Future of NLP

Zhiyuan Liu

liuzy@tsinghua.edu.cn

THUNLP

Outline

- Graph Neural Networks in NLP Applications
- Compress Pre-trained Language Models
- Continual Learning
- Grammatical Error Correction
- Automatic Chinese Poetry Generation
- Textual Adversarial Attack and Defense
- Cross-Modal Learning

Graph Neural Networks in NLP Applications

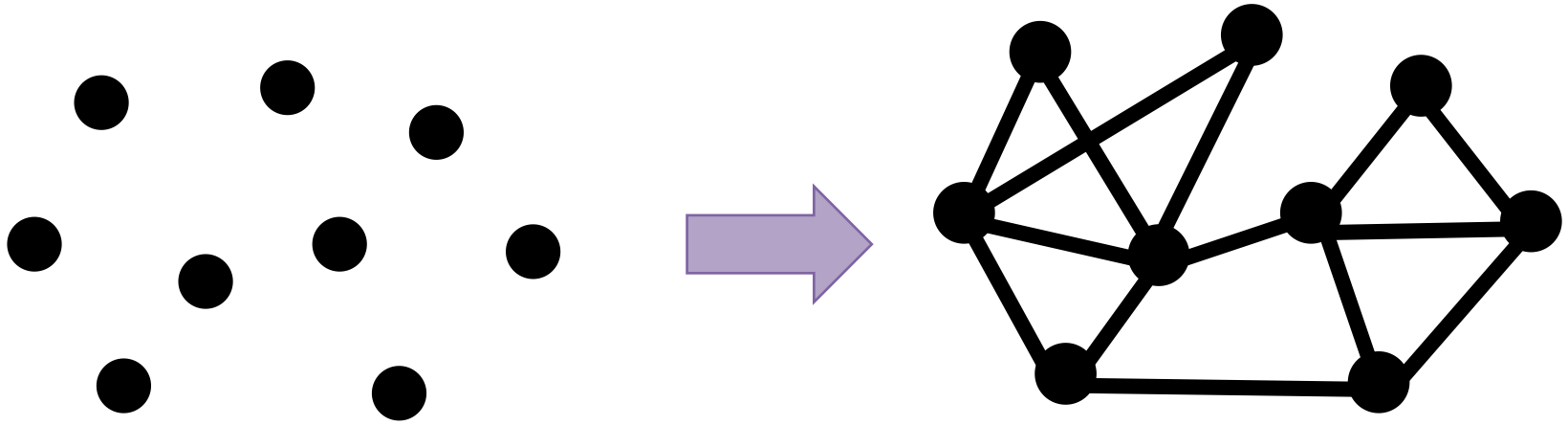
THUNLP

Outline

- Graph Neural Networks
- GNN in NLP Applications

Graph

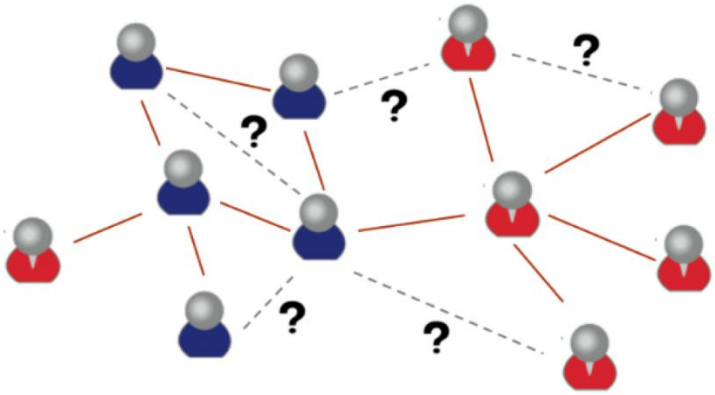
- Graph models a set of objects (nodes) and their relationships (edges).



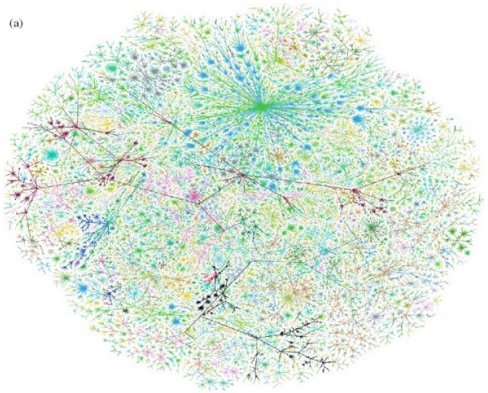
Independent Objects

Graph (Network)

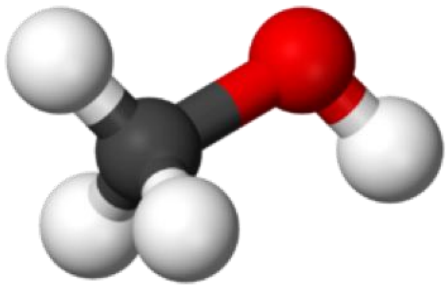
Graphs are Ubiquitous

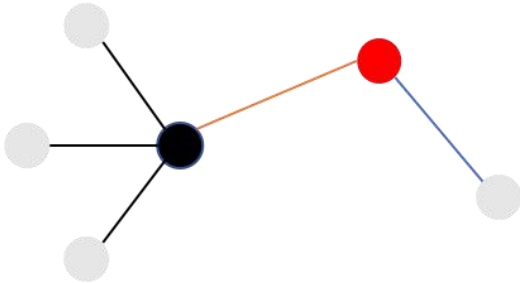


Social network



Internet

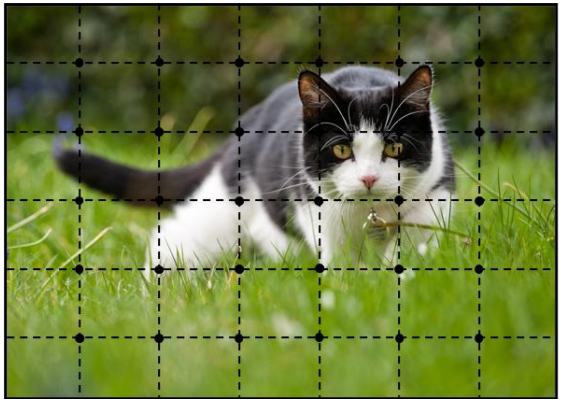




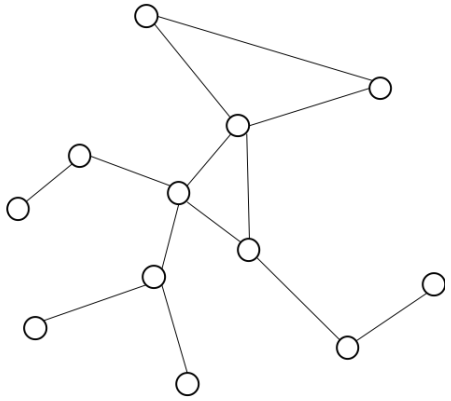
Molecular graph

Graphs are Different

- Unlike texts or images, graphs are typical non-Euclidean data.
- Traditional neural networks (CNNs, RNNs) cannot directly work on graphs.
- Graph neural networks (GNNs) are deep learning tools designed for graph data.



2-D Euclidean data



Non-Euclidean data

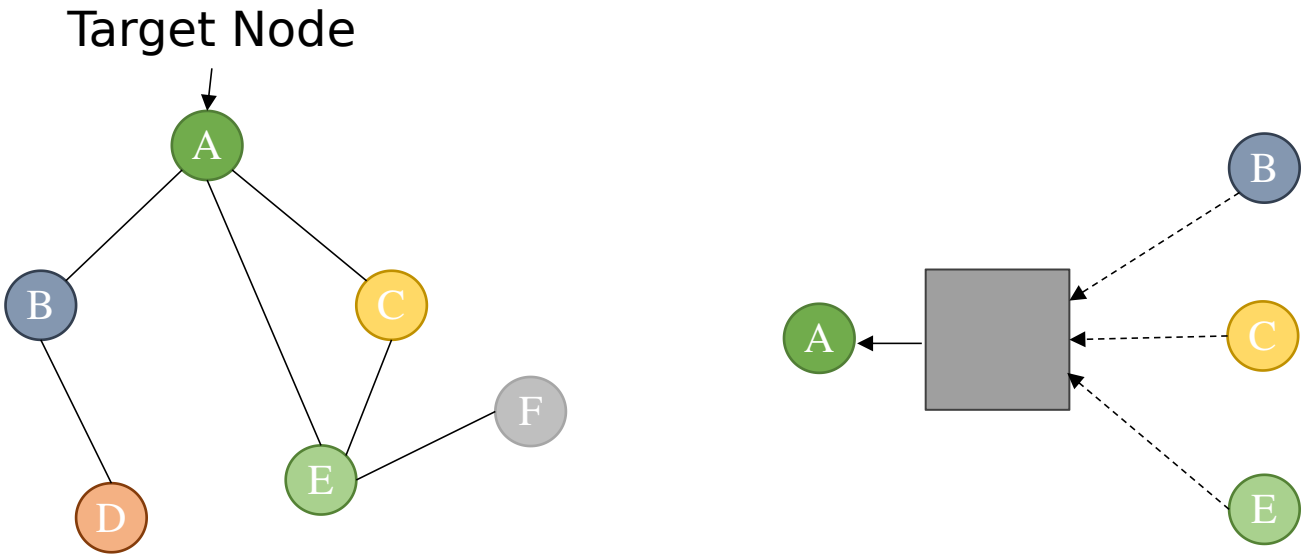
Graph Neural Networks

- Notations
  - $G = (V, E)$
  - Node Embeddings: $h_i \in R^{F_h}$
  - Edge Embeddings: $e_{ij} \in R^{F_e}$
  - The input features can be task-specific, for example:
    - User profiles in a social network
    - Paper titles in a citation network
    - ...

Graph Neural Networks

- Key ideas
  - In each layer, nodes aggregate information from neighbors.

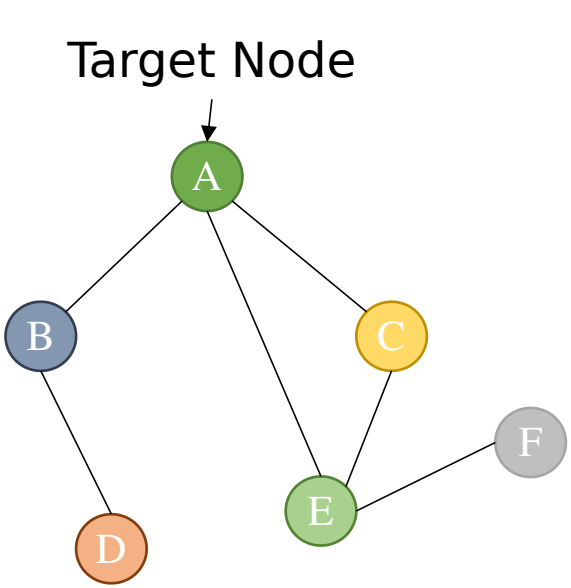
$$h_v = f\left(\text{Agg}(\{h_u,\ \forall u\ \epsilon\ N(v)\})\ ,\ h_v\right)$$

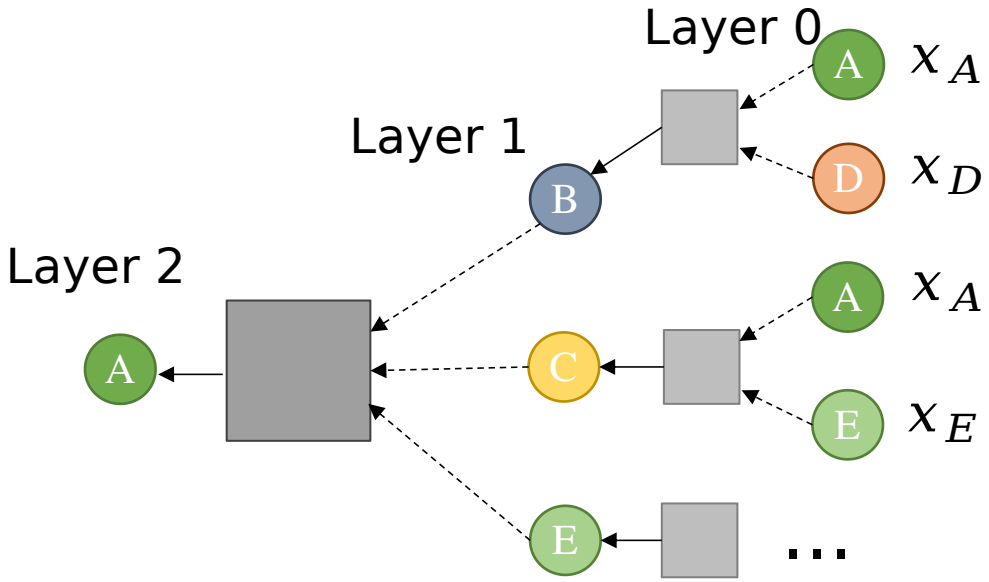


Graph Neural Networks

- Key ideas
  - In each layer, nodes aggregate information from neighbors.
  - The models can have arbitrary layers to collect information from multi-hop neighbors.

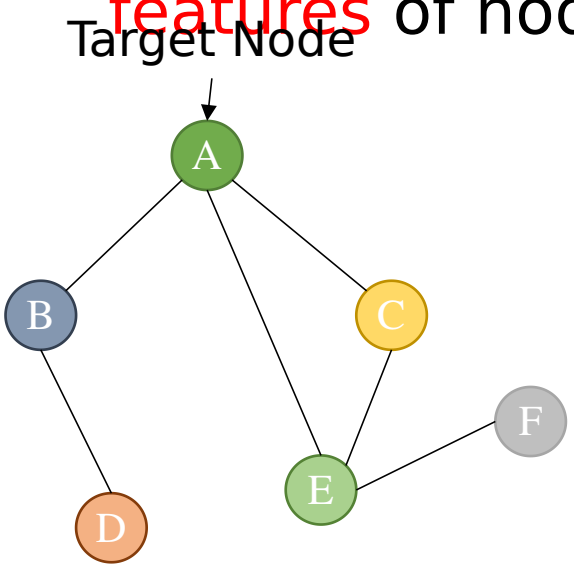
$$h_v^k = f\left(\mathrm{Agg}\left(\{h_u^{k-1}, \forall u \in N(v)\}\right), h_v^{k-1}\right)$$

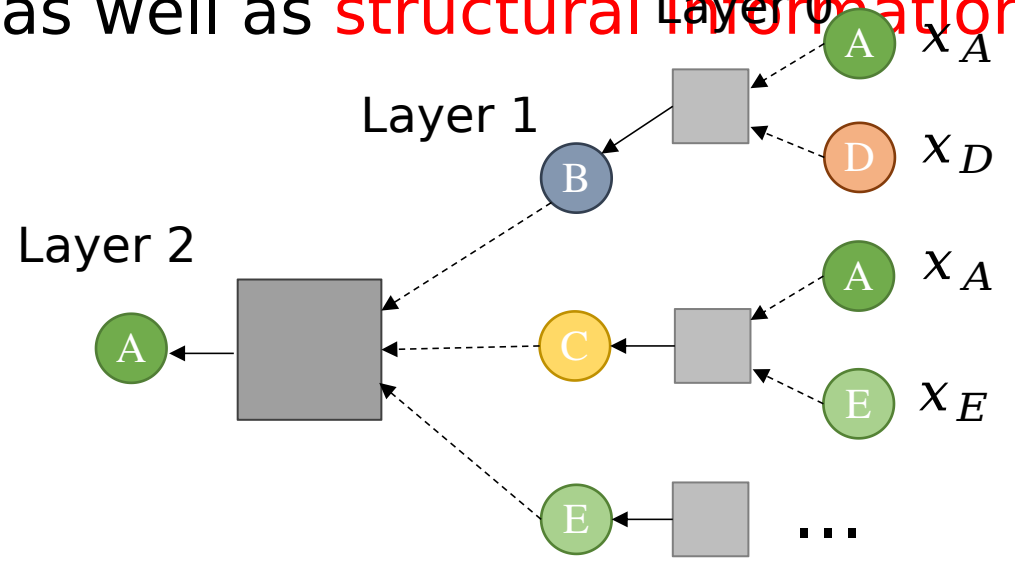




Graph Neural Networks

- Key ideas
  - In each layer, nodes aggregate information from neighbors.
  - The models can have arbitrary layers to collect information from multi-hop neighbors.
  - To learn better representations based on own features of nodes as well as structural information.





Graph Neural Networks

- GNN Variants
  - GCN

$$h_v^k = \sigma\left(W_k \sum_{u \epsilon N(v) \bigcup v} \frac{h_u^{k-1}}{\sqrt{|N(u)|\,|N(v)|}}\right)$$

  - GAT

$$\alpha_{vu} = \frac{\exp(\text{LeakyReLU}(a^T[Wh_v^{k-1} \| Wh_u^{k-1}]))}{\sum_{i \epsilon N(v)} \exp(\text{LeakyReLU}(a^T[Wh_v^{k-1} \| Wh_i^{k-1}]))}$$

$$h_v^k = \sigma\left(\sum_{u \epsilon N(v) \bigcup v} \alpha_{vu} W_k h_u^{k-1}\right)$$

  - GraphSAGE

$$h_v^k = \sigma([W_k \bullet \text{AGG}(\{h_u^{k-1},\ \forall u\ \epsilon\ N(v)\}) \,\|\, B_k h_v^{k-1}])$$

Semi-Supervised Classification with Graph Convolutional Networks. ICLR 2017.
Graph Attention Networks. ICLR 2018.
Inductive Representation Learning on Large Graphs. NIPS 2017.

GNN in NLP Application

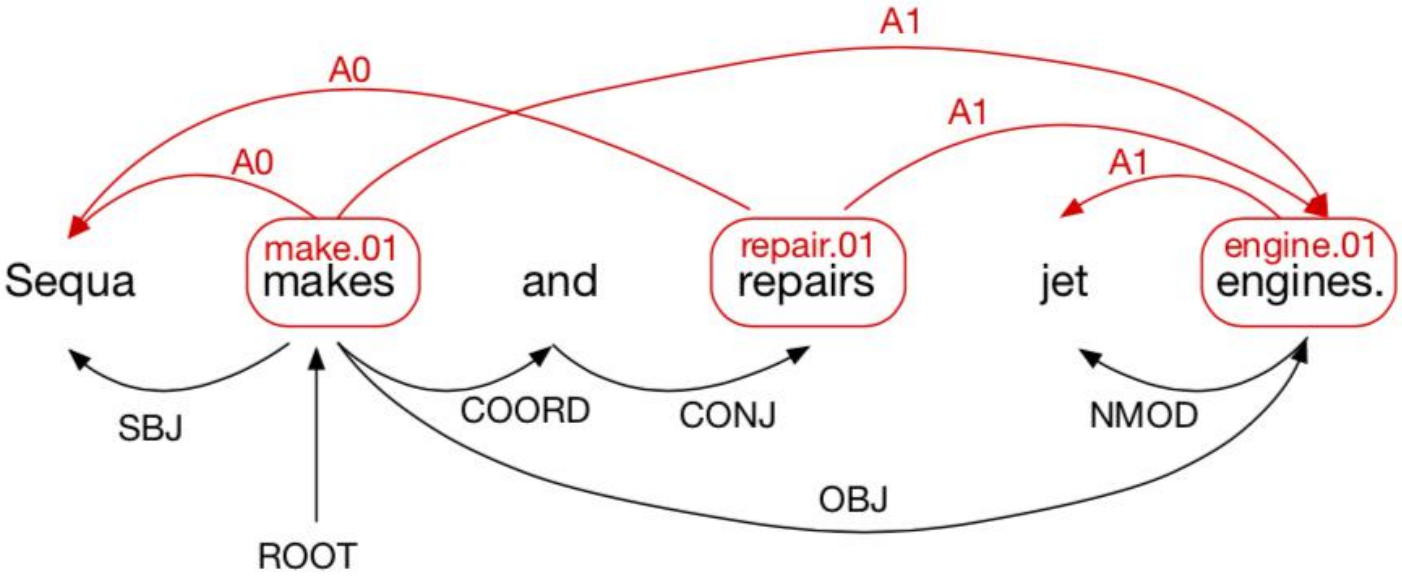
- NLP tasks usually deal with text...
- However, there is no explicit graph structure in text.
- Two trends of using GNNs in NLP applications:
  - Incorporating graphs into text
  - Building graphs from text

Incorporating Graphs into Text

- This kind of methods tries to incorporate additional information with graph structure into text.
- Syntactic GCN for Machine Translation
- Contextualized GCN for Relation Extraction

Syntactic GCN

- Incorporating syntactic or semantic information into machine translation.

