CS11-747 Neural Networks for NLP

Document Level Models

Zhengzhong Liu (Hector)

Carnegie Mellon University

Language Technologies Institute

Site

https://phontron.com/class/nn4nlp2017/

NN and some NLP tasks

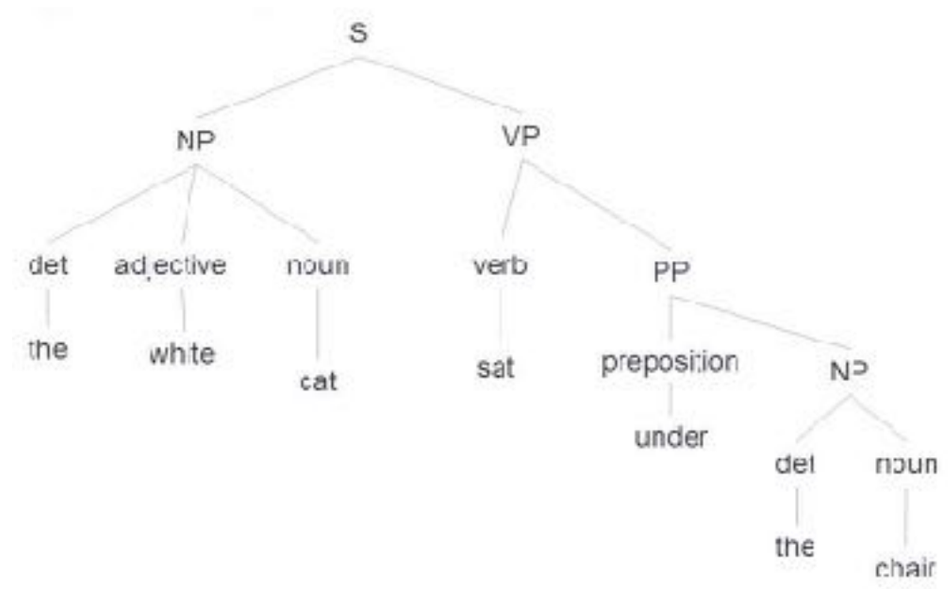
Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'

$P(w_{i+1}= \text{of} \mid w_i=\text{tired}) = 1$
$P(w_{i+1}= \text{of} \mid w_i=\text{use}) = 1$
$P(w_{i+1}= \text{sister} \mid w_i=\text{her}) = 1$
$P(w_{i+1}= \text{beginning} \mid w_i=\text{was}) = 1/2$
$P(w_{i+1}= \text{reading} \mid w_i=\text{was}) = 1/2$

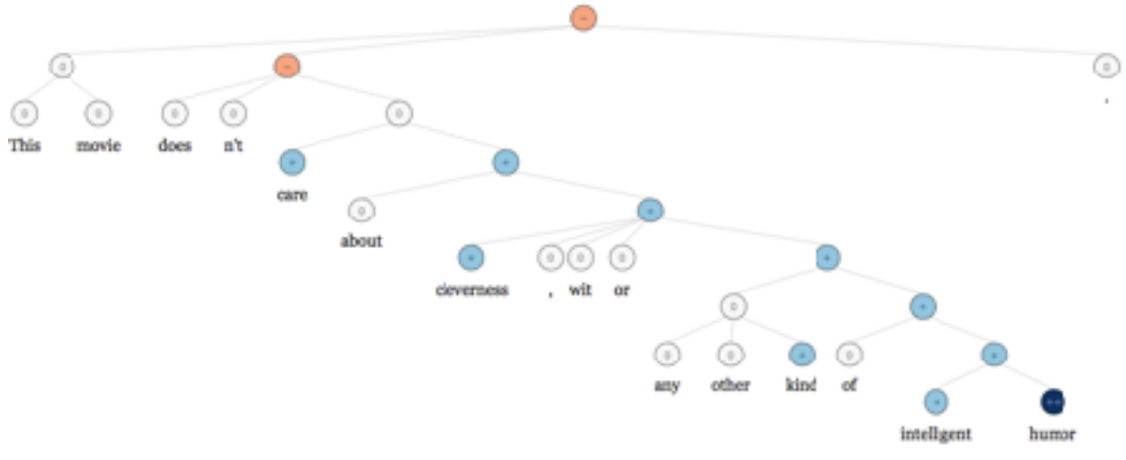
$P(w_{i+1}= \text{bank} \mid w_i=\text{the}) = 1/3$
$P(w_{i+1}= \text{book} \mid w_i=\text{the}) = 1/3$
$P(w_{i+1}= \text{use} \mid w_i=\text{the}) = 1/3$

$P(X|Y) \quad \frac{P(X,Y)}{P(Y)}$

Language Models



Parsing



Classification

Germany's representative to the European Union's veterinary committee Werner Zwingman said on Wednesday consumers should …

Entity Tagging

Their Counter-part in Documents

Recovers structures of documents

	Sentence	Document	
Entity	Entity Tagging	Coreference	70%
Parsing	Semantic Parsing; Syntactic Parsing	Discourse Parsing	15%
Language Model	Word Prediction	Sentence/Discourse Element Prediction	15%
Classification	Sentence Classification	Document Classification	

Document Problems: Entity Coreference

Queen Elizabeth set about transforming her husband,King George VI, into *a viable monarch*.
A renowned speech therapist was summoned to help the King overcome his *speech impediment*...

Example from Ng, 2016

- Step 1: Identify Noun Phrases mentioning an entity (note the difference from *named* entity recognition).

- Step 2: Cluster noun phrases (**mentions**) referring to the same underlying world **entity**.

Mention(Noun Phrase) Detection

A renowned speech therapist was summoned to help the King overcome his *speech impediment*…

A renowned speech therapist was summoned to help the King overcome his *speech impediment*…

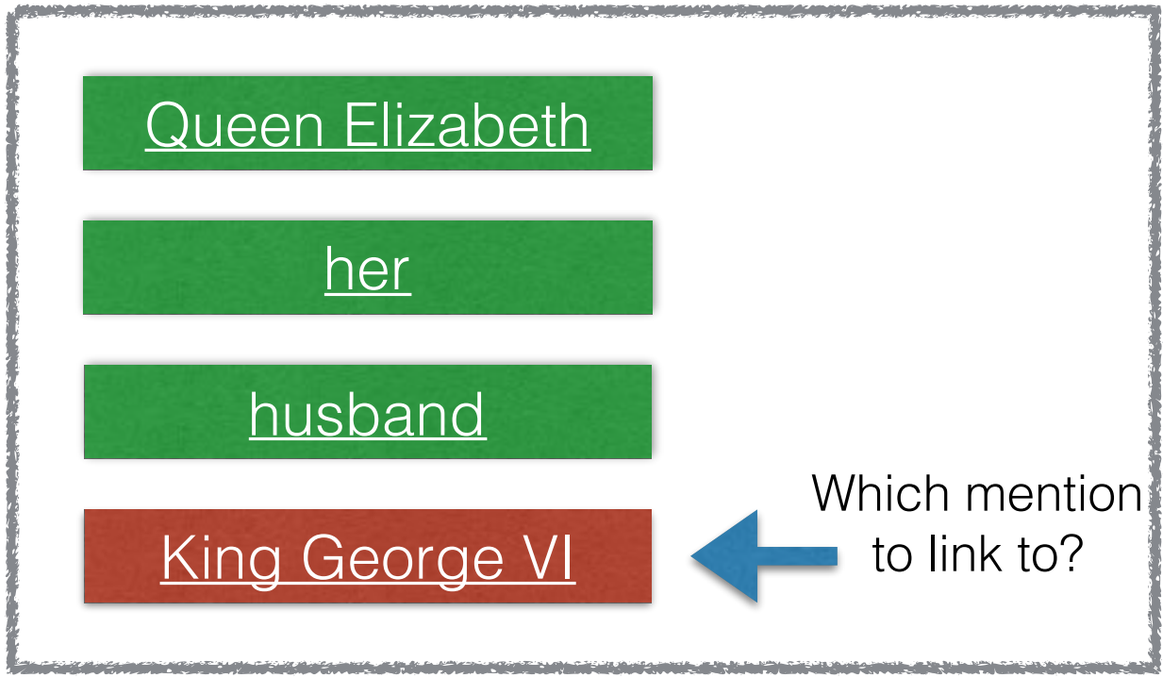
- One may think coreference is simply a clustering problem of given Noun Phrases.
  - Detecting relevant noun phrases is a difficult and important step.
  - Knowing the correct noun phrases affect the result a lot.
  - Normally done as a preprocessing step.

Components of a Coreference Model

- Like a traditional machine learning model:
    - We need to know the **instances** (e.g. shift-reduce operations in parsing).
    - We need to design the **features**.
    - We need to optimize towards the **evaluation metrics**.
    - Search algorithm for structure (covered in later lectures).

Coreference Models:Instances

- Coreference is a structured prediction problem:
  - Possible cluster structures are in exponential number of the number of mentions. (Number of partitions)
- Models are designed to approximate/explore the space, the core difference is the way each instance is constructed:
  - Mention-Pair Model
  - Entity-Mention Model
  - Mention-Ranking Model
  - Latent Tree Models
- Mimic the cluster creation process of human.



