COMP5046
Natural Language Processing

Lecture 9: Information Extraction I
Named Entity Recognition and Coreference Resolution

Semester 1, 2019
School of Computer Science
The University of Sydney, Australia

THE UNIVERSITY OF SYDNEY

Caren Han
Caren.Han@sydney.edu.au

0

LECTURE PLAN

Lecture 9: Named Entity Recognition and Coreference Resolution

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

1 Information Extraction

Information Extraction

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

Here are some questions..

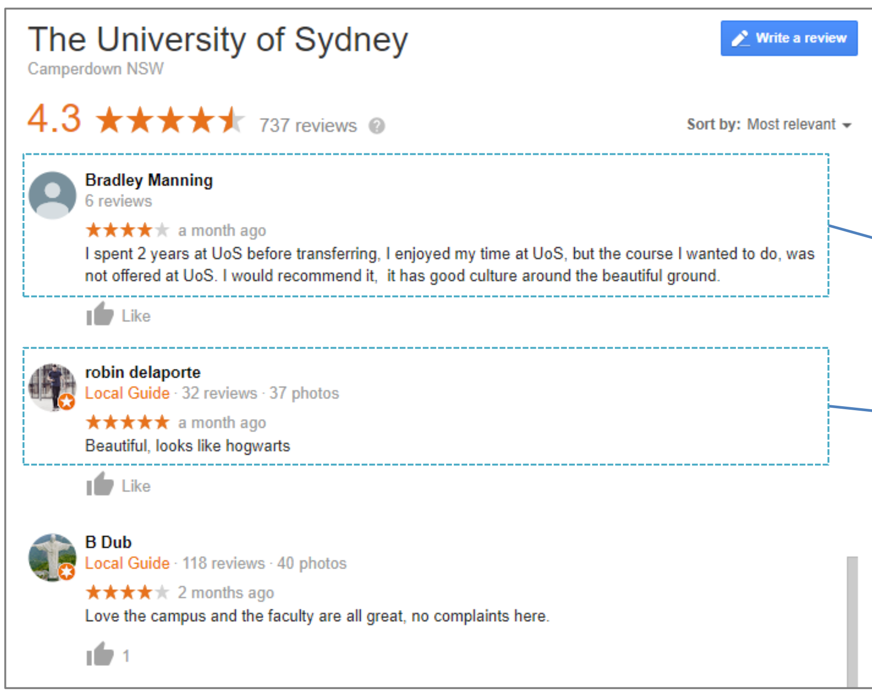
- How to allow computation to be done on the unstructured data
- How to extract clear, factual information
- How to put in a semantically precise form that allows further inferences to be made by computer algorithms

1 Information Extraction

How to extract the structured clear, factual information

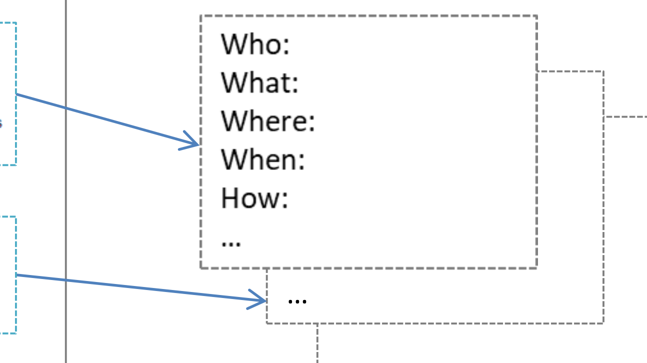
- Find and understand limited relevant parts of texts
- Gather information from many pieces of text
- Produce a structured representation of relevant information

relations (in the database sense) or a knowledge base



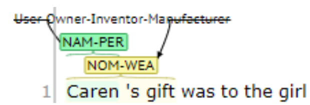
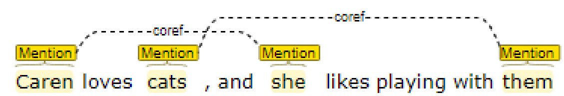
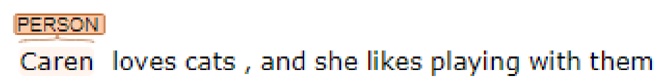
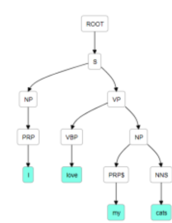
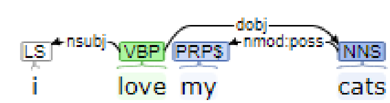
"5W1H"

who, what, where, when, why, how



Information Extraction Pipeline with NLP

Understanding

	Task	Input	Output
Application	***Sentiment Analysis***	Caren loves cats, and she likes playing with them	**[positive: 90.10%] [neutral: 4.70%] [negative: 5.10%]**
	Relation Extraction	Caren's gift was to the girl.	
	Coreference Resolution	Caren loves cats, and she likes playing with them	
NLP Stack	***Entity Extraction***	Caren loves cats, and she likes playing with them	
	Parsing	I love my cats	 
	PoS Tagging	I love my cats	**[I/JJ] [love/VBP] [my/PRP] [cats/NNS]**
	Stemming	I love my cats	**[I] [love] [my] [cat]**
	Tokenisation	I love my cats	**[I] [love] [my] [cats]**

Named Entity Recognition (NER)

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

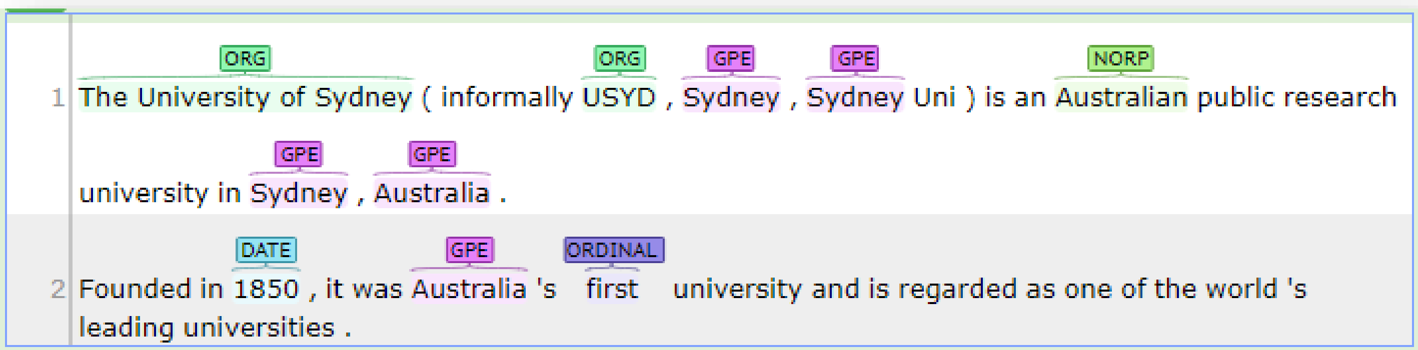
**classes can be different based on annotated dataset*

Named Entity Recognition (NER)