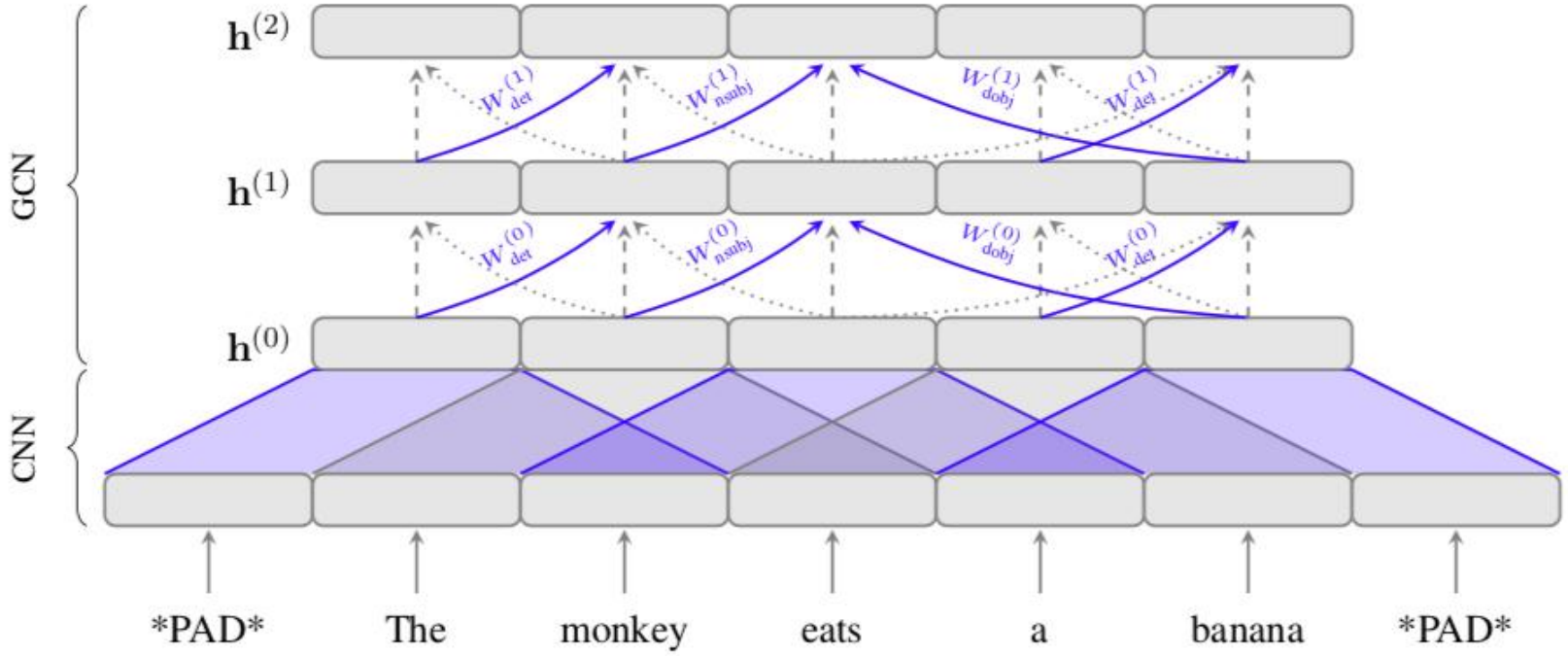


Graph convolutional encoders for syntax-aware neural machine translation. EMNLP 2017

Exploiting semantics in neural machine translation with graph convolutional networks. ACL 2018

Syntactic GCN

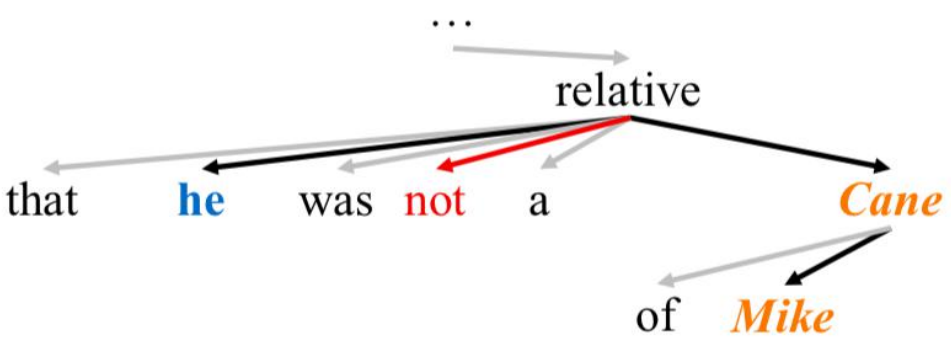
- Incorporating syntactic or semantic information into the encoder.
- The additional information helps the understanding of the source sentence.



Contextualized GCN

- An example of Relation Extraction based on the dependency path.

I had an e-mail exchange with Benjamin Cane of Popular Mechanics which showed that **he** was not a relative of ***Mike Cane***.



Prediction from dependency path: *per:other_family*
Gold label: *no_relation*

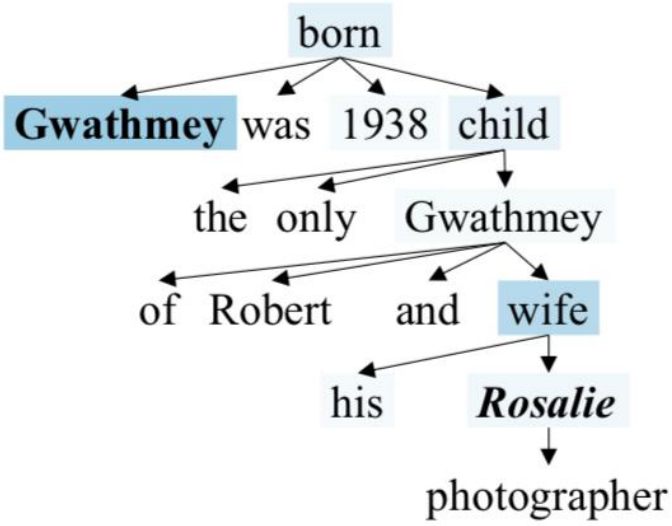
Graph Convolution over Pruned Dependency Trees Improves Relation Extraction.

Contextualized GCN

- Hypothesis: The expected relation exists on the subtree rooted in the least common ancestor (LCA) of the two entities.

Relation: *per:parents*

Gwathmey was born in 1938, the only child of painter Robert Gwathmey and his wife, ***Rosalie***, a photographer.

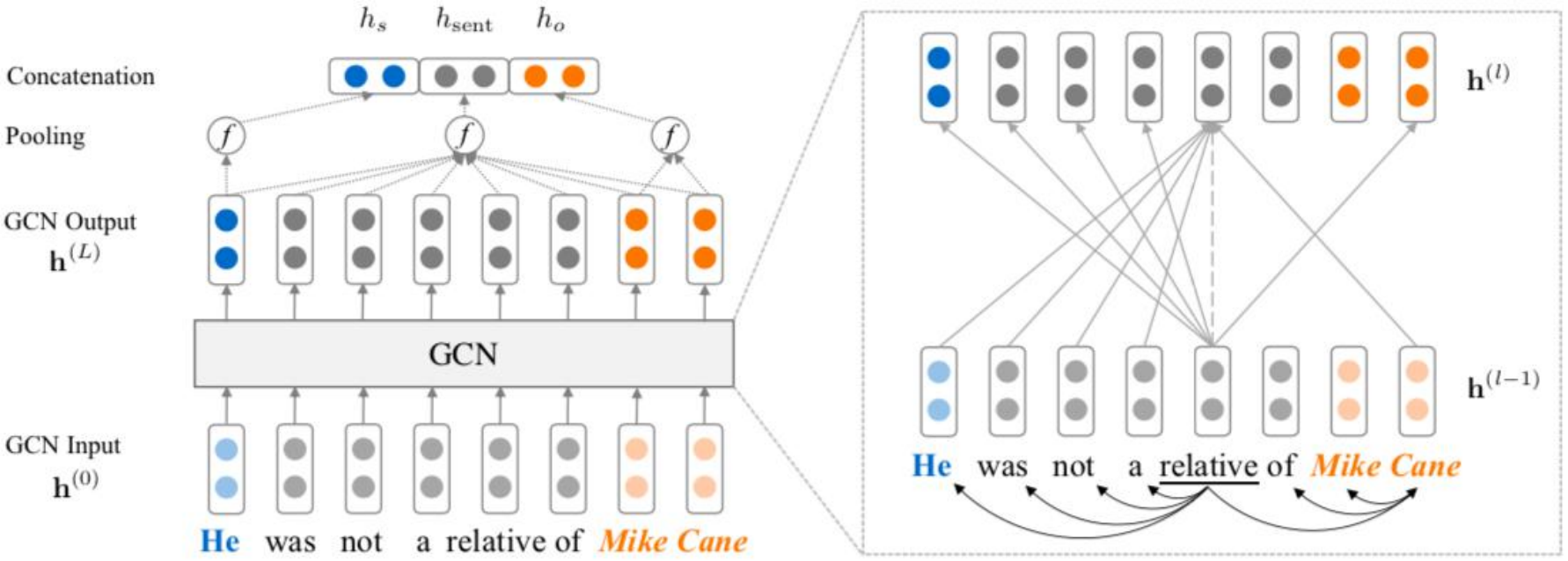


Subtree pruning strategy:

Only preserve the nodes that are k steps away from the dependency path. K=1 in the example.

Contextualized GCN

- Model

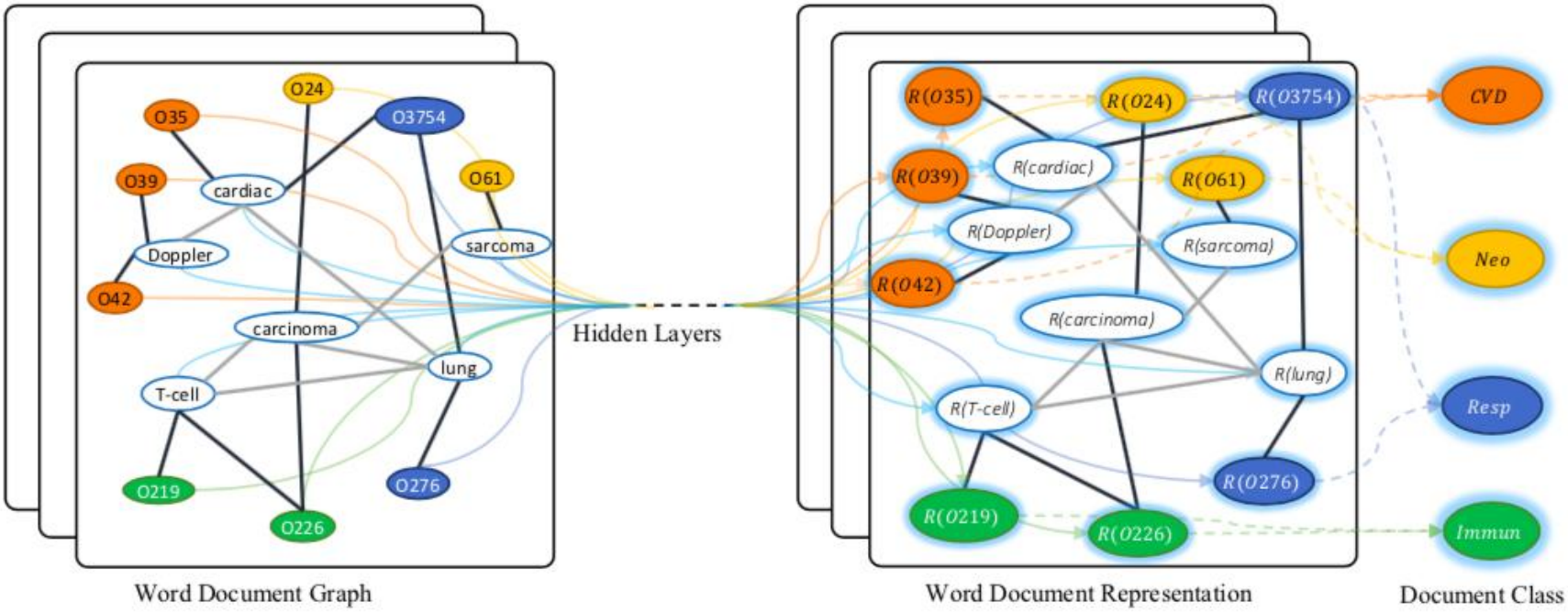


Building Graphs from Text

- This kind of methods builds graphs directly from the corpus and then applies GNNs on them to learn representations.
- Text GCN for Text Classification
- Sentence LSTM for Text Representation

Text GCN

- Text GCN
  - Doc nodes and word nodes.
  - One-hot input features for each node.



Graph Convolutional Networks for Text Classification. AAAI 2019.

Text GCN

- Text GCN
  - The adjacency matrix

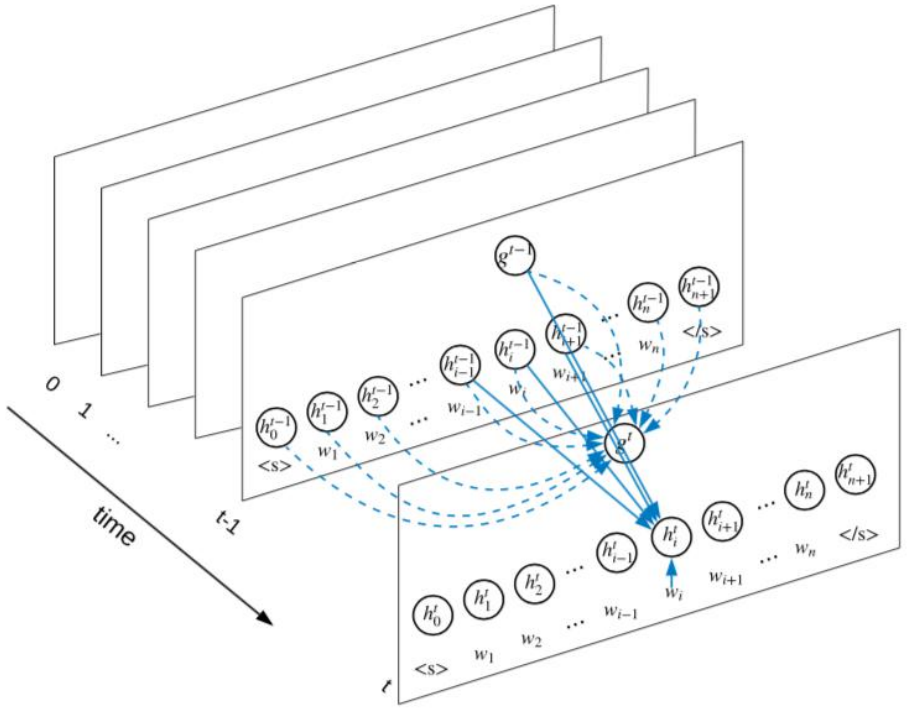
$$A_{ij} = \begin{cases} PMI(i,\ j) & \text{i and j are word nodes and } PMI(i,\ j) > 0 \\ TFIDF(i,\ j) & \text{i is doc node and j is word node} \\ 1 & i = j \\ 0 & \text{others} \end{cases}$$

$$PMI(i,j) = log\frac{p(i,j)}{p(i)p(j)} p(i,j) = \frac{\#W(i,j)}{\#W} \quad p(i) = \frac{\#W(i)}{\#W}$$

  - Then it applies a two-layer GCN to do node classification on doc nodes.

Sentence LSTM

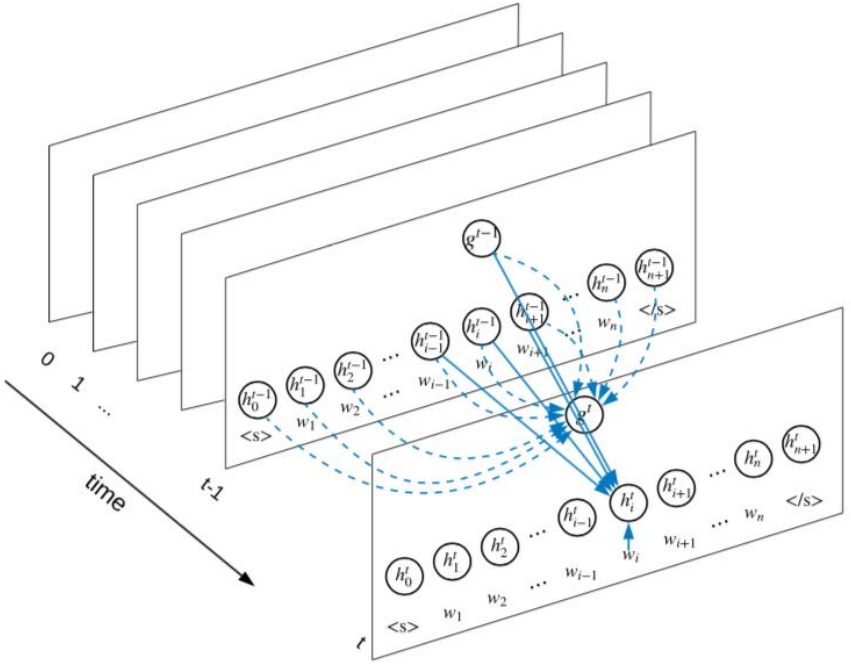
- Sentence LSTM
  - Word node. Connects with its neighbors in a window and the global node from last step.
  - Global node. Connects with all words from last time st



Sentence-State LSTM for Text Representation. ACL 2018.

Sentence LSTM

- Sentence LSTM
  - Word node.
    - Capture contextual information.
    - Solve word-level tasks, such as NER and POS.
  - Global node.
    - Provide long-range dependency information.
    - Solve sentence-level tasks, such as text classification.



Summary

- GNNs are widely applied to NLP tasks.
  - Text Classification, Relation Extraction, Machine Translation…
- GNNs are useful in NLP tasks.
  - Incorporating additional information.
  - Discovering relational information from text.
- How to use GNNs in NLP tasks is also a open problem…
  - Task-dependent -> General methods?

Compress Pre-trained Language Models

THUNLP

What are PLMs

- PLMs are the language models having powerful transferability for other NLP tasks
  - word2vec, GPT, BERT, ...

- BERT-BASE: 110M
- BERT-LARG: 330M
- T5-LARGE: 770M
- T5-11B: 11,000M

Too large!

Compression Methods

- Pruning
  - Remove unnecessary parts of the network
- Knowledge Distillation
  - Trains a much smaller Transformer from scratch on the pre-training / downstream-data
- Weight Sharing
  - Some weights in the model share the same value as other parameters in the model

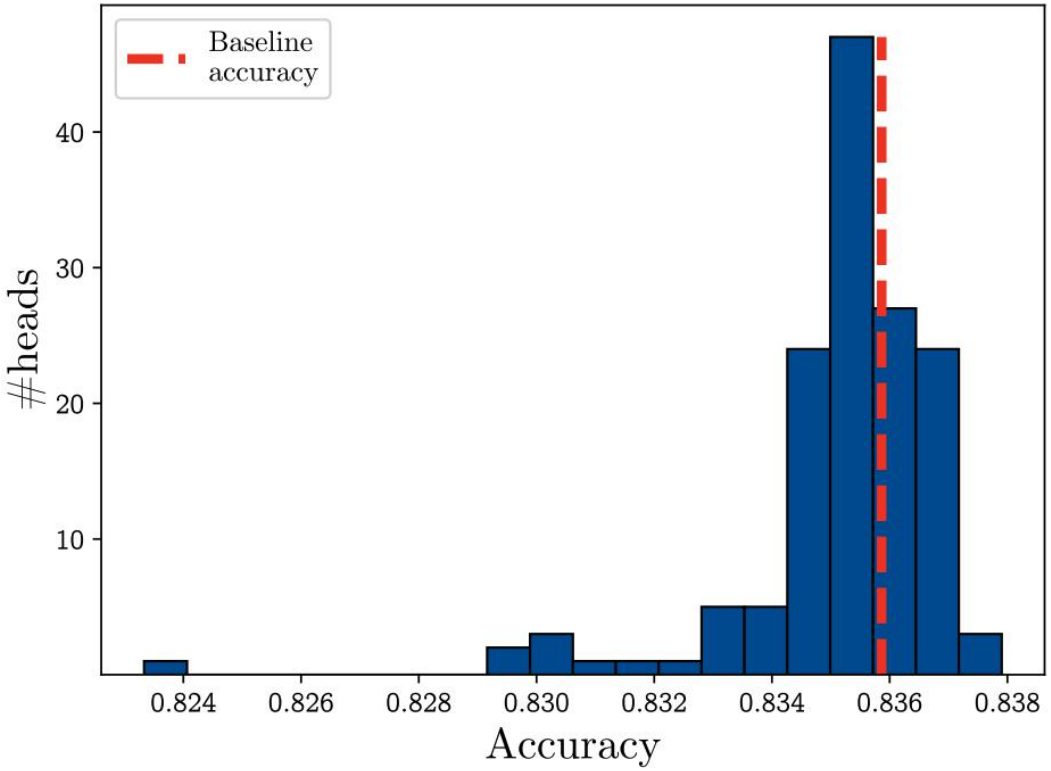
Pruning

- Recall: Attention is an important mechanism in PLMs
- Are all attention heads useful in the task?

