Discourse Analysis

Zhiyuan Liu

liuzy@tsinghua.edu.cn

THUNLP

Discourse Analysis

- **Discourse** is a **coherent structured** group of sentence
  - Monologue
  - Dialogue

- Important problems in discourse analysis
  - Text segmentation
  - Coreference resolution
  - Coherence analysis

Outline

- Coreference Resolution
- Coherence Analysis
- Dialogue Systems

Outline

- **Coreference Resolution**
  - **Introduction**
  - Unsupervised Coreference Resolution
  - Supervised Coreference Resolution
- Coherence Analysis
- Dialogue Systems

Introduction

- Coreference resolution:

  Identify all noun phrases (mentions) that refer to the same entity in a text

Chinese national football team has a match at 10pm today. Their opponent is the Iran national football team. In the past three years, they have always lost to the Iran team.

Introduction

- Coreference resolution:

  Identify all noun phrases (mentions) that refer to the same entity in a text

Chinese national football team has a match at 10pm today. Their opponent is the Iran national football team. In the past three years, they have always lost to the Iran team.

Introduction

- Coreference resolution:

  Identify all noun phrases (mentions) that refer to the same entity in a text

Chinese national football team has a match at 10pm today. Their opponent is the Iran national football team. In the past three years, they have always lost to the Iran team.

Introduction

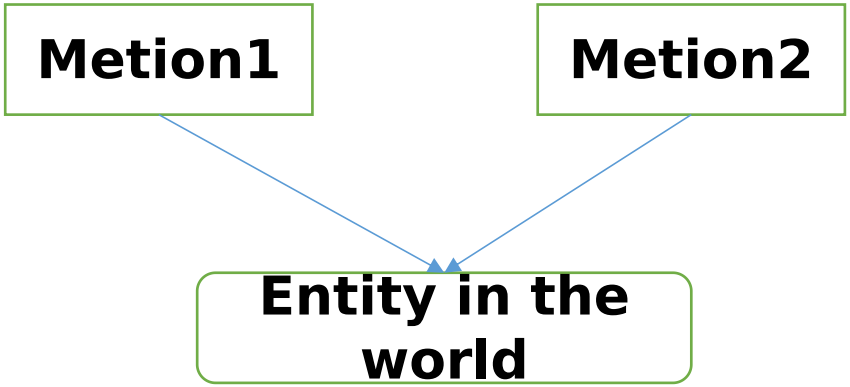
- Some noun phrases, which refers to entities, nest inside others

Jack Ma , the co-founder of Alibaba Group , claims that he is always a rural teacher rather than a CEO of the company .

Kinds of referring

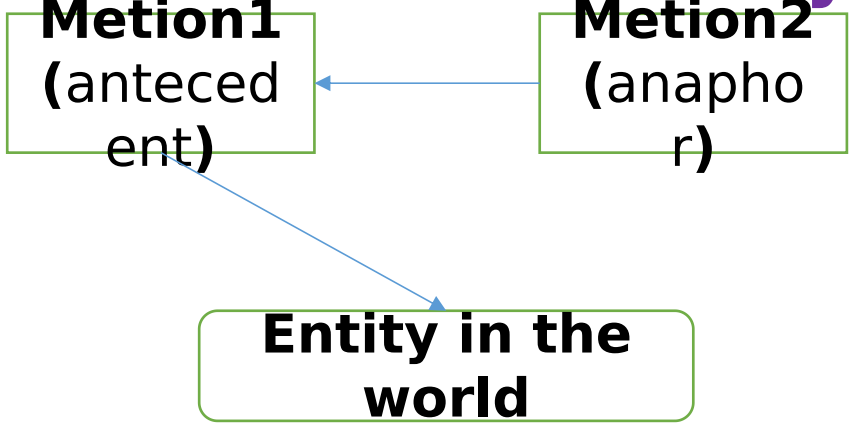
- Co-referring noun phrase
  - Two or more noun phrases refer to the same entity in the world

- E.g. **Shaquille O'Neal** is a NBA superstar, who won 4 champions. In 2011, **the big shark** announced retirement.



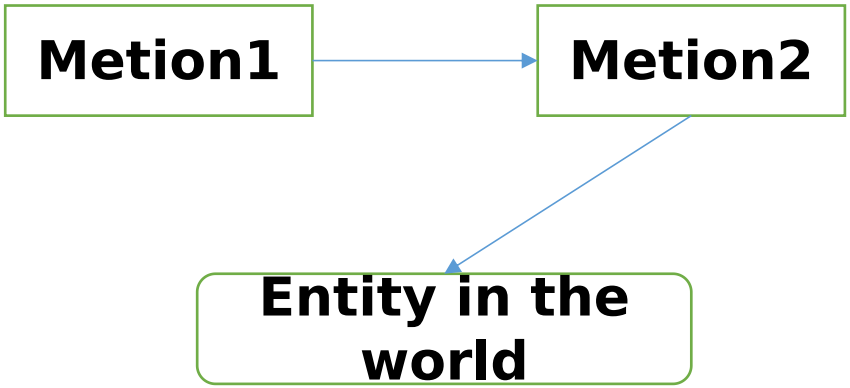
Kinds of referring

- Anaphora
  - A term (anaphor) refers to another term (antecedent)
  - Derives interpretation of the anaphor from the previously expressed antecedent
  - `Ana' in Greek = backwards
- E.g. Look at the **stars**. **They** are so bright.



Kinds of referring

- Cataphora
  - A term refers to a later term
  - Derives interpretation of the former term from the later expressed term
  - `Cata' in Greek = forwards
- E.g. Leaving **his** hometown, **Forrest** pursues the American dream.



Applications

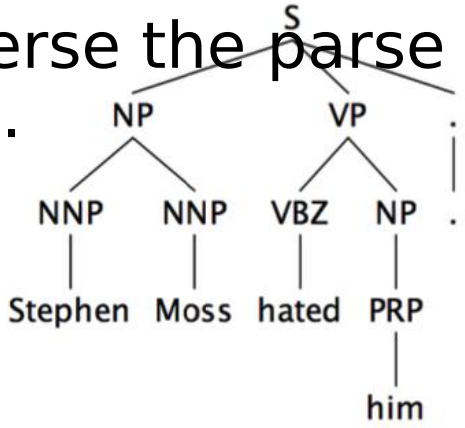
- Full-text understanding:
  - Machine translation
  - Text summarization
  - Information extraction
  - Question answering
  - ……

Outline

- **Coreference Resolution**
  - Introduction
  - **Unsupervised Coreference Resolution**
  - Supervised Coreference Resolution
- Coherence Analysis
- Dialogue Systems

Hobbs Algorithm

- Rule-based pronominal anaphora resolution
- Traverse on the parse tree
- Some rules:
  - Go up tree to the first NP or S. Called this X, and the path p.
  - Left-to-right, breadth-first branches below X to the left of p, to find the proper NP antecedent.
  - If X is S and find no NP antecedents in previous step, traverse the parse tree of the previous sentences.
  - ……



Coreference as Clustering

- Define the distance between mentions
- Clustering mentions referring to the same entity
- **Features for representing a mention pair:**
- (1) Strong Coreference indicator:
  - String match
  - Alias
  - Appositive

Unsupervised Models for Coreference Resolution. EMNLP 2008.

Coreference as Clustering

- Define the distance between mentions
- Clustering mentions referring to the same entity
- **Features for representing a mention pair:**
- (2) Linguistic Constraints
  - Whether the mentions agree in Gender:
    He ->John; She->Mary; it -> car
  - Whether the mentions agree in Number
    Singular pronouns (he/she/it/his/…) refer to singular entities
    Plural pronouns (we/they/us/them) refer to plural entities
  - Have the same semantic class

Features are complicated

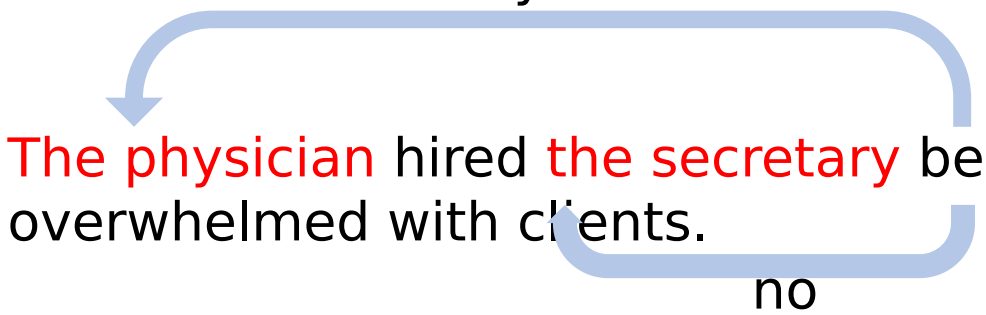
- Common nouns can differ in number but be coreferent:
  - **A patrol** … **the soldiers**
- Common nouns can refer to proper nouns
  - **Jeff Bezos** … **the richest man of the world**
- Gendered pronouns can refer to inanimate things
  - **India** withdrew **her** ambassador from the Commonwealth.
- Split antecedence

Outline

- **Coreference Resolution**
  - Introduction
  - Unsupervised Coreference Resolution
  - **Supervised Coreference Resolution**
- Coherence Analysis
- Dialogue Systems

Supervised Method

- Given a mention pair, classify whether one mention refers to the other
- Identify entities and pronouns
  - A syntactic parser or mention detector
- Obtain positive examples from training data and generate negative examples from unmatched mentions
  - binary classification task / ranking task



Feature-based Model

- Build mention features $\varphi_a(x)$ as mention representation:

$$\boldsymbol{h}_{\mathrm{a}}(x) \triangleq \tanh(\boldsymbol{W}_{\mathrm{a}}\,\boldsymbol{\phi}_{\mathrm{a}}(x) + \boldsymbol{b}_{\mathrm{a}})$$

- Build pairwise features $\varphi_p(x, y)$ as pairwise representation:

$$\boldsymbol{h}_{\mathrm{p}}(x, y) \triangleq \tanh(\boldsymbol{W}_{\mathrm{p}}\,\boldsymbol{\phi}_{\mathrm{p}}(x, y) + \boldsymbol{b}_{\mathrm{p}})$$

- Compute the coreference score for the mention pair:

$$s(x, y) \triangleq \begin{cases} \boldsymbol{u}^{\mathsf{T}}\boldsymbol{g}\left(\begin{bmatrix} \boldsymbol{h}_{\mathrm{a}}(x) \\ \boldsymbol{h}_{\mathrm{p}}(x,y) \end{bmatrix}\right) + u_0 & \text{if } y \neq \epsilon \\ \boldsymbol{v}^{\mathsf{T}}\boldsymbol{h}_{\mathrm{a}}(x) + v_0 & \text{if } y = \epsilon \end{cases}$$

Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution. ACL 2015

Feature-based Model

	Mention Features (ϕ_a)	
	Feature	Value Set
	Mention Head	$\mathcal{V}$
	Mention First Word	$\mathcal{V}$
	Mention Last Word	$\mathcal{V}$
	Word Preceding Mention	$\mathcal{V}$
	Word Following Mention	$\mathcal{V}$
	# Words in Mention	$\{1, 2, \ldots\}$
	Mention Synt. Ancestry	see BCS (2013)
	Mention Type	$\mathcal{T}$
+	Mention Governor	$\mathcal{V}$
+	Mention Sentence Index	$\{1, 2, \ldots\}$
+	Mention Entity Type	NER tags
+	Mention Number	{sing.,plur.,unk}
+	Mention Animacy	{an.,inan.,unk}
+	Mention Gender	{m,f,neut.,unk}
+	Mention Person	{1,2,3,unk}

	Pairwise Features (ϕ_p)	
	Feature	Value Set
	BASIC features on Mention	see above
	BASIC features on Antecedent	see above
	Mentions between Ment., Ante.	{0...10}
	Sentences between Ment., Ante.	{0...10}
	i-within-i	{T,F}
	Same Speaker	{T,F}
	Document Type	{Conv.,Art.}
	Ante., Ment. String Match	{T,F}
	Ante. contains Ment.	{T,F}
	Ment. contains Ante.	{T,F}
	Ante. contains Ment. Head	{T,F}
	Mention contains Ante. Head	{T,F}
	Ante., Ment. Head Match	{T,F}
	Ante., Ment. Synt. Ancestries	see above
+	BASIC+ features on Ment.	see above
+	BASIC+ features on Ante.	see above
+	Ante., Ment. Numbers	see above
+	Ante., Ment. Genders	see above
+	Ante., Ment. Persons	see above
+	Ante., Ment., Entity Types	see above
+	Ante., Ment. Heads	see above
+	Ante., Ment. Types	see above

Feature-based Model

- Evaluation metrics
  - *MUC: Link based; Counts the number of common links and computes F-measure*
  - B^3: *Precision & recall for entities in a reference chain*
  - *CEAF: Entity based; two variants*

System	MUC			B^3			$CEAF_e$			
	P	R	F_1	P	R	F_1	P	R	F_1	CoNLL
BCS (2013)	74.89	67.17	70.82	64.26	53.09	58.14	58.12	52.67	55.27	61.41
Prune&Score (2014)	81.03	66.16	**72.84**	66.9	51.10	57.94	68.75	44.34	53.91	61.56
B&K (2014)	74.3	67.46	70.72	62.71	54.96	58.58	59.4	52.27	55.61	61.63
D&K (2014)	72.73	69.98	71.33	61.18	56.60	58.80	56.20	54.31	55.24	61.79
This work ($\boldsymbol{g}_2$)	76.96	68.10	72.26	66.90	54.12	59.84	59.02	53.34	56.03	62.71
This work ($\boldsymbol{g}_1$)	76.23	69.31	72.60	66.07	55.83	**60.52**	59.41	54.88	**57.05**	**63.39**

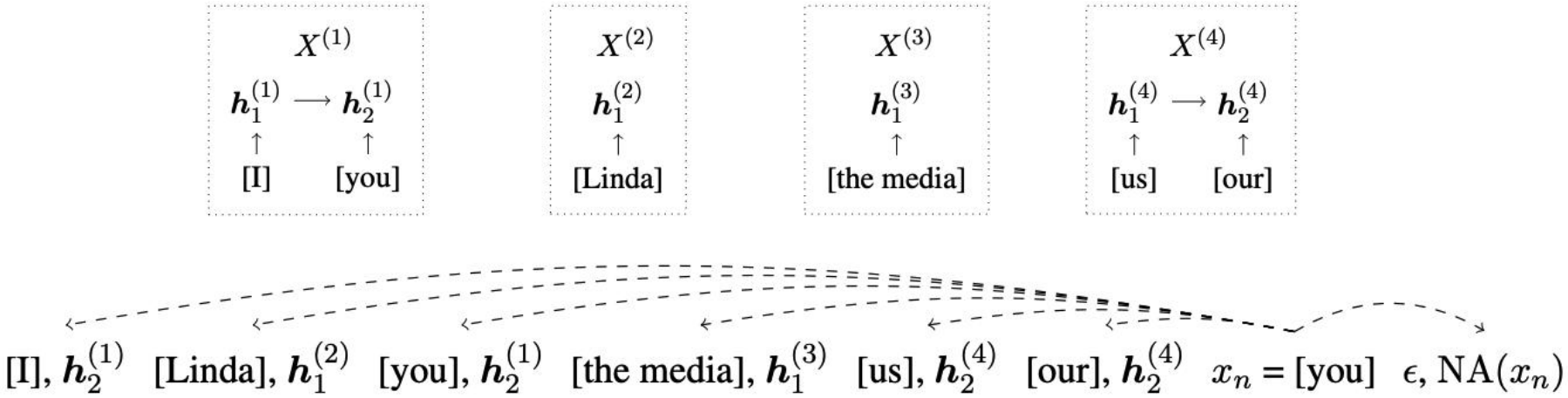
Global Features

Speaker1: um and [I] think that is what's - Go ahead Linda.

