Natural Language Processing with Deep Learning

CS224N/Ling284

Christopher Manning and Richard Socher

Lecture 15: Coreference Resolution

Lecture Plan

1. What is coreference? A worked example (25 mins)
2. Kinds of reference and anaphora (10 mins)
3. *Research highlight: Summarizing Source Code*
4. Introduction to coreference resolution (15 mins)
5. Neural coreference resolution (Clark and Manning 2016) (20m)

Reminders/comments:

Keep working on final projects and Assignment 4

Dynamic memory networks on Thu very useful to Ass4

CodaLab videos up; Ass4 leaderboard; one dev submit/day

What is Coreference Resolution ?

- Identify all noun phrases (**mentions**) that refer

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

What is Coreference Resolution ?

- Identify all noun phrases (**mentions**) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

What is Coreference Resolution ?

- Identify all noun phrases (**mentions**) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

What is Coreference Resolution ?

- Identify all noun phrases (**mentions**) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday

had foreign affairs experience

What is Coreference Resolution ?

- Identify all noun phrases (**mentions**) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

What is Coreference Resolution ?

- Identify all noun phrases (**mentions**) that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chos
had foreign affairs experience as a forr

A couple of years later, Vanaja met Akhila at the local park. Akhila's son Prajwal was just two months younger than her son Akash, and they went to the same school. For the pre-school play, Prajwal was chosen for the lead role of the naughty child Lord Krishna. Akash was to be a tree. She resigned herself to make Akash the best tree that anybody had ever seen. She bought him a brown T-shirt and brown trousers to represent the tree trunk. Then she made a large cardboard cutout of a tree's foliage, with a circular opening in the middle for Akash's face. She attached red balls to it to represent fruits. It truly was the nicest tree.

From The Star by Shruthi Rao, with some shortening.

Reference Resolution

- Noun phrases refer to entities in the world, many pairs of noun phrases co-refer, some nested inside others

John Smith, CFO of Prime Corp. since 1986,

saw his pay jump 20% to $1.3 million

as the 57-year-old also became

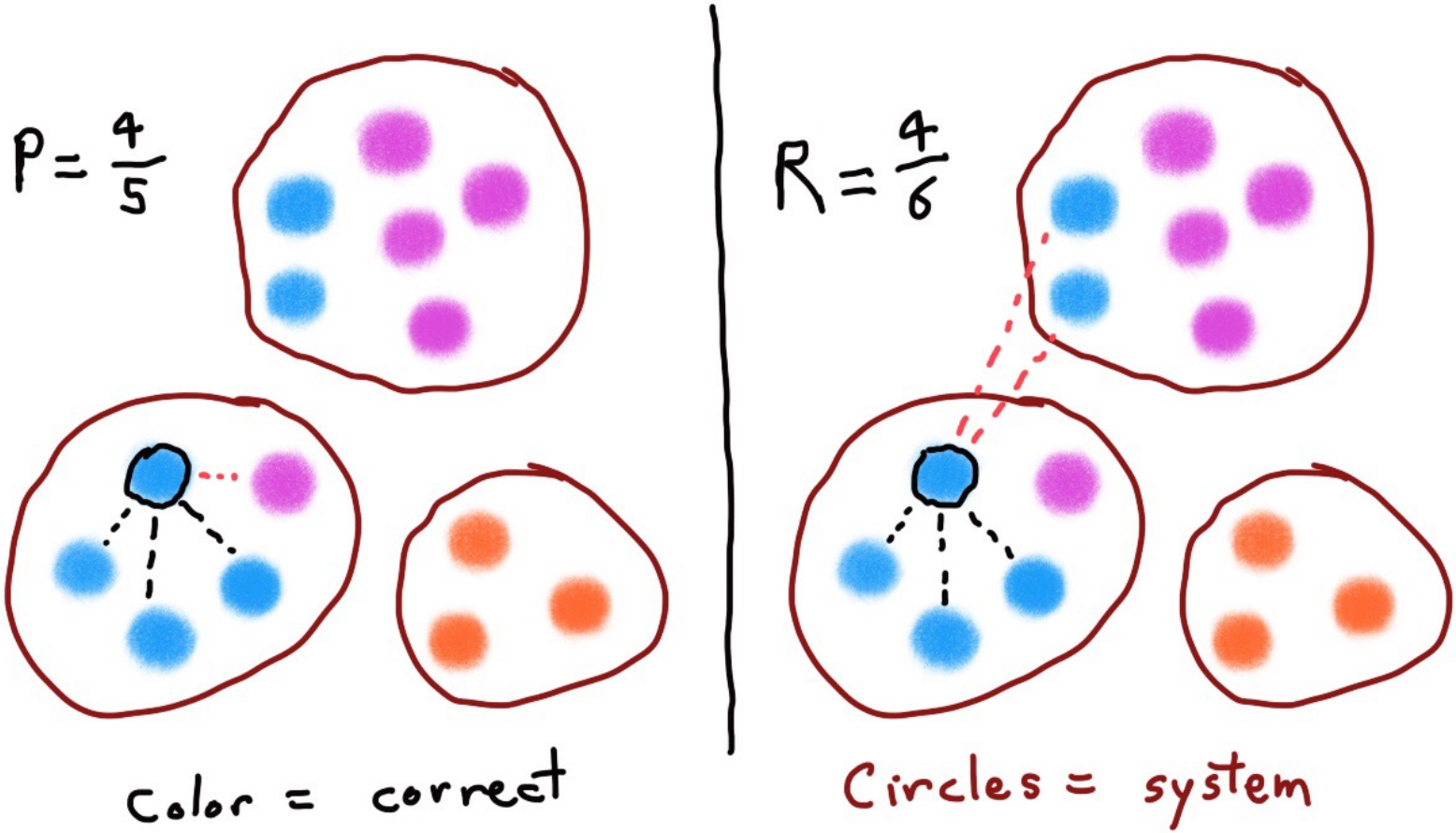
the financial services co.'s president.

Applications

- Full text understanding:
  - understanding an extended discourse
- Machine translation (if languages have different features of gender, number, etc.)
- Text summarization, including things like web snippets
- Tasks like information extraction and question answering
  - Correctly answering often involves resolving anaphora
  - *He married Claudia Ross in 1971.*

Coreference Evaluation

B^3 (B cubed) evaluation metric



Coreference Evaluation

- B^3 (B-CUBED) algorithm for evaluation
  - Precision & recall for *entities in a reference chain*
  - Precision: % of elements in a hypothesized reference chain that are in the true reference chain
  - Recall: % of elements in a true reference chain that are in the hypothesized reference chain
  - Overall precision & recall are the (size-weighted) average of per-chain precision & recall
  - Optimizing chain-chain pairings is a hard problem
    - In the computational NP-hard sense
  - Greedy matching is done in practice for evaluation

Other coreference evaluation metrics

- MUC Score (Vilain et al., 1995)
  - Link based: Counts the number of common links and computes f-measure. Sort of defective.
- CEAF (Luo 2005); entity based
- BLANC (Recasens and Hovy 2011) Cluster RAND-index
- …

- All of them are sort of evaluating getting coreference links/clusters right and wrong, but the differences can be important….

Kinds of Reference

- Referring expressions
  - *John Smith*
  - *President Smith*
  - *the president*
  - *the company's new executive*

More common in newswire, generally harder in practice

- Free variables
  - Smith saw *his pay* increase
- Bound variables
  - The dancer hurt *herself.*

More interesting grammatical constraints, more linguistic theory, easier in practice

"anaphora resolution"

Not all NPs are referring!

- *Every dancer* twisted *her knee*.
- *No dancer* twisted *her knee.*

- There are three NPs in each of these sentences; because the first one is non-referential, the other two aren't either.

Coreference, anaphors, cataphors

- Coreference is when two mentions refer to the same entity in the world
- The relation of anaphora is when a term (anaphor) refers to another term (antecedent) and the interpretation of the anaphor is in some way determined by the interpretation of the antecedent

  ... and traditionally the antecedent came first

Cataphora

"From the corner of the divan of Persian saddle-bags on which he was lying, smoking, as was his custom, innumerable cigarettes, Lord Henry Wotton could just catch the gleam of the honey-sweet and honey-coloured blossoms of a laburnum…"

(Oscar Wilde – The Picture of Dorian Gray)

Cataphora

"From the corner of the divan of Persian saddle-bags on which he was lying, smoking, as was his custom, innumerable cigarettes, Lord Henry Wotton could just catch the gleam of the honey-sweet and honey-coloured blossoms

laburnum…"

(Oscar Wilde –

Anaphora vs. coreference

- Not all anaphoric relations are coreferential

We went to see a concert last night. The tickets were really expensive.

- This is referred to as bridging anaphora.

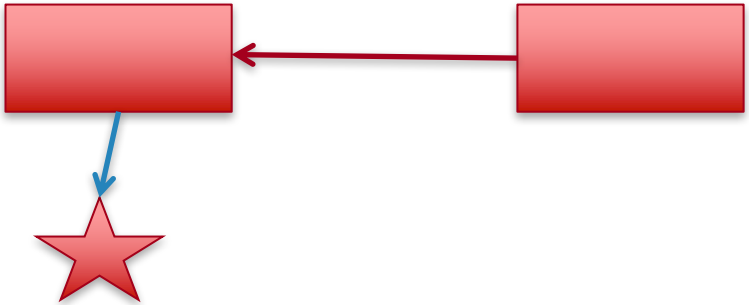
- Conversely, multiple identical full NP references are typically coreferential but not anaphoric.