Mention Pair Models

- The simplest one: Mention Pair Model:
  - Classify the coreference relation between every 2 mentions.
- Simple but many drawbacks:
  - May result in conflicts in transitivity.
  - Too many negative training instances.
  - Do not capture **entity/cluster level** features.
  - No ranking of instances.

Queen Elizabeth set about transforming her husband,King George VI, into *a viable monarch*. *A renowned speech therapist* was summoned to help the King overcome his *speech impediment*...

✓: Queen Elizabeth <-> her
✗: Queen Elizabeth <-> husband
✗: Queen Elizabeth <-> King George VI
✗: Queen Elizabeth <-> a viable monarch
.....

Entity Models: Entity-Mention Models

- Entity-Mention Models
  - Create an instance between a mention and a previous* cluster.

Daume & Marcu (2005); Cullotta et al. (2007)

* This process often follows the natural discourse order, so we can refer to partial build clusters.

Example Cluster Level Features:

- Are the genders all compatible?
- Is the cluster containing pronouns only?
- Most of the entities are the same gender?????
- Size of the clusters?

Problems:

- No ranking between the antecedents.
- Cluster level features are difficult to design.

Entity Models: Entity-Centric Models

Clark and Manning (2015)

- Entity Centric Models
  - Create an instance between two clusters.
  - Allow building a entity representation.

Problems:
- Cluster level features are difficult to design. (recurring problem)
- No direct guidance of entity creation process

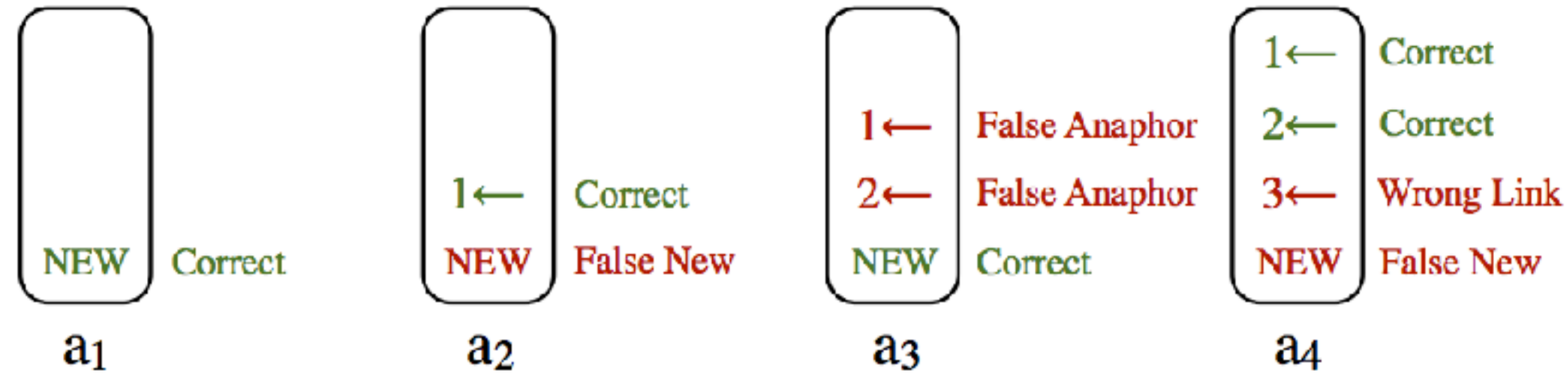
Learning Algorithm
- Build up clusters during learning (normally agglomerative)
- No cluster creation gold standard!!
  - "**Create**" gold standard to guide the clusters.
  - Train with RL: Clark and Manning (2015) trained it with DAgger.

Ranking Models

- Added relative importance to antecedents.
  - Easy-first intuition, some decisions are easier than the others.
  - Help deal with imbalance between positive and negative.
    - Anaphora problem: what if a mention does not have an antecedent? (Create a NULL mention)
- Mention Ranking (Currently more popular)
  - Ranking previous mentions. (Durrett & Klein 2013, Ma et.al 2016)
- Entity Ranking
  - Rank preceding clusters, not individual mentions. (Rahman & Ng, 2009)

Ranking Model: Mention Ranking

(Durrett and Klein, 2013)



A **Log-Linear probabilistic** Model

- Create a antecedent structure (a1, a2, a3, a4): where each mention need to decide a ranking of the antecedents
- Problem: No Gold Standard antecedent structure?
  - **Sum over** all possible structures licensed by the gold cluster

Ranking Model: Entity Ranking

(Rahman & Ng, 2009)

Features describing m_j, a candidate a		
1	PRONOUN_1	Y if m_j is a pro
2	SUBJECT_1	Y if m_j is a sub
3	NESTED_1	Y if m_j is a nes
Features describing m_k, the mention		
4	NUMBER_2	SINGULAR or PL
5	GENDER_2	MALE, FEMALE, common first n
6	PRONOUN_2	Y if m_k is a pro
7	NESTED_2	Y if m_k is a nes
8	SEMCLASS_2	the semantic cl NIZATION, DATI mined using W nizer (Finkel, C
9	ANIMACY_2	Y if m_k is deter recognizer; else
10	PRO_TYPE_2	the nominative feature value fo

Features describing the relationship between m_j, a candidate antecedent and m_k, the mention to be resolved (continued from the previous page)		
30	SEMCLASS	C if the mentions have the same semantic class (where the set of semantic classes considered here is enumerated in the description of the SEMCLASS_2 feature); I if they don't; NA if the semantic class information for one or both mentions cannot be determined
31	ALIAS	C if one mention is an abbreviation or an acronym of the other; else I
32	DISTANCE	binned values for sentence distance between the mentions
Additional features describing the relationship between m_j, a candidate antecedent and m_k, the mention to be resolved		
33	NUMBER'	the concatenation of the NUMBER_2 feature values of m_j and m_k. E.g., if m_j is *Clinton* and m_k is *they*, the feature value is SINGULAR-PLURAL, since m_j is singular and m_k is plural
34	GENDER'	the concatenation of the GENDER_2 feature values of m_j and m_k
35	PRONOUN'	the concatenation of the PRONOUN_2 feature values of m_j and m_k
36	NESTED'	the concatenation of the NESTED_2 feature values of m_j and m_k
37	SEMCLASS'	the concatenation of the SEMCLASS_2 feature values of m_j and m_k
38	ANIMACY'	the concatenation of the ANIMACY_2 feature values of m_j and m_k
39	PRO_TYPE'	the concatenation of the PRO_TYPE_2 feature values of m_j and m_k

cn m_j, a candidate antecedent and m_k,

e the same head noun; else I
the same string; else I
substring of the other; else I
e pronominal and are the same string; else I
e proper names and are the same string; else I
s are both non-pronominal and are the same

the same modifiers; NA if one of both of them
; else I
pronominal and are either the same pronoun
respect to case; NA if at least one of them is

ree in number; I if they disagree; NA if the
h mentions cannot be determined
e is gender; I if they disagree; NA if the gender
ler cannot be determined
e is both gender and number; I if they disagree
nder; else NA
ith in animacy; I if they don't; NA if the
ch mentions cannot be determined
e pronouns; I if neither are pronouns; else NA
e proper names; I if neither are proper names;

does not have the same maximal NP projec-

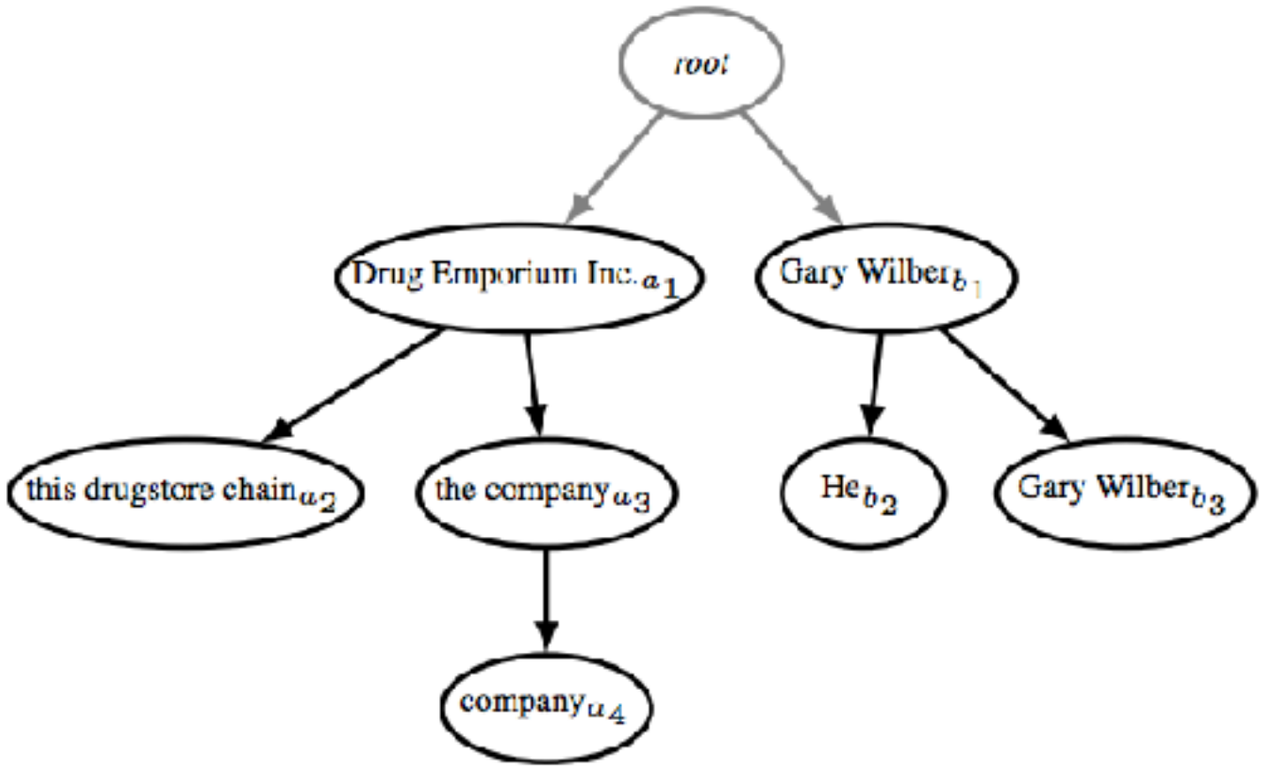
nent the other; else I
e NP and is not in an apposition relationship;

