

# COMP5046

# Natural Language Processing

## Lecture 9: Named Entity Recognition and Coreference Resolution

Semester 1, 2020  
School of Computer Science  
The University of Sydney, Australia

## Lecture 9: Named Entity Recognition and Coreference

1. Information Extraction
2. Named Entity Recognition (NER) and Evaluation
3. Traditional NER
4. Sequence Model for NER
5. Coreference Resolution
6. Coreference Model
7. Coreference Evaluation
8. Preview

## Information Extraction

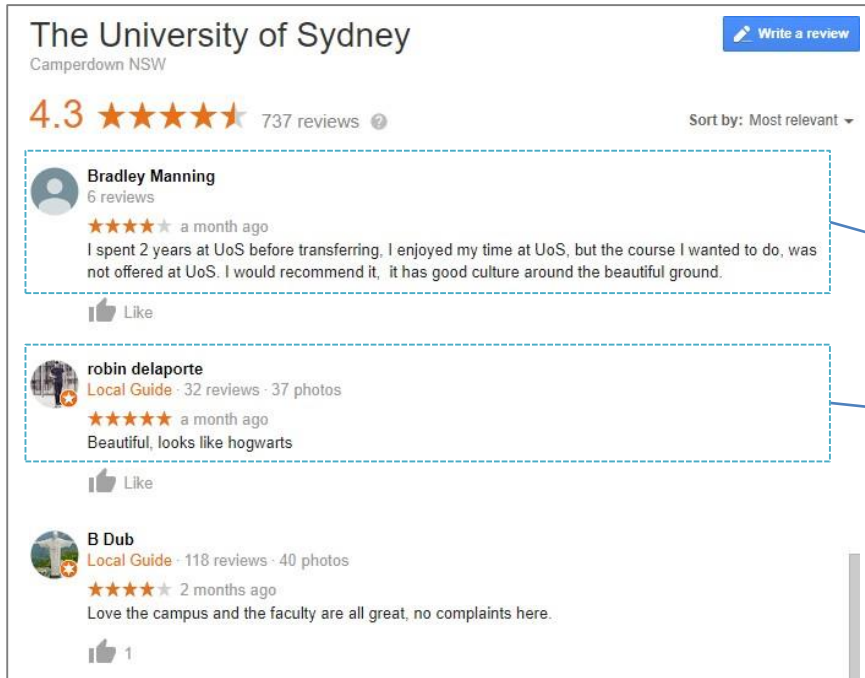
*“The task of automatically **extracting structured information** from unstructured and/or semi-structured machine-readable documents”*

Here are some questions..

- How to allow computation to be done on the unstructured data
- How to extract clear, factual information
- How to put in a semantically precise form that allows further inferences to be made by computer algorithms

## How to extract the structured clear, factual information

- Find and understand limited relevant parts of texts
- Gather information from many pieces of text
- Produce a structured representation of relevant information  
*relations (in the database sense) or a knowledge base*



The University of Sydney  
Camperdown NSW

4.3 ★★★★★ 737 reviews

Sort by: Most relevant

**Bradley Manning**  
6 reviews  
★★★★★ a month ago  
I spent 2 years at UoS before transferring. I enjoyed my time at UoS, but the course I wanted to do, was not offered at UoS. I would recommend it, it has good culture around the beautiful ground.

**robin delaporte**  
Local Guide · 32 reviews · 37 photos  
★★★★★ a month ago  
Beautiful, looks like hogwarts

**B Dub**  
Local Guide · 118 reviews · 40 photos  
★★★★★ 2 months ago  
Love the campus and the faculty are all great, no complaints here.

**“5W1H”**

*who, what, where, when, why, how*

Who:  
What:  
Where:  
When:  
How:

...

...

## How to extract the structured clear, factual information

- Find and understand limited relevant parts of texts
- Gather information from many pieces of text
- Produce a structured representation of relevant information  
*relations (in the database sense) or a knowledge base*



**Sydney**  
City in New South Wales

Sydney, capital of New South Wales and one of Australia's largest cities, is best known for its harbourfront Sydney Opera House, with a distinctive sail-like design. Massive Darling Harbour and the smaller Circular Quay port are hubs of waterside life, with the arched Harbour Bridge and esteemed Royal Botanic Garden nearby. Sydney Tower's outdoor platform, the Skywalk, offers 360-degree views of the city and suburbs.

Area: 12,368 km<sup>2</sup>

Weather: 19 °C, Wind S at 19 km/h, 38% Humidity

Local time: Monday 1:08 pm

Population: 4.452 million (2014) United Nations

State electorate(s): various (49)

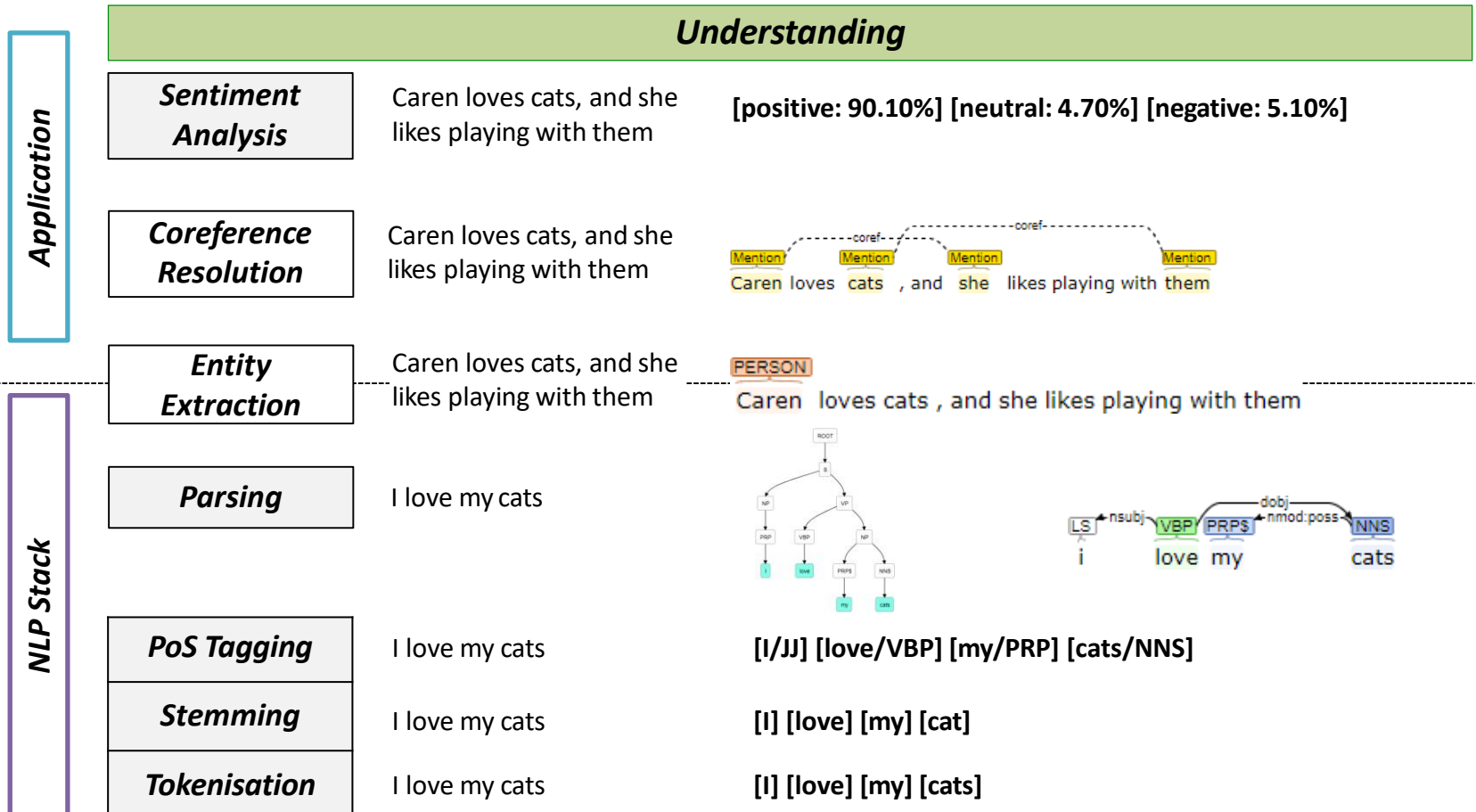


Subject	Relation	Object
Sydney	IS-A	Capital of NewSouth Wales
Sydney	IS-A	Australia's largest cities
Sydney	KNOWN FOR	Sydney Opera House
...	...	...

*Textual abstract: Summary for human*

*Structured information: Summary for machine*

## Information Extraction Pipeline with NLP



## What is Named Entity Recognition?

*“The subtask of information extraction that seeks to locate and classify named entity mentions in unstructured text into pre-defined categories such as the person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.”*

## Why recognise Named Entities?

- Named entities can be indexed, linked off, etc.
- Sentiment can be attributed to companies or products
- A lot of relations are associations between named entities
- For question answering, answers are often named entities.

# Named Entity Recognition (NER)

## How to recognize Named Entities?

**Identify** and **classify** names in text

- The University of Sydney (informally **USYD**, **Sydney**, **Sydney Uni**) is an **Australian** public research university in **Sydney**, **Australia**. Founded in **1850**, it was **Australia**'s first university and is regarded as one of the world's leading universities. (Wikipedia, University of Sydney)*

Different types of named entity classes

Type	Classes
3 class	Location, Person, Organization
4 class	Location, Person, Organization, Misc
7 class	Location, Person, Organization, Money, Percent, Date, Time

*\*classes can be different based on annotated dataset*



# Named Entity Recognition (NER)

## How to recognize Named Entities?

**Identify** and **classify** names in text

**Upenn CogComp-NLP** <http://macniece.seas.upenn.edu:4004/>

1	The University of Sydney ( informally USYD , Sydney , Sydney Uni ) is an Australian public research university in Sydney , Australia .
2	Founded in 1850 , it was Australia 's first university and is regarded as one of the world 's leading universities .

**Stanford CoreNLP 3.9.2** <http://nlp.stanford.edu:8080/corenlp/process>

1	The University of Sydney (informally USYD, Sydney, Sydney Uni) is an Australian public research university in Sydney, Australia.
2	Founded in 1850, it was Australia's first university and is regarded as one of the world's leading universities.

## 2

# Named Entity Recognition (NER)

How to evaluate the NER performance?

The goal: *predicting entities in a text*

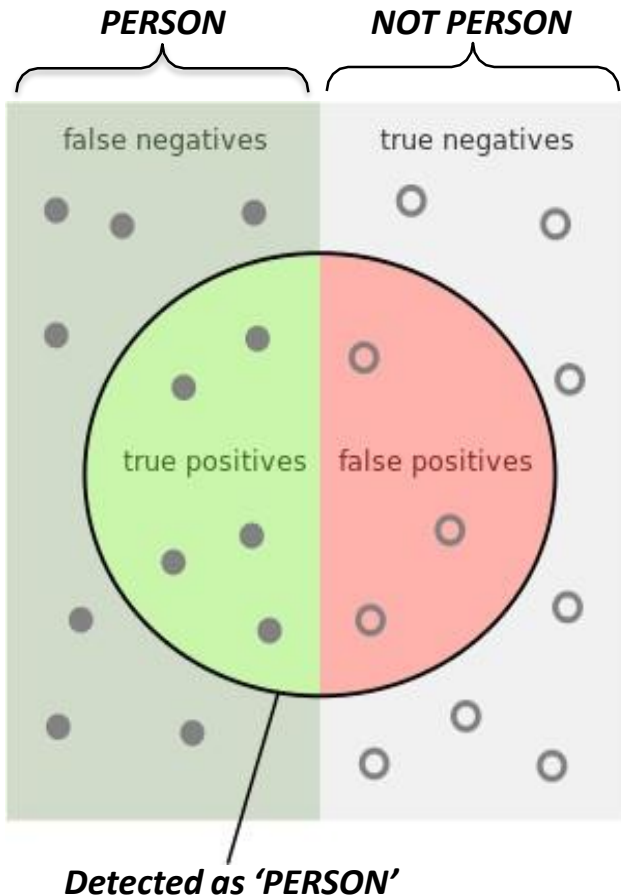
\*Standard evaluation is per entity, not per token

Caren Soyeon Han is working at Google at Sydney, Australia

gold	PER	PER	PER	O	O	O	ORG	O	LOC	LOC
predicted	O	O	O	O	O	O	ORG	O	LOC	LOC

# Named Entity Recognition (NER)

How to evaluate the NER performance? Precision and recall

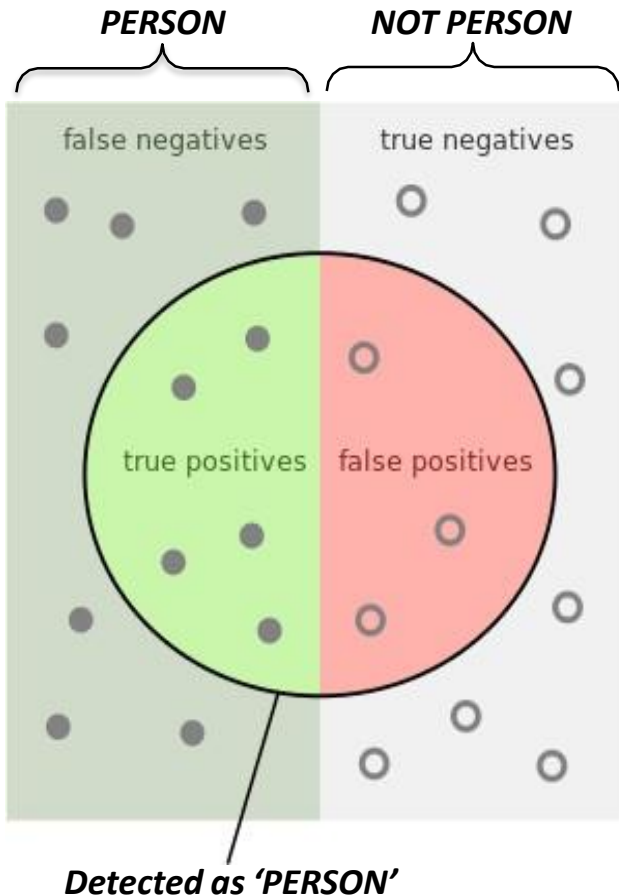


$$\text{Precision} = \frac{\text{Detected as 'PERSON' correctly}}{\text{Total number of detected as 'PERSON'}}$$

$$\text{Recall} = \frac{\text{Detected as 'PERSON' correctly}}{\text{Total number of actual 'PERSON' entities}}$$

# Named Entity Recognition (NER)

How to evaluate the NER performance? Precision and recall



**True positives:** The 'PERSON's that the model detected as 'PERSON'

**False positives:** The NOT 'PERSON's that the model detected as 'PERSON'

**False negatives:** The 'PERSON's that the model detected as NOT 'PERSON'

**True negatives:** The 'NOT PERSON's that the model detected as NOT 'PERSON'

## 2

# Named Entity Recognition (NER)

How to evaluate the NER performance?

