Applied Natural Language Processing

Info 256

Lecture 21: NER (Nov 6, 2023)

David Bamman, UC Berkeley

Named entity recognition

[Mrs Oedipa Maas]$_{\text{PER}}$ came home from a Tupper-ware party

- Identifying spans of text that correspond to typed entities that are proper names.

Named entity recognition

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	The **Mt. Sanitas** loop is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the **Golden Gate Bridge**.
Vehicles	VEH	planes, trains, automobiles	It was a classic **Ford Falcon**.

Figure 17.1 A list of generic named entity types with the kinds of entities they refer to.

ACE NER categories (+weapon)

Named entity recognition

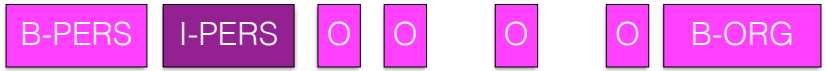
- GENIA corpus of MEDLINE abstracts (biomedical)

We have shown that [interleukin-1]$_{\text{PROTEIN}}$ ([IL-1]$_{\text{PROTEIN}}$) and [IL-2]$_{\text{PROTEIN}}$ control [IL-2 receptor alpha (IL-2R alpha) gene]$_{\text{DNA}}$ transcription in [CD4-CD8- murine T lymphocyte precursors]$_{\text{CELL LINE}}$

protein
cell line
cell type
DNA
RNA

http://www.aclweb.org/anthology/W04-1213

BIO notation

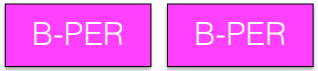


tim cook is the ceo of apple

- **B**eginning of entity
- **I**nside entity
- **O**utside entity

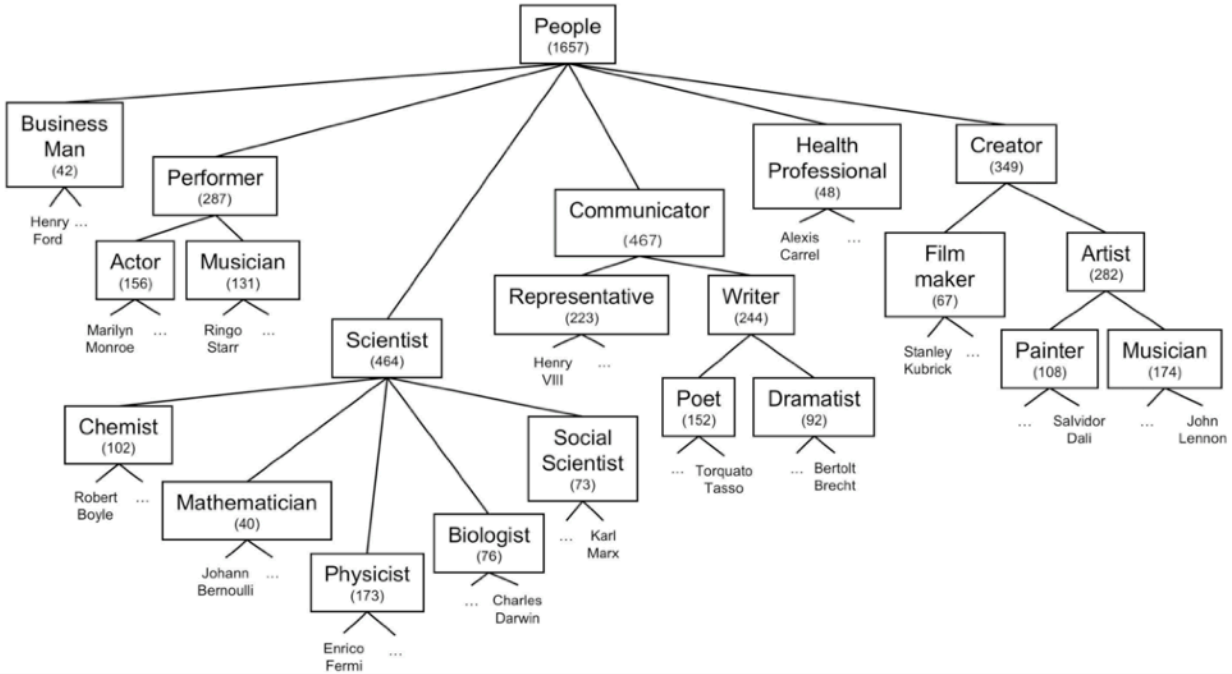
[tim cook]$_{PER}$ is the ceo of [apple]$_{ORG}$

Named entity recognition



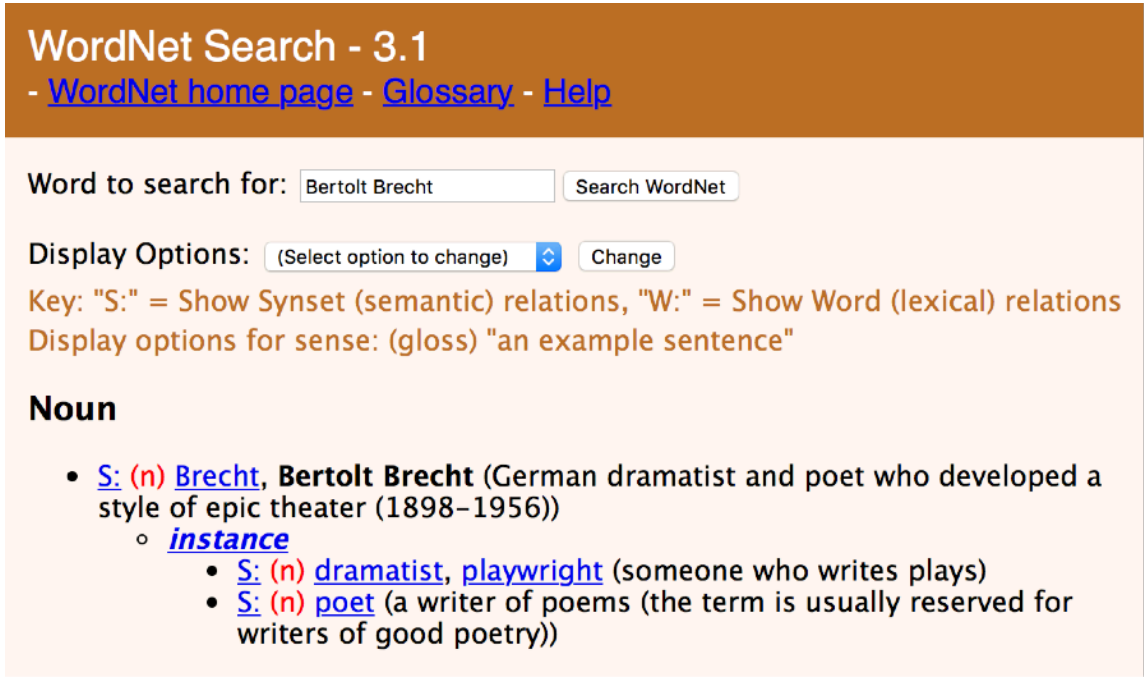
After he saw Harry Tom went to the store

Fine-grained NER



Giuliano and Gliozzo (2008)

Fine-grained NER



Entity recognition

Person	… named after [the daughter of a Mattel co-founder] …
Organization	[The Russian navy] said the submarine was equipped with 24 missiles
Location	Fresh snow across [the upper Midwest] on Monday, closing schools
GPE	The [Russian] navy said the submarine was equipped with 24 missiles
Facility	Fresh snow across the upper Midwest on Monday, closing [schools]
Vehicle	The Russian navy said [the submarine] was equipped with 24 missiles
Weapon	The Russian navy said the submarine was equipped with [24 missiles]

ACE entity categories

Named entity recognition

- Most named entity recognition datasets have flat structure (i.e., non-hierarchical labels).

  ✔ [The University of California]$_{ORG}$

  ✖ [The University of [California]$_{GPE}$]$_{ORG}$

- Mostly fine for named entities, but more problematic for general entities:

  [[John]$_{PER}$'s mother]$_{PER}$ said …

Nested NER

named	after	the	daughter	of	a	Mattel	co-founder
						B-ORG	
					B-PER	I-PER	I-PER
		B-PER	I-PER	I-PER	I-PER	I-PER	I-PER

Sequence labeling

$$x = \{x_1, \ldots, x_n\}$$

$$y = \{y_1, \ldots, y_n\}$$

- For a set of inputs x with n sequential time steps, one corresponding label y_i for each x_i
- Model correlations in the labels y.

