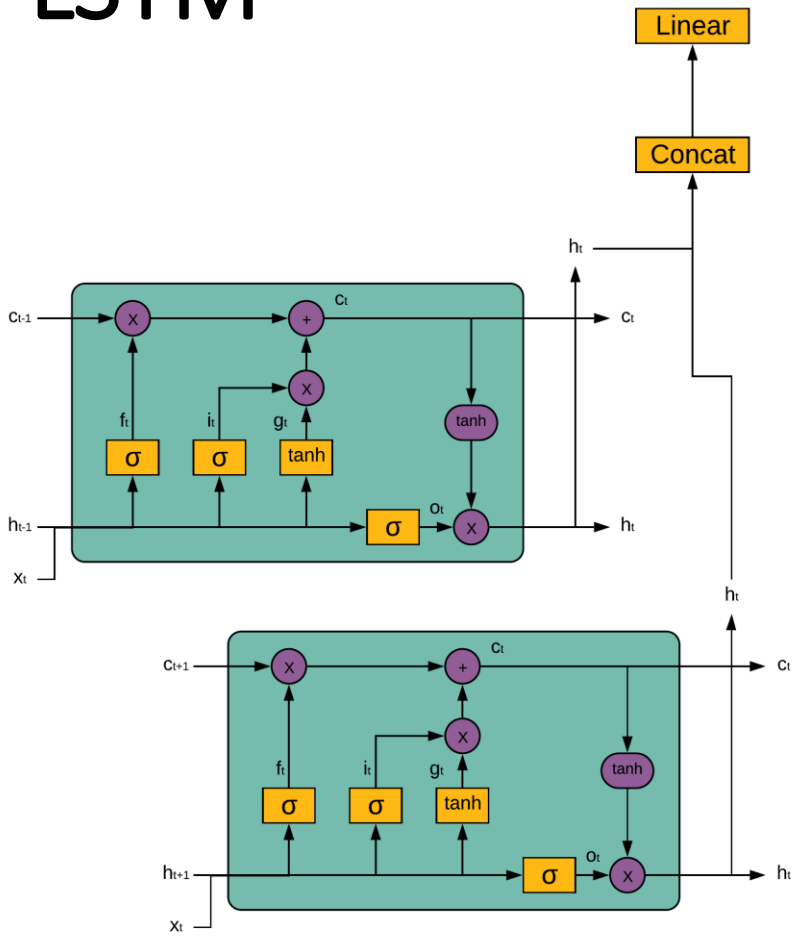
σ = sigmoid
tanh = hyperbolic tangent
* = Hadamard product (not matrix multiplication)

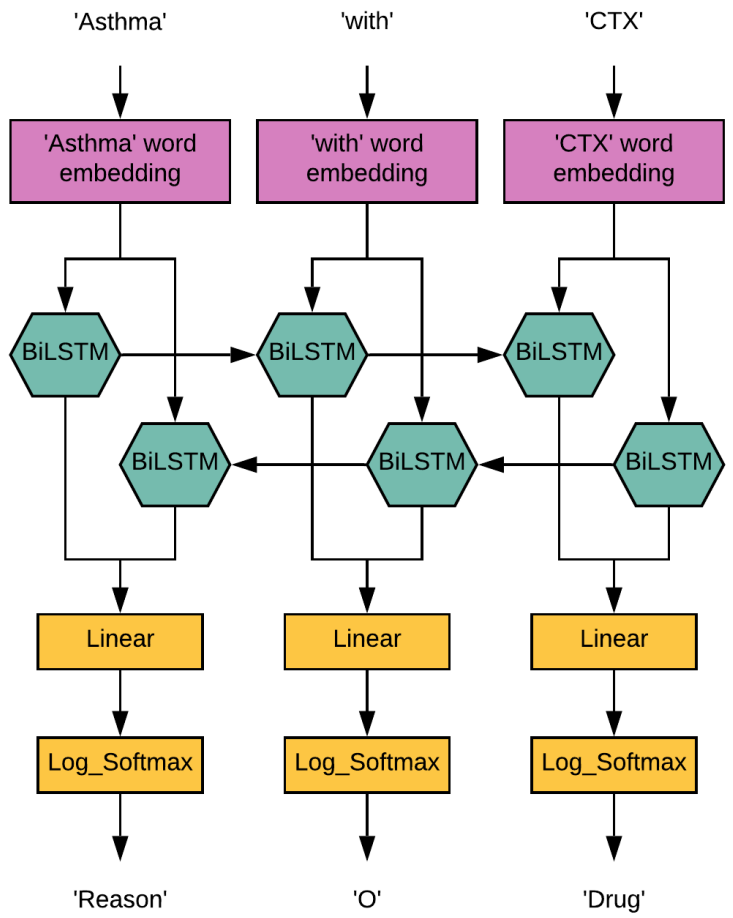
i = input gate
f = forget gate
g = cell gate
o = output gate
*Gates are always W**x** + b + W**h** + b*



$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{(t-1)} + b_{hi})$$
$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{(t-1)} + b_{hf})$$
$$g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{(t-1)} + b_{hg})$$
$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{(t-1)} + b_{ho})$$
$$c_t = f_t * c_{(t-1)} + i_t * g_t$$
$$h_t = o_t * \tanh(c_t)$$

Bi-LSTM





Input

Word Embeddings

LSTM Layer

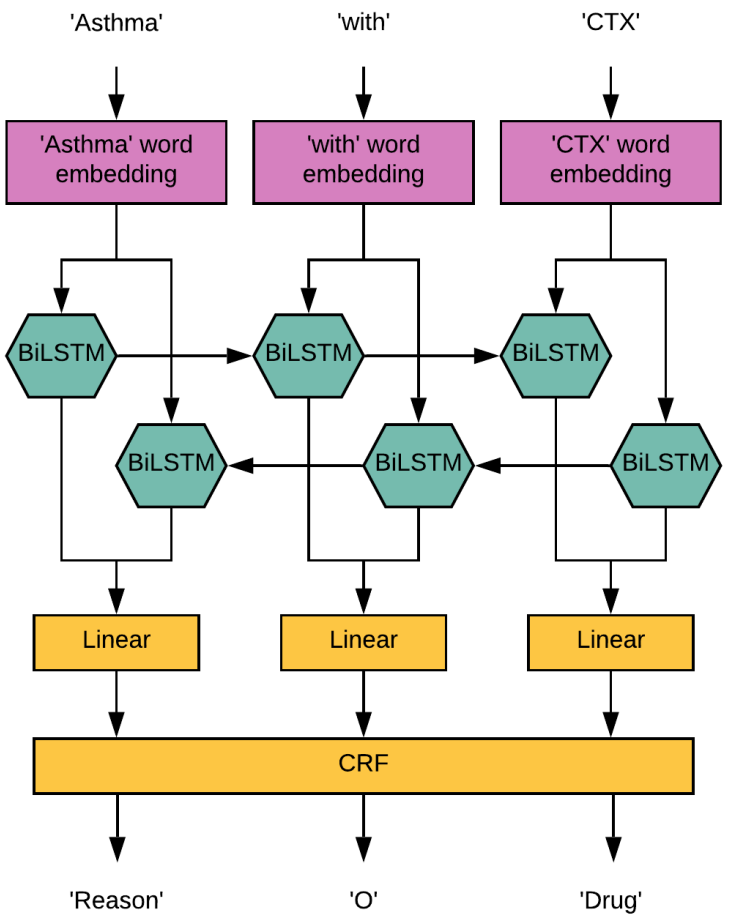
Linear Layer

LogSoftmax

Output

BiLSTM+CRF

$$
\begin{aligned}
P(\mathbf{y} \mid \mathbf{X}) &= \prod_{k=1}^{\ell} P(y_k \mid \mathbf{x}_k) \\
&= \prod_{k=1}^{\ell} \frac{\exp(U(\mathbf{x}_k, y_k))}{Z(\mathbf{x}_k)} \\
&= \frac{\exp\left(\sum_{k=1}^{\ell} U(\mathbf{x}_k, y_k)\right)}{\prod_{k=1}^{\ell} Z(\mathbf{x}_k)}
\end{aligned}
$$



Input

Word Embeddings

LSTM Layer

Linear Layer

CRF

Output

Feature representations

- Feature-based
- Featureless

What are some feature-based features that you think would be good for NER?