Are Sixteen Heads Really Better than One? NeurIPS 2019.

Ablating One Head

- To understand the contribution of a particular attention head, they evaluate the model's performance while masking that head



Ablating All Heads but One

Layer		Layer	
1	-0.01%	7	0.05%
2	0.10%	8	-0.72%
3	-0.14%	9	-0.96%
4	-0.53%	10	0.07%
5	-0.29%	11	-0.19%
6	-0.52%	12	-0.12%

Table 3: Best delta accuracy by layer when only one head is kept in the BERT model. None of these results are statistically significant with $p < 0.01$.

Head Importance Score

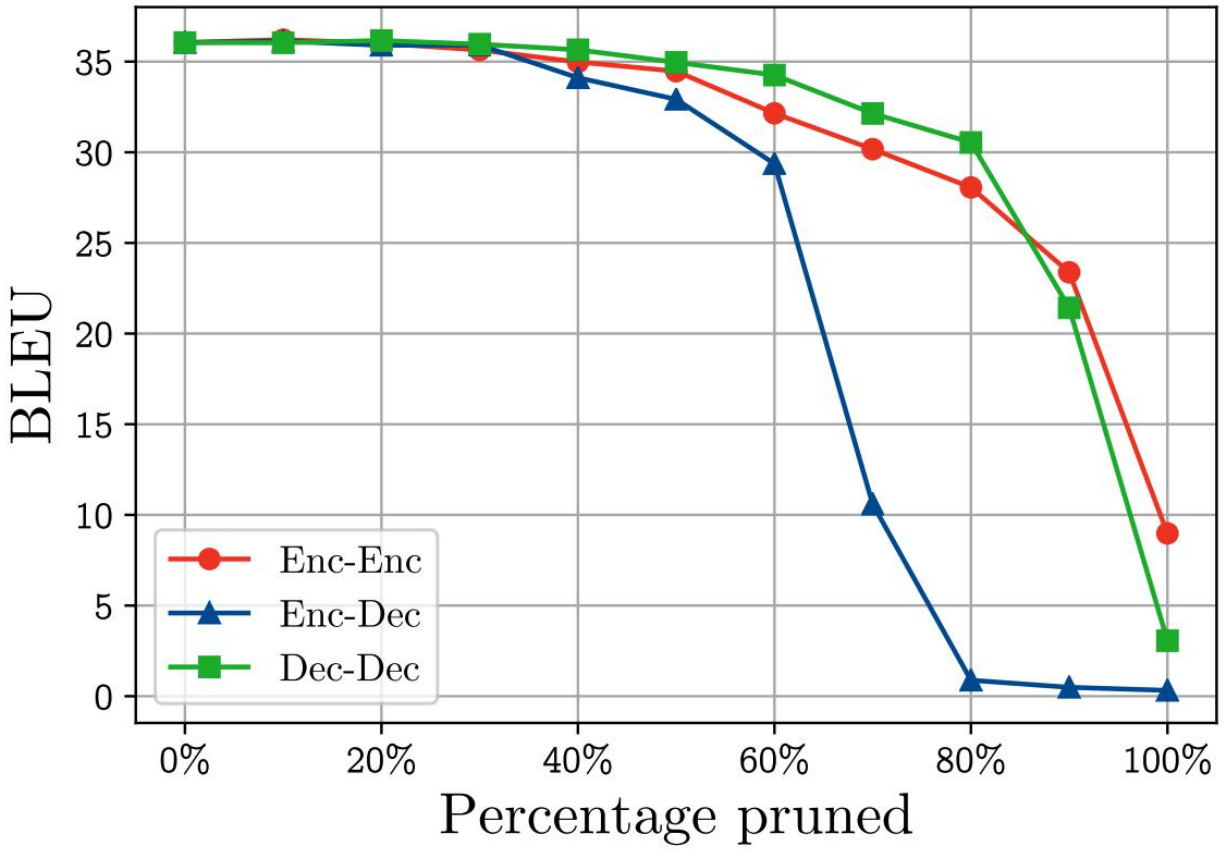
- Look at the expected sensitivity of the model to the mask

$$I_h = \mathbb{E}_{x \sim X} \left| \mathrm{Att}_h(x)^T \frac{\partial \mathcal{L}(x)}{\partial \mathrm{Att}_h(x)} \right|$$

- High values mean these heads have a large effect on the model

Experimental Result

- Machine translation

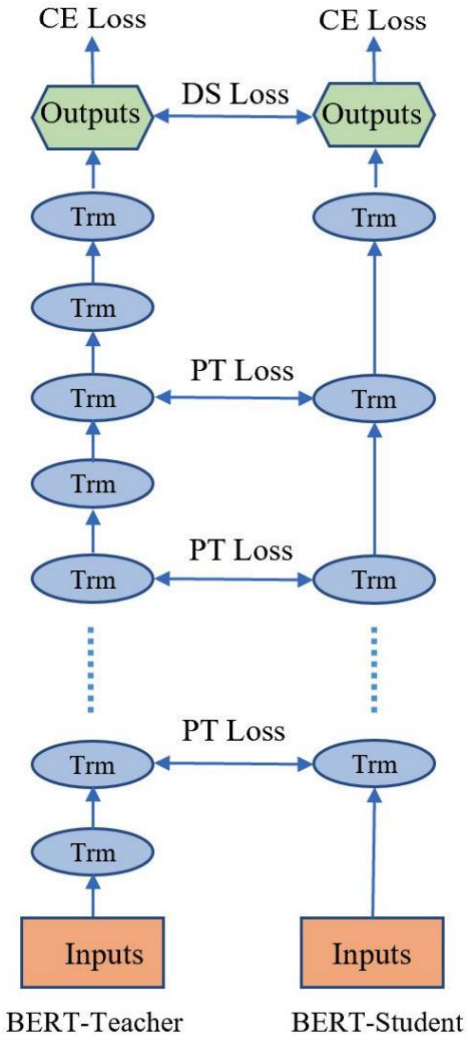


Knowledge Distillation

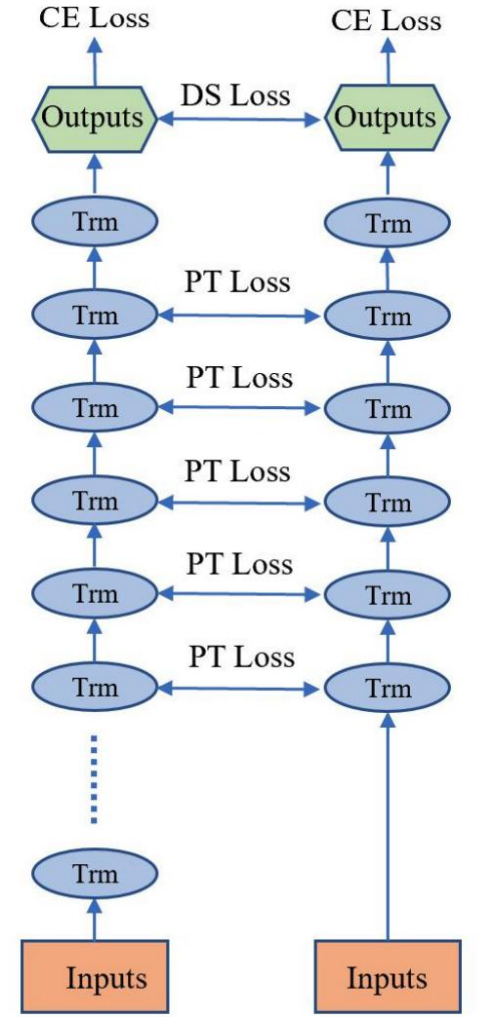
- Conventional knowledge distillation
  - Use the output from the last layer of the teacher network for distillation
- Patient Knowledge Distillation
  - patiently learn from multiple intermediate layers of the teacher model for incremental knowledge extraction

Patient Knowledge Distillation for BERT Model Compression.

Patient Knowledge Distillation



(a) PKD-*Skip*



(b) PKD-*Last*

Patient Knowledge Distillation

- Compress downstream models
- Three steps

- Fine-tune a full model as the teacher
- Patient knowledge distillation from teacher
- Efficient inference

Patient Knowledge Distillation

- Slight degradation

Model	SST-2 (67k)	MRPC (3.7k)	QQP (364k)	MNLI-m (393k)	MNLI-mm (393k)	QNLI (105k)	RTE (2.5k)
$BERT_{12}$ (Google)	93.5	88.9/84.8	71.2/89.2	84.6	83.4	90.5	66.4
$BERT_{12}$ (Teacher)	94.3	89.2/85.2	70.9/89.0	83.7	82.8	90.4	69.1
$BERT_6$-FT	90.7	85.9/80.2	69.2/88.2	80.4	79.7	86.7	63.6
$BERT_6$-KD	91.5	**86.2/80.6**	70.1/88.8	80.2	79.8	88.3	64.7
$BERT_6$-PKD	**92.0**	85.0/79.9	**70.7/88.9**	**81.5**	**81.0**	**89.0**	**65.5**
$BERT_3$-FT	86.4	80.5/**72.6**	65.8/86.9	74.8	74.3	84.3	55.2
$BERT_3$-KD	86.9	79.5/71.1	67.3/87.6	75.4	74.8	84.0	56.2
$BERT_3$-PKD	**87.5**	**80.7**/72.5	**68.1/87.8**	**76.7**	**76.3**	**84.7**	**58.2**

ALBERT

- Factorized embedding parameterization
  - Token embedding
  - O(V × H) to O(V × E + E × H).

- Cross-layer parameter sharing
  - share all parameters across layers

ALBERT: A Lite BERT for Self-supervised Learning of Language

ALBERT

- The reduction of parameters y ALBERT

Model		Parameters	Layers	Hidden	Embedding	Parameter-sharing
BERT	base	108M	12	768	768	False
	large	334M	24	1024	1024	False
ALBERT	base	12M	12	768	128	True
	large	18M	24	1024	128	True
	xlarge	60M	24	2048	128	True
	xxlarge	235M	12	4096	128	True

ALBERT

- Better performance
- Smaller model size
- But, more computation!

Model		Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg	Speedup
BERT	base	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3	4.7x
	large	334M	92.2/85.5	85.0/82.2	86.6	93.0	73.9	85.2	1.0
ALBERT	base	12M	89.3/82.3	80.0/77.1	81.6	90.3	64.0	80.1	5.6x
	large	18M	90.6/83.9	82.3/79.4	83.5	91.7	68.5	82.4	1.7x
	xlarge	60M	92.5/86.1	86.1/83.1	86.4	92.4	74.8	85.5	0.6x
	xxlarge	235M	**94.1/88.3**	**88.1/85.1**	**88.0**	**95.2**	**82.3**	**88.7**	0.3x

Summary

- Introduce three representative works on compressing pre-trained language models
- Small models with slightly worse results
- Use them when the computation resource is limited

Continue Learning

THUNLP

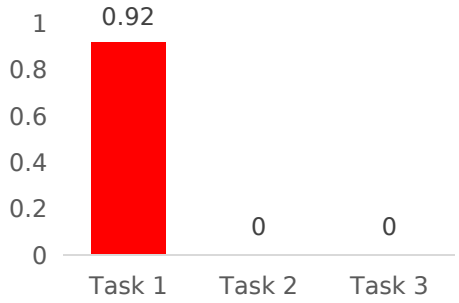
Continual Learning

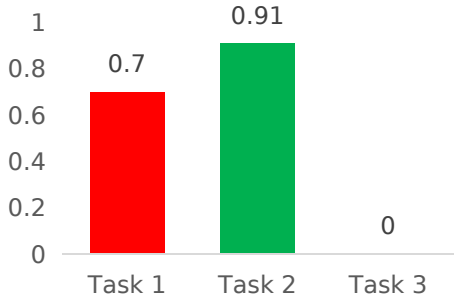
- Continual Learning
  - also known as Lifelong Learning / Incremental Learning
  - The model is trained on a sequence of tasks
  - After training the model on the k-th task, the model should handle all known tasks (1~k-th tasks)

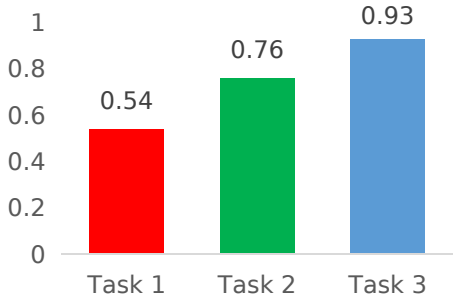
Aljundi et al. Memory aware synapses: Learning what (not) to forget. ECCV.

Catastrophic Forgetting

- Catastrophic Forgetting
  - Continual Learning models suffer from the catastrophic forgetting problem
  - Because they overfit new tasks and forget old







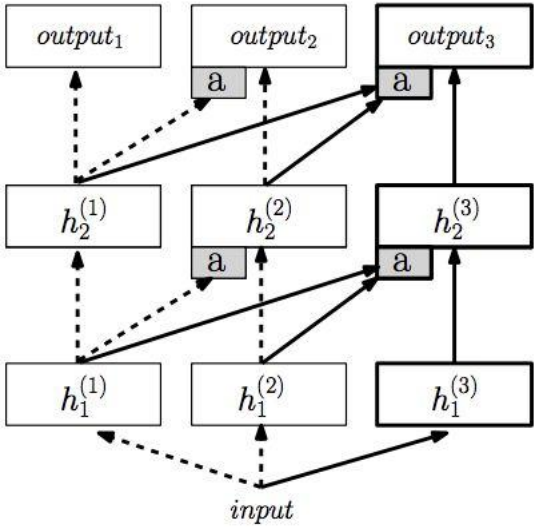
Aljundi et al. Memory aware synapses: Learning what (not) to forget. ECCV.

Continual Learning Models

- Continual Learning Models
  - Dynamic Architecture Methods
  - Consolidation-based Methods
  - Memory-based Methods

Dynamic Architecture Methods

- Dynamic Architecture Methods
  - Progressive Neural Networks
    - For each new task, new parameters are set
    - The old parameters will be concatenated with the new parameters for training the model on the new task

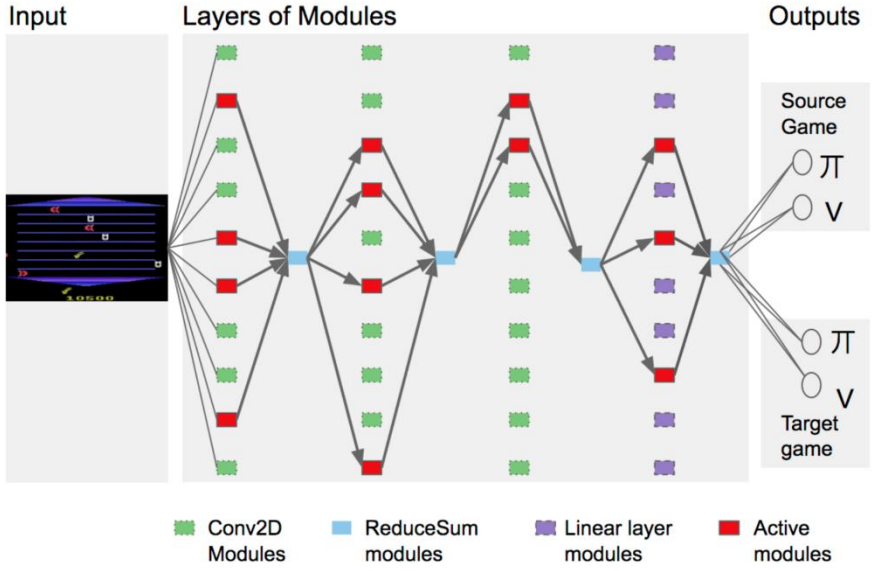


$$h_i^{(k)} = f\left(W_i^{(k)} h_{i-1}^{(k)} + \sum_{j<k} U_i^{(k:j)} h_{i-1}^{(j)}\right),$$

Rusu et al. Progressive neural networks. 2016.

Dynamic Architecture Methods

- Dynamic Architecture Methods
  - PathNet
    - The number of neural modules are fixed.
    - For each new task, the system will select several modules to construct a corresponding model
    - Only the selected modules will be changed



Fernando et al. Evolution channels gradient descent in super neural networks. 2017

Continual Learning Models

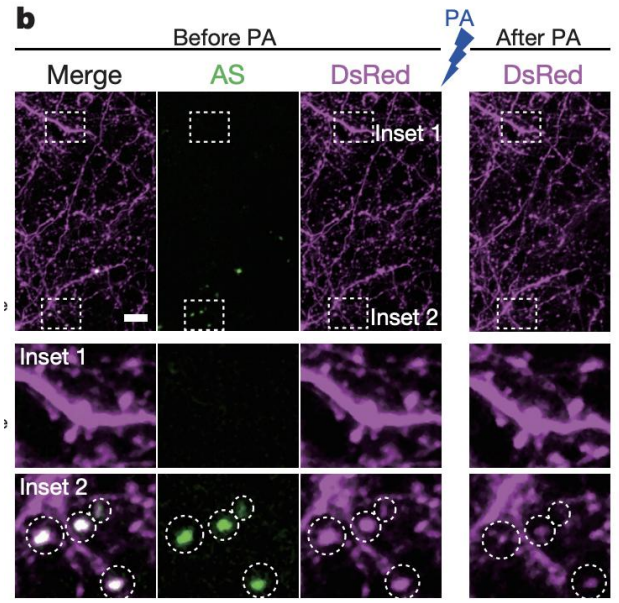
- Continual Learning Models
  - Dynamic Architecture Methods
  - Consolidation-based Methods
  - Memory-based Methods

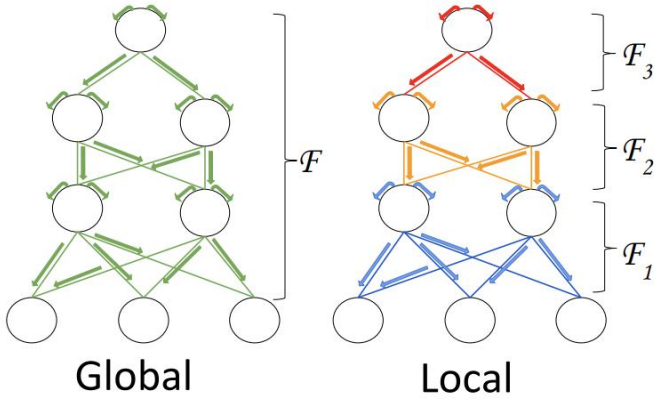
Consolidation-based Methods

- Consolidation-based Methods
  - Human has a consolidation mechanism for long-term memory
  - Design models similar to the consolidation mechanism in human brains

Consolidation-based Methods

- Consolidation-based Methods
  - Synapse-based methods
    - The synapses working for long-term memory are more stable than others
    - Accordingly, we hope those parameters used for old tasks change slowly
    - Compute the parameter weights of according to the gradient flows

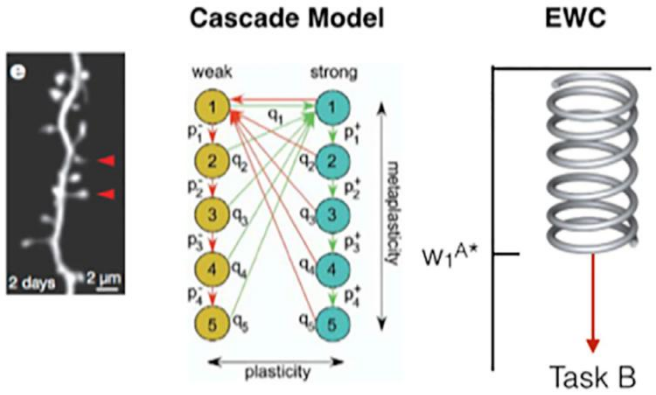


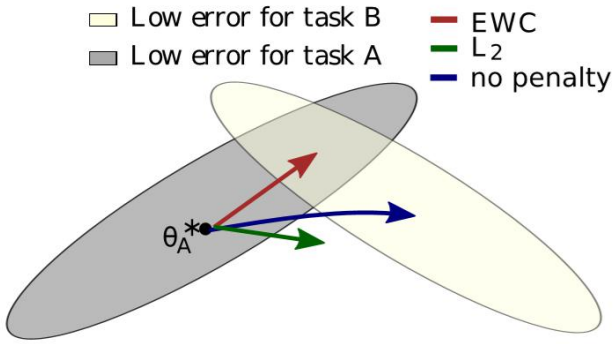


Hayashi-Takagi et al. Labelling and optical erasure of synaptic memory traces in the motor cortex. Nature. 2015.
Aljundi et al. Memory aware synapses: Learning what (not) to forget. ECCV. 2018.