Speaker2: Well and uh thanks goes to [you] and to the media to help us... So our hat is off to all of [you] as well.

- [you] is coreferent with the singular [I]
- [you], as part of the phrase "all of you," is evidently plural
- Previous feature-based strategy may predict [you] to be an antecedent of [you]
- Have access to the history of predictions ?

Global Features



- Apply RNN to learn global representations of entity clusters from their mentions
  - Map mention's feature to vector-space representation:

$$\boldsymbol{h}_{\mathrm{c}}(x_n) \triangleq \tanh(\boldsymbol{W}_{\mathrm{c}}\,\boldsymbol{\phi}_{\mathrm{a}}(x_n) + \boldsymbol{b}_{\mathrm{c}}) \qquad \boldsymbol{h}_j^{(m)} \leftarrow \mathbf{RNN}(\boldsymbol{h}_{\mathrm{c}}(X_j^{(m)}), \boldsymbol{h}_{j-1}^{(m)}; \boldsymbol{\theta})$$

- When testing, apply greedy search clusters

Entity-level representation

- Fewer hand-engineered features ?
- Easy to extend to multiple language ?

Entity-level representation

- Fewer hand-engineered features ?
- Easy to extend to multiple language ?

- Word embedding !

  e.g. word2vec

Entity-level representation

- Fewer hand-engineered features ?
- Easy to extend to multiple language ?

- Word embedding !

  e.g. word2vec

- More complex neural network based on pre-trained word embedding
- Don’t need to design features for each language

Entity-level representation

- Cluster {Bill Clinton} and Cluster {C , she}

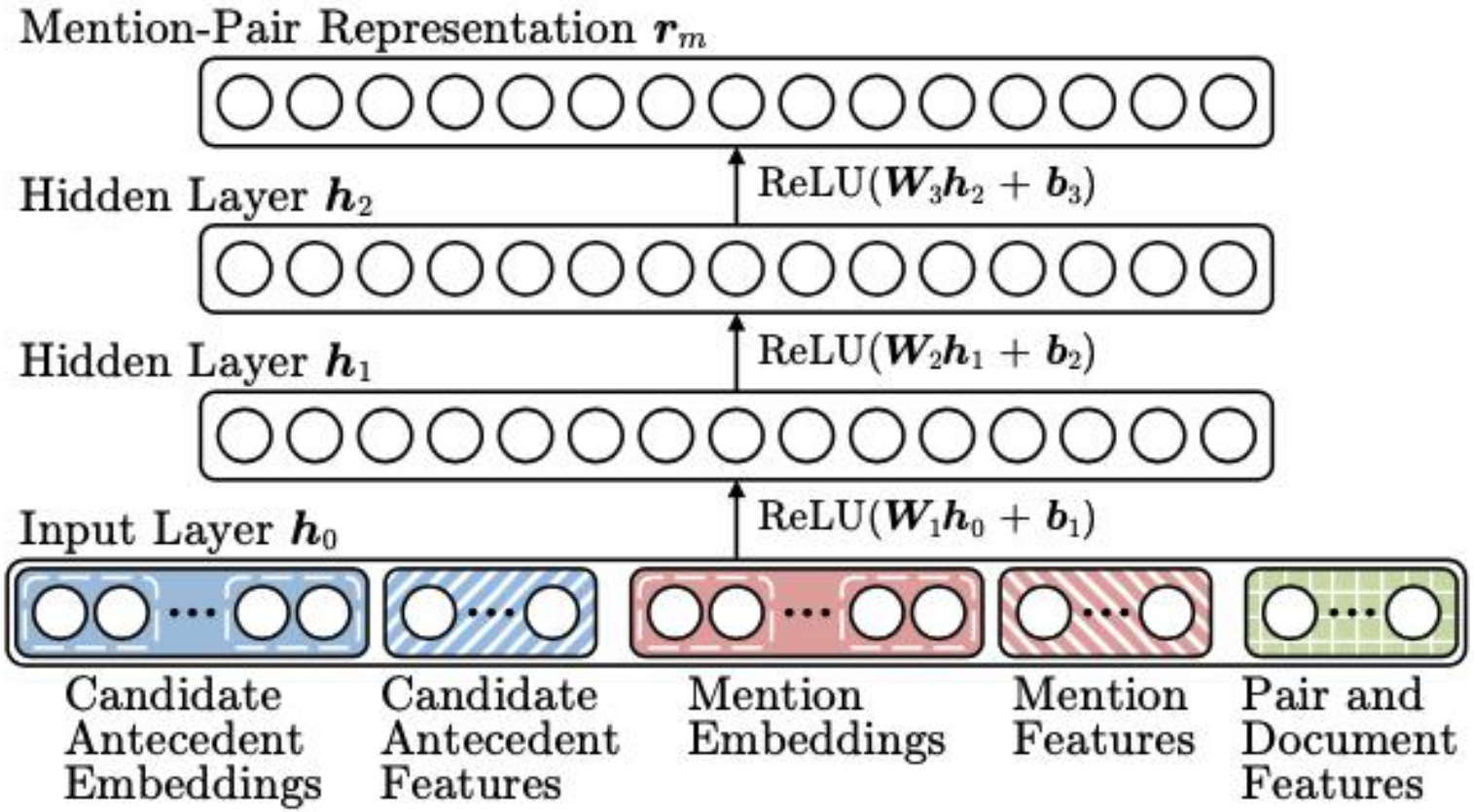
- Cluster {Bill Clinton} and Cluster {Clinton}

- Strategy: Merge cluste gradually, considering mutual relation between cross-

Improving Coreference Resolution by Learning Entity-Level Distributed Representations. *ACL 2016*

Entity-level representation

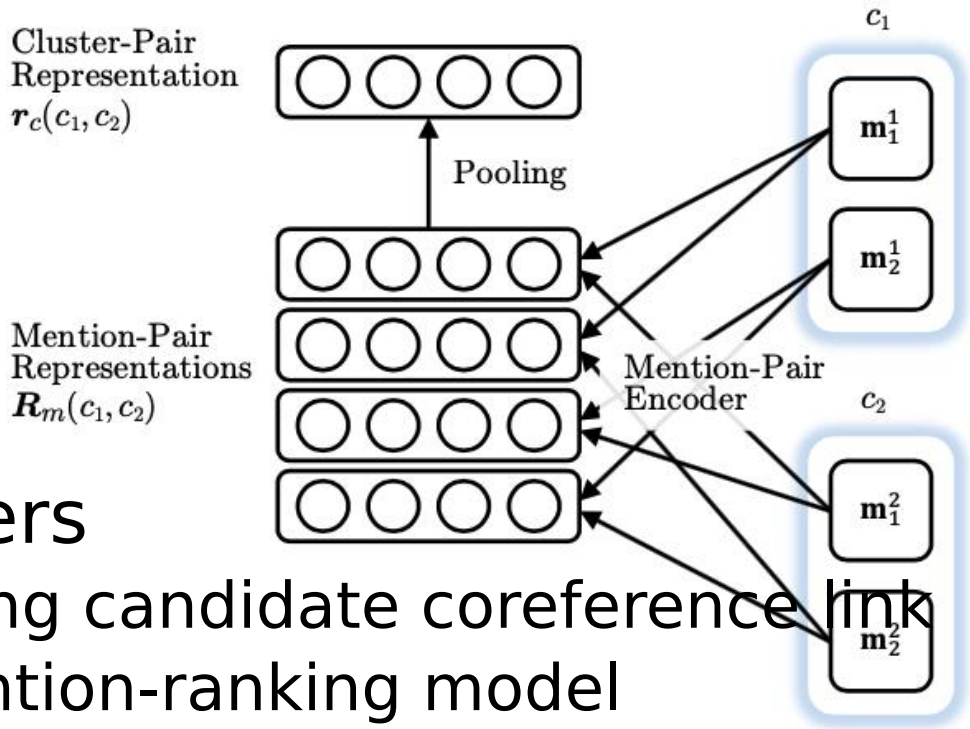
- Mention-pair Encoder
  - Produce a distributed representation for mention pairs



Entity-level representation

- Cluster-pair Encoder
  - Consider distance between two clusters
- Learn to merge clusters
  - Sort the highest scoring candidate coreference link according to the mention-ranking model
  - Action: Merge | Pass
  - Train policy network for select each action that increases B^3 metric each step
  - Further apply Reinforcement Learning

(*Deep Reinforcement Learning for Mention-Ranking Coreference Models, EMNLP2016*)



Entity-level representation

- Evaluation
  - Benefit from general word embedding, model can be applied on multiple languages
  - Cluster ranker significantly improves performance

	MUC			B^3			$CEAF_{\phi4}$			
	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Avg. F_1
CoNLL 2012 English Test Data										
Clark and Manning (2015)	76.12	69.38	72.59	65.64	56.01	60.44	59.44	52.98	56.02	63.02
Peng et al. (2015)	–	–	72.22	–	–	60.50	–	–	56.37	63.03
Wiseman et al. (2015)	76.23	69.31	72.60	66.07	55.83	60.52	59.41	54.88	57.05	63.39
Wiseman et al. (2016)	77.49	69.75	73.42	66.83	56.95	61.50	62.14	53.85	57.70	64.21
NN Mention Ranker	79.77	69.10	74.05	69.68	56.37	62.32	63.02	53.59	57.92	64.76
NN Cluster Ranker	78.93	69.75	**74.06**	70.08	56.98	**62.86**	62.48	55.82	**58.96**	**65.29**
CoNLL 2012 Chinese Test Data										
Chen & Ng (2012)	64.69	59.92	62.21	60.26	51.76	55.69	51.61	58.84	54.99	57.63
Björkelund & Kuhn (2014)	69.39	62.57	65.80	61.64	53.87	57.49	59.33	54.65	56.89	60.06
NN Mention Ranker	72.53	65.72	68.96	65.49	56.87	60.88	61.93	57.11	59.42	63.09
NN Cluster Ranker	73.85	65.42	**69.38**	67.53	56.41	**61.47**	62.84	57.62	**60.12**	**63.66**

End-to-end Model

- Previous mention detectors:

(1)Using a syntactic parser

(2)Hand-engineered mention detector

(3) Loss_{mention detector} + Loss_{coreference detector}

End-to-end Model

- End-to-end training: learn a conditional probability distribution to produce the correct clustering

$$P(y_1, \ldots, y_N \mid D) = \prod_{i=1}^{N} P(y_i \mid D)$$

$$= \prod_{i=1}^{N} \frac{\exp(s(i, y_i))}{\sum_{y' \in \mathcal{Y}(i)} \exp(s(i, y'))}$$

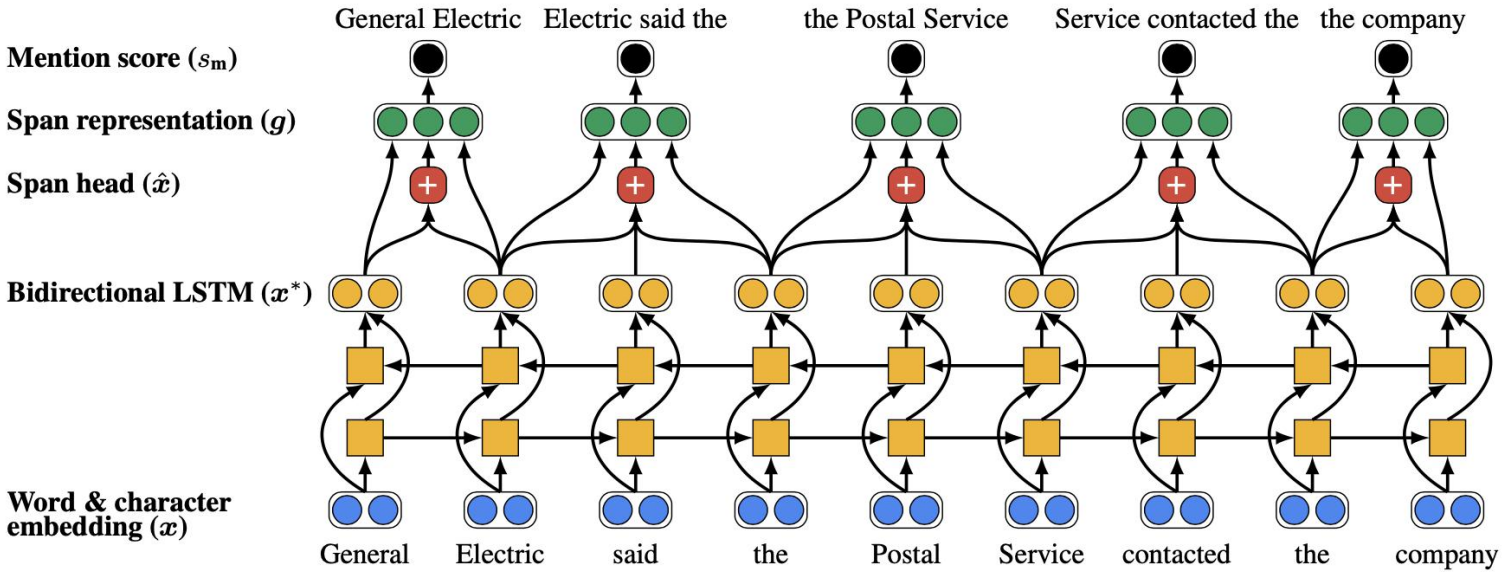
- Coreference Score $s(i, j)$ depends on:

(1) span i is a mention (2) span j is a mention

(3) span j is an antecedent of span i

$$s(i,j) = \begin{cases} 0 & j = \epsilon \\ s_{\mathrm{m}}(i) + s_{\mathrm{m}}(j) + s_{\mathrm{a}}(i,j) & j \neq \epsilon \end{cases}$$

End-to-end Model



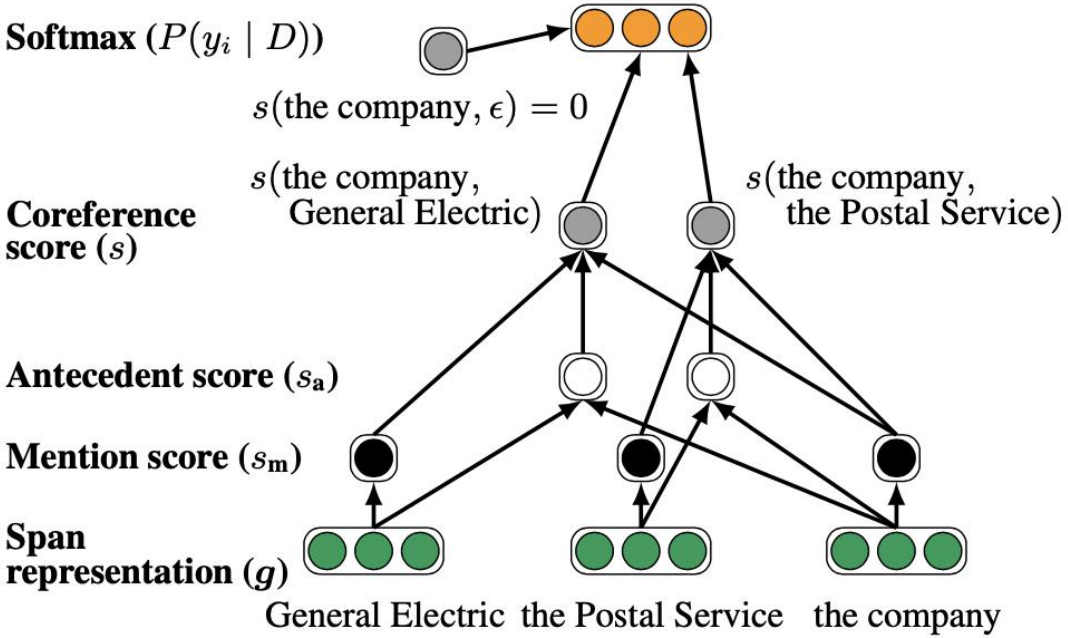
- Obtain the representation of span i:

$$g_i = [x^*_{START(i)}, x^*_{END(i)}]$$

- Low-scoring spans are pruned

End-to-end Neural Coreference Resolution. EMNLP 2017.