in an apposition relationship; else I
in a copular construction; else I

Rank previous clusters for a given mention.
Similarly, a NULL cluster is added to the antecedents.
Rahman & Ng use a complex set of features (39 feature templates)

Latent Tree Models

(Bjorkelund and Kuhn, 2014)



Latent Tree Model share some similarities with the mention ranking models.

Each subtree under the root represent a cluster.

Trained as **structured perceptron**

- Create a antecedent structure (as a tree), where each mention need to decide which antecedent to linked to (similar to a ranking)
- Problem: No Gold Standard antecedent tree? (Hence called the Latent Tree)
  - **Pick the highest scored tree structure within** all possible structures licensed by the gold cluster

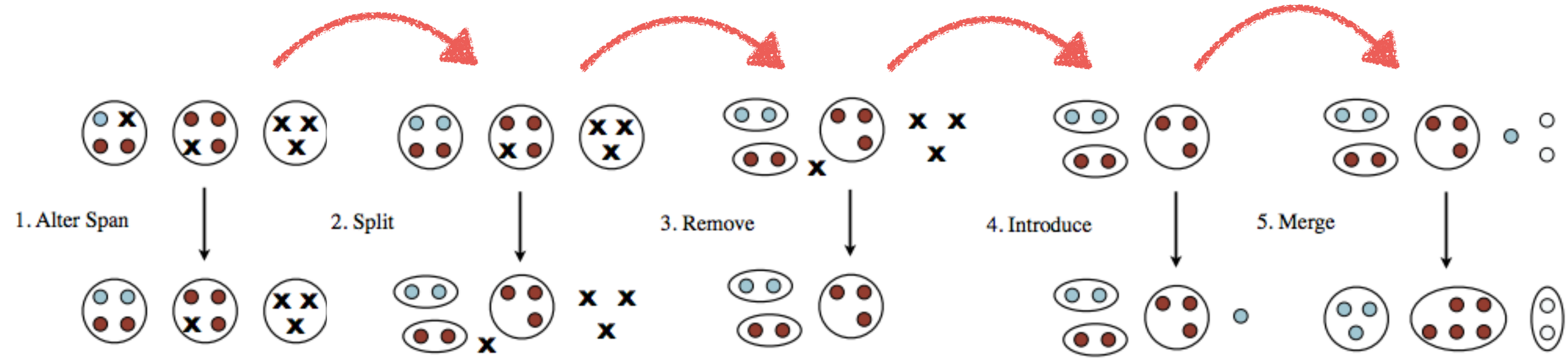
What’s the role of Neural Networks here?

Problems in Coreference: revisited

- **Instance** Problem
  - We've introduced 4 different modeling methods, many seem to work in their own settings.
- **Feature** Problem
  - The core of the success may still be the feature problem. For example, Bjorkelund and Kuhn use a decision tree for feature induction. Durrett and Klein conduct careful feature engineering and selection.
- **Metric** Problem: clustering metric is (very) difficult to compute (any thoughts?)

Error Driven Analysis

(Kummerfeld and Klein, 2013)



- Five types of operation to transform coreference decisions.
- The combination of the operations creates 7 types of errors.

Error Driven Analysis

(Kummerfeld and Klein, 2013)

System	Metric F-Scores Mention	MUC	B^3	Span Error	Conflated Entities	Extra Mention	Extra Entity	Divided Entity	Missing Mention	Missing Entity
PUBLICLY AVAILABLE SYSTEMS										
BERKELEY	75.57	66.43	66.17							
IMS	72.96	64.71	64.73							
STANFORD-T	71.21	61.40	63.06							
STANFORD	58.56	48.37	56.42							
RECONCILE	46.45	49.40	54.90							
BART	56.61	46.00	52.56							
UIUC	50.60	45.21	52.88							
CHERRYPICKER	41.10	40.71	51.39							

- Five types of operation to transform coreference decisions.
- The combination of the operations creates 7 types of errors.

Easy Victories & Uphill Battles

- A mention ranking model (We've actually covered its model in previous slides).

- Error type based loss in cost function:

- Trained with softmax-margin cost (a way to add cost sensitive training to log-linear models).

- Combined loss: $l(a, C^*) = \alpha_{\mathrm{FA}}\mathrm{FA}(a, C^*) + \alpha_{\mathrm{FN}}\mathrm{FN}(a, C^*) + \alpha_{\mathrm{WL}}\mathrm{WL}(a, C^*)$

- FA (False Anaphora), FN (False New), WL(Wrong Link)

Easy Victories & Uphill Battles

(Durrett and Klein, 2013)

- Easy Victories from Surface (lexical) Features:
    - Ignore all many complex features, all replaced with surface features.
    - **Data driven features** beat Heuristic driven (Sounds familiar?).
    - Many heuristic features can be captured (implicitly) by surface features:
        - Number, gender, person can be encoded in pronouns.
        - Centering theory: verb before or after can indicate subj, obj.
        - Definiteness: first word of a mention will encode that.

Easy Victories & Uphill Battles

Feature name	Count
Features on the current mention	
[ANAPHORIC] + [HEAD WORD]	41371
[ANAPHORIC] + [FIRST WORD]	18991
[ANAPHORIC] + [LAST WORD]	19184
[ANAPHORIC] + [PRECEDING WORD]	54605
[ANAPHORIC] + [FOLLOWING WORD]	57239
[ANAPHORIC] + [LENGTH]	4304
Features on the antecedent	
[ANTECEDENT HEAD WORD]	57383
[ANTECEDENT FIRST WORD]	24239
[ANTECEDENT LAST WORD]	23819
[ANTECEDENT PRECEDING WORD]	53421
[ANTECEDENT FOLLOWING WORD]	55718
[ANTECEDENT LENGTH]	4620
Features on the pair	
[EXACT STRING MATCH (T/F)]	47
[HEAD MATCH (T/F)]	46
[SENTENCE DISTANCE, CAPPED AT 10]	2037
[MENTION DISTANCE, CAPPED AT 10]	1680

Feature name	Count
Features of the SURFACE system	418704
Features on the current mention	
[ANAPHORIC] + [CURRENT ANCESTRY]	46047
Features on the antecedent	
[ANTECEDENT ANCESTRY]	53874
[ANTECEDENT GENDER]	338
[ANTECEDENT NUMBER]	290
Features on the pair	
[HEAD CONTAINED (T/F)]	136
[EXACT STRING CONTAINED (T/F)]	133
[NESTED (T/F)]	355
[DOC TYPE] + [SAME SPEAKER (T/F)]	437
[CURRENT ANCESTRY] + [ANT. ANCESTRY]	2555359