Consolidation-based Methods

- Consolidation-based Methods
  - Elastic Weight Consolidation
    - Design a method similar to the synaptic plasticity
    - Compute the parameter weights of according to the fisher information





Fusi et al. Cascade models of synaptically stored memories. Neuron. 2005.
Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. NAS. 2017.

Continual Learning Models

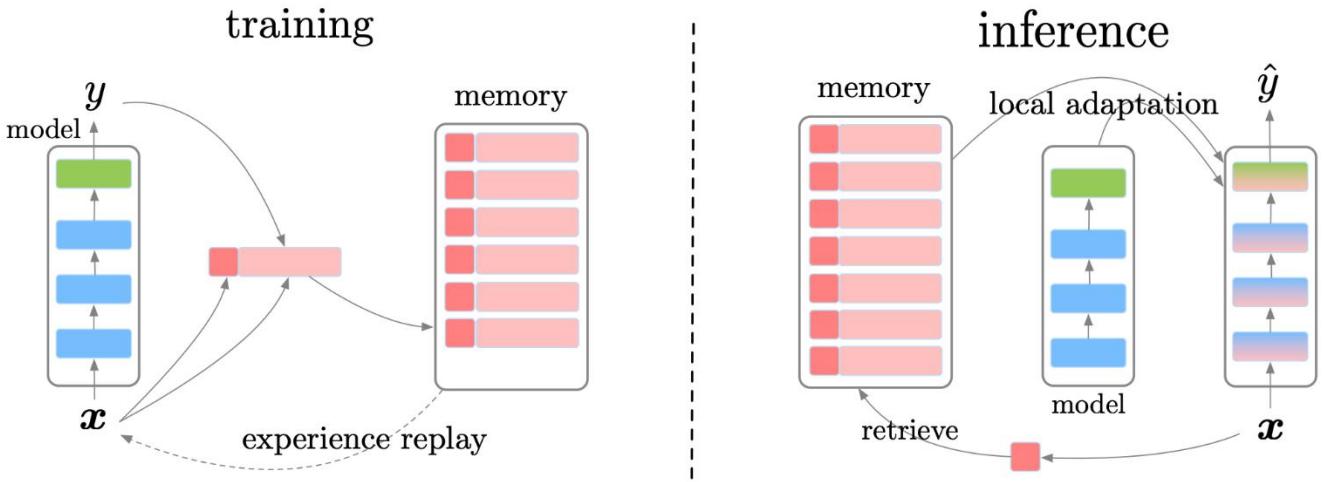
- Continual Learning Models
  - Dynamic Architecture Methods
  - Consolidation-based Methods
  - Memory-based Methods

Memory-based Methods

- Memory-based Methods
  - Human has a experience replay mechanism for long-term memory
  - Design models similar to the experience replay mechanism in human brains
  - The memory-based methods have been proven to be the most promising for NLP tasks

Memory-based Methods

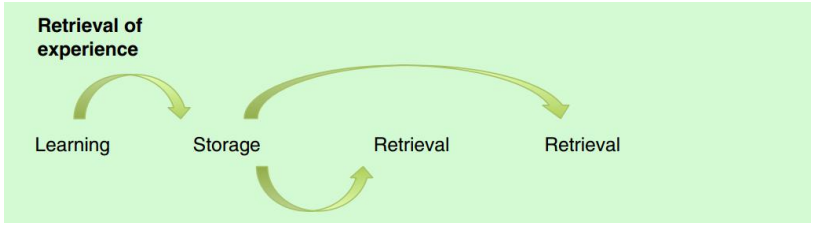
- Memory-based Methods
  - Episodic Memory Replay
    - Remember a few examples in old tasks
    - continually learn them with emerging new tasks to alleviate catastrophic forgetting



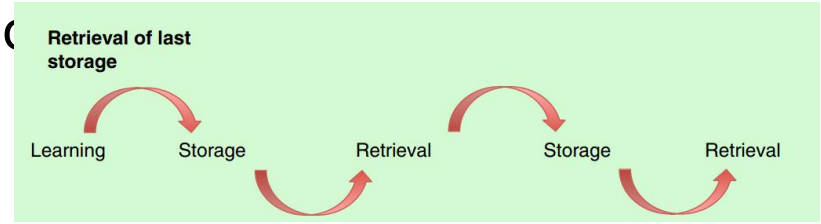
Lopez-Paz et al. Gradient episodic memory for continual learning. In Proceedings of NeurIPS 2017.
d'Autume et al. Episodic Memory in Lifelong Language Learning. In Proceedings of NeurIPS 2019.

Memory-based Methods

- Memory Model in Neuroscience
  - The conventional view
    - memories are stored once
    - each time the memory is activated, a trace of the original experience is retrieved
  - The reconsolidation view
    - memories are susceptible to change each time they are retrieved.
    - The next time the memory is activated **the version stored during the last retrieval**, rather than the original version



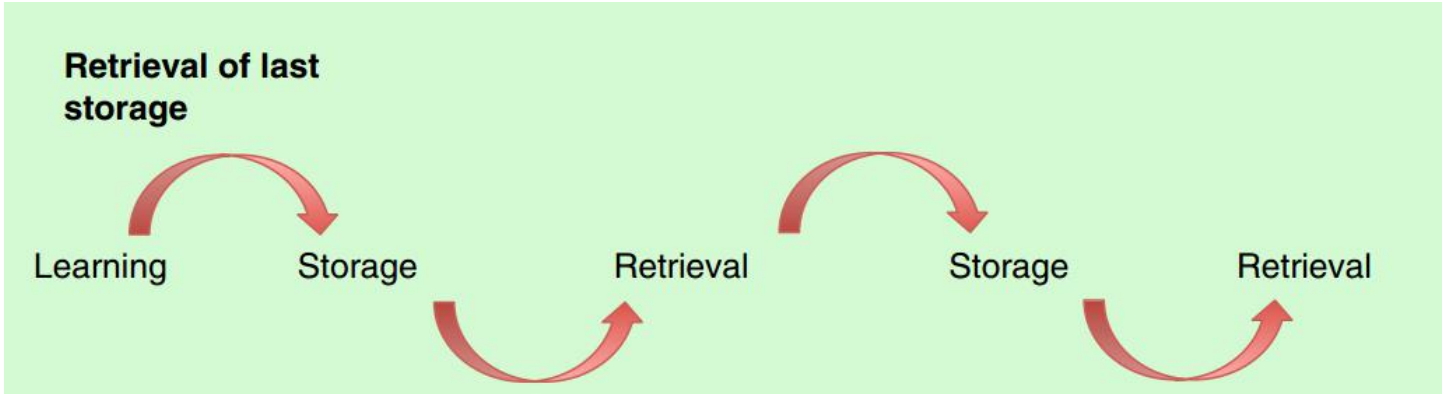
is ac



Alberini et al. Memory reconsolidation[J]. Current Biology, 2013, 23(17): R746-R750.

Memory-based Methods

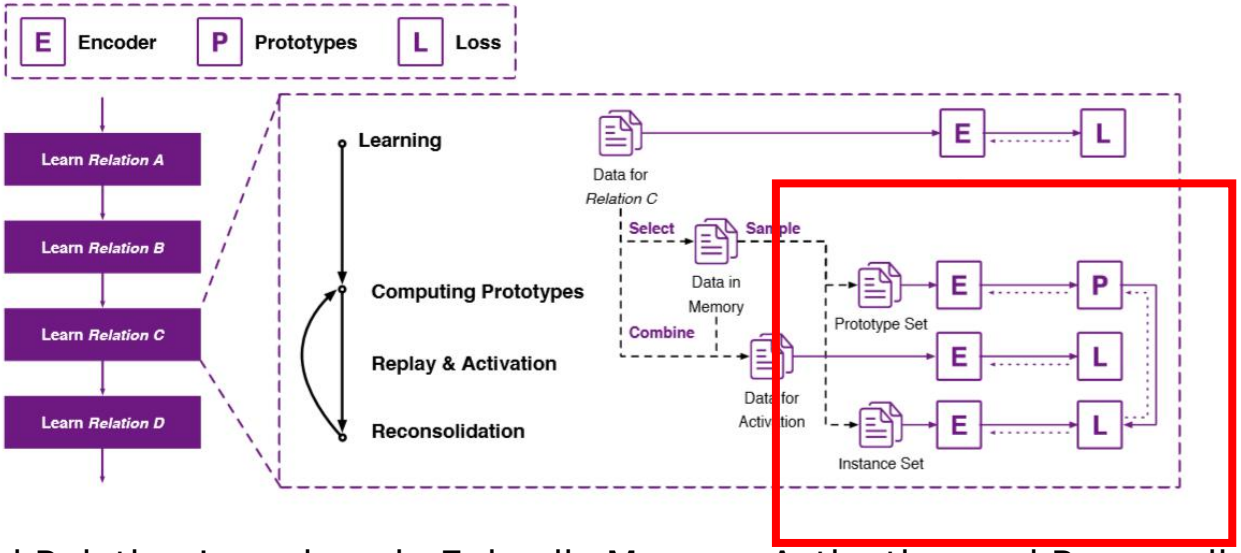
- Memory Model in Neuroscience
  - Reactivation & econsolidation
    - The **reactivation** (Retrieval) of consolidated memory will **trigger** a **reconsolidation** stage
    - The **reconsolidation** stage can **help** continually **maintain memory** (Memory Storage)
    - Memory is weak in the **reconsolidation** stage (~6 hours), and easy to be changed or erased
    - i.e., we can **edit memory** in the this stage, or **apply reconsolidation exercises** to help brain go through this stage



Alberini et al. Memory reconsolidation[J]. Current Biology, 2013, 23(17): R746-R750.

Memory-based Methods

- Memory-based Methods
  - Reconsolidation exercise
    - Compute prototypes with memorized examples for all classes
    - Conduct memory replay & activation, and use prototypes to classify memorized examples to help reconsolidate memory



Han et al. Continual Relation Learning via Episodic Memory Activation and Reconsolidation. ACL 2020.

Grammatical Error Correction

THUNLP

Outline

- Grammatical Error Correction
- Sequence to sequence models
- Model Ensemble
- Non-autoregressive models

Grammatical Error Correction

- Background
  - Grammatical Error Correction (GEC) aims to correct writing errors
  - The a Mobile phone is a marvelous invention to charge the world
  - The Mobile phone is a marvelous invention to change the world.

Grammatical Error Correction

- Background
  - For English, there are 28 kinds of grammatical errors

Type	Description	Example
ArtOrDet	Article or determiner	It is obvious to see that [*internet* → *the internet*] saves people time and also connects people globally.
Wci	Wrong collocation/idiom	Early examination is [*healthy* → *advisable*] and will cast away unwanted doubts.
Rloc−	Redundancy	It is up to the [*patient's own choice* → *patient*] to disclose information.
Nn	Noun number	A carrier may consider not having any [*child* → *children*] after getting married.
Vt	Verb tense	Medical technology during that time [*is* → *was*] not advanced enough to cure him.
Mec	Spelling, punctuation, capitalization, etc.	This knowledge [*maybe relavant* → *may be relevant*] to them.
Pref	Pronoun reference	It is everyone's duty to ensure that [*he or she* → *they*] undergo regular health checks.
Wform	Word form	The sense of [*guilty* → *guilt*] can be more than expected.

Grammatical Error Correction

- Background
  - There are also some troubles in existing GEC systems

This seemingly important events brought us [*a news*].

This seemingly important events brought us [*a very good and exciting news*].

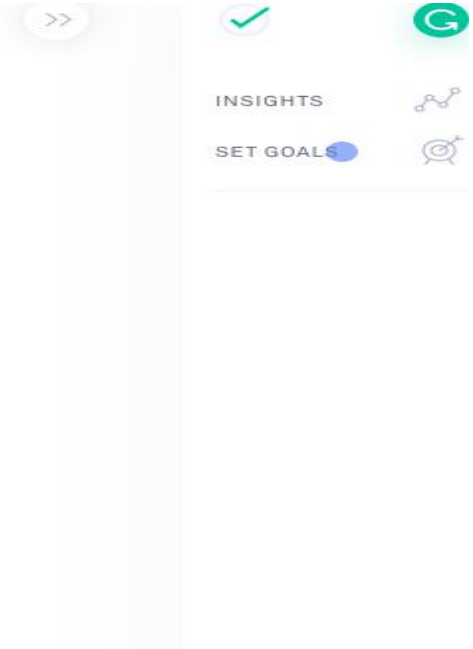
Grammatical Error Correction

- Background
  - There are also some troubles in existing GEC systems

    He becomes [*the*] better man.

    The procedure called genetic testing is [*an expensive procedure*].

To err is human; to edit, divine.



Grammatical Error Correction

- Datasets
  - CoNLL-2014
  - FCE
  - BEA19
  - JFLEG
  - MQR
  - Lang8

Grammatical Error

Fluency

Query Rewriting

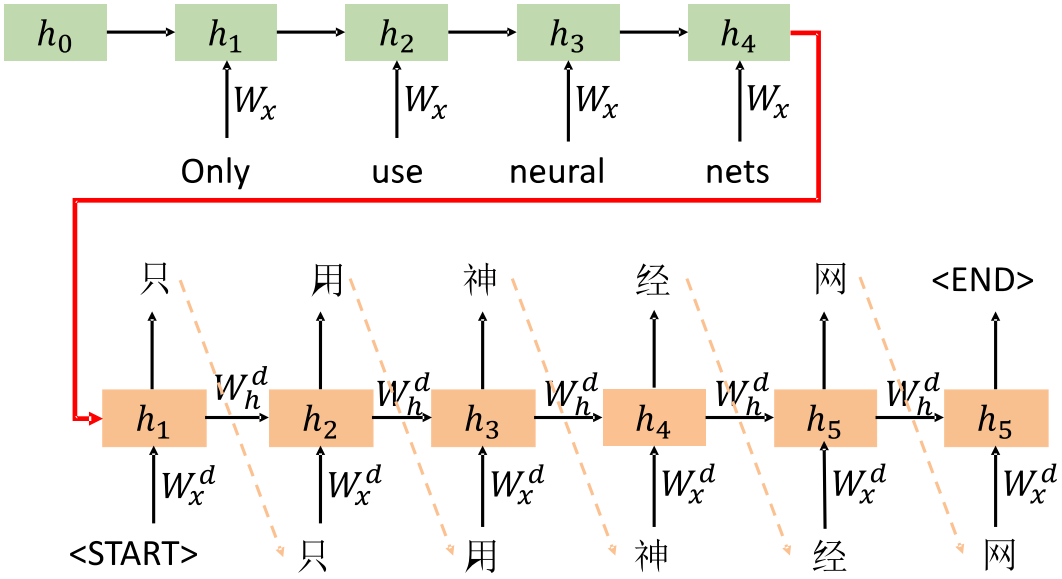
Grammatical Error Correction (Crowdsourcing from lang8 website)

Outline

- Grammatical Error Correction
- Sequence to sequence models
- Model Ensemble
- Non-autoregressive models

Sequence to sequence models

- Following architectures can be chose
  - CNN
  - RNN
  - Transformers



Sequence to sequence models

- Following architectures can be chose
  - CoNLL-2014
    - Transformer is better

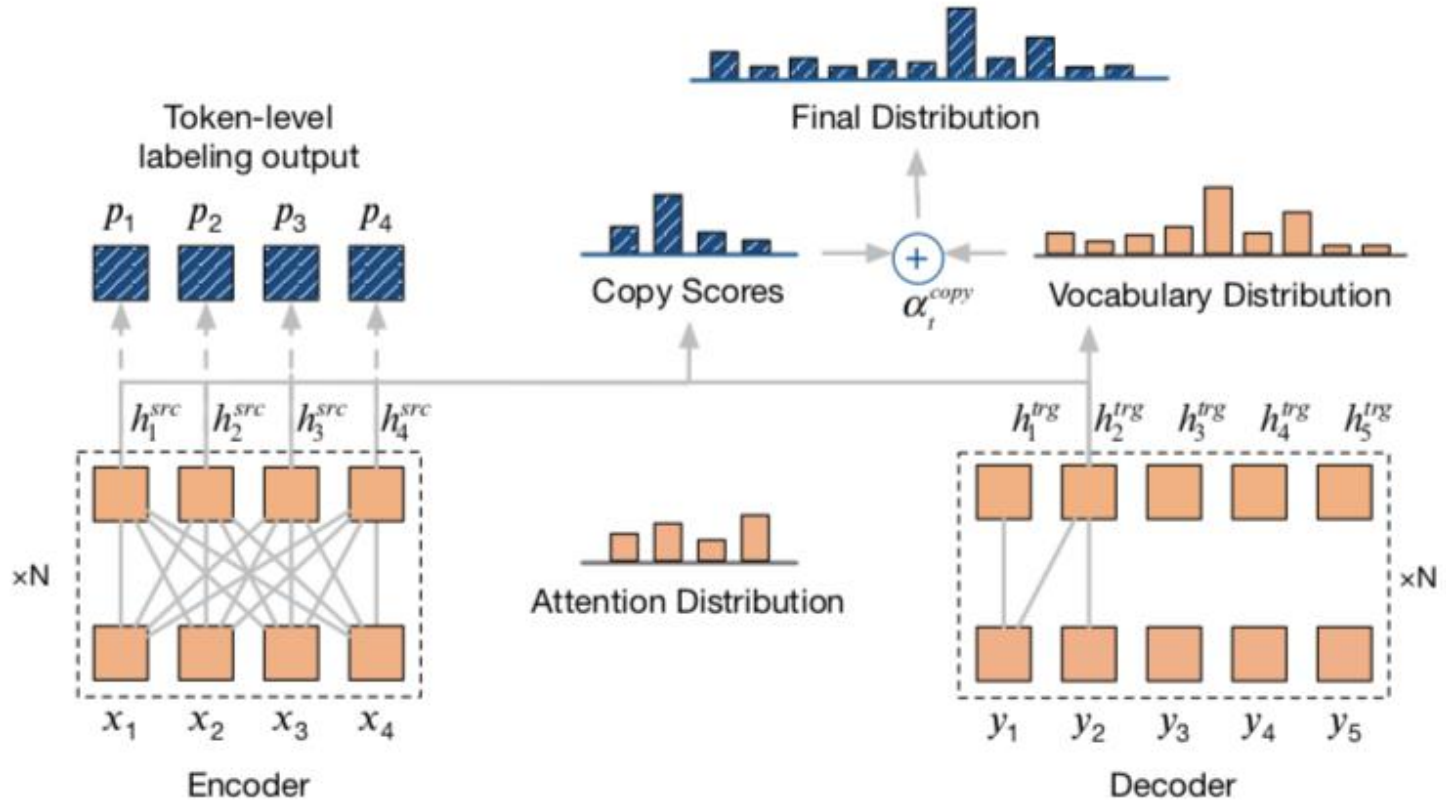
Model	Prec.	Recall	F0.5
SMT	39.71	30.10	37.33
CNN	59.68	23.15	45.36
Transformer	**55.96**	**30.73**	**48.07**

Sequence to sequence models

- Consider this feature in GEC
  - Most words are reserved
    - The a Mobile phone is a marvelous invention to charge the world
    - The Mobile phone is a marvelous invention to change the world.
    - Unchanged words: 10
    - Changed words: 3

Sequence to sequence models

- Copy Mechanism in GEC



Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. Zhao et a

Sequence to sequence models

- Copy Mechanism in GEC
  - Achieve better performance

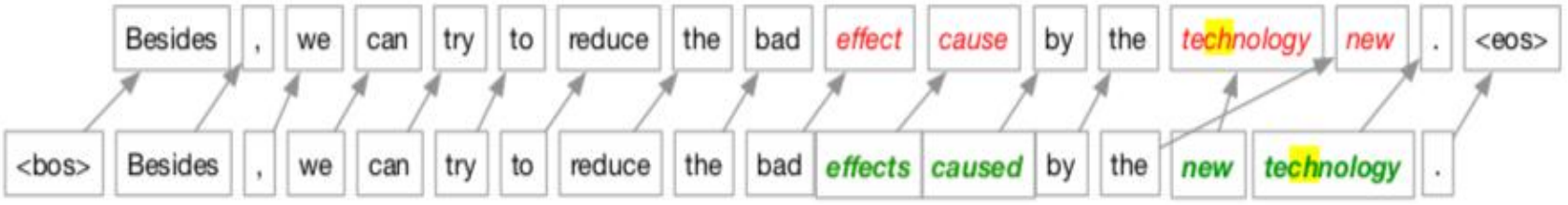
Model	Pre.	Rec.	$F_{0.5}$	Imp.
Transformer	55.96	30.73	48.07	-
+ Copying	65.23	33.18	**54.67**	+6.60
Ignoring UNK words as edits				
Transformer	65.26	30.63	53.23	-
+ Copying	65.54	33.18	54.85	+1.62
+ Pre-training				
Copy-Augmented Transformer	65.23	33.18	54.67	-
+ Pre-training Decoder (partially pre-trained)	68.02	34.98	57.21	+2.54
+ Denosing Auto-encoder (fully pre-trained)	68.97	36.98	**58.80**	+4.13
+ Multi-tasks				
Copy-Augmented Transformer	67.74	40.62	**59.76**	-

Table 5: Single Model Ablation Study on CoNLL 2014 Test Data Set.