The goal: *predicting entities in a text*

\*Standard evaluation is per entity, not per token

Caren Soyeon Han is working at Google at Sydney, Australia

<i>gold</i>	PER	PER	PER	○	○	○	ORG	○	LOC	LOC
<i>predicted</i>	○	○	○	○	○	○	ORG	○	LOC	LOC

	correct	not correct
selected	True Positive (TP)	False Positive (FP)
not selected	False Negative (FN)	True Negative (TN)

## 2

# Named Entity Recognition (NER)

How to evaluate the NER performance?

The goal: *predicting entities in a text*

\*Standard evaluation is per entity, not per token

	Caren Soyeon Han is working at Google at Sydney, Australia									
<i>gold</i>	PER	PER	PER	O	O	O	ORG	O	LOC	LOC
<i>predicted</i>	O	O	O	O	O	O	ORG	O	LOC	LOC

*Precision and Recall are straightforward for text categorization or web search, where there is only one grain size (documents)*

## Named Entity Recognition (NER)

### Quick Exercise: F measure Calculation

Let's calculate Precision, Recall, and F-measure together!

$$P = ??$$

$$R = ??$$

$$F_1 = ??$$

$$F1 = 2 * \frac{P * R}{P + R}$$

	correct	not correct
selected	2 (TP)	0 (FP)
not selected	1 (FN)	0 (TN)

## Data for learning named entity

- Training counts joint frequencies in a corpus
- The more training data, the better
- Annotated corpora are small and expensive

Corpora	Source	Size	Class Type
muc-7	New York Times	164k tokens	per, org, loc, dates, times, money, percent <a href="https://aclweb.org/aclwiki/MUC-7_(State_of_the_art)">https://aclweb.org/aclwiki/MUC-7_(State_of_the_art)</a>
conll-03	Reuters	301k	per, org, loc, misc
bbn	Wall Street Journal	1174k	<a href="https://catalog.ldc.upenn.edu/docs/LDC2005T33/BBN-Types-Subtypes.html">https://catalog.ldc.upenn.edu/docs/LDC2005T33/BBN-Types-Subtypes.html</a>



## Named Entity Recognition (NER)

### Data for learning named entity

- Models trained on one corpus perform poorly on others

<i>train</i>	<i>F-score</i>		
	<i>muc</i>	<i>conll</i>	<i>bbn</i>
<i>muc</i>	82.3	54.9	69.3
<i>conll</i>	69.9	86.9	60.2
<i>bnn</i>	80.2	58.0	88.0

## CoNLL 2003 NER dataset

- Performance measure:  $F = 2 * \text{Precision} * \text{Recall} / (\text{Recall} + \text{Precision})$

RANK	METHOD	F1	EXTRA TRAINING DATA	PAPER TITLE	YEAR	PAPER	CODE
1	CNN Large + fine-tune	93.5	✓	<a href="#">Cloze-driven Pretraining of Self-attention Networks</a>	2019		
2	GCDT + BERT-L	93.47	✓	<a href="#">GCDT: A Global Context Enhanced Deep Transition Architecture for Sequence Labeling</a>	2019		
3	I-DARTS + Flair	93.47	✓	<a href="#">Improved Differentiable Architecture Search for Language Modeling and Named Entity Recognition</a>	2019		
4	LSTM-CRF+ELMo+BERT+Flair	93.38	✓	<a href="#">Neural Architectures for Nested NER through Linearization</a>	2019		
5	Hierarchical + BERT	93.37	✗	<a href="#">Hierarchical Contextualized Representation for Named Entity Recognition</a>	2019		
6	BERT-MRC+DSC	93.33	✓	<a href="#">Dice Loss for Data-imbalanced NLP Tasks</a>	2019		
7	Flair embeddings + Pooling	93.18	✓	<a href="#">Pooled Contextualized Embeddings for Named Entity Recognition</a>	2019		
8	Flair embeddings	93.09	✓	<a href="#">Contextual String Embeddings for Sequence Labeling</a>	2018		
9	BERT-MRC	93.04	✓	<a href="#">A Unified MRC Framework for Named Entity Recognition</a>	2019		
10	BERT Large	92.8	✓	<a href="#">BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding</a>	2018		

## Datasets for NER in English

The following table shows the list of datasets for English entity recognition.

Dataset	Domain	License	Reference
CONLL 2003	News	DUA	Sang and Meulder, 2003
NIST-IEER	News	None	NIST 1999 IE-ER
MUC-6	News	LDC	Grishman and Sundheim, 1996
OntoNotes 5	Various	LDC	Weischedel et al., 2013
BBN	Various	LDC	Weischedel and Brunstein, 2005
GMB-1.0.0	Various	None	Bos et al., 2017
GUM-3.1.0	Wiki	Several (*2)	Zeldes, 2016
wikigold	Wikipedia	CC-BY 4.0	Balasuriya et al., 2009
Ritter	Twitter	None	Ritter et al., 2011
BTC	Twitter	CC-BY 4.0	Derczynski et al., 2016
WNUT17	Social media	CC-BY 4.0	Derczynski et al., 2017
i2b2-2006	Medical	DUA	Uzuner et al., 2007
i2b2-2014	Medical	DUA	Stubbs et al., 2015
CADEC	Medical	CSIRO	Karimi et al., 2015
AnEM	Anatomical	CC-BY-SA 3.0	Ohta et al., 2012
MITRestaurant	Queries	None	Liu et al., 2013a
MITMovie	Queries	None	Liu et al., 2013b
MalwareTextDB	Malware	None	Lim et al., 2017
re3d	Defense	Several (*1)	DSTL, 2017
SEC-filings	Finance	CC-BY 3.0	Alvarado et al., 2015
Assembly	Robotics	X	Costa et al., 2017

**DUA:** Data Use Agreement

**LDC:** Linguistic Data Consortium

**CC-BY 4.0:** Creative Commons Attribution 4.0

# Traditional NER

## Three standard approaches to NER

- Rule-based NER
  - Classifier-based NER
  - Sequence Model for NER
- ] Traditional Approaches

## Rule-based NER

- Entity references have internal and external language cues  
Mr. [per Scott Morrison] flew to [loc Beijing]
- Can recognise names using lists (or gazetteers):
  - Personal titles: Mr, Miss, Dr, President
  - Given names: Scott, David, James
  - Corporate suffixes: & Co., Corp., Ltd.
  - Organisations: Microsoft, IBM, Telstra
- and rules:
  - personal title X  $\Rightarrow$  per
  - X, location  $\Rightarrow$  loc or org
  - travel verb to X  $\Rightarrow$  loc
- Effectively regular expressions, PoS Tagger

## Rule-based NER

- Determining which person holds what office in what organization
  - [person] , [office] of [org]
    - Michael Spence, the vice-chancellor and principal of the University of Sydney
  - [org] (named, appointed, etc.) [person] Prep [office]
    - WHO appointed Tedros Adhanom as Director-General
- Determining where an organization is located
  - [org] in [loc]
    - Google headquarters in California
  - [org] [loc] (division, branch, headquarters, etc.)
    - Google London headquarters

## Statistical approaches are more portable

- Learn NER from annotated text
  - weights ( $\approx$  rules) calculated from the corpus
  - same machine learner, different language or domain
- Token-by-token classification (with any machine learning)
- Each token may be:
  - not part of an entity (tag o)
  - beginning an entity (tag b-per, b-org, etc.)
  - continuing an entity (tag i-per, i-org, etc.)
- What about N-gram model?

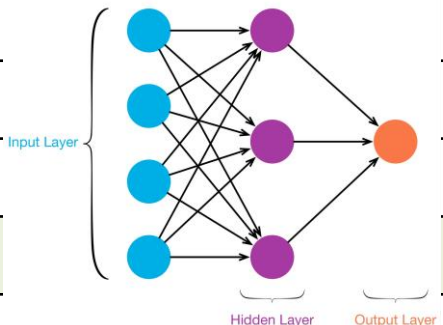
## Various features for statistical NER

Unigram	Mr.	Scott	Morrison	flew	to	Beijing
Lowercase unigram	mr.	scott	morrison	flew	to	beijing
POS tag	nnp	nnp	nnp	vbd	to	nnp
length	3	5	4	4	2	7
In first-name gazetteer	no	yes	no	no	no	no
In location gazetteer	no	no	no	no	no	yes
3-letter suffix	Mr.	ott	son	lew	-	ing
2-letter suffix	r.	tt	on	ew	to	ng
1-letter suffix	.	t	n	w	o	g
Tag predictions	O	B-per	I-per	O	O	B-loc



## Various features for statistical NER

Unigram	Mr.	Scott	Morrison	flew	to	Beijing
Lowercase unigram	mr.	scott	morrison	flew	to	beijing
POS tag	nnp	nnp	nnp	vbd	to	nnp
length	3	5	4	4	2	7
In first-name gazetteer	no	yes	no	no	no	no
In location gazetteer	no	no	no	no	no	yes
3-letter suffix	Mr.			lew	-	ing
2-letter suffix	r.			ew	to	ng
1-letter suffix	.			w	o	g
Tag predictions	O			O	O	B-loc



Input Layer      Hidden Layer      Output Layer

Mr. Scott Morrison lives in Sydney

---->

**Predictive Model**

---->

O B-PER I-PER O O B-LOC

## Traditional NER Approaches - Pros and Cons

### *Rule-based approaches*

- Can be high-performing and efficient
- Require experts to make rules
- Rely heavily on gazetteers that are always incomplete
- Are not robust to new domains and languages

### *Statistical approaches*

- Require (expert-)annotated training data
- May identify unforeseen patterns
- Can still make use of gazetteers
- Are robust for experimentation with new features
- Are largely portable to new languages and domains

## Sequence Model (N to N)

ADV VERB DET NOUN NOUN

*Output: Part of Speech*

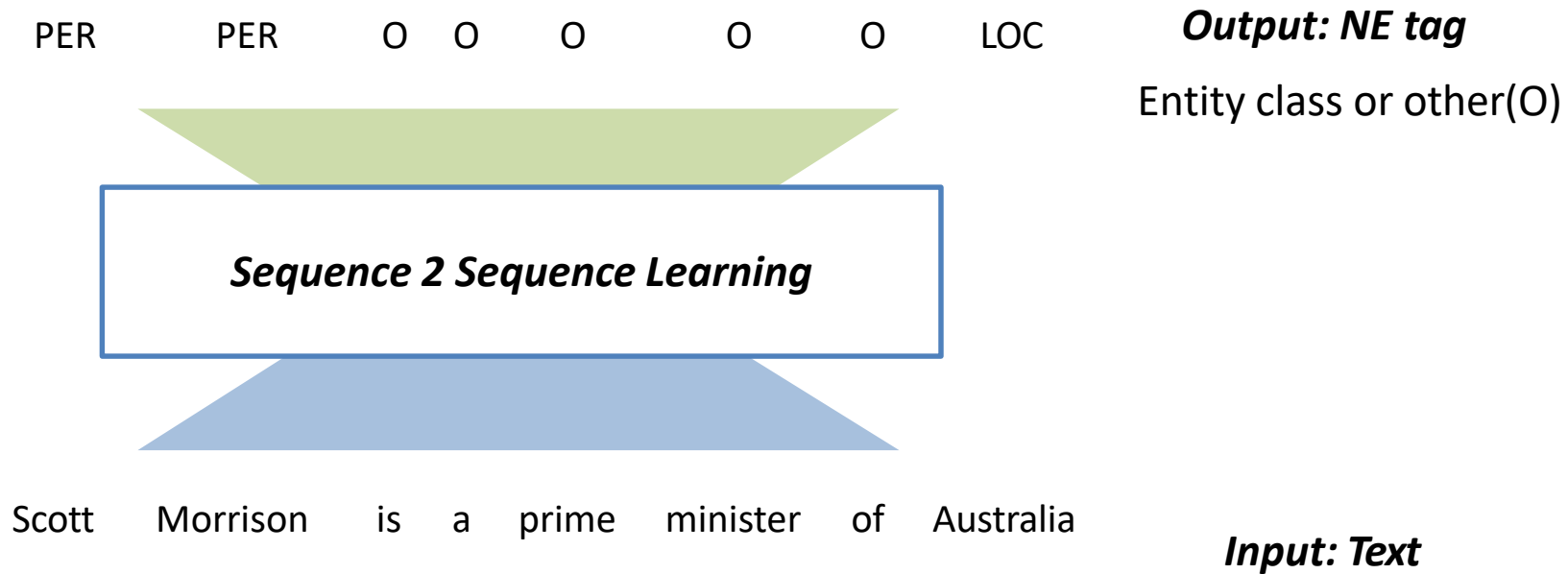


*Sequence 2 Sequence Learning*

How is the weather today

*Input: Text*

## Sequence Model



## Encoding classes for sequence labeling

PER      PER      O   O      O      O      O      LOC

**Output: NE tag**

Entity class or other(O)



*Sequence 2 Sequence Learning*

Scott    Morrison    is    a    prime    minister    of    Australia

**Input: Text**

*The IOB (short for inside, outside, beginning) is a common tagging format*

- I- prefix before a tag indicates that the tag is inside a chunk.
- B- prefix before a tag indicates that the tag is the beginning of a chunk.
- An O tag indicates that a token belongs to no chunk (outside).

