Question Answering with Heterogeneous Data

Duyu Tang, Nan Duan
Microsoft Research Asia
2019-07-14

Agenda

- Question Answering over Table, KB and Image
- Pre-training
- Summary

Question Answering over Table, KB and Image

Duyu Tang

Table-based QA (TBQA)

Question 1: Which city hosted the Summer Olympics in 2008?

Answer: Beijing

Question 2: How many nations participate that year?

Answer: 204

Question 3: How about 2004?

Answer: 201

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Pairs	France	24
...	...	...	...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

Knowledge-based QA (KBQA)

- Answer natural language questions based on given knowledge bases

Question 1: Which city hosted the Summer Olympics in 2008?

Answer: Honolulu

Question 2: Who is his wife?

Answer: Michelle Obama

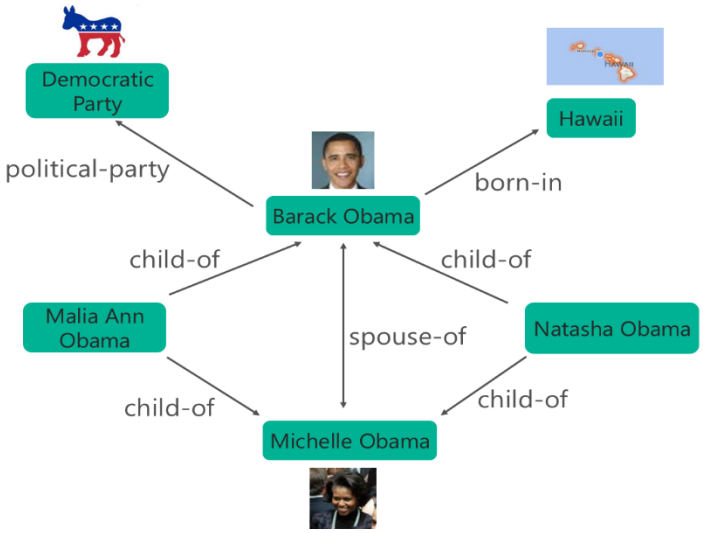


Image courtesy [Yih. 2017]

Image-based QA

- Visual Question Answering
  - Input: NL question + Image
  - Output: NL Answer

Q: What color is the drink?

A: Pink

Q: What is on the tray?

A: Food

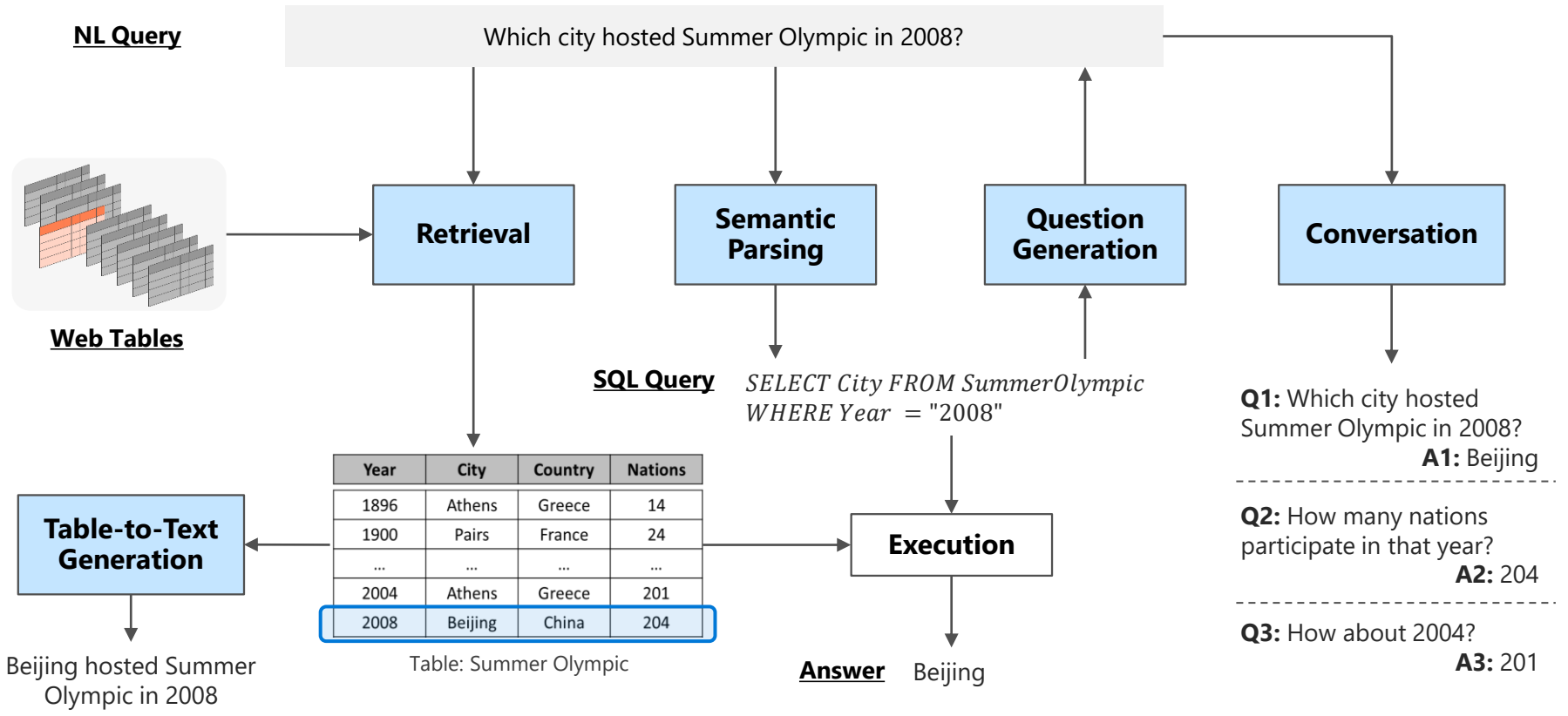
Q: What color is the thing under the food left of the little girl?

A: Red

Outline

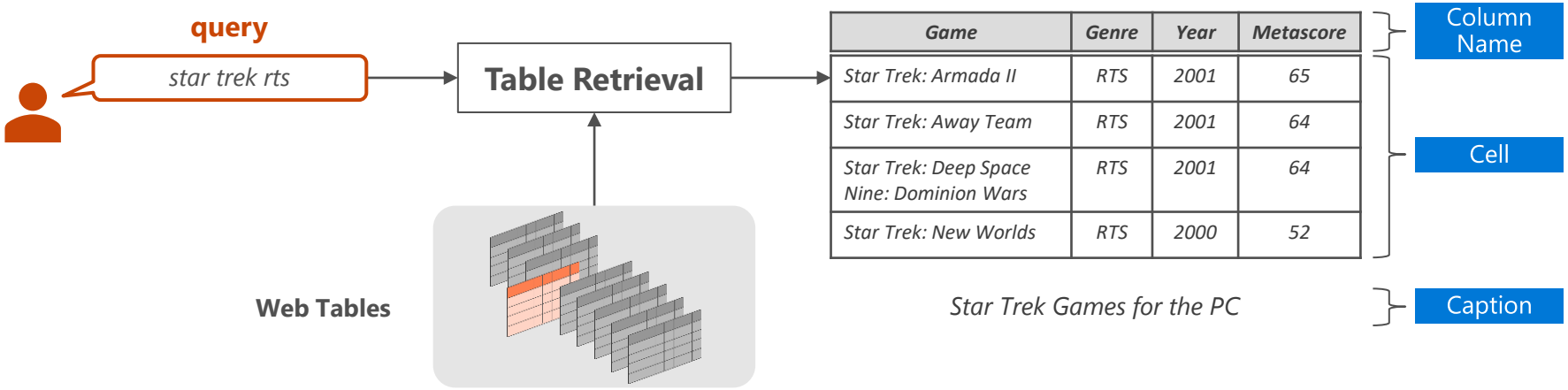
- Table-based QA
  - Retrieval
  - Semantic Parsing
  - Question Generation
  - Conversational Question Answering
  - Table-to-Text Generation
- Conversational KBQA
- Image-based QA