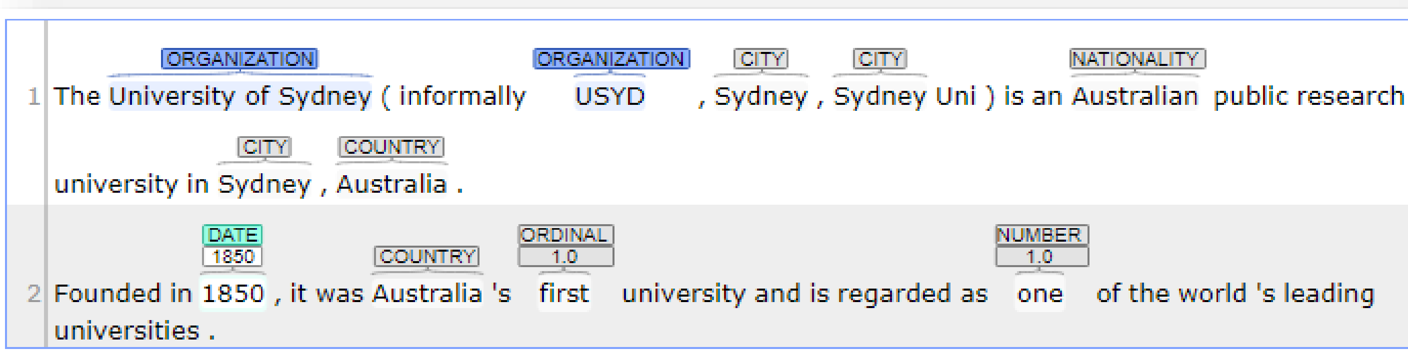
How to recognize Named Entities?

Identify and **classify** names in text

Upenn CogComp-NLP



Stanford CoreNLP 3.9.2



2 Named Entity Recognition (NER)

How to evaluate the NER performance?

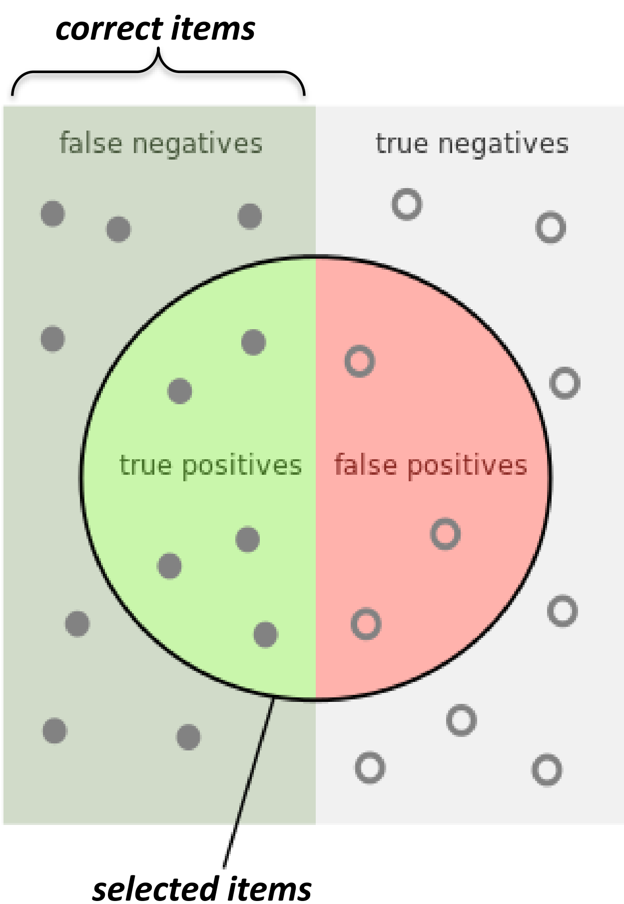
The goal: ***predicting entities in a text***

***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**

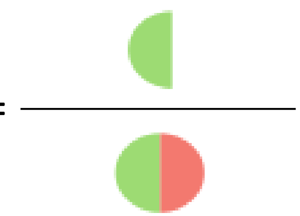
Named Entity Recognition (NER)

How to evaluate the NER performance? Precision and recall



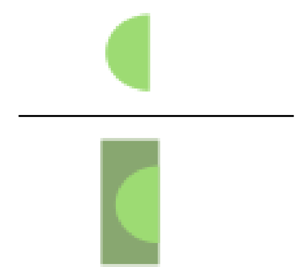
How many selected items are correct?

Precision =



How many correct items are selected?

Recall =



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

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

2 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)

2 Named Entity Recognition (NER)

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
conll-03	Reuters	301k	per, org, loc, misc
bbn	Wall Street Journal	1174k	https://catalog.ldc.upenn.edu/docs/LDC2005T33/BBN-Types-Subtypes.html

- Different genre and style
- Different Annotation Schema

Named Entity Recognition (NER)

Data for learning named entity

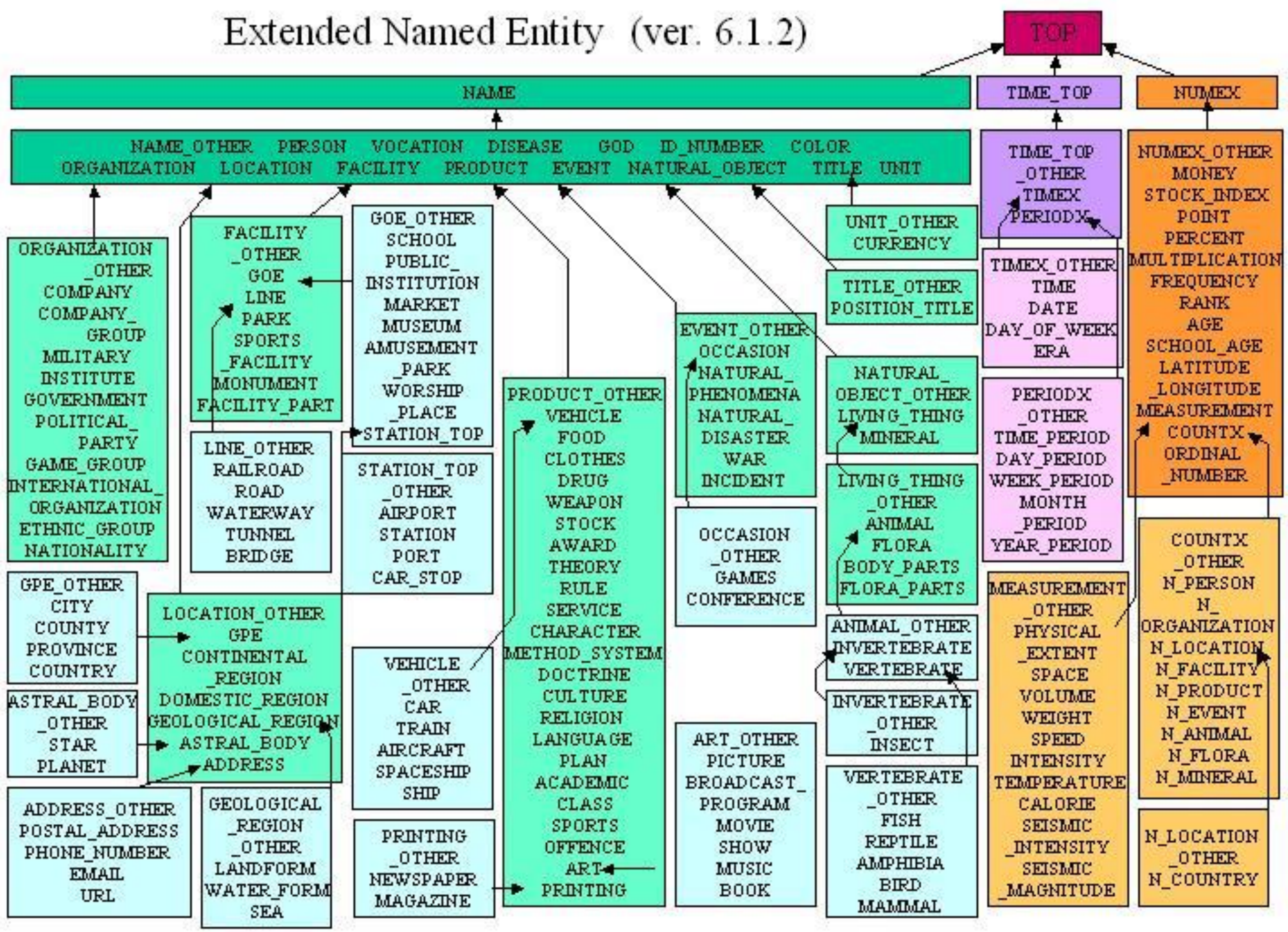
- Models trained on one corpus perform poorly on others