End-to-end Model



- Assign to each span i and antecedent $y_i \in \{\epsilon, 1..., i-1\}$
- Compute the pairwise linking score from pairs of span representations

$$s_a(i,j) = w_a \cdot FFNN_a(g_i,\ g_j,\ g_i \circ g_j,\ \varphi(i,j))$$

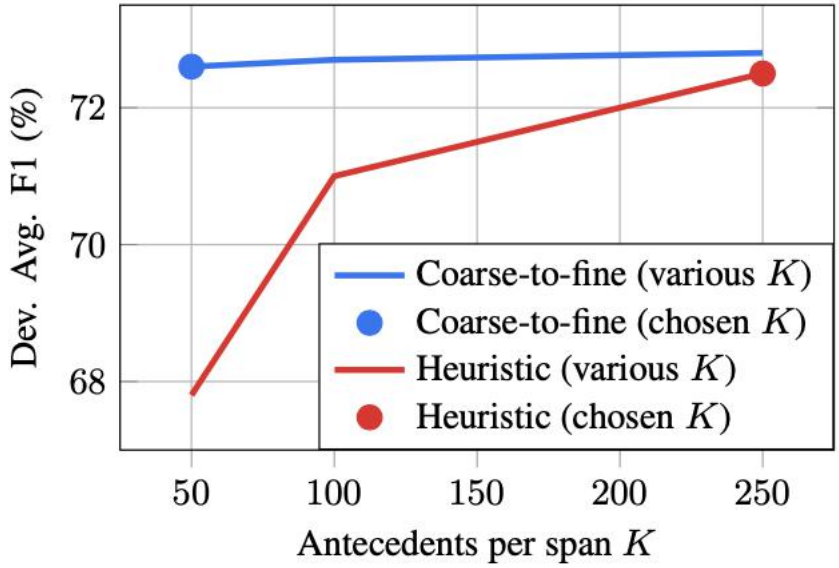
Higher-order, Coarse-to-fine

Speaker 1: Um and **[I]** think that is what's - Go ahead Linda.
Speaker 2: Well and uh thanks goes to **[you]** and to the media to help us... So our hat is off to **[all of you]** as well.

- Higher-order representation (likes an antecedent GNN):
  - g_i^n is computed with an attention mechanism that averages over previous representations g_j^{n-1} weighted according to how likely each mention j is to be an antecedent for i

Higher-order Coreference Resolution with Coarse-to-fine Inference. NAACL 2018.

Higher-order, Coarse-to-fine



- Heuristic:

  Select the top nearest K mention as antecedents to compute scores for saving time

- Coarse-to-fine inference
  - Incorporates a less accurate but more efficient bilinear scorer, enabling more aggressive pruning without hurting accuracy

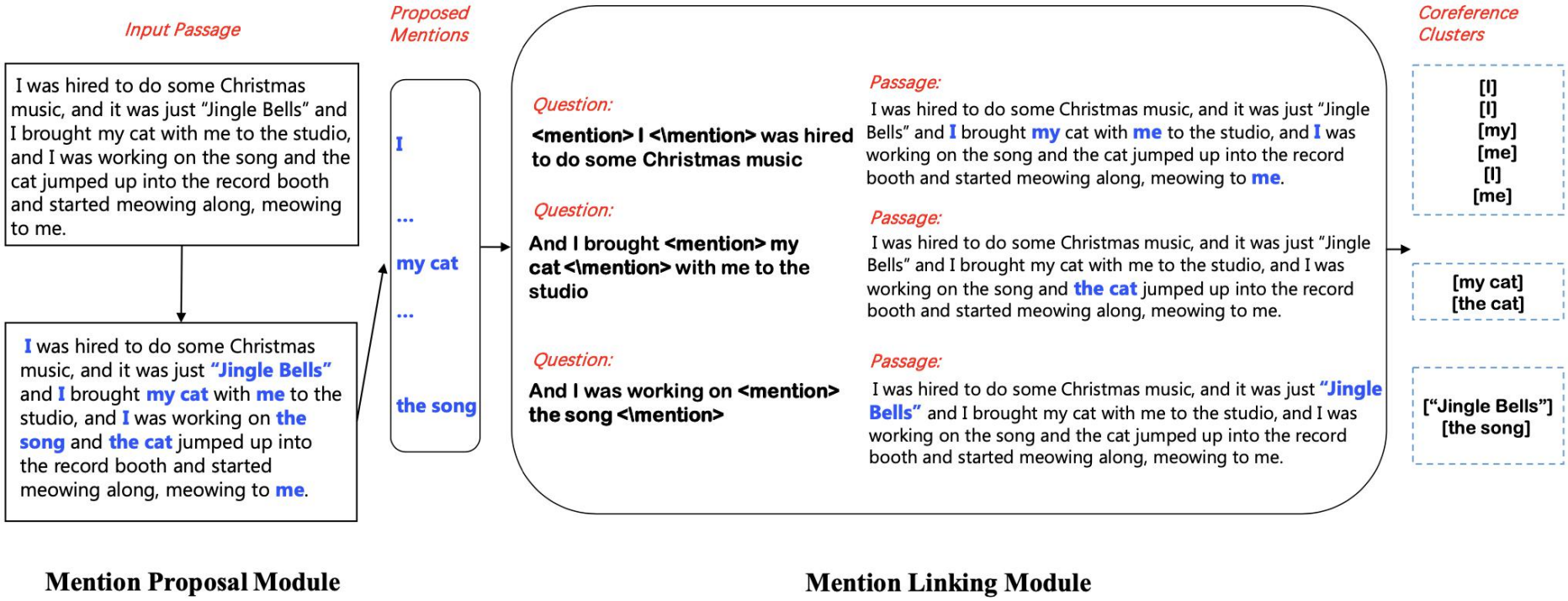
Higher-order Coreference Resolution with Coarse-to-fine Inference. NAACLs 2018.

Higher-order, Coarse-to-fine

- Evaluation
  - End-to-end model outperforms previous methods with separate mention detector and coreference linker
  - Higher-order representation & Coarse-to-fine inference work

	MUC			B^3			$CEAF_{\phi_4}$			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Avg. F1
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Lee et al. (2017)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
+ ELMo (Peters et al., 2018)	80.1	77.2	78.6	69.8	66.5	68.1	66.4	62.9	64.6	70.4
+ hyperparameter tuning	80.7	78.8	79.8	71.7	68.7	70.2	67.2	66.8	67.0	72.3
+ coarse-to-fine inference	80.4	**79.9**	80.1	71.0	**70.0**	70.5	67.5	**67.2**	67.3	72.6
+ second-order inference	**81.4**	79.5	**80.4**	**72.2**	69.5	**70.8**	**68.2**	67.1	**67.6**	**73.0**

Coreference as QA



- Transfer coreference resolution task to machine reading comprehension

Coreference Resolution as Query-based Span Prediction. 2019.

Coreference as QA

	MUC			B^3			$CEAF_{\phi_4}$			
	P	R	F1	P	R	F1	P	R	F1	Avg. F1
e2e-coref(Lee et al., 2017)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
c2f-coref + ELMo (Lee et al., 2018)	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
EE + BERT-large (Kantor and Globerson, 2019)	82.6	84.1	83.4	73.3	76.2	74.7	72.4	71.1	71.8	76.6
c2f-coref + BERT-large (Joshi et al., 2019b)	84.7	82.4	83.5	76.5	74.0	75.3	74.1	69.8	71.9	76.9
c2f-coref + SpanBERT-large (Joshi et al., 2019a)	85.8	84.8	85.3	78.3	77.9	78.1	76.4	74.2	75.3	79.6
CorefQA + SpanBERT-base	85.2	87.4	86.3	78.7	76.5	77.6	76.0	75.6	75.8	79.9 (+0.3)
CorefQA + SpanBERT-large	**88.6**	**87.4**	**88.0**	**82.4**	**82.0**	**82.2**	**79.9**	**78.3**	**79.1**	**83.1** (+3.5)

Evaluation results on the English CoNLL-2012 shared task

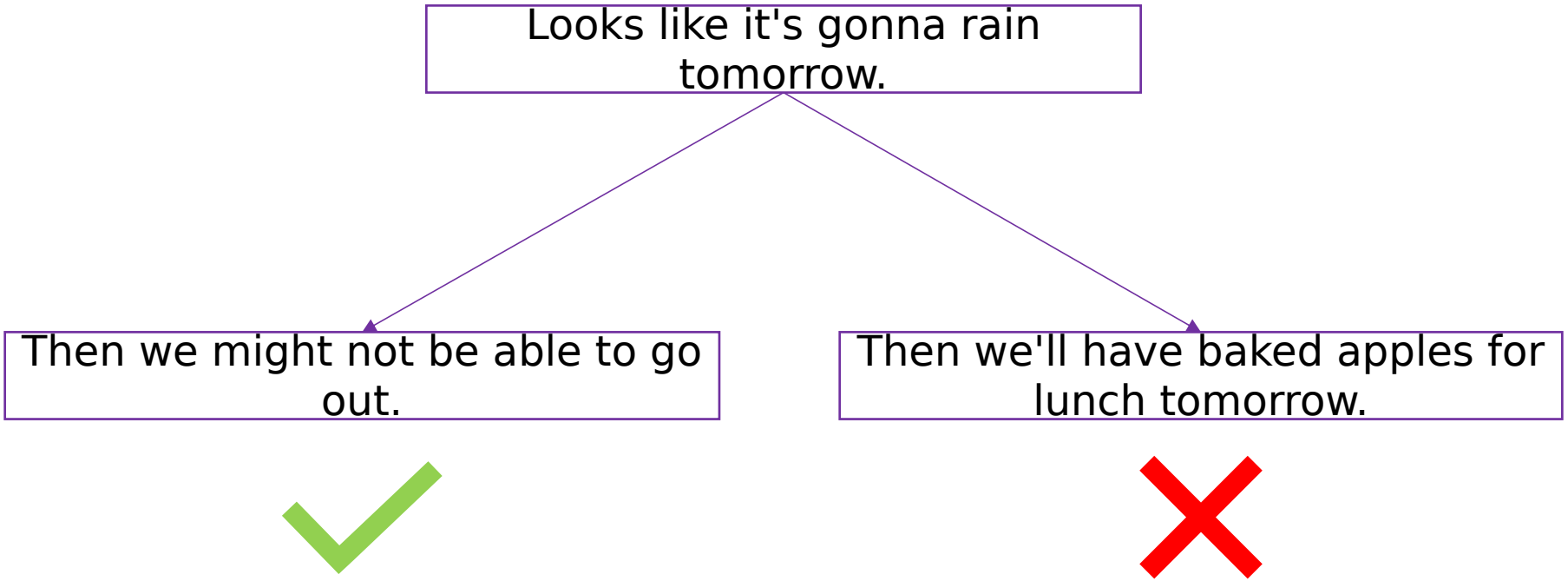
- Since CorefQA adopts a question answering framework, existing machine comprehension datasets can be used for **data augmentation** to improve model generalization capability

Outline

- Coreference Resolution
- **Coherence Analysis**
  - **Introduction**
  - Explicit Symbols, Implicit Relations and Feature Engineering
  - Deep Learning Based Methods
- Dialogue Systems

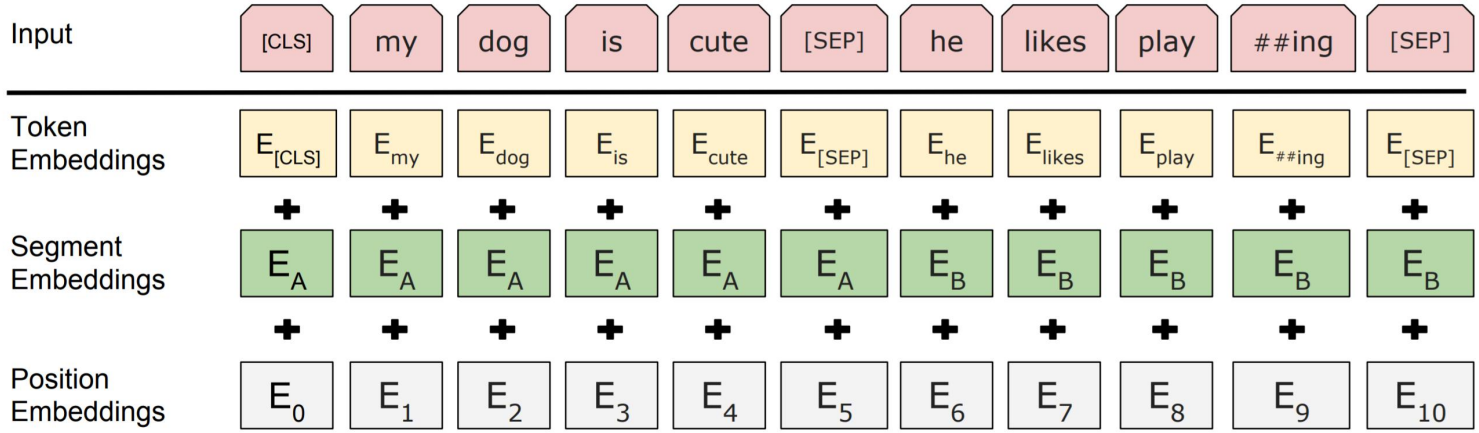
Coherence

- Coherence: Judging whether a context can be connected smoothly



Coherence in BERT

- Two pre-training tasks in BERT:
  - Recover masked words (Learn words and phrases)
  - Predict whether two sentences are contextual (Learn coherence)



Discourse Relation Classification

- Discourse Relation Classification: Predict the relationship between sentences

Income from continuing operations was up 26%

Expansion (Implicit)