Maximum Entropy Markov Model (MEMM)

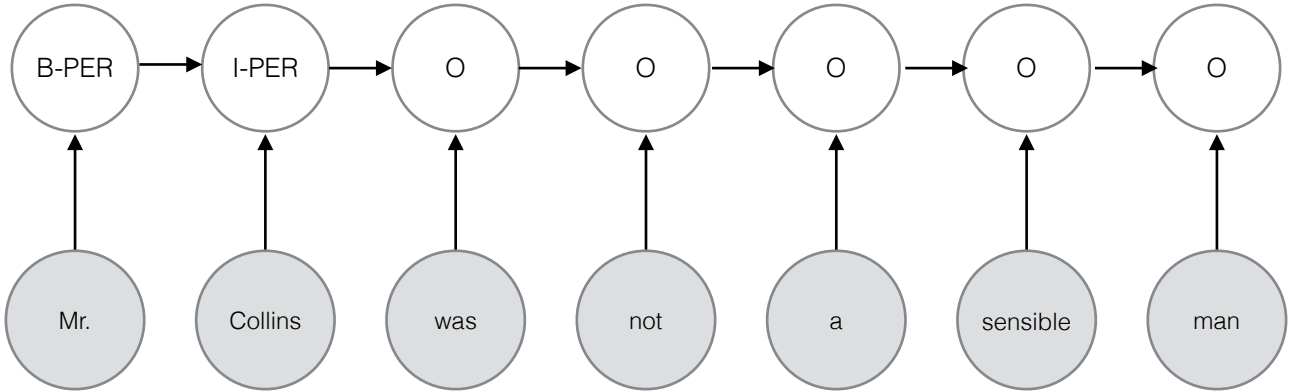
General maximum entropy form (e.g., logistic regression)

$$\arg\max_y P(y \mid x, \beta)$$

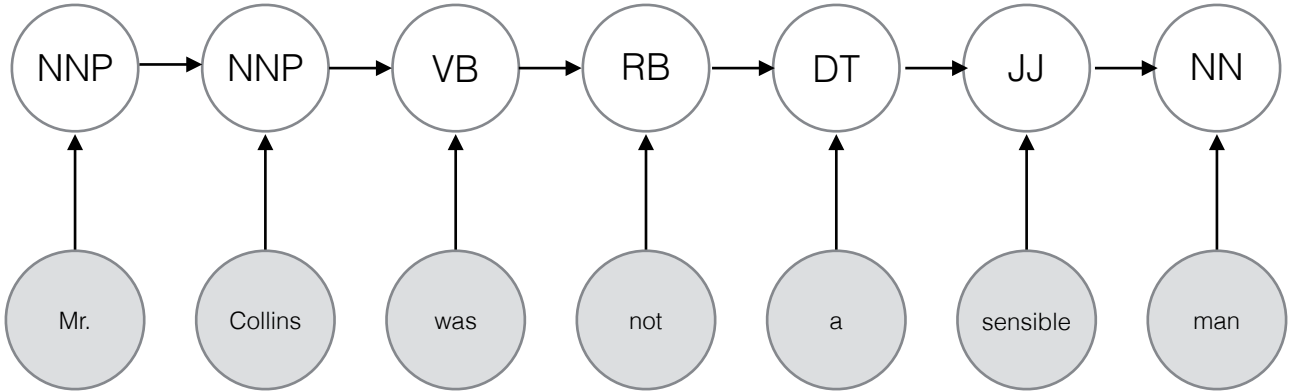
Maxent with Markov assumption: Maximum Entropy Markov Model

$$\arg\max_y \prod_{i=1}^{n} P(y_i \mid y_{i-1}, x)$$

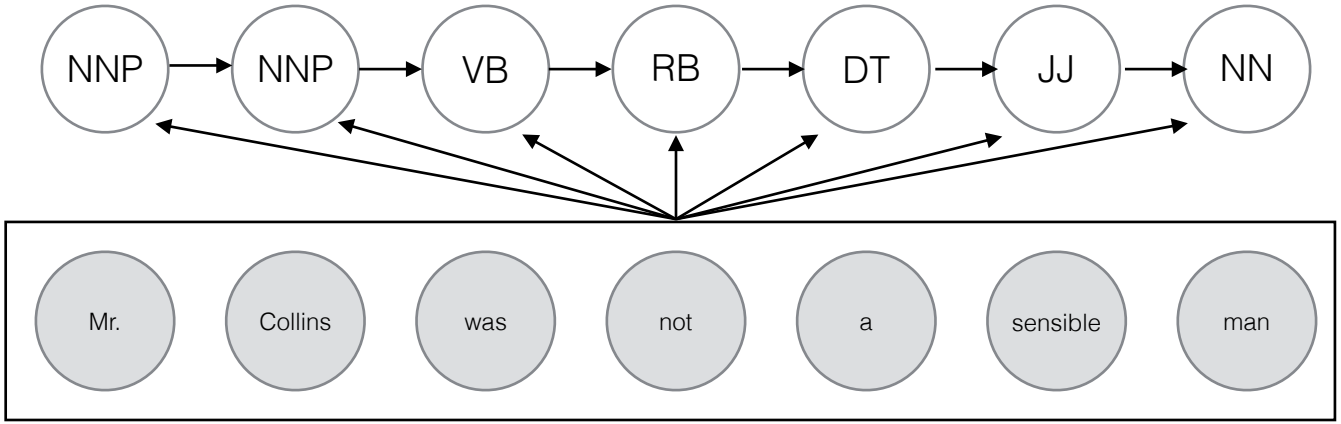
MEMM



MEMM

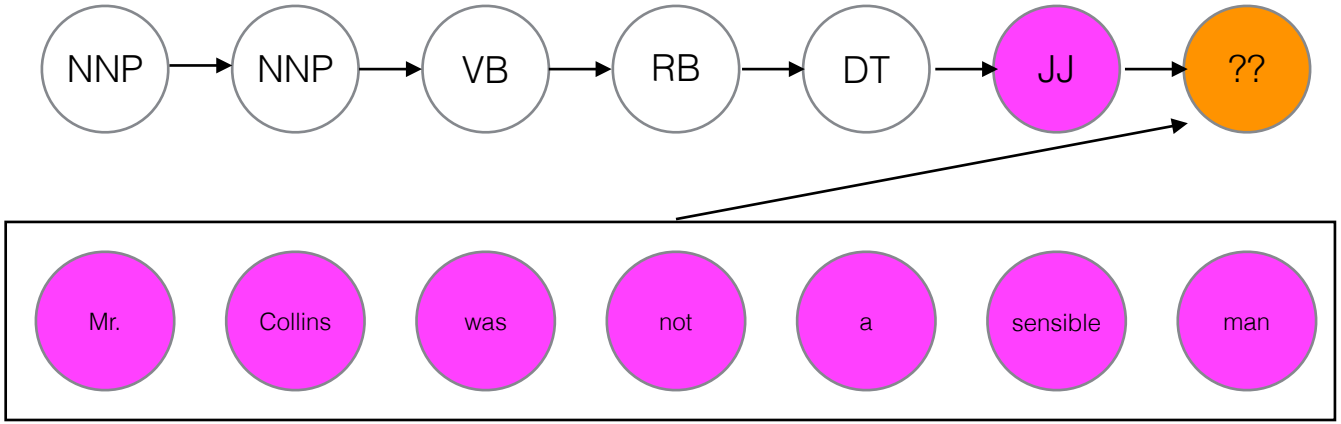


MEMM



MEMMs condition on the *entire* input

MEMM



Features

$$f(y_i, y_{i-1}; x_1, \ldots, x_n)$$

Features are scoped over the previous predicted tag and the entire observed input

feature	example
x_i = man	1
y_{i-1} = JJ	1
i=n (last word of sentence)	1
x_i ends in -ville	0

NER sequence labeling

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
base-phrase syntactic chunk label of w_i and neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
w_i is all upper case
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
presence of hyphen

Figure 17.5 Typical features for a feature-based NER system.