Table Intelligence



Retrieval

- Input
  - A user query q
  - A table collection $T=\{t_1,...,t_n\}$
- Output
  - A table that is most relevant to q



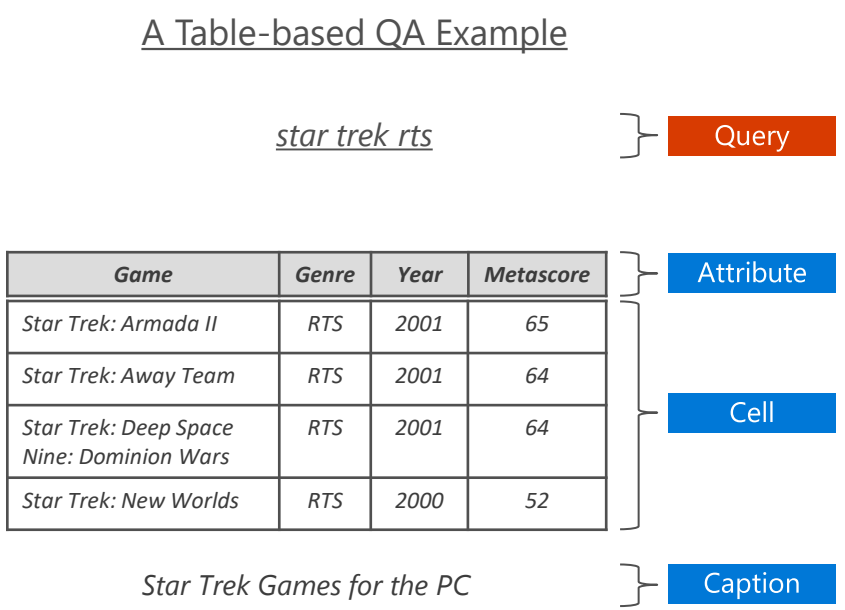
Basic Features

- Literal similarity between a query and a table
  - Query-Attributes
  - Query-Cells
  - Query-Caption

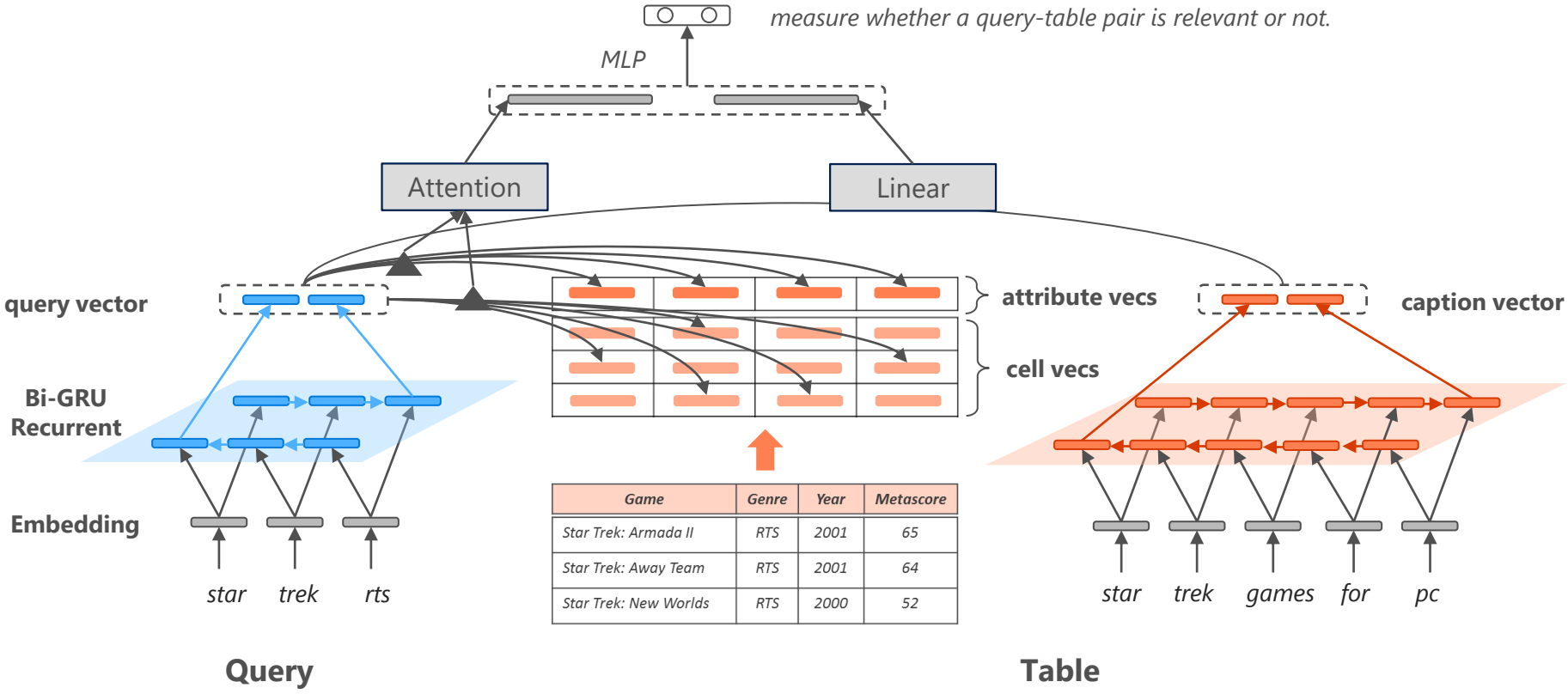
$JaccardSimilarity$ (*star trek rts*, *game genre year metascore*)

$CDSSM$ (*star trek rts*, *game genre year metascore*)

- Static table features
  - NumOfColumns
  - NumOfRows
  - TablePositionInDocument
  - …



Retrieval: Approach



Game	Genre	Year	Metascore
Star Trek: Armada II	RTS	2001	65
Star Trek: Away Team	RTS	2001	64
Star Trek: New Worlds	RTS	2000	52

Retrieval: Evaluation

- WebQueryTable dataset
  - 21,113 questions
  - 273,816 web tables
  - https://github.com/tangduyu/Table-Intelligence

Method	MAP	P@1
BM25	58.23	47.12
Feature	61.02	47.79
NeuralNet	61.94	49.02
Feature + NeuralNet	67.18	54.15

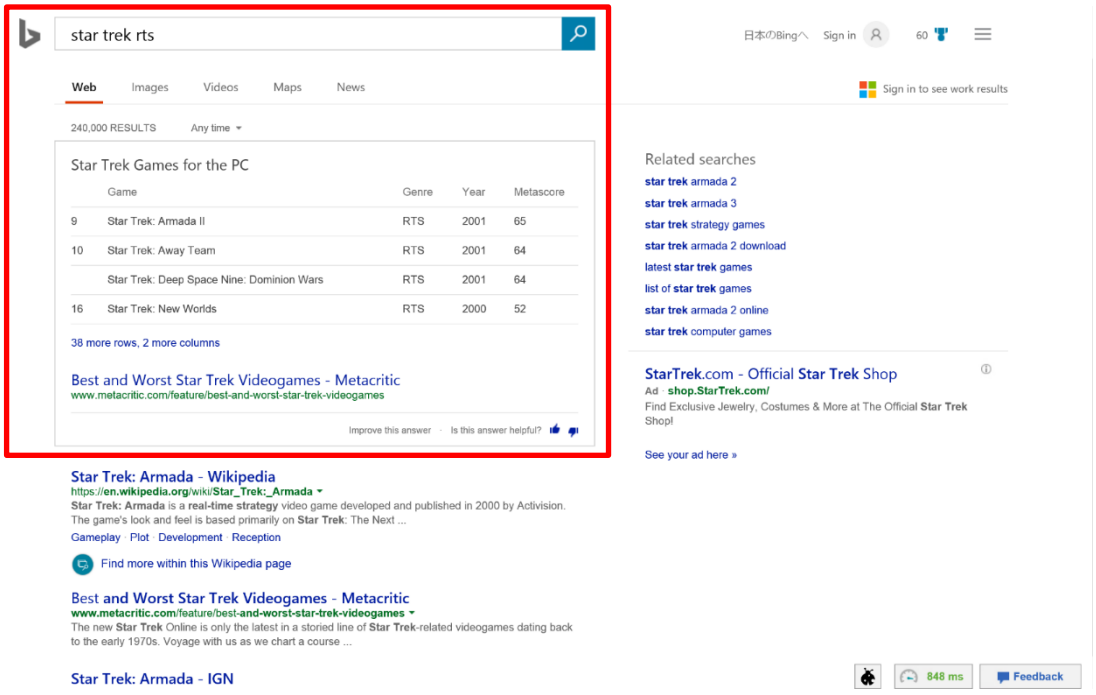
Query: *ramadan dates malaysia 2016*

Retrieved result:

Weekday	Date	Year
Tue	9-Jul	2013
Sun	29-Jun	2014
Thu	18-Jun	2015
Tue	7-Jun	2016

Table. Ramadan begins in Malaysia

Table Retrieval in Bing

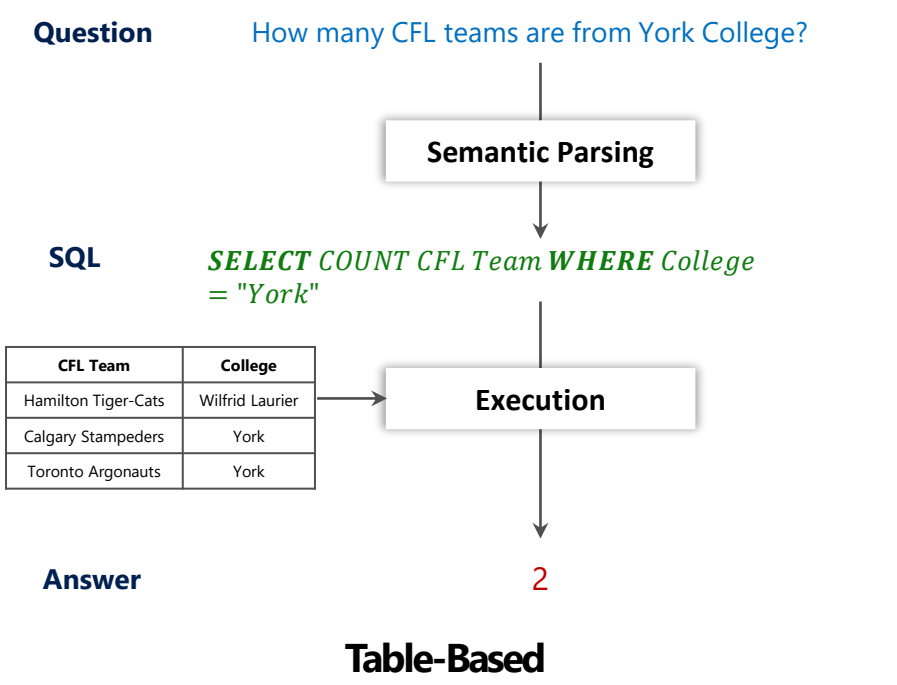


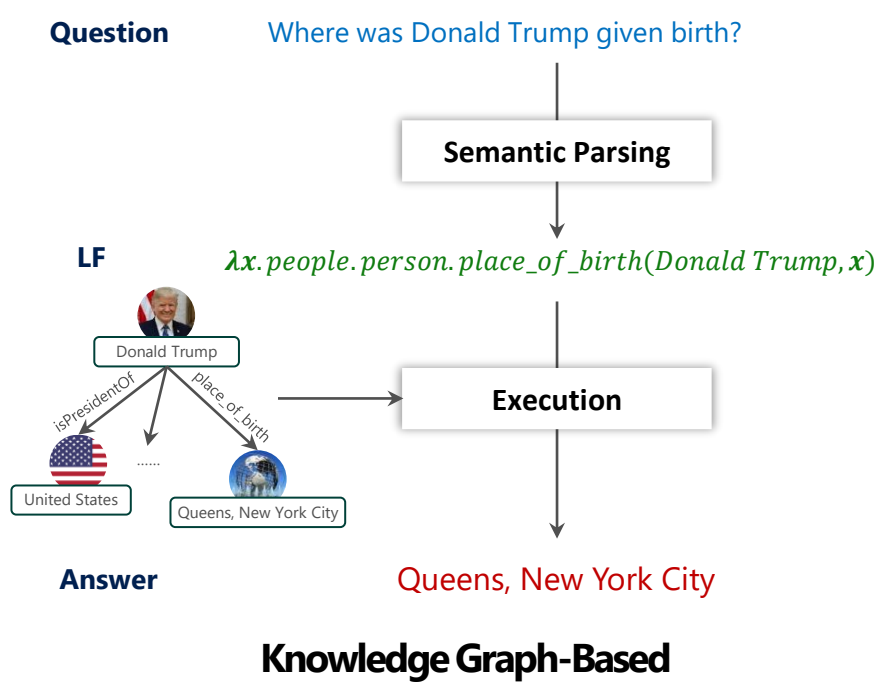
Outline

- Table-based QA
  - Retrieval
  - Semantic Parsing
  - Question Generation
  - Conversational Question Answering
  - Table-to-Text Generation
- Conversational KBQA
- Image-based QA

Semantic Parsing

- Map NL questions into machine executable logical forms based on a knowledge graph/web table





Evaluation on WikiSQL Dataset

- WikiSQL Dataset
  - Zhong, Victor, Caiming Xiong, and Richard Socher. "Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning." *arXiv:1709.00103* (2017).
  - **87,726** human annotated question-SQL pairs distributed across **26,375** tables from Wikipedia

	# of <Q, SQL, T, A> tuples
Train set	61,297
Dev set	9,145
Test set	17,284

Question: How many CFL teams are from York College?

Table:

Pick #	CFL Team	Player	Position	College
27	Hamilton Tiger-Cats	Connor Healy	DB	Wilfrid Laurier
28	Calgary Stampeders	Anthony Forgone	OL	York
29	Toronto Argonauts	Frank Hoffman	DL	York