## Encoding classes for sequence labeling

The IO and IOB (inside, outside, beginning) is a common tagging format

	Josiah	tells	Caren	John	Smith	is	a	student	
IO encoding	PER	O	PER	PER	PER	O	O	O	<b>n+1</b>
IOB encoding	B-PER	O	B-PER	B-PER	I-PER	O	O	O	<b>2n+1</b>
			<i>even</i>	B-PER	I-PER	I-PER			

### *IO encoding vs IOB encoding*

- *Computation Time?*
- *Efficiency?*

## Features for sequence labeling

### Words

- Current word (essentially like a learned dictionary)
- Previous/next word (context)

### Other kinds of inferred linguistic classification

- Part-of-speech tags

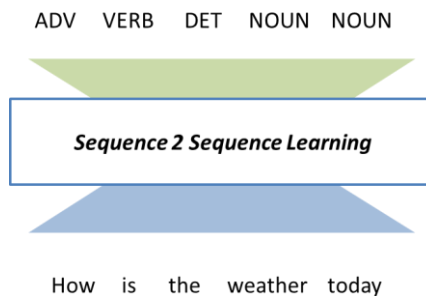
### Label context

- Previous (and perhaps next) label

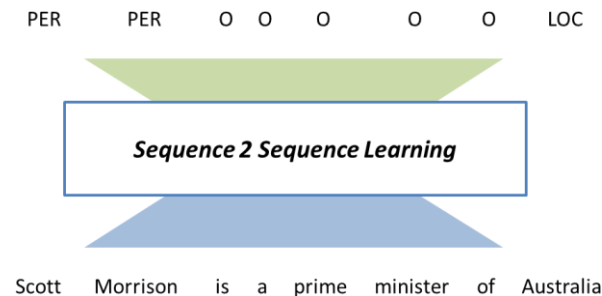
## N to N Sequence model

- There are different NLP tasks that used N to N sequence model

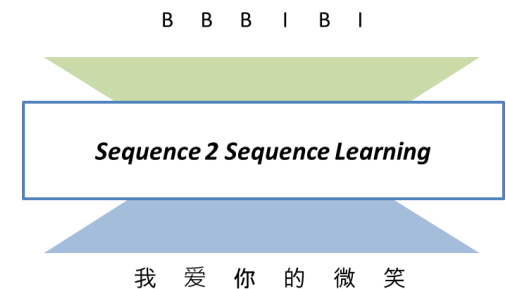
### *POS tagging*



### *Named Entity Recognition*

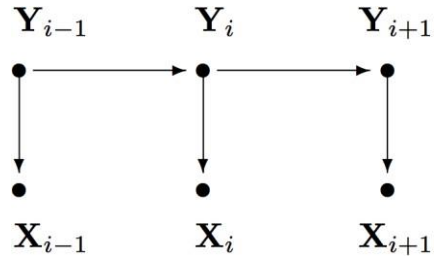


### *Word Segmentation*

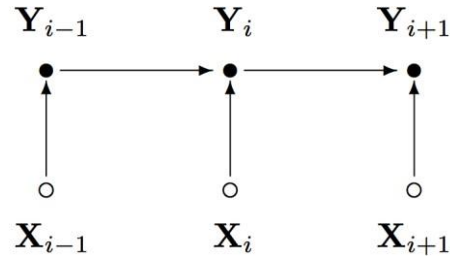




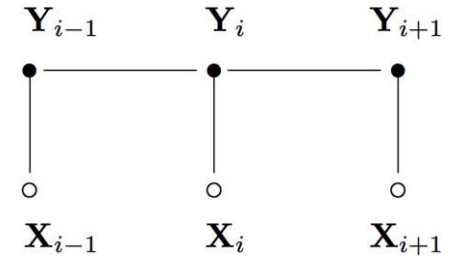
## Sequence Model (MEMM, CRF)



HMM



MEMM



CRF

## Sequence Inference for NER

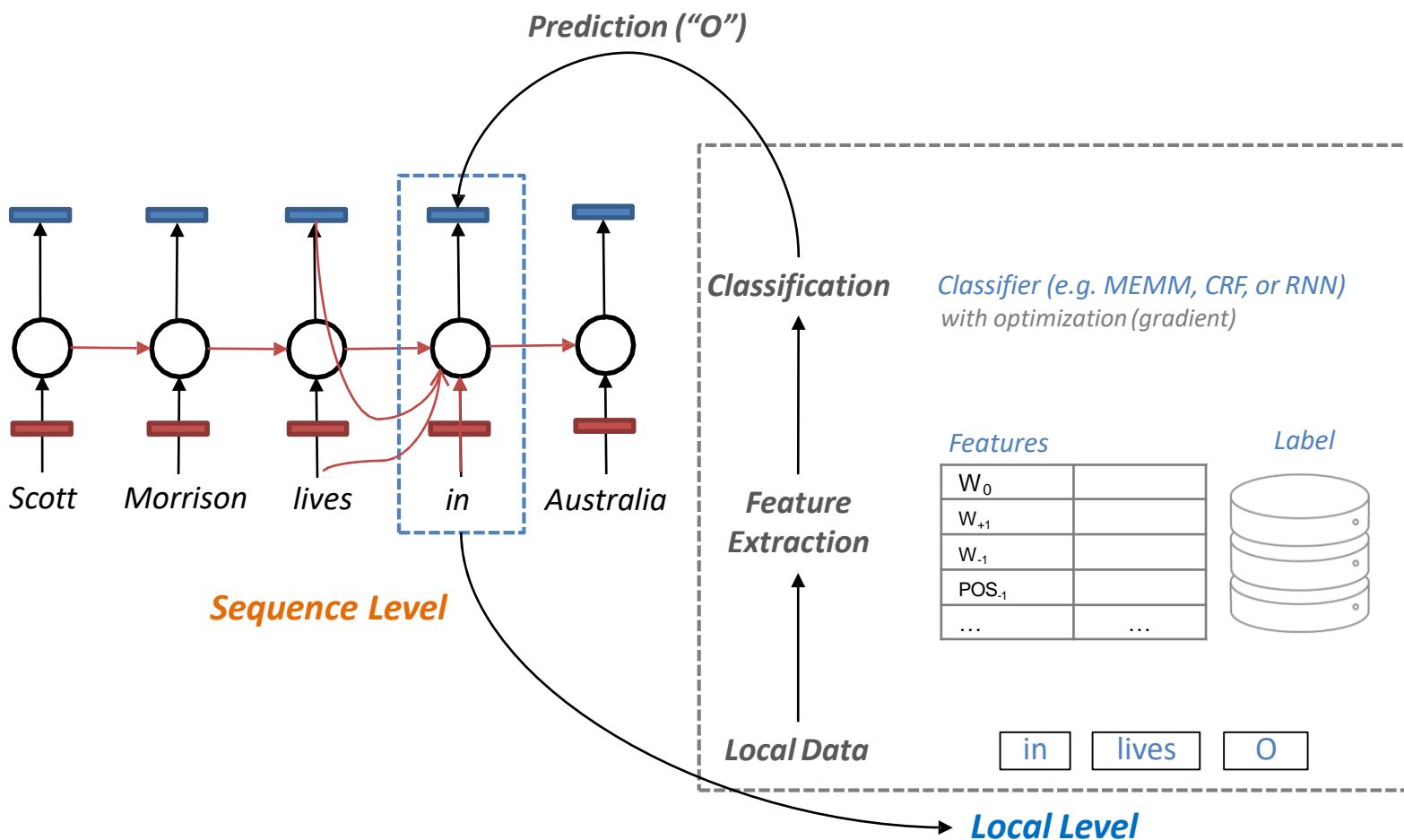
- For a Maximum Entropy Markov Model (MEMM), the classifier makes a single decision at a time, conditioned on evidence from observations and previous decisions

<b>-3</b>	<b>-2</b>	<b>-1</b>	<b>0</b>	<b>+1</b>
Scott	Morrison	lives	in	Australia
<b>NN</b>	<b>NN</b>	<b>VBZ</b>	<b>IN</b>	<b>NN</b>

### Features

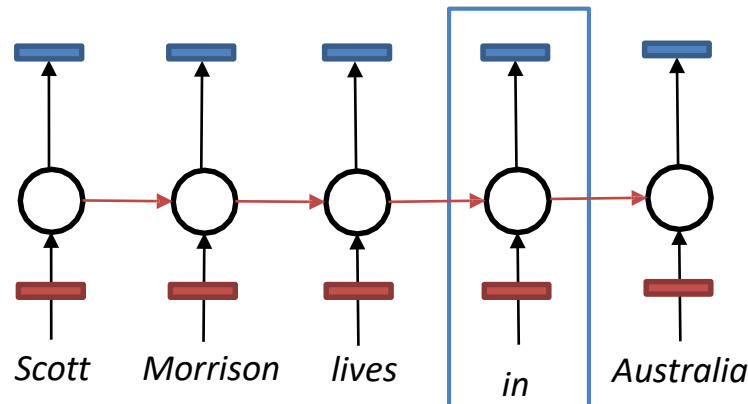
$W_0$	in
$W_{+1}$	Australia
$W_{-1}$	lives
$POS_{-1}$	VBZ
$POS_{-2}-POS_{-1}$	NN - VBZ
hasDigit?	0
...	...

## Sequence Inference for NER



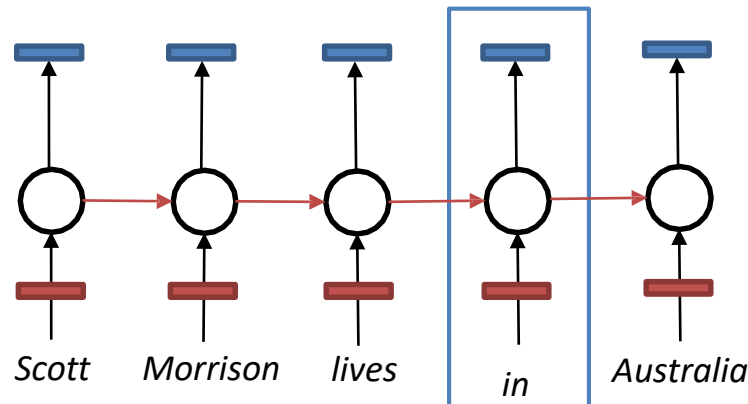
## Greedy Inference

- Greedy inference:
  - We just start at the left, and use our classifier at each position to assign a label
  - The classifier can depend on previous labeling decisions as well as observed data
- Advantages:
  - Fast, no extra memory requirements
  - Very easy to implement
  - With rich features including observations to the right, it may perform quite well
- Disadvantage:
  - Greedy. We make commit errors we cannot recover from



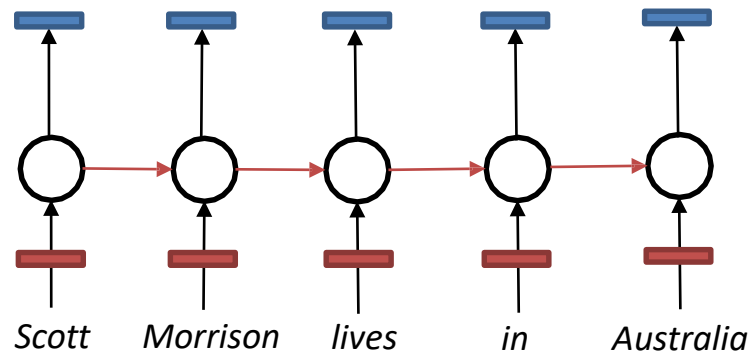
## Beam Inference

- Beam inference:
  - At each position keep the top k complete sequences.
  - Extend each sequence in each local way.
  - The extensions compete for the k slots at the next position.
- Advantages:
  - Fast; beam sizes of 3–5 are almost as good as exact inference in many cases.
  - Easy to implement (no dynamic programming required).
- Disadvantage:
  - Inexact: the globally best sequence can fall off the beam.



## Viterbi Inference

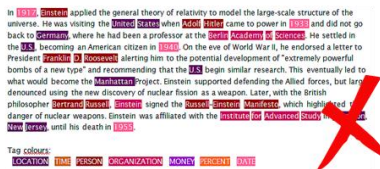
- Viterbi inference:
  - Dynamic programming or memorisation.
  - Requires small window of state influence (e.g., past two states are relevant).
- Advantage:
  - Exact: the global best sequence is returned.
- Disadvantage:
  - Harder to implement long-distance state-state interactions



## What if there is a language that do not have any annotation? NER in Low Resource Language


### Extreme Low Resource Language Scenario

1




**No annotated NE dataset  
for low resource target language**

2



**No parallel corpus between  
source and target language**

3



**No or few bilingual dictionary  
for target language**

### Current State of the Art model: Han et al. 2019 from Usyd NLP Research Group

TASK	DATASET	MODEL	METRIC NAME	METRIC VALUE	GLOBAL RANK	COMPARE
Low Resource Named Entity Recognition	CONLL 2003 Dutch	Low Resource Named Entity Recognition using Contextual Word Representation and Neural Cross-Lingual Knowledge Transfer	F1 score	75.10	# 1	<a href="#">See all</a>
Low Resource Named Entity Recognition	CONLL 2003 German	Low Resource Named Entity Recognition using Contextual Word Representation and Neural Cross-Lingual Knowledge Transfer	F1 score	58.63	# 1	<a href="#">See all</a>
Low Resource Named Entity Recognition	Conll 2003 Spanish	Low Resource Named Entity Recognition using Contextual Word Representation and Neural Cross-Lingual Knowledge Transfer	F1 score	75.34	# 1	<a href="#">See all</a>
Low Resource Named Entity Recognition	Uyghur Unsequestered Set	Low Resource Named Entity Recognition using Contextual Word Representation and Neural Cross-Lingual Knowledge Transfer	F1 score	42.88	# 1	<a href="#">See all</a>

## NER and Coreference Resolution

NER only produces a list of entities in a text.