train	*F-score*		
	muc	*conll*	*bbn*
muc	82.3	54.9	69.3
conll	69.9	86.9	60.2
bnn	80.2	58.0	88.0

- Domain-specific:
  - Health: disease, drug, ward
  - Molecular biology: protein, gene, virus
  - Astronomy: galaxy, telescope, moon

2 Named Entity Recognition (NER)

Tagsets (Sekine et al. 2010)



https://nlp.cs.nyu.edu/ene/version7_1_0Beng.html

3 Traditional NER

Three standard approaches to NER (and IE)

- Rule-based NER
- Statistical-based NER
- Sequence Model for NER

Traditional Approaches

Traditional NER

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 ⇒ per
  - X, location ⇒ loc or org
  - travel verb to X ⇒ loc
- Effectively regular expressions

3 Traditional NER

Rule-based NER

- Determining which person holds what office in what organization
  - [person] , [office] of [org]
    - Vuk Draskovic, leader of the Serbian Renewal Movement
  - [org] (named, appointed, etc.) [person] Prep [office]
    - NATO appointed Wesley Clark as Commander in Chief
- Determining where an organization is located
  - [org] in [loc]
    - NATO headquarters in Brussels
  - [org] [loc] (division, branch, headquarters, etc.)
    - KFOR Kosovo headquarters

3 Traditional NER

Statistical approaches are more portable

- Learn NER from annotated text
  - weights (≈ rules) calculated from the corpus
  - same machine learner, different language or domain

- Token-by-token classification
- 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.)

- N-gram model:

$$tn = \underset{t \in T}{\arg\max}\ p(t | wn, wn-1, wn-2)$$

3 Traditional NER

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

Traditional NER

Traditional NER Approaches - Pros and Cons

Rule-based approaches

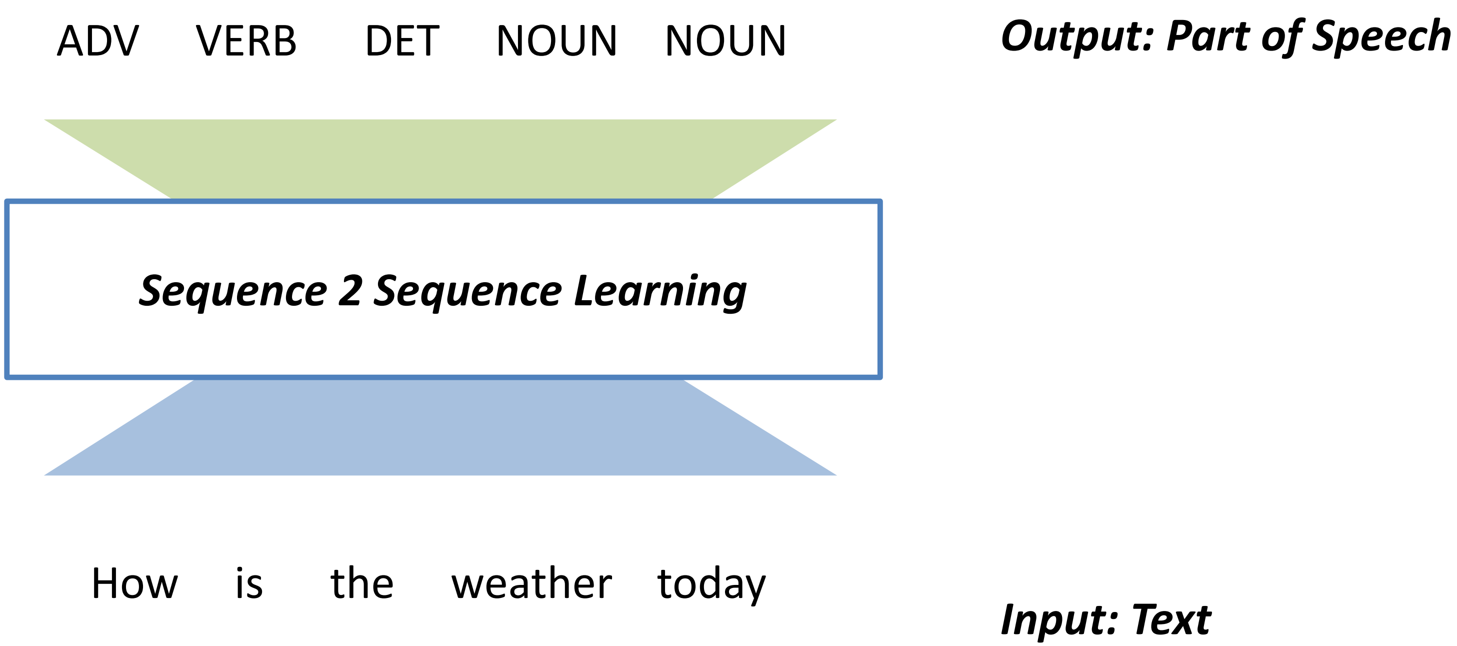
- Can be high-performing and efficient
- Require experts to make rules
- Rely heavily on gazetteers that are always incomplete
- Are not robust to new domains and languages