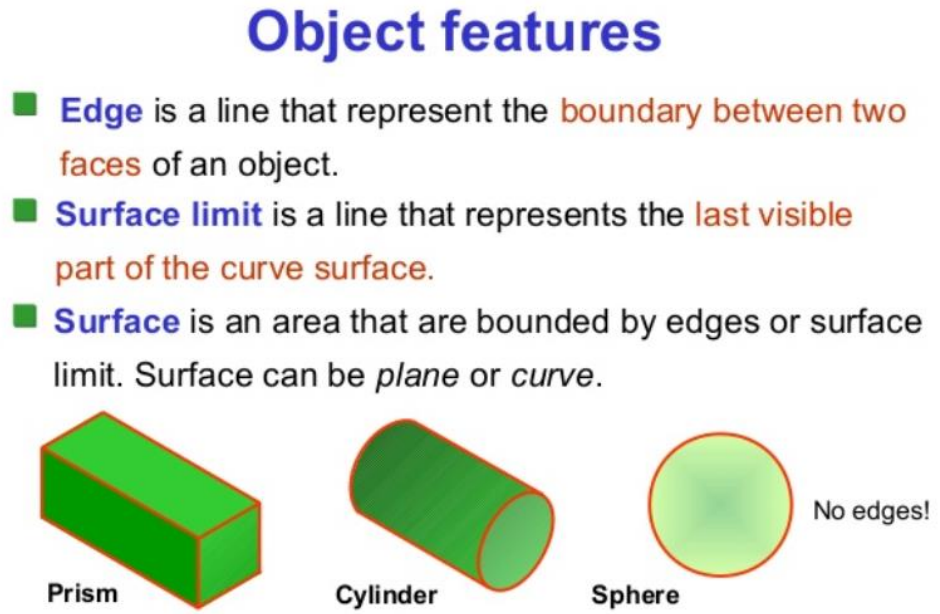
$f_j(s, i, l_i, l_{i-1})$: feature function j

Traditional Features:

- Orthographic
- Morphological
- Lexical
- Syntactic
- Semantic
- Predictions
- Triggers

Orthographic : what does the word look like?

- initial capital letter
- all capital letters
- roman number
- acronym
- single character
- all digits
- contains digits
- contains hyphen
- etc….



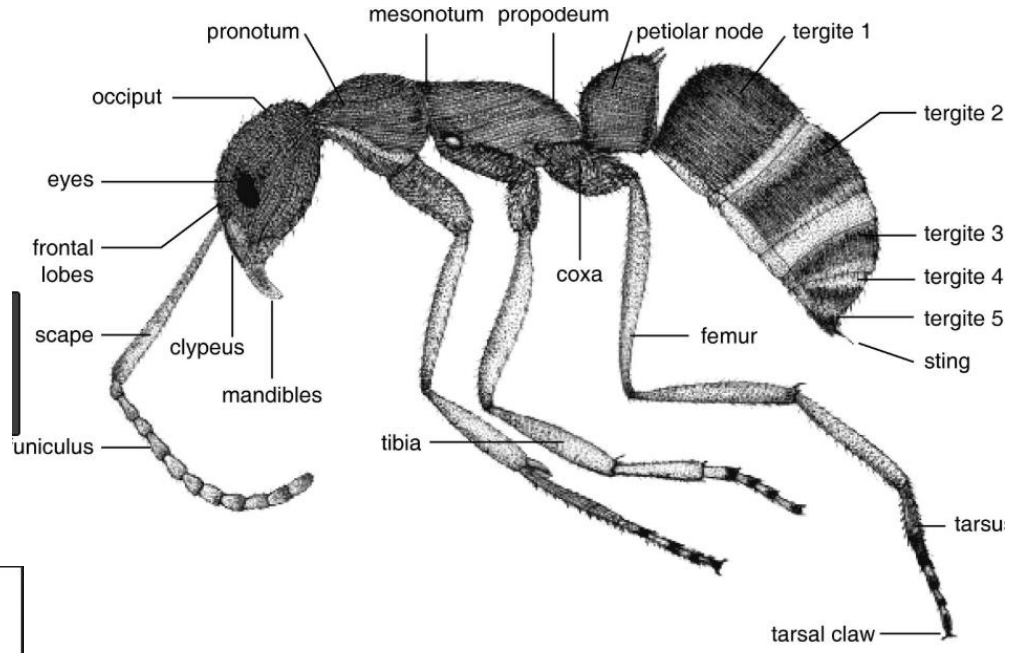
325	mg	aspirin	p.o.	4x/day
Dose	DoseUnit	Drug	RA	Freq

Morphological: what is the word made up of?

- suffix
- prefix

325	mg	aspirin	p.o.	4x/day
Dose	DoseUnit	Drug	RA	Freq

ACE INHIBITORS	ALBUMIN	ASPIRIN	ATENOLOL
AVASTIN	BONIVA	NORVASC	CELEXA
CIPRO	COUMADIN	ERYTHROPOIETIN	GLEEVEC
HEPARIN	HIRUDIN	HUMIRA	LASIX
LIPITOR	MORPHINE	METOCLOPRAMIDE	METHOTREXATE
NEXIUM	NITROGLYCERIN	NSAID	PREDNISONE
PROTAMINE SULFATE	RED BLOOD CELLS	TRASYLOL	TYLENOL
PLATELETS	VANCOMYCIN	VERSED	VIOXX
YAZ	VITAMIN A		

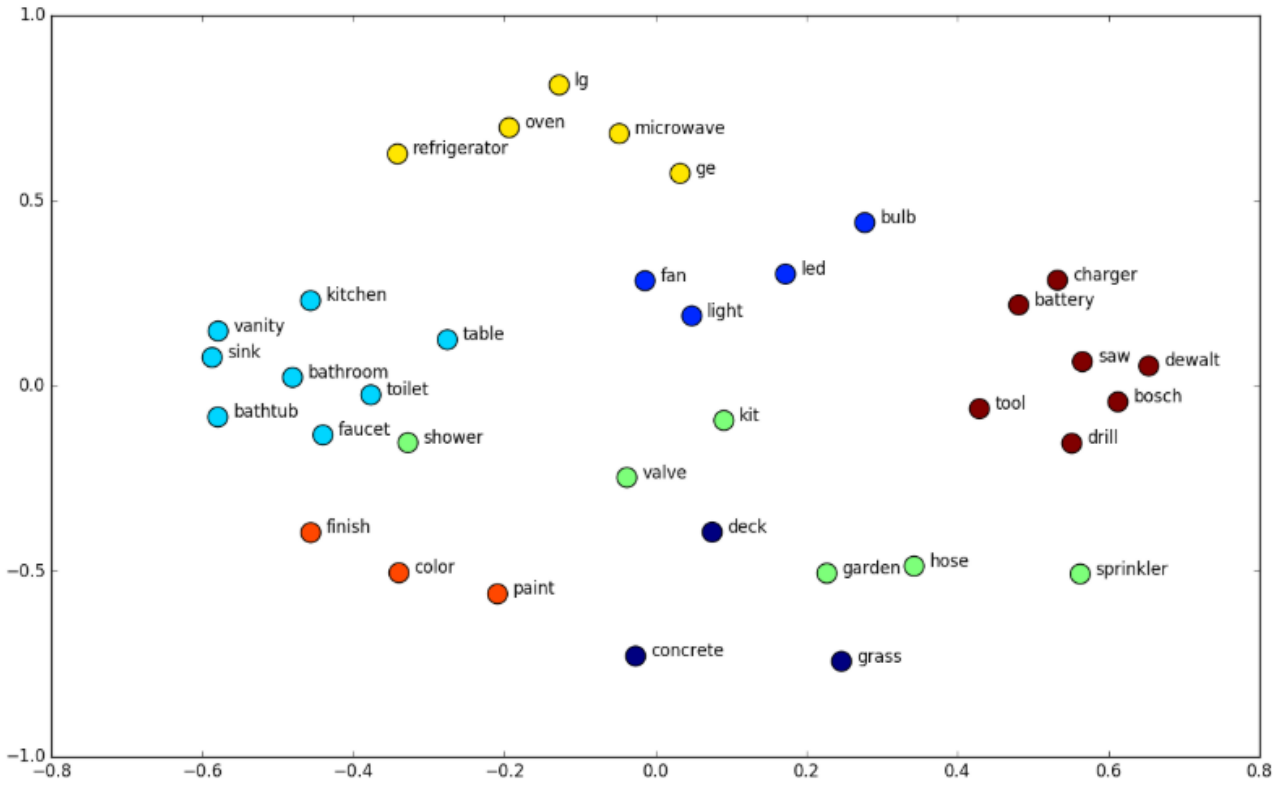


Lexical

"You shall know a word by the company it keeps"

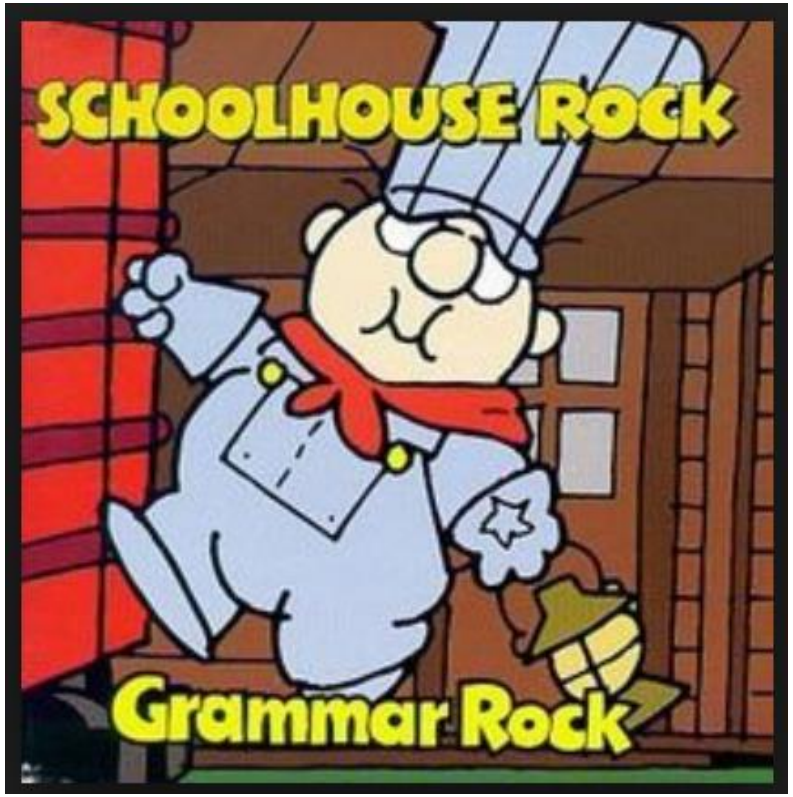
–J. R. Firth 1957

- how can we represent the meaning of a word?