- “I voted for **Scott** because he was most aligned with my values”

### Then, How to trace it?

**Coreference Resolution** is the task of finding all expressions that refer to the same entity in a text

- “**I** voted for **Scott** because **he** was most aligned with **my** values”
  - **Scott** ← **he**
  - **I** ← **my**



## What is Coreference Resolution?

Finding all mentions that refer to the same entity

Donald Trump said he considered nominating Ivanka Trump to be president of the World Bank because “she is very good with numbers,”



## What is Coreference Resolution?

Finding all mentions that refer to the same entity

**Donald Trump** said **he** considered nominating Ivanka Trump to be president of the World Bank because “she is very good with numbers,”



## What is Coreference Resolution?

Finding all mentions that refer to the same entity

Donald said he considered nominating **Ivanka Trump** to be **president of the World Bank** because “**she** is very good with numbers,”



## How to conduct Coreference Resolution?

### 1. Detect the mentions

*\* Mention: span of text referring to same entity*

- Pronouns

e.g. I, your, it, she, him, etc.

- Named entities

e.g. people, places, organisation etc.

- Noun phrases

e.g. a cat, a big fat dog, etc.

## The difficulty in coreference resolution

### 1. Detect the mentions

*\* Mention: span of text referring to same entity*

### Tricky mentions...

- **It** was very interesting
- **No staff**
- **The best university in Australia**

*How to handle this tricky mentions?      **Classifiers!***

## How to conduct Coreference Resolution?

### *1. Detect the mentions*

Donald Trump said he considered nominating Ivanka Trump to be president of the World Bank because “she is very good with numbers,”

### *2. Cluster the mentions*

Donald Trump said he considered nominating Ivanka Trump to be president of the World Bank because “she is very good with numbers,”

## How to cluster the mentions and find the coreference

### Coreference

It occurs when two or more expressions in a text refer to the same person or thing.

- “**Donald Trump** is a president of the United States. **Trump** was born and raised in the New York City borough of Queens”

### Anaphora

The use of a word referring back to a word used earlier in a text or conversation. Mostly noun phrases

- a word (anaphor) refers to another word (antecedent)
- “Donald Trump is a president of the United States. Before entering politics, he was a businessman and television personality”

*antecedent*

*anaphor*

## Coreference vs Anaphora

### *Coreference*

Donald Trump

Trump



### *Anaphora*

Donald Trump

↑  
he





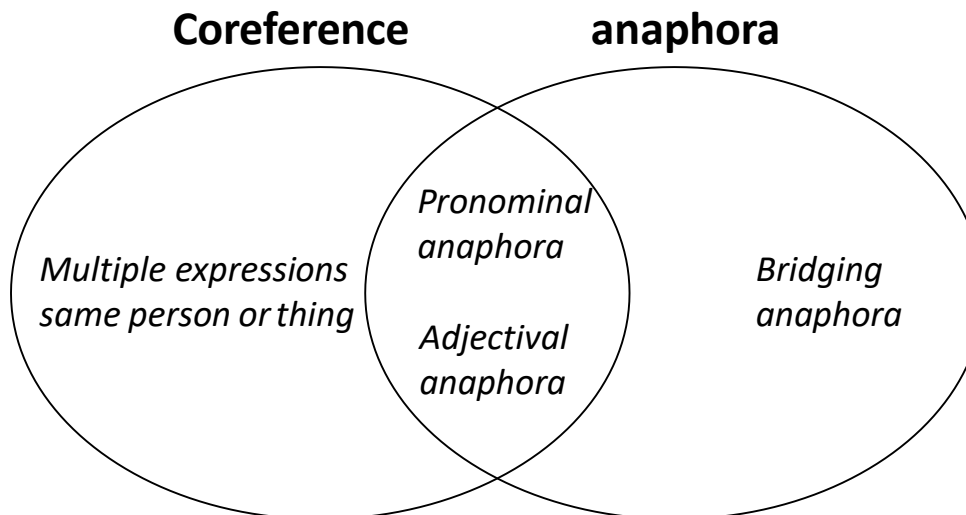
## Not all anaphoric relations are coreferential

### 1. Not all noun phrases have reference

- Every student like his speech
- No student like his speech

### 2. Not all anaphoric relations are co-referential (bridging anaphora)

- I attended **the meeting** yesterday. **The presentation** was awesome!



### cataphora

I almost stepped on **it**.  
It was a big **snake**...

## How to Cluster Mentions?

After detecting this all mentions in a text, we need to cluster them!

*Ivanka*

*Donald*

*he*

*her*

*she*

**Ivanka** was happy that **Donald** said **he** considered nominating **her** because **she** is very good with numbers

## How to Cluster Mentions?

After detecting this all mentions in a text, we need to cluster them!

*Ivanka*

*Donald*

*he*

*her*

*she*

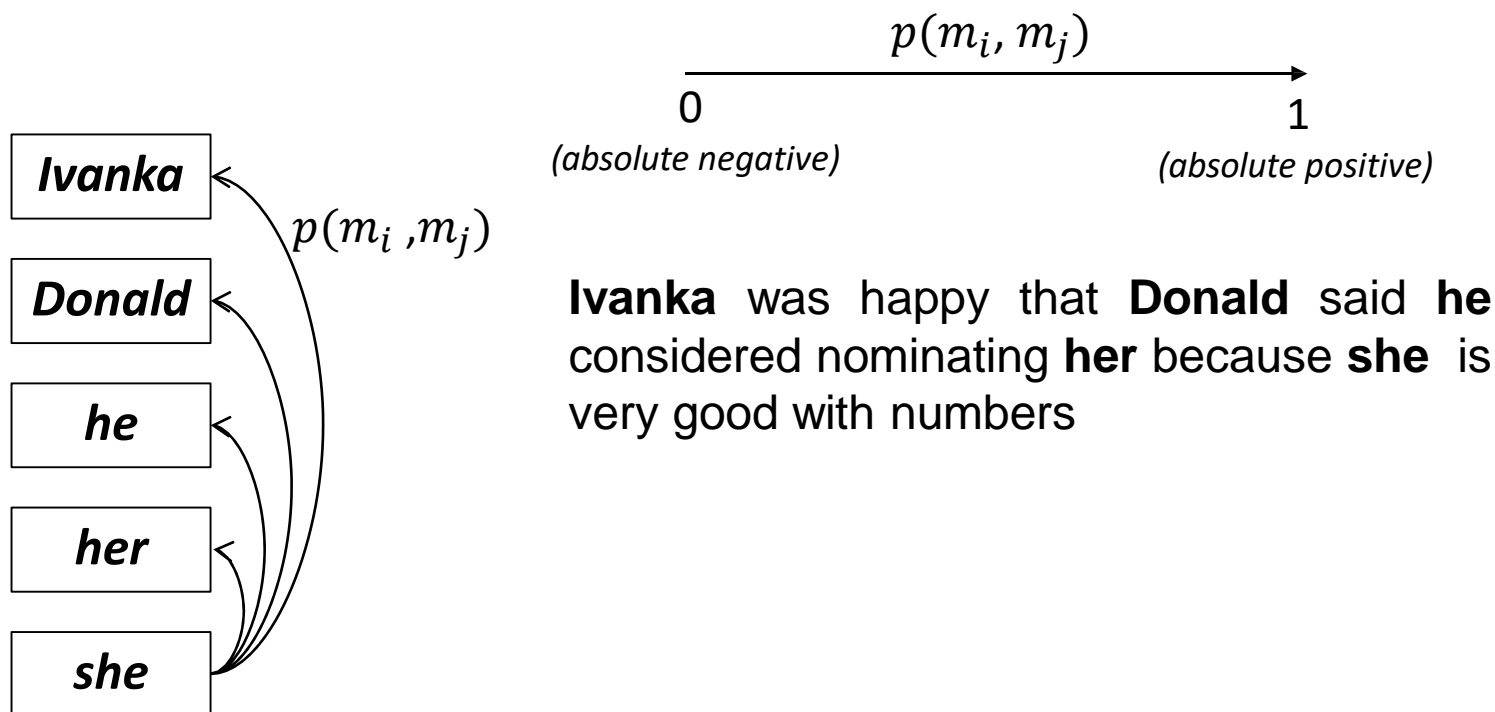
**Ivanka** was happy that **Donald** said **he** considered nominating **her** because **she** is very good with numbers

*Gold cluster 1*

*Gold cluster 2*

## How to Cluster Mentions?

- Train a binary classifier that assigns every pair of mentions a probability of being coreferent:  $p(m_i, m_j)$



## Mention Pair Training

- N mentions in a document
- $y_{ij} = 1$  if mentions  $m_i$  and  $m_j$  are coreferent, -1 if otherwise
- Just train with regular cross-entropy loss (looks a bit different because it is binary classification)

$$J = - \sum_{i=2}^N \sum_{j=1}^i y_{ij} \log p(m_j, m_i)$$

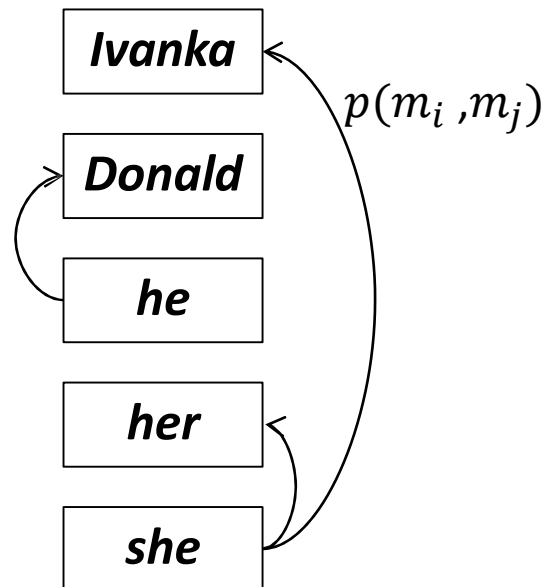
Coreferent mentions pairs should get high probability, others should get low probability

Iterate through  
mentions

Iterate through candidate  
antecedents (previously  
occurring mentions)

## Mention Pair Testing

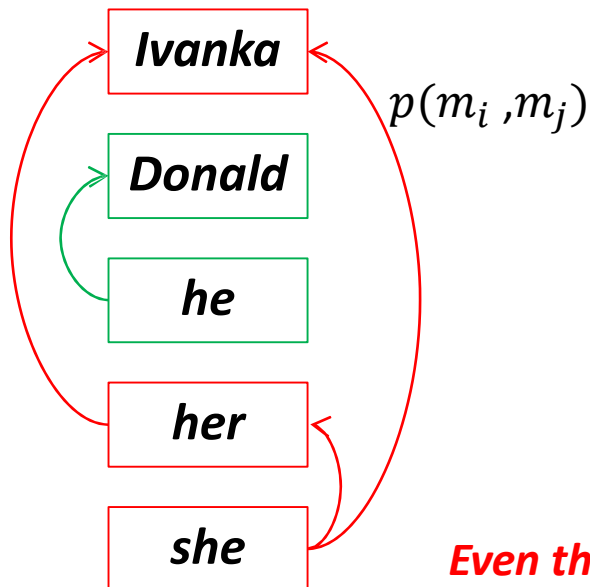
- Coreference resolution is a clustering task, but we are only scoring pairs of mentions... what to do?
- Pick some threshold (e.g., 0.5) and add coreference links between mention pairs where  $p(m_i, m_j)$  is above the threshold



**Ivanka** was happy that **Donald** said **he** considered nominating **her** because **she** is very good with numbers

## Mention Pair Testing

- Pick some threshold (e.g., 0.5) and add coreference links between mention pairs where  $p(m_i, m_j)$  is above the threshold
- Take the transitive closure to get the clustering

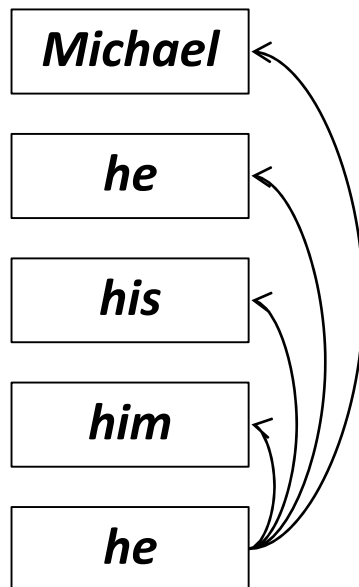


**Ivanka** was happy that **Donald** said **he** considered nominating **her** because **she** is very good with numbers

*Even though the model did not predict this coreference link, Ivanka and her are coreferent due to transitivity*

## Mention Pair Testing: Issue

- Assume that we have a **long document** with the following mentions
- Michael... he ... his ... him ... <several paragraphs>
- ... won the game because he ...



Many mentions only have one clear antecedent but we are asking the model to predict all of them

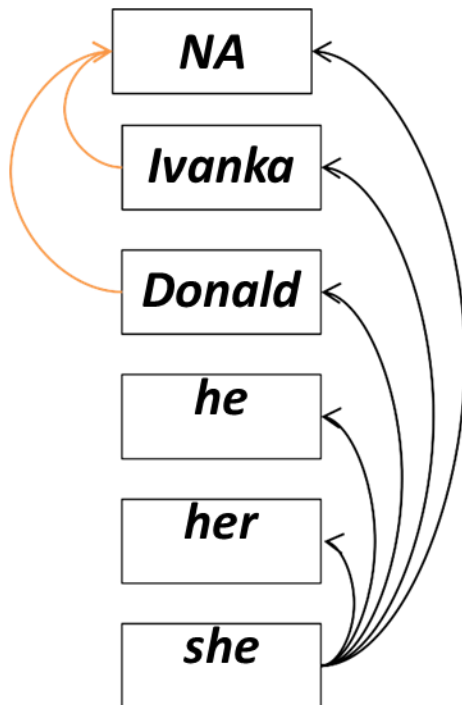
**Alternative solution:** instead train the model to predict only one antecedent for each mention

***Mention Ranking***



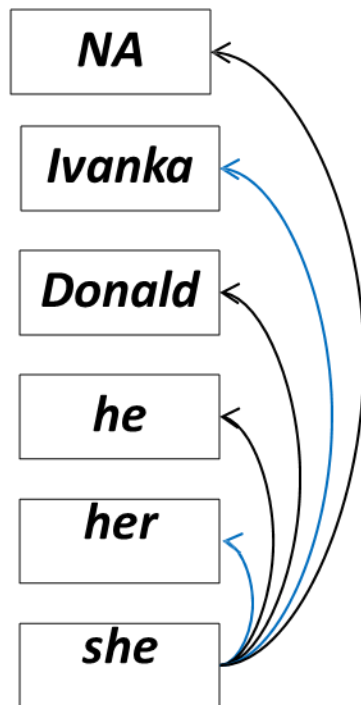
## Mention Ranking

- Assign each mention its highest scoring candidate antecedent according to the model
- Dummy NA mention allows model to decline linking the current mention to anything (“singleton” or “first” mention)



## Mention Ranking

- Assign each mention its highest scoring candidate antecedent according to the model
- Dummy NA mention allows model to decline linking the current mention to anything (“singleton” or “first” mention)

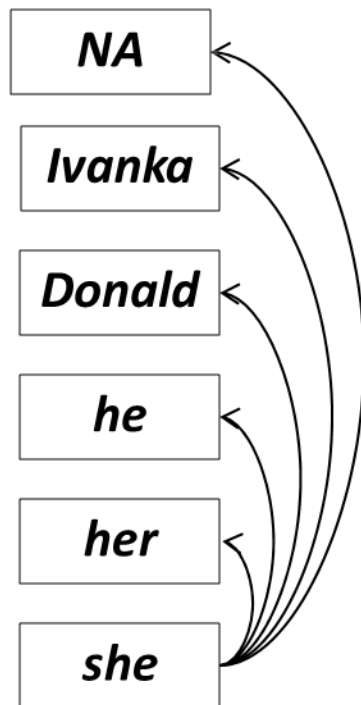


What can be the best antecedent for she?