Gazetteers

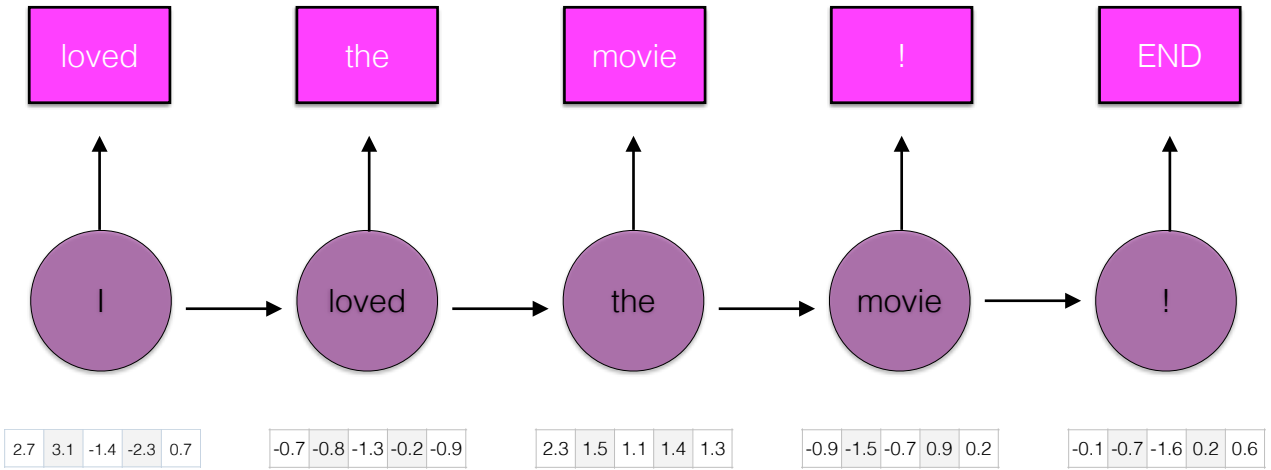
- List of place names; more generally, list of names of some typed category
- GeoNames (GEO), US SSN (PER), Getty Thesaurus of Geographic Placenames, Getty Thesaurus of Art and Architecture

Cliff
Bun Crannchа
Dromore West
Dromore
Youghal Harbour
Youghal Bay
Youghal
Eochaill
Yellow River
Yellow Furze
Woodville
Wood View
Woodtown House
Woodstown
Woodstock House
Woodsgift House
Woodrooff House
Woodpark
Woodmount
Wood Lodge
Woodlawn Station
Woodlawn
Woodlands Station
Woodhouse
Wood Hill
Woodfort
Woodford River
Woodford
Woodfield House
Woodenbridge Junction Station
Woodenbridge
Woodbrook House
Woodbrook
Woodbine Hill
Wingfield House
Windy Harbour
Windy Gap

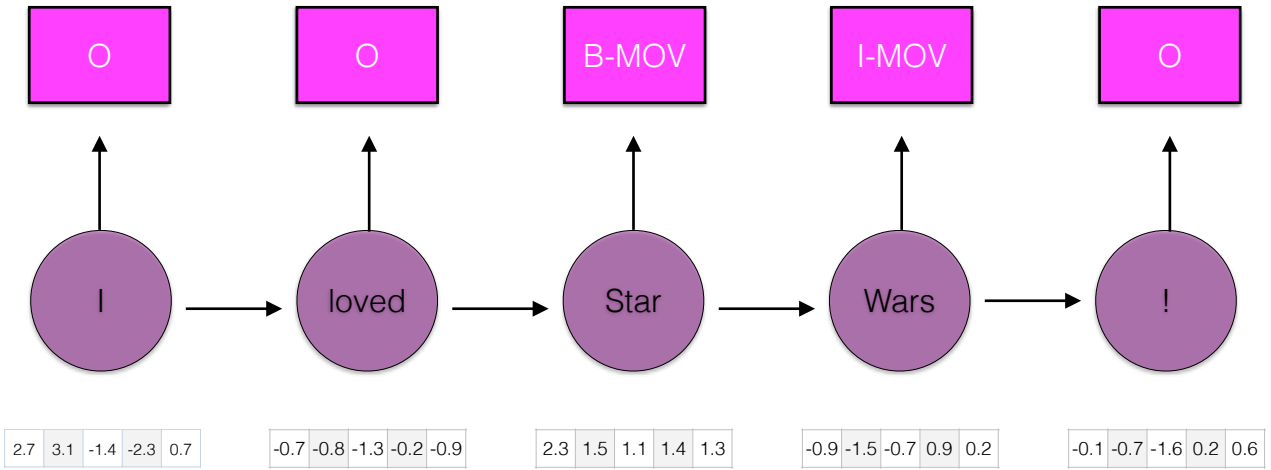
Recurrent neural network

- RNNs allow arbitarily-sized conditioning contexts and condition on the entire sequence history.

RNNs for language modeling are already performing a kind of sequence labeling: at each time step, predict the word from $\mathscr{V}$ conditioned on the context