Syntactic

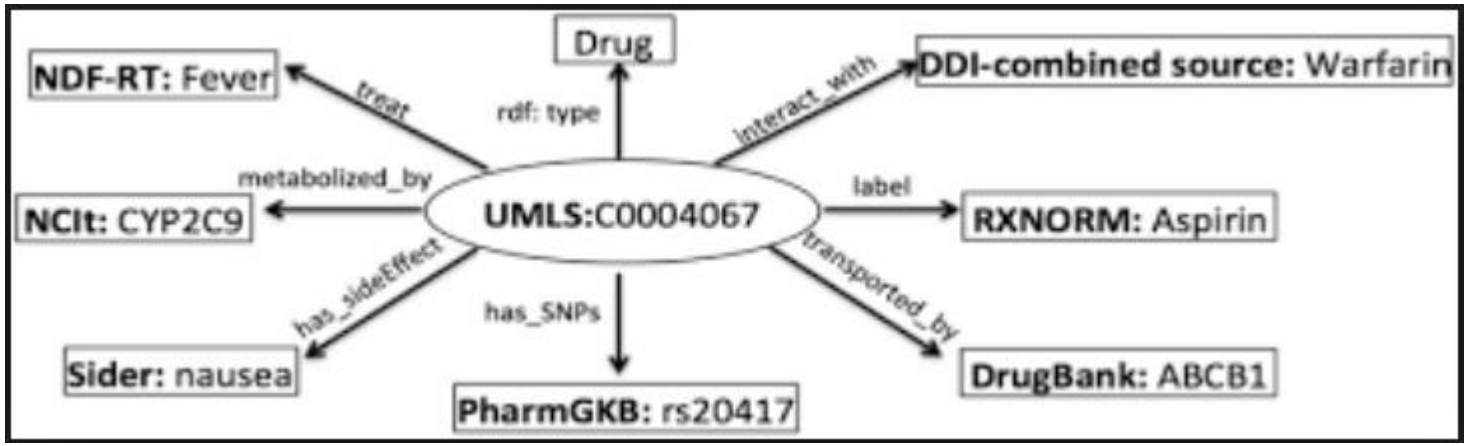
- Current words POS
- POS of surrounding word
  - Patterns:
    - DET NN NN
  - Part of a term



CD	CD	NN	NN	CD	Prep	CD	NN
325	mg	aspirin	p.o.	4x/day	for	2	weeks
Dose	DoseUnit	Drug	RA	Freq	Freq	Freq	Freq

Semantic

- concept from a taxonomy (e.g. WordNet, UMLS)
- semantic group or type
- semantic relationship



325	**mg**	**aspirin**	**p.o.**	**4x/day**
Dose	DoseUnit	Drug	RA	Freq

Triggers

- Trigger words
  - for person (Mr., Mrs., Dr.)
  - for location (city, street, avenue)
  - for organizations (Inc, Co., Ltd.)

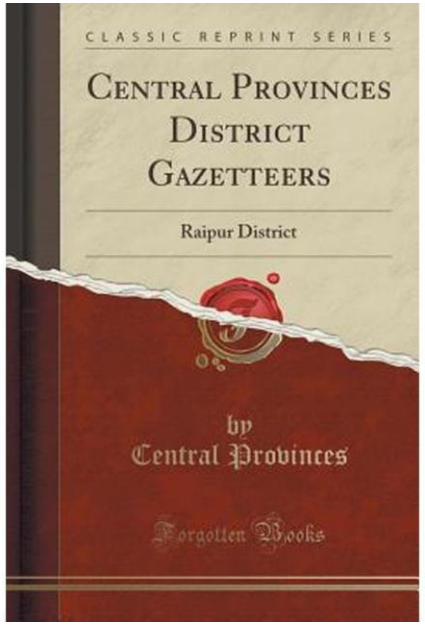
What do you need to know in order to identify triggers?

Gazeteers

- Gazetteers
  - names of cities, countries, villages, streets
  - names of organizations and companies
  - census
    - people's first names
    - people's last names
  - Drug lists

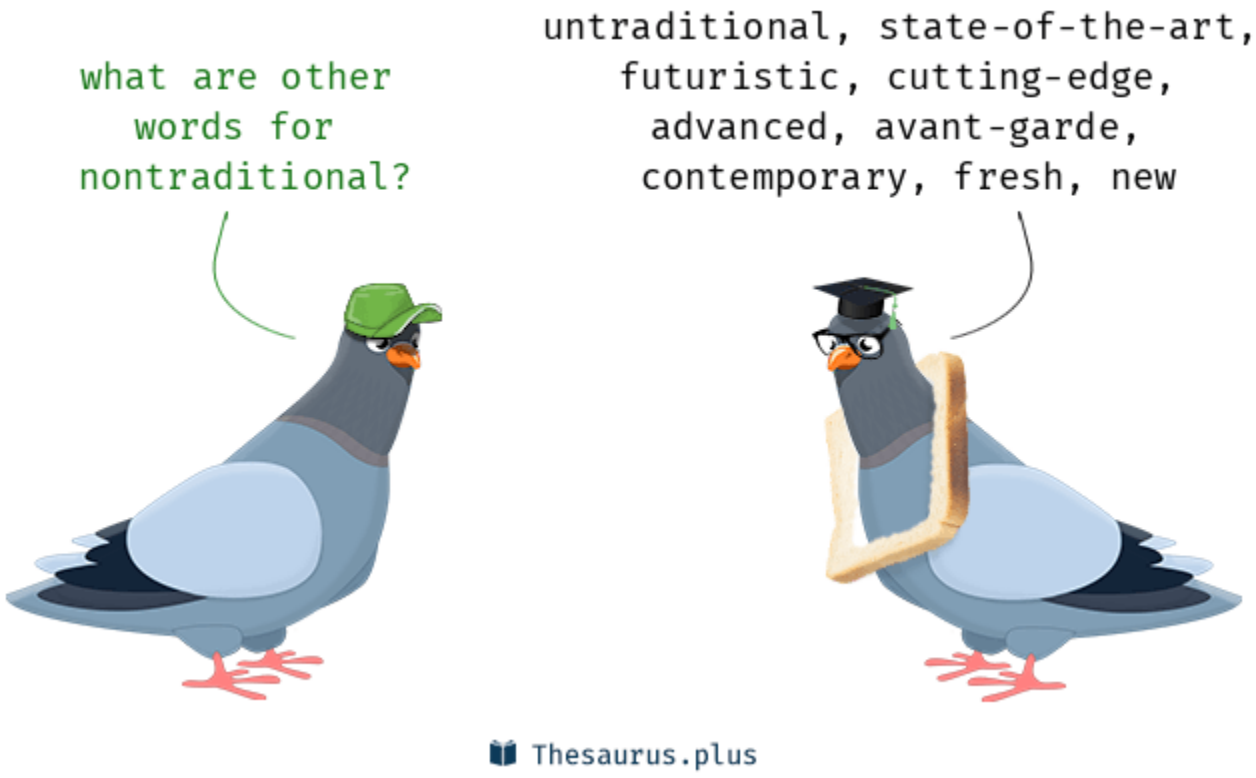
- Where do these Gazeteers come from:
  - Previously:
    - Census data
    - Lists of companies

  - Also: Wikipedia
    - Artwork: novels, books, paintings, operas, plays
    - Named Objects: aircraft tanks, rifles, weapons
    - Events: playoffs, championships, races



Non traditional features

- What other features do you think could be incorporated?

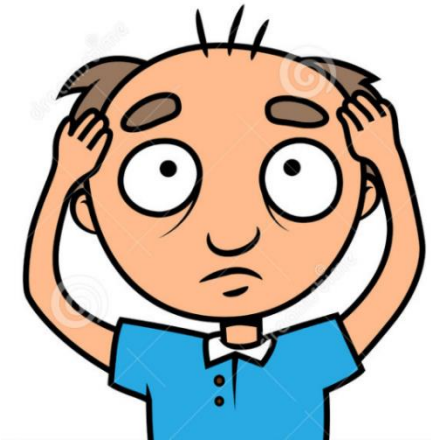


Feature-based

What difficulty do we have with these feature-based representations?