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

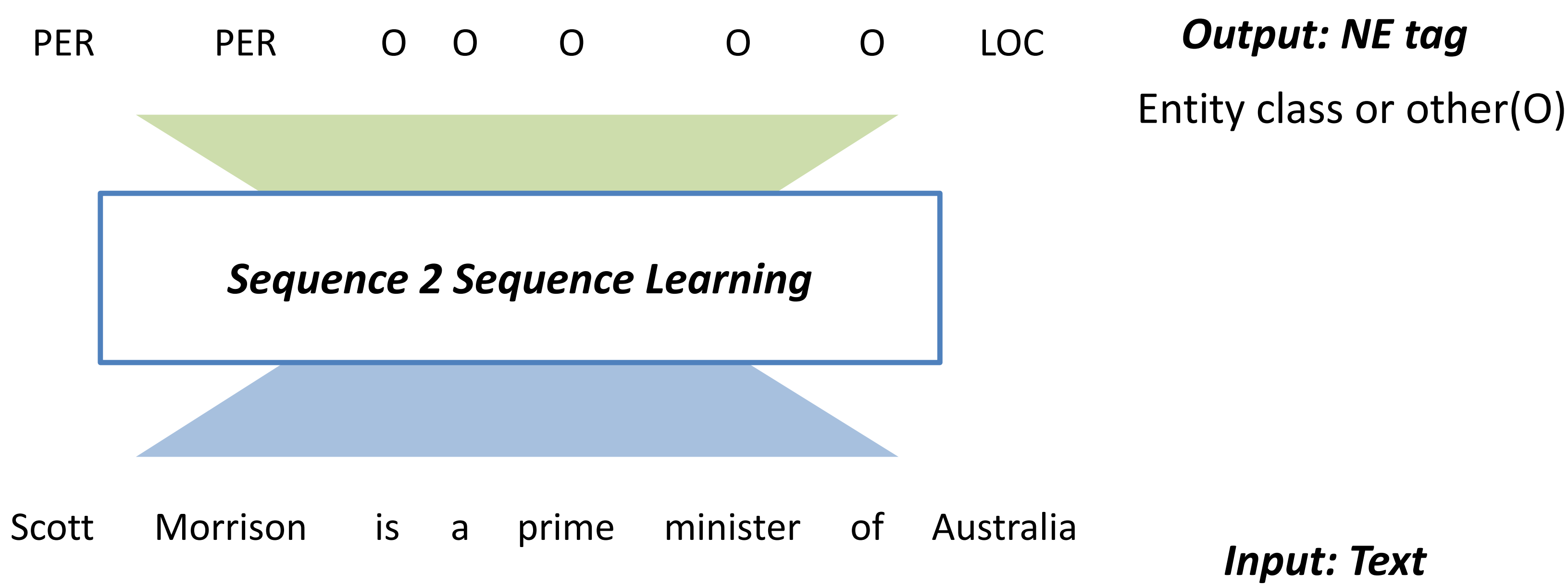
Sequence Model for NER

Sequence Model



4 Sequence Model for NER

Sequence Model



4 Sequence Model for NER

Encoding classes for sequence labeling

	Josiah	tells	Caren	John	Smith	is	a	student	
IO encoding	PER	O	PER	PER	PER	O	O	O	**n+1**
IOB encoding	B-PER	O	B-PER	B-PER	I-PER	O	O	O	**2n+1**
		even	B-PER	I-PER	I-PER				

IO encoding vs IOB encoding

- *Computation Time?*
- *Efficiency?*

4 Sequence Model for NER

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

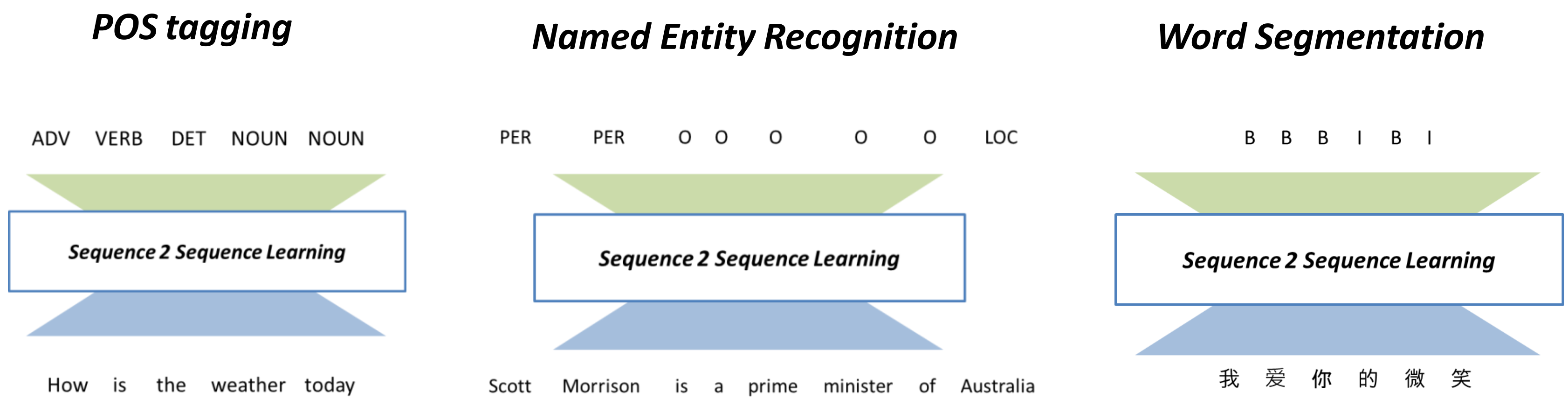
Label context

- Previous (and perhaps next) label

4 Sequence Model for NER

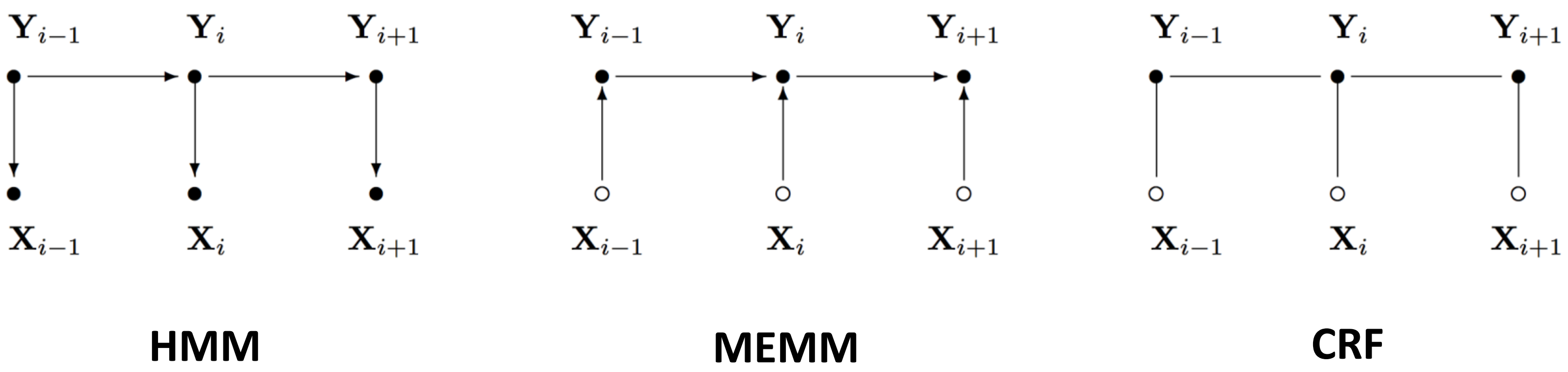
N to N Sequence model

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



Sequence Model for NER

Sequence Model (MEMM, CRF)