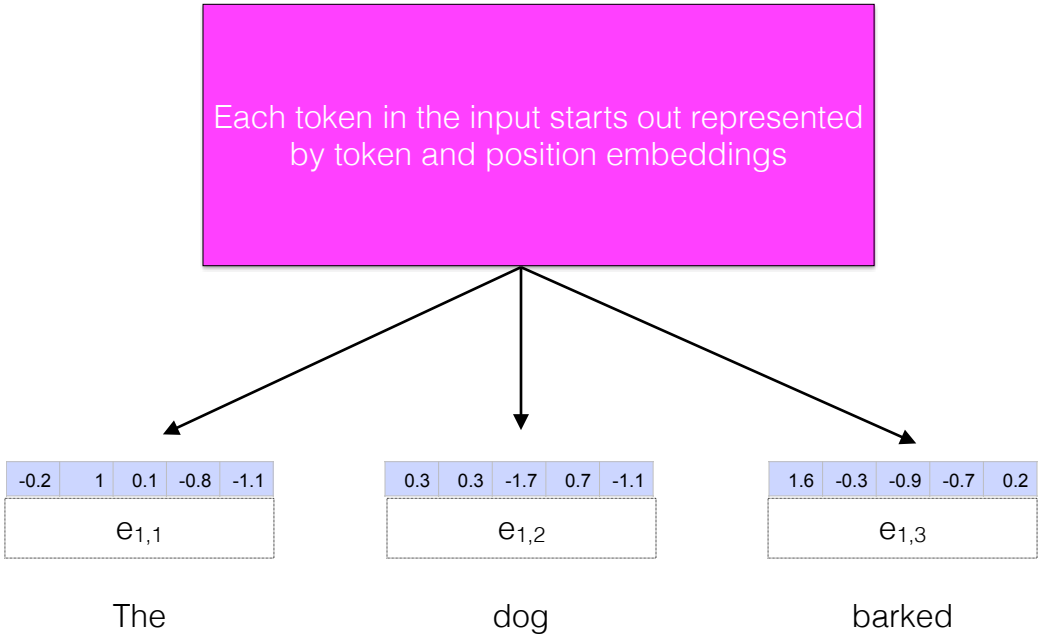


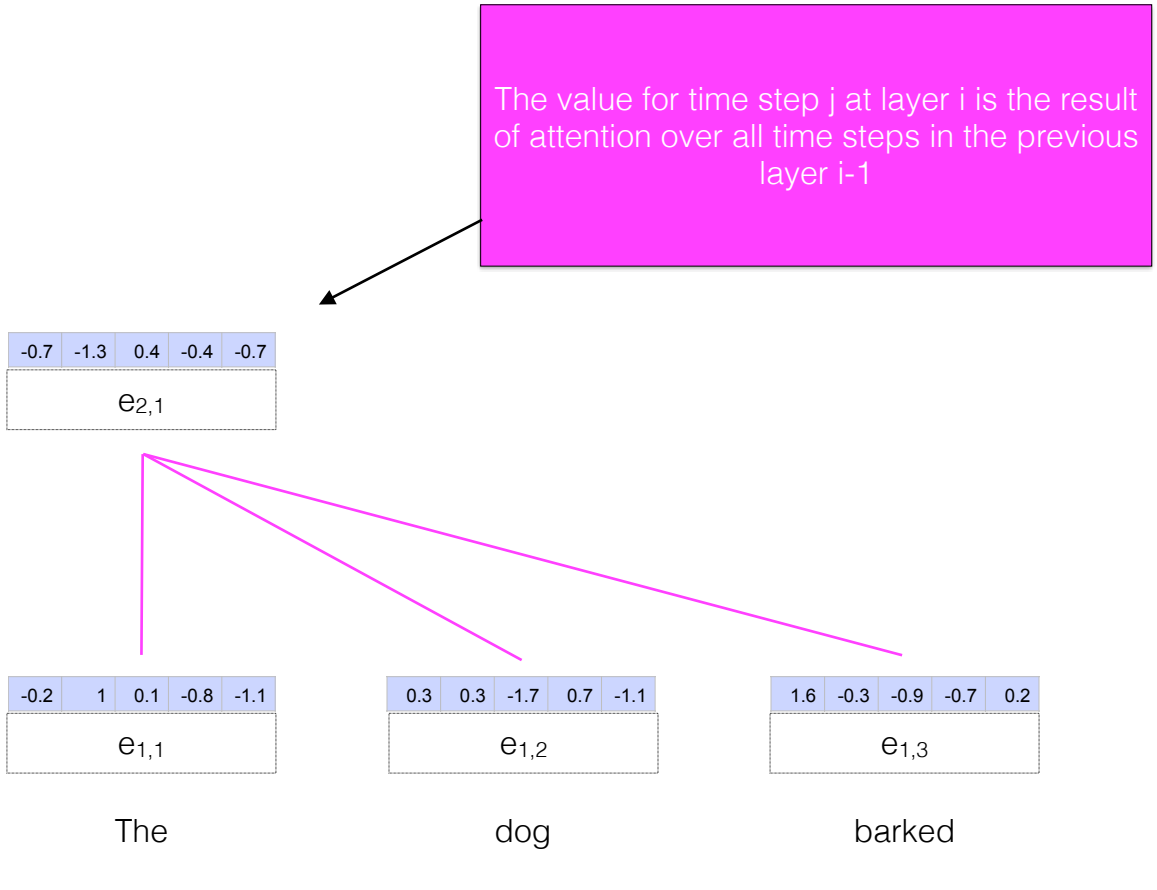
For NER, predict the tag from $\mathscr{Y}$ conditioned on the context

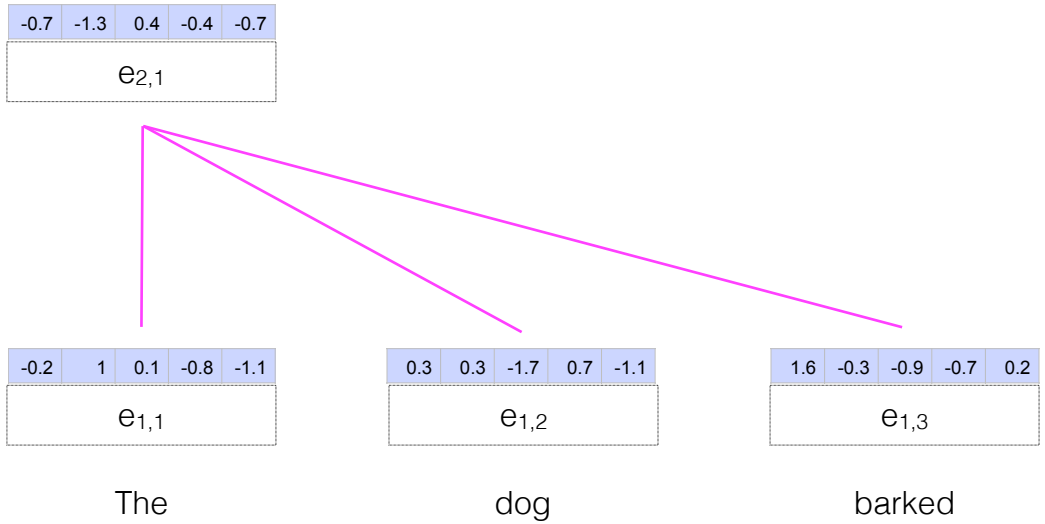


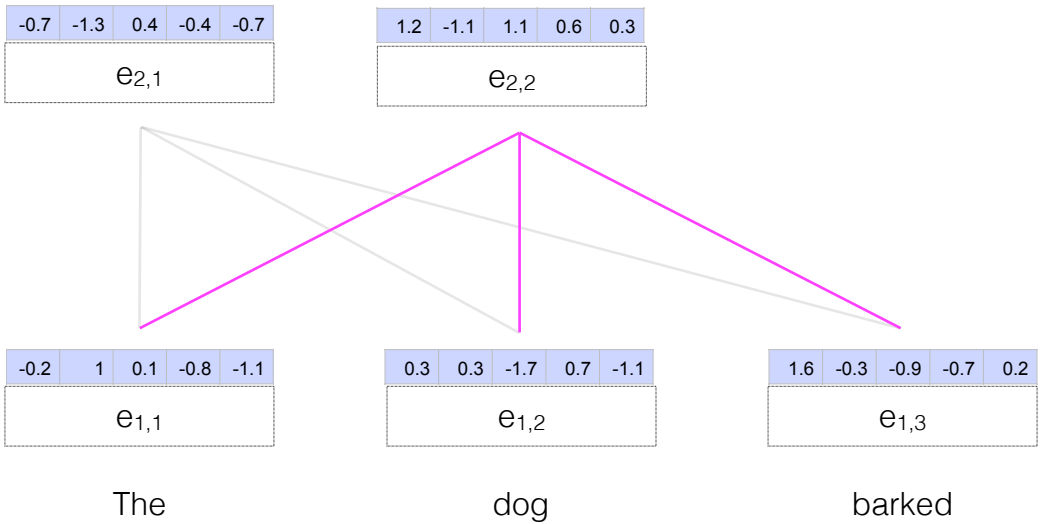
BERT

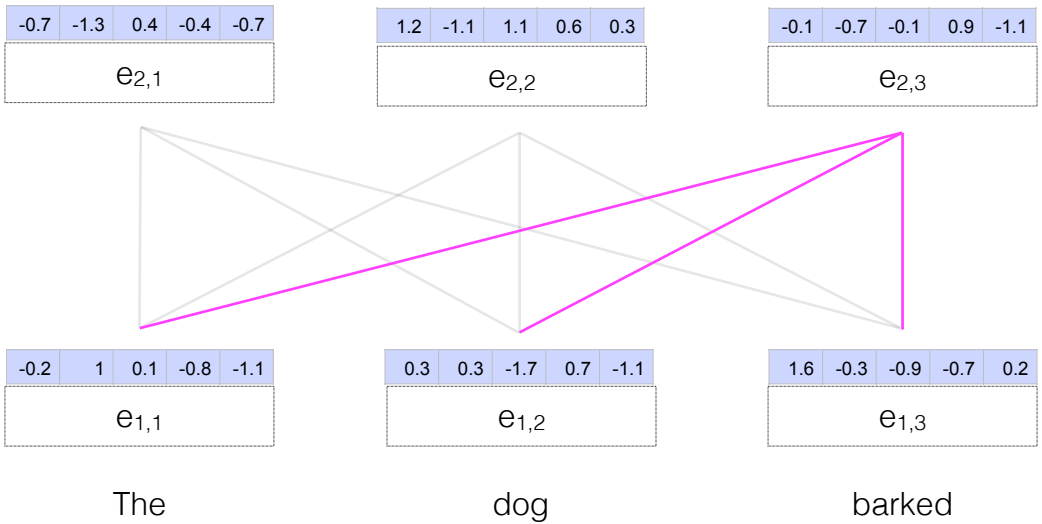
- Transformer-based model (Vaswani et al. 2017) to predict masked word using bidirectional context + next sentence prediction.
- Generates multiple layers of representations for each token sensitive to its context of use.