Two different things…

- Anaphora

  Text

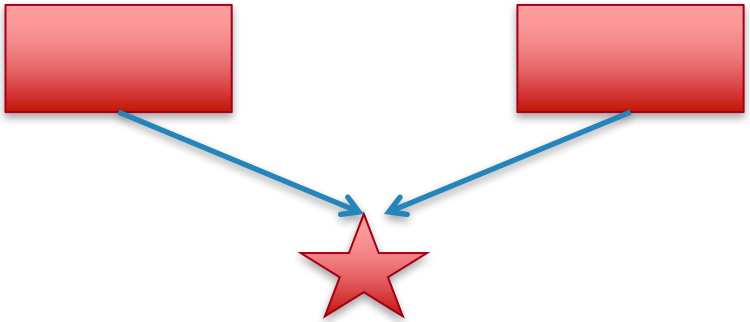
  World



- (Co)Reference

  Text

  World



Two different things…

- Something you might like to think about:
  - Do various models treat these two cases the same or differently?
  - Should we do more to treat them differently?

Summarizing Source Code using a Neural Attention Model

Srinivasan Iyer, Ionnis Konstas, Alvin Cheung, Luke Zettlemoyer
University of Washington CSE
In proceedings of ACL ‘16

Task and Dataset

Task: Generate sentences to describe C# code snippets and SQL queries.

Dataset: StackOverflow posts and responses tagged with C#, SQL, database, or oracle.

[Cleaning] *Remove posts where the title/question text is irrelevant to the code.*

[Parsing] *Replace literals, table/column names and remove inline comments.*

1. Source Code (C#):

```
public int TextWidth(string text) {
  TextBlock t = new TextBlock();
  t.Text = text;
  return
    (int)Math.Ceiling(t.ActualWidth);
}
```

Descriptions:

a. Get rendered width of string rounded up to the nearest integer

b. Compute the actual textwidth inside a textblock

3. Source Code (SQL):

```
SELECT Max(marks) FROM stud_records
  WHERE marks <
  (SELECT Max(marks) FROM stud_records);
```

Descriptions:

a. Get the second largest value of a column

b. Retrieve the next max record in a table

Task and Methodology

1. **Text generation**

   Given an input code sequence, generate a sentence that maximizes the scoring function.

2. **Information retrieval**

   Given an input question, find the code snippet in the corpus that maximizes the scoring function.

Probability of current word in summary given previous words.

$$s(c, n) = \prod_{i=1}^{l} p(n_i | n_1, \ldots, n_{i-1})$$

$$p(n_i | n_1, \ldots, n_{i-1}) \propto \mathbf{W} \tanh(\mathbf{W_1 h_i} + \mathbf{W_2 t_i})$$

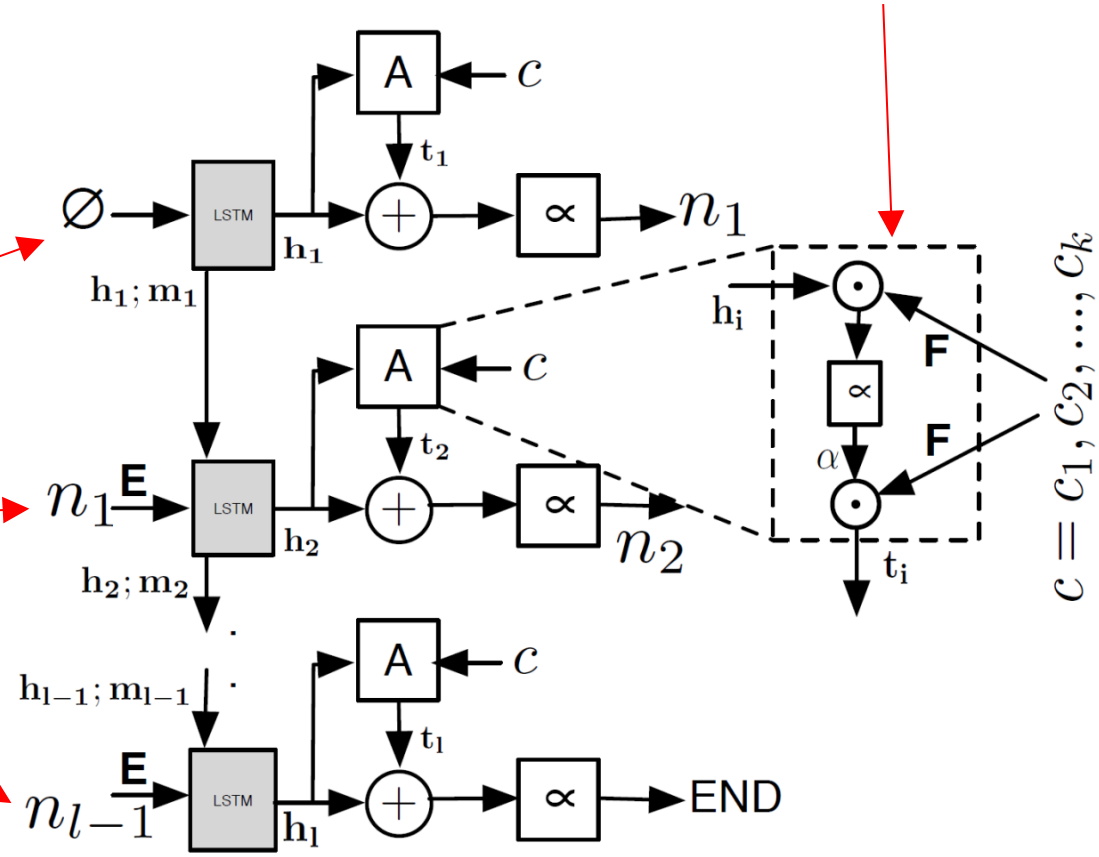
Hidden state from LSTM

Attention on source code

CODE-NN Text Generation

[2] Attention over code token embeddings

[1] Natural Language token embeddings



Evaluation

1. **Text generation**

   MT metrics: METEOR, BLEU

   User study

	Model	METEOR	BLEU-4
C#	IR	7.9 (6.1)	13.7 (12.6)
	MOSES	9.1 (9.7)	11.6 (11.5)
	SUM-NN	10.6 (10.3)	19.3 (18.2)
	CODE-NN	**12.3 (13.4)**	**20.5 (20.4)**
SQL	IR	6.3 (8.0)	13.5 (13.0)
	MOSES	8.3 (9.7)	15.4 (15.9)
	SUM-NN	6.4 (8.7)	13.3 (14.2)
	CODE-NN	**10.9 (14.0)**	**18.4 (17.0)**

Table 3: Performance on EVAL for the GEN task. Performance on DEV is indicated in parentheses.

	Model	Naturalness	Informativeness
C#	IR	3.42	2.25
	MOSES	1.41	2.42
	SUM-NN	**4.61***	1.99
	CODE-NN	**4.48**	**2.83**
SQL	IR	3.21	2.58
	MOSES	2.80	2.54
	SUM-NN	4.44	2.75
	CODE-NN	**4.54**	**3.12**

Table 4: Naturalness and Informativeness measures of model outputs. Stat. sig. between CODE-NN and others is computed with a 2-tailed Student's t-test; $p < 0.05$ except for *.

Evaluation

1. **Text generation**

   MT metrics: METEOR, BLEU

   User study

2. **Information retrieval**

   Mean Reciprocal Rank (MRR)

	Model	MRR
C#	RET-IR	0.42 ± 0.02 (0.44 ± 0.01)
	CODE-NN	$\mathbf{0.58 \pm 0.01}$ $(\mathbf{0.66 \pm 0.02})$
SQL	RET-IR	0.28 ± 0.01 (0.4 ± 0.01)
	CODE-NN	$\mathbf{0.44 \pm 0.01}$ $(\mathbf{0.54 \pm 0.02})$

Table 5: MRR for the RET task. Dev set results in parentheses.

	Model	MRR
L to C	Allamanis	0.182 ± 0.009
	CODE-NN	$\mathbf{0.590 \pm 0.044}$
C to L	Allamanis	0.434 ± 0.003
	CODE-NN	$\mathbf{0.461 \pm 0.046}$

Table 6: MRR values for the Language to Code (L to C) and the Code to Language (C to L) tasks using the C# dataset of Allamanis et al. (2015b)

Example Outputs

C# code	`foreach (string pTxt in xml.parent) {`
	`  TreeNode parent = new TreeNode();`
	`  foreach (string cTxt in xml.child) {`
	`  TreeNode child = new TreeNode();`
	`  parent.Nodes.Add(child);`
	`  }`
	`}`

Gold	Adding childs to a treenode dynamically in C#
IR	How to set the name of a tabPage programmatically
MOSES	How can TreeView nodes from XML parentText string to a treeview node
SUM-NN	How to get data from xml file in C#
CODE-NN	How to get all child nodes in TreeView?

Example Outputs

SQL Query	`SELECT * FROM table` `ORDER BY Rand() LIMIT 10`
Gold	Select random rows from mysql table
IR	How to select a random record from a mysql database?
MOSES	How to select all records in mysql ?
SUM-NN	How can I select random rows from a table
CODE-NN	How to get random rows from a mysql database?