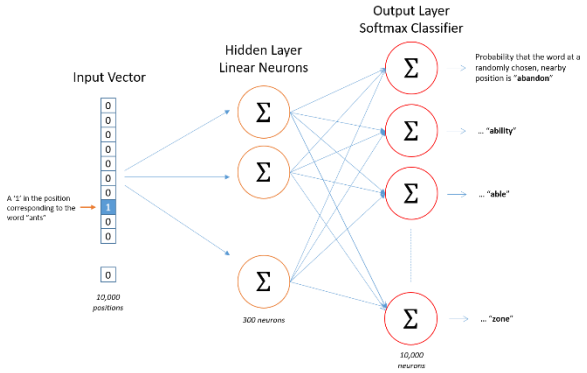
Feature-based

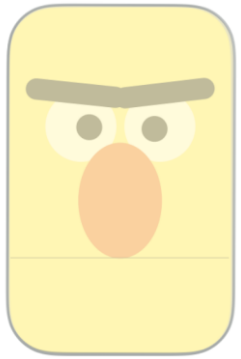
What difficulty do we have with these feature-based representations?

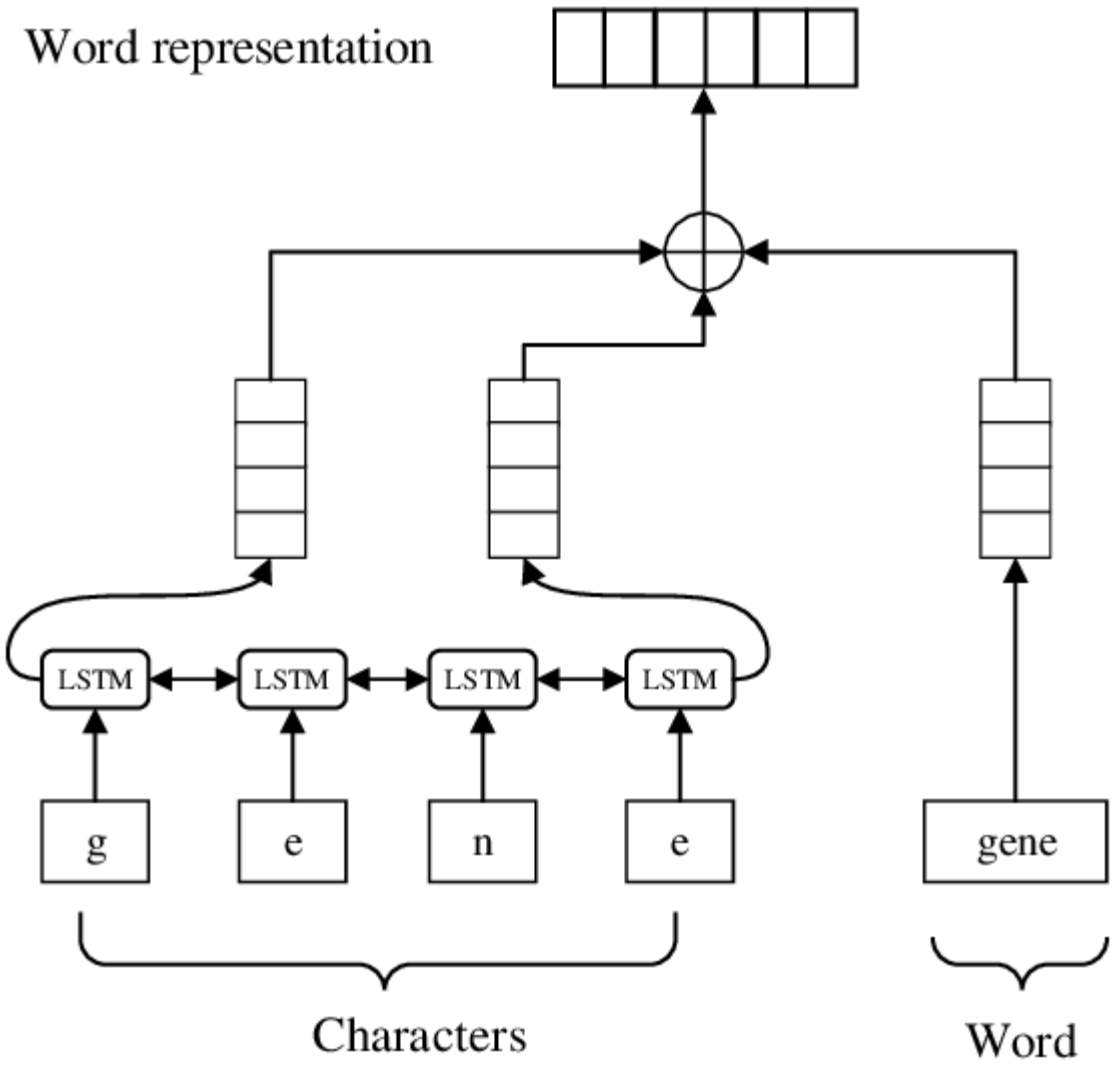


Featureless

- Features:
  - word embeddings
    - Glove
    - W2V
    - BERT
    - ELMO
    - ERNIE
  - character embeddings







Featureless

What is the problem though with these featureless representations?

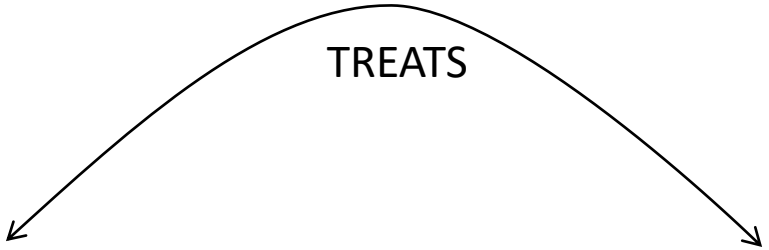
Featureless

What is the problem though with these featureless representations?

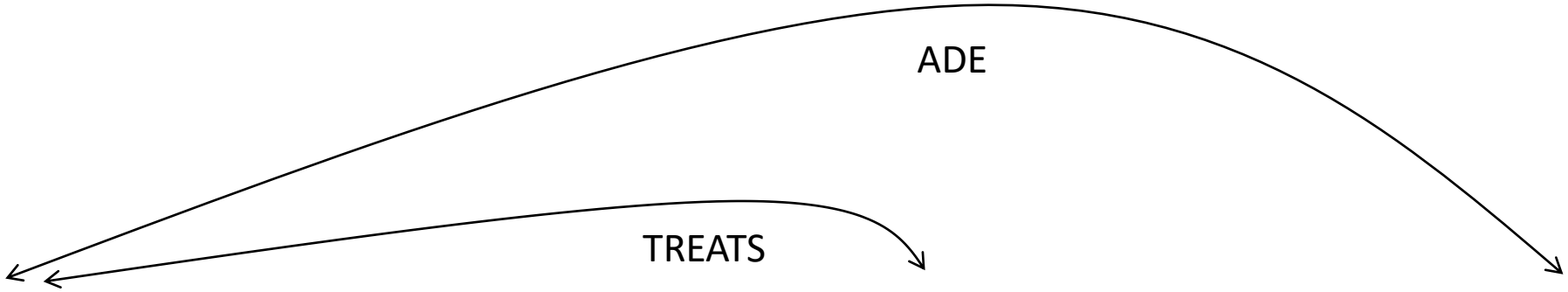
Are we losing structural knowledge?

Relationship Extraction

Relationship extraction



The	doctor	prescribed	him	aspirin	for	his	headache
O	O	O	O	Drug	O	O	Reason



The	Accupril	prescribed	for	his	high	blood	pressure	gave	him	a	headache
O	Drug	O	O	O	Reason	Reason	Reason	O	O	O	Adverse reaction

Heuristics

The	doctor	prescribed	him	aspirin	for	his	headache
O	O	O	O	Drug	O	O	Reason

The	Accupril	prescribed	for	his	high	blood	pressure	gave	him	a	headache
O	Drug	O	O	O	Reason	Reason	Reason	O	O	O	Adverse reaction

Drug names appear prior to their dosages, modes, frequencies, durations, and reasons in more than 90% of the cases

in addition

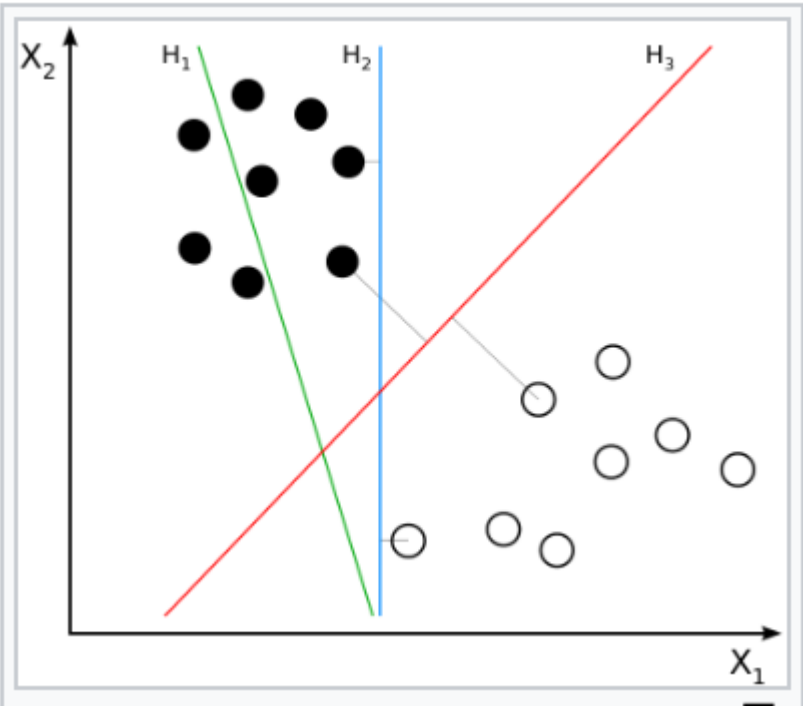
they are more likely to occur in the same sentence together

Supervised machine learning algorithms

- Feature-based:
  - Support Vector Machines (SVMs)
- Featureless:
  - Convolutional Neural Networks (cNN)
    - sentence cNN
    - segment cNN

Support Vector Machines

In machine learning, **support vector machines** (**SVMs**, also **support vector networks**[1]) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.