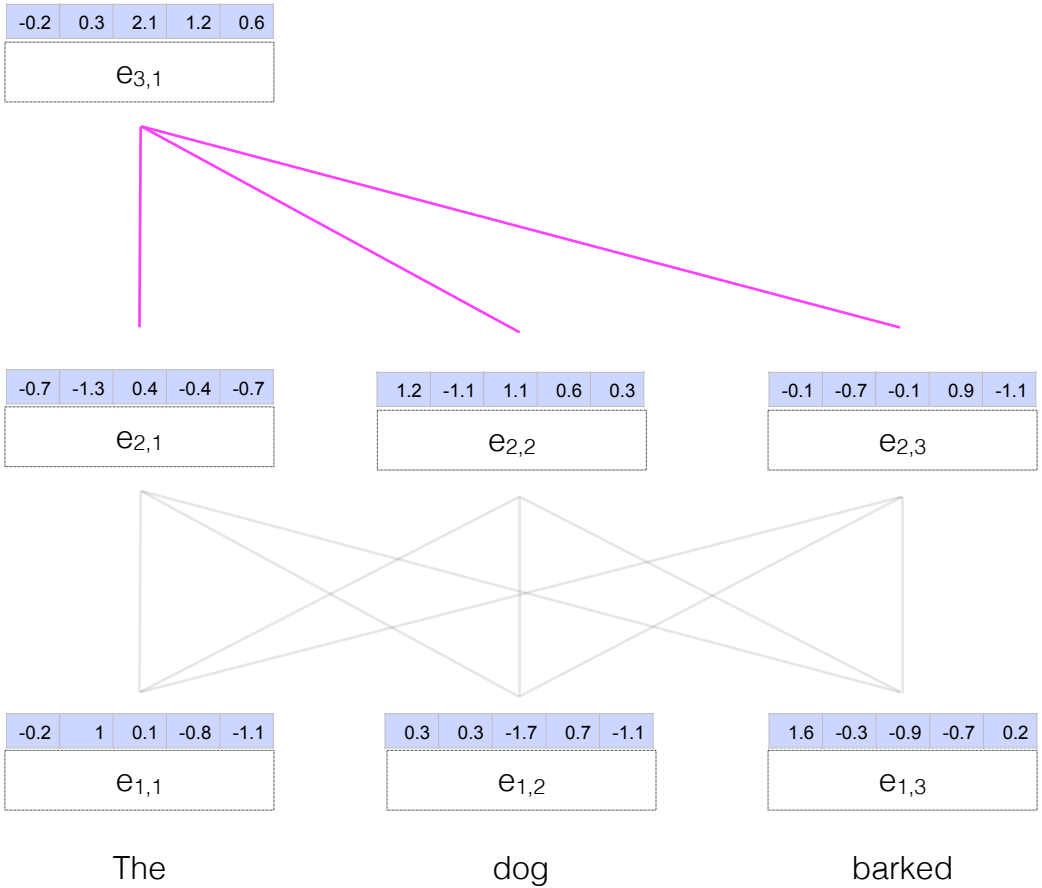

Each token in the input starts out represented by token and position embeddings
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked

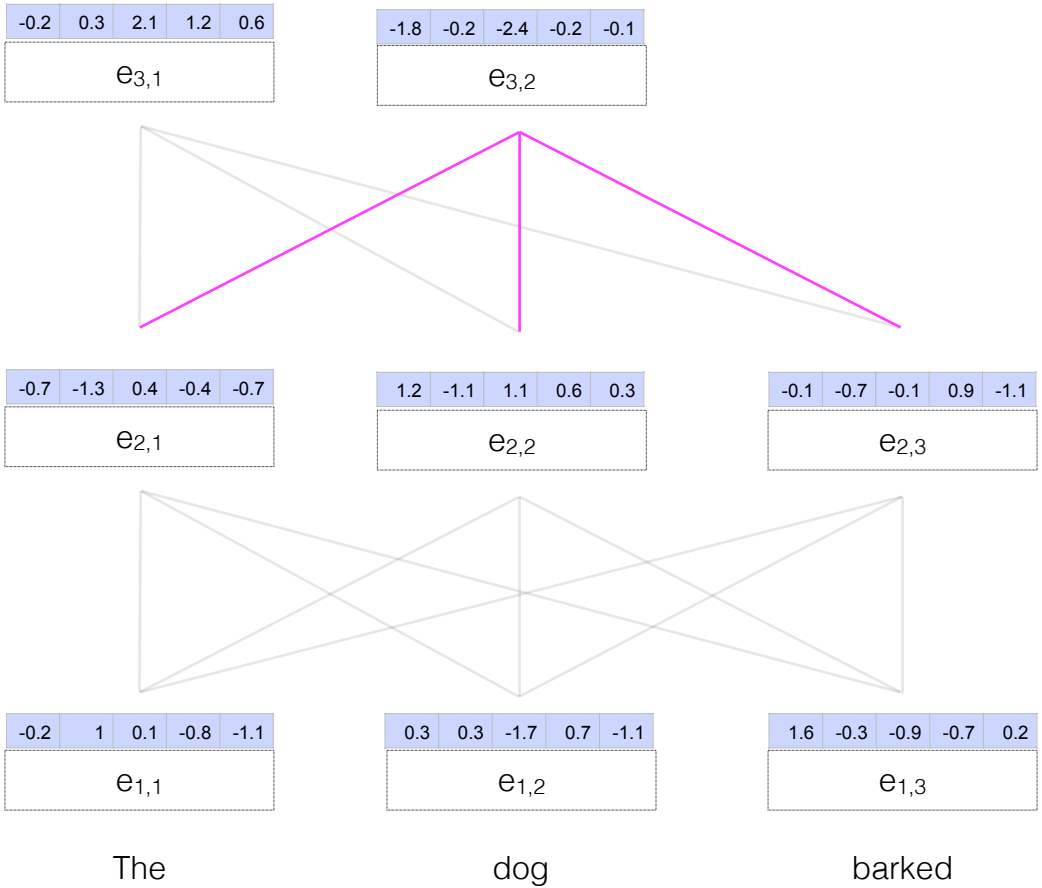

The value for time step j at layer i is the result of attention over all time steps in the previous layer i-1
-0.7 -1.3 0.4 -0.4 -0.7
e2,1
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked

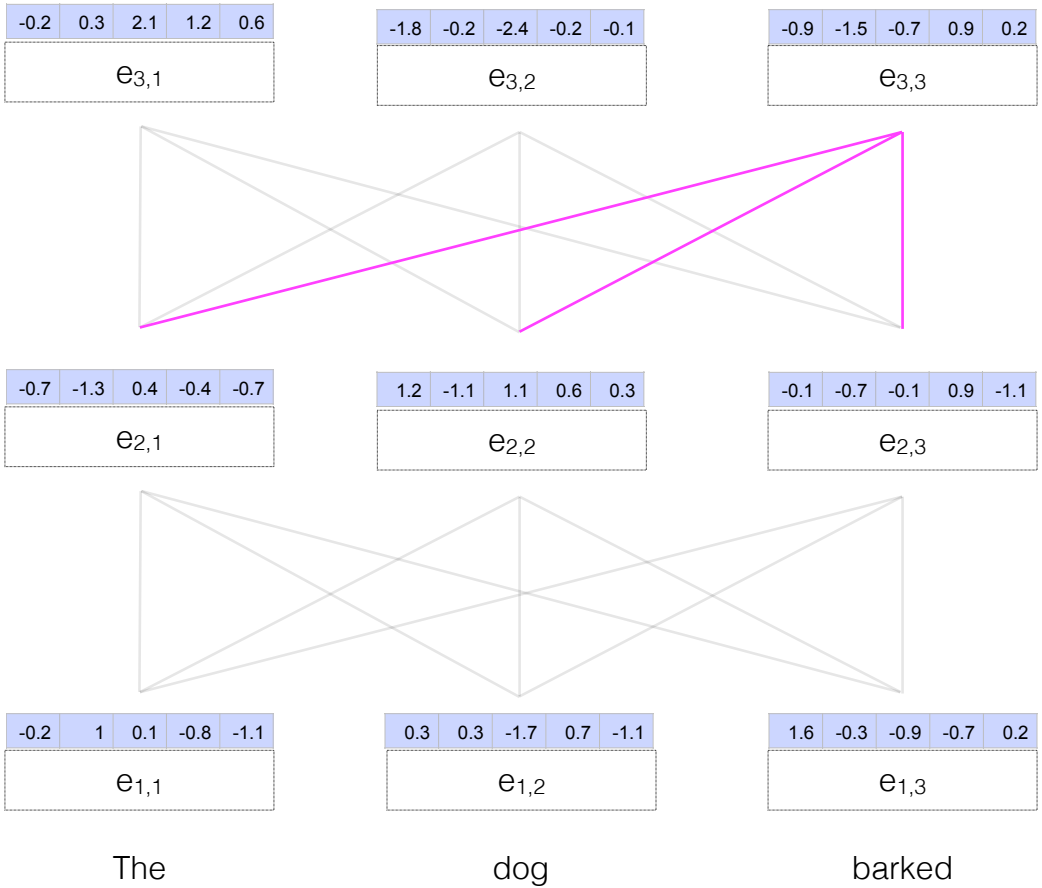

-0.7 -1.3 0.4 -0.4 -0.7
e2,1
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked