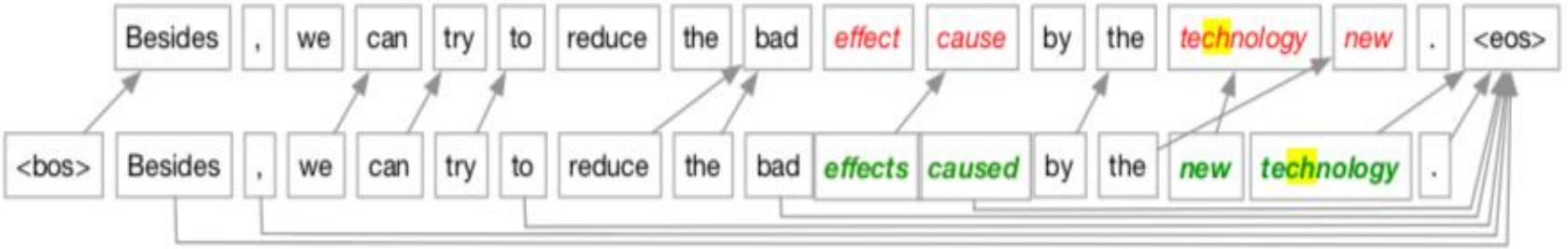
Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. Zhao et a

Sequence to sequence models

- Copy Mechanism in GEC
  - Better alignments



(a) Copy Alignment



(b) Encoder-Decoder Atttention Alignment

Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. Zhao et a

Sequence to sequence models

- Copy Mechanism in GEC
  - Achieve better performance

Model	Pre.	Rec.	$F_{0.5}$	Imp.
Transformer	55.96	30.73	48.07	-
+ Copying	65.23	33.18	**54.67**	+6.60
Ignoring UNK words as edits				
Transformer	65.26	30.63	53.23	-
+ Copying	65.54	33.18	54.85	+1.62
+ Pre-training				
Copy-Augmented Transformer	65.23	33.18	54.67	-
+ Pre-training Decoder (partially pre-trained)	68.02	34.98	57.21	+2.54
+ Denosing Auto-encoder (fully pre-trained)	68.97	36.98	**58.80**	+4.13

Pre-train is effectiveness for GEC models

Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. Zhao et a

Sequence to sequence models

- Data Augmentation
  - Week Supervision Corpus
    - Wikipedia edit history (Lichtarge et al
    - Github typos (Hagiwara and Mita, 20
    - Word confusion set (Grundkiewicz et

Word	Confusion set
has	Haas HS Hans hats gas had Ha ha As as
is	IRS ISO OS US us Si its
island	islands Iceland slant
issued	issues issue used issuers eased sued assumed assured missed
student	students strident stunt
walking	talking whaling
large	larger lag lake barge Lodge lodge
largest	latest longest

(Artillery in 1941 and was medically dis-charged) : (Artillery in 1941 he was later medically discharged with)

(Special terms have been coined to denote many imfortant technical concepts in the game of Go. Such technical) : (Players of the game of Go often use jargon terms to describe situations on the board and surrounding the game. Such technical)

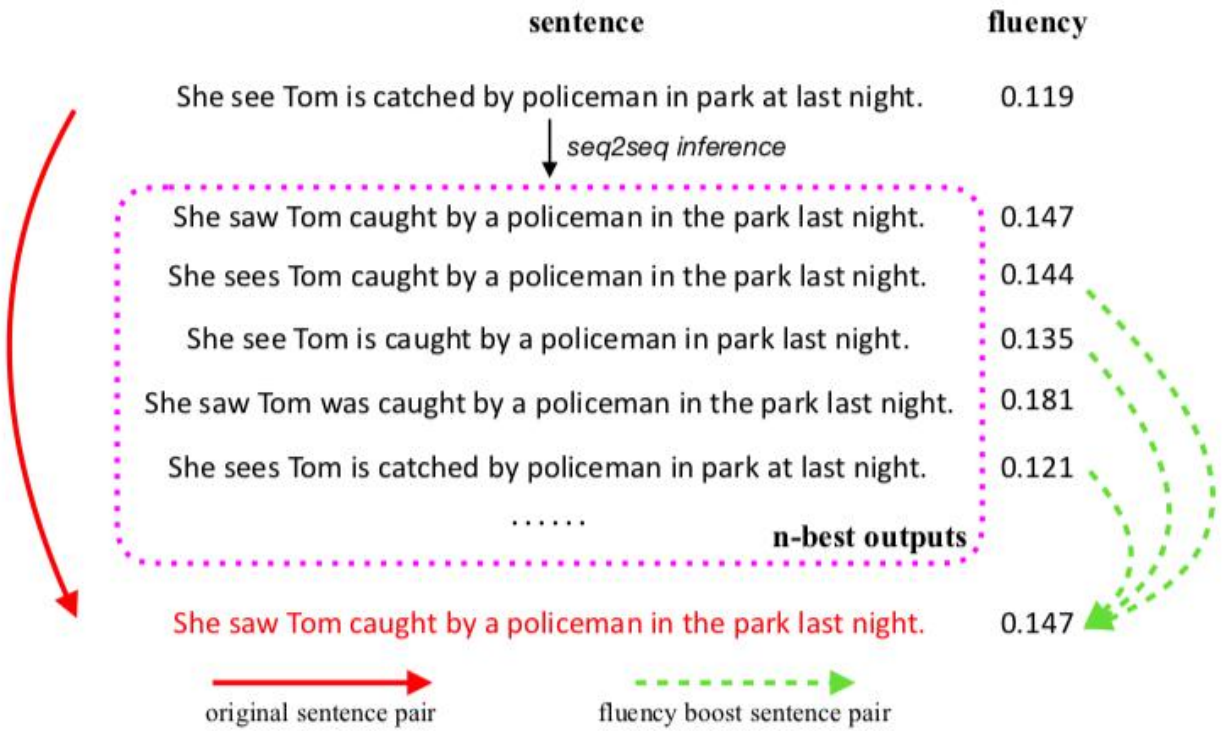
(The County of Fitzroy is a county in Queensland,) : (The County of Fitzroy is a county (a cadastral division) in Queensland,)

Sequence to sequence models

- Data Augmentation
  - Generate week supervision data
    - Round trip NMT (Lichtarge et al., 2019)

Sequence to sequence models

- Data Augmentation
  - Generate week supervision data
    - Round trip NMT (Lichtarge et al., 2019)
    - N-best hypotheses from GEC model (Ge et al., 2018)



Sequence to sequence models

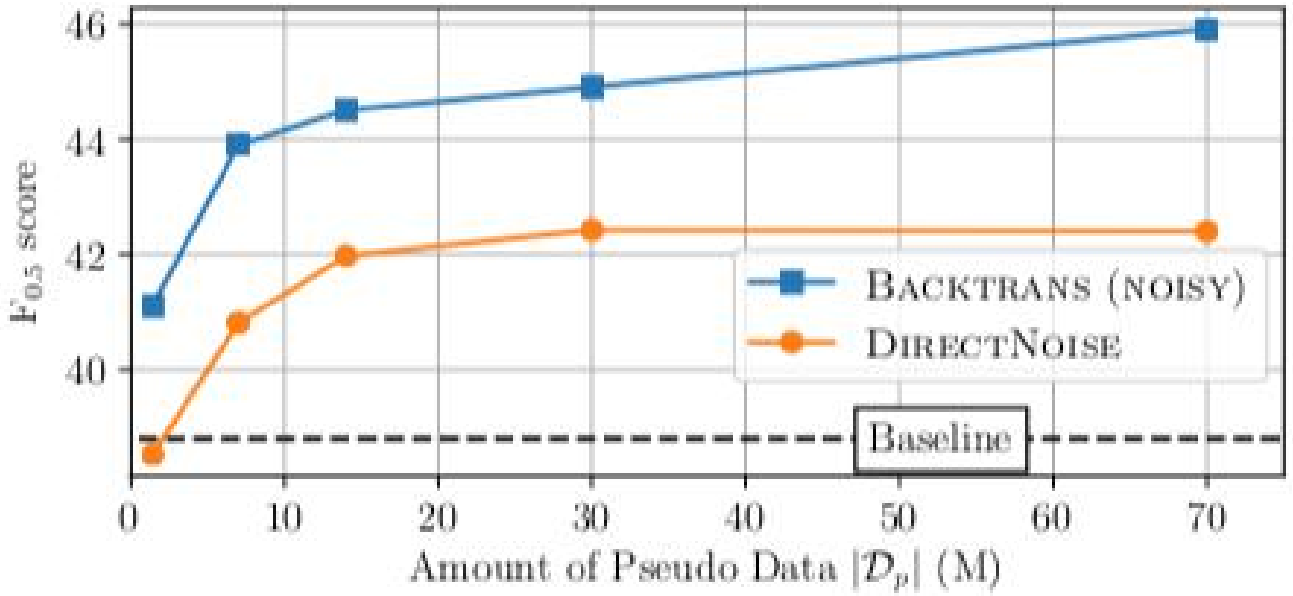
- Data Augmentation
  - Performance
    - Transformer

Model	Ensemble	CoNLL-2014 (M^2 scorer) Prec.	Rec.	$F_{0.5}$	CoNLL-2014 (ERRANT) Prec.	Rec.	$F_{0.5}$	JFLEG GLEU	BEA-test (ERRANT) Prec.	Rec.	$F_{0.5}$
Chollampatt and Ng (2018)		60.9	23.7	46.4	-	-	-	51.3	-	-	-
Junczys-Dowmunt et al. (2018)		-	-	53.0	-	-	-	57.9	-	-	-
Grundkiewicz and Junczys-Dowmunt (2018)		66.8	34.5	56.3	-	-	-	61.5	-	-	-
Lichtarge et al. (2019)		65.5	37.1	56.8	-	-	-	61.6	-	-	-
Chollampatt and Ng (2018)	✓	65.5	33.1	54.8	-	-	-	57.5	-	-	-
Junczys-Dowmunt et al. (2018)	✓	61.9	40.2	55.8	-	-	-	59.9	-	-	-
Lichtarge et al. (2019)	✓	66.7	43.9	60.4	-	-	-	**63.3**	-	-	-
Zhao et al. (2019)	✓	71.6	38.7	61.2	-	-	-	61.0	-	-	-
Grundkiewicz et al. (2019)	✓	-	-	64.2	-	-	-	61.2	72.3	60.1	69.5
PRETLARGE		67.9	44.1	61.3	61.2	42.0	56.0	59.7	65.5	59.4	64.2

An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction . Kiyono et al. EMNLP 2019.

Sequence to sequence models

- Data Augmentation
  - Performance
    - Transformer



An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction . Kiyono et al. EMNLP 2019.

Outline

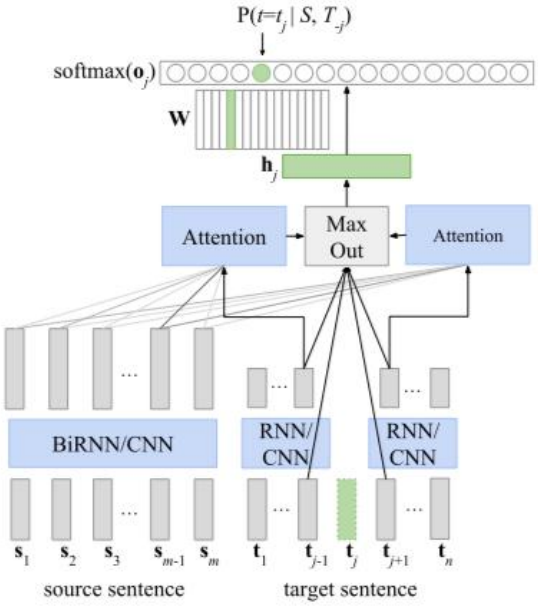
- Grammatical Error Correction
- Sequence to sequence models
- Model Ensemble
- Non-autoregressive models

Model Ensemble

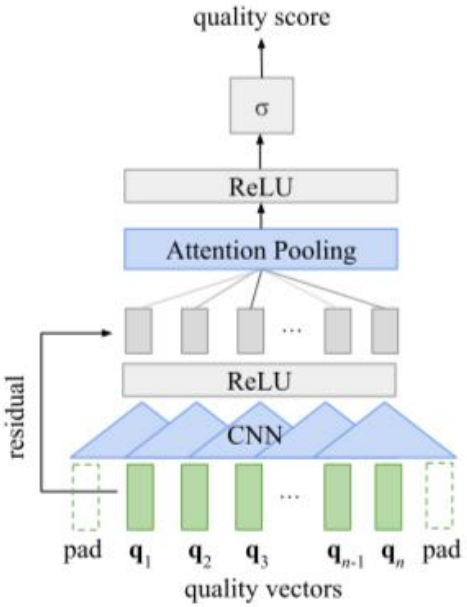
- Language model based models with BERT
  - Perplexity
  - BERT based sentence label prediction
    - Use [CLS] hidden state to predict the sentence correctness
  - BERT base token detection
    - Predict token correctness
    - Average all token correctness probability

Model Ensemble

- Neural quality estimation
  - GEC Model (Learn language features)
  - Quality Estimation Model (Prediction F0.5 score)



GEC Model



Quality Estimation Model

Neural Quality Estimation of Grammatical Error Correction. Chollampatt et al. EMNLP 2018.

Model Ensemble

- Neural quality estimation
  - Performance

	FCE	CoNLL-2014
Best published results		
G&J (2018) w/ SpellCheck	–	56.25
JGGH (2018)	–	55.8
C&N (2018) w/ SpellCheck	–	54.79
Re-scorer trained with 5.4k sents. from NUCLE		
Base GEC	47.53	55.86
+ SpellCheck	47.79	56.43
Re-scorer trained with FCE+CoNLL dev set		
Base GEC	47.29	55.72
+ $\mathbf{M^2}$ ($\mathbf{NQE_{ALL}}$)	**48.47***	**55.97***
+ SpellCheck	48.70	56.52
Base GEC + M^2 (Oracle)	76.70	80.74

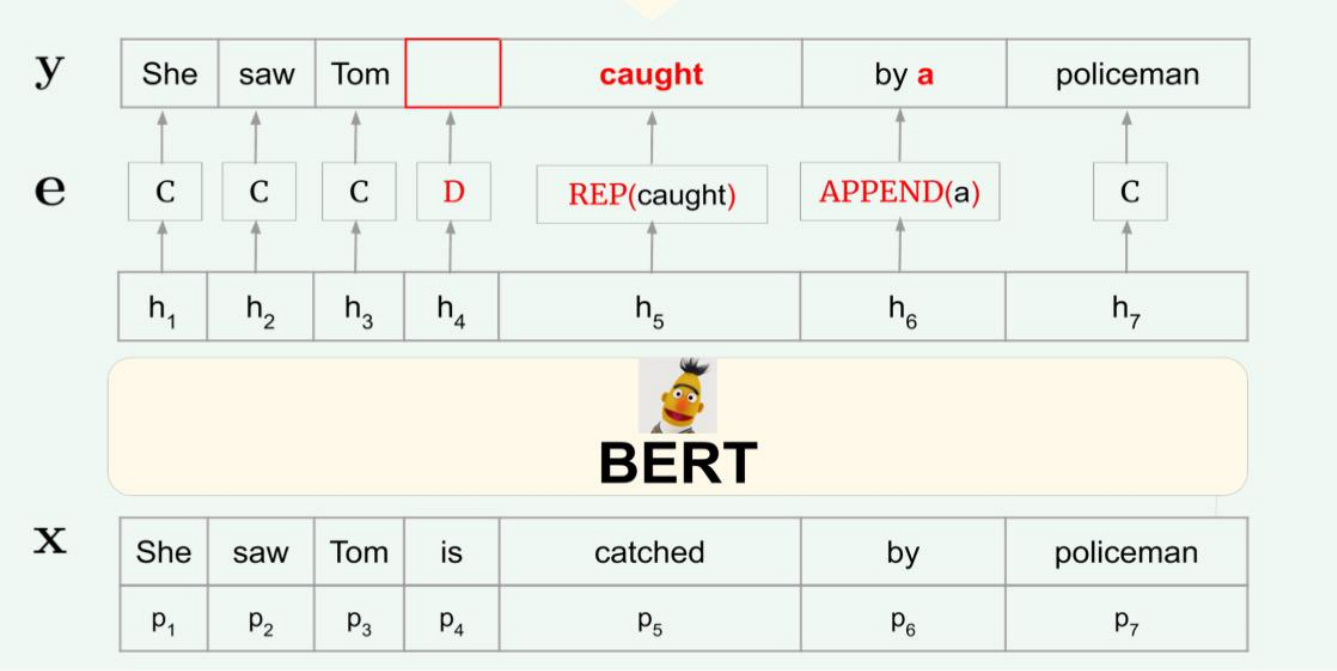
Neural Quality Estimation of Grammatical Error Correction. Chollampatt et al. EMNLP 2018.

Outline

- Grammatical Error Correction
- Sequence to sequence models
- Model Ensemble
- Non-autoregressive models

Non-autoregressive models

- We can correct sentences iteratively
  - Sentences can be edited with three operations
  - Insert; Delate; Replace



Parallel Iterative Edit Models for Local Sequence Transduction. Awasthi et al. EMNLP 2019.

Non-autoregressive models

- We can correct sentences iteratively
  - Parallel Iterative Edit Models (PIE)

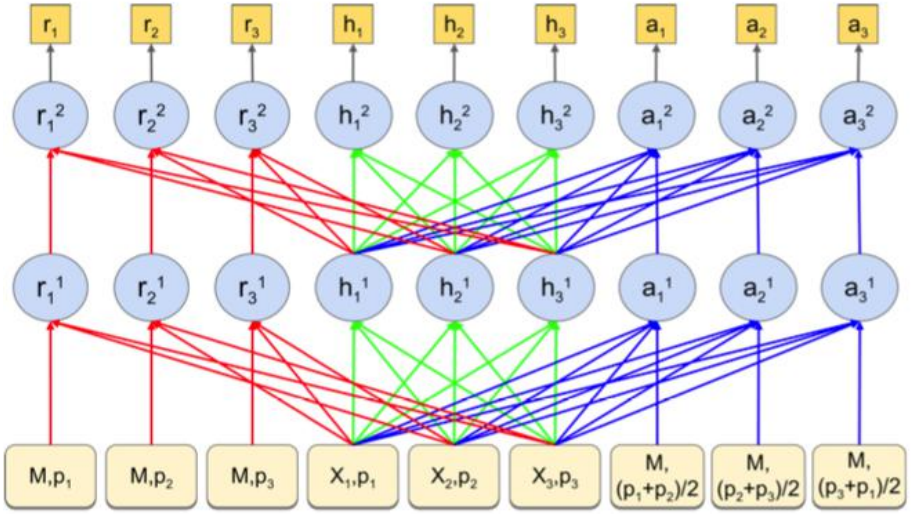


Figure 2: A Parallel Edit Architecture based on a 2-layer bidirectional transformer. Input sequence length is 3. Arrows indicate attention mask for computation of $\mathbf{h}_i^l, \mathbf{r}_i^l, \mathbf{a}_i^l$ at position i for layer l.