SQL: $SELECT\ COUNT\ CFL\ Team\ WHERE\ College\ =\ "York"$

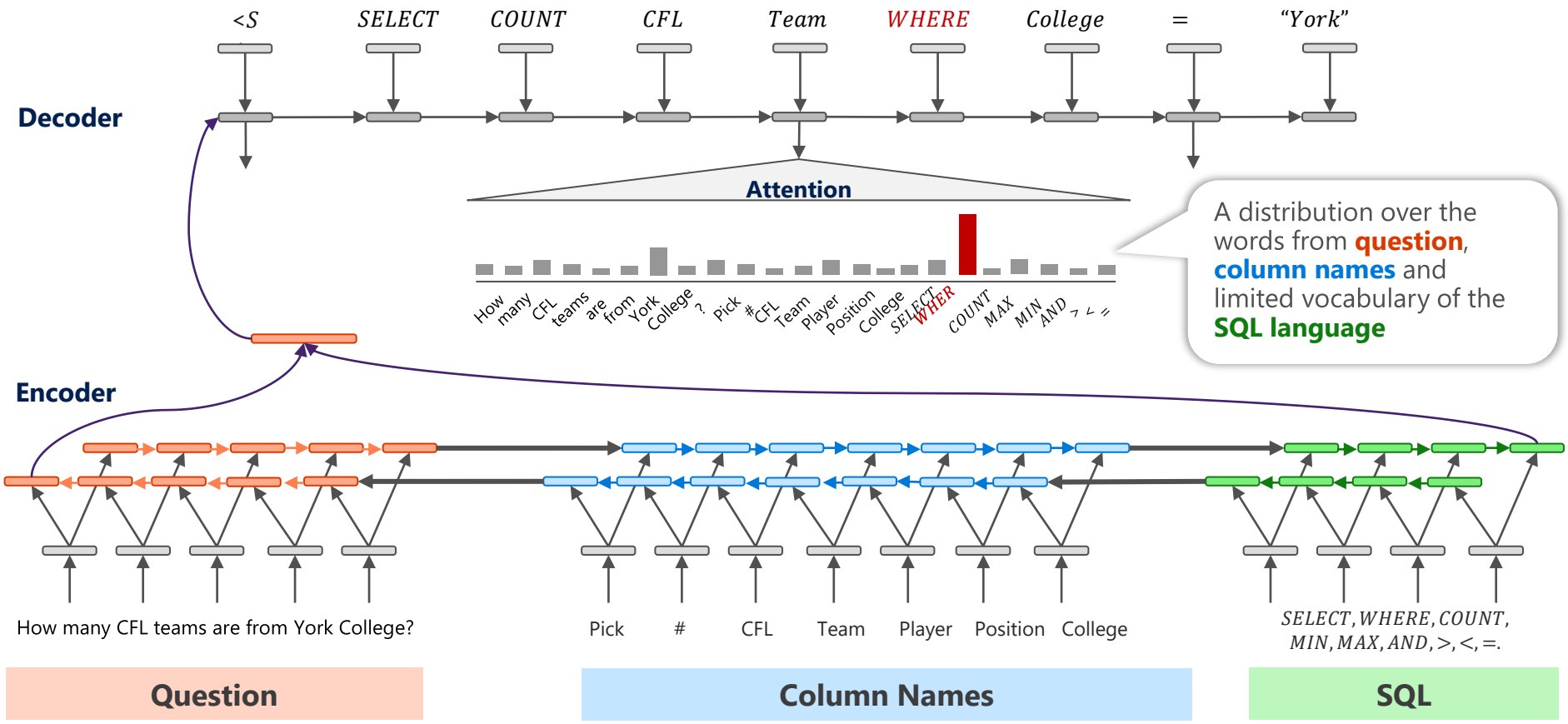
Answer: 2

Two evaluation metrics

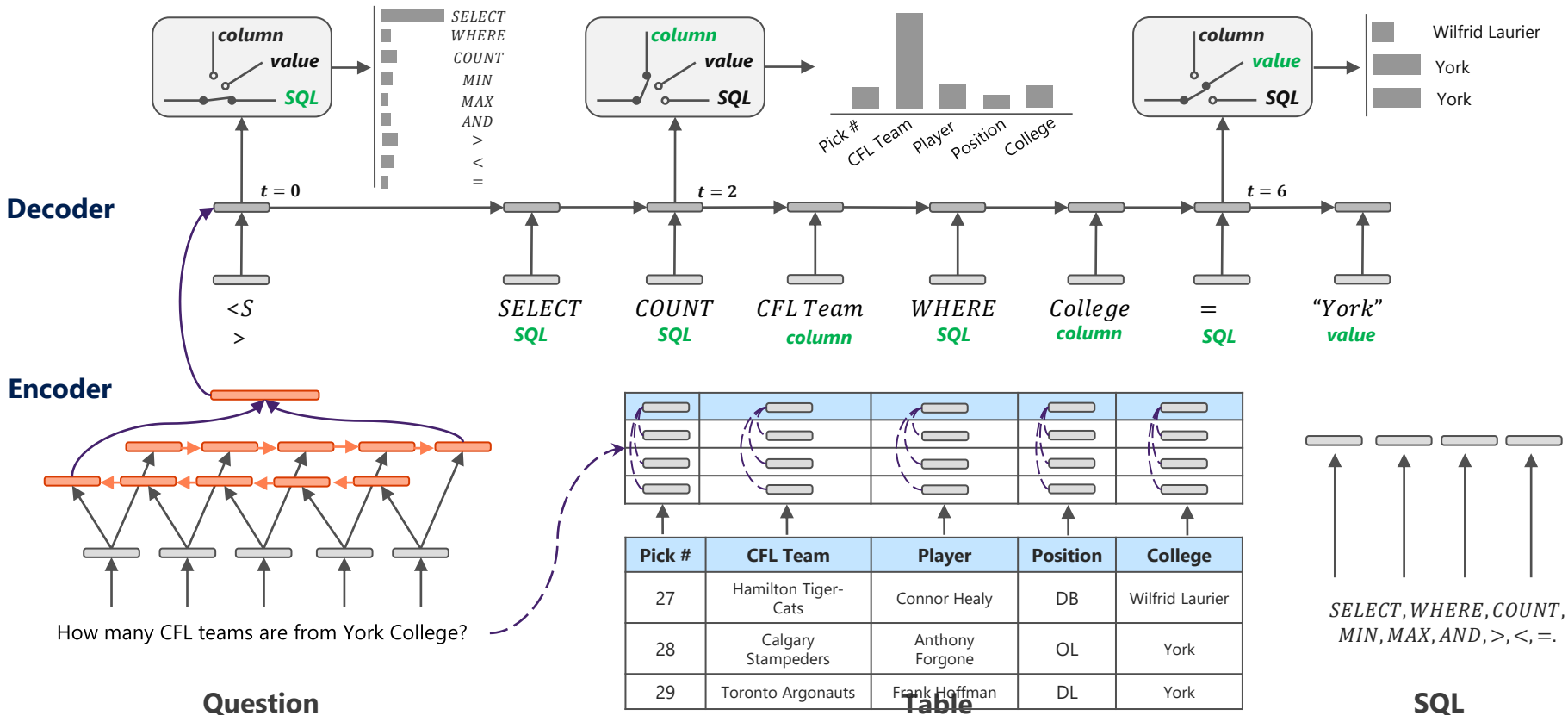
- **SQL Accuracy**
- **Execution Accuracy**

Use SQL sketch	Method	Execution Accuracy	Comments
N	Seq2Seq (Dong 2016)	35.9%	Sequence-to-Sequence
	Seq2SQL (Zhong 2017)	59.4%	Seq2Seq + PointNet + SELECT column and agg
	Wang 2017	66.8%	Seq2Seq + type decoder
	Huang 2018	68.0%	Seq2Seq + type decoder + meta-learning
	MAPO (Liang 2018)	72.6%	Denotation + fine-grained actions + improved RL
	Our End2End approach (Sun 2018)	**74.4%**	**MSRA NLC @ACL-2018**
	MQAN (McCann et al. 2018)	81.4%	Natural Language Decathlon (multi-task)
Y	SQLNet (Xu 2017)	68.0%	Predict WHERE column, then op and value
	Guo 2018	69.0%	SQLNet + charemb + bi-attention
	Our On-going	72.8%	word-token dictionary + iterative back-translation
	Coarse2Fine (Dong 2018)	78.5%	First decode SQL sketch, then tokens
	TypeSQL (Yu 2018)	82.6%	Predict fine-grained input types w/ rule + Freebase
	IncSQL (Shi 2018)	83.7%	Seq2Action + execution-oriented column modeling
	Coarse2Fine + EG Decoding (Wang 2018)	83.8%	Use partially generated output to guide the decoding
	Our SF approach	**85.5%**	**MSRA NLC slot-filling based model**
	IncSQL + EG Decoding (Shi 2018)	87.1%	IncSQL + execution-guided decoding

Seq2SQL with Pointer Network (Zhong et al. 2017)



Model 1: **End2End** Approach



Yibo Sun, Duyu Tang, Nan Duan, Jianshu Ji, Guihong Cao, Xiaocheng Feng, Bing Qin, Ting Liu and Ming Zhou. "*Semantic Parsing with Syntax- and Table-Aware SQL Generation.*" ACL-2018 full paper.