Sequence Model for NER

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

-3	-2	-1	0	+1
Scott	Morrison	lives	in	Australia
NN	**NN**	**VBZ**	**IN**	**NN**

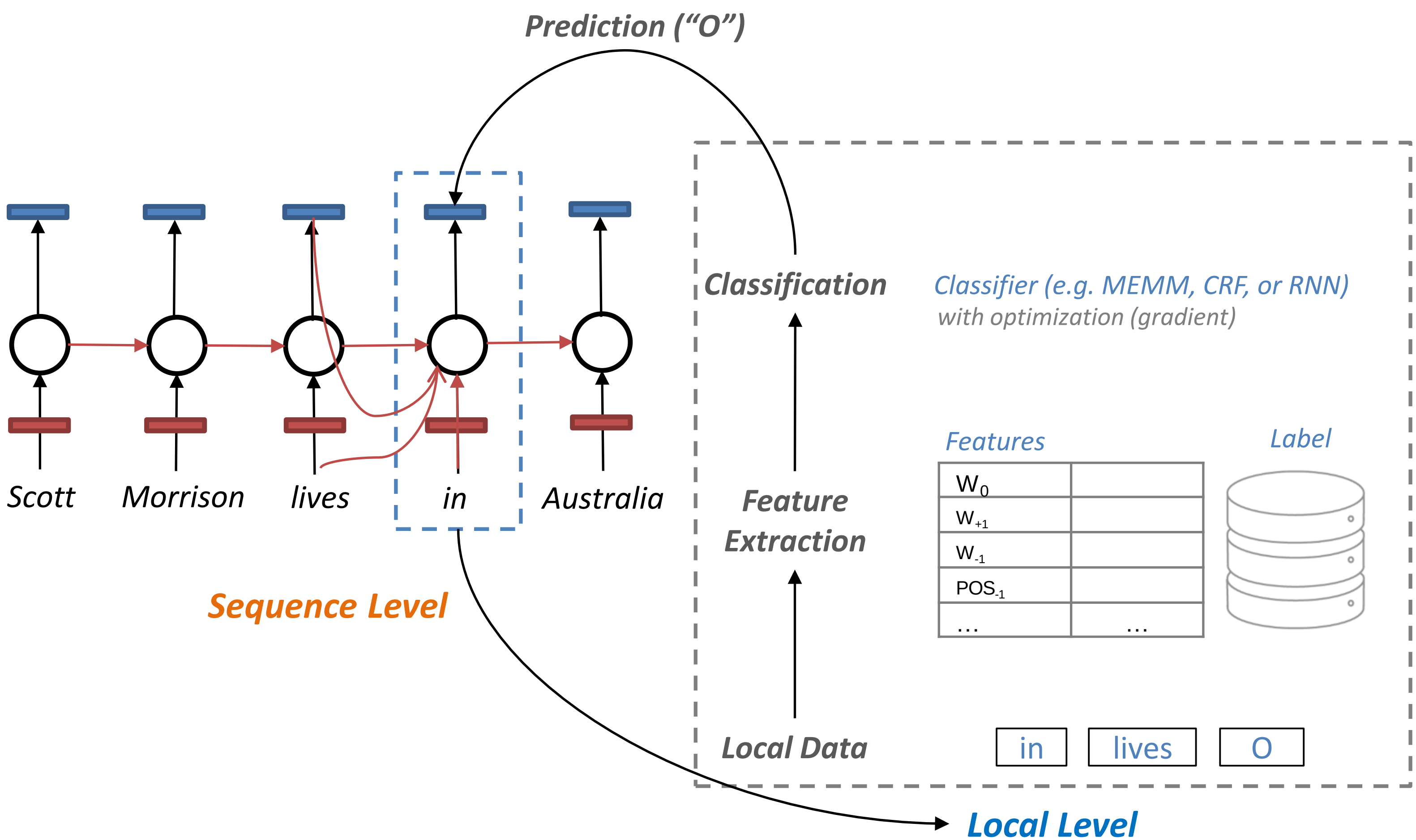
Features

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

(Toutanova et al. 2003, etc.)

4 Sequence Model for NER

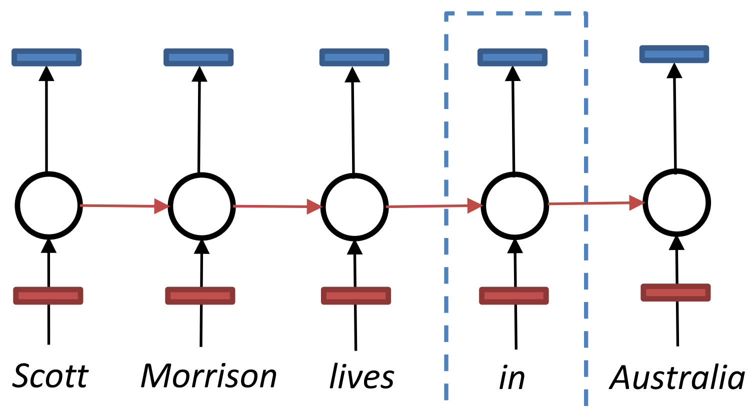
Sequence Inference for NER



Sequence Model for NER

Greedy Inference

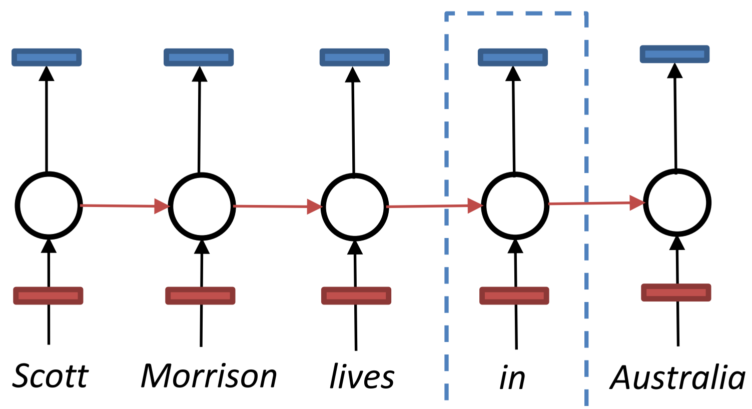
- Greedy inference:
  - We just start at the left, and use our classifier at each position to assign a label
  - The classifier can depend on previous labeling decisions as well as observed data
- Advantages:
  - Fast, no extra memory requirements
  - Very easy to implement
  - With rich features including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



4 Sequence Model for NER

Beam Inference

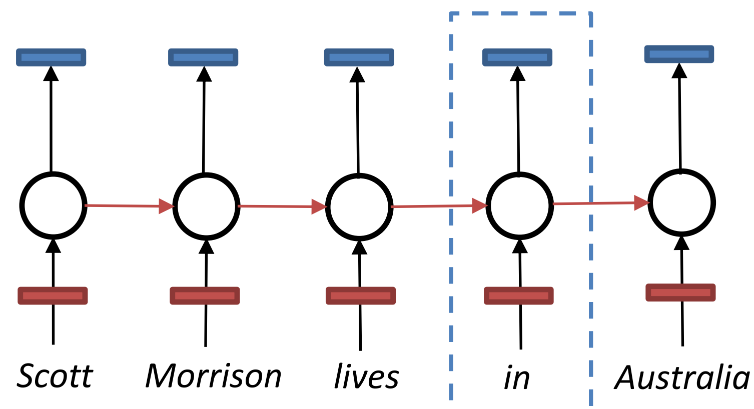
- Beam inference:
  - At each position keep the top k complete sequences.
  - Extend each sequence in each local way.
  - The extensions compete for the k slots at the next position.
- Advantages:
  - Fast; beam sizes of 3–5 are almost as good as exact inference in many cases.
  - Easy to implement (no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



4 Sequence Model for NER

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



Coreference Resolution

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**

Coreference Resolution

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,” according to a new interview.