Traditional pronominal anaphora resolution: Hobbs' naive algorithm

1. Begin at the NP immediately dominating the pronoun
2. Go up tree to first NP or S. Call this X, and the path p.
3. Traverse all branches below X to the left of p, left-to-right, breadth-first. Propose as antecedent any NP that has a NP or S between it and X
4. If X is the highest S in the sentence, traverse the parse trees of the previous sentences in the order of recency. Traverse each tree left-to-right, breadth first. When an NP is encountered, propose as antecedent. If X not the highest node, go to step 5.

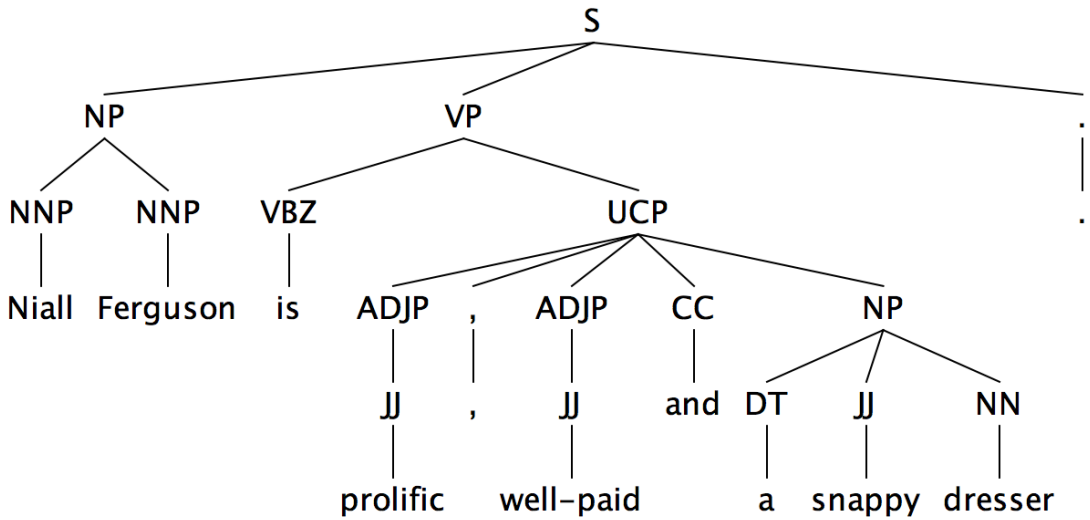
Hobbs' naive algorithm (1976)

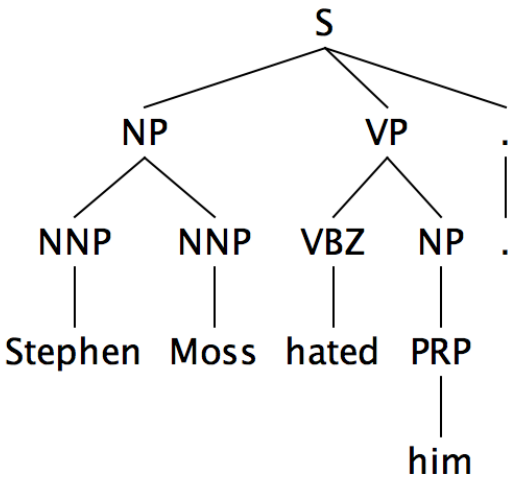
5. From node X, go up the tree to the first NP or S. Call it X, and the path p.
6. If X is an NP and the path p to X came from a non-head phrase of X (a specifier or adjunct, such as a possessive, PP, apposition, or relative clause), propose X as antecedent

   (The original said "did not pass through the N' that X immediately dominates", but the Penn Treebank grammar lacks N' nodes….)
7. Traverse all branches below X to the left of the path, in a left-to-right, breadth first manner. Propose any NP encountered as the antecedent
8. If X is an S node, traverse all branches of X to the right of the path but do not go below any NP or S encountered. Propose any NP as the antecedent.
9. Go to step 4

Actually still often used as a feature in ML systems!

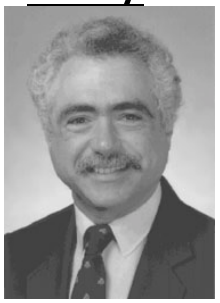
Hobbs Algorithm Example



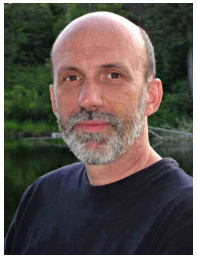


Knowledge-based Pronominal Coreference

- [The city council] refused [the women] a permit because they feared violence.
- [The city council] refused [the women] a permit because they advocated violence.
  - Winograd (1972)



- See: Hector J. Levesque "On our best behaviour" IJCAI 2013. http://www.cs.toronto.edu/~hector/Papers/ijcai-13-paper.pdf
  - Winograd Schema Challenge @ CommonSense 2015
    - http://commonsensereasoning.org/winograd.html



Hobbs' algorithm: commentary

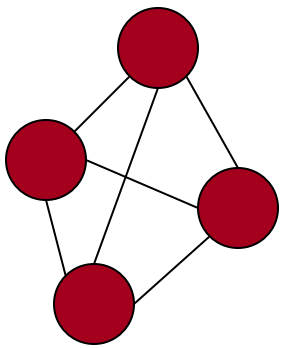
"… *the naïve approach is quite good. Computationally speaking, it will be a long time before a semantically based algorithm is sophisticated enough to perform as well, and these results set a very high standard for any other approach to aim for.*

"Yet there is every reason to pursue a semantically based approach. The naïve algorithm does not work. Any one can think of examples where it fails. In these cases it not only fails; it gives no indication that it has failed and offers no help in finding the real antecedent."

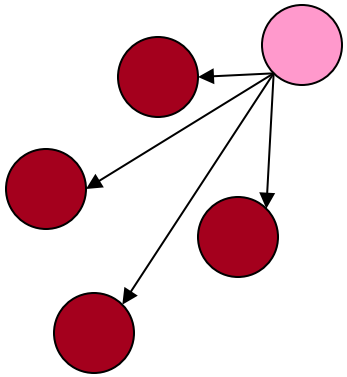
— Hobbs (1978), *Lingua,* p. 345

Kinds of Coreference Models

- Mention Pair models
  - Treat coreference chains as a collection of pairwise links
  - Make independent pairwise decisions
  - Reconcile them in some deterministic way (e.g., transitivity or greedy partitioning)



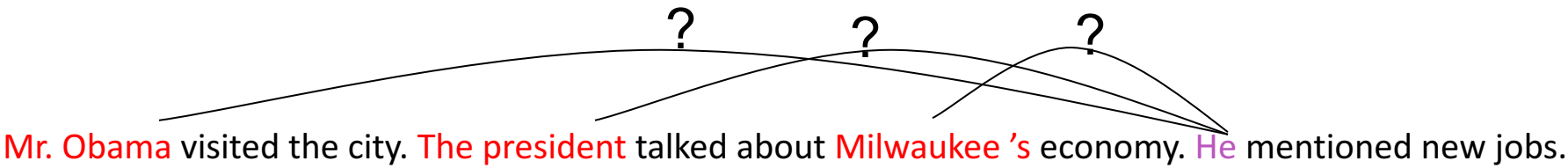
- Mention ranking models
  - Explicitly rank all candidate antecedents for a mention



- Entity-Mention models
  - A cleaner, but less studied, approach
  - Posit single underlying entities
  - Each mention is linked to a discourse entity [Pasula et al. 03], [Luo et al. 04]
  - Explicitly cluster mentions of the same discourse entity

Supervised Mention-Pair Model

- Given a mention and earlier mentions, classify whether the pronoun refers to each earlier entity or not given the surrounding context (yes/no)



- Use any classifier, obtain positive examples from training data, generate negative examples by pairing each mention with other (incorrect) mentions
- This is naturally thought of as a binary classification task

Features for Coreference Resolution

- Person/Number/Gender agreement
  - Jack gave Mary a gift. She was excited.
- Semantic compatibility
  - … the mining conglomerate … the company …
- Certain syntactic constraints
  - John bought him a new car. [him can not be John]
- More recently mentioned entities preferred for referenced
  - John went to a movie. Jack went as well. He was not busy.
- Grammatical Role: Prefer entities in the subject position
  - John went to a movie with Jack. He was not busy.
- Parallelism:
  - John went with Jack to a movie. Joe went with him to a bar.
- …

Neural Coreference Models

- Wiseman, Rush, Shieber, and Weston (ACL 2015)
  - Mention-pair model. Only partially neural network system over conventional, categorical coreference features
- Wiseman, Rush, and Shieber (NAACL 2016)
  - Uses RNNs to learn global representations of entity clusters from mentions
- Clark and Manning (ACL 2016)
  - An entity-mention model based around clustering using distributed representations of mentions and entity clusters
- Clark and Manning (EMNLP 2016)
  - Explores deep reinforcement learning to improve a mention-pair model

Deep Reinforcement Learning for Mention-Ranking Coreference Models (Clark & Manning 2016)

- Coreference resolution is a document-level structured prediction task

*"**I** voted for **Nader** because **he** was most aligned with **my** values," **she** said.*