Model output of **End2End** approach

Episode #	Country	City	Martial Art/Style	Masters	Original Airdate
1.1	China	Dengfeng	Kung Fu (Wushu ; Sanda)	Shi De Yang, Shi De Cheng	28-Dec-07
1.2	Philippines	Manila	Kali	Leo T. Gaje Jr. Cristino Vasquez	4-Jan-08
1.3	Japan	Tokyo	Kyokushin Karate	Yuzo Goda, Isamu Fukuda	11-Jan-08
1.4	Mexico	Mexico City	Boxing	Ignacio "Nacho" Beristáin Tiburcio Garcia	18-Jan-08
1.5	Indonesia	Bandung	Pencak Silat	Rita Suwanda Dadang Gunawan	25-Jan-08
1.7	South Korea	Seoul	Hapkido	Kim Nam Je, Bae Sung Book Ju Soong Weo	8-Feb-08
1.8	Brazil	Rio de Janeiro	Brazilian Jiu-Jitsu	Breno Sivak, Renato Barreto Royler Gracie	15-Feb-08
1.9	Israel	Netanya	Krav Maga	Ran Nakash Avivit Oftek Cohen	22-Feb-08

Question #1: how many masters fought using a boxing style ?
Aug.PntNet: select count masters from table where style = boxing
STAMP: select count masters from table where martial art/style = boxing

Question #2: when did the episode featuring a master using brazilian jiu-jitsu air ?
Aug.PntNet: select original airdate from table where masters = brazilian jiu-jitsu
STAMP: select original airdate from table where martial art/style = brazilian jiu-jitsu

Model 2: Slot-Filling Approach

Question How many CFL teams are from York College?

SQL sketch

SELECT $select-aggregator $select-column
WHERE $where-column$_1$ $where-operator$_1$ $where-value$_1$
... ...
AND $where-column$_n$ $where-operator$_n$ $where-value$_n$

Step 1. $where-value

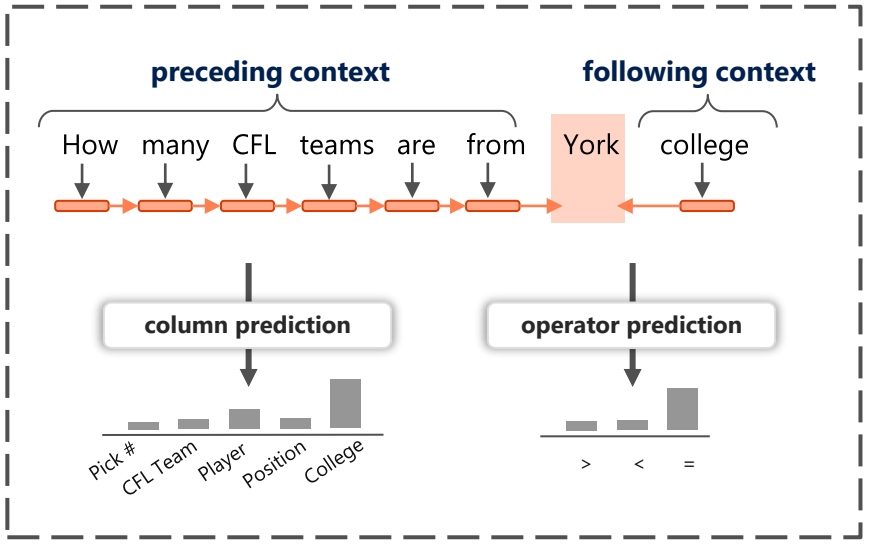
SELECT $select-aggregator $select-column
WHERE $where-column$_1$ $where-operator$_1$ York

Step 2. $where-column/operator

SELECT $select-aggregator $select-column
WHERE College = York

Step 3. $select-column/aggregator

SELECT COUNT *CFL Team*
WHERE College = York

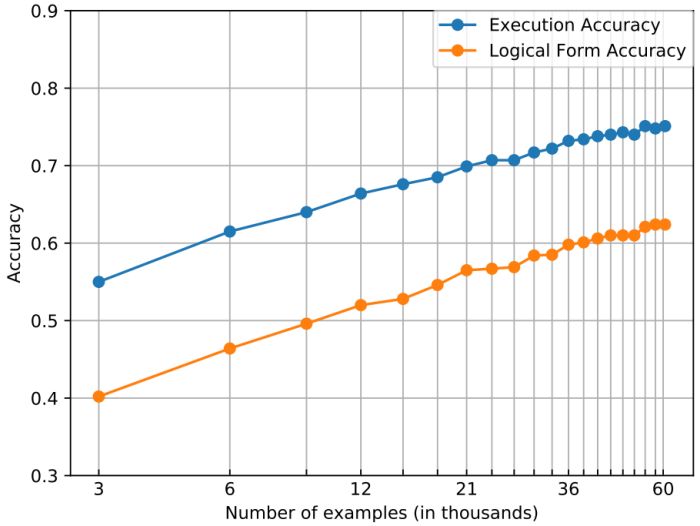


Outline

- Table-based QA
  - Retrieval
  - Semantic Parsing
  - Question Generation
  - Conversational Question Answering
  - Table-to-Text Generation
- Conversational KBQA
- Image-based QA

Question Generation: Motivation

amount of training data ← **Logarithmic relationship** → **semantic parsing accuracy**



The horizontal axis is the training data size in log-scale

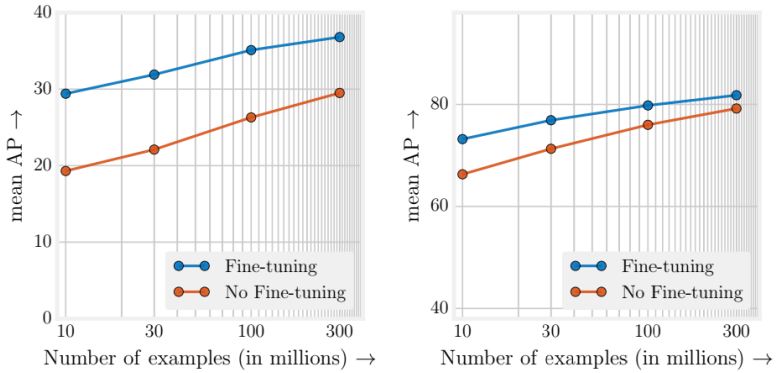


Figure 4. Object detection performance when initial checkpoints are pre-trained on different subsets of JFT-300M from scratch.

Consistent with Google Research & CMU 's observation in ICCV 2017 paper, entitled "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era"

Data Augmentation with Question Generation

Rank	Nation	Gold	Silver	Bronze	Total
1	France	4	1	3	8
2	Great Britain	2	0	1	3
3	Germany	1	1	2	4
4	Belgium	1	1	1	3
5	Japan	0	2	1	3
6	Australia	0	1	1	2
6	United States	0	1	1	2
6	Norway	0	1	1	2
9	Netherlands	0	0	3	3
10	Austria	0	0	1	1
10	Italy	0	0	1	1

Table

SQL Sampler

SQL: SELECT MAX bronze WHERE silver > 2.0

Supervised data

Question Generation

Question:

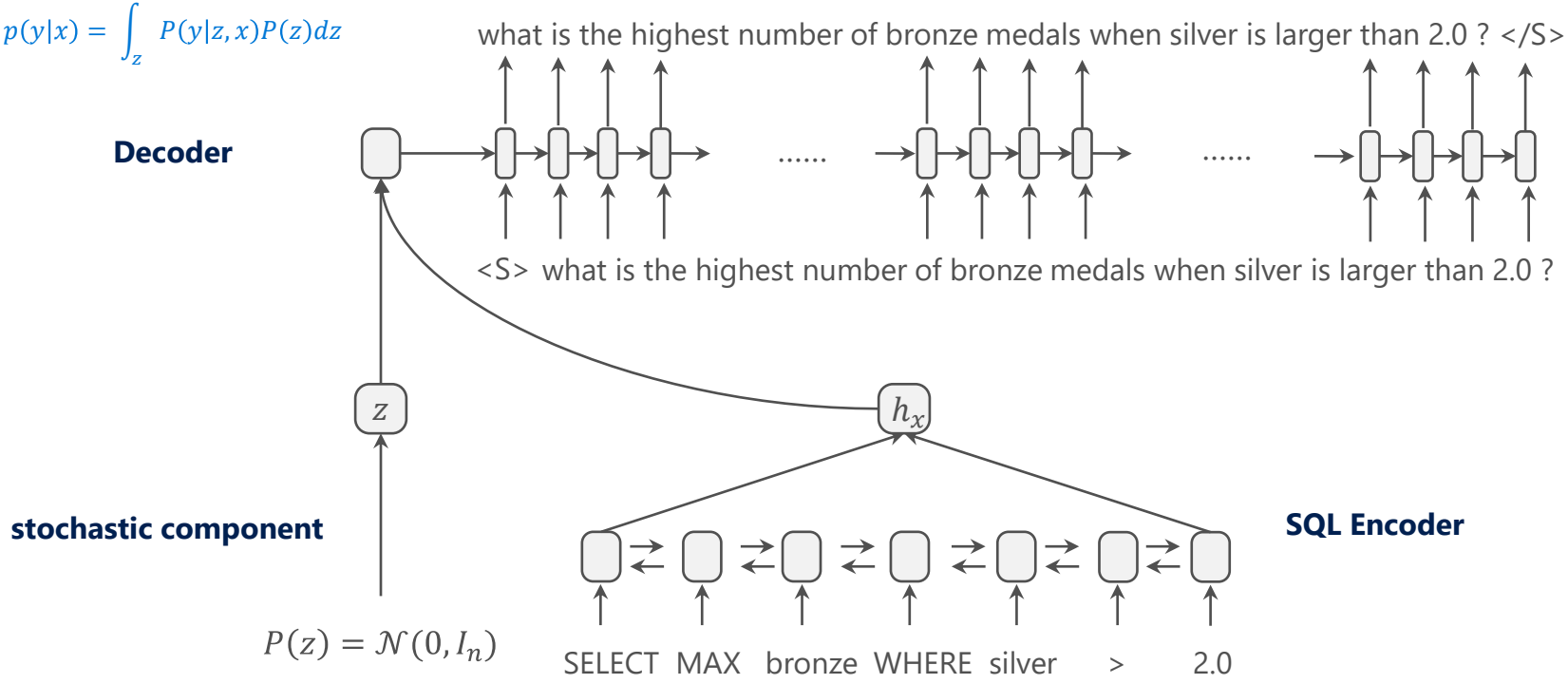
1. what is the highest number of bronze medals when silver is larger than 2.0 ?
2. if the silver is larger than 2.0 , what is the highest amount of bronze medals ?
3. .....

Supervised data + Generated data → Model Training → Semantic parser (seq2SQL model)

Daya Guo, Yibo Sun, Duyu Tang, Nan Duan, Ming Zhou and Jian Yin. "Question Generation from SQL Queries Improves Neural Semantic Parsing." EMNLP-2018 full paper.

Seq2Seq with Latent Variable

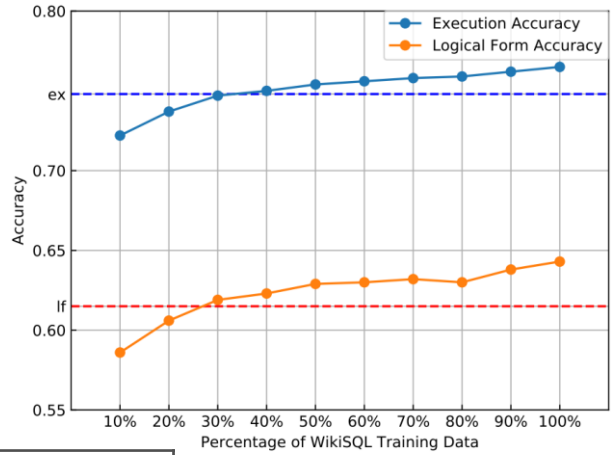
$$p(y|x) = \int_z P(y|z,x)P(z)dz$$



Evaluation on WikiSQL

- QG with **30%** supervised training data reaches parity with **100%** supervised training data

Method	Scale	Execution Accuracy
Seq2Seq (Dong et al., 2016)	100%	35.9%
Seq2SQL (Zhong et al., 2017)	100%	59.4%
Wang et al., 2017	100%	65.1%
STAMP	**100%**	**74.4%**
	30%	68.9%
STAMP + QG	**30%**	**73.9%**
	100%	75.5%



SQL	SELECT COUNT 2nd leg WHERE aggregate = 7-2
Question (ground truth)	what is the total number of 2nd leg where aggregate is 7-2
Question (s2s + cp)	how many 2nd leg with aggregate being 7-2
Question (s2s + cp + lv)	(1) what is the total number of 2nd leg when the aggregate is 7-2 ? (2) how many 2nd leg with aggregate being 7-2 (3) name the number of 2nd leg for 7-2

Outline

- Table-based QA
  - Retrieval
  - Semantic Parsing
  - Question Generation
  - Conversational Question Answering
  - Table-to-Text Generation
- Conversational KBQA
- Image-based QA

Conversation

- **Question1:** Which city hosted the Summer Olympics in 2008?
- **SQL1**: SELECT Character WHERE Year = 2008
- **Answer1:** Beijing

- **Question2**: How many nations participate that year?
- **SQL2**: SELECT Nations WHERE Year = 2008
- **Answer2:** 204

- **Question3**: How about 2004?
- **SQL3**: SELECT Nations WHERE Year = 2004
- **Answer3:** 201

1st turn

2nd turn (coreference)

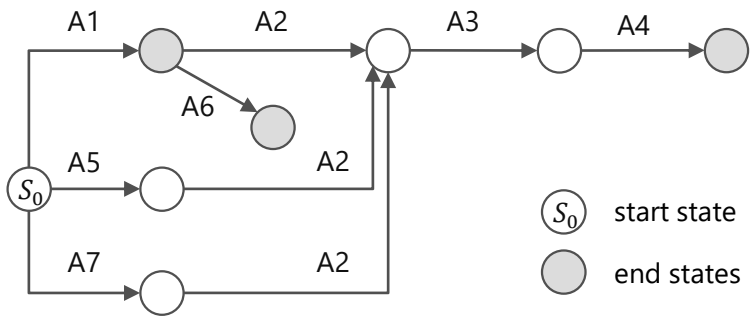
3rd turn (ellipsis)

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Pairs	France	24
…	…	…	…
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

Grammar & Actions

Action	Operation
A1	SELECT
A2	WHERE-Col
A3	WHERE-Op
A4	WHERE-Val
A5	Copy SELECT
A6	Copy WHERE
A7	Copy SELECT+ WHERE