Who has taken fuzzy logic?

SVM: Feature-based representation

The	doctor	prescribed	him	aspirin	for	his	headache
O	O	O	O	Drug	O	O	Reason

The	Accupril	prescribed	for	his	high	blood	pressure	gave	him	a	headache
O	Drug	O	O	O	Reason	Reason	Reason	O	O	O	Adverse reaction

What features?

SVM: Feature-based representation

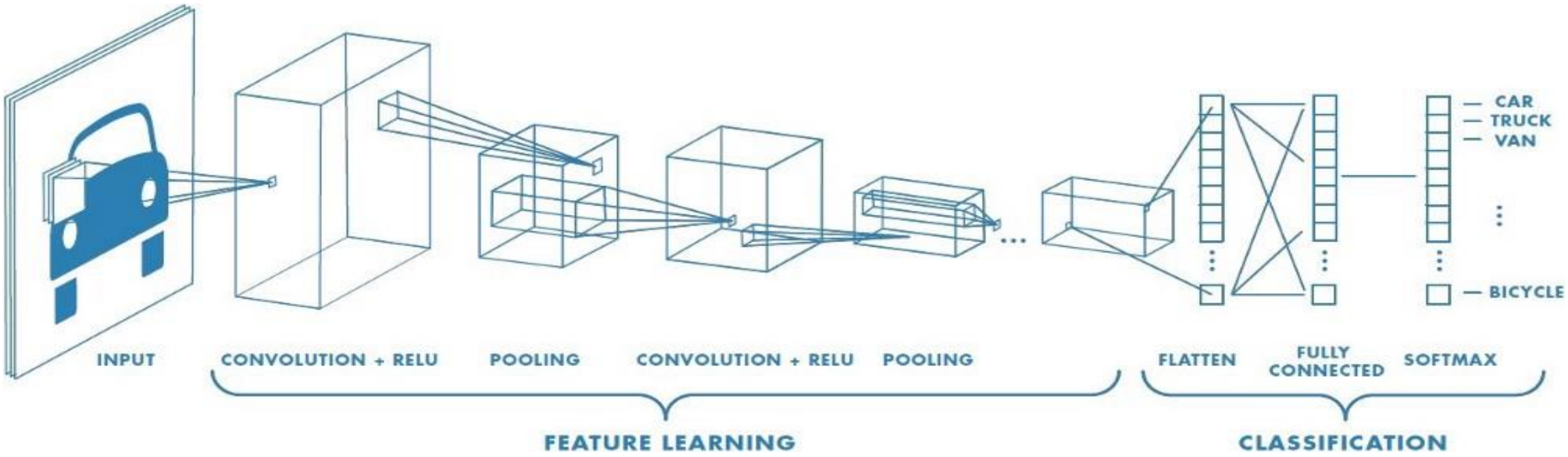
The	doctor	prescribed	him	aspirin	for	his	headache
O	O	O	O	Drug	O	O	Reason

The	Accupril	prescribed	for	his	high	blood	pressure	gave	him	a	headache
O	Drug	O	O	O	Reason	Reason	Reason	O	O	O	Adverse reaction

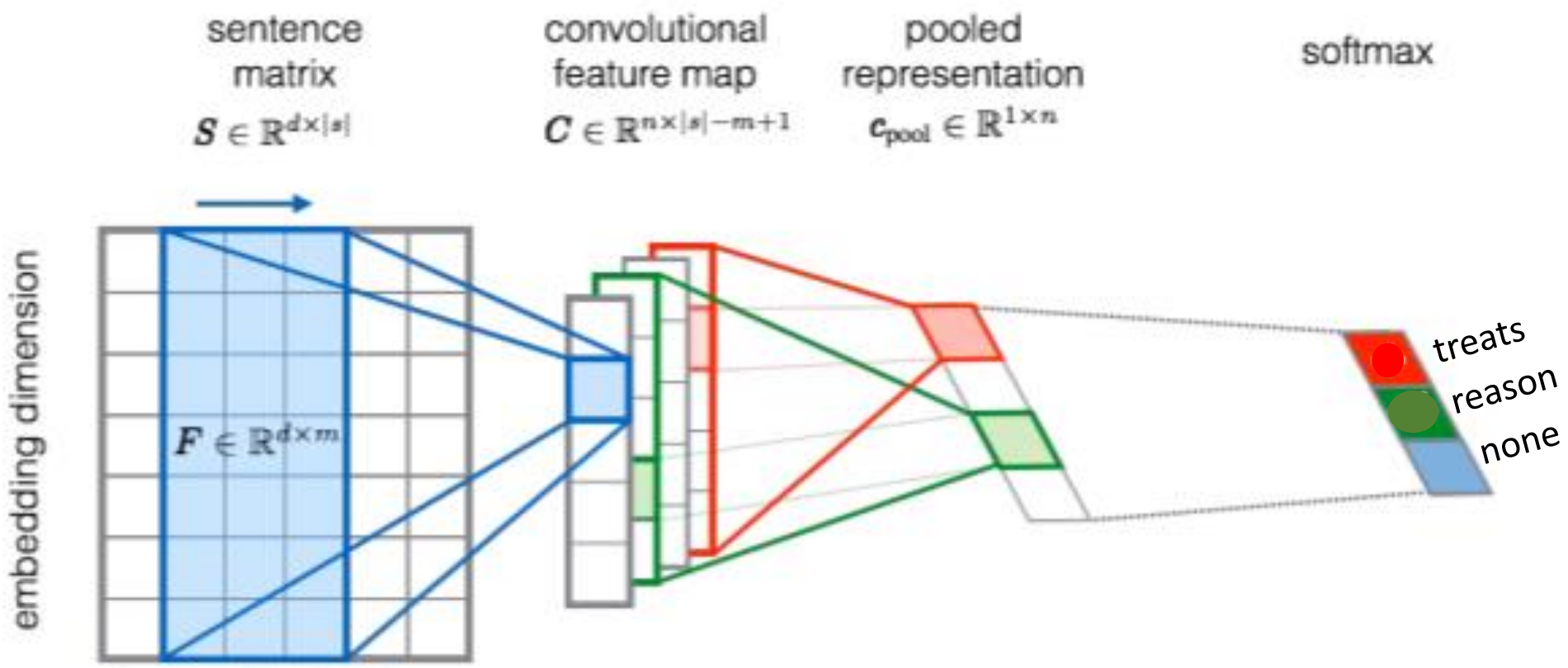
Possible Features:

- Entity types
- Context between and around the entities
- Syntactic information

cNN: Featureless representation



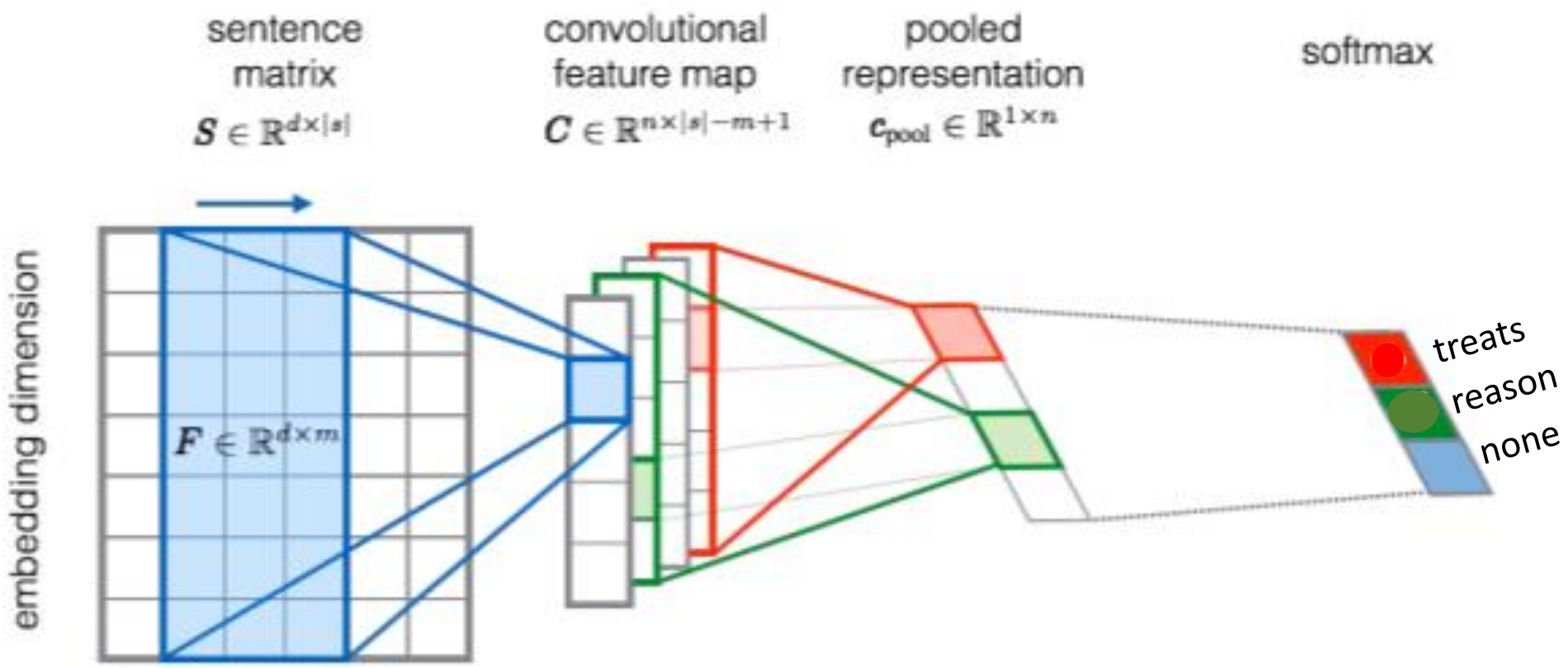
sentence cNN



I took aspirin for my headache

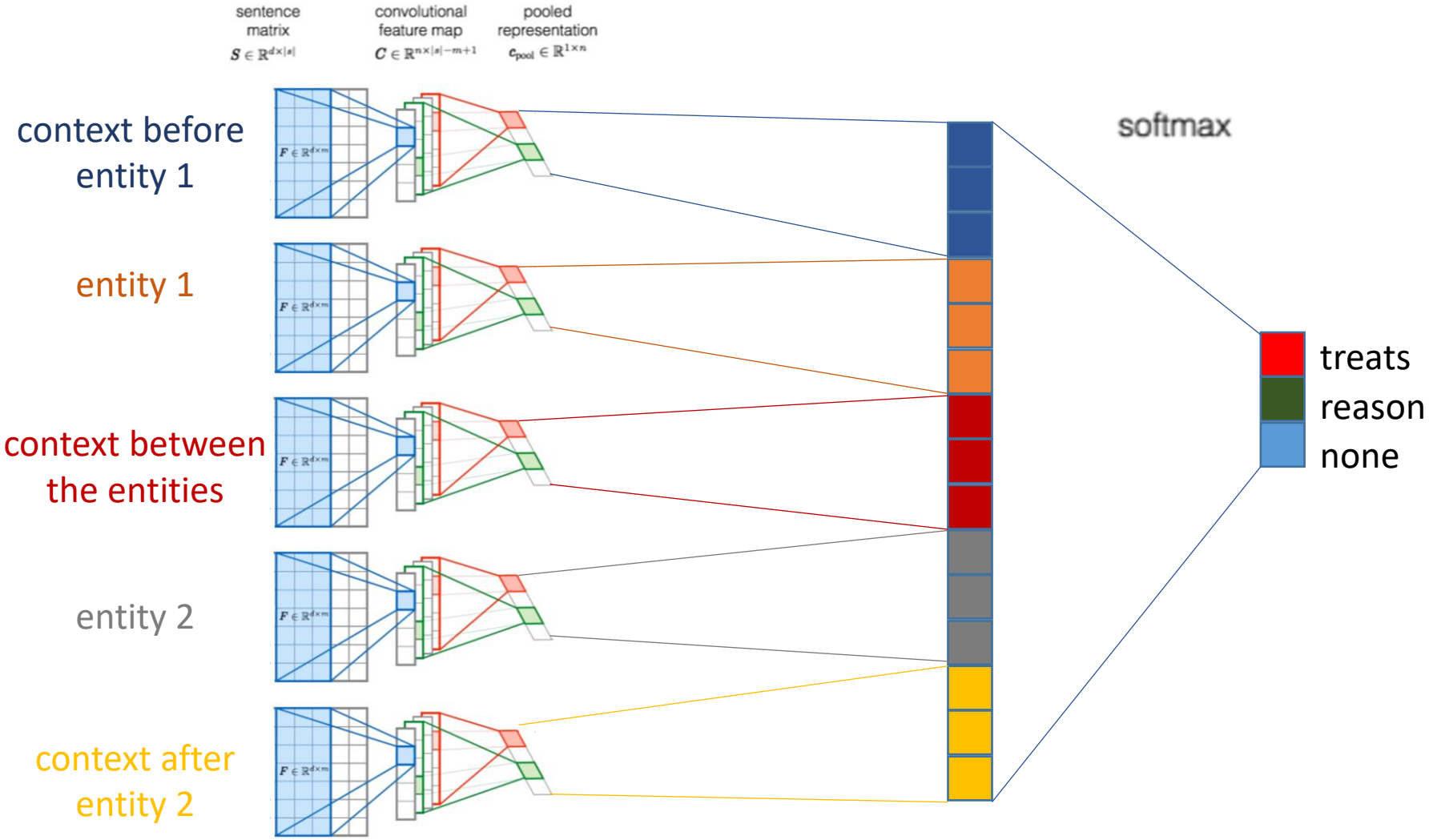
sentence cNN

What potential difficulty do you see with sentence cNN?