-0.7 -1.3 0.4 -0.4 -0.7
e2,1
1.2 -1.1 1.1 0.6 0.3
e2,2
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked

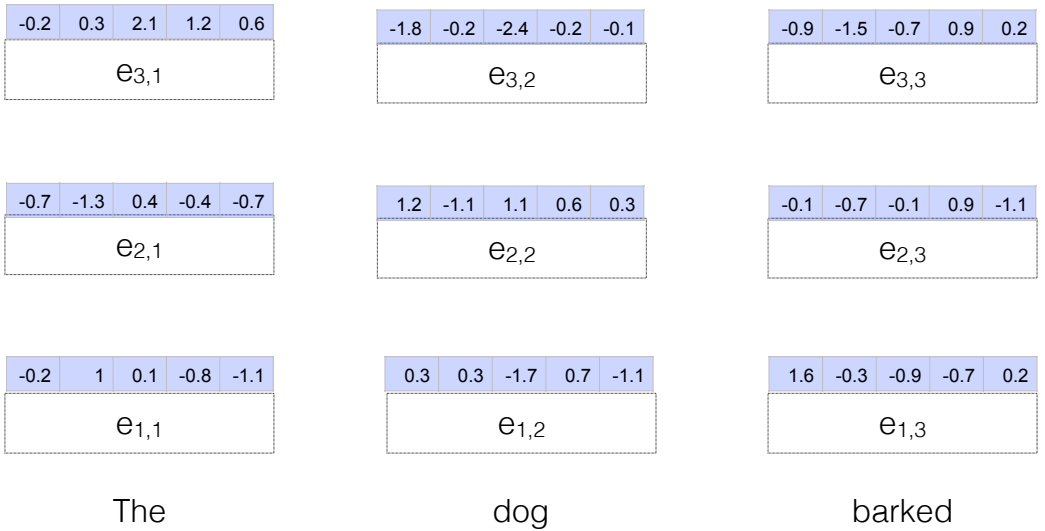

-0.7 -1.3 0.4 -0.4 -0.7
e2,1
1.2 -1.1 1.1 0.6 0.3
e2,2
-0.1 -0.7 -0.1 0.9 -1.1
e2,3
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked


-0.2 0.3 2.1 1.2 0.6
e3,1
-0.7 -1.3 0.4 -0.4 -0.7
e2,1
1.2 -1.1 1.1 0.6 0.3
e2,2
-0.1 -0.7 -0.1 0.9 -1.1
e2,3
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked


-0.2 0.3 2.1 1.2 0.6
e3,1
-1.8 -0.2 -2.4 -0.2 -0.1
e3,2
-0.7 -1.3 0.4 -0.4 -0.7
e2,1
1.2 -1.1 1.1 0.6 0.3
e2,2
-0.1 -0.7 -0.1 0.9 -1.1
e2,3
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked


-0.2 0.3 2.1 1.2 0.6
e3,1
-1.8 -0.2 -2.4 -0.2 -0.1
e3,2
-0.9 -1.5 -0.7 0.9 0.2
e3,3
-0.7 -1.3 0.4 -0.4 -0.7
e2,1
1.2 -1.1 1.1 0.6 0.3
e2,2
-0.1 -0.7 -0.1 0.9 -1.1
e2,3
-0.2 1 0.1 -0.8 -1.1
e1,1
0.3 0.3 -1.7 0.7 -1.1
e1,2
1.6 -0.3 -0.9 -0.7 0.2
e1,3
The
dog
barked

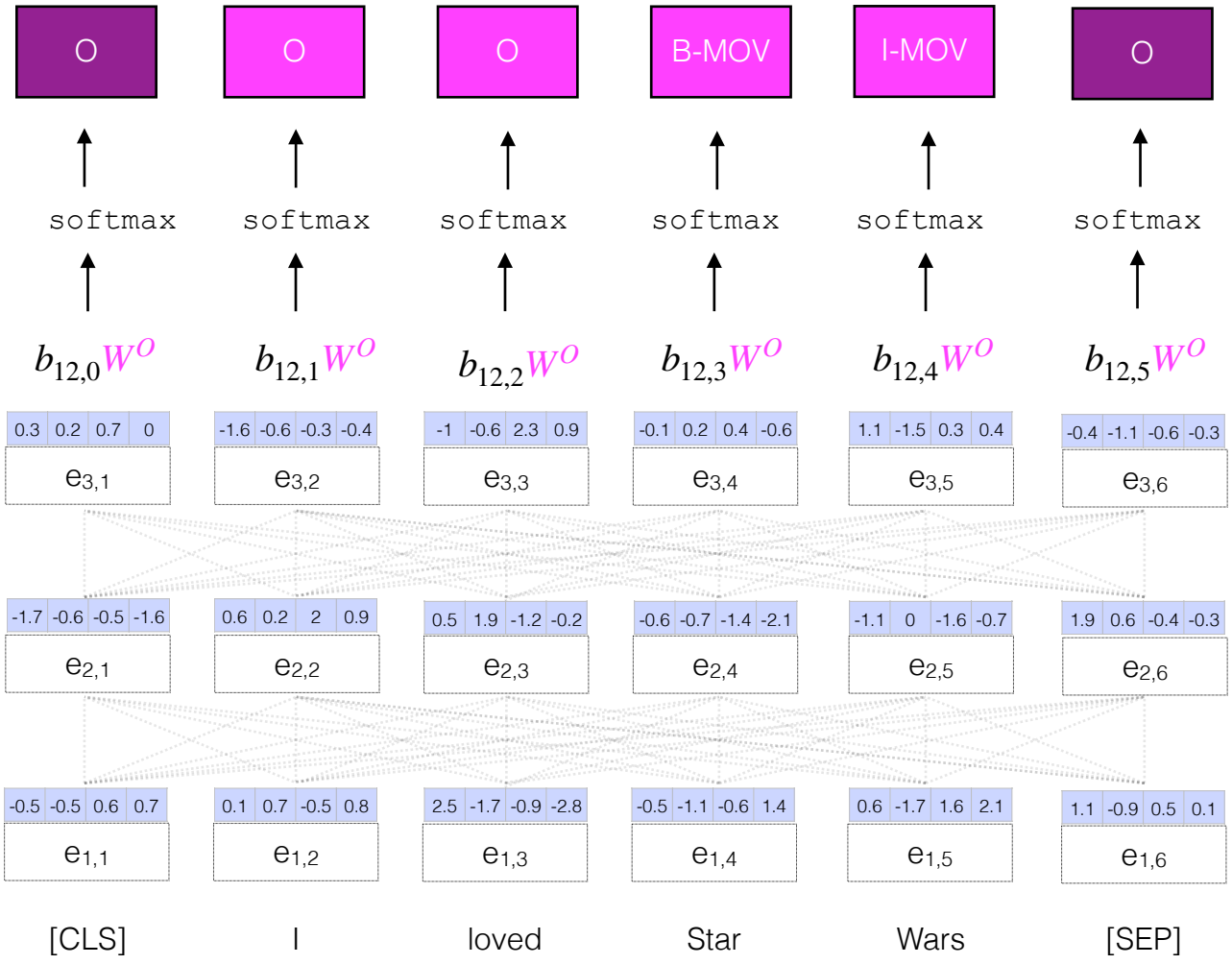
At the end of this process, we have one representation for each layer for each token