Parallel Iterative Edit Models for Local Sequence Transduction. Awasthi et al. EMNLP 2019.

Non-autoregressive models

- We can correct sentences iteratively
  - Parallel Iterative Edit Models (PIE)
  - Performance

Work	CONLL-14			JFLEG
	P	R	$F_{0.5}$	GLEU$^+$
Zhao et al. (2019)	67.7	40.6	59.8	-
Lichtarge et al. (2019)	65.5	37.1	56.8	61.6
Chollampatt and Ng (2018a)	69.9	23.7	46.4	51.3
PIE (This work)	66.1	43.0	59.7	60.3

Parallel Iterative Edit Models for Local Sequence Transduction. Awasthi et al. EMNLP 2019.

Non-autoregressive models

- We can correct sentences iteratively
  - Parallel Iterative Edit Models (PIE)
  - Performance

x	I started invoving into Facebook one years ago .
PIE1	I started ***involving in*** Facebook one ***year*** ago .
PIE2	I started ***involved*** in Facebook one year ago .
PIE3	I started ***getting*** involved in Facebook one year ago .
x	Since ancient times , human interact with others face by face .
PIE1	Since ancient times , humans ***interacted*** with others face ***to*** face .
PIE2	Since ancient times , humans ***have*** interacted with others face to face .
x	However , there are two sides of stories always .
PIE1	However , there are ***always*** two sides ***to*** stories always .
PIE2	However , there are always two sides to ***the*** stories .
PIE3	However , there are always two sides to the ***story*** .

Parallel Iterative Edit Models for Local Sequence Transduction. Awasthi et al. EMNLP 2019.

Conclusion

- GEC can be regarded as a generative task
- Source and target sentences share most tokens
- Copy mechanism and Non-autoregressive model help better edit sentence
- GEC task is usually regarded as a low resource task

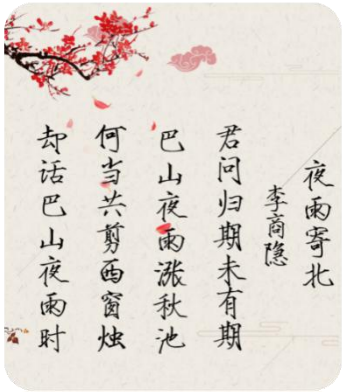
Outline

- Background & Overview
- Models for Quality Improvement
- Models for Attribute Control
- Summary

Outline

- Background & Overview
- Models for Quality Improvement
- Models for Attribute Control
- Summary

Background & Overview

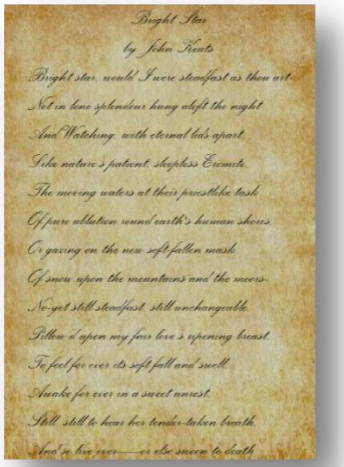


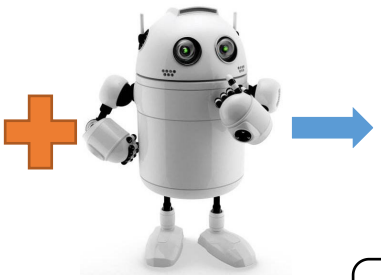
Poetry

Elegant Expressions

Colorful Contents

Diverse Styles





Exploration

- Human Writing Mechanism
- Computational Creativity
- Entertainment
- Humanizing AIs

Application

- Advertising
- Poetry

Background & Overview

Poetry Generation ➡ Conditional Autoregressive Sequence Generation

x: a poem with n lines

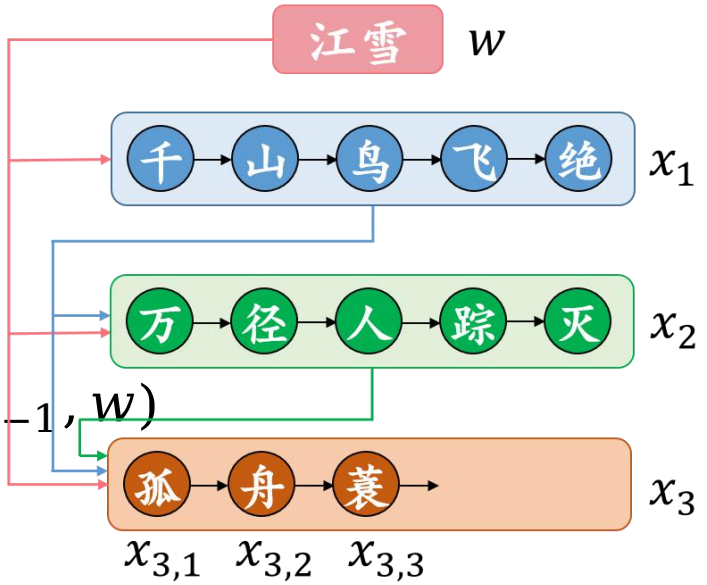
$x = x_1, \ldots, x_n$ $x_{i,j}$ j-th word in i-th line

w user-specified topic (keywords / title)

$$p(x|w) = \prod_{i=1}^{n} p(x_i|x_{1:i-1}, w)$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{|x_i|} p(x_{i,j}|x_{i,1:j-1}, x_{1:i-1}, w)$$

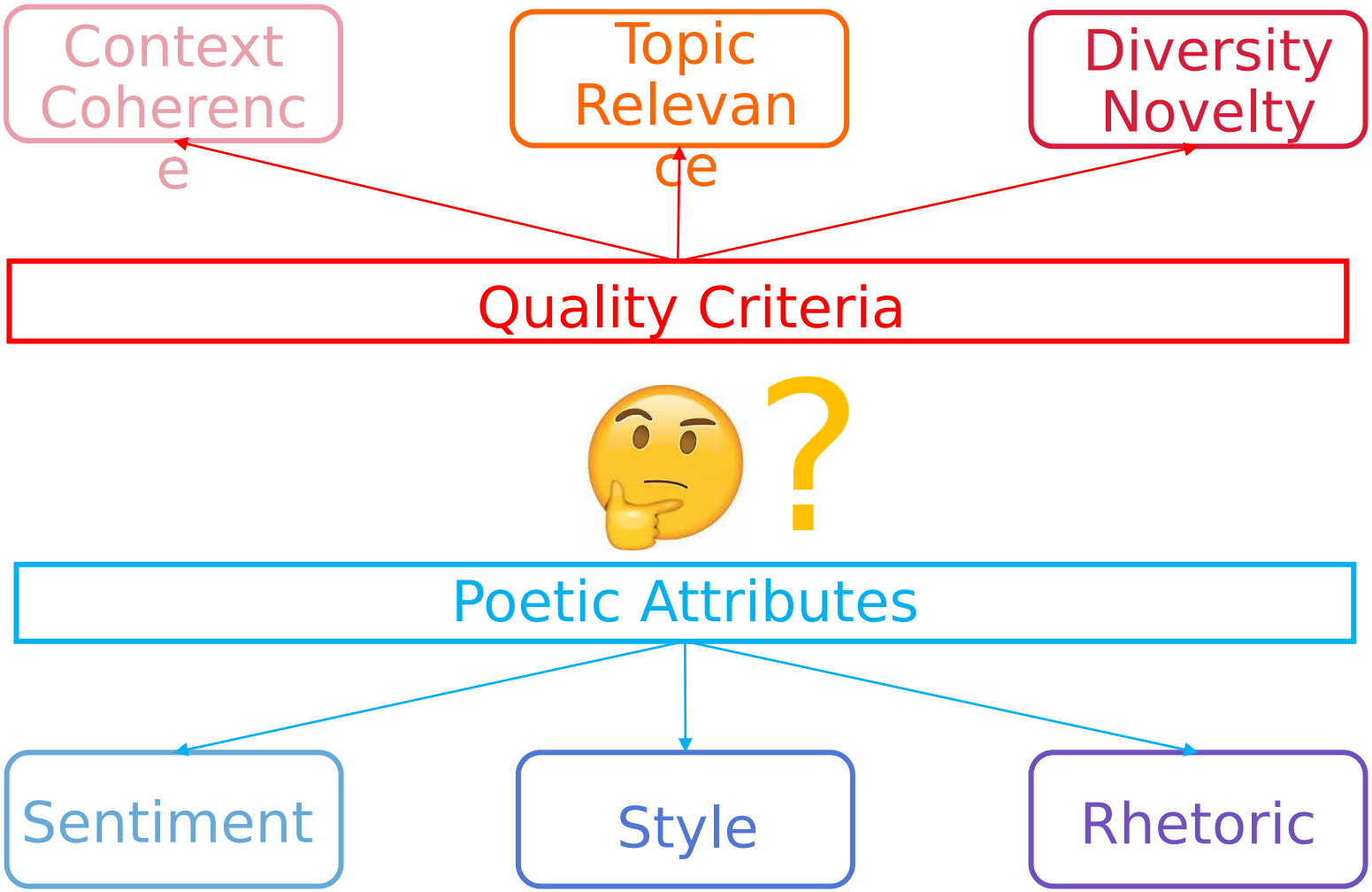
$$x^* = \arg\max_x \sum_i \sum_j \log p_\theta(x_{i,j}|x_{i,1:j-1}, x_{1:i-1}, w)$$



RNN/GRU/LSTM + Attention

(Zhang and Lapata, 2014; Yan 2016; Wang et al., 2016; Yi et al., 2017)

Background & Overview



Outline

- Background & Overview
- Models for Quality Improvement
- Models for Attribute Control
- Summary

Models for Quality Improvement

Improve Context Coherence

闺怨
Sorrow of a Young Bride in Her Boudoir
王昌龄

闺中少妇不知愁。
The young bride in her boudoir does not know what grieves,

春日凝妆上翠楼。
She mounts the tower, gaily dressed, on a spring day.

忽见陌头杨柳色，
Suddenly seeing by roadside green willow leaves,

悔教夫婿觅封侯。
How she regrets her lord seeking fame far away!

Close Connection

Natural Transition

Consistent Theme & Aesthetic

Poetry: Discourse-Level T

春风
Spring Breeze

江上春风吹绿杨，
Spring breeze blows the green willows on riverbank.

月明天地白皑皑。
Bright moonlight makes the sky and the ground turn white.

百年功业无消息，
The war, which has lasted for one century, won't be over.

万古英雄事已灰。
The hero died and his corpse has already become dust.

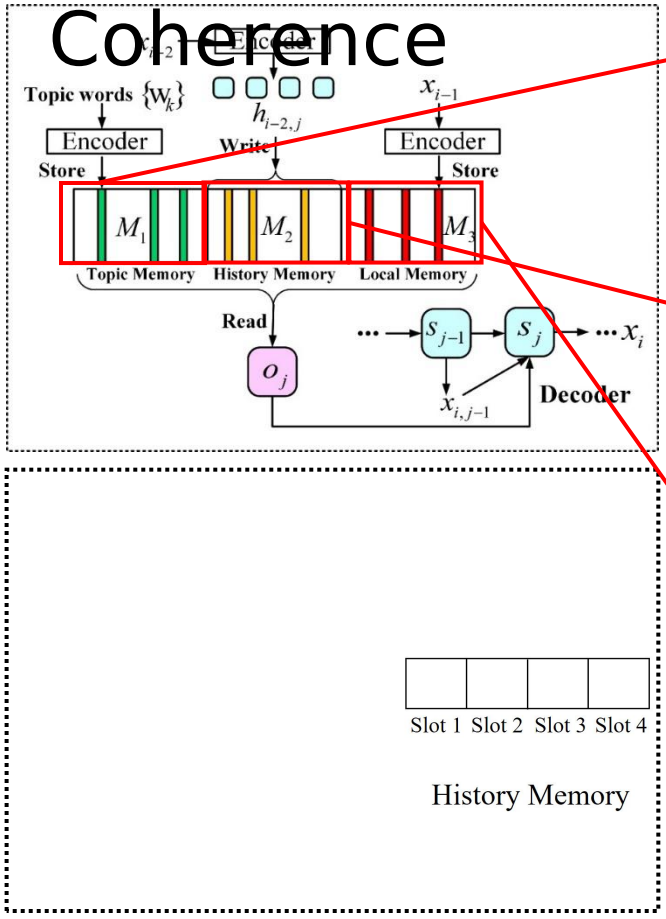
Inconsistent Topics

No Transition

A poem generated by RNNPG (Zhang and Lapata, 2014)

Models for Quality Improvement

Improve Context Coherence —WM Model(Yi et al., 2018)



Topic Memory maintains each M_i explicitly and independently.

Memory reading: $M = [M_1; M_2; M_3]$

$$\alpha_r = A_r(M, [s_{t-1}; v_{i-1}]),$$

$$o_t = \sum_{k=1}^{K} \alpha_r[k] * M[k],$$

Local Memory is dynamic and informative characters, which is dynamically read and written.

Memory writing:

$$\alpha_w = A_w(\tilde{M}_2, [h_t; v_{i-1}]),$$

$$\beta[k] = I(k = \arg\max_j \alpha_w[j]),$$

$$\tilde{M}_2[k] \leftarrow (1 - \beta[k]) * \tilde{M}_2[k] + \beta[k] * h_t,$$

Short-term memory is important for generating couplet lines.

白日依山尽，
黄河入海流。

Yi et al. Chinese Poetry Generation with a Working Memory Model. IJCAI 2018

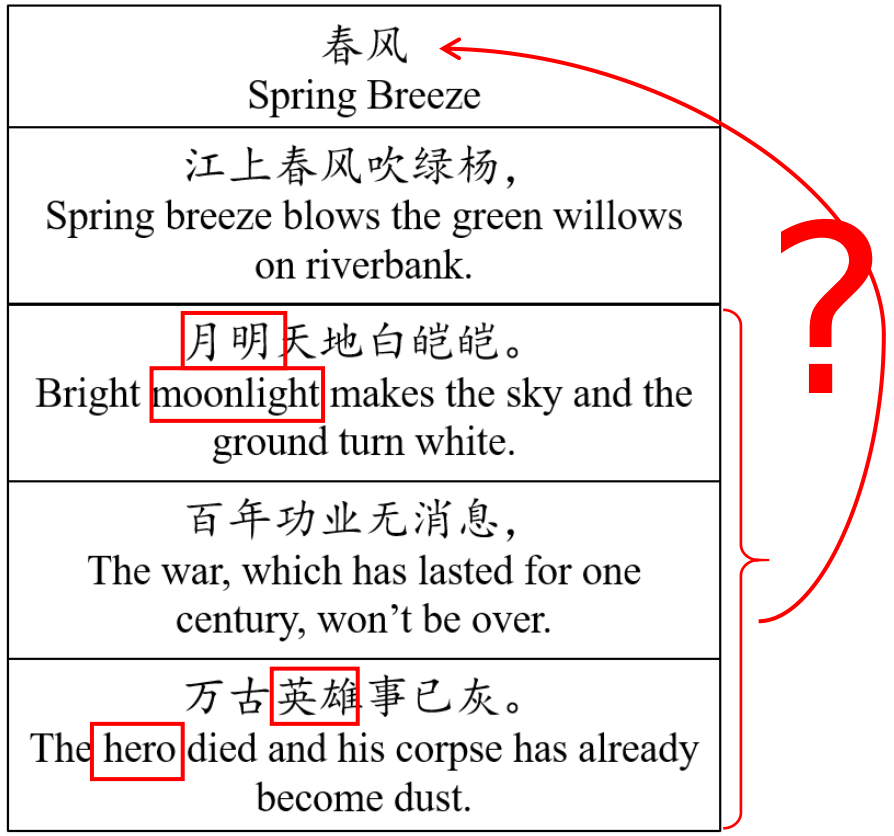
Models for Quality Improvement

Improve Topic Relevance

闺怨 Sorrow of a Young Bride in Her Boudoir 王昌龄
闺中少妇不知愁。 The young bride in her boudoir does not know what grieves,
春日凝妆上翠楼。 She mounts the tower, gaily dressed, on a spring day.
忽见陌头杨柳色， Suddenly seeing by roadside green willow leaves,
悔教夫婿觅封侯。 How she regrets her lord seeking fame far away!

Conditional Generation

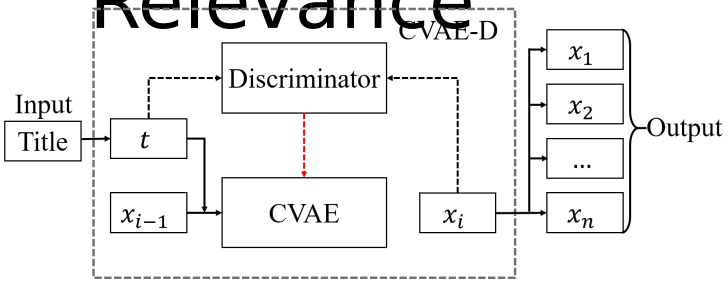
春风 Spring Breeze
江上春风吹绿杨， Spring breeze blows the green willows on riverbank.
月明天地白皑皑。 Bright moonlight makes the sky and the ground turn white.
百年功业无消息， The war, which has lasted for one century, won't be over.
万古英雄事已灰。 The hero died and his corpse has already become dust.



A poem generated by RNNPG (Zhang and Lapata, 2014)

Models for Quality Improvement

Improve Topic Relevance



—CVAE-D (Li et al., 2018)

Condition: $c = (x_{i-1}, t)$ Ground Truth: x_i^*

Discriminator $D(x, c)$: consistent with the title t?

$$L_{CVAE} = -E_{q(z|x_i,\ c)}[\log p(x_i|z,c)] + KL[q(z|x_i,c)||p(z|c)]$$

$$L_D = \log D(x_i, t) + \log(1 - D(x_i^*, t))$$

- Train the generator to minimize $L_{CVAE-D} = L_{CVAE} - L_D$
- Train the discriminator to minimize L_D

Li et al. Generating Classical Chinese Poems via Conditional Variational Autoencoder and Adversarial Training. EMNLP 2018.

Models for Quality Improvement

Improve Diversity and Novelty

Different Input Keywords/Titles

Generator

Diverse and Novel Generated Poems

Poetry: Literary Text

keyword: Desolation (萧条)

萧条风雨夜，
I see desolation on a stormy night.
寂寞夕阳边。
I feel lonely at sunset.
何处堪惆怅，
Where can I place my sadness?
无人问钓船。
No one cares about the fishing ship's course.

keyword: autumn lake (秋水)

山中秋水阔，
In autumn, the lake in the mountains becomes broad.
门外夕阳斜。
Through the door, I see the sunset.
何处堪惆怅，
Where can I place my sadness?
西风起暮鸦。
At dusk, along with the westerly wind, crows start to dance.

Two poems generated by a basic model with two different keywords as input.

Maximum Likelihood Estimation (MLE)

$$\theta^* = \arg\max_{\theta} \sum_{k=1}^{N} \log p_\theta(x^k|w^k)$$

Token-Level Cross Entropy

Tendency for common patterns

(Zhang et al., 2017)

e.g. high-frequency phrases

Models for Quality Improvement

Improve Diversity and Novelty

- Fluency Reward (sentence-level)

$$R_1(x) = \frac{1}{n}\sum_{i=1}^{n} e^{-ma\times(|p_{lm}(x_i)-\mu|-\delta*\sigma,\ 0)}$$

- Coherence Reward (sentence-level)

$$R_2(x) = \frac{1}{n-1}\sum_{i=2}^{n}[\log p_{s2s}(x_i|x_{1:i-1}) - \lambda \log p_l$$

- TF-IDF Reward (token-level)

$$R_3(x) = \frac{1}{n}\sum_{i=1}^{n} F_\varphi(x_i)$$

- Classification Reward (discourse-level)

$$R_4(x) = \sum_{k=1}^{n_{cl}} P_{cl}(k|x) * k$$

—MRL (Yi et al., 2018)

$$L = L_{MLE} - \beta * \sum_{M} E_{x\sim p_\theta(x|w)}[\sum_{4} \alpha_j * R_j(O)]$$

Mem

三十年前事已非，
Thirty years have passed, and everything has changed.
敢言吾道岂无违。
I dare to say that my road is not the same as before.
可怜万里归来晚，
It is a pity to come back late from tens of thousands miles away,
一片青山眼底飞。
and green hills are flying under my eyes.