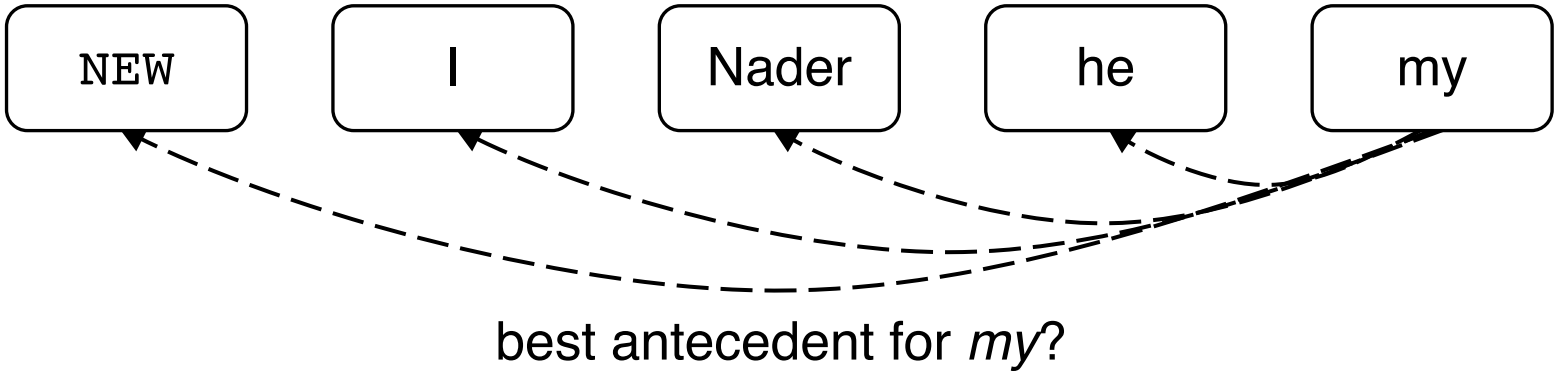
I	Nader	he	my	she

Coreference Cluster 1

Coreference Cluster 2

Mention-Ranking Models

- Dominant approach to coreference resolution in recent years
- Assign each mention its highest scoring candidate antecedent according to the model