**Positive examples:** model has to assign a high probability to either one (but not necessarily both)

## Mention Ranking

- Assign each mention its highest scoring candidate antecedent according to the model
- Dummy NA mention allows model to decline linking the current mention to anything (“singleton” or “first” mention)



What can be the best antecedent for she?

Apply a **softmax** over the scores for candidate antecedents so probabilities sum to 1

- $p(\text{NA}, \text{she}) = 0.1$
- $p(\text{Ivanka}, \text{she}) = 0.5$
- $p(\text{Donald}, \text{she}) = 0.1$
- $p(\text{he}, \text{she}) = 0.1$
- $p(\text{her}, \text{she}) = 0.2$

## Mention Ranking

- Assign each mention its highest scoring candidate antecedent according to the model
- Dummy NA mention allows model to decline linking the current mention to anything (“singleton” or “first” mention)

**NA**

**Ivanka**

**Donald**

**he**

**her**

**she**

What can be the best antecedent for she?

Apply a **softmax** over the scores for candidate antecedents so probabilities sum to 1

- $p(\text{NA}, \text{she}) = 0.1$
- **$p(\text{Ivanka}, \text{she}) = 0.5$**
- $p(\text{Donald}, \text{she}) = 0.1$
- $p(\text{he}, \text{she}) = 0.1$
- $p(\text{her}, \text{she}) = 0.2$

*only add highest scoring  
coreference link*

## Coreference Models: Training

- The current mention  $m_j$  should be linked to any one of the candidate antecedents it's coreferent with.
- Mathematically, maximize this probability:

$$\sum_{j=1}^{i-1} \mathbb{1}(y_{ij} = 1) p(m_i, m_j)$$

Iterate through candidate antecedents (previously occurring mentions)

For ones that are coreferent to  $m_j$ ...

...we want the model to assign a high probability

## Coreference Models: Training

- The current mention  $m_j$  should be linked to any one of the candidate antecedents it's coreferent with.
- Mathematically, maximize this probability:

$$\sum_{j=1}^{i-1} \mathbb{1}(y_{ij} = 1) p(m_i, m_j)$$

Iterate through candidate  
antecedents (previously  
occurring mentions)

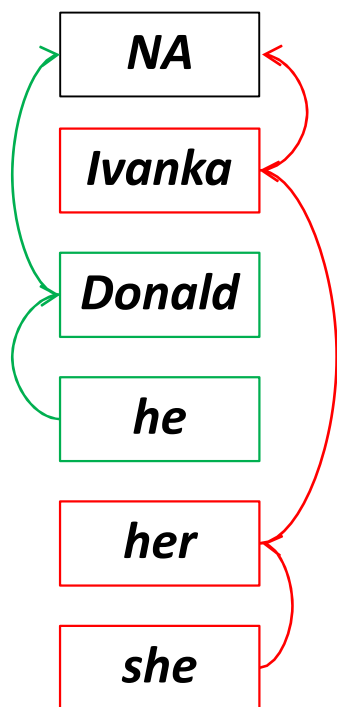
For ones that  
are coreferent  
to  $m_j$ ...

...we want the model to  
assign a high probability

*The model could produce 0.9 probability for one of the correct antecedents and low probability for everything else, and the sum will still be large*

## Mention Ranking Models: Test Time

- Similar to mention-pair model except each mention is assigned only one antecedent



How do we compute the probabilities?

- Non-neural statistical classifier
- Simple neural network
- More advanced model using LSTMs, attention

**How do we compute the probabilities?**

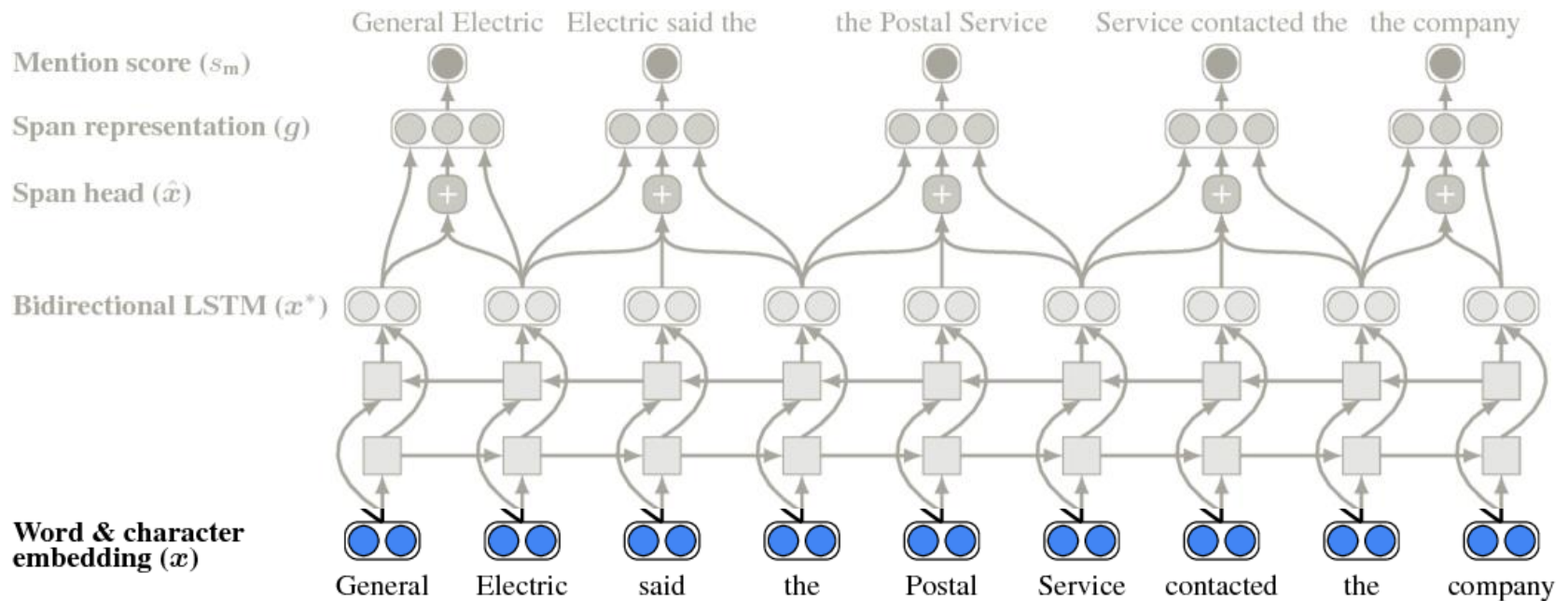
**End to End Model (Lee et al., 2017)**

- Current state-of-the-art model for coreference resolution (before 2019)
- Mention ranking model
- Improvements over simple feed-forward NN
  - Use an LSTM
  - Use attention (will learn about this in Lecture 10)
  - Do mention detection and coreference end-to-end
    - No mention detection step



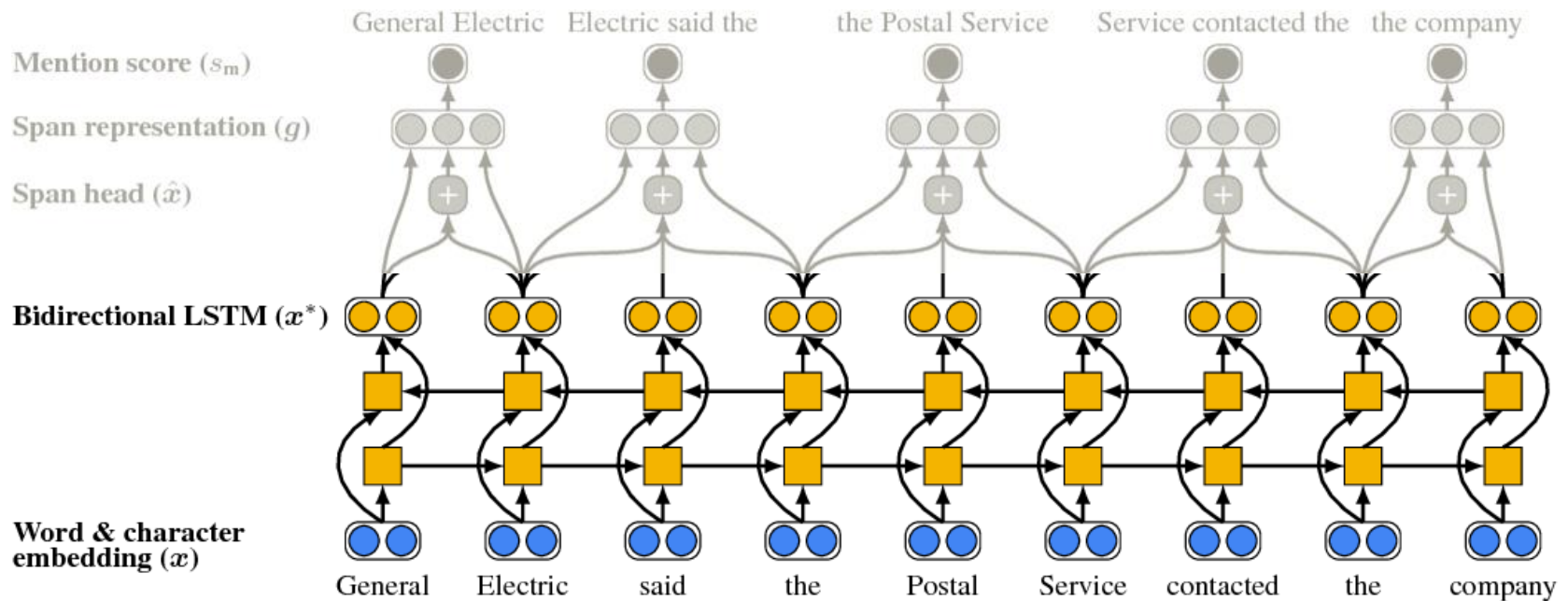
## End to End Model (Lee et al., 2017)

- First embed the words in the document using a word embedding matrix and a character-level embedding



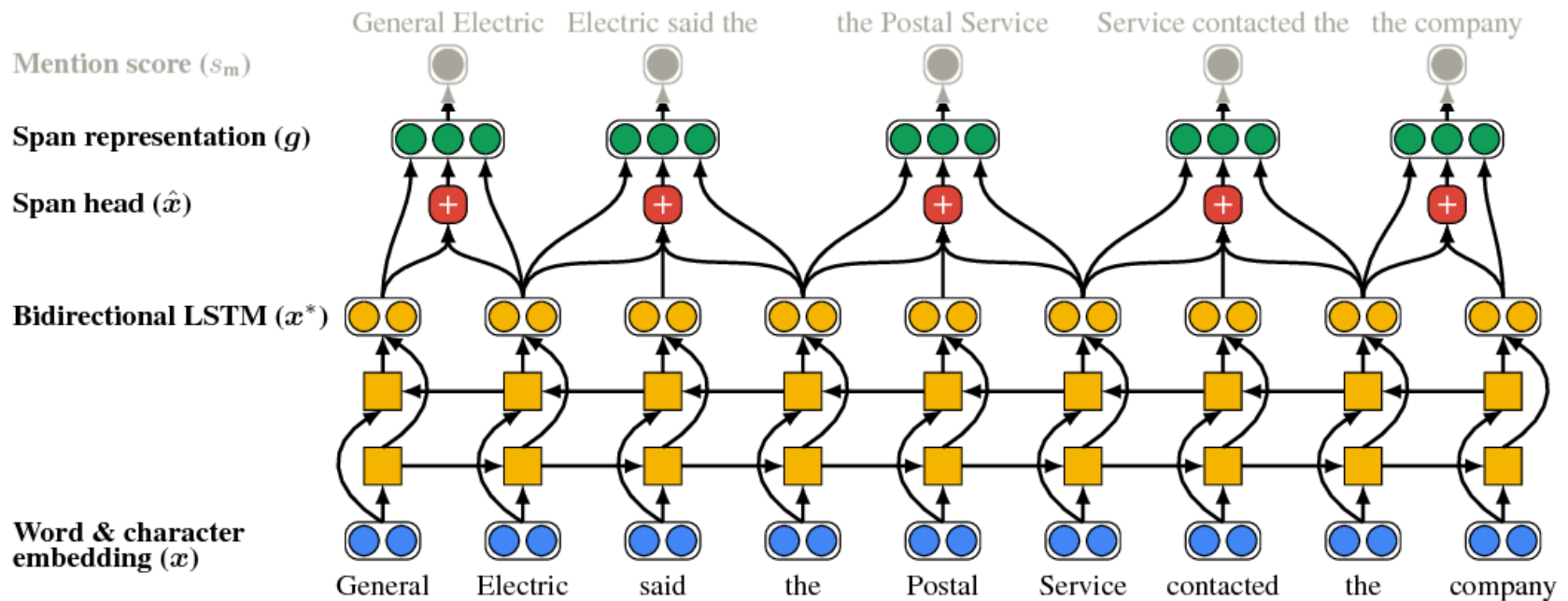
## End to End Model (Lee et al., 2017)