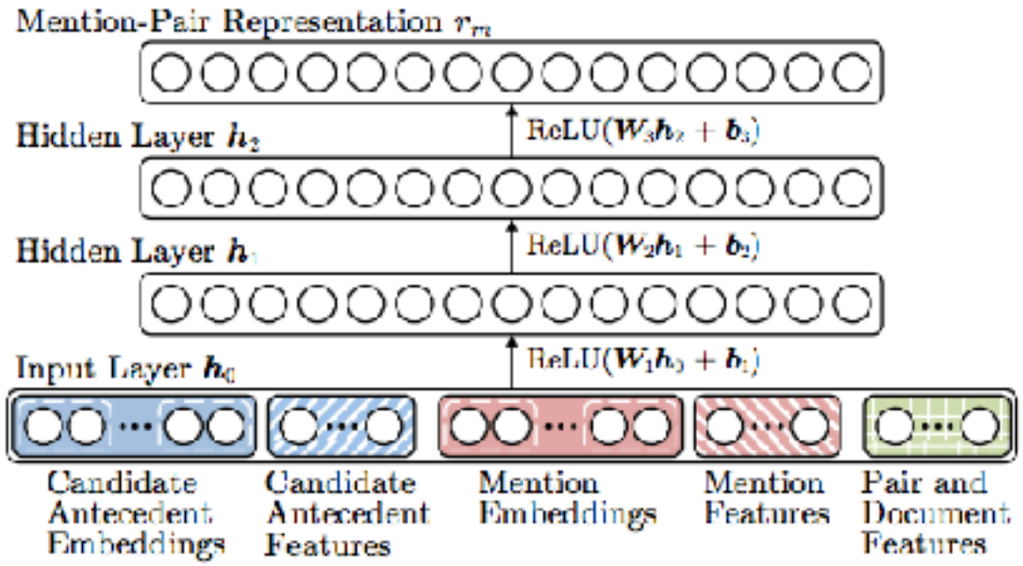
Final Feature Set

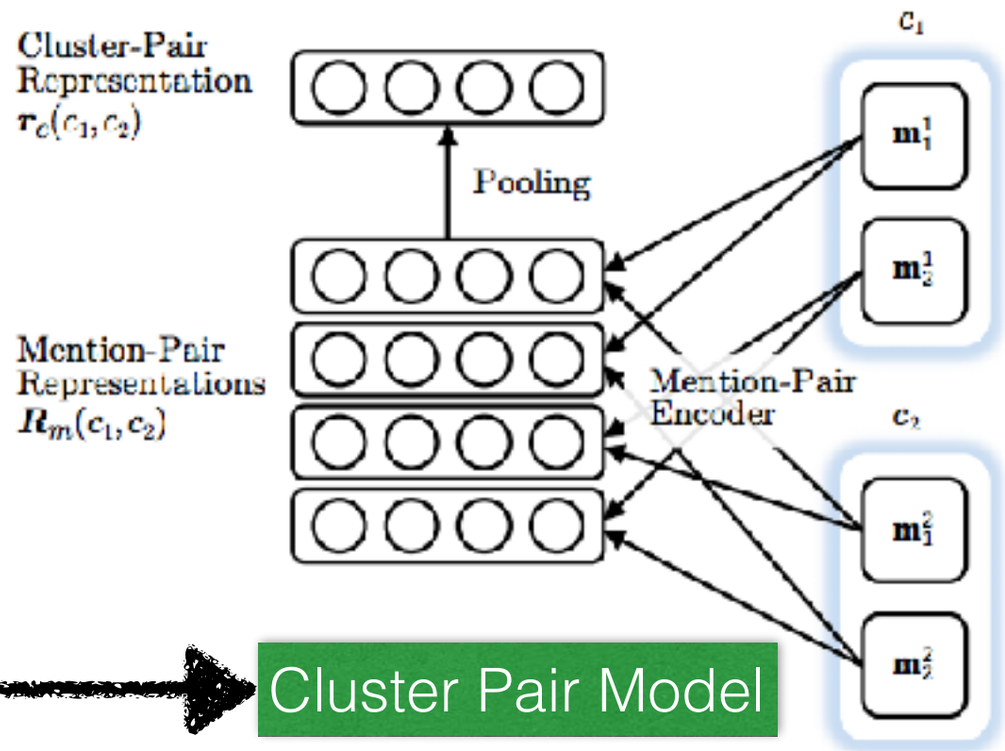
Some Possible Improvements w/ NN

- Train towards the metric using Deep RL.
- Learn the features with embeddings since most of them can be captured by surface features.
- Can some features be captured better with NN?
- Train the full system to reduce specific error types:
  - which errors specifically?

Coreference Resolution w/ Entity-Level Distributed Representations

Clark & Manning (2015)





Mention Pair Model → Cluster Pair Model

- Mention Pair Model and Cluster Pair model to capture representation
- Typical Coreference Features are used as embeddings or on-hot features
- Mention Pair Features are fed to the cluster pair features, followed by pooling

Feature

- Heuristic Max-Margin as in Wiseman et al.(2015) and Durrett & Klein (2013)

Objective

- Cluster merging as with Policy Network (MERGE or PASS)
- Trained with SEARN (Daume III et al., 2009)

Training

Deep Reinforcement Learning for Mention-Ranking Coreference Models

Clark & Manning (2016)

- A continuous of the previous model:
  - Same features and structure.
- Objective changed: reinforcement learning
  - Choosing which previous antecedent is considered as an action of the agent.
  - The final reward is one of the 4 main evaluation metric in coreference (B-Cubed).
  - Best model is reward-rescaled reinforcement method.

Cluster Features w/ Neural Network

Wiseman et.al (2016)

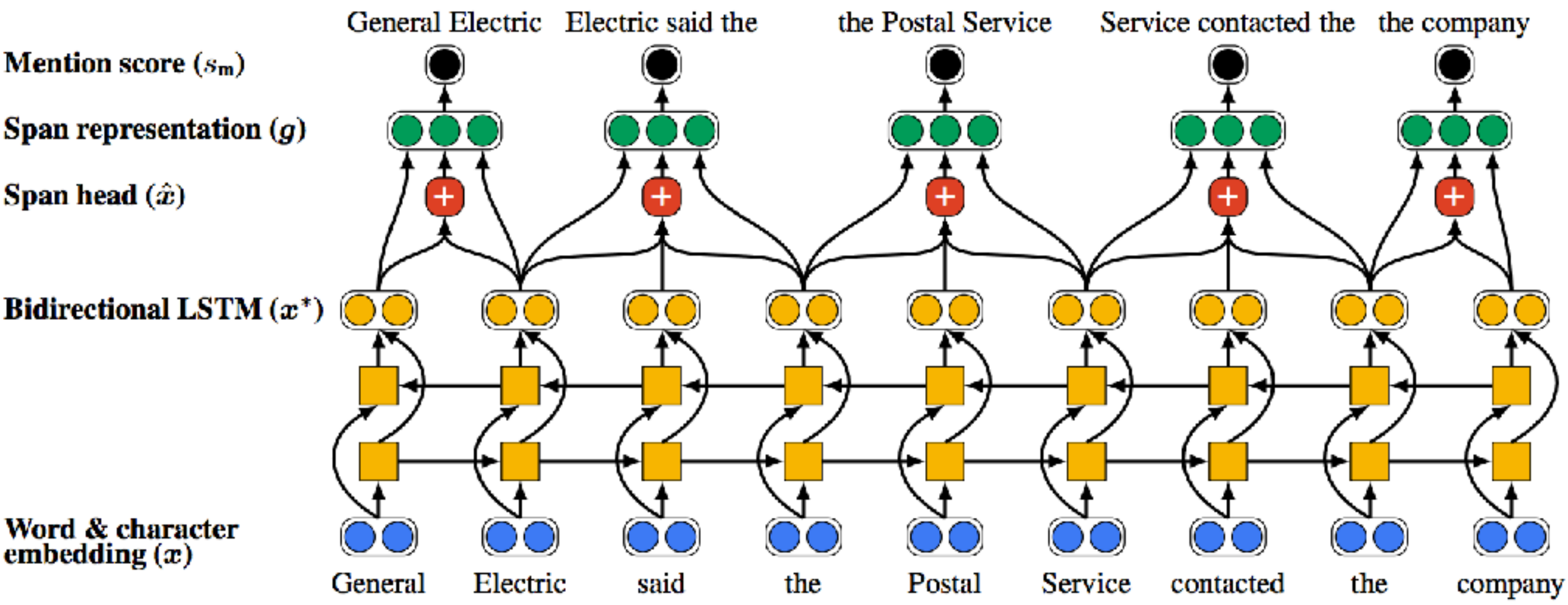
- Cluster level features are difficult to capture.
- Example cluster level features:
  - most-female=true (how to define most?).
  - Pronoun sequence: C-P-P = true.
- Use RNN to embed features from multiple mentions into a single representation.
  - No hand designed cluster level feature templates.