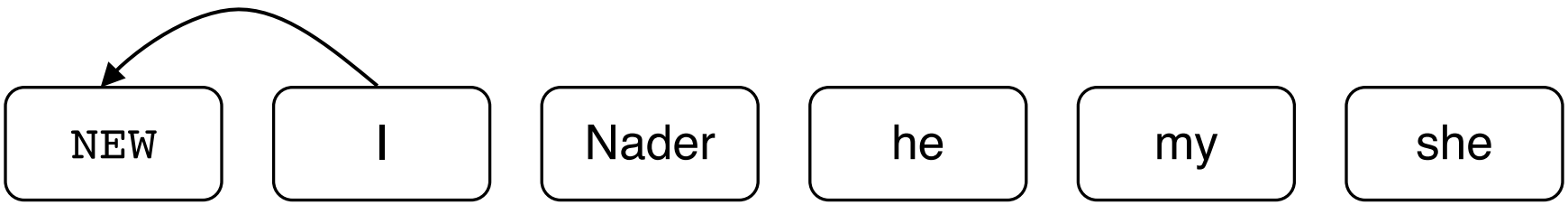
Mention-Ranking Models

- Infer global structure by making a sequence of local decisions

NEW	I	Nader	he	my	she

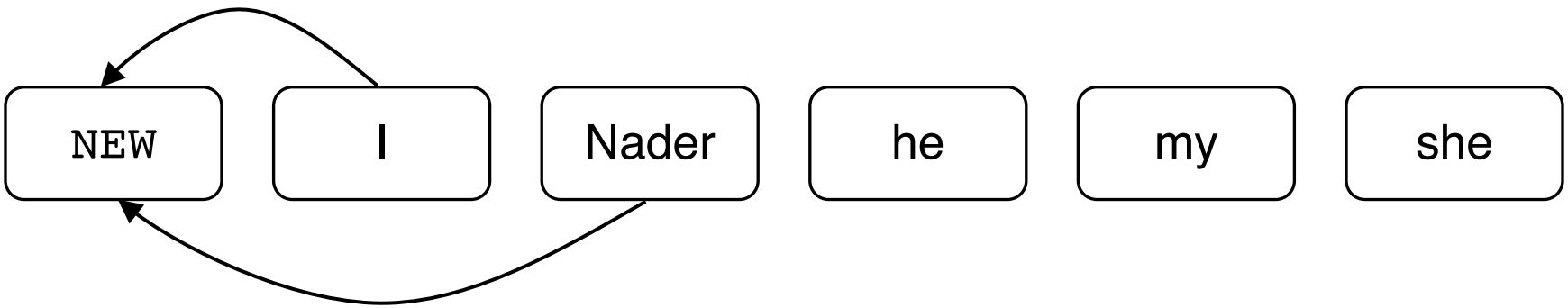
Mention-Ranking Models

- Infer global structure by making a sequence of local decisions



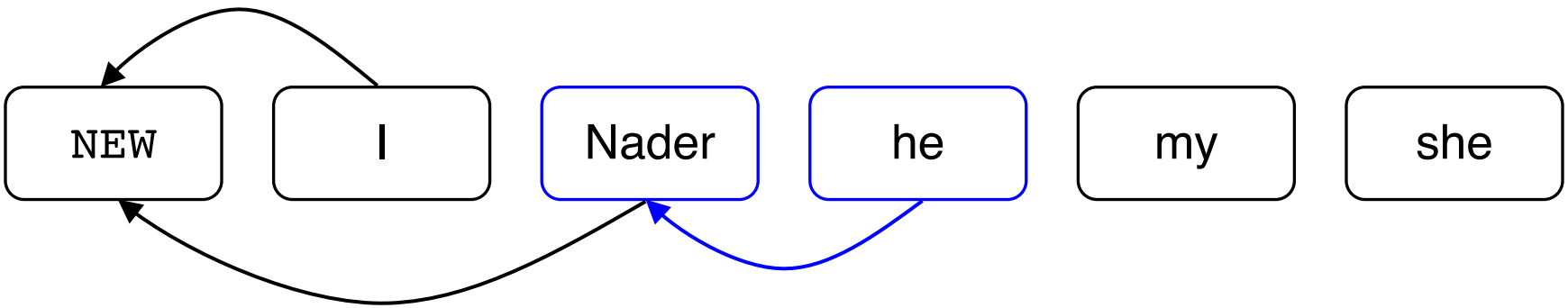
Mention-Ranking Models

- Infer global structure by making a sequence of local decisions



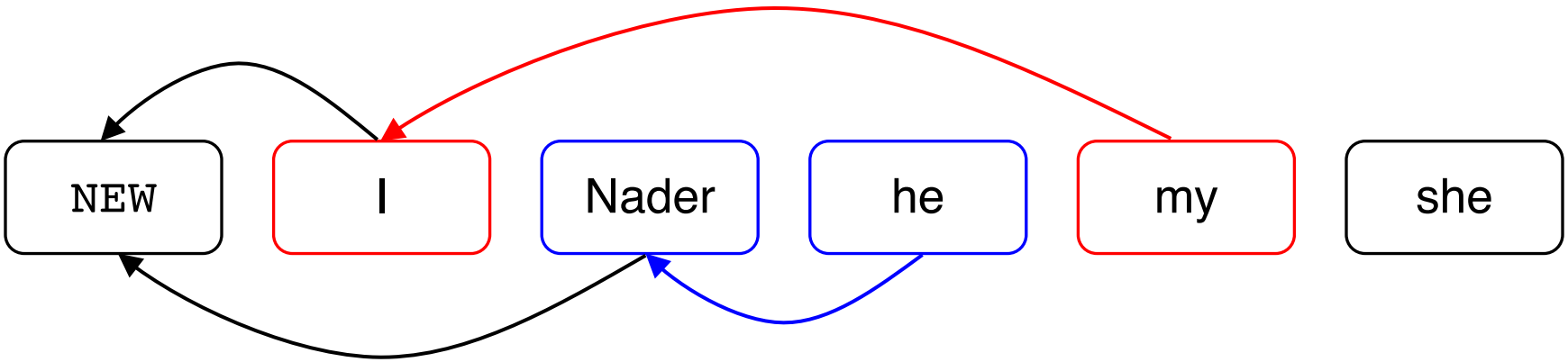
Mention-Ranking Models

- Infer global structure by making a sequence of local decisions



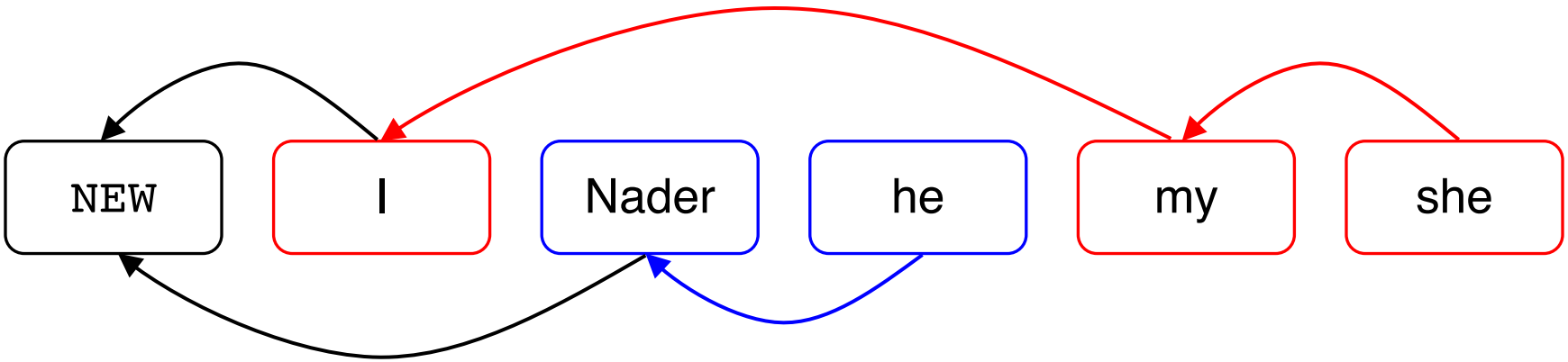
Mention-Ranking Models

- Infer global structure by making a sequence of local decisions



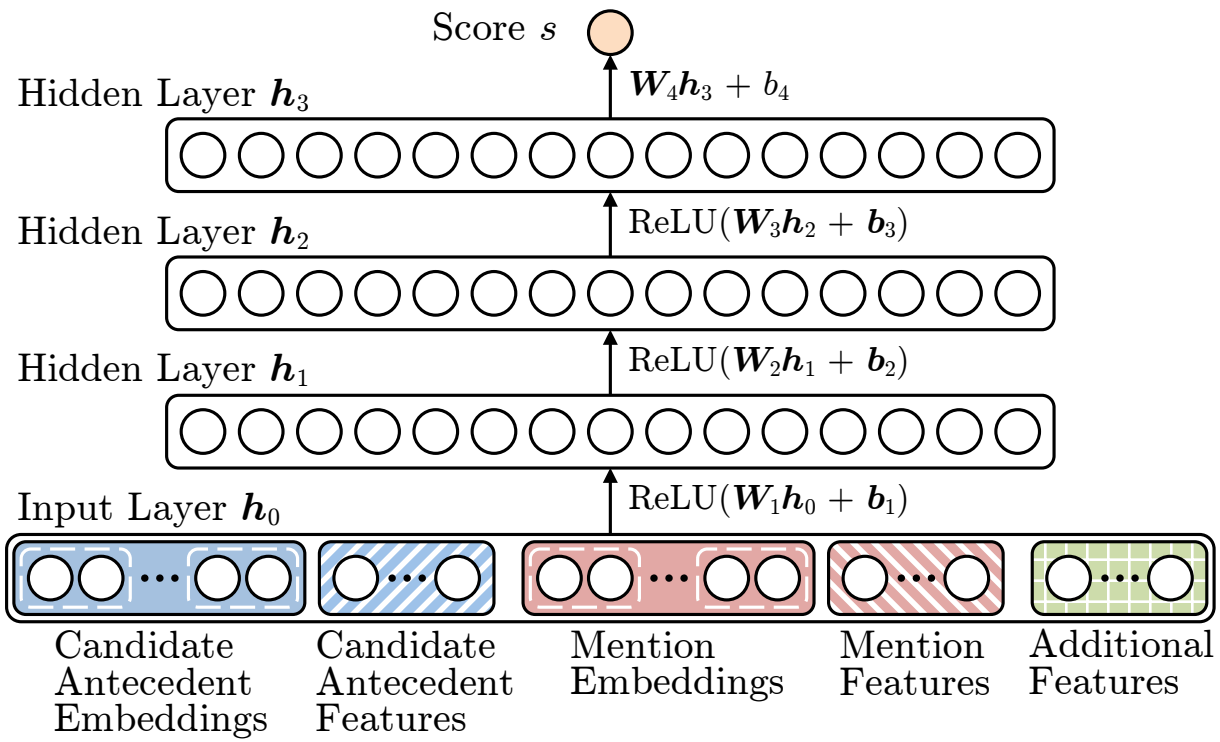
Mention-Ranking Models

- Infer global structure by making a sequence of local decisions



Neural Mention-Pair Model

- Standard feed-forward neural network
  - From (Clark and Manning, 2016); similar to Wiseman et al. (2015)
  - Input layer: word embeddings and a few categorical features



Mention-Pair Representations: Features

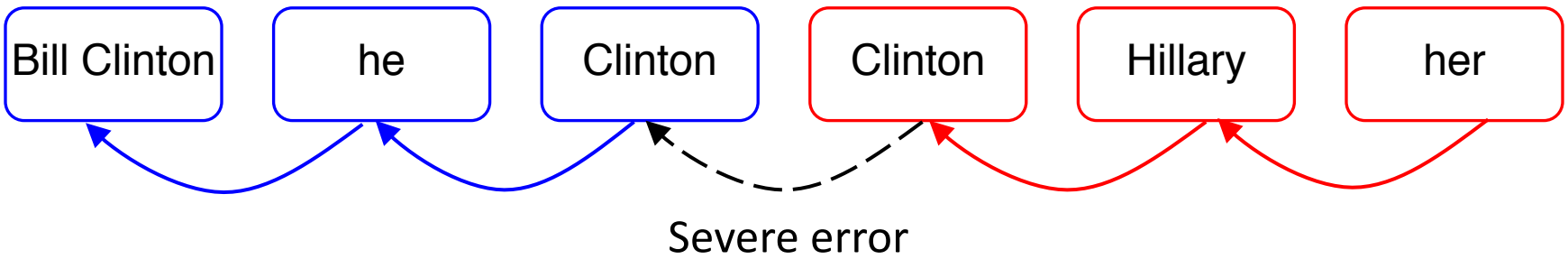
- Word embedding features (first word, head word, etc.)
- Small number of hand-crafted features (distance, speaker, string-matching, etc.)
  - Latter turn out to still be very important!

Less Hand-Engineering in Coreference Resolvers	
Stanford Deterministic System (Raghunathan et al., 2010)	Rule-based system
Stanford Statistical System (Clark and Manning, 2015)	>100 hand-crafted features
First Neural Coref System (Wiseman et al., 2015)	29 hand-crafted features
This work	13 hand-crafted features

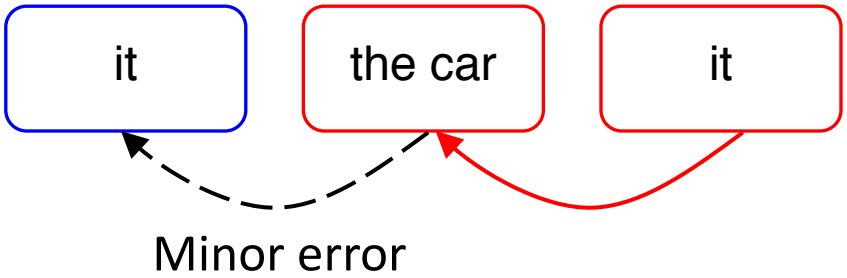
Neural coreference system details

- Uses pretrained word embeddings
- No RNNs: Just take certain words and use feed-forward network
- Use deep network (3 hidden layers)
- Dropout

Challenge: Some Local Decisions Matter More than Others



*"**it** was raining, but **the car** stayed dry because **it** was under cover"*



Reinforcement Learning to the Rescue!

- **Prior work:** heuristically defined the importance of a coreference decision
  - Requires careful tuning with hyperparameters

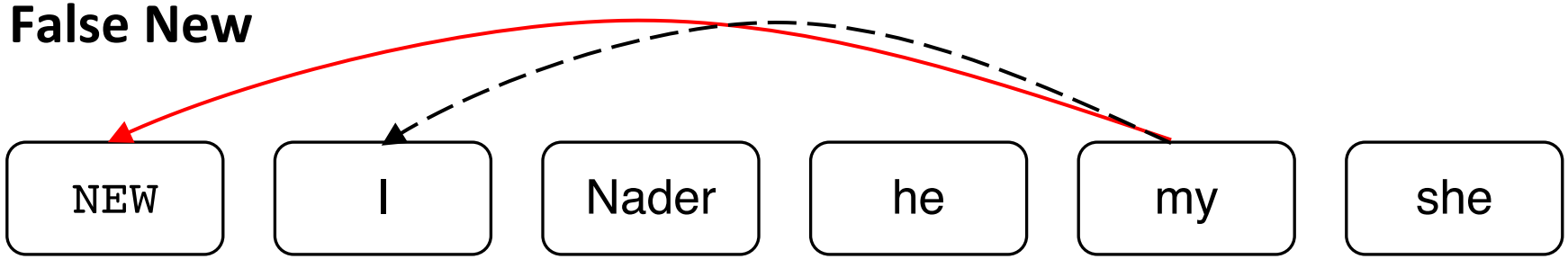
- **Clark and Manning (EMNLP 2016):** instead use RL to learn which local decisions lead to a good clustering
  - No pesky hyperparameter search
  - Small boost in accuracy