Revenue rose 24% to $6.5 billion from $5.23 billion

as in summarily sacking exchange controls

Expansion (Explicit)

and in particular slashing the top rate of income taxation to 40%

Coherent & Discourse Relation Classification

- Similarities:
  - Two sentences as input
  - Consider context
  - …
- Differences:
  - Coherence is a scoring task or binary classification
  - Discourse relation classification is multilabel classification

Discourse Relation Classification & Pre-trained Language Model

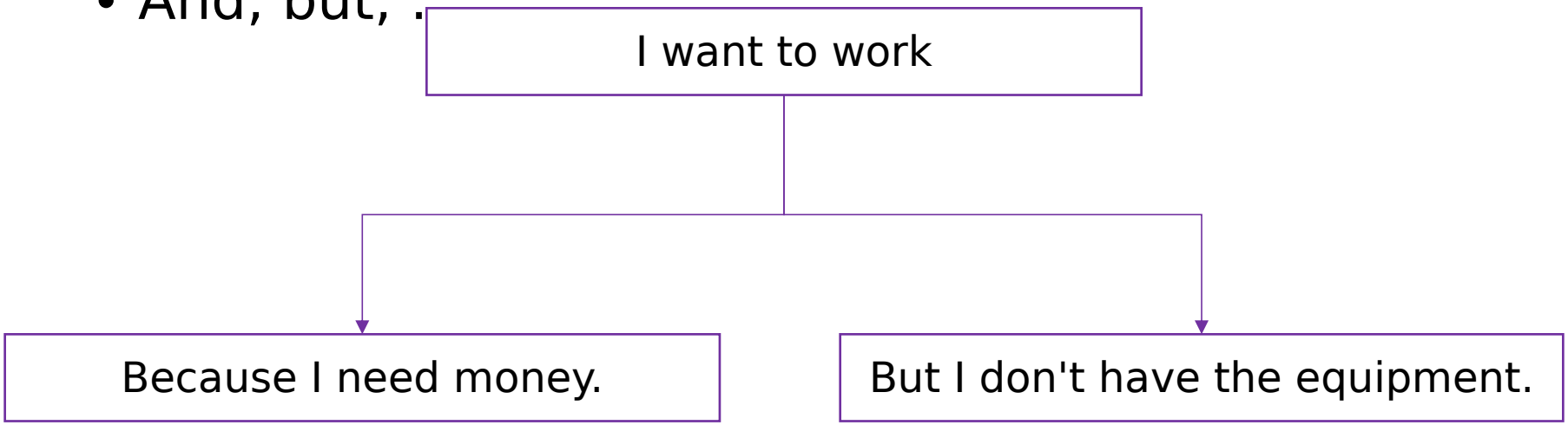
- Use coherence tasks in pre-trained language models
- What about using the task of predicting discourse relations in pre-training?
- Major challenge: unlabeled data

Outline

- Coreference Resolution
- **Coherence Analysis**
  - Introduction
  - **Explicit Symbols, Implicit Relations and Feature Engineering**
  - Deep Learning Based Methods
- Dialogue Systems

Explicit Symbols

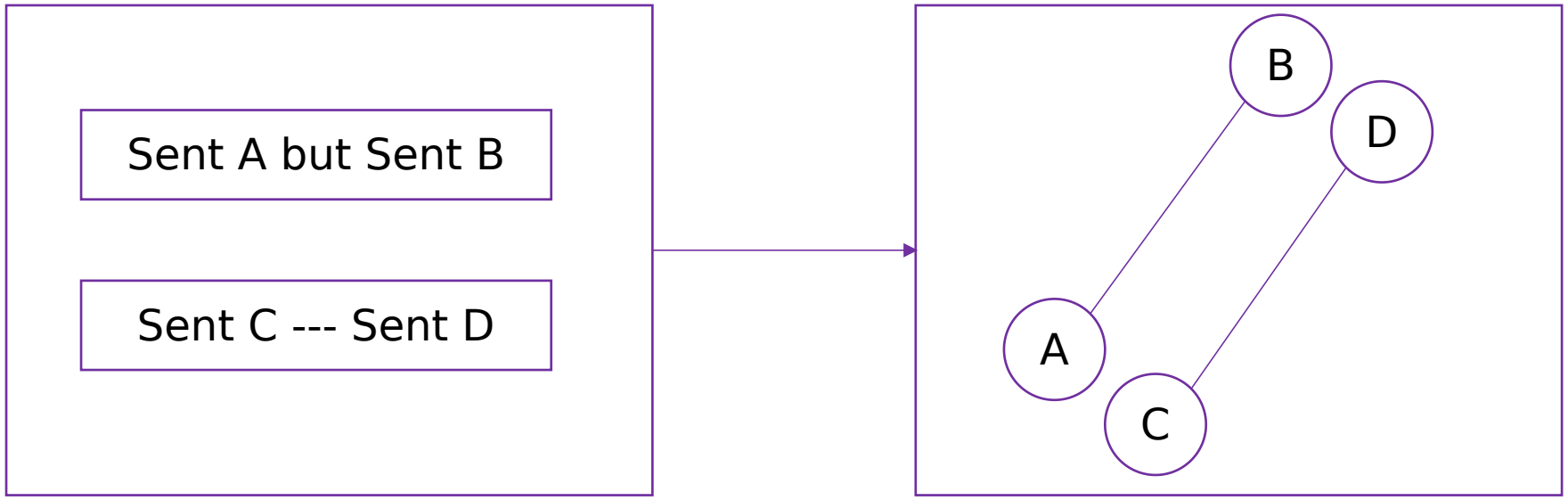
- Explicit Symbols:
  - The symbols that explicitly represent contextual relationships
  - And, but, ..



- With explicit symbols, existing methods can reach over 93% on accuracy and 90% on F1

Explicit Symbols

- We can use explicit symbols to build labels for explicit discourse relations to help pre-training



Implicit Relations

- How to deal with implicit relations?

- Possible Methods:
  - Embedding based methods
  - Keyword extraction
  - Feature engineering
  - ……

Cross Product of Words

- Cross product of words:
  - Sentence pair: (s_1, s_2)
  - Cross product of words:
    - $\forall w_1 \in s_1, w_2 \in s_2, \{(w_1, w_2)\}$

The recent explosion of country funds mirrors the "closed-end fund mania" of the 1920s, Mr. Foot says, when narrowly focused funds grew wildly **popular**. *They fell into* **oblivion** *after the 1929 crash.*

Automatic sense prediction for implicit discourse relations in text. ACL 2009.

Cross Product of Words

- $S_1 = (w_1, w_2, \cdots, w_n)$
- $S_2 = (w_{n+1}, w_{n+2}, \cdots, w_m)$
- $(S_1, S_2) = \{(w_i,\ w_j) | 1 \le i \le n < j \le m\}$
- What we want to learn:
- $\Pr[r|(S_1, S_2)]$
- Using a naïve Bayesian:
- $\Pr[r|(S_1, S_2)] = \frac{\Pr[r] \times \Pr[(S_1, S_2)|r]}{\sum \Pr[r'] \times \Pr[(S_1, S_2)|r']}$
- $\Pr[(S_1, S_2)|r] = \prod \Pr[(w_1, w_2)|r]$

Cross Product of Words

- Using pre-trained word embeddings (word2vec, glove, ...)

- Feature engineering:
  - The maximal distance between word pairs
  - The average distance between word pairs
  - Use feed-forward network to predict the relation
  - ...

Cross Product of Words

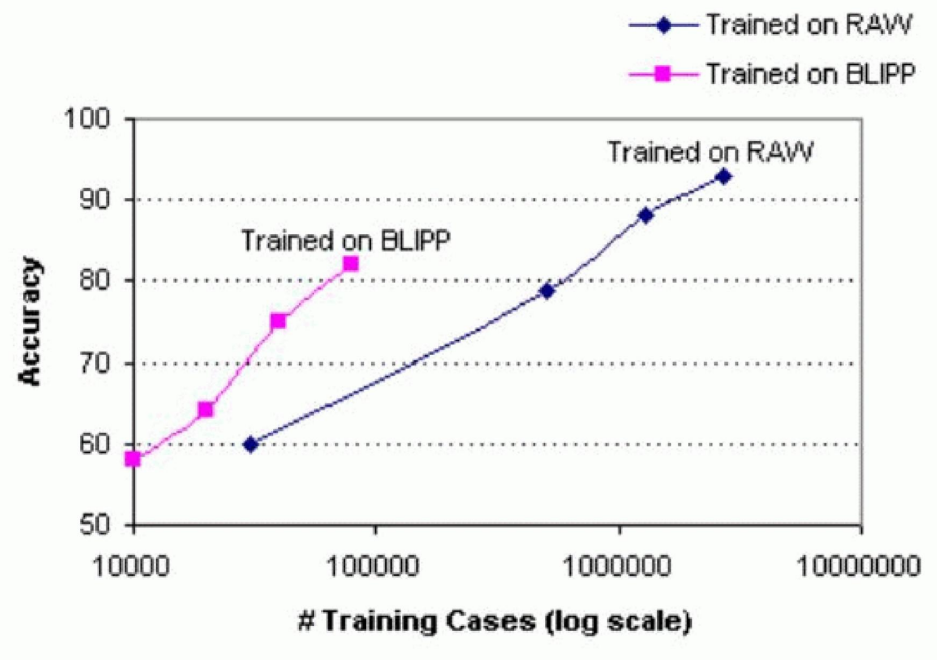
- Shortcomings of cross product of words:
  - The assumptions have not been verified yet
  - The number of features of cross product of words is numerous, which may lead to sparsity problems

- As a result, feature elimination and selection are required

Automatic sense prediction for implicit discourse relations in text. ACL 2009.

Cross Product of Words

- Why do we need to consider all the word pairs?
- Of course not!
- noun-noun, verb-verb, cue phrase-cue phrase



An unsupervised approach to recognizing discourse relations. ACL 2001.

Cross Product of Words

- Some other phenomena:
  - Verb pairs are more informative than noun pairs and adjective pairs. (Inferring sentence-internal temporal relations)
  - Stemming is useful
  - Only using high frequency words is useful
  - Removing stop-words will make it less effective (Building and Refining Rhetorical-Semantic Relation Models. NAACL 2007)
  - …

Polarity Tag

- Sentiment of words:
  - Positive
  - Negative
  - Both
  - Neutral

Sentence 1: Executives at Time Inc. Magazine Co., a subsidiary of Time Warner, have said the joint venture with Mr. Lang wasn't a good one.

Sentence 2:The venture, formed in 1986, was supposed to be Time's low-cost, safe entry into women's magazines.

Automatic sense prediction for implicit discourse relations in text. ACL 2009.

Inquirer Tag

- General Inquirer Lexicon (The General Inquirer: A Computer Approach to Content Analysis)
  - Rise vs. Fall
  - Pleasure vs. Pain
  - Understate

Automatic sense prediction for implicit discourse relations in text. ACL 2009.

Money & Percent & Number

- The emergence of numbers in a particular field also has special significance
- Last year my income was 1000 RMB, up 100% from the year before
- Such feature is probably genre-dependent

Automatic sense prediction for implicit discourse relations in text. ACL 2009.

Modality

- Modal words:
  - Can
  - Should
  - May
  - …
- If I were a wealthy man, I wouldn’t have to work hard.

Automatic sense prediction for implicit discourse relations in text. ACL 2009.

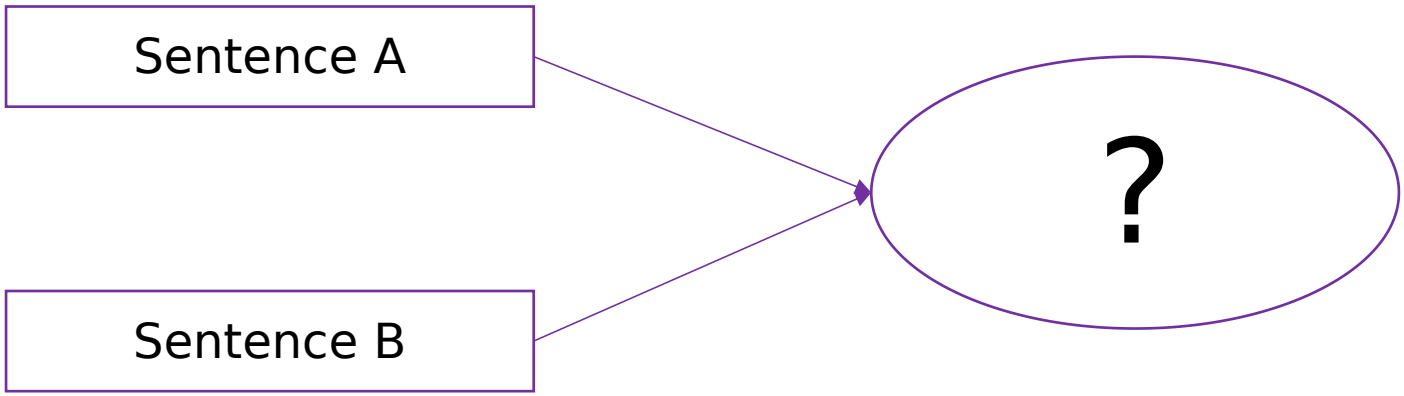
Outline

- Coreference Resolution
- **Coherence Analysis**
  - Introduction
  - Explicit Symbols, Implicit Relations and Feature Engineering
  - **Deep Learning Based Methods**
- Dialogue Systems

Basic Challenge

- The two sentences, which form the context, are two separate individuals

- The basic challenge
  - Map the context into the same vector space



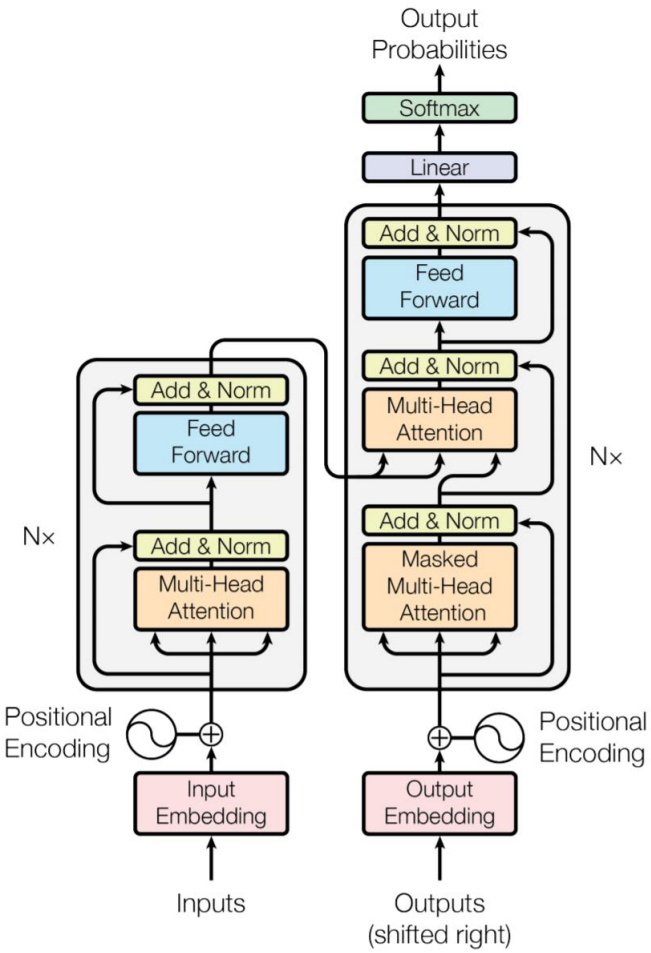
Basic Idea

- How to combine the information of two sentences?