BERT

- BERT can be used not only as a language model to generate contextualized word representations, but also as a predictive model whose parameters are fine-tuned to a task.


neutral
softmax
$b_{12,0}W^O$
0.3 0.2 0.7 0
$e_{3,1}$
-1.6 -0.6 -0.3 -0.4
$e_{3,2}$
-1 -0.6 2.3 0.9
$e_{3,3}$
-0.1 0.2 0.4 -0.6
$e_{3,4}$
1.1 -1.5 0.3 0.4
$e_{3,5}$
-0.4 -1.1 -0.6 -0.3
$e_{3,6}$
-1.7 -0.6 -0.5 -1.6
$e_{2,1}$
0.6 0.2 2 0.9
$e_{2,2}$
0.5 1.9 -1.2 -0.2
$e_{2,3}$
-0.6 -0.7 -1.4 -2.1
$e_{2,4}$
-1.1 0 -1.6 -0.7
$e_{2,5}$
1.9 0.6 -0.4 -0.3
$e_{2,6}$
-0.5 -0.5 0.6 0.7
$e_{1,1}$
0.1 0.7 -0.5 0.8
$e_{1,2}$
2.5 -1.7 -0.9 -2.8
$e_{1,3}$
-0.5 -1.1 -0.6 1.4
$e_{1,4}$
0.6 -1.7 1.6 2.1
$e_{1,5}$
1.1 -0.9 0.5 0.1
$e_{1,6}$
[CLS]
The
dog
bark
#ed
[SEP]


O
O
O
B-MOV
I-MOV
O
softmax
softmax
softmax
softmax
softmax
softmax
$b_{12,0}W^O$
$b_{12,1}W^O$
$b_{12,2}W^O$
$b_{12,3}W^O$
$b_{12,4}W^O$
$b_{12,5}W^O$
0.3 0.2 0.7 0
$e_{3,1}$
-1.6 -0.6 -0.3 -0.4
$e_{3,2}$
-1 -0.6 2.3 0.9
$e_{3,3}$
-0.1 0.2 0.4 -0.6
$e_{3,4}$
1.1 -1.5 0.3 0.4
$e_{3,5}$
-0.4 -1.1 -0.6 -0.3
$e_{3,6}$
-1.7 -0.6 -0.5 -1.6
$e_{2,1}$
0.6 0.2 2 0.9
$e_{2,2}$
0.5 1.9 -1.2 -0.2
$e_{2,3}$
-0.6 -0.7 -1.4 -2.1
$e_{2,4}$
-1.1 0 -1.6 -0.7
$e_{2,5}$
1.9 0.6 -0.4 -0.3
$e_{2,6}$
-0.5 -0.5 0.6 0.7
$e_{1,1}$
0.1 0.7 -0.5 0.8
$e_{1,2}$
2.5 -1.7 -0.9 -2.8
$e_{1,3}$
-0.5 -1.1 -0.6 1.4
$e_{1,4}$
0.6 -1.7 1.6 2.1
$e_{1,5}$
1.1 -0.9 0.5 0.1
$e_{1,6}$
[CLS]
I
loved
Star
Wars
[SEP]

Evaluation

- We evaluate NER with precision/recall/F1 over typed chunks.

Evaluation

	1	2	3	4	5	6	7
	tim	cook	is	the	CEO	of	Apple
gold	B-PER	I-PER	O	O	O	O	B-ORG
system	B-PER	B-PER	O	O	B-PER	O	B-ORG

<start, end, type>

Precision	1/4
Recall	1/2

gold

<1,2,PER>
<7,7,ORG>

system

<1,1,PER>
<5,5,PER>
<7,7,ORG>
<2,2,PER>

Tools for NER

- Spacy (15 languages, blogs, news, comments)
  https://spacy.io
- Stanza (9 languages, News + Wikipedia)
  https://stanfordnlp.github.io/stanza/
- LitBank (English, literature)
  https://github.com/dbamman/litbank

LitBank

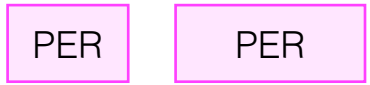
- 100 books from Project Gutenberg
- Mix of high literary style (e.g., Edith Wharton's *Age of Innocence,* James Joyce's *Ulysses)* and popular pulp (Haggard's *King Solomon's Mines*, Alger's *Ragged Dick).*
- Select first 2000 words from each text

Data

Cat	Count	Examples
PER	9,383	my mother, Jarndyce, the doctor, a fool, his companion
FAC	2,154	the house, the room, the garden, the drawing-room, the library
LOC	1,170	the sea, the river, the country, the woods, the forest
GPE	878	London, England, the town, New York, the village
VEH	197	the ship, the car, the train, the boat, the carriage
ORG	130	the army, the Order of Elks, the Church, Blodgett College

Metaphor

- Only annotate copular phrases whose types denotes an entity class.



John is a doctor

the young man was not really a poet; but surely he was a poem

Chesterton, *The Man Who Was Thursday*

Personification

- Person includes characters who engage in dialogue or have reported internal monologue, regardless of human status (includes aliens and robots as well).

> As soon as I was old enough to eat grass my mother used to go out to work in the daytime, and come back in the evening.

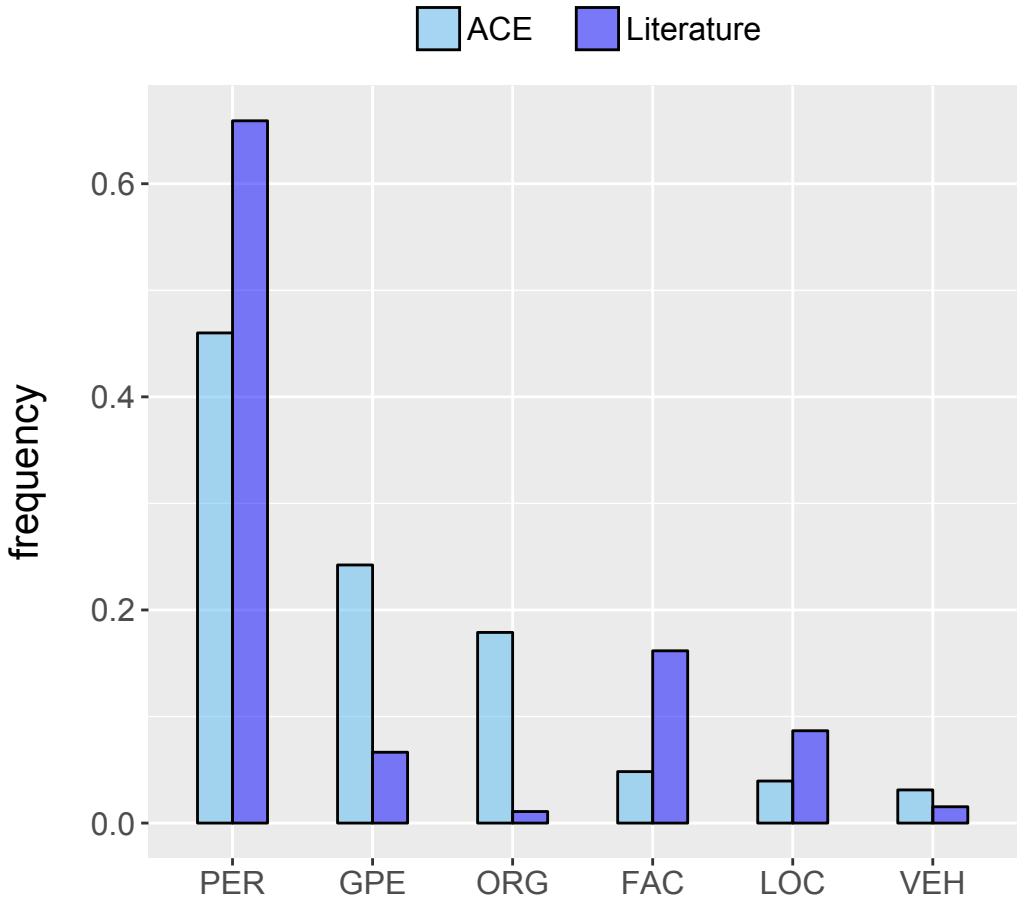
Sewell, *Black Beauty*

Prediction

How well can find these entity mentions in text as a function of the training domain?