MRL

老去无心听管弦，
I don't like listening to music anymore when getting old.
一杯浊酒已醺然。
Just a cup of cheap wine makes me drunk.
诗成桦烛灯前夜，
In the light of candles, I write a poem at night,
梦到西窗月满船。
and dream that through the west window, I see the boat is filled with moonlight.

Yi et al. Automatic Poetry Generation with Mutual Reinforcement Learning. EMNLP

Outline

- Background & Overview
- Models for Quality Improvement
- Models for Attribute Control
- Summary

Models for Quality Improvement

Sentiment —SCPG (Chen et al., 2019)

Expressing diverse sentiments !

Sadness of ageing

Happiness of feasting

Holistic sentiment: implicit negative

向晚意不适， At dusk my heart is filled with glooms;	美人卷珠帘， The **beauty** rolls up the curtain and thousands of miles away she stares worriedly at;
驱车登古原。 I drive my cab to ancient tomb.	深坐颦蛾眉。 With lovely eyebrows frowning sadly, she has been seated still for a long while after that.
夕阳无限好， The setting sun seems so sublime;	美人成列抹朱弦， The **beauties** are playing the Chinese lute for celebration in a line;
只是近黄昏。 But it is near its dying time.	劝得嘉宾醉满筵。 The guest around the banquet all get drunk pleasantly after great wine.

negative | implicit negative | neutral | implicit positive | positive

Keyword: beauty
Sentiment: negative

Keyword: beauty
Sentiment: positive

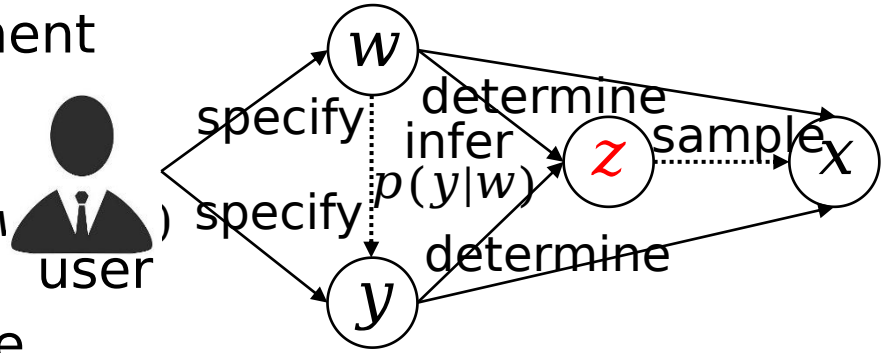
Chen et al. Sentiment-Controllable Chinese Poetry Generation. IJCAI 2019.

Models for Quality Improvement

Sentiment Control —SCPG (Chen et al., 2019)

x: poem w: keyword y: holistic sentiment

z: latent sentiment variable



$$p(x, y, z|w) = p(y|w) * p(z|w, y) * p(x|w, z, y)$$

- Holistic Sentiment Control Module

$$\begin{aligned}\log\, p(x, y|w) &\geq \mathbb{E}_{q(z|x,w,y)}[\log\, p(x|z, w, y)] \\ &\quad - KL[q(z|x, w, y)||p(z|w, y)] + \log\, p(y|w) \\ &= -\mathcal{L}(x, y, w)\end{aligned}$$

$$\begin{aligned}\log p(x|w) &= \iint q(y, z|x, w) \log\, p(x|w) dydz \\ &\geq \mathbb{E}_{q(y|x,w)}[-\mathcal{L}(x, y, w) - \log\, q(y|x, w)] \\ &= -\mathcal{U}(x, w).\end{aligned}$$

$$\mathcal{S}_1 = \mathbb{E}_{p_l(x,w,y)}[\mathcal{L}(x, y, w) - \log\, q(y|x, w)] + \mathbb{E}_{p_u(x,w)}[\mathcal{U}(x, w)]$$

- Temporal Sentiment Control Module

Consider each line x_i and its sent

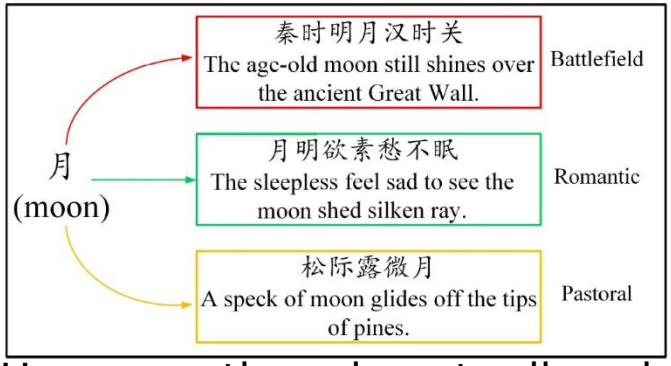
$$\mathcal{S}_2 = \mathbb{E}_{p_l(x,w,y,y_{1:n})} \sum_{i=1}[\mathcal{L}(x_{1:i}, y_{1:i}, w) - \log\, q(y_i|x_{1:i}, w)]$$

$$+ \mathbb{E}_{p_u(x,w)} \sum_{i=1}^{n} \mathcal{U}(x_{1:i}, w).$$

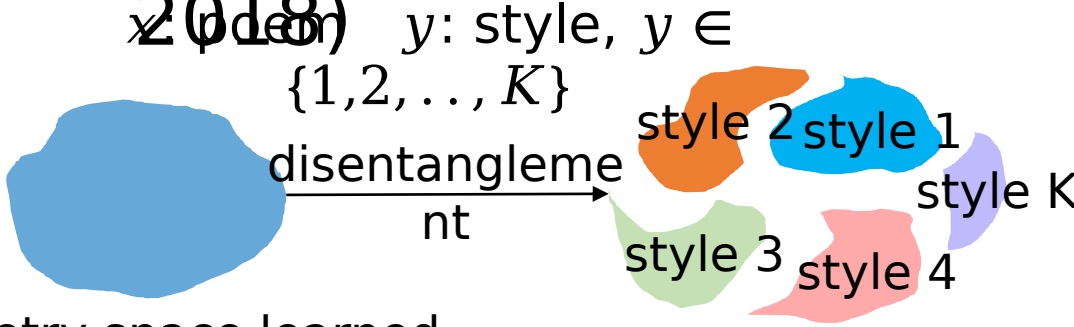
Chen et al. Sentiment-Controllable Chinese Poetry Generation. IJCAI 2019.

Models for Quality Improvement

Unsupervised Style —USPG (Yang et al., 2018)



Human-authored poetry lines in diverse styles under the same keyword.



poetry space learned by a simple model $p(x)$

Stylistic subspaces $p(x|y)$

LSTM decoder

$$s_{i,j} = LSTM(s_{i,j-1}, [e(x_{i,j-1}); a_{i,j-1}])$$

$$p(x_{i,j}|x_{i,1:j-1}, x_{i-1}, y) = softmax(Ws_{i,j})$$

initial decoder state

$$s_{i,0} = [onehot(y);\ h_{|x_{i-1}|}]$$

No Labelled Data!

Yang et al. Stylistic Chinese Poetry Generation via Unsupervised Style

Models for Quality Improvement

Unsupervised Style Control —USPG (Yang et al., 2018)

Use Mutual Information (MI) to build dependency between input y the generated x_i!

maximi

ze:

$$MI(p(x_i|x_{i-1}),\ p(y)) = \frac{1}{K}$$

$$= \int \sum_{i=1}^{K} p(y=k|x_i)\log p(y=k|x_i)\mathrm{d}x_i + \log K$$

variational lower bound

$$\geq \sum_{k=1}^{K} p(y=k)\int \log q(y=k|x_i)\mathrm{d}x_i$$

estimate the integration

$$\approx \frac{1}{K}\sum_{k=1}^{K}\log\{softmax(W\frac{1}{T}\sum_{j=1}^{T} expect(j;k$$

$$= L_{reg}$$

maximize $\log p(x_i|x_{i-1}) + L_{reg}$

Style irrelevant likelihood

style regularization

浊酒一杯聊酩酊，
After a cup of unstrained wine,
I have been a little drunk
白云千里断鸿濛。
I saw the cloud split the sky apart.
马蹄踏破青山路，
On horseback, I pass through every road across the mountain,
惆怅斜阳落日红。
but can only watch the red sun falling down with sorrow.

(a) Style 1: "loneliness, melancholy"

浊酒一杯聊酩酊，
After a cup of unstrained wine,
I have been a little drunk
扁舟何处问渔樵。
With a narrow boat, where could I find the hermits?
行人莫讶归来晚，
Friends, don't be surprised that I come back so late,
万里春风到海潮。
I have seen the great tide and the grand spring breeze.

(b) Style 4: "hermit, rural scenes"

Yang et al. Stylistic Chinese Poetry Generation via Unsupervised Style

Models for Quality Improvement

Semi-Supervised Style Control —MixPoet (Yi et al., 2020)

Style

Mixture of different factors

Historical background
Living experience
Love experience
Gender
School of literary
factors

Style of some poems authored by Li Bai: romantic poems created in the glorious age of Tang dynasty by a male poet who experienced ups and downs of his official career.

M factors $y_1, \ldots, y_i, \ldots, y_M$

each factor y_i is discretized into K_i classes

$\prod_{i=1}^{m} K_i$ factor mixtures → styles

- Annotation on each factor
- Avoid data sparsity

Semi-Supervised CVAE:

$$p(x, y, z|w) = p(y|w) * p(z|w, y) * p(x|w$$

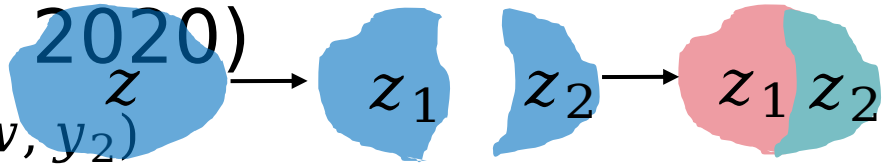
Disentanglement of the latent space rather than the poetry space !

Yi et al. MixPoet: Diverse Poetry Generation via Learning Controllable Mixed

Models for Quality Improvement

Semi-Supervised Style Control —MixPoet (Yi et al., 2020)

$z = [z_1, z_2, ..., z_M]$ e.g., $M = 2$



$p(z|w, y_1, y_2) = p(z_1|w, y_1)p(z_2|w, y_2)$

Universal Approximator (Makhzani et al., 2015)

$q(z|c, \eta) = \delta(z - f(c, \eta)), \eta \sim N(0,1)$

No analytical form of the KL term !

Density Ratio Loss (Rosca et al. 2017)

$$KL[q(z|x, w, y_1, y_2)||p(z_1|w, y_1)p(z_2|w, y_2)]$$

$$\approx E_{q(z|x,w,y_1,y_2)}[\log \frac{C(z, y_1, y_2)}{1 - C([z_1; z_2], y_1, y_2)}]$$

Conditional latent discriminator

Yi et al. MixPoet: Diverse Poetry Generation via Learning Controllable Mixed

Models for Quality Improvement

Rhetoric Control —(Liu et al., 2019)

Poetry: Literary Text

Rhetoric
- Metaphor
- Personification
- Irony
- Pun
- Hyperbole
- ...

A modern Chinese poetry with metaphor and personification

Liu et al. Rhetorically Controlled Encoder-Decoder for Modern Chinese Poetry Generatio

Models for Quality Improvement

Rhetoric Control —(Liu et al., 2019)

CVAE $\log p(x|c) \geq \mathcal{L}(x;c) = E_{q(z|x,c)}[\log p(x|z, c)] - KL[q(z|x,c)||p$

condition c: topic words c and rhetoric label r

Rhetoric Predictor $p(r_i|x_{i-1})$

Simple Seq2Seq Model | Working Memory | Rhetorically Controlled Model

Simple Seq2Seq Model	Working Memory	Rhetorically Controlled Model
不管有多少风雨 (No matter how much wind and rain)	那些岁月里的美好时光 (Good times in those years)	青春有你有我的世界里 (Youth is in your and my world)
我愿意为你 (I'd like to do it for you)	我们都在寻觅 (We're all looking for it)	它像个孩子一样微笑甜蜜 (It smiles like a child)
守护在青春岁月里 (Guard in youth)	你的心已变得陌生 (Your heart has become unfamiliar)	我的故事写在那个岁月里 (My story is written in those years)
愿意为你 (Willing to anything for you)	爱变得不能相聚 (Love becomes impossible to embrace)	静静地睡去 (Sleep quietly)
不要问我为何 (Don't ask me why)	我会在等你 (I will be waiting for you)	但永远被铭记 (But be remembered forever)

Poems generated with topic words {青春 Youth, 爱情 Love, 岁月 Years} by different

Liu et al. Rhetorically Controlled Encoder-Decoder for Modern Chinese Poetry Generatio

Outline

- Background & Overview
- Models for Quality Improvement
- Models for Attribute Control
- Summary

Summary

Poetry Generation

- Quality Improvement
  - Context Coherence
  - Topic Relevance
  - Diversity and Novelty
- Attribute Control
  - Sentiment
  - Style
  - Rhetoric

Methods : Memory Network, Reinforcement Learning, Adversarial Training, Variational Autoencoder, ...

Structures : RNN with GRU/LSTM Cells, CNN, Transformer

Summary

Jiuge(九歌)

A Chinese poetry generation system developed by THUNLP.

Online system
https://jiuge.thunlp.cn/
https://jiuge.thunlp.org/

GitHub
https://github.com/thunlp-aipoet

A paper list for the interdisciplinary field of AI and poetry, including automatic poetry generation, analysis, translation, etc.
https://github.com/THUNLP-AIPoet/PaperList

Textual Adversarial Attack and Defense

THUNLP

Outline

- What is Adversarial Attack?
- Why do we study Adversarial Attack?
- Generation of Adversarial Examples
- Evaluation of Adversarial Attack
- Research Highlights
- Challenges and Future Direction

What is Adversarial Attack?

Adversarial Attack is aimed at generating **adversarial examples** to impair the performance of machine learning (ML) models.

Szegedy, Christian, et al. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199* (2013).
Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *ICL 2015*

What is Adversarial Attack? (cont.)

- Textual Adversarial Attack Example

Original Text Prediction = **Negative**. (Confidence = 78.0%)
*This movie had **terrible** acting, **terrible** plot, and **terrible** choice of actors. (Leslie Nielsen ...come on!!!) the one part I **considered** slightly funny was the battling FBI/CIA agents, but because the audience was mainly **kids** they didn't understand that theme.*
Adversarial Text Prediction = **Positive**. (Confidence = 59.8%)
*This movie had **horrific** acting, **horrific** plot, and **horrifying** choice of actors. (Leslie Nielsen ...come on!!!) the one part I **regarded** slightly funny was the battling FBI/CIA agents, but because the audience was mainly **youngsters** they didn't understand that theme.*

Alzantot, Moustafa, et al. "Generating Natural Language Adversarial Examples." *EMNLP 2018*.

Why do we study Adversarial Attack?

1. Find the **vulnerability** of ML models

   Which kind of adversarial attack is more likely to succeed?

2. Evaluate the **security** of ML models

   Under which scenario we should deploy the model? What is the latent risk?

3. Improve the **interpretability** of ML models

   What patterns do models care or not? What do they try to learn?

4. Develop more **robust** ML models

   Try to fix exposed bugs of ML models using generated attack examples

Generation of Adversarial Examples

- Black-box Setting
  - Only **confidence** and **predicted labels** are accessible
  - Can be implemented efficiently
  - Can be transferred to various victim models
  - Synonym-based word substitution, Irrelevant sentences appendment, Word misspelling, etc.
- White-box Setting
  - Have **full access** to victim models, including **internal parameters** and **gradients**
  - Heavy computational burden (e.g., multiple BPs for a single perturbation)
  - Limited usages for attacking victim models
  - Perturbations are made toward the biggest confidence change w.r.t. gradients

Jia, Robin, and Percy Liang. "Adversarial examples for evaluating reading comprehension systems." *EMNLP 2017*

Ebrahimi, Javid, et al. "Hotflip: White-box adversarial examples for text classification." *ACL 2018*

Evaluation of Adversarial Attack

- Is attack valid?

  Groundtruth label should be the same as original input

- Is attack effective?

  Success Rate, Confidence Reduction Rate, etc.

- Are changes perceivable? Are they readable?

  Edit Distance, BLEU, Human Evaluation, etc.

- What can attack help for models?

  Find clues for debugging models, Data Augmentation, etc.

Ribeiro, Tulio, et al. "Semantically equivalent adversarial rules for debugging NLP models." *ACL 2018*
Michel, Paul, et al. "On Evaluation of Adversarial Perturbations for Sequence-to-Sequence

How can we defend Adversarial Attack?

- Mainstream Method: Adversarial Training (Data Augmentation)
  a. Simply add the attack examples to the training set
  b. Not always effective, may cause the model to **overfit** the **noise** introduced by attack examples
  c. Weak **transferability** across different tasks and models
- Effective alternates
  a. Spelling Check (Especially powerful for misspelling attacks)
  b. Rule-based Attack Recovery (e.g., try to recover the grammatic structure)
  c. Novel Loss (e.g., character-level loss, Trust Region-based loss)

Eger, Steffen, et al. "Text Processing Like Humans Do: Visually Attacking and Shielding NLP Systems." *NAACL-HLT 2019*
Goren, Kurland, et al. "Ranking robustness under adversarial document manipulations." *SIGIR 2018*

Research Highlights (1)

AddSent / AddAny (EMNLP 2017)

1. Perturbation:
   Sentence-level
2. Method:
   Black-box
   Adding irrelevant sentences
3. Downstream Task:
   Reading Comprehension (RC)

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Jia, Robin, and Percy Liang. "Adversarial examples for evaluating reading comprehension systems." *EMNLP 2017*

Research Highlights (2)

Hotflip (ACL 2018)

1. Perturbation:
   Character-level
2. Method:
   White-box
   Gradient-based
3. Downstream Task:
   Classification
   Machine Translation

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism.
57% **World**

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mooP of optimism.
95% **Sci/Tech**

Ebrahimi, Javid, et al. "Hotflip: White-box adversarial examples for text classification." *ACL 2018*

Research Highlights (3)

Continuous-Space Perturbation
(ICLR 2018)

1. Perturbation:
   Embedding Space
2. Method:
   Black-box
   GAN, Stochastic Search
3. Downstream Task:
   Image, Machine Translation,
   Textual Entailment

Classifiers	Sentences	Label
Original	**p** : The man wearing blue jean shorts is grilling. **h** : The man is walking his dog.	Contradiction
Embedding	$\mathbf{h}'$: The man is walking by the dog.	Contradiction → Entailment
LSTM	$\mathbf{h}'$: The person is walking a dog.	Contradiction → Entailment
TreeLSTM	$\mathbf{h}'$: A man is winning a race.	Contradiction → Neutral

Zhao, Zhengli, Dheeru Dua, and Sameer Singh. "Generating natural adversarial examples." *ICLR 2018*

Research Highlights (4)

Using Paraphrase (NAACL-HLT 2019)

1. Perturbation:
   Sentence-level
2. Method:
   Black-box
   Word Swapping, Back-translation
3. Downstream Task:
   Paraphrase Identification

Word Swapping

	Sentence 1	Sentence 2	Generation Type
(1)	Can a **bad** person become **good**?	Can a **good** person become **bad**?	Adjective swap
(2)	**Jerry** looks over **Tom**'s shoulder and gets punched.	**Tom** looks over **Jerry**'s shoulder and gets punched.	Named entity swap
(3)	The team also toured in Australia **in 1953**.	**In 1953,** the team also toured in Australia.	Temporal phrase swap
(4)	Erikson **formed** the rock band Spooner with two fellow musicians.	Erikson **founded** the rock band Spooner with two fellow musicians.	Word replacement

Back-translation

Zhang, Yuan, Jason Baldridge, and Luheng He. "PAWS: Paraphrase Adversaries from Word Scrambling." *NAACL-HLT 2019*

Research Highlights (5)

Using Genetic Algorithm (EMNLP 2018)

1. Perturbation:
   Word-level
2. Method:
   Black-box
   Genetic Algorithm
   Word Substitution
3. Downstream Task:
   Textual Entailment
   Sentiment Analysis

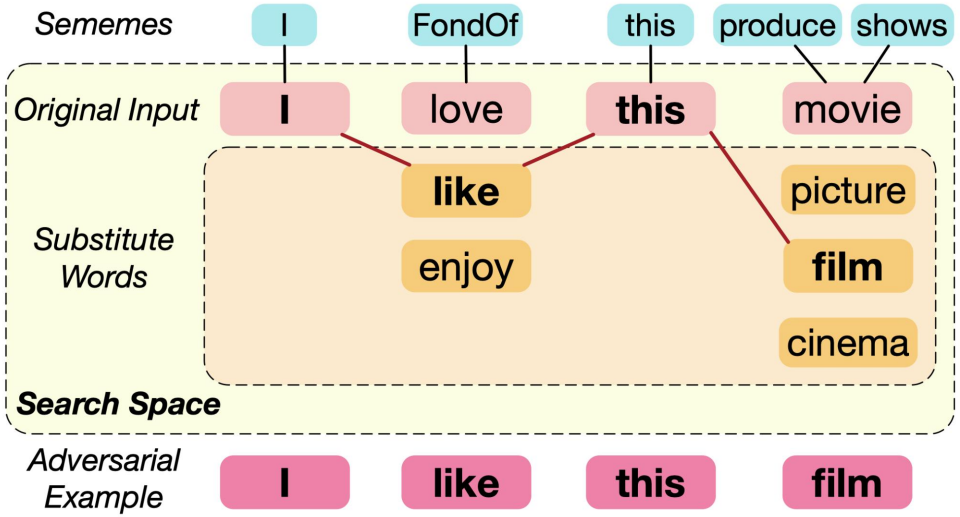
Original Text Prediction: **Entailment** (Confidence = 86%)
Premise: *A runner wearing purple strives for the finish line.* **Hypothesis:** *A **runner** wants to head for the finish line.*
Adversarial Text Prediction: **Contradiction** (Confidence = 43%)
Premise: *A runner wearing purple strives for the finish line.* **Hypothesis:** *A **racer** wants to head for the finish line.*

Alzantot, Moustafa, et al. "Generating Natural Language Adversarial Examples." *EMNLP 2018*.