- Attention!

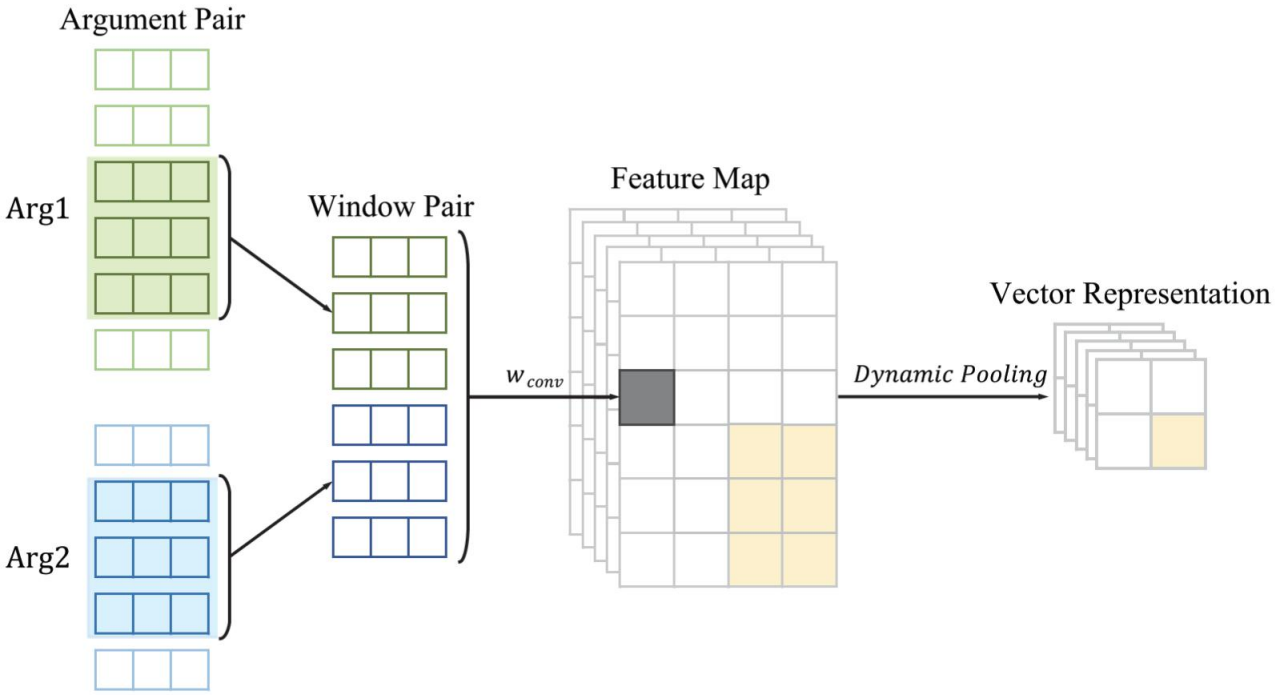
Basic Idea

- Attention -> Transformer -> BE
- Sentence 1 [SEP] Sentence 2
- Use the attention in transforme to obtain the information betwe two sentences



Similar Approach to Cross Product

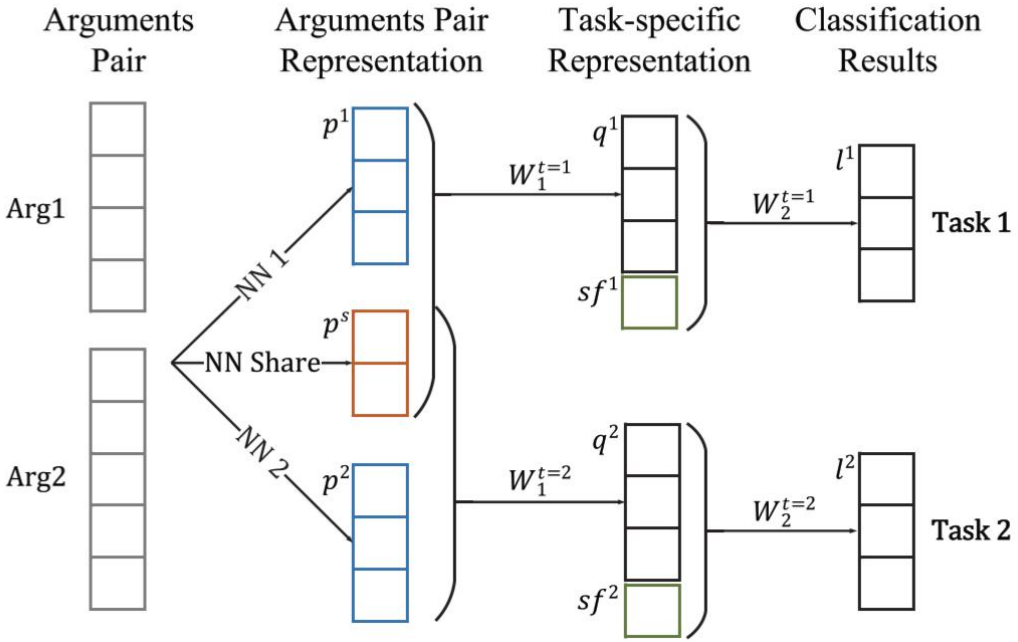
- Word pairs -> phrase (w words) pairs
- Use CNN to extract the features of phrases



Implicit Discourse Relation Classification via Multi-Task Neural Networks. AAAI 2016

Multi-task Learning

- Problems for applying deep learning methods:
  - Few datasets
  - No enough data
  - The tasks of different datasets are still different



Implicit Discourse Relation Classification via Multi-Task Neural Networks. AAAI 2016

Multi-task Learning

- Experimental Results

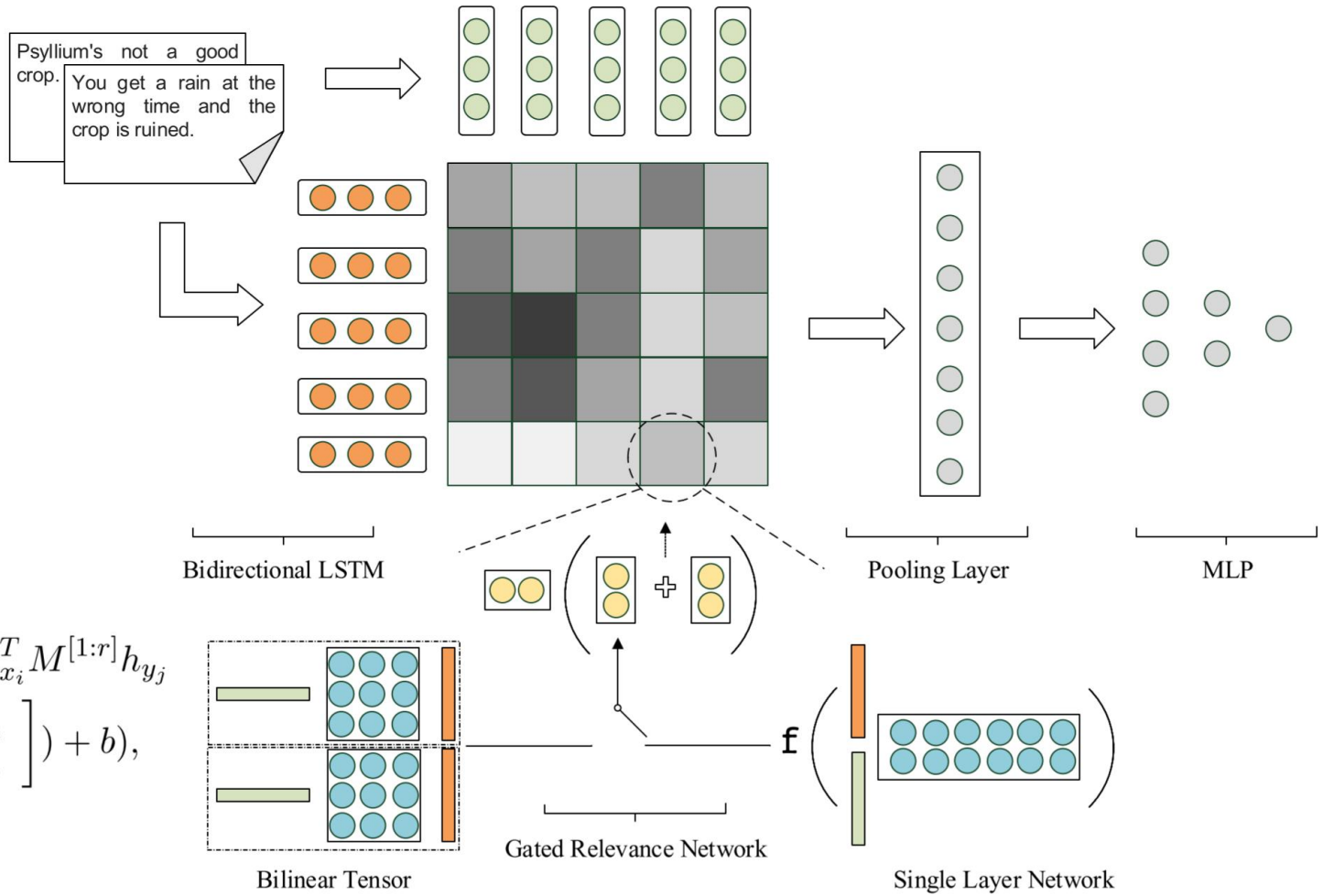
System	Comp.	Cont.	Expa.	Temp.
(Zhou et al. 2010)	31.79	47.16	-	20.30
(Park and Cardie 2012)	31.32	49.82	-	26.57
(Ji and Eisenstein 2015)	35.93	52.78	-	27.63
(R&X 2015)	**41.00**	53.80	69.40	33.30
Proposed STL	37.10	51.73	67.53	29.38
Proposed MTL	37.91	**55.88**	**69.97**	**37.17**

Learn Word Pairs by Bilinear

- Attention: Use matrix multiplication to learn features between words in sentences
- CNN: Learn features between phrases
- What else?
- Bilinear: Another form of attention to learn features between words

Implicit Discourse Relation Detection via a Deep Architecture with Gated Relevance Network. ACL 2016

Learn Word Pairs by Bilinear



$$s(h_{x_i}, h_{y_j}) = u^T(g \odot h_{x_i}^T M^{[1:r]} h_{y_j} + (1-g) \odot f(V \begin{bmatrix} h_{x_i} \\ h_{y_j} \end{bmatrix}) + b),$$

Implicit Discourse Relation Detection via a Deep Architecture with Gated Relevance Network. ACL 2016

Learn Word Pairs by Bilinear

	Comparison	Contingency	Expansion	Temporal
(Pitler et al., 2009)	21.96%	47.13%	76.42%	16.76%
(Zhou et al., 2010)	31.79%	47.16%	70.11%	20.30%
(Park and Cardie, 2012)	31.32%	49.82%	79.22%	26.57%
(Rutherford and Xue, 2014)	39.70%	54.42%	80.44%	28.69%
(Ji and Eisenstein, 2015)	35.93%	52.78%	80.02%	27.63%
LSTM	31.78%	45.39%	75.10%	19.65%
Bi-LSTM	31.97%	45.66%	75.13%	20.02%
Word+NTN	32.18%	46.45%	77.64%	21.60%
LSTM+NTN	36.82%	50.09%	79.88%	26.54%
Bi-LSTM+NTN	39.36%	53.74%	80.02%	28.41%
Word+GRN	32.67%	46.52%	77.68%	21.21%
LSTM+GRN	38.13%	52.25%	79.96%	27.15%
Bi-LSTM+GRN	**40.17%**	**54.76%**	**80.62%**	**31.32%**

Word Embedding

- In word pair method:
  - Good and great should be similar
  - Bad should be in the opposite direction of these two words
- In traditional word embedding methods:
  - Both good, great and bad will be close to each other
- What we want:
  - The embeddings of good and great are close
  - The embedding of bad is far away from them

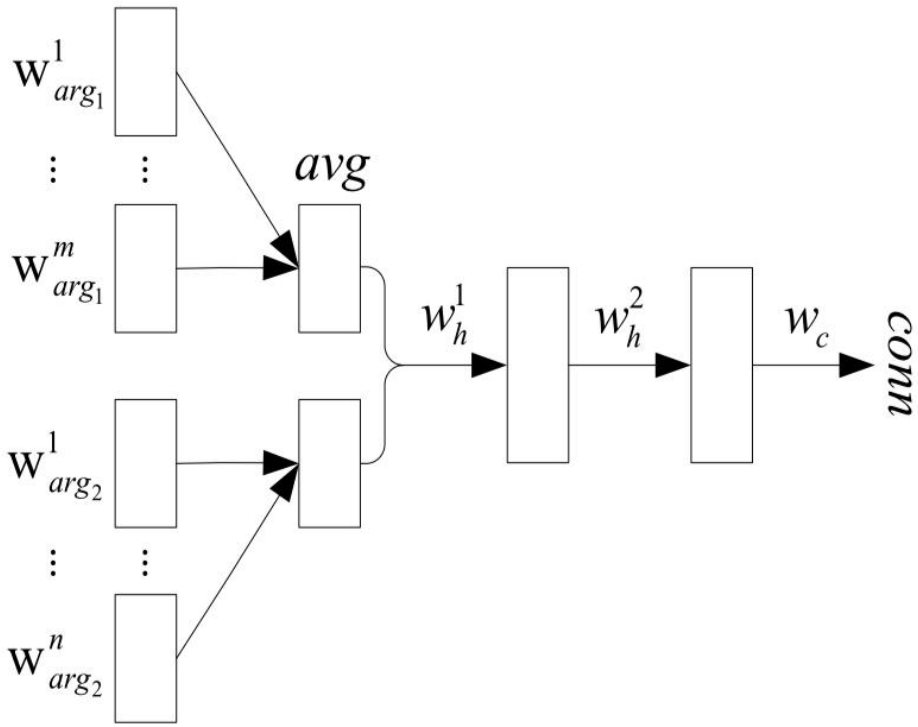
Word Embedding

- What are the main challenges for training word embeddings for discourse relation classification?
  - What kind of the model we want?
  - How to train the model from unlabeled data?

Word Embedding

- Explicit markers!