End-to-End Neural Coreference

Lee et.al (2017)

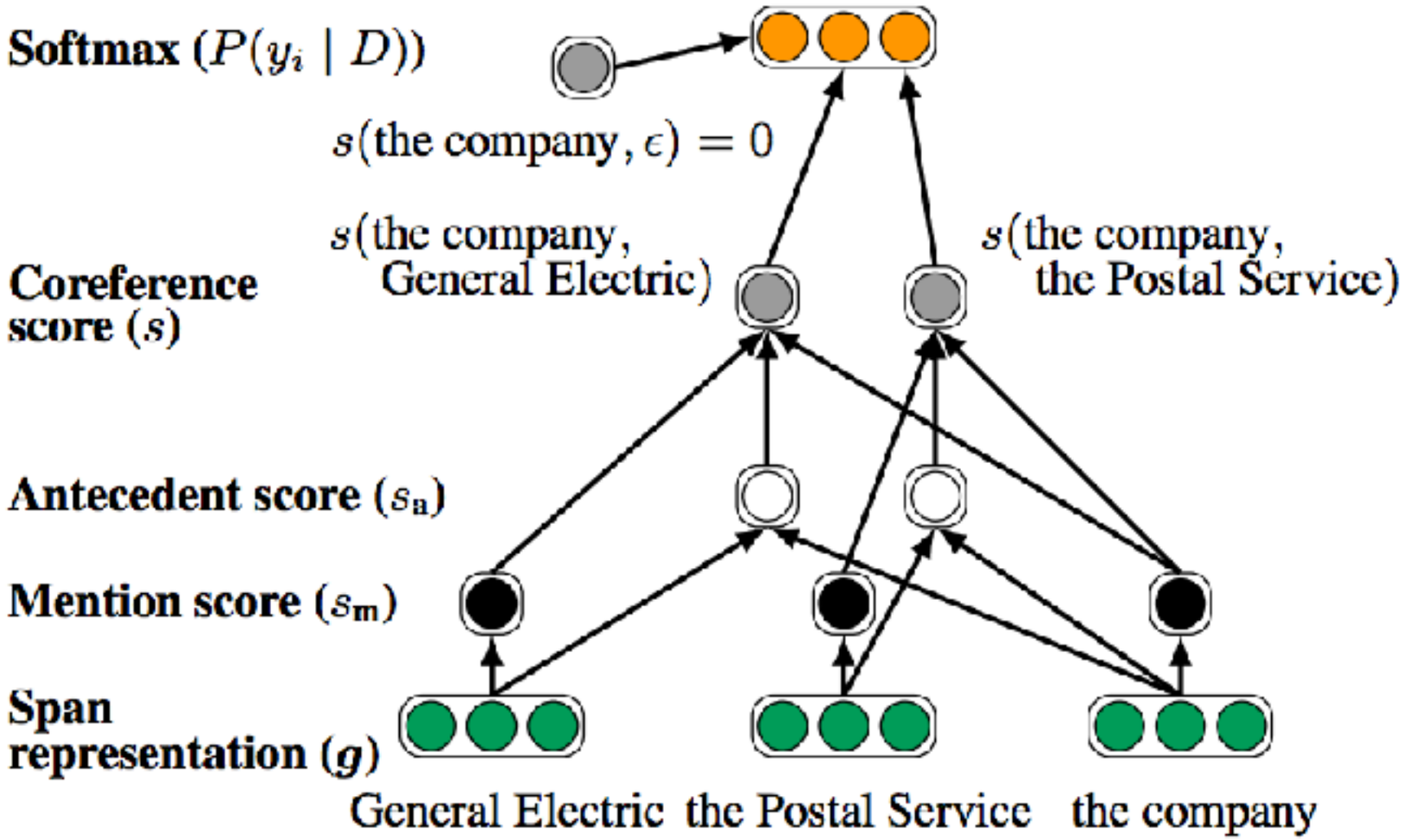
- 2 main contributions by this paper:
  - Can we represent all features with a more typical neural network embedding way?
  - Can neural network allow errors to flow end-to-end? All the way to mention detection?
    - This solves another type of error (span error), which is not previously handled.

End-to-End Neural Coreference (Span Model)



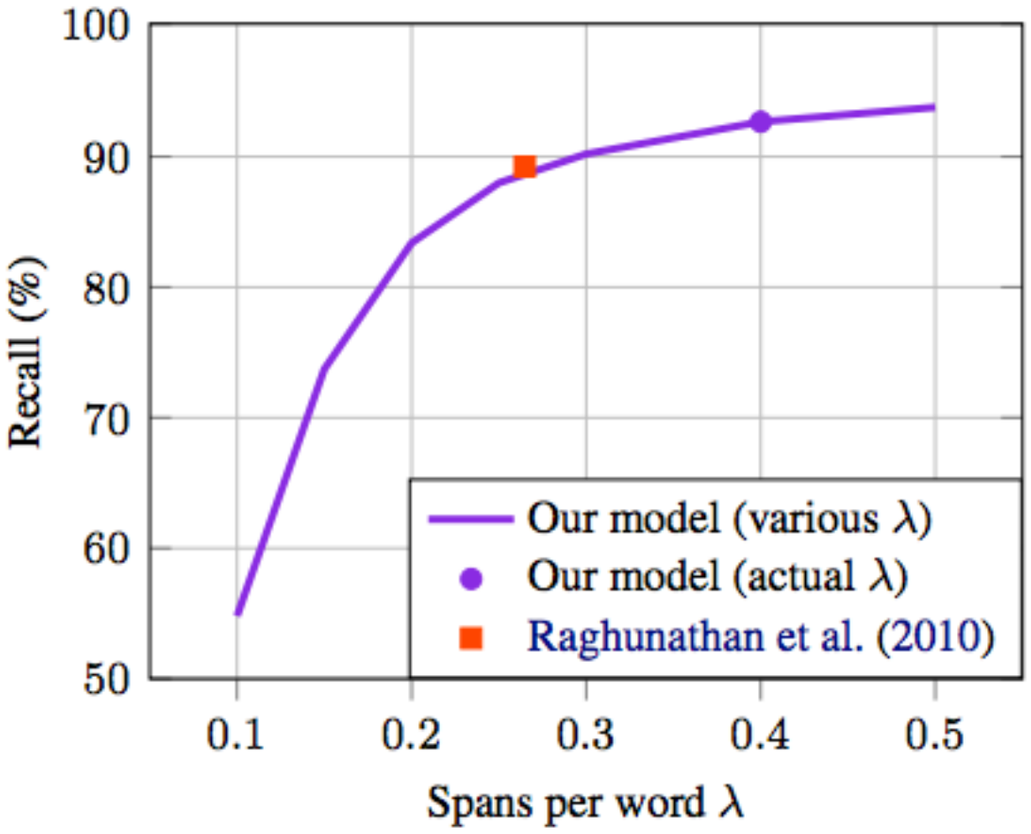
- Build mention representation from word representation (all possible spans)
- Head extracted by self-attention.

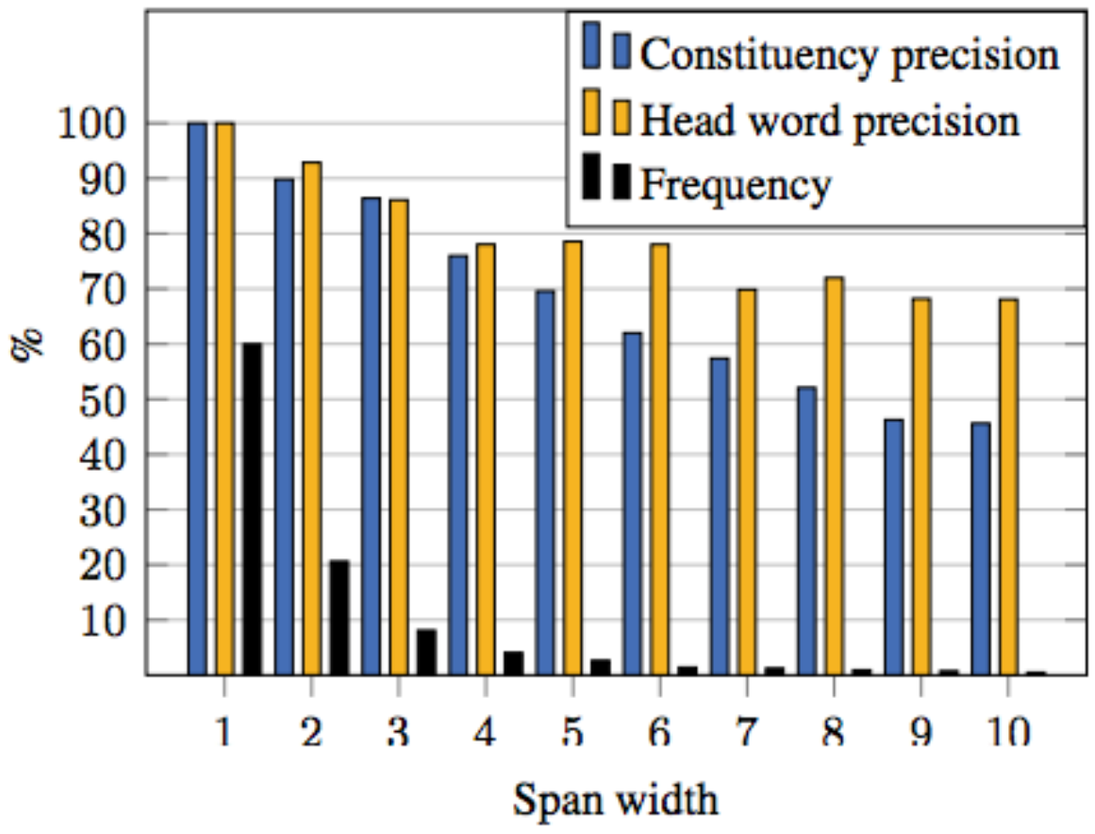
End-to-End Neural Coreference (Coreference Model)



- Coreference model is similar to a mention ranking.
- Coreference score consist of multiple scores.
- Simple max-likelihood (not the cost sensitive method by Durrett, why?)