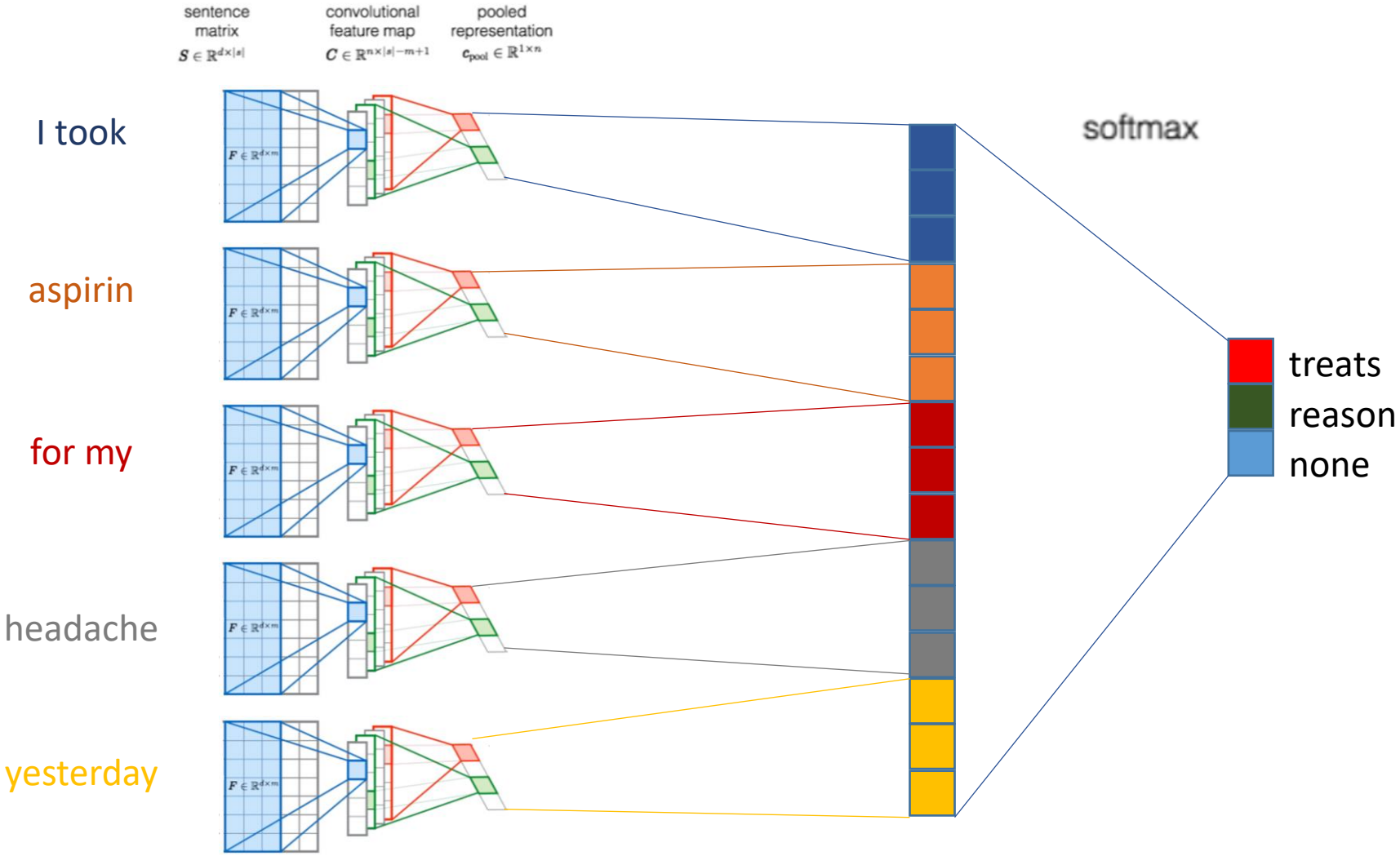


I took aspirin for my headache

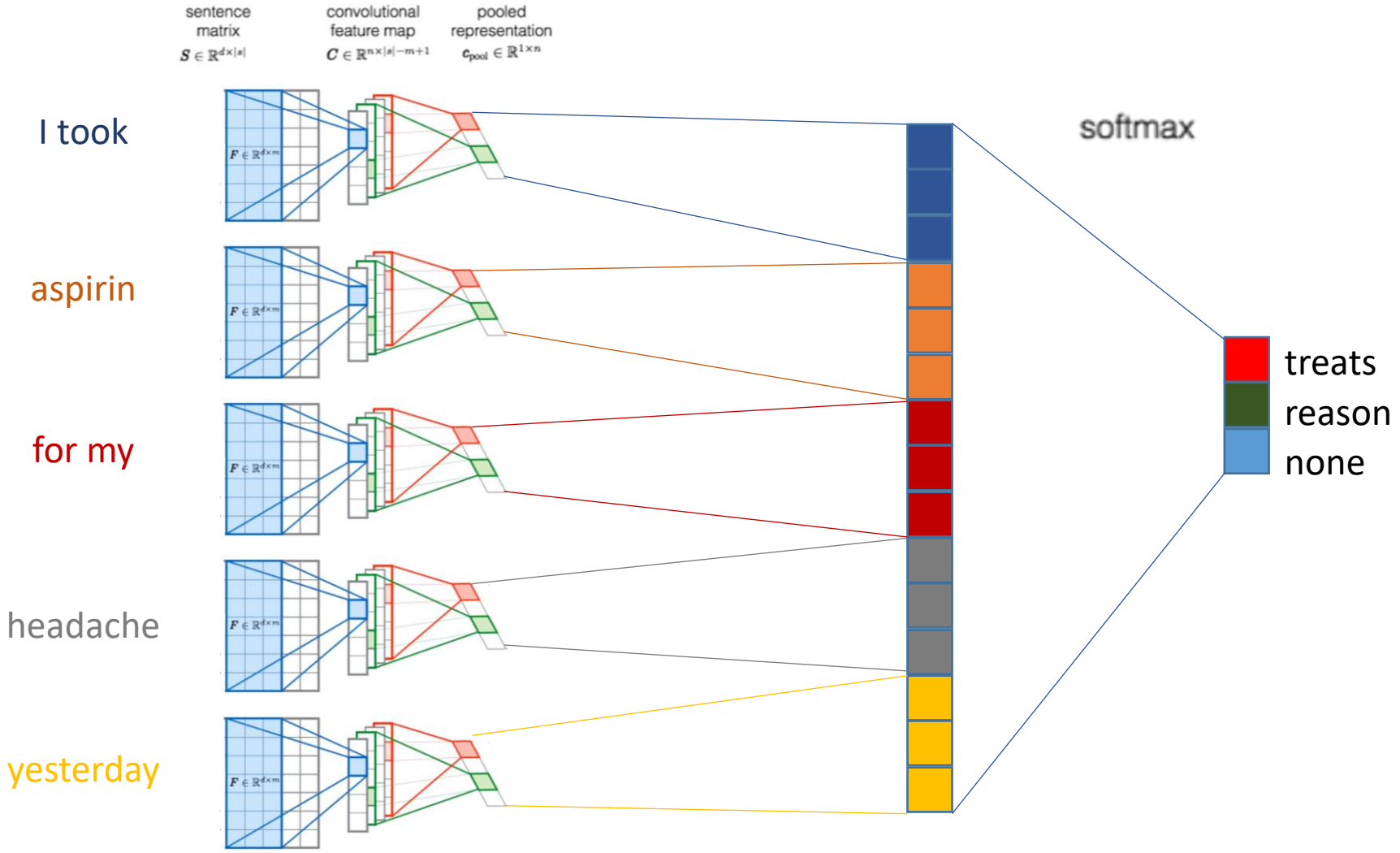
segment cNN


sentence matrix
$S \in \mathbb{R}^{d \times |s|}$
convolutional feature map
$C \in \mathbb{R}^{n \times |s|-m+1}$
pooled representation
$c_{pool} \in \mathbb{R}^{1 \times n}$
context before entity 1
entity 1
context between the entities
entity 2
context after entity 2
softmax
treats
reason
none

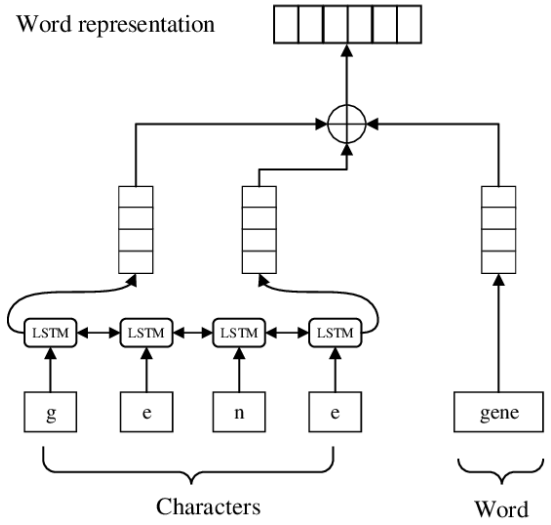
segment cNN


sentence matrix
$S \in \mathbb{R}^{d \times |s|}$
convolutional feature map
$C \in \mathbb{R}^{n \times |s|-m+1}$
pooled representation
$c_{pool} \in \mathbb{R}^{1 \times n}$
I took
aspirin
for my
headache
yesterday
$F \in \mathbb{R}^{d \times m}$
softmax
treats
reason
none

segment cNN



- Features:
  - word embeddings
    - Glove
    - W2V
    - BERT
    - ELMO
    - ERNIE
  - character embeddings



What are featureless respresentations primarily incorporating?

What are featureless respresentations primarily incorporating?

Context

A woman with a history of congestive heart failure, dementia, gastroesophageal reflux disease, degenerative joint disease and arthritis who presents with upper gastrointestinal bleed thought secondary to NSAID use

Drug
Disease
Adverse Reaction

A woman with a history of congestive heart failure, dementia, gastroesophageal reflux disease, degenerative joint disease and arthritis who presents with upper gastrointestinal bleed thought secondary to NSAID use

Drug
Disease
Adverse Reaction

Question: what is the Drug-Disease relationship in this sentence?

A woman with a history of congestive heart failure, dementia, gastroesophageal reflux disease, degenerative joint disease and arthritis who presents with upper gastrointestinal bleed thought secondary to NSAID use

Drug
Disease
Adverse Reaction

Context doesn't always help us in this case

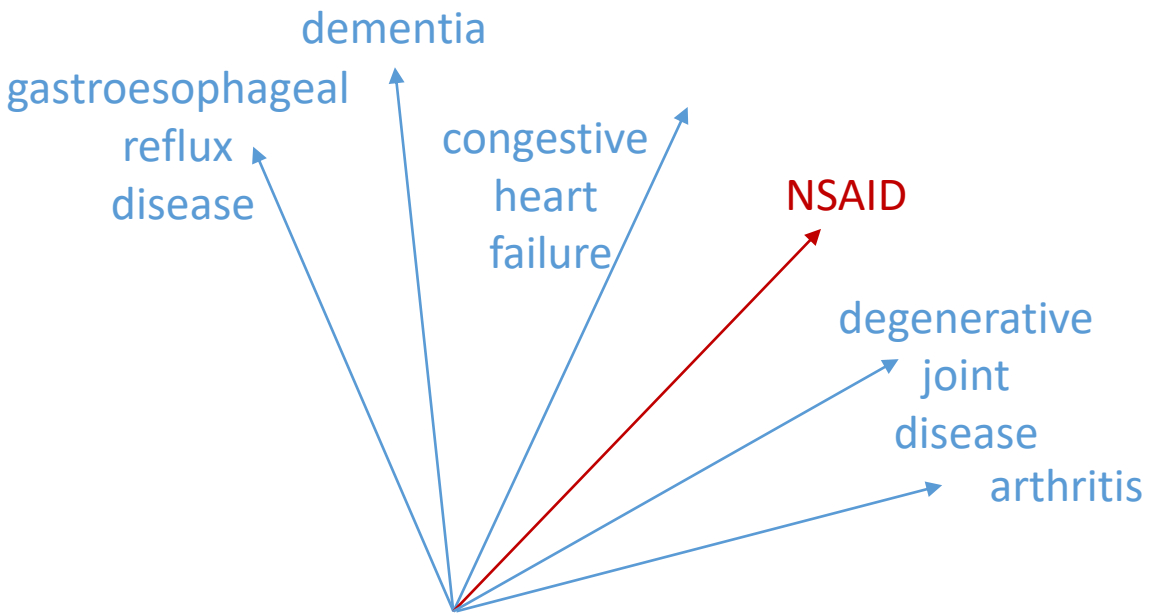
Outside knowledge

Nonsteroidal anti-inflammatory drugs, or NSAIDs (pronounced en-saids), are the most prescribed medications for treating conditions such as arthritis. Most people are familiar with over-the-counter, nonprescription NSAIDs, such as **aspirin** and **ibuprofen**. NSAIDs are more than just **pain relievers**.