- Obtain Explicit Data
- Train on Explicit Data



Improving Implicit Discourse Relation Recognition with Discourse-specific Word Embeddings. ACL 2017

Word Embedding

CDRR		+GloVe	+word2vec	+DSWE
Temp	*P*	36.00	27.03	31.58
	R	16.36	18.18	21.82
	F_1	22.50	21.74	25.81
Comp	*P*	53.97	50.00	43.00
	R	23.45	20.00	29.66
	F_1	32.69	28.57	35.10
Cont	*P*	44.90	51.81	55.29
	R	40.29	36.63	42.12
	F_1	42.47	42.92	47.82
Expa	*P*	60.47	60.72	63.91
	R	76.21	81.60	79.00
	F_1	67.43	69.63	70.66
Accuracy		55.68	57.17	**58.85**
Macro F_1		41.27	40.71	**44.84**

not		*good*	
word2vec	*DSWE*	*word2vec*	*DSWE*
do	no	great	great
did	n't	bad	lot
anymore	never	terrific	very
necessarily	nothing	decent	better
anything	neither	nice	success
anyway	none	excellent	well
does	difficult	fantastic	happy
never	nor	better	certainly
want	refused	solid	respect
neither	impossible	lousy	fine
if	limited	wonderful	import
know	declined	terrible	positive
anybody	nobody	Good	help
yet	little	tough	useful
either	denied	best	welcome

Improving Implicit Discourse Relation Recognition with Discourse-specific Word Embeddings, ACL 2017

Adversarial

- Connective words are important!
- Accuracy with explicit markers: 85%
- Accuracy without explicit markers: 50%
- We need connective words to reach a better performance

[*Arg1*]: Never mind.

[*Arg2*]: You already know the answer.

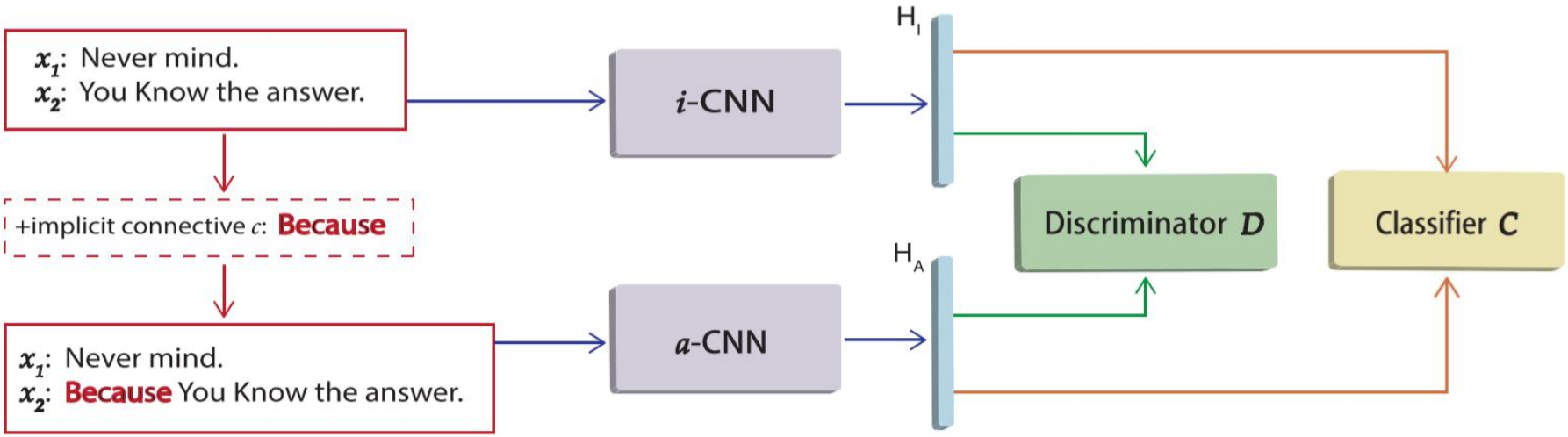
[*Implicit connective*]: Because

[*Discourse relation*]: Cause

Adversarial Connective-exploiting Networks for Implicit Discourse Relation Classification, ACL 2017

Adversarial

- Hard to predict without connective words
- We can use adversarial to solve the problem



Adversarial Connective-exploiting Networks for Implicit Discourse Relation Classification, ACL 2017

Adversarial

- Training Process:

  - While not converge:

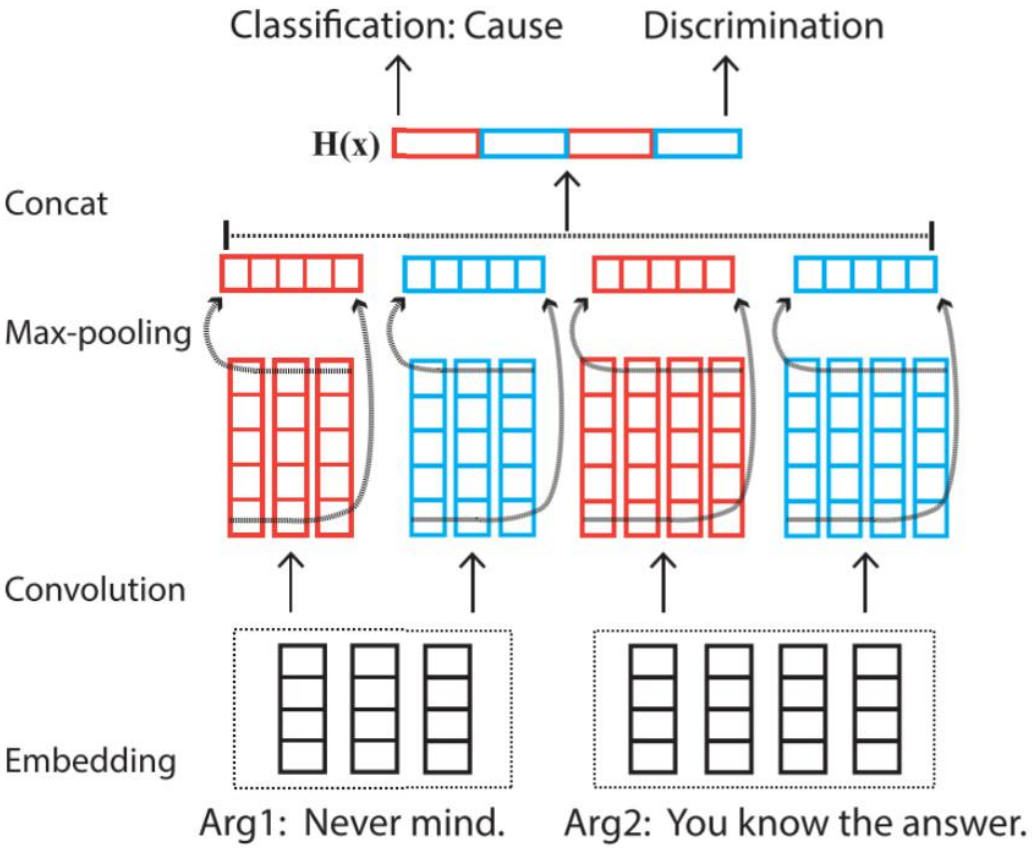
    Train the parameters of i-CNN, a-CNN and D

    Train the parameters of i-CNN, a-CNN and C

    output i-CNN and C as final models

Adversarial

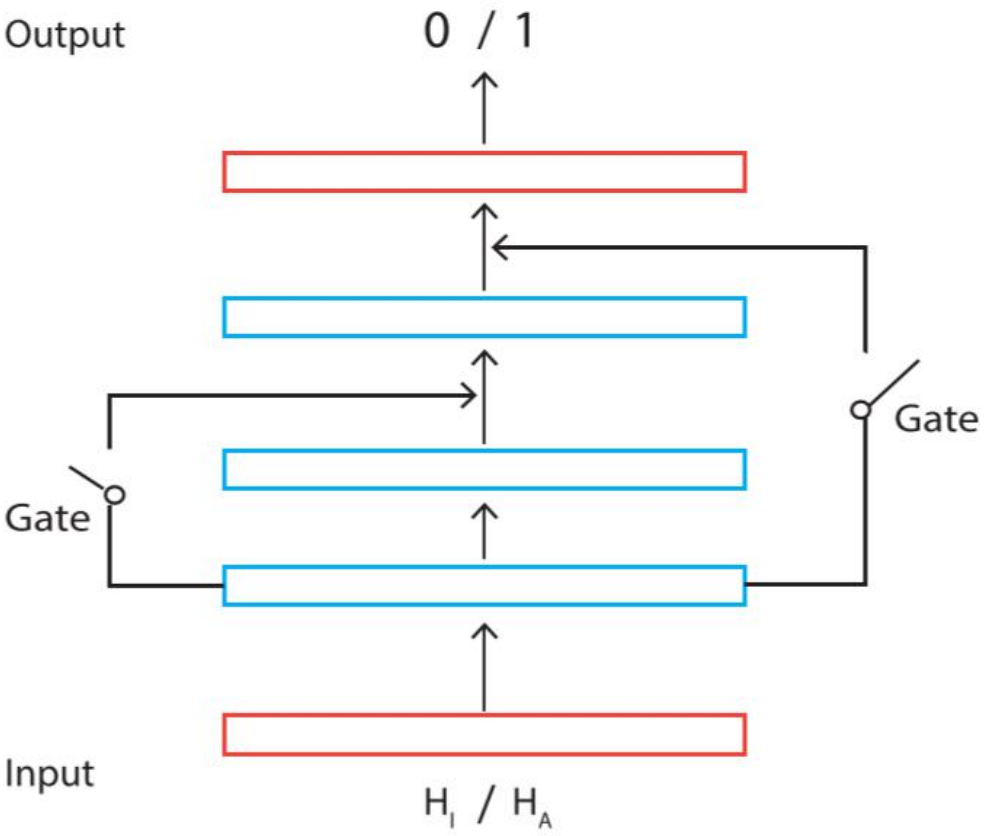
- i-CNN
- a-CNN



Adversarial Connective-exploiting Networks for Implicit Discourse Relation Classification. ACL 2017

Adversarial

- Discriminator



Adversarial Connective-exploiting Networks for Implicit Discourse Relation Classification, ACL 2017

Adversarial

	Model	PDTB-Lin	PDTB-Ji
1	Word-vector	34.07	36.86
2	CNN	43.12	44.51
3	Ensemble	42.17	44.27
4	Multi-task	43.73	44.75
5	ℓ_2-reg	44.12	45.33
6	Lin et al. (2009)	40.20	-
7	Lin et al. (2009) +Brown clusters	-	40.66
8	Ji and Eisenstein (2015)	-	44.59
9	Qin et al. (2016a)	43.81	45.04
10	Ours	**44.65**	**46.23**

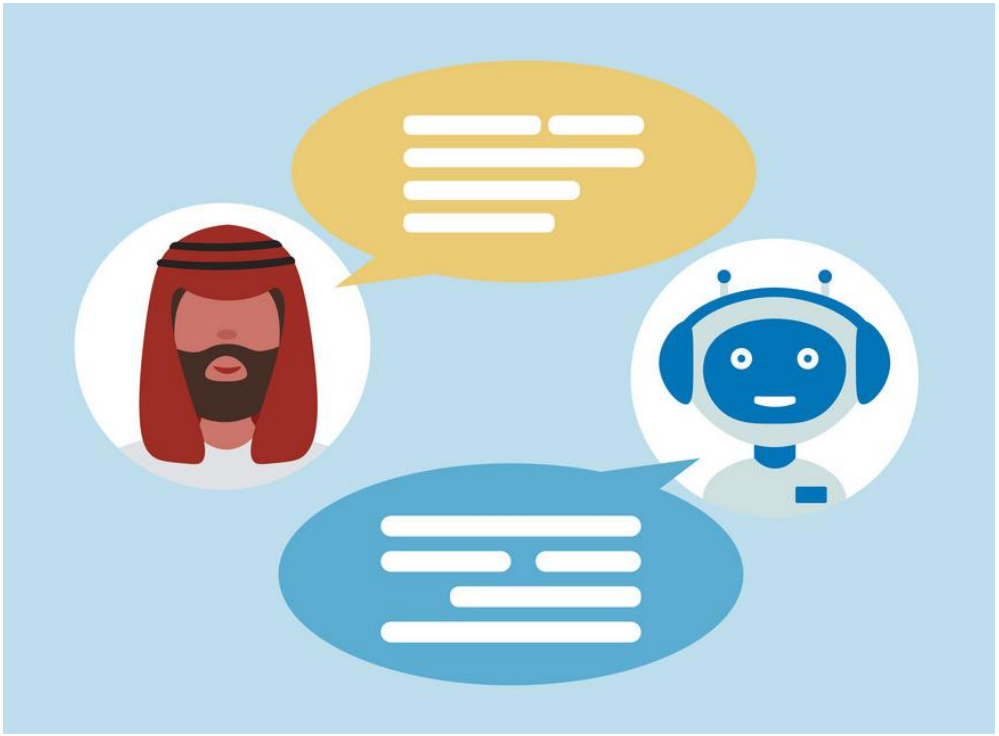
Adversarial Connective-exploiting Networks for Implicit Discourse Relation Classification, ACL 2017

Outline

- Coreference Resolution
- Coherence Analysis
- **Dialogue Systems**
  - **Introduction**
  - Goal-oriented Dialogue Systems
  - Non-goal-oriented Dialogue Systems
  - Evaluation

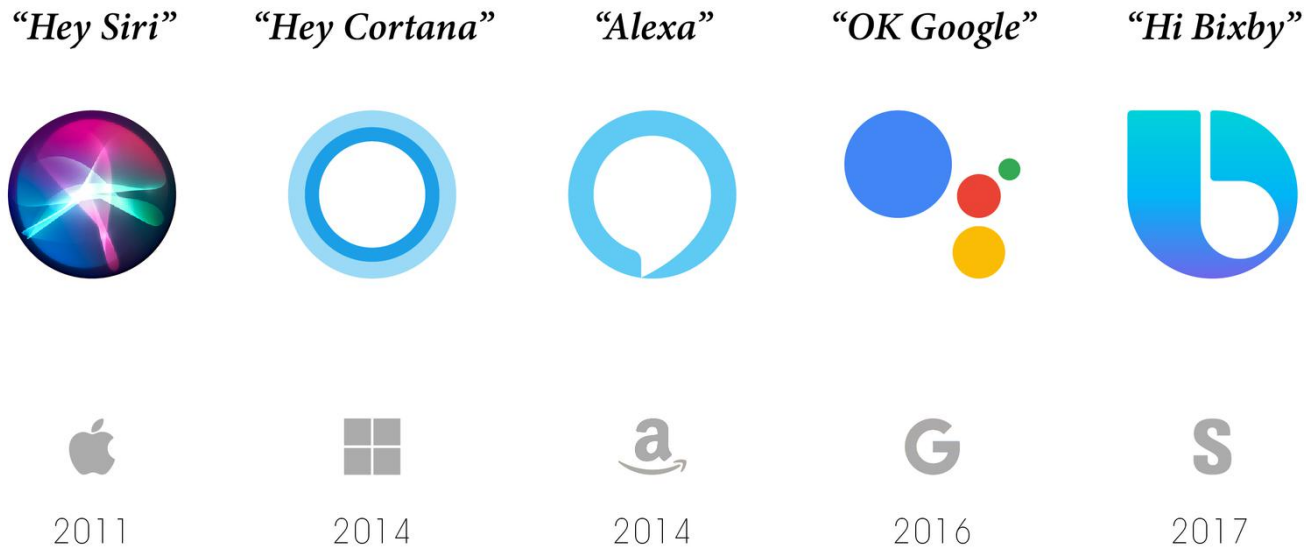
Dialogue Systems

- A **dialogue system**, or **conversational agent**, is a computer system intended to converse with a human



Dialogue Systems

- Many virtual assistants are available nowadays
  - Help with specific tasks
  - Chat for entertainment