Prior Work: Error Types

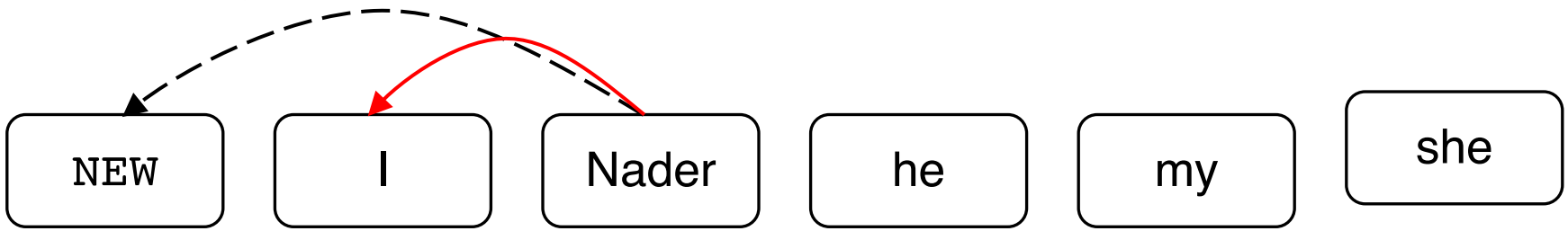
Correct Decision

Model Decision

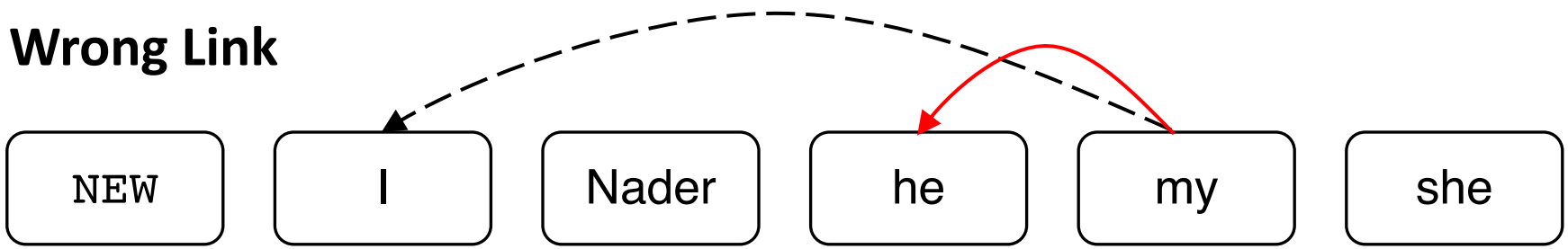
1. **False New**



2. **False Anaphoric**

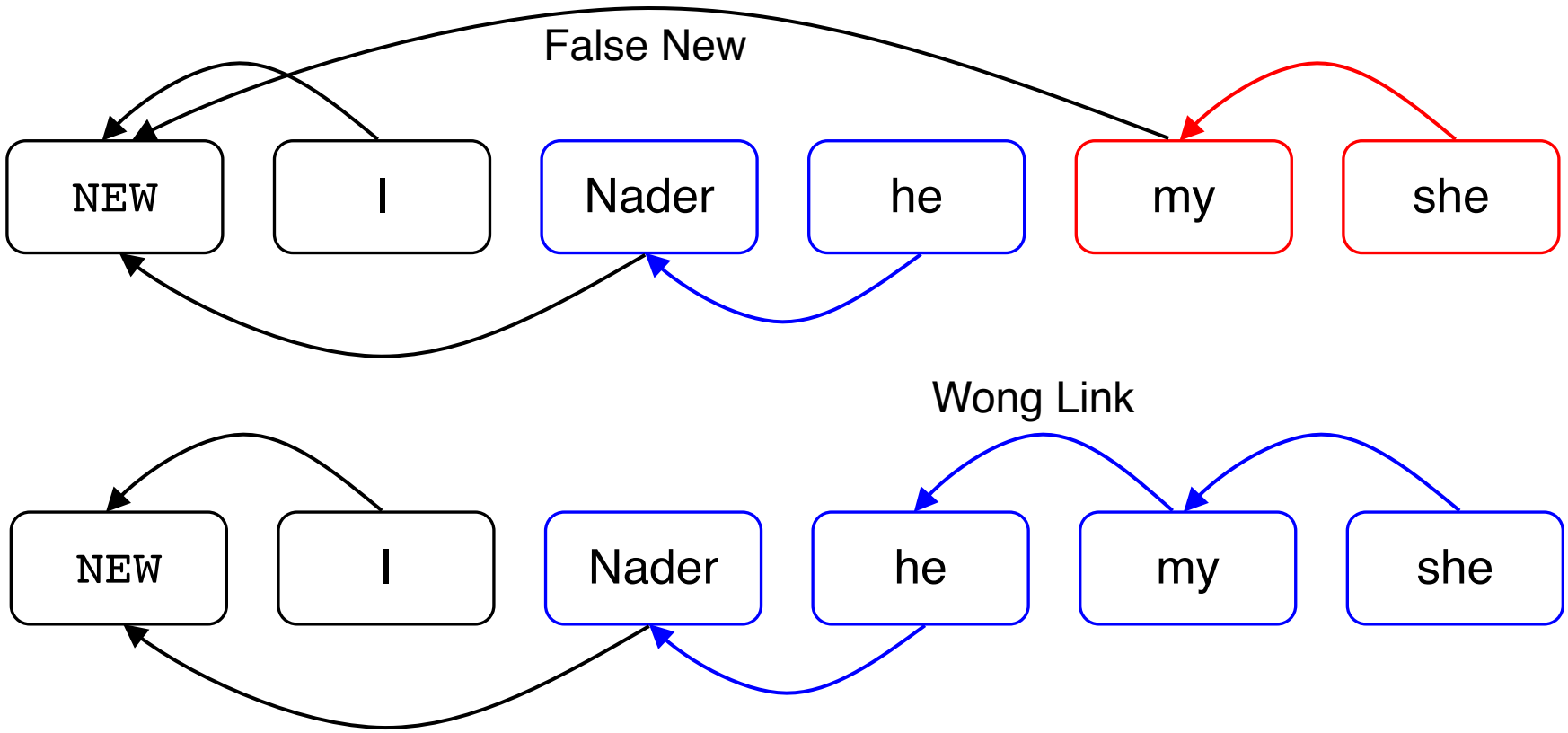


3. **Wrong Link**



Prior Work: Error Types

- Captures a bit about which coreference decisions are more important
  - e.g., a Wrong Link is worse than a False New



Prior Work: Heuristic Loss Function

- Max-margin loss

$$\max_{c \in \mathcal{C}(m_i)} \Delta_h(c, m_i)(1 + s(c, m_i) - s(\hat{t}_i, m_i))$$

max over candidate corefence decisions

cost for this coref decision

loss for scoring this decision too highly

- With heuristic cost function

$$\Delta_h(c, m_i) = \begin{cases} 0 & \text{if } c \text{ and } m_i \text{ are coreferent} \\ \alpha_{\text{FN}} & \text{if false new error} \\ \alpha_{\text{FA}} & \text{if false anaphoric error} \\ \alpha_{\text{WL}} & \text{if wrong link error} \end{cases}$$

Prior Work: Heuristic Loss Function

- Heuristic error costs are widely used

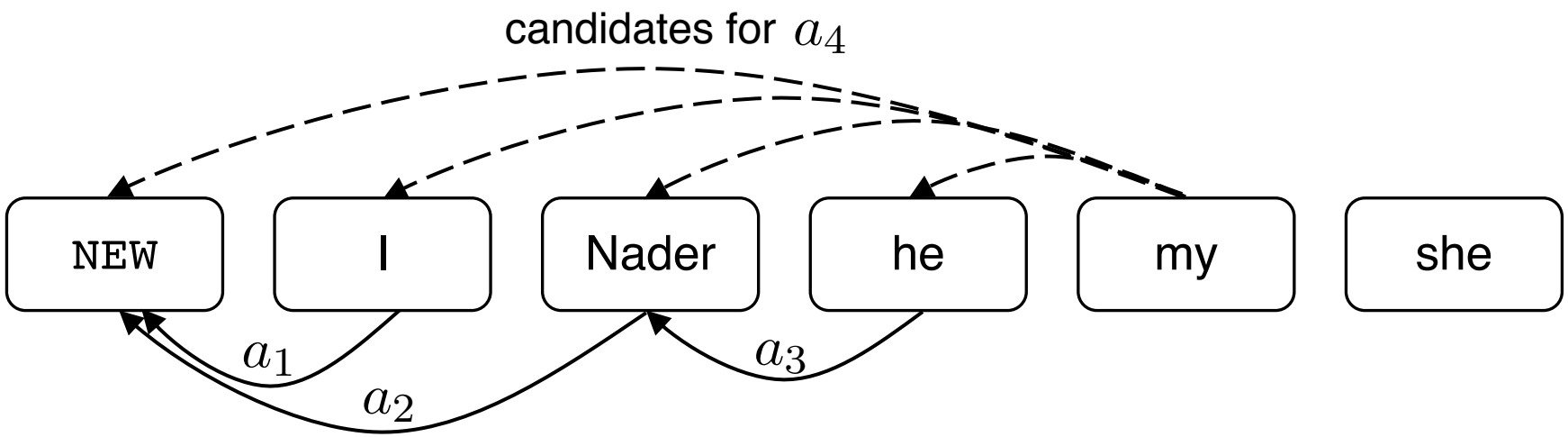
  (Fernandes et al., 2012; Durrett et al., 2013; Durrett and Klein., 2013; Björkelund and Kuhn, 2014; Wiseman et al., 2015; Martschat and Strube, 2015; Wiseman et al., 2016)

- Grid search over $(\alpha_{\mathrm{FN}}, \alpha_{\mathrm{FA}}, \alpha_{\mathrm{WL}})$ requires training many instances of the same model
  - Has to be redone for different languages/datasets/metrics