Copying



$S_0 \rightarrow A1$ ⟶ predict SELECT from the current question

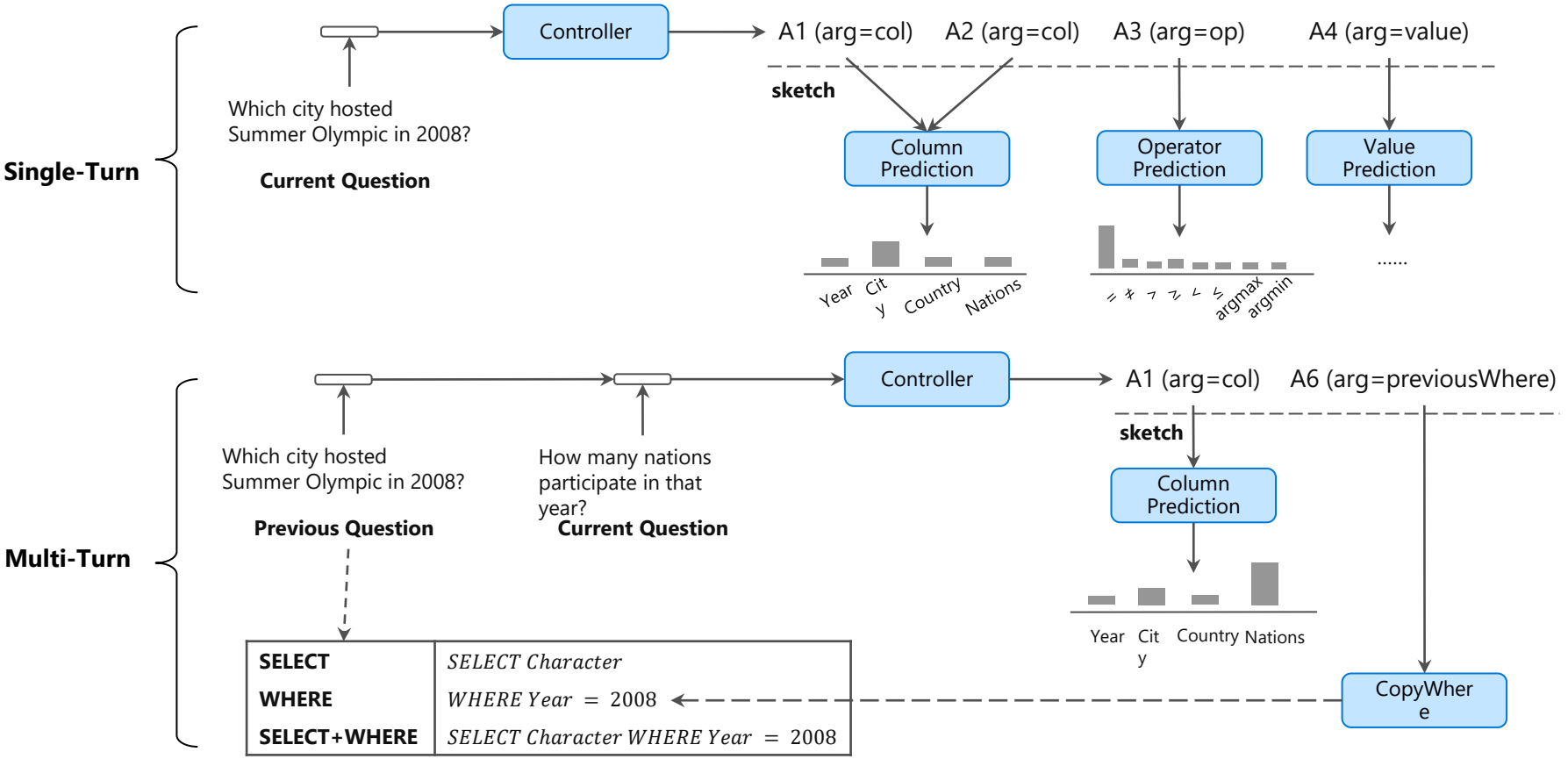
$S_0 \rightarrow A1 \rightarrow A2 \rightarrow A3 \rightarrow A4$ ⟶ predict SELECT+WHERE from the current question

$S_0 \rightarrow A5 \rightarrow A2 \rightarrow A3 \rightarrow A4$ ⟶ use previous SELECT, add new WHERE from the current question

$S_0 \rightarrow A1 \rightarrow A6$ ⟶ predict SELECT from the current question, add previous WHERE

$S_0 \rightarrow A7 \rightarrow A2 \rightarrow A3 \rightarrow A4$ ⟶ use previous SELECT+WHERE, add new WHERE from the current question

NL2Action Approach



Yibo Sun, Duyu Tang, Nan Duan, Jingjing Xu, Xiaocheng Feng, Bing Qin. "*Knowledge-Aware Conversational Semantic Parsing Over Web Tables.*" arXiv:1809.04271.

Evaluation on SequentialQA

- SequentialQA dataset
  - 6,066 sequences, 17,533 questions
  - Released by Iyyer et al. @ACL-2017

Model	All	Seq	Pos 1	Pos 2	Pos 3
FP (Pasupat&Liang 2015)	33.2	7.7	51.4	22.2	22.3
NP (Neelakantan+ 2017)	40.2	11.8	60.0	35.9	**25.5**
DynSP (Iyyer+ 2017)	44.7	12.8	70.4	41.1	23.6
CAMP	45.0	11.7	71.3	42.8	21.9
CAMP + TU	**45.5**	12.0	**71.9**	**43.2**	22.5
CAMP + TU + LM	**45.5**	**13.2**	70.3	42.6	24.8

TU : Table Understanding, LM: Language Modeling

Model	controller	SELECT	WHERE	Operator	Value
CAMP	83.5	82.5	35.0	69.7	21.2
CAMP + TU	83.5	83.4	35.9	69.7	21.2
CAMP + TU + LM	84.8	83.7	36.6	70.2	21.5

Date	Time	TV	Attendance
1-Sep	2:30 PM	BTN	14
8-Sep	3:00 PM	FX	24
…	…	…	…
24-Nov	2:30 PM	ESPN 2	201
1-Dec	7:00 PM	FOX	204
1-Jan	4:10 PM	ESPN	204

Question1: On what dates did the football team play?
SQL1: **SELECT** Date
Action1: A1(Date)

Question2: Of these games, which had an attendance over 90,000?
SQL2: **SELECT** Date **WHERE** Attendance >= 90000.0
Action2: A7 A2(Attendance) A3(>=) A4(90000.0)

Question3: What were the exact attendance numbers for those games?
SQL3: **SELECT** Attendance **WHERE** Attendance >= 90000.0
Action3: A1(Attendance) A6

Question4: Which was the best attendance on the chart?
SQL4: **SELECT** Attendance **WHERE** Attendance **is Max**
Action4: A1(Attendance) A2(Attendance) A3(Max)

Outline

- Table-based QA
  - Retrieval
  - Semantic Parsing
  - Question Generation
  - Conversational Question Answering
  - Table-to-Text Generation
- Conversational KBQA
- Image-based QA

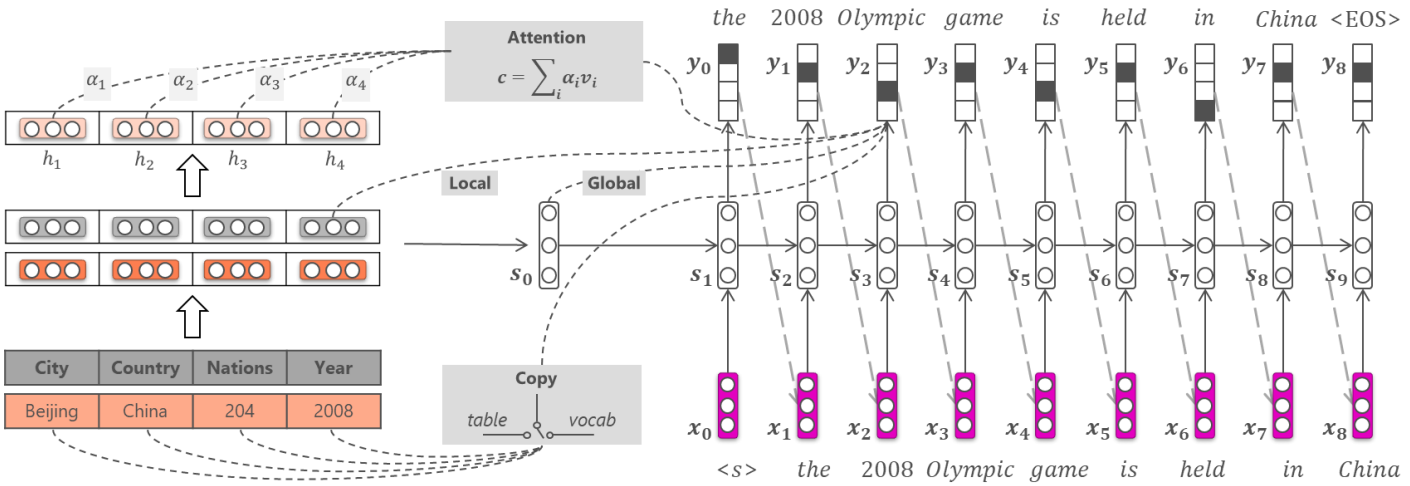
Table-to-Text Generation

- Input: A table and a selected row from it
- Output: A natural language text describing the selected row.

Host cities for Olympic Games

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Pairs	France	24
...	...	...	...
2008	Beijing	China	204
2012	London	UK	204

The 2008 Olympic game is held in Beijing, China.



Junwei Bao, Duyu Tang, Nan Duan, Zhao Yan, Yuanhua Lv, Ming Zhou, Tiejun Zhao. Table-to-Text: Describing Table Region with Natural Language. AAAI-2018 full paper.