- Then run a bidirectional LSTM over the document



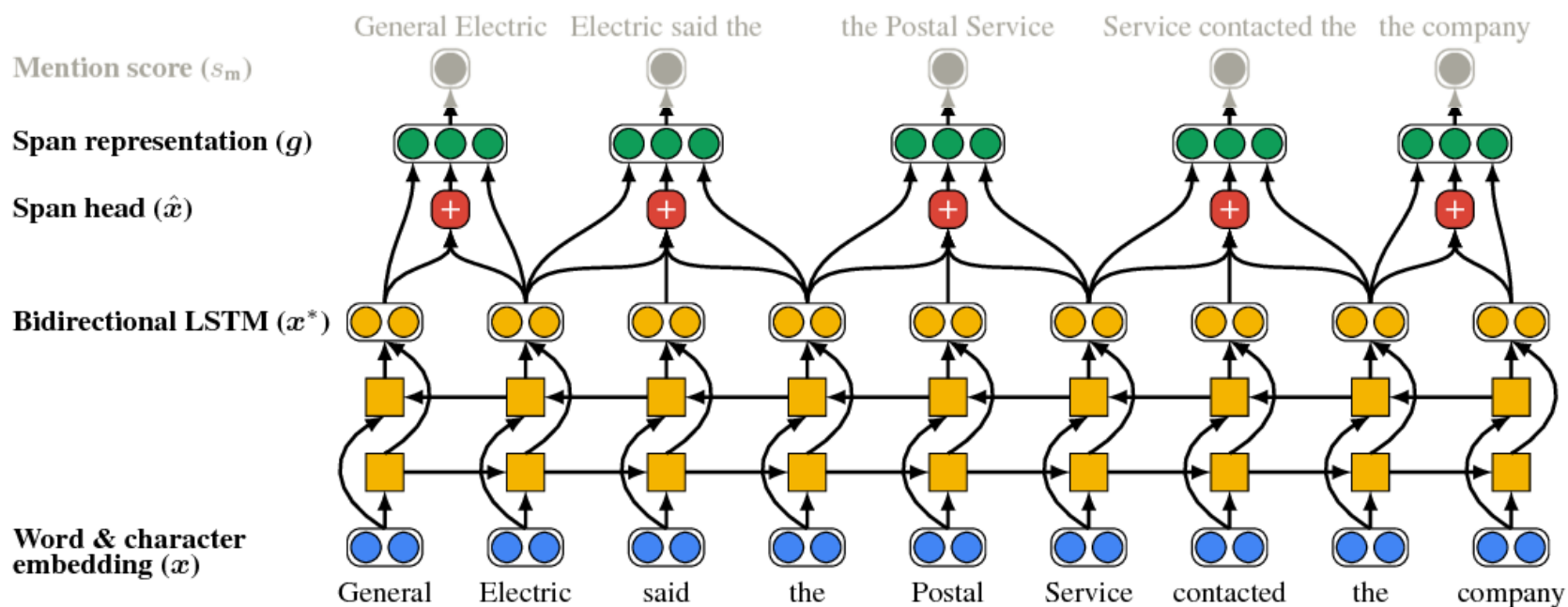
## End to End Model (Lee et al., 2017)

- Next, represent each span of text  $i$  going from  $START(i)$  to  $END(i)$  as a vector



## End to End Model (Lee et al., 2017)

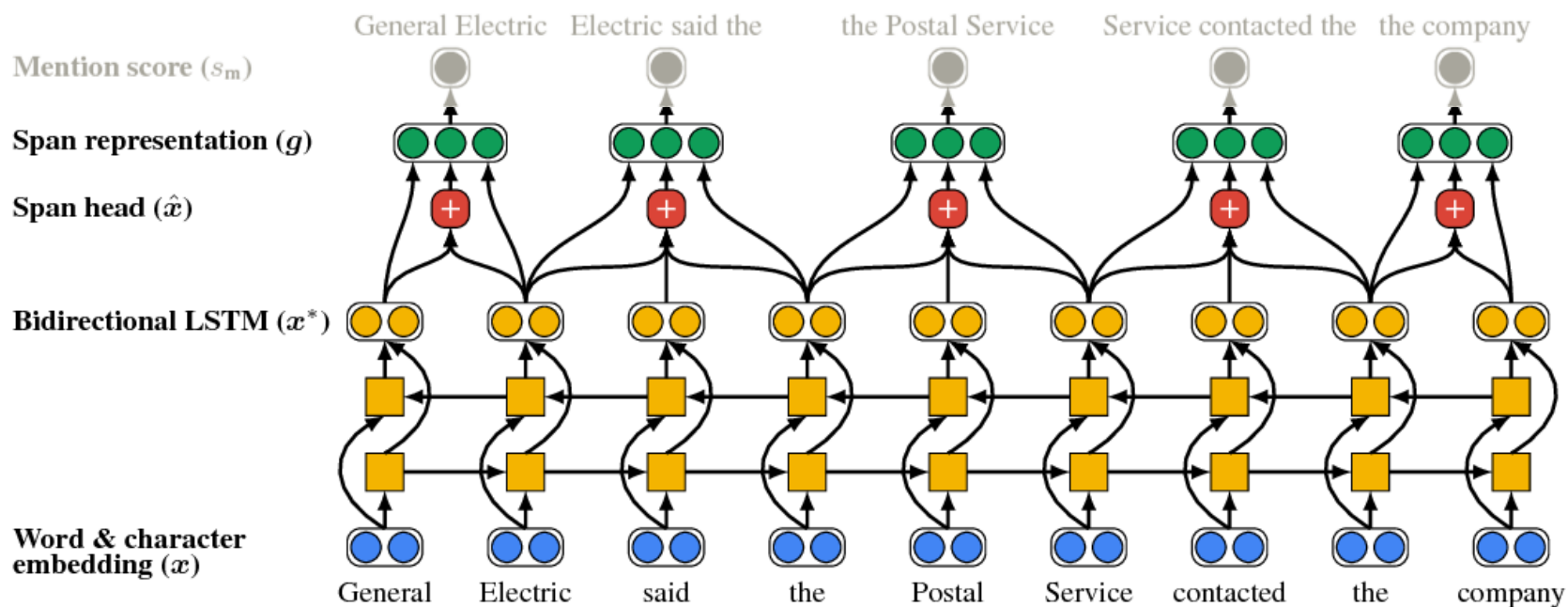
- Next, represent each span of text  $i$  going from  $START(i)$  to  $END(i)$  as a vector



*General, General Electric, General Electric said, ... Electric, Electric said, ... will all get its own vector representation*

## End to End Model (Lee et al., 2017)

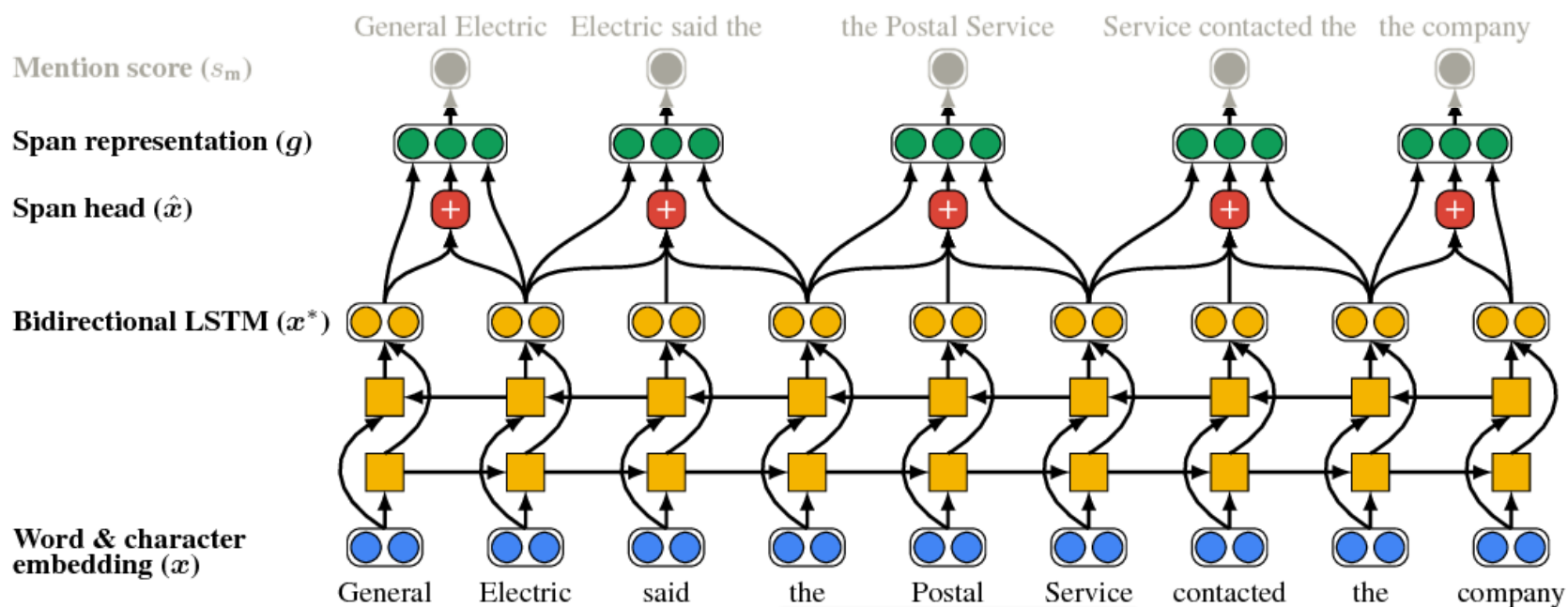
- Next, represent each span of text  $i$  going from  $START(i)$  to  $END(i)$  as a vector



$$\text{Span Representation: } g_i = [x_{\text{START}(i)}^*, x_{\text{END}(i)}^*, \hat{x}_i, \phi(i)]$$

## End to End Model (Lee et al., 2017)

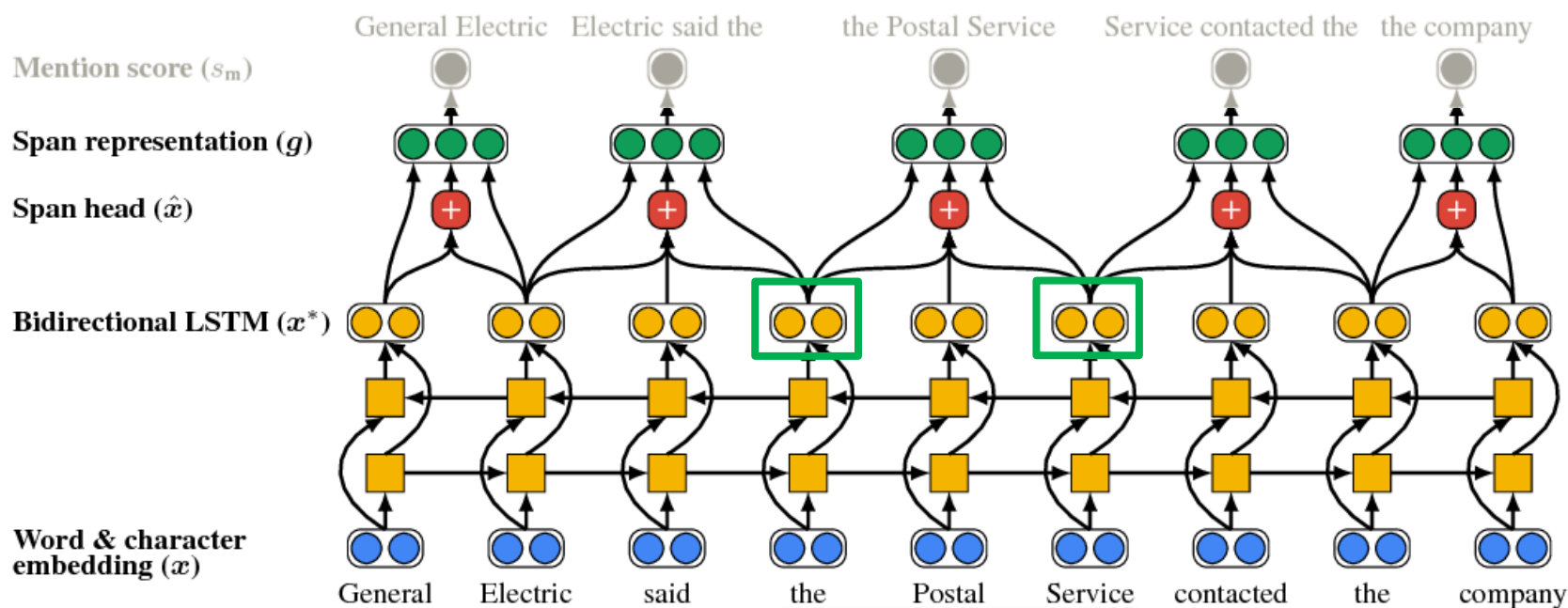
- Next, represent each span of text  $i$  going from  $START(i)$  to  $END(i)$  as a vector. *For example, for “the postal service”*



$$\text{Span Representation: } g_i = [x_{\text{START}(i)}^*, x_{\text{END}(i)}^*, \hat{x}_i, \phi(i)]$$

## End to End Model (Lee et al., 2017)

- Next, represent each span of text  $i$  going from  $START(i)$  to  $END(i)$  as a vector. *For example, for “the postal service”*