Research Highlights (6)

PSO-based Word Substitution (ACL 2020)

1. Perturbation:
   Word-level
2. Method:
   Black-box
   Sememe-based Word Substitution
3. Downstream Task:
   Sentiment Analysis, Natural Language Inference



Challenges and Future Directions

- Challenges
  a. Trade-off between **Diversity** and **Semantics-preserving**
  b. Many generated examples lack **Readability**
  c. Indirect **Interpretability** of target models
- Future Directions
  a. Attacks beyond sentence-level (e.g., paragraph-level)
  b. Design defense models more effective than data augmentation (e.g., novel loss, attack recognition)
  c. Improve readability

Cross-Modal Learning

THUNLP

Cross-Modal Learning

- A modality is the classification of a single independent channel of sensory input/output between a computer and a human
  - Text
  - Images
  - Audio
  - Tactile
  - …

Cross-Modal Learning

- A modality is the classification of a single independent channel of sensory input/output between a computer and a human
  - Text
  - Images
  - Audio
  - Tactile
  - ...
- Cross-modal learning refers to any kind of learning that involves information obtained from more than one modality

Cross-Modal Learning

- A modality is the classification of a single independent channel of sensory input/output between a computer and a human
  - Text
  - Images
  - Audio
  - Tactile
  - ...
- Cross-modal learning refers to any kind of learning that involves information obtained from more than one modality

Why Cross-Modal Learning

- Human beings are exposed to multi-modal information every day
  - Integrate information from different modalities
  - Make comprehensive judgments

- Multiple modalities provide complementary information
  - Judgment of a syllable

Cross-Modal Learning Tasks

- Image Captioning
  - Generating a description of an image

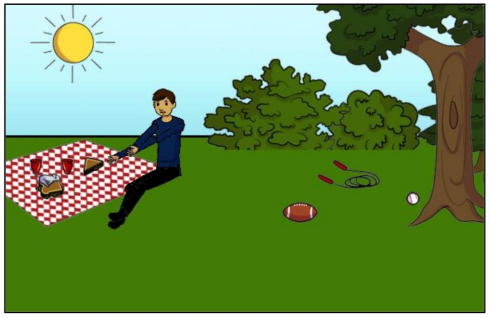
Lin et al. Microsoft COCO: Common Objects in Context. 2014

Cross-Modal Learning Tasks

- Visual Question Answering
  - Input an image and a natural language question
  - Output a natural language answer

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



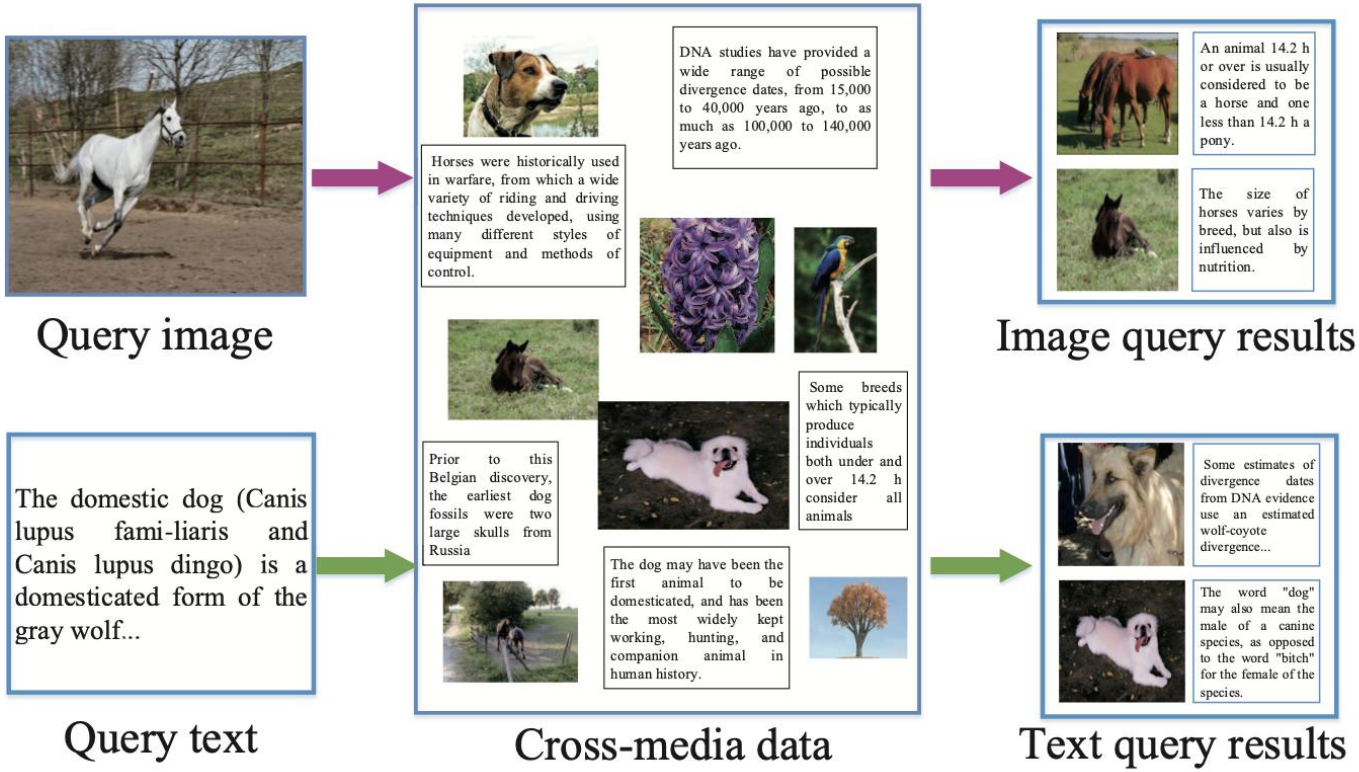
Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Antol et al. Visual Question Answering. ICCV 2015

Cross-Modal Learning Tasks

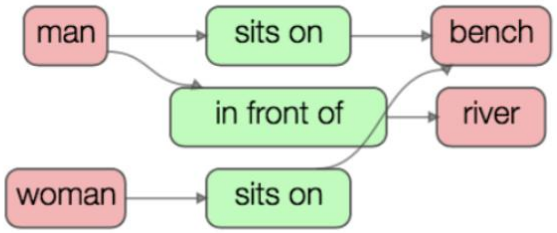
- Cross-Modal Retrieval
  - Retrieve the results with various media types by submitting one query of any media type.



Huang et al. Deep Cross-Media Knowledge Transfer. CVPR 2018

Cross-Modal Learning Tasks

- Scene Graph Generation
  - Generating a graph depicting the objects and their realtions

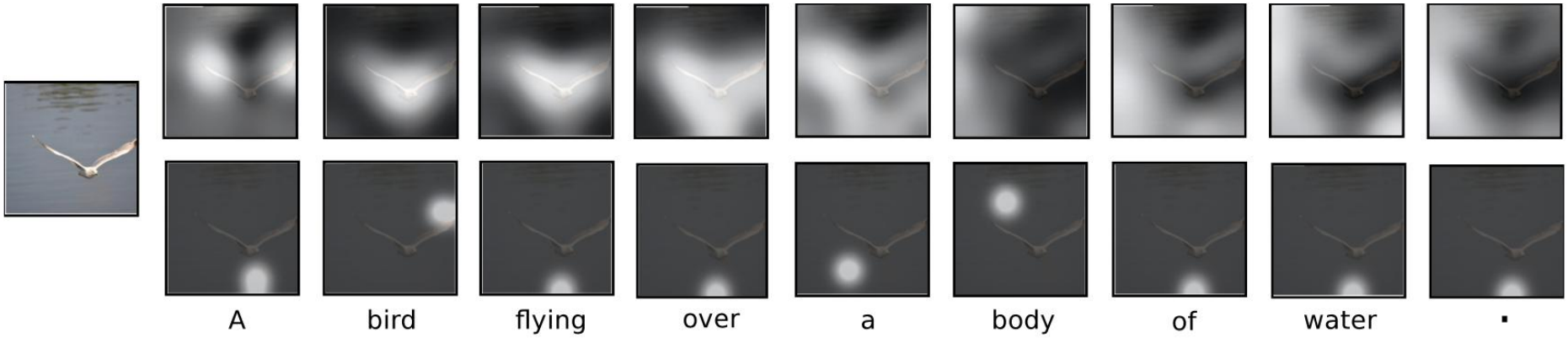


A man and a woman sit on a park bench along a river.

Ranjay et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. 2016

Cross-Modal Learning Paradigms

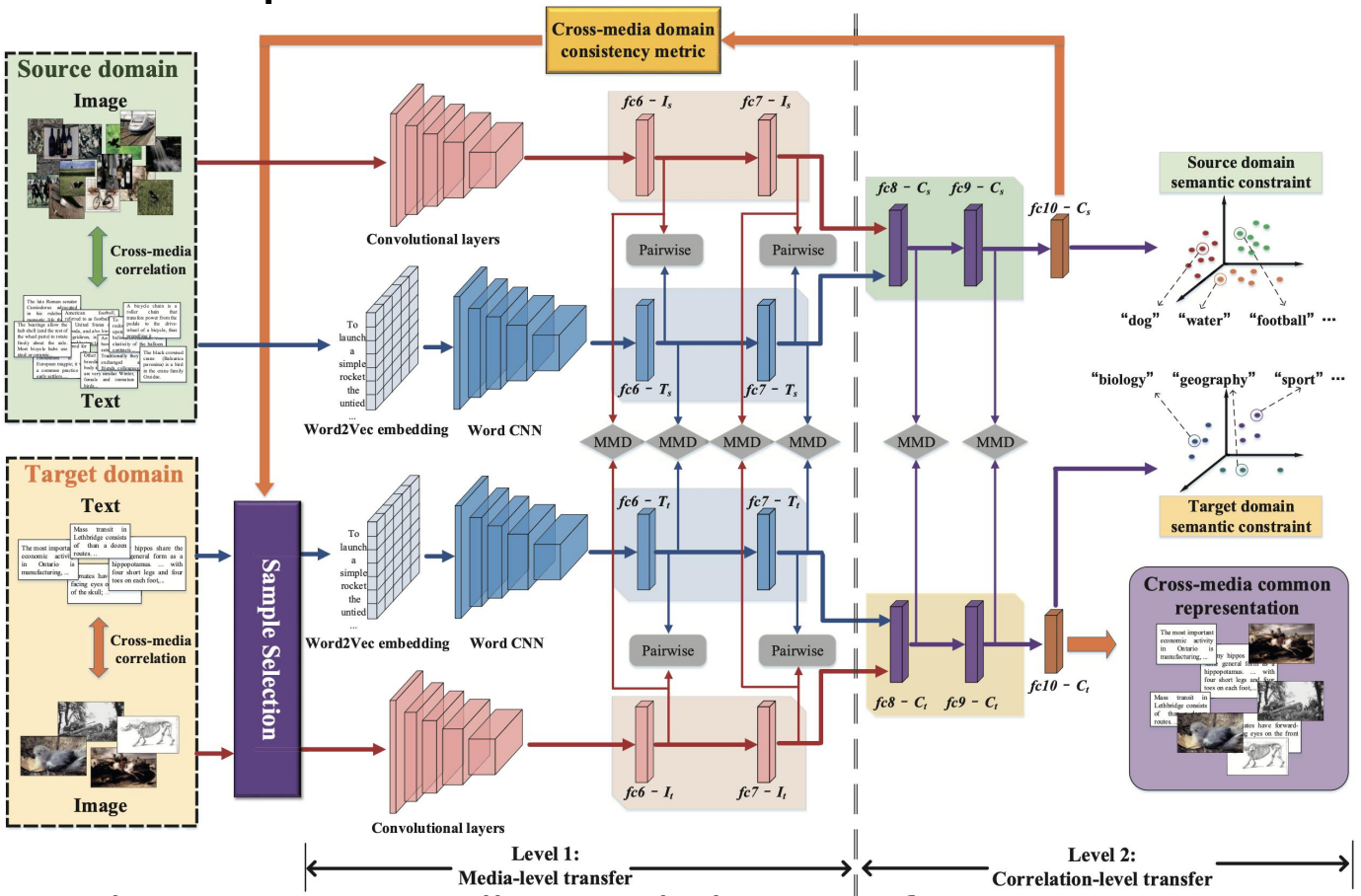
- Cross-Modal Attention
  - Capture interactions between word symbols and visual regions
  - Perform cross-modal information fusion



Xu et al. Show, Attend and Tell. 2015

Cross-Modal Learning Paradigms

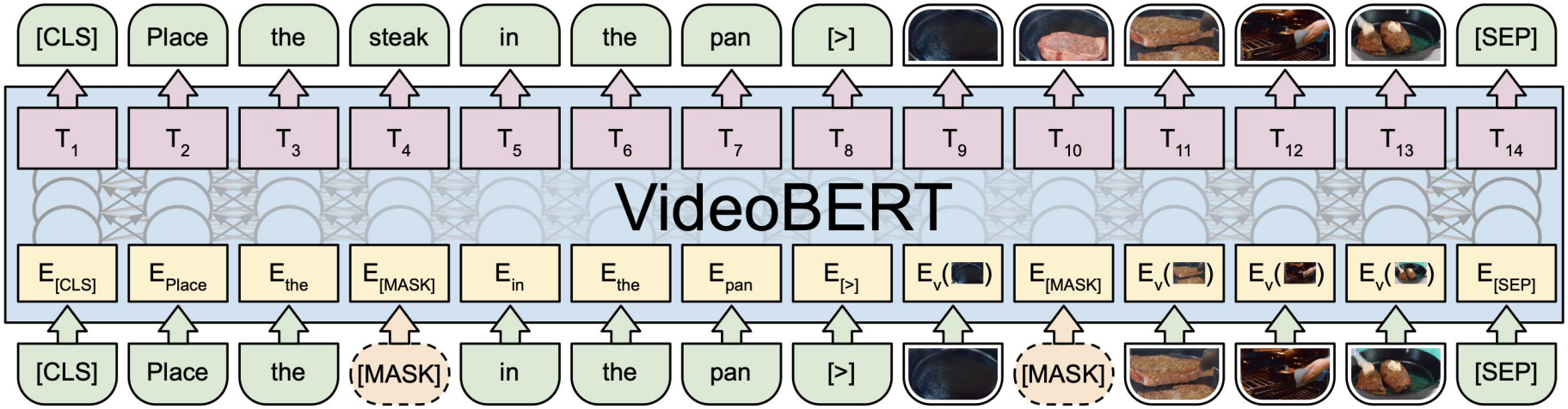
- Cross-Modal Representation Learning
  - Learn representations shared in different



Huang et al. Deep Cross-Media Knowledge Transfer. CVPR 2018

Cross-Modal Learning Trends

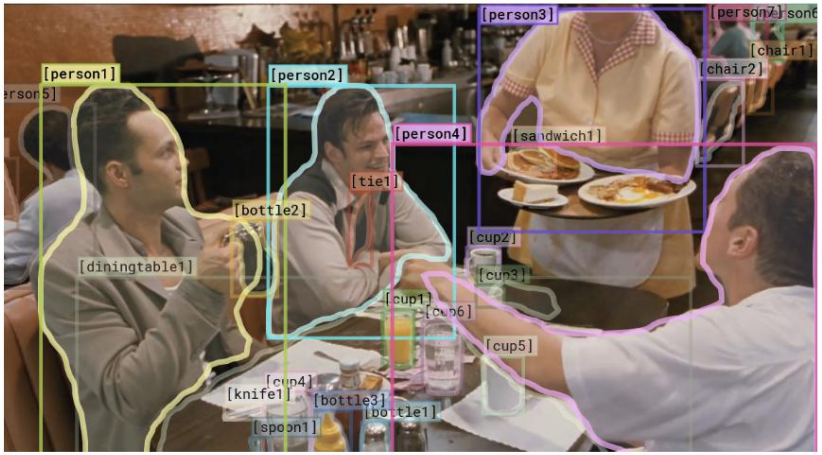
- Cross-Modal Pretraining
  - Learn visual grounding from large-scale unsupervised data
  - Transfer to downstream tasks

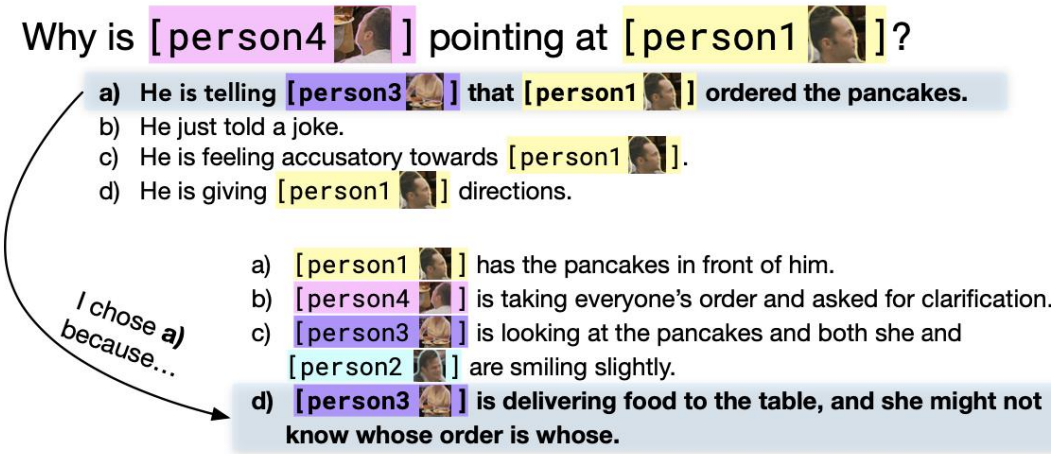


Sun et al. VideoBERT: A Joint Model for Video and Language Representation

Cross-Modal Learning Trends

- Visual Commonsense Reasoning
  - From recognition to cognition
  - Infer the likely intents, goals, and social dynamics of people





Zellers et al. From Recognition to Cognition: Visual Commonsense Reasoning.


TSINGHUA UNIVERSITY
-1911-


Tsinghua University
NLP
Natural Language Processing

THUNLP