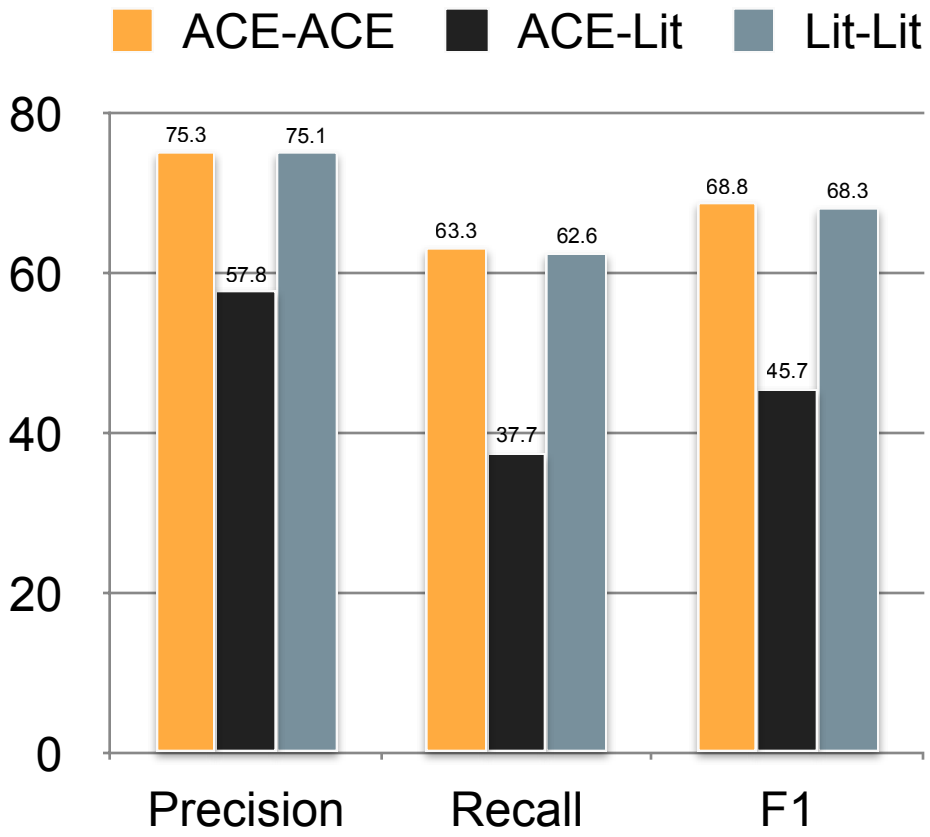
Data

- ACE (2005) data from newswire, broadcast news, broadcast conversation, weblogs



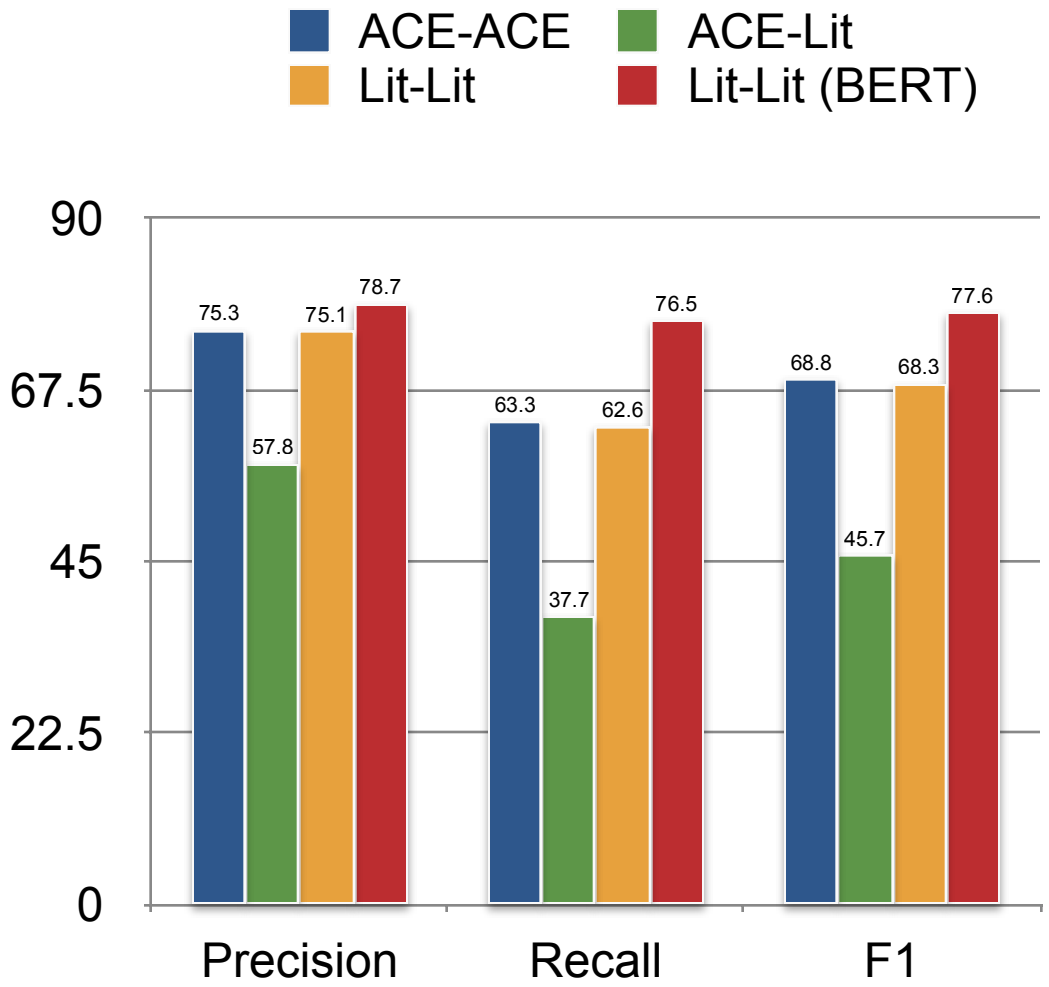
Prediction

- Ju et al. (2018): layered BiLSTM-CRF; state-of-the-art on ACE 2005.
- Evaluate performance difference when altering the training/test domain.



Prediction

- Ju et al. (2018): layered BiLSTM-CRF; state-of-the-art on ACE 2005.
- Evaluate performance difference when altering the training/test domain.
- Adding BERT contextual embeddings (Devlin et al. 2019) yields +9.3 F1 score



Toponym resolution

He thought back to their first meeting, four years earlier at a lecture hall in Cambridge, where a group of Bengali poets were giving a recital.

Jhumpa Lahiri, *Interpreter of Maladies*

He thought back to their first meeting, four years earlier at a lecture hall in Cambridge, where a group of Bengali poets were giving a recital.

Jhumpa Lahiri, *Interpreter of Maladies*

He thought back to their first meeting, four years earlier at a lecture hall in Cambridge, where a group of Bengali poets were giving a recital.

Jhumpa Lahiri, *Interpreter of Maladies*

Toponym resolution

- Given a gazetteer of place names (paired with latitude/longitude coordinates), identity the physical location of a place mentioned in text.
- Example of the more general task of entity linking (e.g., disambiguating mentions of “Michael Jordan” to the specific referent).

Methods

- Smith and Crane (2001): Each document has a geographic centroid calculated from unambiguous places; referents for each ambiguous place are then scored relative to their distance to this document centroid (and other factors).

- Speriosu and Baldridge (2013): Use non-geographic markers in text to predict geographic location (e.g., “lobster” near Portland → Portland, ME, not Portland, OR or Portland, MI).

Activity

`13.ner/ToponymResolution.ipynb`

- Run NER and toponym resolution to extract place names and map them for a selection of Wikipedia texts and *Innocents Abroad* (a travelogue by Mark Twain).

- Select your own text from Project Gutenberg (ideally one you know) and run it through that pipeline—does the spatial distribution align with what you expect? Be prepared to share your screen showing your results for the rest of the class.