A woman with a history of congestive heart failure, dementia, gastroesophageal reflux disease, degenerative joint disease and arthritis who presents with upper gastrointestinal bleed thought secondary to NSAID use

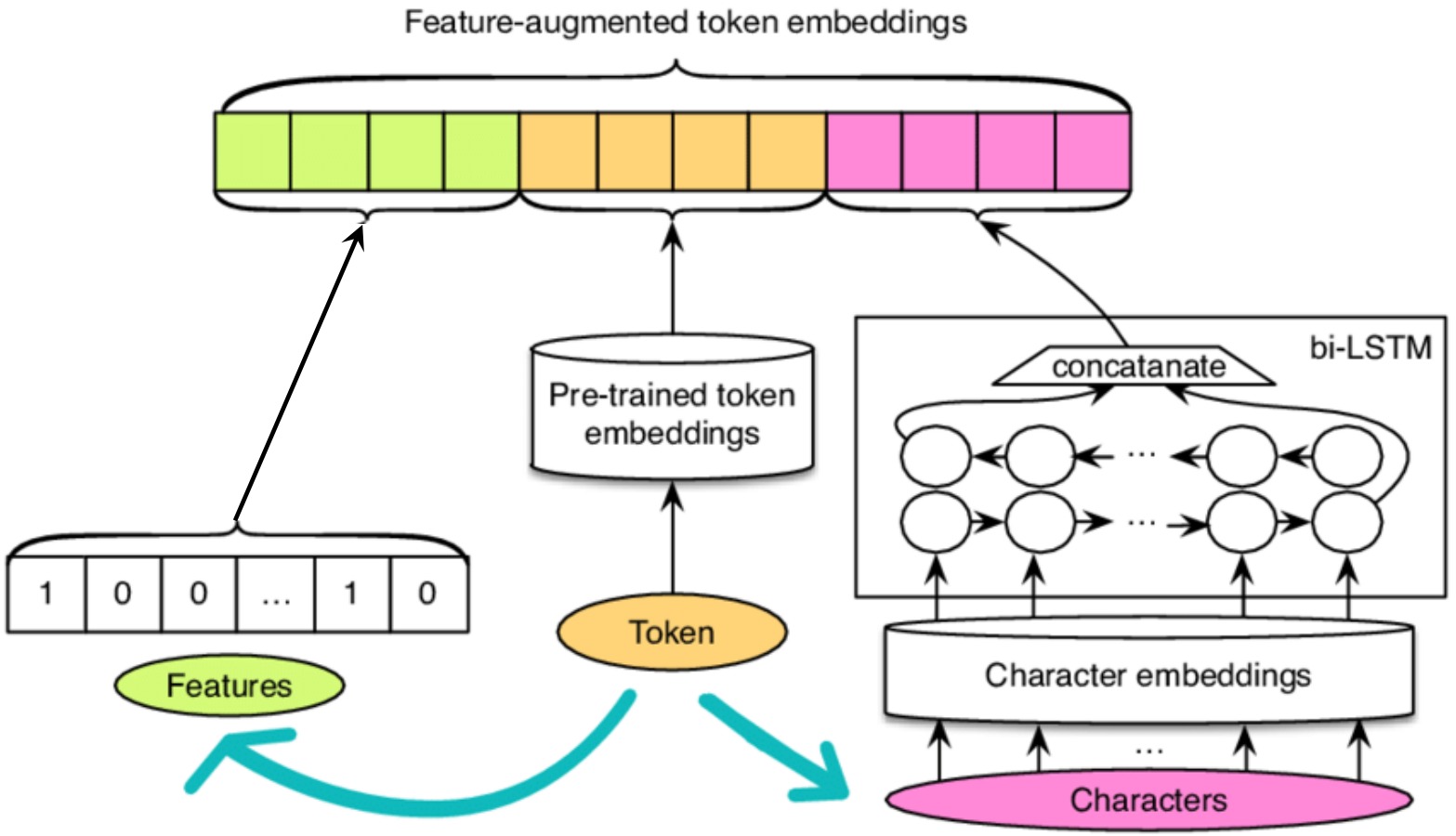
Drug
Disease
Adverse Reaction

Outside knowledge

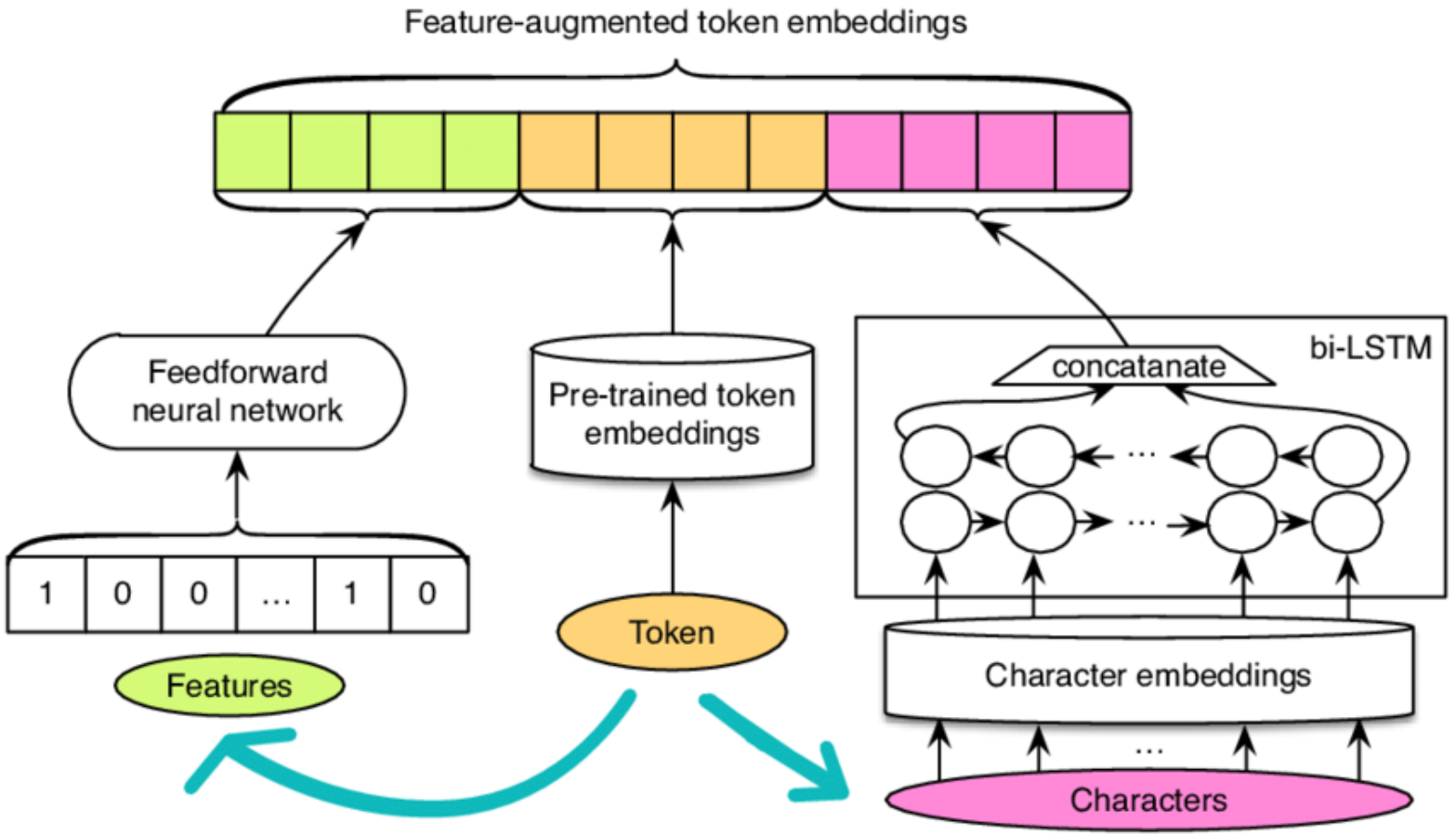


- Metrics:
  - Association measures
  - Relatedness measures
- Corpus:
  - FDA Drug Labels
  - Medline
  - Clinical notes

Combining feature-based and featureless representations



Combining feature-based and featureless representations



Questions?