- At best, loss is correlated with evaluation metrics

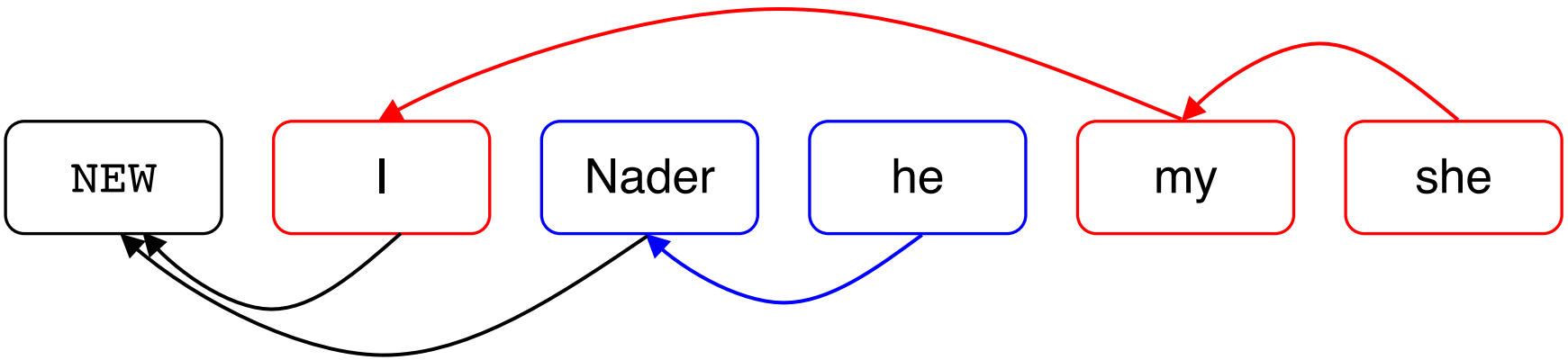
Coreference Resolution with Reinforcement Learning

- Model takes a sequence of actions $a_{1:T} = a_1, a_2, ..., a_T$
  - an action $a_i = (c, m_i)$ adds a coreference link between the *i*th mention and a candidate antecedent c



Coreference Resolution with Reinforcement Learning

- After completing a sequence of actions, the model receives a reward (B^3 coreference metric in our case)



$$R(a_{1:5}) = 100$$

Coreference Resolution with Reinforcement Learning

- Clark and Manning (2016) explore using two RL methods:
  1. REINFORCE algorithm (Williams, 1992)
  2. Reward-rescaling (novel)

1. The REINFORCE Algorithm

- Define probability distribution over actions:

  $p_\theta((c,m)) \propto e^{s(c,m)}$ for any action $a = (c,m)$

- Maximize expected reward

  $$J(\theta) = \mathbb{E}_{[a_{1:T} \sim p_\theta]} R(a_{1:T})$$

  - Sample trajectories of actions to approximate gradient

1. The REINFORCE Algorithm

- Competitive with the heuristic loss
- But a small disadvantage over heuristic loss: REINFORCE maximizes performance in expectation
  - Really we only need the highest scoring action to be correct

- **Reward Rescaling:** incorporate rewards into the max-margin objective's slack rescaling

$$\Delta_h(c, m_i) = \begin{cases} 0 & \text{if } c \text{ and } m_i \text{ are coreferent} \\ \alpha_{\text{FN}} & \text{if false new error} \\ \alpha_{\text{FA}} & \text{if false anaphoric error} \\ \alpha_{\text{WL}} & \text{if wrong link error} \end{cases}$$

1. The REINFORCE Algorithm

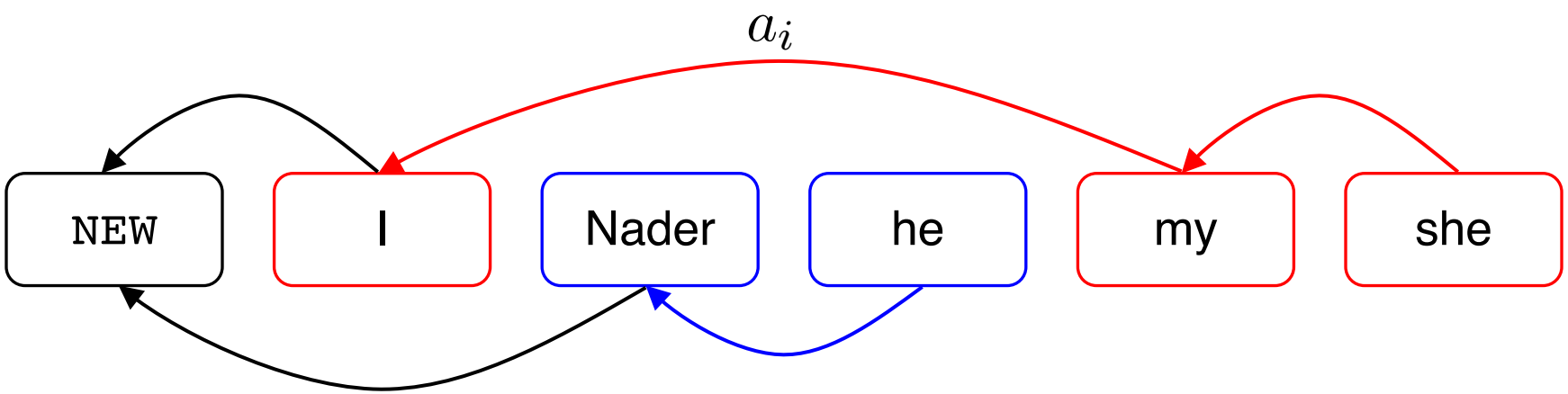
- Competitive with the heuristic loss
- But a small disadvantage over heuristic loss: REINFORCE maximizes performance in expectation
  - Really we only need the highest scoring action to be correct

- **Reward Rescaling:** incorporate rewards into the max-margin objective's slack rescaling

$$\Delta_h(c, m_i) = \begin{cases} 0 & \text{if } c \text{ and } m_i \text{ are coreferent} \\ \alpha_{\text{FN}} & \text{if false new error} \\ \alpha_{\text{FA}} & \text{if false anaphoric error} \\ \alpha_{\text{WL}} & \text{if wrong link error} \end{cases}$$

2. Reward-Rescaling

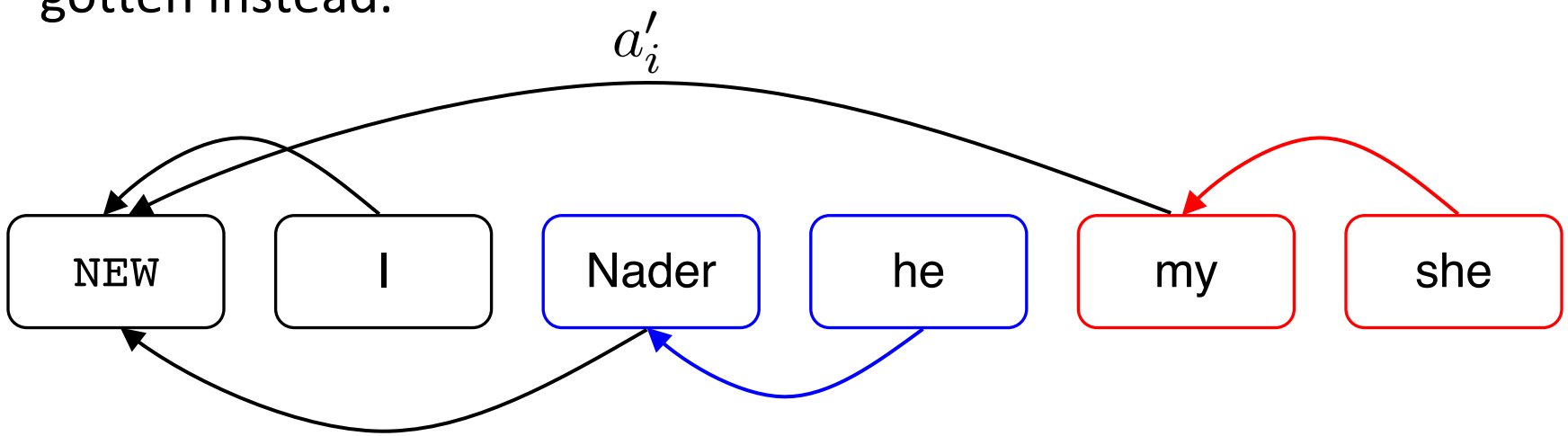
- Since actions are independent, we can change change an action a_i to a different one a_i' and see what reward we would have gotten instead.



Reward = 100

2. Reward-Rescaling

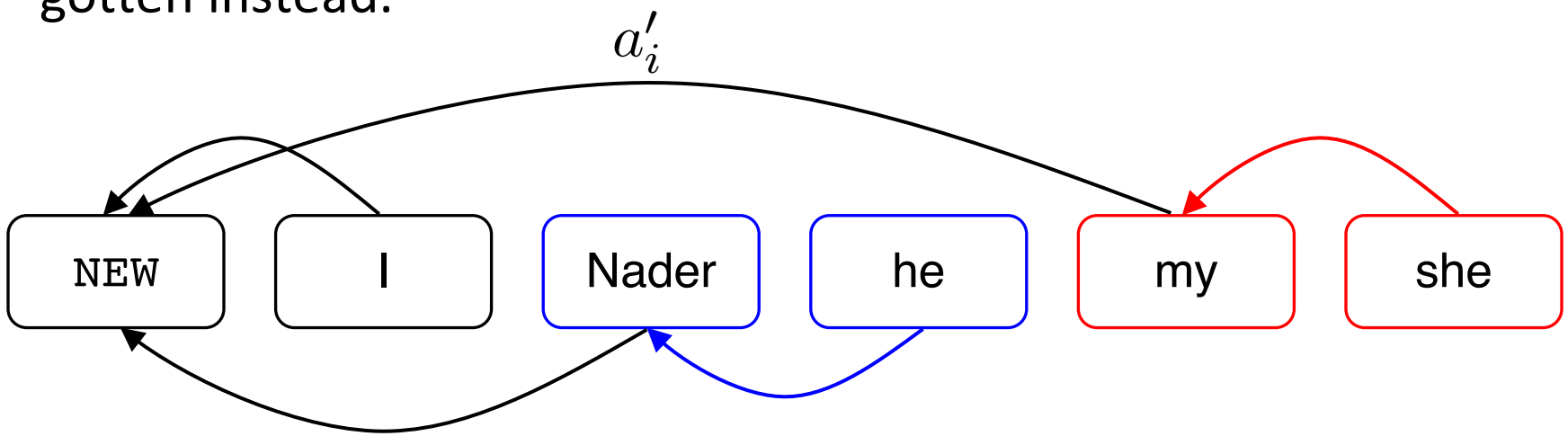
- Since actions are independent, we can change change an action a_i to a different one a_i' and see what reward we would have gotten instead.