5 Coreference Resolution

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,” according to a new interview.

5 Coreference Resolution

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,” according to a new interview.

5 Coreference Resolution

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.

5 Coreference Resolution

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!***

5

Coreference Resolution

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," according to a new interview.

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," according to a new interview.

5 Coreference Resolution

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***

5 Coreference Resolution

Coreference vs Anaphora

Coreference

Donald Trump

Trump

Anaphora

Donald Trump

he

Coreference Resolution

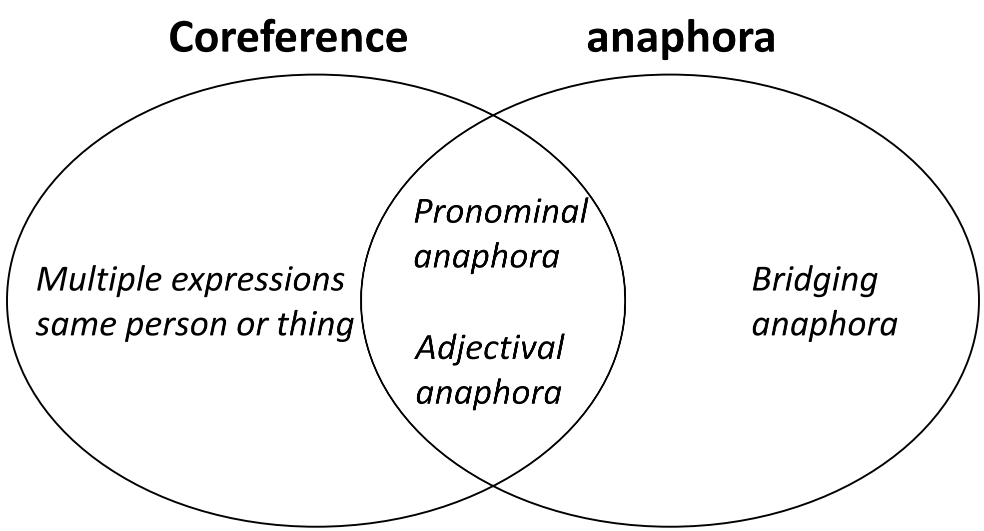
Not all anaphoric relations are coreferential

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**…

6 Coreference Model

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

Coreference Model

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

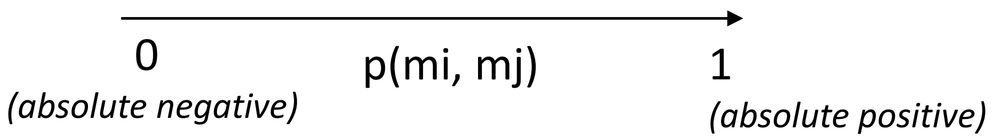
Gold cluster 1

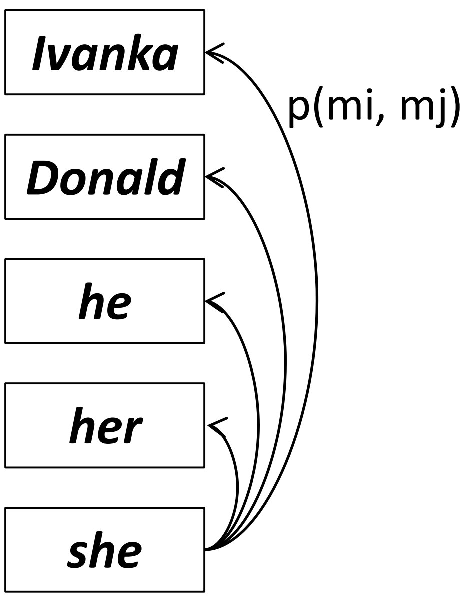
Gold cluster 2

6 Coreference Model

How to Cluster Mentions?

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





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

Coreference Model

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)

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

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

Iterate through mentions

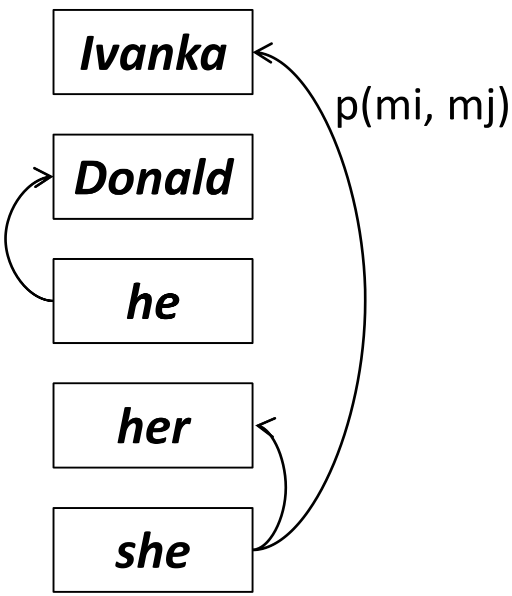
Iterate through candidate antecedents (previously occurring mentions)

6

Coreference Model

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(mi, mj) is above the threshold

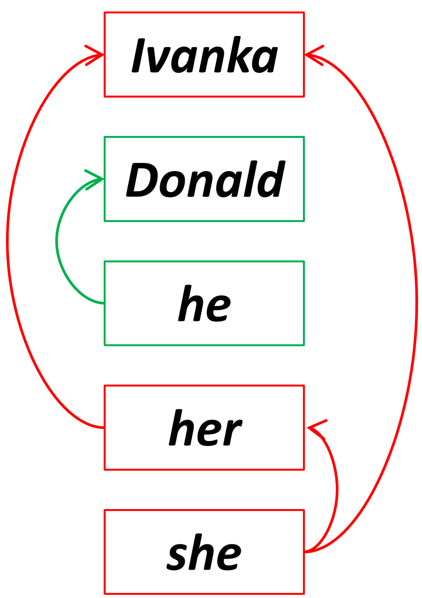


Ivanka was happy that **Donald** said **he** considered nominating **her** because **she** is very good with numbers

6 Coreference Model

Mention Pair Testing

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



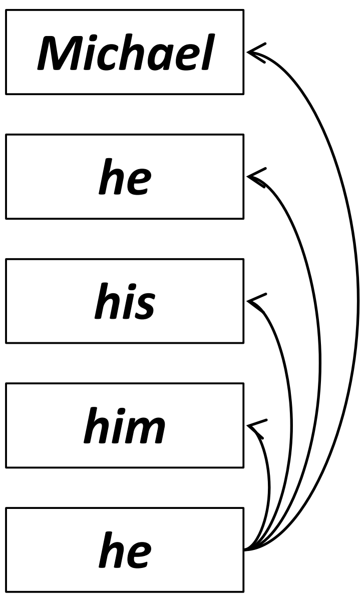
Ivanka was happy that **Donald** said **he** considered nominating **her** because **she** is very good with numbers