Quality of Mentions





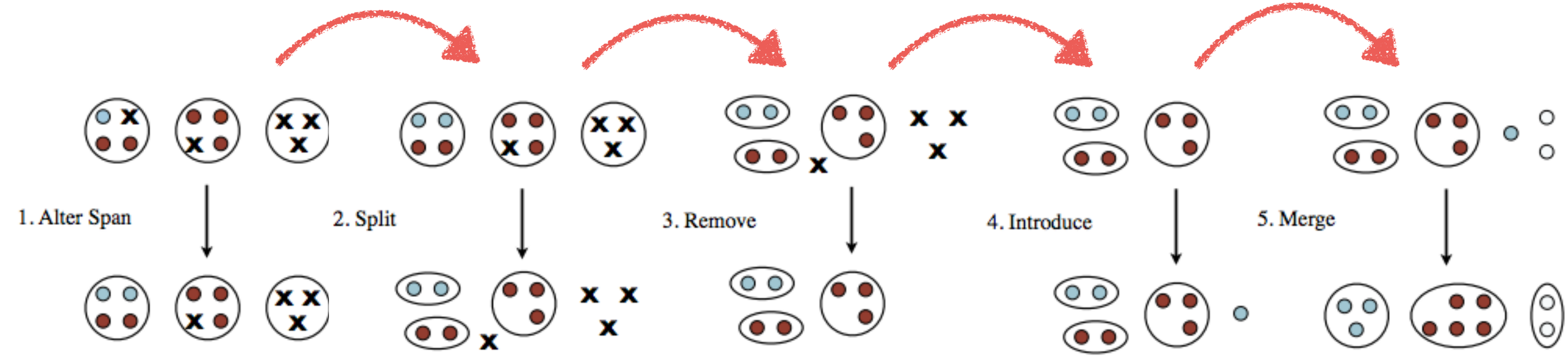
- Build mention representation from word representation (all possible spans)
- Head extracted by self-attention.

Ablations of modules

	Avg. F1	Δ
Our model (ensemble)	69.0	+1.3
Our model (single)	67.7	
– distance and width features	63.9	-3.8
– GloVe embeddings	65.3	-2.4
– speaker and genre metadata	66.3	-1.4
– head-finding attention	66.4	-1.3
– character CNN	66.8	-0.9
– Turian embeddings	66.9	-0.8

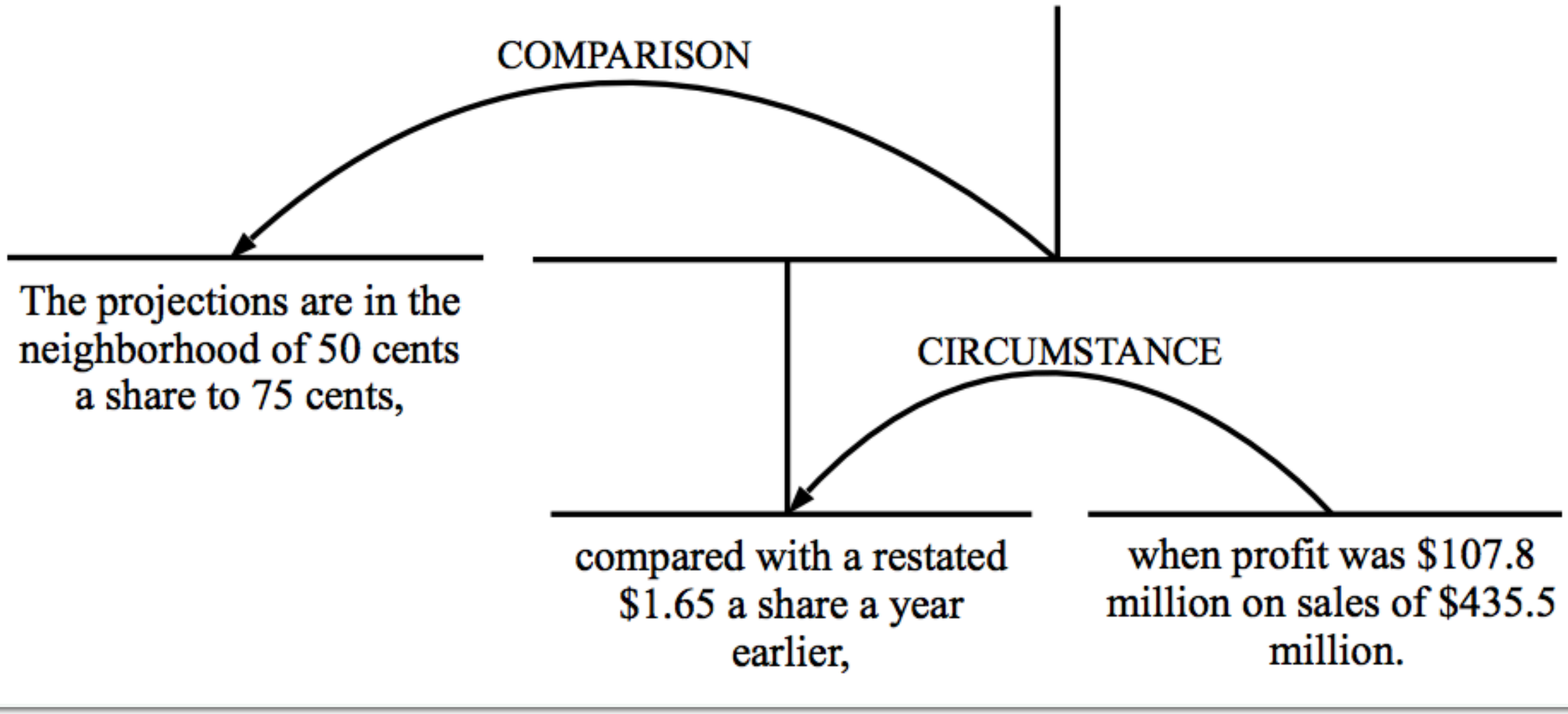
Table 2: Comparisons of our single model on the development data. The 5-model ensemble provides a 1.3 F1 improvement. The head-finding attention, features, and all word representations contribute significantly to the full model.

Error Type Revisited



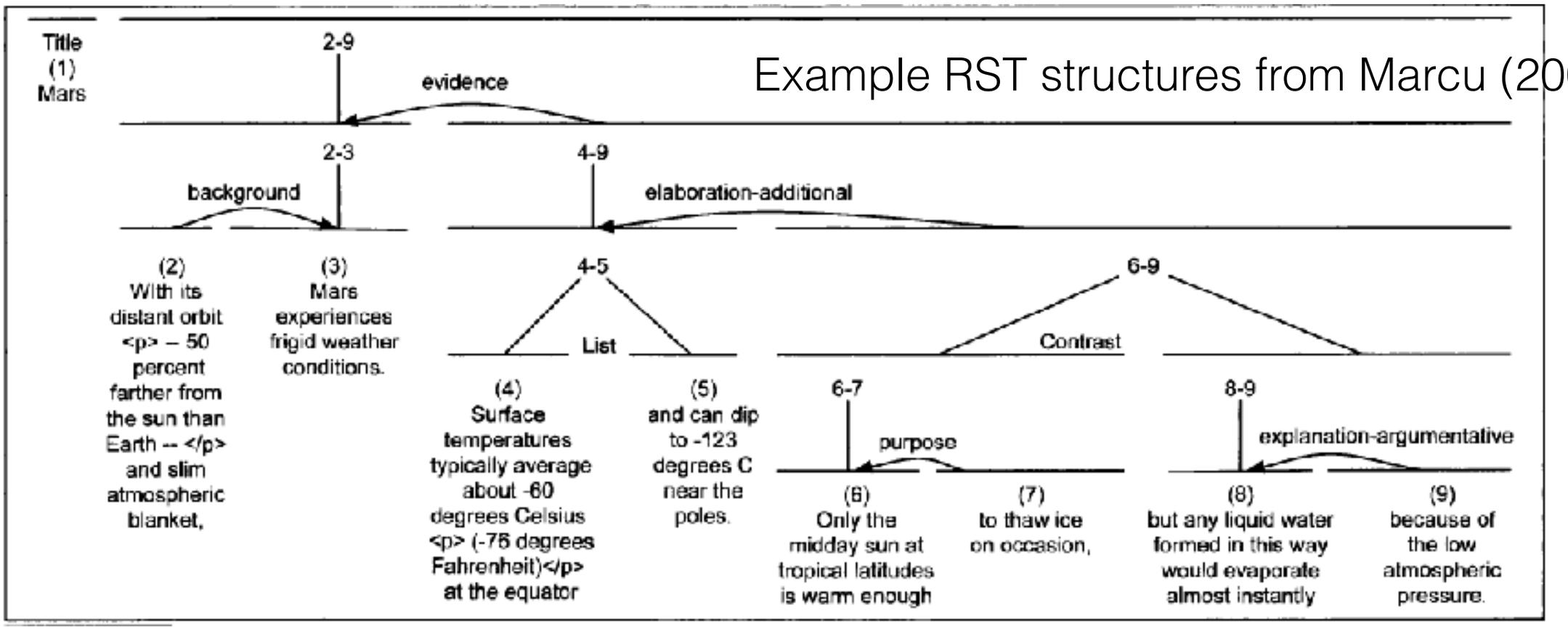
System	Metric F-Scores Mention	MUC	B^3	Span Error	Conflated Entities	Extra Mention	Extra Entity	Divided Entity	Missing Mention	Missing Entity
PUBLICLY AVAILABLE SYSTEMS										
BERKELEY	75.57	66.43	66.17							
IMS	72.96	64.71	64.73							
STANFORD-T	71.21	61.40	63.06							
STANFORD	58.56	48.37	56.42							
RECONCILE	46.45	49.40	54.90							
BART	56.61	46.00	52.56							
UIUC	50.60	45.21	52.88							
CHERRYPICKER	41.10	40.71	51.39							