Reward = 85

2. Reward-Rescaling

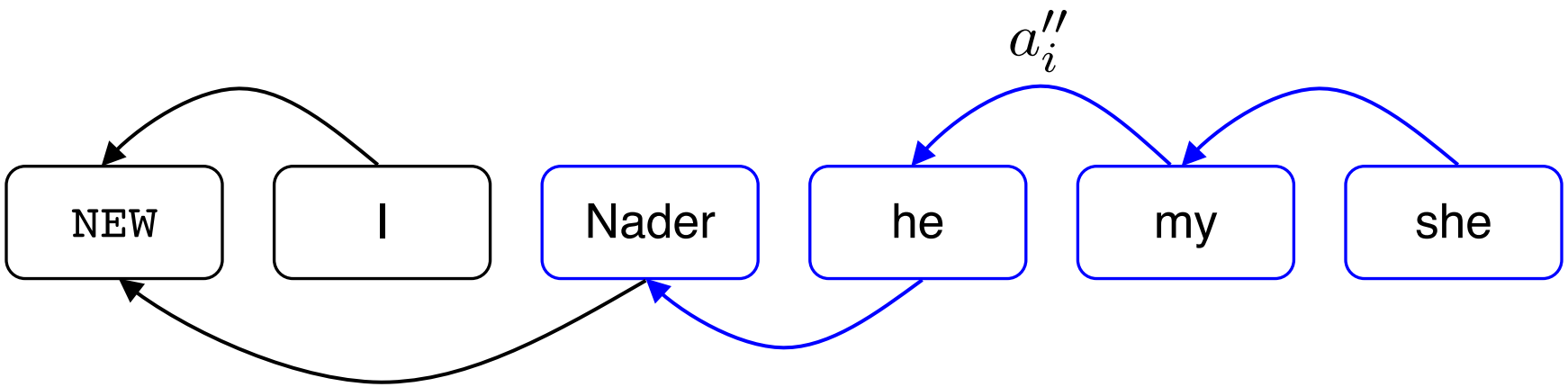
- Since actions are independent, we can change change an action a_i to a different one a_i' and see what reward we would have gotten instead.



Reward = 85
Regret = 15

2. Reward-Rescaling

- Since actions are independent, we can change change an action a_i to a different one a_i' and see what reward we would have gotten instead.



Reward = 66
Regret = 34

2. Reward-Rescaling

- Use this idea to do reward-based slack-rescaling

$$\Delta_r(c, m_i) = \max_{a_i' \in \mathcal{A}_i} R(a_1, ..., a_i', ..., a_T)$$ reward for best action

$$- R(a_1, ..., (c, m_i), ..., a_T)$$ reward for current action

- Cost is the regret of taking the action
  - Replaces the heuristic cost, otherwise use the same max-margin loss function

Experimental Setup

- English and Chinese Portions of the CoNLL 2012 Shared Task dataset
- Predicted mentions from the Stanford rule-based system (Lee et al., 2011)
- Scores are CoNLL F_1 scores
  - Average of MUC, B-cubed, and CEAFe metrics

System Performance

Model	English	Chinese
Heuristic Loss	65.36	63.54
REINFORCE	65.41	63.64
Reward Rescaling	**65.73**	**63.88**

- Small but statistically significant ($p < 0.05$) improvement from reward rescaling

Clark and Manning (EMNLP 2016) results

CoNLL average, CoNLL 2012 data, scorer v8.01

Model	English	Chinese
Chen & Ng (2012) [CoNLL 2012 Chinese winner]	54.52	**57.63**
Fernandes (2012) [CoNLL 2012 English winner]	**60.65**	51.46
Björkelund & Kuhn. (2014) Best previous Chinese system]	61.63	**60.06**
Wiseman et al. (2016) [Best previous English system]	**64.21**	—
Clark & Manning (ACL 2016)	65.29	63.66
(Clark & Manning, EMNLP 2016)	**65.73**	**63.88**

Where do neural scoring models help?

- Especially with nominals with no head match
  - 18.9 F_1 vs 10.7 F_1 on this type compared to 68.7 vs 66.1 F_1 overall when compared with Clark & Manning (2015).

Example Wins

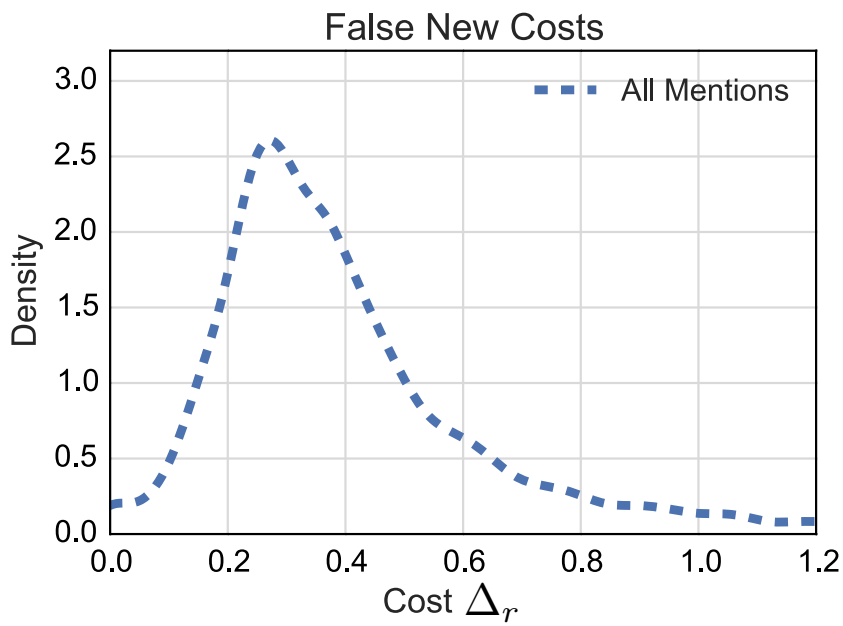
Anaphor	Antecedent
the country's leftist rebels	the guerillas
the company	the New York firm
216 sailors from the ``USS cole''	the crew
the gun	the rifle

Error Breakdown

- Reward-rescaling model actually makes more errors!
- However, the errors are less severe
  - ≈0.7% lower cost on average

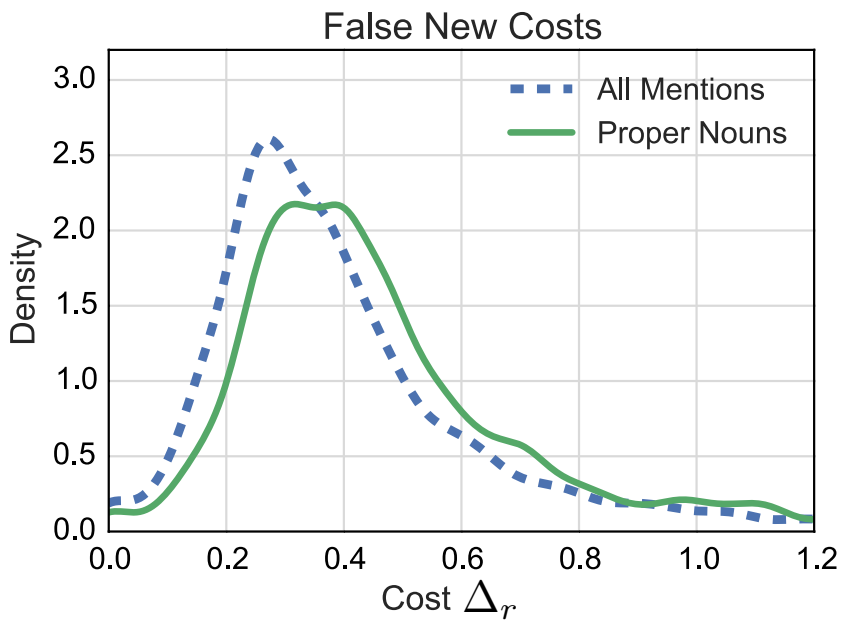
Model	# False Anaphoric	# False New	# Wrong Link
Heuristic Loss	1956	1719	1258
Reward Rescaling	1994	1725	1247

Examining Reward-Based Costs



- High variance in costs for a given error type

Examining Reward-Based Costs



- High variance in costs for a given error type
- One example: “False New” costs for proper nouns are higher on average

Error Breakdown: Proper Nouns

- Reward rescaling gets fewer “false new” errors with proper nouns than the heuristic loss

Model	# False Anaphoric	# False New	# Wrong Link
Heuristic Loss	597	**403**	221
Reward Rescaling	660	**334**	233