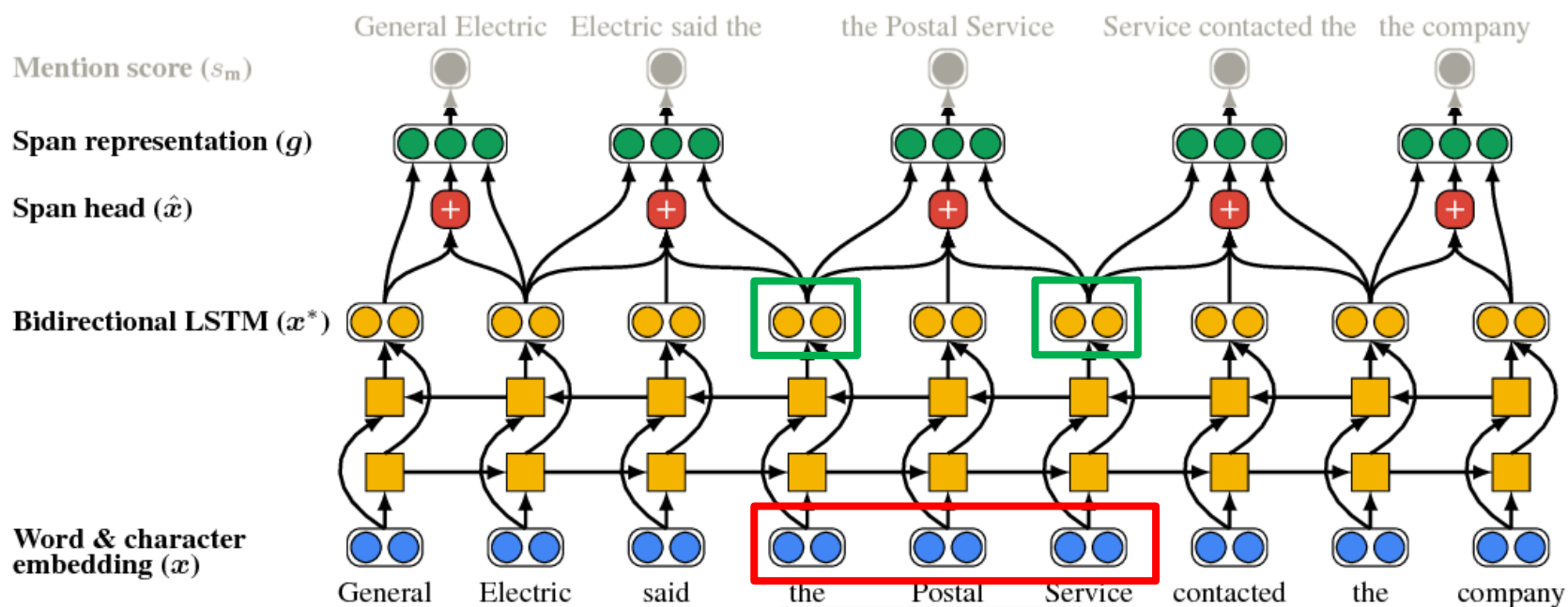
$$\text{Span Representation: } g_i = [\underline{x_{\text{START}(i)}^*}, \underline{x_{\text{END}(i)}^*}, \hat{x}_i, \phi(i)]$$

*Bi-LSTM hidden states for span's start and end*



## End to End Model (Lee et al., 2017)

- Next, represent each span of text  $i$  going from  $START(i)$  to  $END(i)$  as a vector. *For example, for “the postal service”*



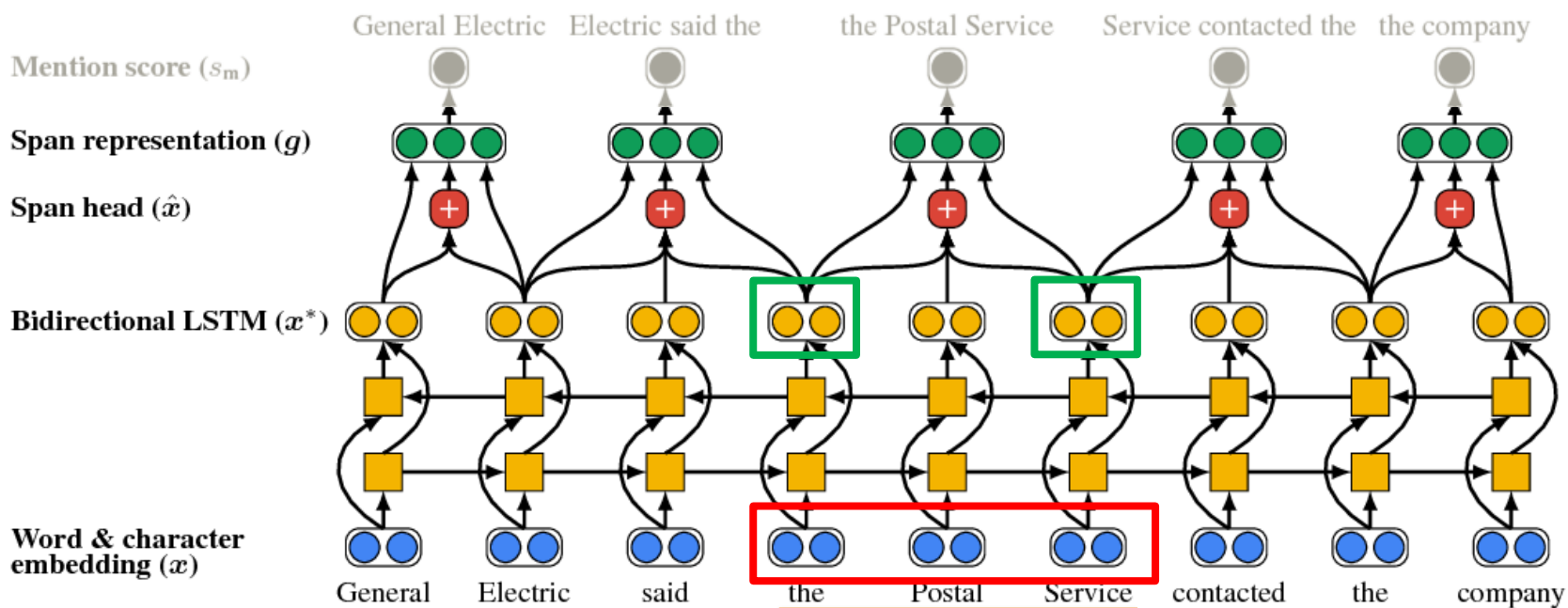
$$\text{Span Representation: } g_i = [\underline{x_{\text{START}(i)}^*}, \underline{x_{\text{END}(i)}^*}, \underline{\hat{x}_i}, \phi(i)]$$

*Attention-based representation (Lecture 10)*



## End to End Model (Lee et al., 2017)

- Next, represent each span of text  $i$  going from  $START(i)$  to  $END(i)$  as a vector. *For example, for “the postal service”*



$$\text{Span Representation: } g_i = [\underline{x_{\text{START}(i)}^*}, \underline{x_{\text{END}(i)}^*}, \underline{\hat{x}_i}, \underline{\phi(i)}]$$

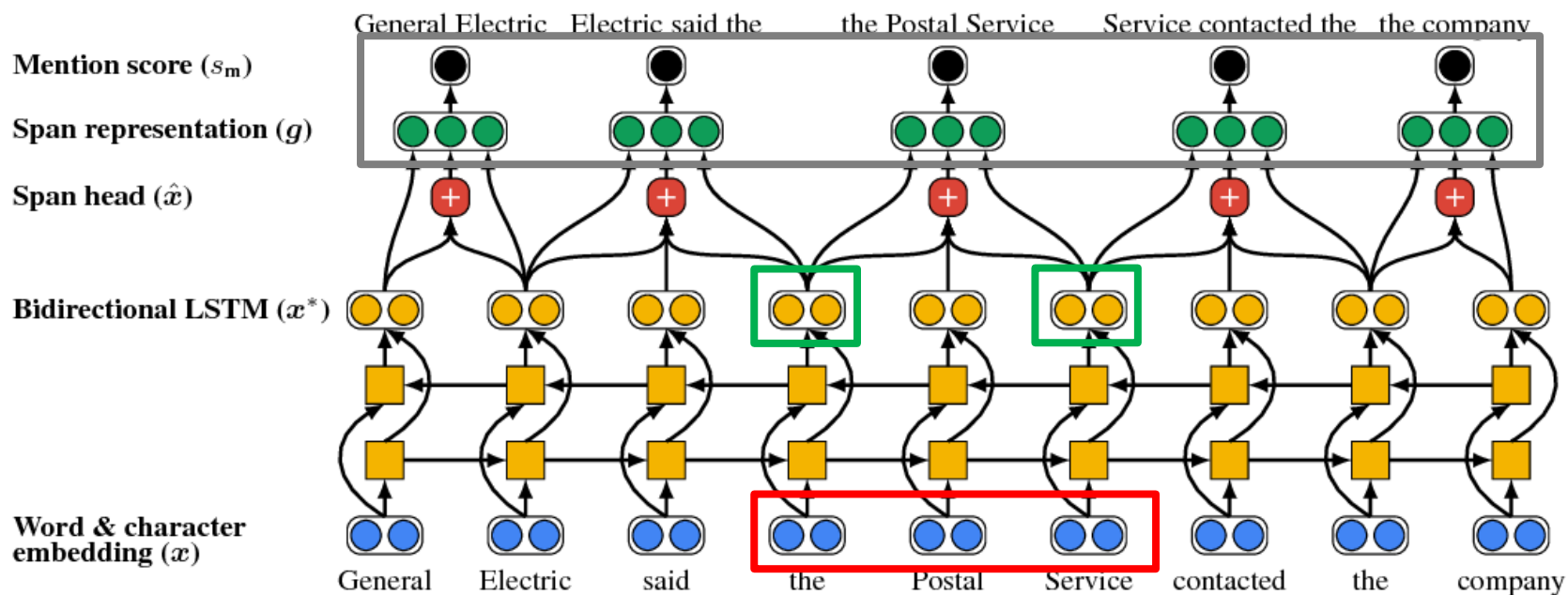
*Additional features*

## End to End Model (Lee et al., 2017)

- Next, represent each span of text  $i$  going from  $START(i)$  to  $END(i)$  as a vector.

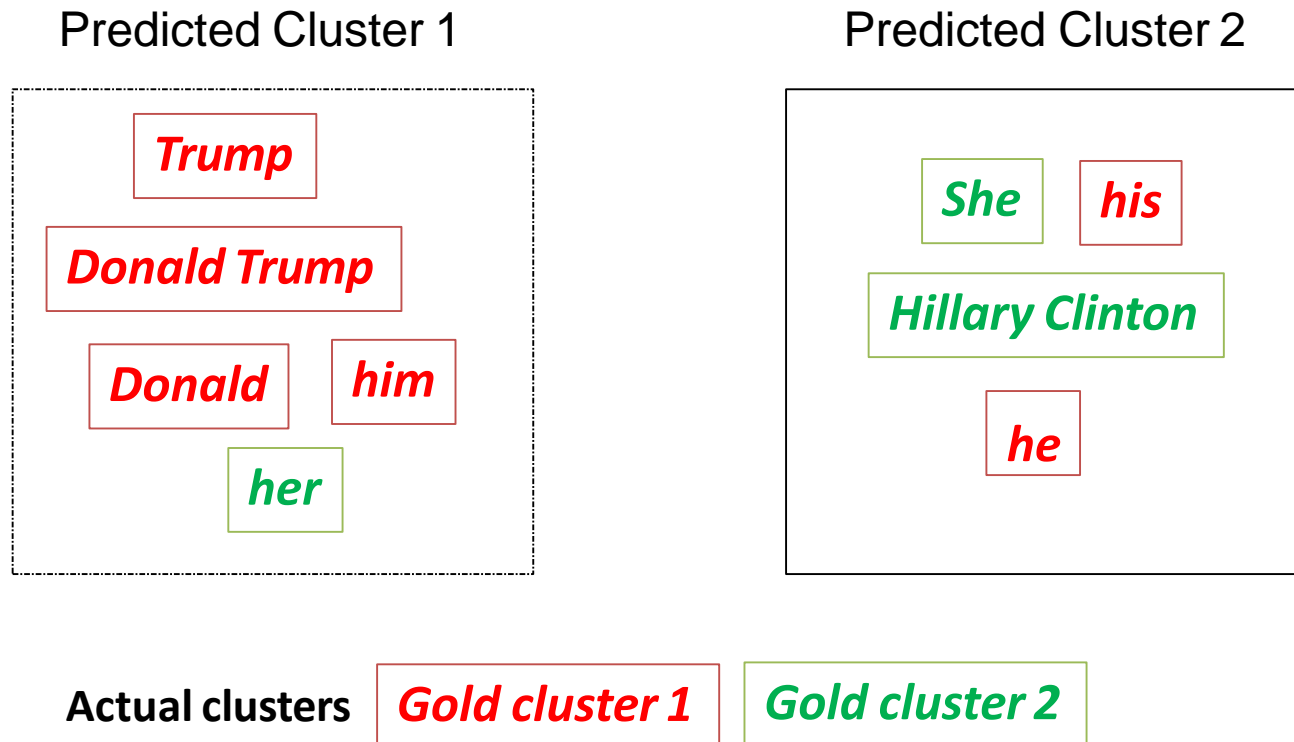
$$s(i, j) = s_m(i) + s_m(j) + s_a(i, j)$$

Are spans  $i$  and  $j$  coreference mentions? Is  $i$  a mention? Is  $j$  a mention? Do they look coreferent?



## How to evaluate coreference?

There are different types of metrics available for evaluating coreference, such as B-CUBED, MUC, CEAF, LEA, BLANC, or Often report the average over a few different metrics

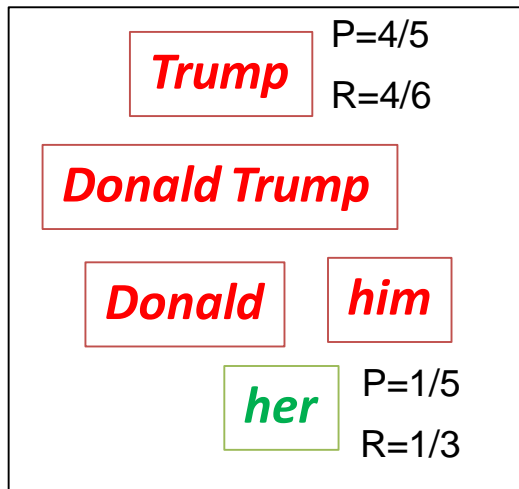


## How to evaluate coreference?

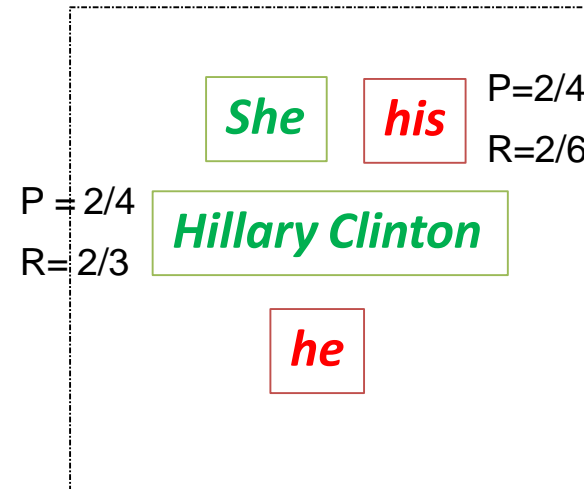
Let's evaluate with B-CUBED metrics

- Compute **P**recision and **R**ecall for each mention.