Kummerfeld and Klein, 2013



Discourse Parsing

Document Problems: Discourse Parsing

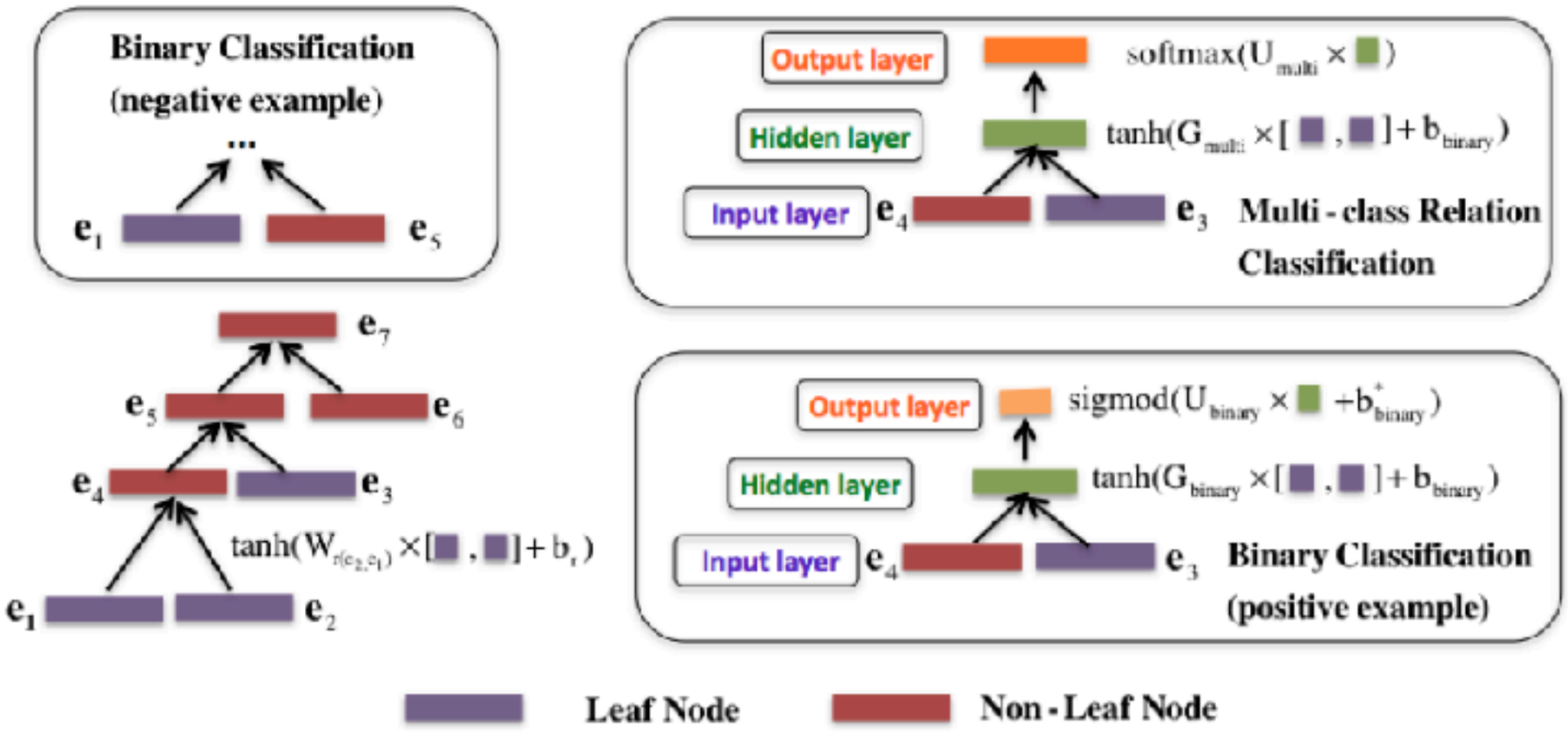


Example RST structures from Marcu (2000)

- Parse a piece of text into a relations between discourse units (EDUs).
- Researchers mainly used the Rhetorical Structure Theory (RST) formalism, which forms a tree of relations.

Recursive Deep Models for Discourse Parsing

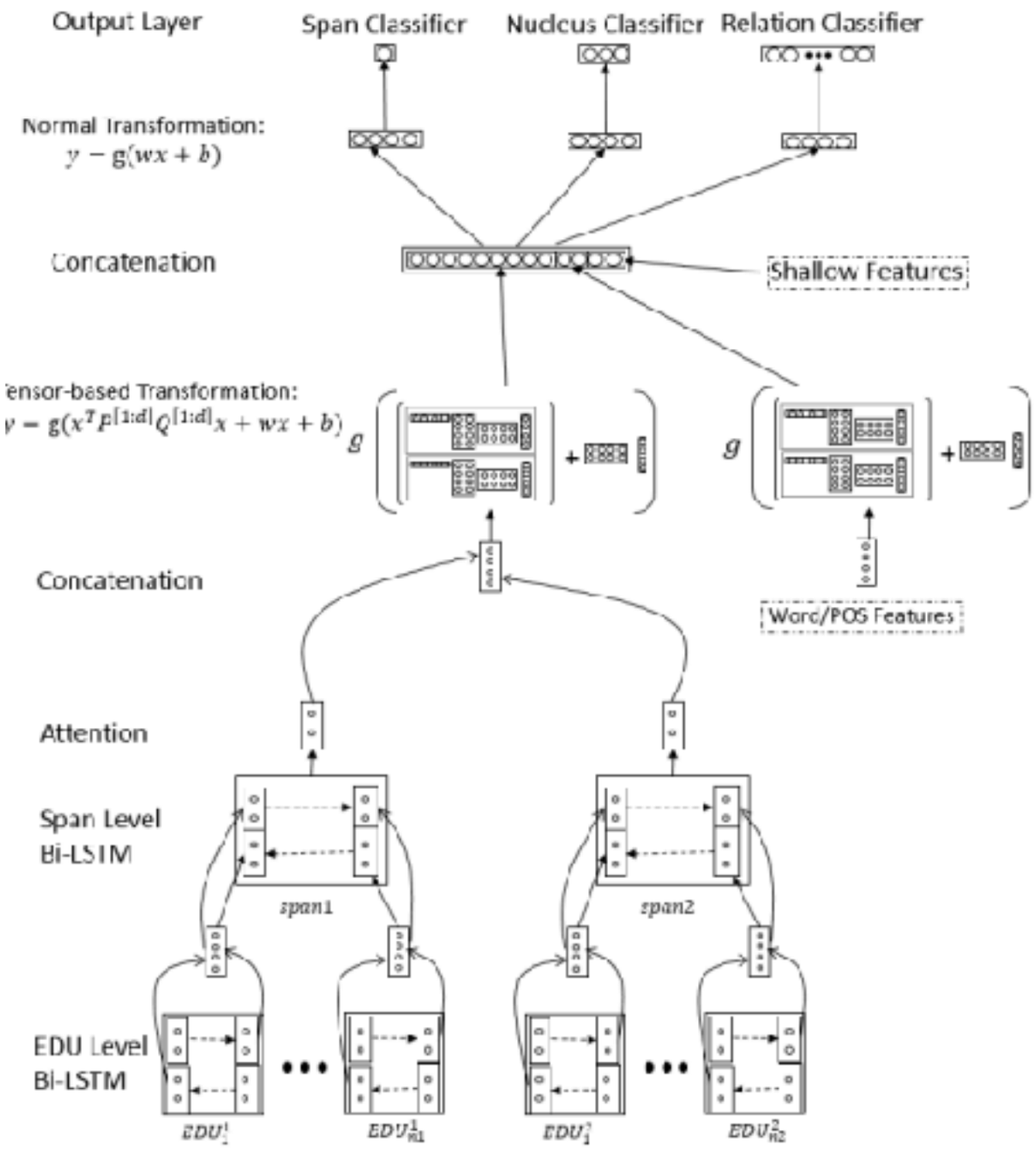
Li et.al (2014)



- Recursive NN for discourse parsing (similar to Socher's recursive parsing)
- First determine whether two spans should be merged (Binary)
- Then determine the relation type