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

6 Coreference Model

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

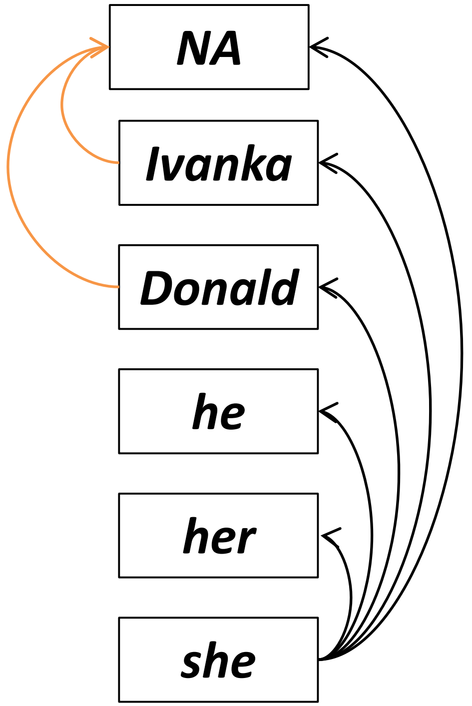
Alternative solution: instead train the model to predict only one antecedent for each mention

Mention Ranking

Coreference Model

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)

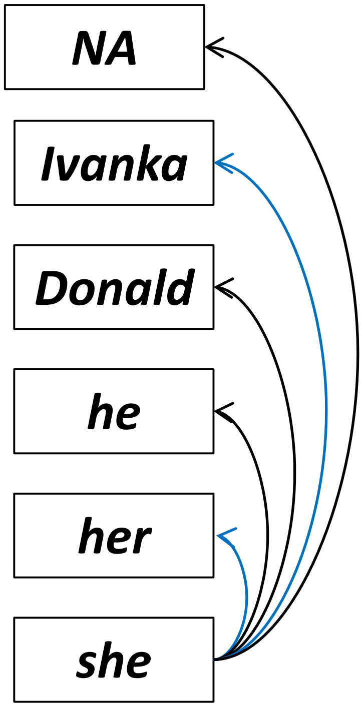


What can be the best antecedent for **she**?

Coreference Model

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?

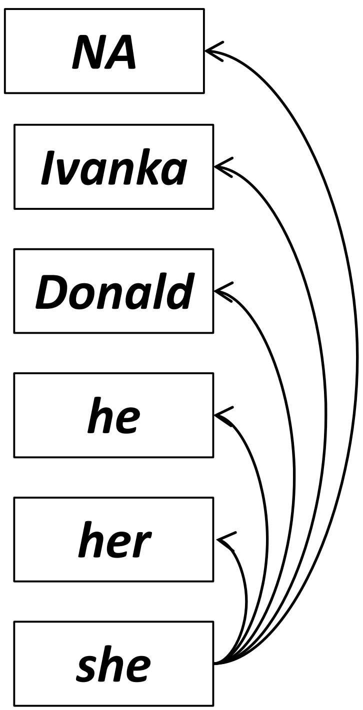
Positive examples: model has to assign a high probability to either one (but not necessarily both)

6

Coreference Model

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?

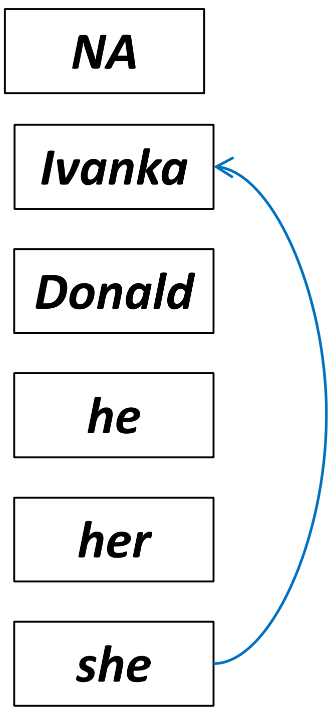
Apply a **softmax** over the scores for candidate antecedents so probabilities sum to 1

- p(NA, she) = 0.1
- p(Ivanka, she) = 0.5
- p(Donald, she) = 0.1
- p(he, she) = 0.1
- p(her, she) = 0.2

6 Coreference Model

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(NA, she) = 0.1
- **p(Ivanka, she) = 0.5**
- p(Donald, she) = 0.1
- p(he, she) = 0.1
- p(her, she) = 0.2

only add highest scoring coreference link

Coreference Model

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:

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

Iterate through candidate antecedents (previously occurring mentions)

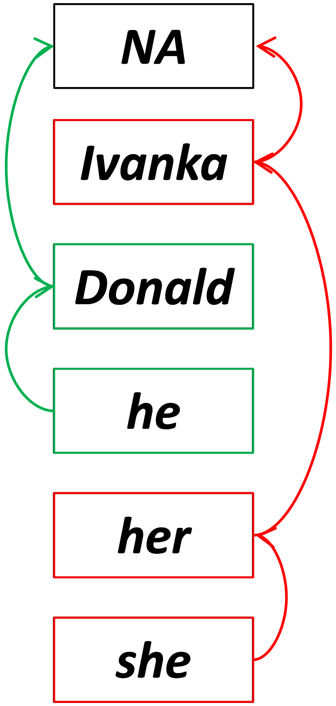
For ones that are coreferent to m_j...