Predicted Cluster 1



Predicted Cluster 2



Actual clusters

**Gold cluster 1**

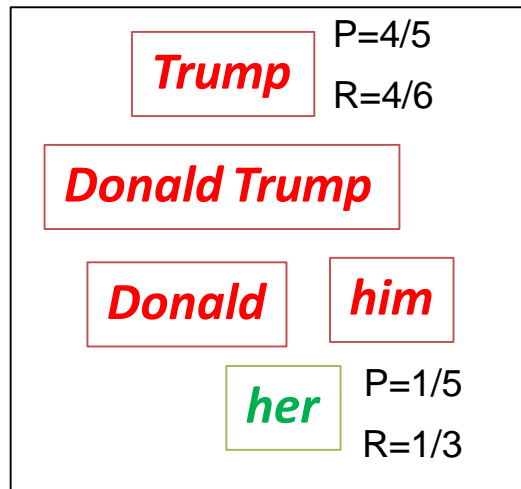
**Gold cluster 2**

## How to evaluate coreference?

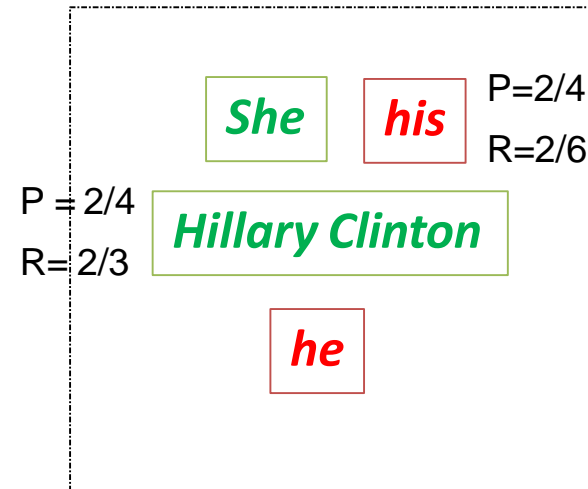
Let's evaluate with B-CUBED metrics

- Compute precision and recall for each mention.
- Average the individual Ps and Rs

Predicted Cluster 1



Predicted Cluster 2



Actual clusters

**Gold cluster 1**

**Gold cluster 2**

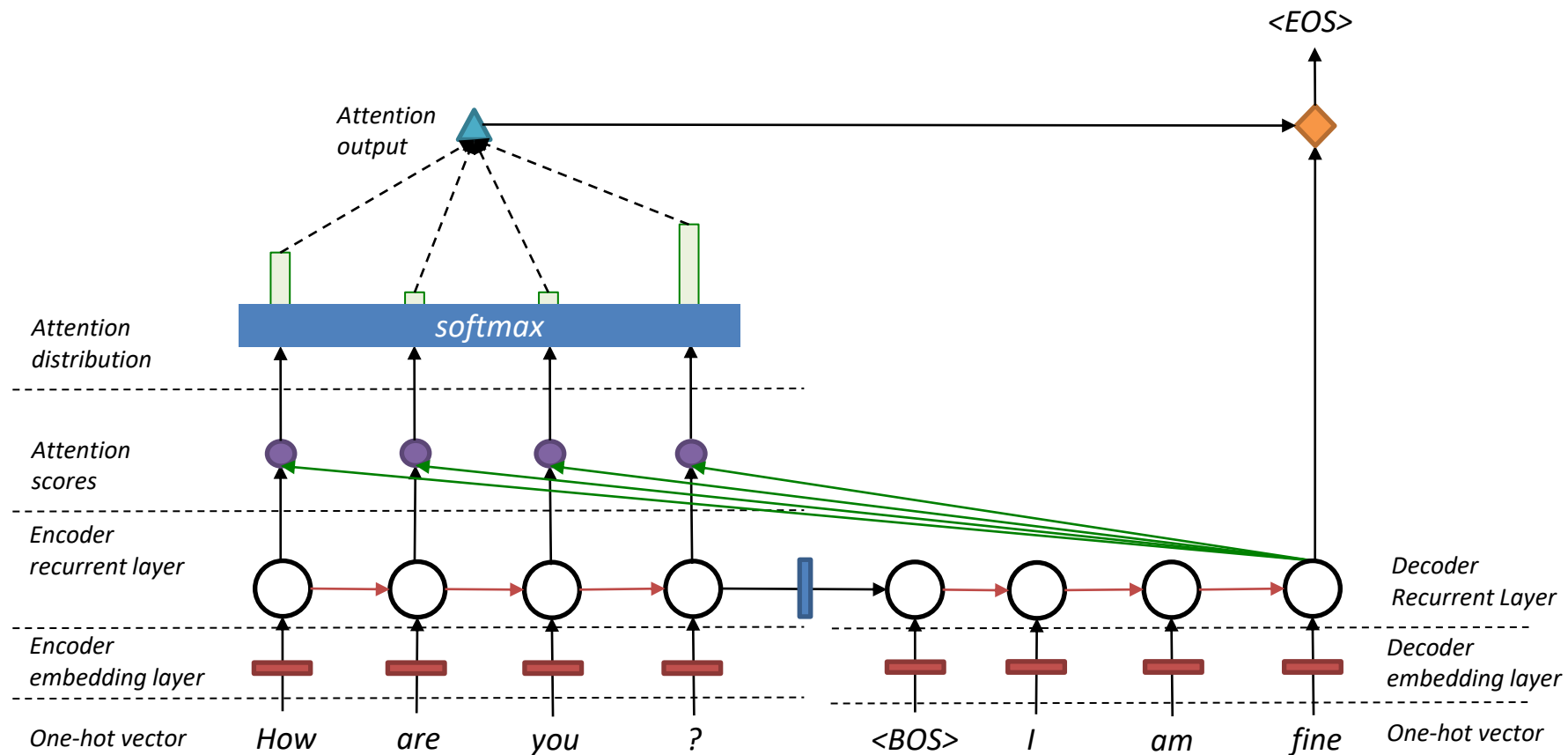
## Performance Comparison

OntoNotes dataset: ~3000 documents labeled by humans

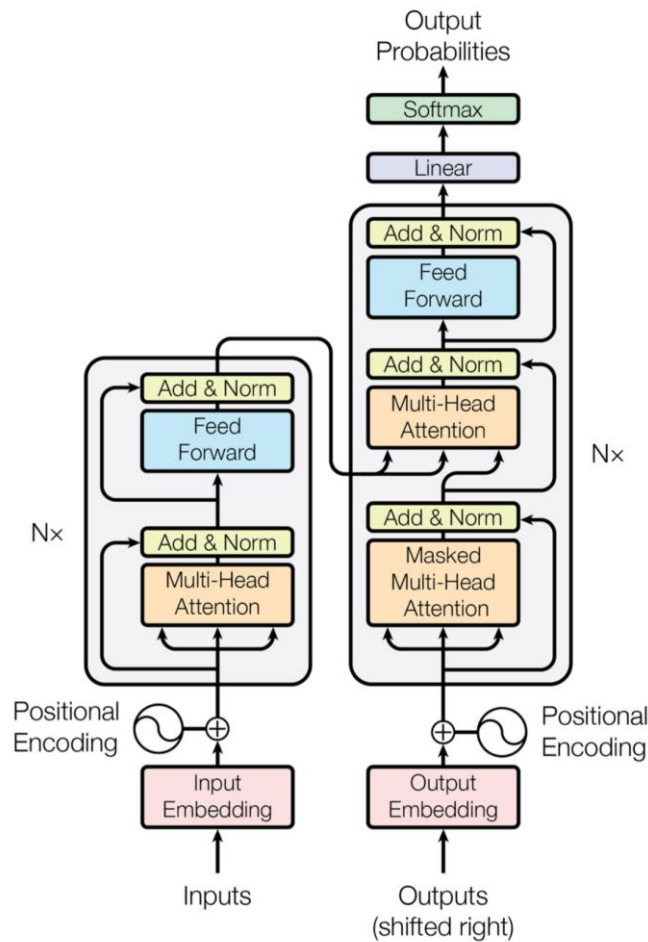
- English and Chinese data

Model	Approach	English	Chinese
Lee et al. (2010)	Rule-based system	~55	~50
Chen & Ng (2012) [CoNLL 2012 Chinese winner]	Non-neural machine learning models	54.5	57.6
Fernandes (2012) [CoNLL 2012 English winner]		60.7	51.6
Wiseman et al. (2015)	Neural mention ranker	63.3	—
Lee et al. (2017)	Neural mention ranker (end- to-end style)	67.2	--
UsydNLP (2019)	Neural mention ranker with lemma cross validation	74.87	--

## Attention and Reading Comprehension



## Transformer and Machine Translation



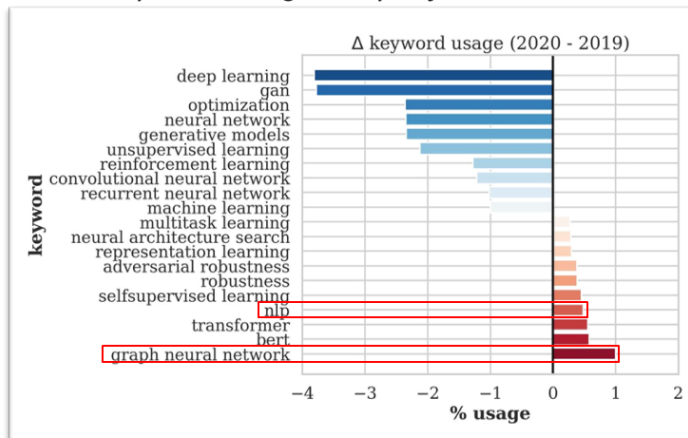
### *Pretrained Model*

**ERNIE****ELMO****BERT**



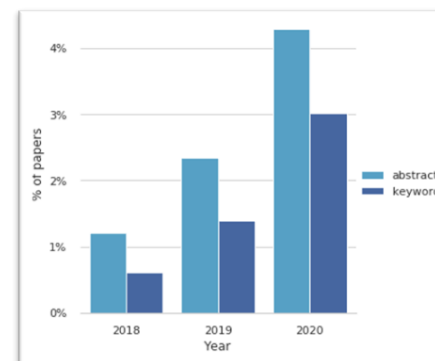
## Graph Neural Network and NLP

Keywords usage analysis for ICLR 2020

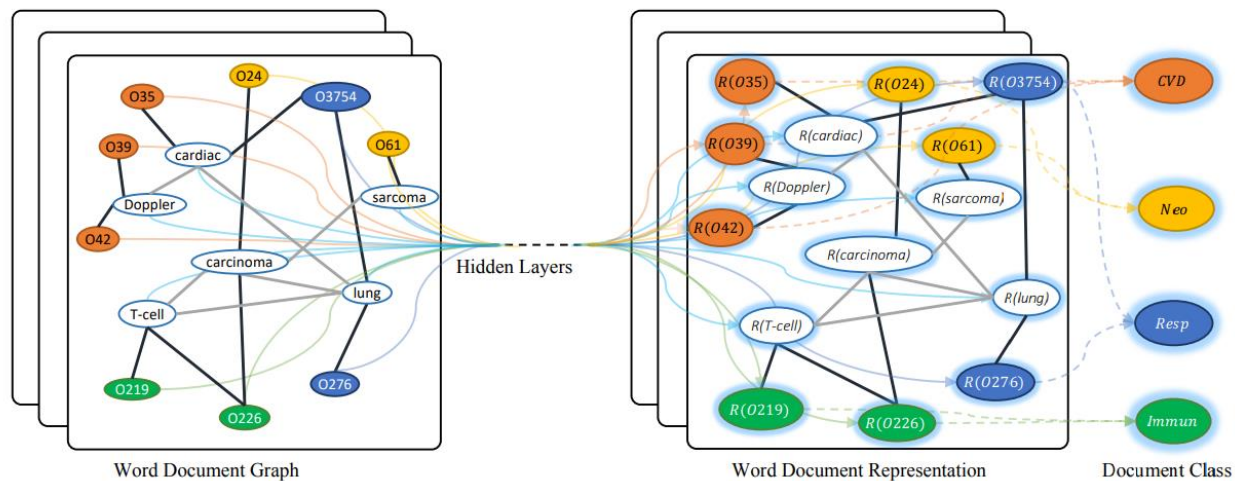
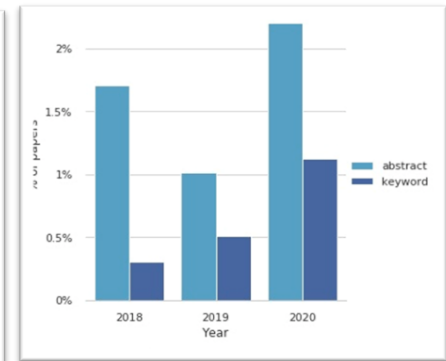


Keywords usage analysis for ICLR, NIPS, CVPR

Graph Neural Networks



Natural Language Processing



## Reference for this lecture

- Deng, L., & Liu, Y. (Eds.). (2018). Deep Learning in Natural Language Processing. Springer.
- Rao, D., & McMahan, B. (2019). Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning. " O'Reilly Media, Inc."
- Manning, C. D., Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.
- Manning, C 2018, Natural Language Processing with Deep Learning, lecture notes, Stanford University
- Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2015). A diversity-promoting objective function for neural conversation models. arXiv preprint arXiv:1510.03055.
- Jiang, S., & de Rijke, M. (2018). Why are Sequence-to-Sequence Models So Dull? Understanding the Low-Diversity Problem of Chatbots. arXiv preprint arXiv:1809.01941.
- Liu, C. W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., & Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. arXiv preprint arXiv:1603.08023.