Discourse Parsing w/ Attention-based Hierarchical Neural Networks

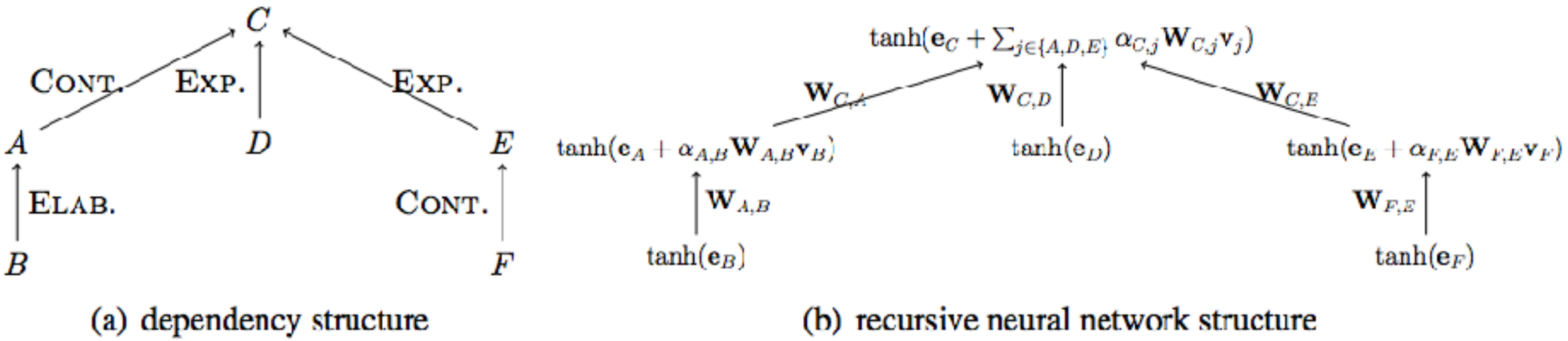
Li et.al (2016)



- Hierarchical bi-LSTM to learn composition scoring.
- Augmented with attention mechanism. (Span is long)
- 2 Bi-LSTMs: first used to capture the representation of a EDU, then combine EDU representation into larger representation
- CKY Parsing

Discourse Structure can help represent documents

Ji and Smith (2017)



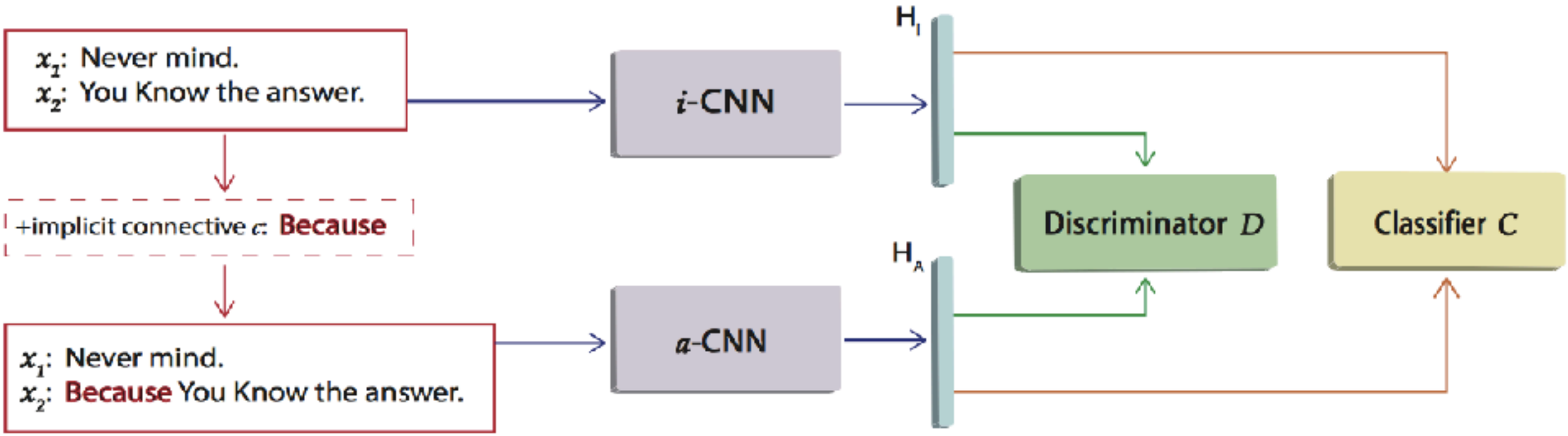
- This work shows that the document representation can be built with discourse structure.
- Similar to a representation of sentence using recursive NN on parse tree.
- They reported better sentiment analysis and document topic classification.

Implicit Discourse Connection Classification w/ Adversarial Objective

(Qin et al. 2017)

- Idea: implicit discourse relations are not explicitly marked, but would like to detect them if they are

- Text with explicit discourse connectives should be the same as text without!



Context	Right Ending	Wrong Ending
Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.	Karen became good friends with her roommate.	Karen hated her roommate.
Jim got his first credit card in college. He didn't have a job so he bought everything on his card. After he graduated he amounted a $10,000 debt. Jim realized that he was foolish to spend so much money.	Jim decided to devise a plan for repayment.	Jim decided to open another credit card.
Gina misplaced her phone at her grandparents. It wasn't anywhere in the living room. She realized she was in the car before. She grabbed her dad's keys and ran outside.	She found her phone in the car.	She didn't want her phone anymore.

The Test	Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. Her teacher stated that the test is postponed for next week. Jennifer felt bittersweet about it.
The Hurricane	Morgan and her family lived in Florida. They heard a hurricane was coming. They decided to evacuate to a relative's house. They arrived and learned from the news that it was a terrible storm. They felt lucky they had evacuated when they did.
Spaghetti Sauce	Tina made spaghetti for her boyfriend. It took a lot of work, but she was very proud. Her boyfriend ate the whole plate and said it was good. Tina tried it herself, and realized it was disgusting. She was touched that he pretended it was good to spare her feelings.

Discourse Prediction

Document Problems: Discourse Unit Prediction

(I)$^{(1)}$ decided to take a (bath)$^{(2)}$ yesterday afternoon after working out . Once (I)$^{(1)}$ got back home , (I)$^{(1)}$ walked to (my)$^{(1)}$ (bathroom)$^{(3)}$ and first quickly scrubbed the (bathroom tub)$^{(4)}$ by turning on the (water)$^{(5)}$ and rinsing (it)$^{(4)}$ clean with a rag . After (I)$^{(1)}$ finished , (I)$^{(1)}$ plugged XXXXXX

Referent Prediction Corpus from (Modi et.al. 2017)

ROS Story corpus (Mostafazade et.al. 2017)

Premise Document	Right Hypothesis	Wrong Hypothesis
Ron started his new job as a landscaper today. He loves the outdoors and has always enjoyed working in it. His boss tells him to re-sod the front yard of the mayor's home. Ron is ecstatic, but **does a thorough job** and **finishes super early**.	His boss **commends** him for a **job well done**.	Ron is immediately **fired** for insubordination.
One day, my sister came over to the house to show us her **puppy**. She told us that she had just gotten the puppy across the street. My sons begged me to get them one. I told them that if they would care for it, they could have it.	My son said they would, so we got a **dog**.	We then grabbed a small **kitten**.

Predicting the next entity/sentence given previous sentences

Predicting Discourse Units are similar to Language Modeling

- Pichotta and Mooney, 2016 use RNN to predict the next event.
  - Basically Sentence-Level Language Models (of events)
- Peng and Roth, 2016 introduced Semantic Language Model
  - *Kevin* was **robbed** by *Robert*, but the *police* mistakenly **arrested** *him*.
  - Frame sequence: [f1, dis1, f2, dis2, …]
  - Entity sequence: [e1, dis1, e2, dis2, …]
  - Applied to coreference resolution and shallow discourse parsing.

Story Completion Task

Context	Incorrect Ending	Correct Ending
He didn't know how the television worked. He tried to fix it, anyway. He climbed up on the roof and fiddled with the antenna. His foot **slipped** on the wet shingles and he went tumbling down.	He decided that was fun and to try tumbling again.	Thankfully, he **recovered**.
Pam thought her front yard looked boring. So she decided to buy several plants. And she placed them in her front yard. She was **proud** of her work.	Pam was **upset** at herself.	Pam was **satisfied**.
Maria smelled the fresh Autumn air and decided to celebrate. She wanted to make candy apples. She picked up the ingredients at a local market and headed home. She cooked the candy and prepared the apples.	Maria's apple **pie** was delicious.	She enjoyed the candy apples.

- Snigdha et.al. (2017) use the Semantic LM learnt by Peng et.al. (2016) as a feature to learn next sentence.
- Cai et.al (2017) use LSTM to encode words as sentences, then encode a series of sentences, to predict next sentence.

Why Discourse LM?

- A normal language model can help predict the next word, very useful in speech recognition, translation, etc.
- A discourse language model help predict the next entity/event, potentially useful for:
  - Information extraction.
  - Entity Coreference (Hey, we just talked about it! Let's elaborate!).

Solving Hard Coreference with LM (The uphill battle!)

Peng et. al. (2015)

- The older students were bullying the younger ones, so we [rescued/punished] **them**.

- Robert was robbed by Kevin , and **he** is [arrested/rescued] by police.

The Winograd Schema Challenge

- Semantic LMs are useful for solving difficult coreference problems.
  - They capture common senses that are not accessible in surface features.
- Peng et. al. (2015, 2016) shows performance improvement of these cases.

Questions?