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

Coreference Model

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

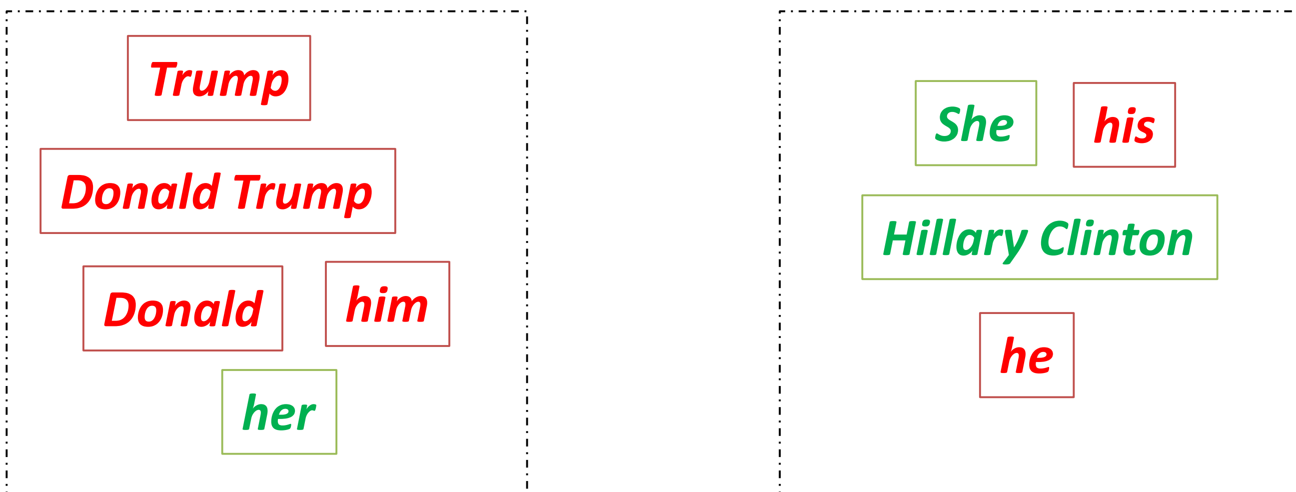
Coreference Evaluation

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

Predicted Cluster 1

Predicted Cluster 2



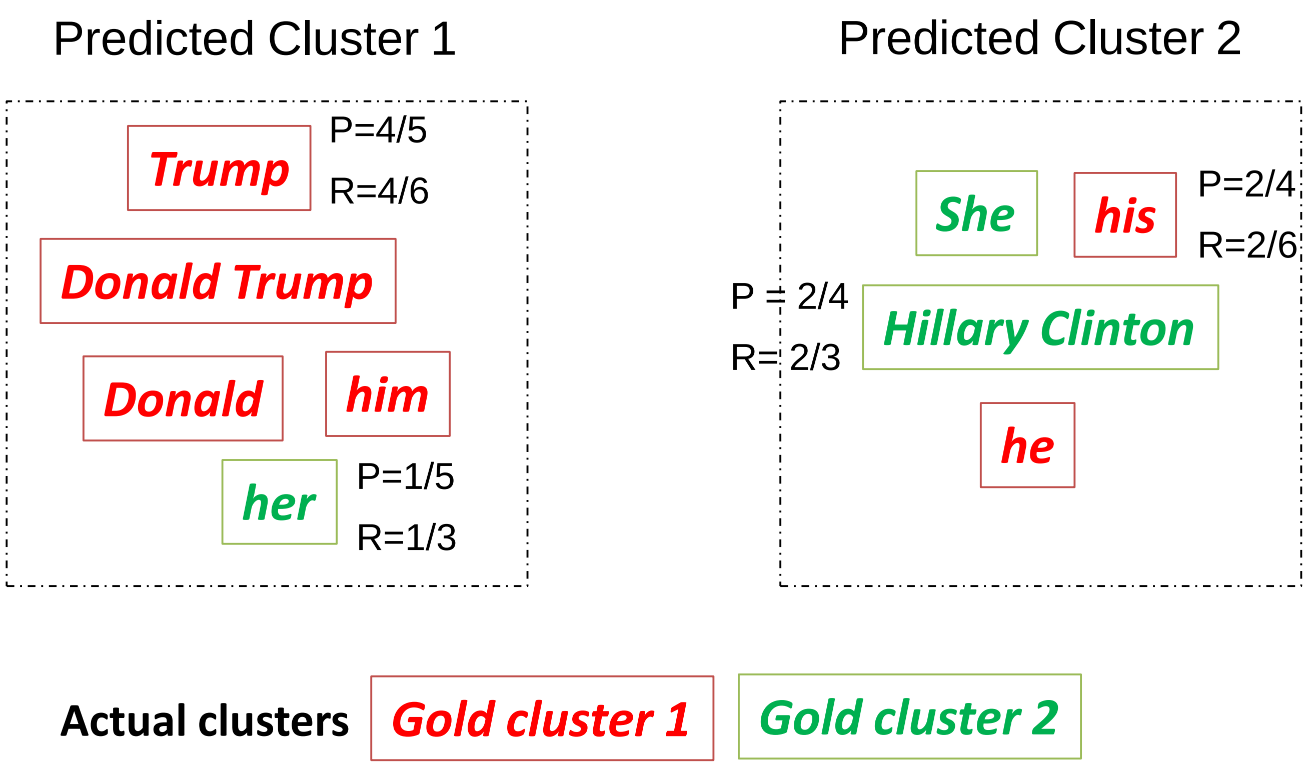
Actual clusters *Gold cluster 1* *Gold cluster 2*

Coreference Evaluation

How to evaluate coreference?

Let's evaluate with B-CUBED metrics

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

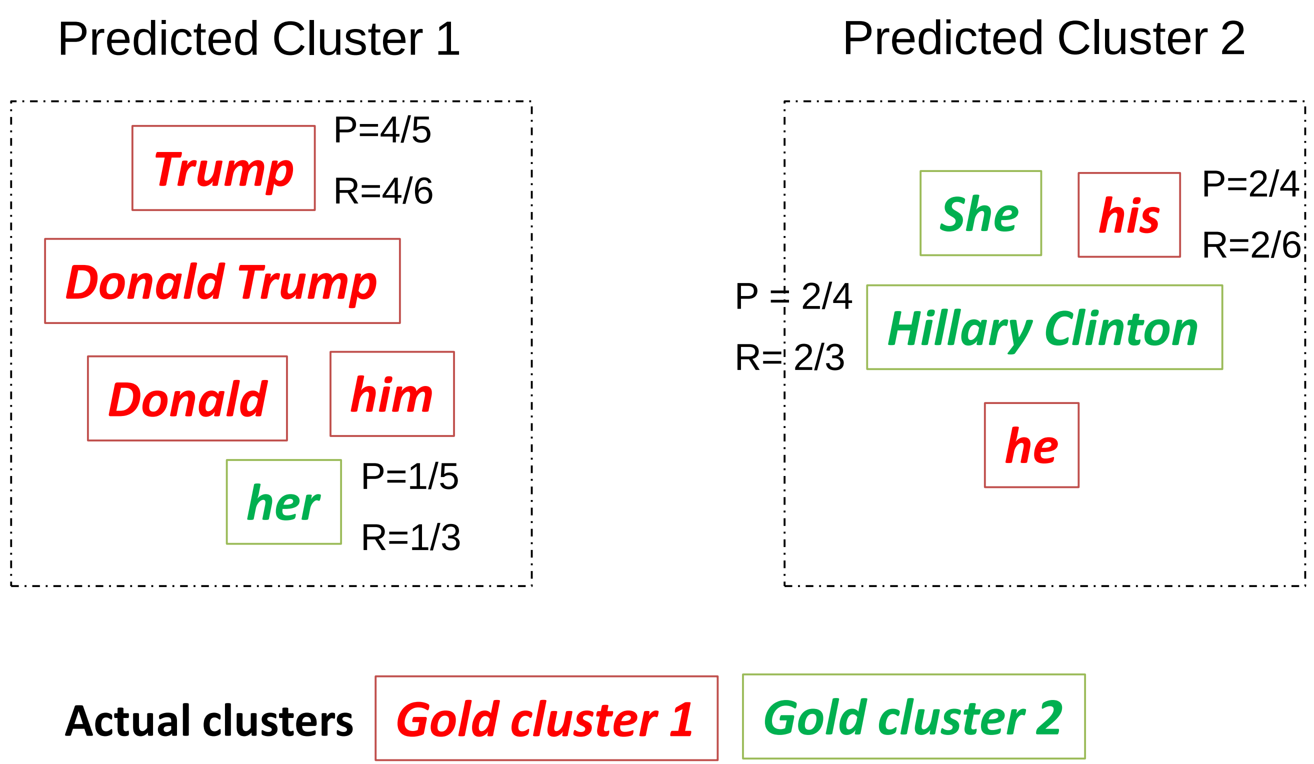


Coreference Evaluation

How to evaluate coreference?

Let’s evaluate with B-CUBED metrics

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



Coreference Evaluation

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	--

/ Reference

THE UNIVERSITY OF SYDNEY

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

- Finkel, J. R., Grenager, T., & Manning, C. (2005, June). Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 363-370). Association for Computational Linguistics.
- Clark, K., & Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models. arXiv preprint arXiv:1609.08667.
- Clark, K., & Manning, C. D. (2016). Improving coreference resolution by learning entity-level distributed representations. arXiv preprint arXiv:1606.01323.
- Lee, H., Recasens, M., Chang, A., Surdeanu, M., & Jurafsky, D. (2012, July). Joint entity and event coreference resolution across documents. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 489-500). Association for Computational Linguistics.