Dialogue Systems

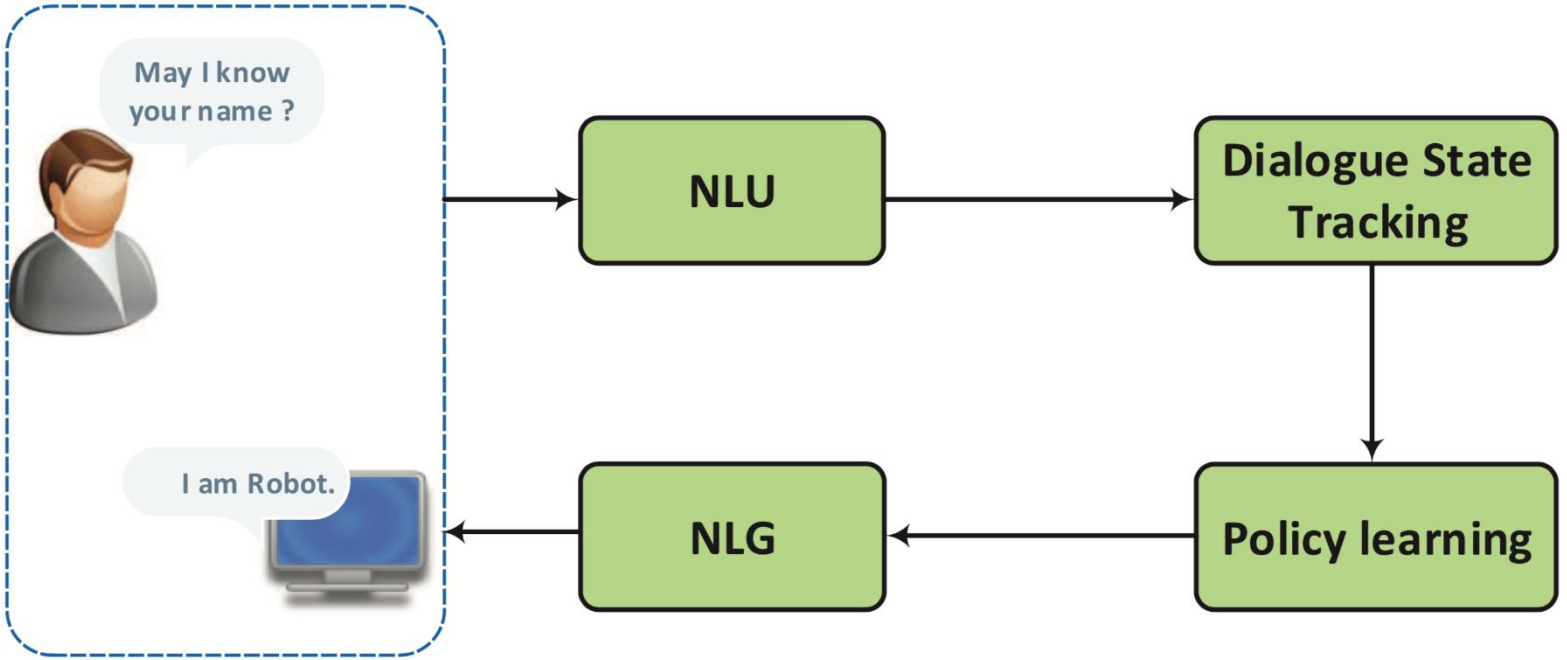
- Goal-oriented
  - Assist the user to complete certain tasks
    - Find products
    - Book hotels and restaurants
- Non-goal-oriented
  - Interact with human to provide reasonable responses
    - Entertainment
    - Companion
    - Psychological treatment

Outline

- Coreference Resolution
- Coherence Analysis
- **Dialogue Systems**
  - Introduction
  - **Goal-oriented Dialogue Systems**
  - Non-goal-oriented Dialogue Systems
  - Evaluation

Goal-oriented Dialogue Systems

- Goal-oriented dialogue system pipeline



Goal-oriented System Pipeline

- Natural Language Understanding
  - Parse the user utterance into predefined semantic slots
  - The slots are task specific and can serve downstream operations

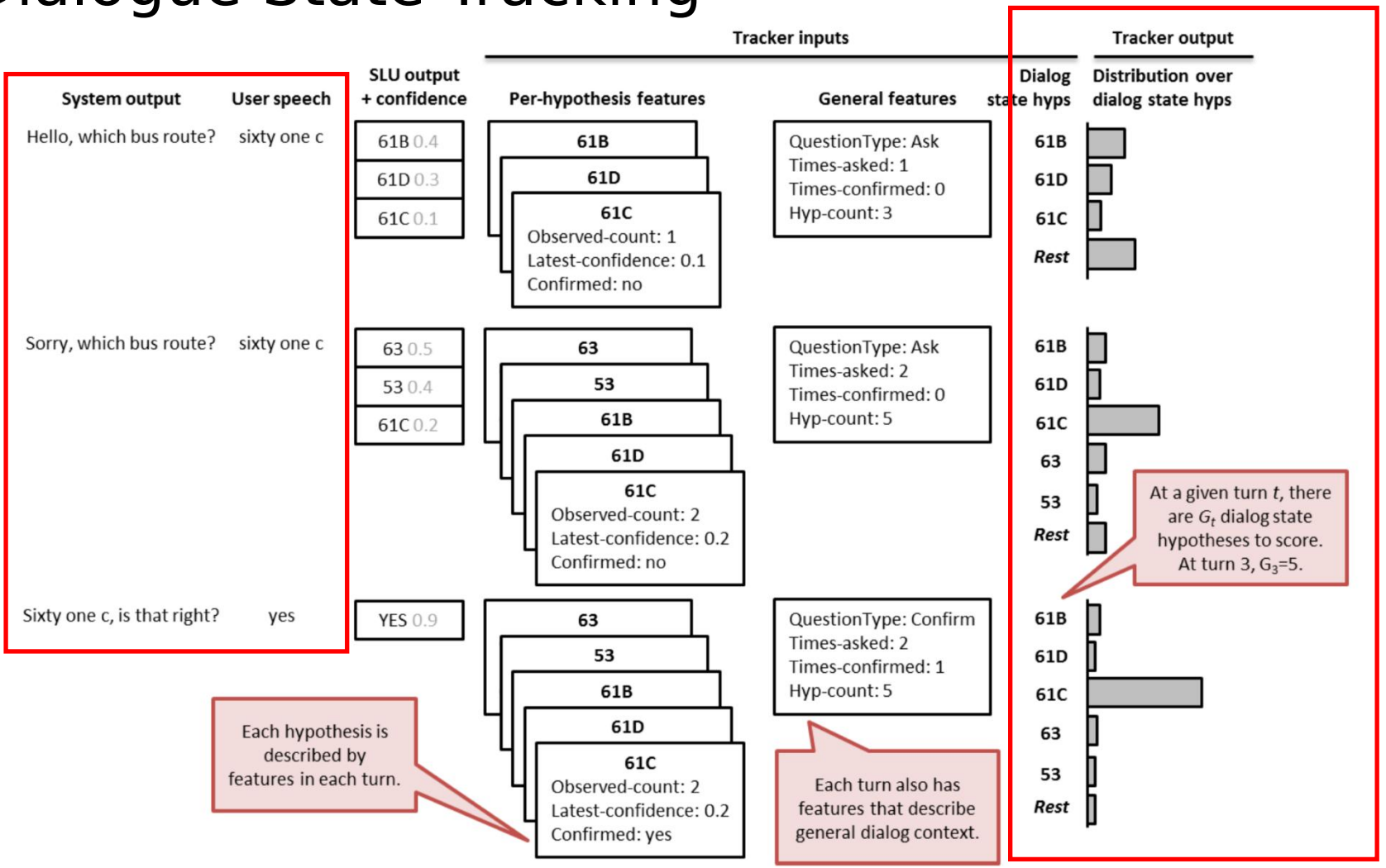
Sentence	show	restaurant	at	New	York	tomorrow
Slots	O	O	O	B-desti	I-desti	B-date
Intent	Find Restaurant					
Domain	Order					

Goal-oriented System Pipeline

- Dialogue State Tracking
  - Track the current dialogue state
  - Estimate the user's goal at every turn of the dialogue

Goal-oriented System Pipeline

- Dialogue State Tracking



Goal-oriented System Pipeline

- Policy Learning
  - Generate the next available system action
  - E.g., recommendation, comparison
  - Conditioned on dialogue state
- Natural Language Generation
  - Conditioned on dialogue action

Goal-oriented System

- Limitations of goal-oriented system pipeline
  - Handcrafting problem
    - Require a lot of domain-specific handcrafting
    - Difficult to adapt to new domains
  - Process interdependence
    - The input of a component in the pipeline is dependent on the output of another component
    - When adapting one component, all the other components need to be adapted

Goal-oriented System

- Limitations of goal-oriented system pipeline
  - Credit assignment problem
    - The end user's feedback is hard to be propagated to each upstream module
    - Determining the source of the error requires tedious error analysis in each module

Goal-oriented System

- End-to-end goal-oriented system
  - Attempt to replace the components in the pipeline with a single module
  - Adopt deep reinforcement learning for end-to-end optimization

Zhao et al. Towards End-to-End Learning for Dialog State Tracking and Management using Deep Rei
SIGDIAL 2016

Outline

- Coreference Resolution
- Coherence Analysis
- **Dialogue Systems**
  - Introduction
  - Goal-oriented Dialogue Systems
  - **Non-goal-oriented Dialogue Systems**
  - Evaluation

Non-goal-oriented System

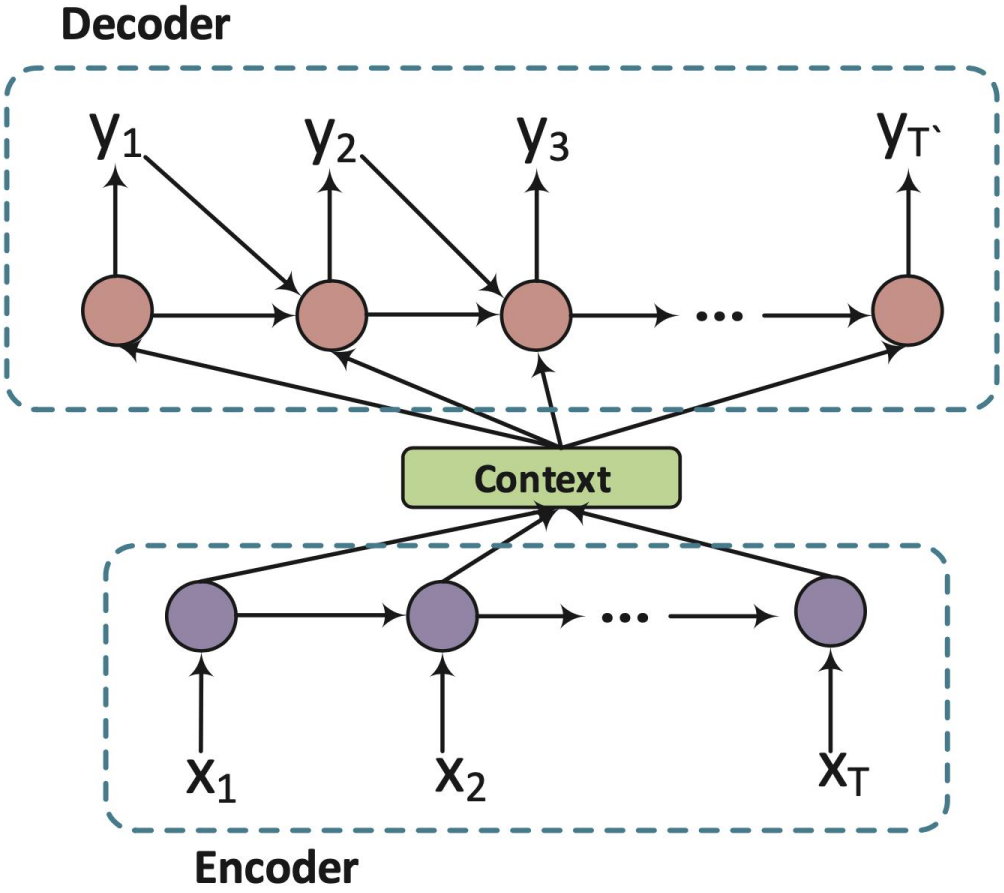
- Non-goal-oriented dialogue system
  - Converse with human in open domains
  - Important even in goal-oriented scenarios
    - Over 80% utterances are chit-chat messages in the online shopping scenario
  - Usually driven by large-scale conversation data on social media websites such as Facebook and Twitter

Non-goal-oriented System

- Generative methods
  - Adopt seq2seq framework for response generation
  - Can generate more proper responses that never appeared in the corpus
- Retrieval-based methods
  - Choose a response from candidate responses
  - The responses are usually more informative and fluent

Generative Dialogue Systems

- Adopt seq2seq framework for response generation



Generative Dialogue Systems

- Challenges in dialogue generation
  - Dialogue Context
  - Response Diversity
  - Topic and Personality
  - External Knowledge Base
  - Interactive Dialogue Learning

Generative Dialogue Systems

- Dialogue Context
  - Important in dialogue generation
  - The phrase “good luck” is plainly motivated by the reference to “your game” in the first utterance

Serban et al. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. AAAI 2016

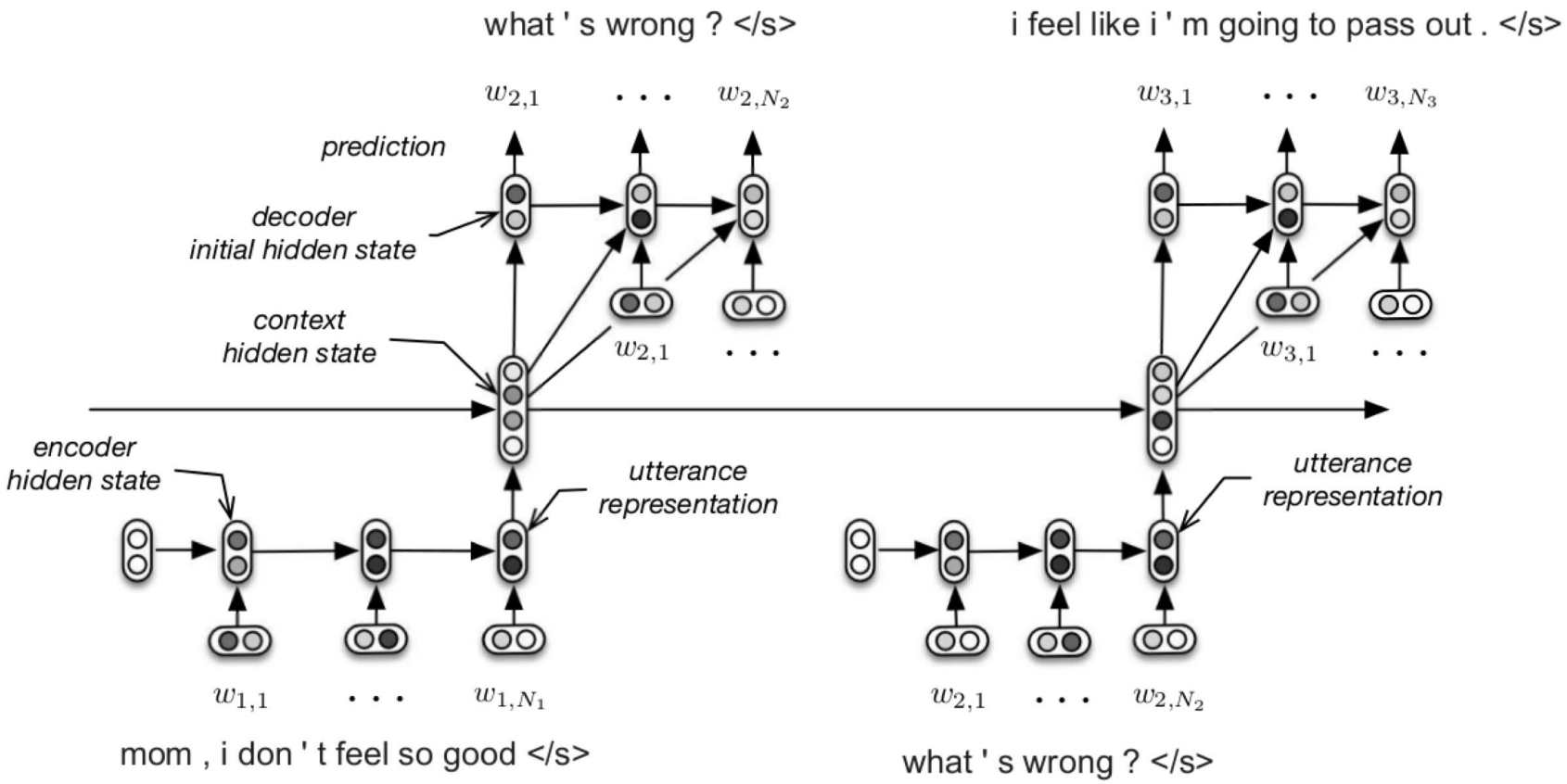
Generative Dialogue Systems

- Hierarchical Recurrent Encoder-Decoder
  - The history of past submitted queries is considered as a sequence at two levels
    - A sequence of words for each query
    - A sequence of queries
  - Model this hierarchy of sequences with two RNNs
    - Local word-level RNN
    - Global query-level RNN

Serban et al. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. AAAI 2016

Generative Dialogue Systems

- Hierarchical Recurrent Encoder-Decoder



Serban et al. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. AAAI 2016

Generative Dialogue Systems

- External Knowledge Base
  - Generating reasonable responses usually requires external knowledge
    - Lexical knowledge
    - World knowledge
    - Commonsense knowledge

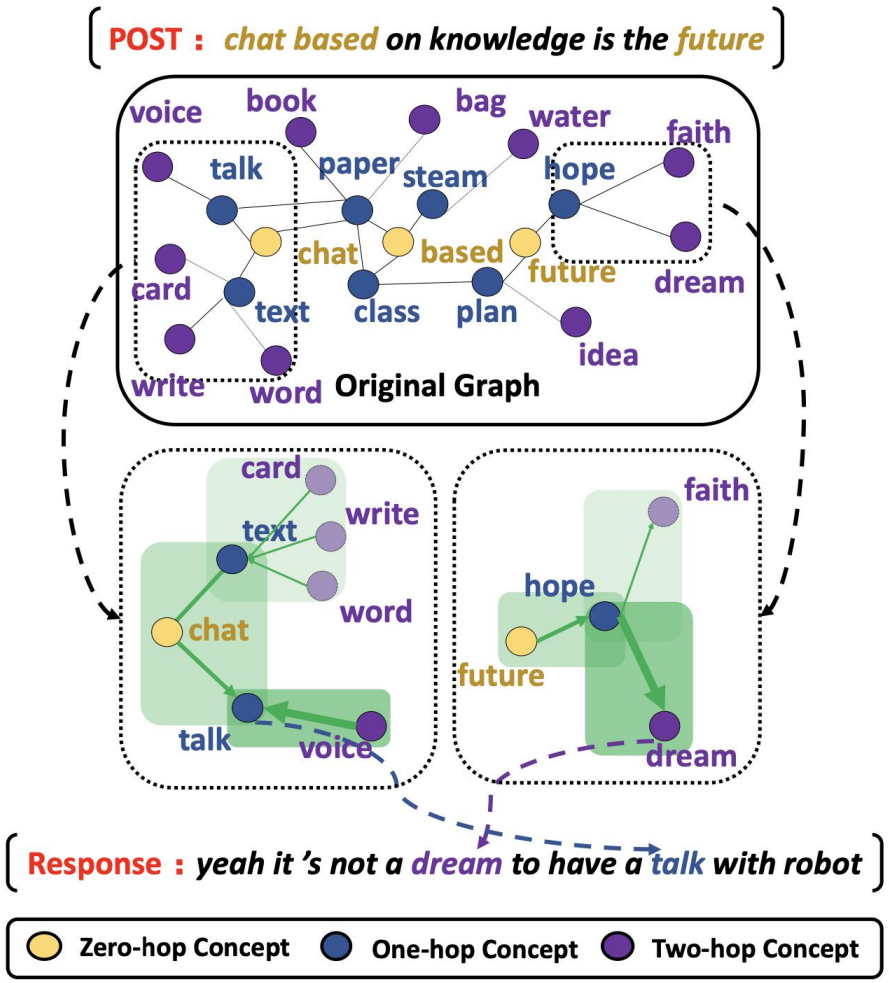
Generative Dialogue Systems

- Dialogue generation with commonsense knowledge
  - Mimic human conversations that evolve around related concepts and hop to distant concepts
  - Framework
    - Ground conversations to the concept space in ConceptNet
    - Traverse commonsense knowledge graphs to explicitly model conversation flows
    - Guided by graph attentions in the concept graph

Zhang et al. Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs. ACL 2020

Generative Dialogue Systems

- Dialogue generation with commonsense knowledg

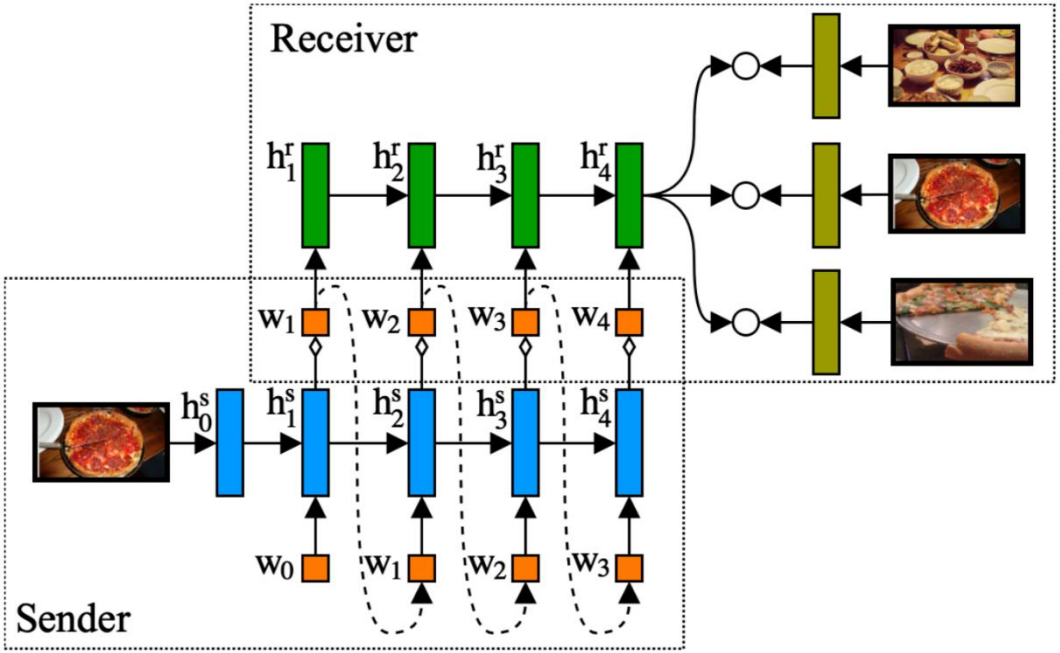


Generative Dialogue Systems

- Interactive Dialogue learning
  - Learn to communicate through interaction
  - Rather than relying on explicit supervision
  - A difficult but important problem
  - Most existing works ask agents to
    - Cooperate to accomplish certain tasks
    - Communicate in the form of a language (i.e., sequences of discrete symbols)

Generative Dialogue Systems

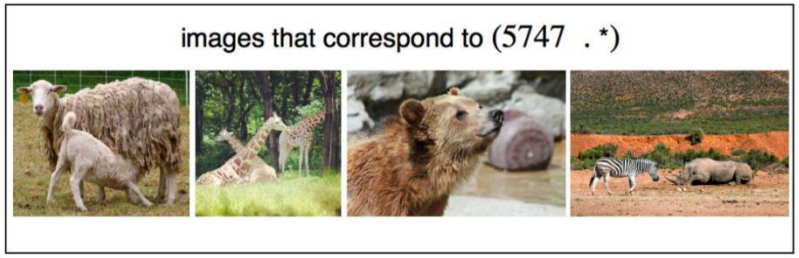
- Referential Game
  - The sender agent describes a target image by sequences of discrete symbols
  - The receiver agent tries to figure out which image to se

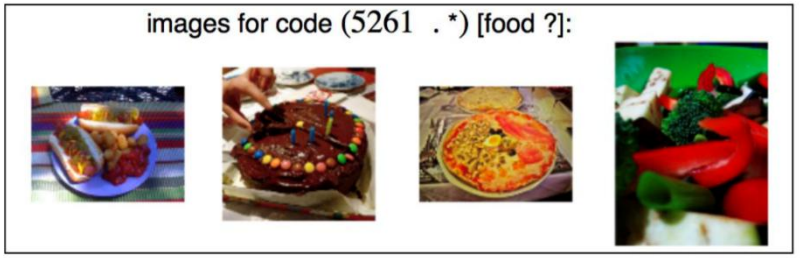


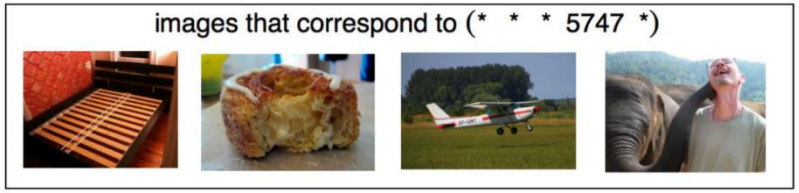
Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols. NIPS 2017

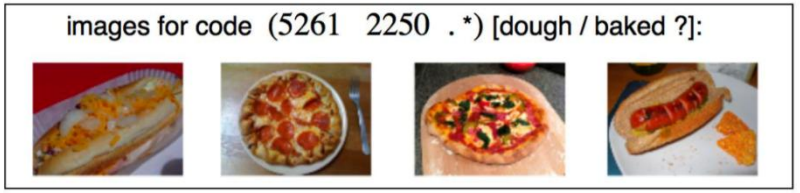
Generative Dialogue Systems

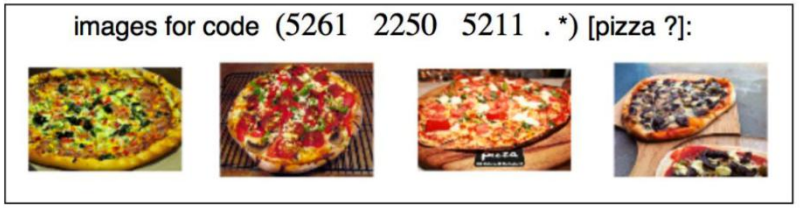
- The language learned by the game exhibits a degree of compositionality and variability
  - (5747 * * * *) corresponds to animals
  - (5747 5747 7125 * *) corresponds to a particular type of bears











Retrieval-based Systems

- Retrieval-based dialogue generation
  - Choose a response from candidate responses
  - The key is post-response matching, i.e., calculating the similarity between post and response
  - Matching algorithms have to overcome semantic gaps between posts and responses

P:	*It is my last day in Paris. So hard to say goodbye.*
R1:	*Enjoy your time in Paris.*
R2:	*Man, I wish I am in London right now.*

Outline

- Coreference Resolution
- Coherence Analysis
- **Dialogue Systems**
  - Introduction
  - Goal-oriented Dialogue Systems
  - Non-goal-oriented Dialogue Systems
  - **Evaluation**

Evaluation of Dialogue Systems

- Evaluation of goal-oriented dialogue systems
  - Can be evaluated based on human-generated supervised signals
    - Task completion tests
    - User satisfaction score

Evaluation of Dialogue Systems

- Evaluation of non-goal-oriented dialogue systems
  - Notoriously hard to evaluate automatically and remain an open question
  - Automatic evaluation metrics in machine translation (e.g., BLEU) are adopted
  - However, researchers have found that that those metrics have either weak or no correlation with human judgements

Evaluation of Dialogue Systems

- Evaluation of non-goal-oriented dialogue systems
  - In the following example, the reasonable model response would receive a BLEU score of 0
  - Due to the intrinsic diversity of valid responses in a dialogue

Context of Conversation
Speaker A: Hey John, what do you want to do tonight?
Speaker B: Why don't we go see a movie?
Ground-Truth Response
Nah, I hate that stuff, let's do something active.
Model Response
Oh sure! Heard the film about Turing is out!

Evaluation of Dialogue Systems

- Evaluation of non-goal-oriented dialogue systems
  - Most non-goal-oriented are evaluated by both
    - Automatic evaluation metrics
    - **Human evaluation**
  - Human evaluation is both time-consuming and labor-intensive
  - Showing the performance of dialogue systems by cherry-picking cases can also be problematic

Summary

- Coherent and Discourse Relation Classification
- Explicit Markers, Cross Product of Words, Feature Engineering
- Deep Learning Based Methods
- Dialog Systems: goal-oriented & non-goal-oriented
- Automatic evaluation and human evaluation

Reading Material

a. Reference in Language & Coreference Resolution
Unsupervised Models for Coreference Resolution. EMNLP 2008. [link]
End-to-end Neural Coreference Resolution. EMNLP 2017. [link]
Coreference Resolution as Query-based Span Prediction. 2019. [link]
b. Coherence & Discourse Relation Classification
Implicit Discourse Relation Classification via Multi-Task Neural Networks. AAAI 2016 [link]
Implicit Discourse Relation Detection via a Deep Architecture with Gated Relevance Network. ACL 2016 [link]
Employing the Correspondence of Relations and Connectives to Identify Implicit Discourse Relations via Label Embeddings. ACL 2019 [link]
Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution. AAAI 2018 [link]

For reading material recommendation of this course, please refer to our github

Reading Material

c. Context Modeling and Conversation

A Survey on Dialogue Systems: Recent Advances and New Frontiers. Hongshen Chen, Xiaorui Liu, Dawei Yin, Jiliang Tang. 2018 [link]

A Diversity-Promoting Objective Function for Neural Conversation Models. Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, Bill Dolan. NAACL 2016 [link]

A Persona-Based Neural Conversation Model. Jiwei Li, Michel Galley, Chris Brockett, Georgios PSpithourakis, Jianfeng Gao, Bill Dolan. ACL 2016 [link]

For reading material recommendation of this course, please refer to our github


TSINGHUA UNIVERSITY
-1911-


Tsinghua University
NLP
Natural Language Processing

THUNLP