Table-to-Text Generation: Evaluation

- WikiTableText dataset
  - 5,000 wiki tables
  - 10,000/1,619/1,619 texts
  - https://github.com/tangduyu/Table-Intelligence

BLEU-4 scores on WikiTableText

Method	Dev	Test
NLM	5.06	5.32
Our model	**35.69**	**37.90**
Our model (w/o caption)	26.21	27.06
Our model (w/o copy)	4.78	5.41
Our model (w/o global)	34.82	36.68
Our model (w/o local)	34.08	36.50

Sophan Sophiaan | Awards and Nominations

Year	Awards	Category	Film	Results
1973	PWI Awards	Runner Up IV Actor (Best Actor/Actress 1972-1973)	Perkawinan	Won
1973	Indonesian Film Festival	Leading Role Actor II	Perkawinan	Won
1990	Indonesian Film Festival	Citra Award for Best Leading Actor	Sesaat Dalam Pelukan	Nominated
1991	Indonesian Film Festival	Citra Award for Best Supporting Actor	Yang Tercinta	Nominated

Model output: Sophan Sophiaan was nominated as the Citra award for best leading actor for Indonesian film festival in 1990 .

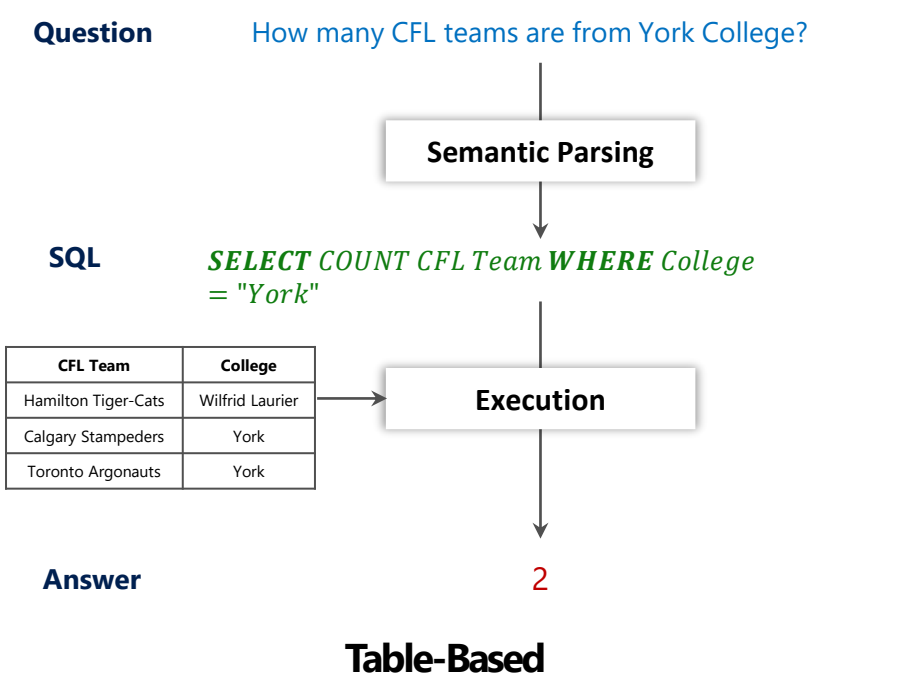
Ground truth: Sophan Sophiaan was nominated to win Indonesian film festival in 1990 .

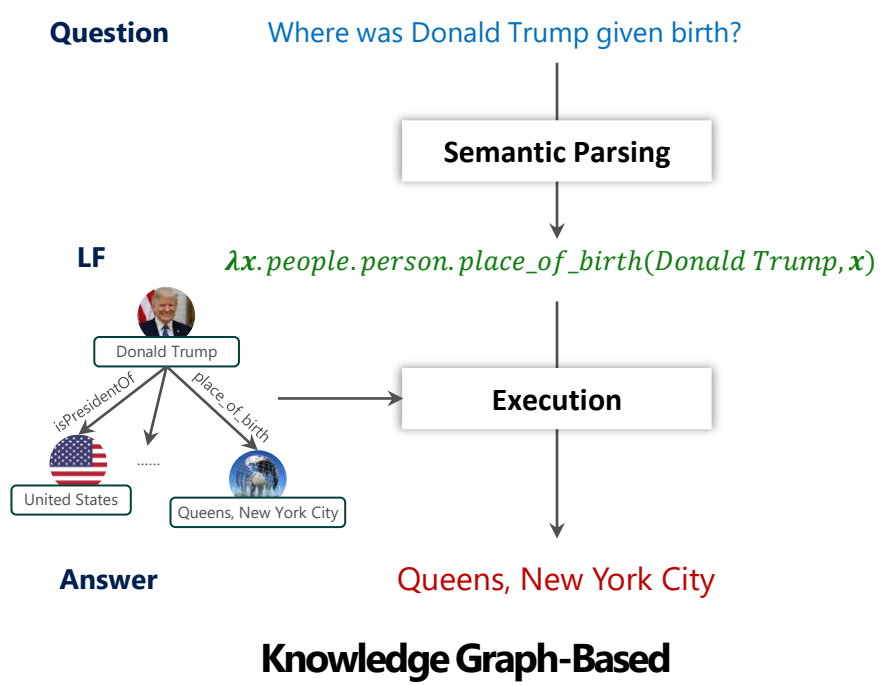
Outline

- Table-based QA
  - Retrieval
  - Semantic Parsing
  - Question Generation
  - Conversational Question Answering
  - Table-to-Text Generation
- Conversational KBQA
  - Dialog-to-Action
  - Coupling Retrieval and Meta-Learning
- Image-based QA

Semantic Parsing

- Map NL questions into machine executable logical forms based on a knowledge graph/web table





Coreference and Ellipsis Phenomena

Question Entity Coreference

- **Q1:** Who is the president of the United States?
- **A1:** Donald Trump
- **Q2:** what is its population?

Answer Entity Coreference

- **Q1:** Who is the president of the United States?
- **A1:** Donald Trump
- **Q2:** How many children does he have?

Question Subsequent Coreference

- **Q1:** Where was the president of the United States born?
- **A1:** New York City
- **Q2:** Where did he graduate from?

Entity Ellipsis

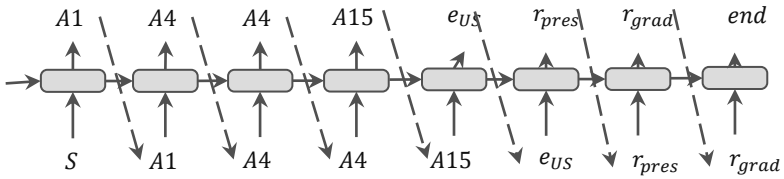
- **Q1:** What movie did Leonardo DiCaprio won an Oscar for?
- **A1:** The Revenant
- **Q2:** who is the director?

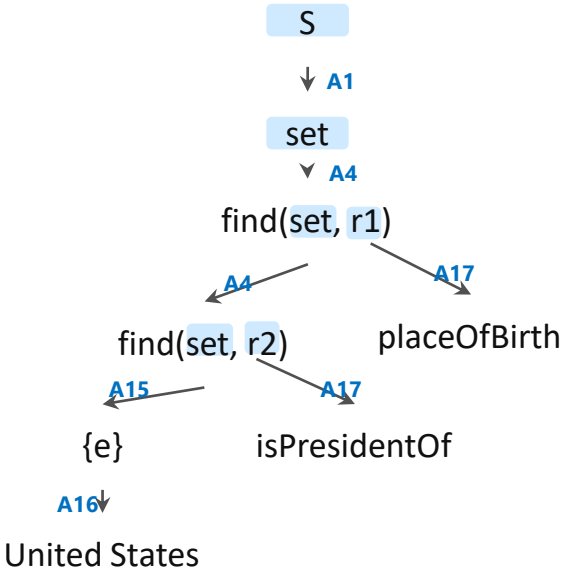
Predicate Ellipsis

- **Q1:** Who is the president of the United States?
- **A1:** Donald Trump
- **Q2:** and also tell me about China?

KBQA with Semantic Parsing (single-turn)

Where was the president of the United States born?





A1: $S \rightarrow set$

A4: $set \rightarrow find(set, r1)$

A4: $set \rightarrow find(set, r2)$

A15: $set \rightarrow \{e\}$

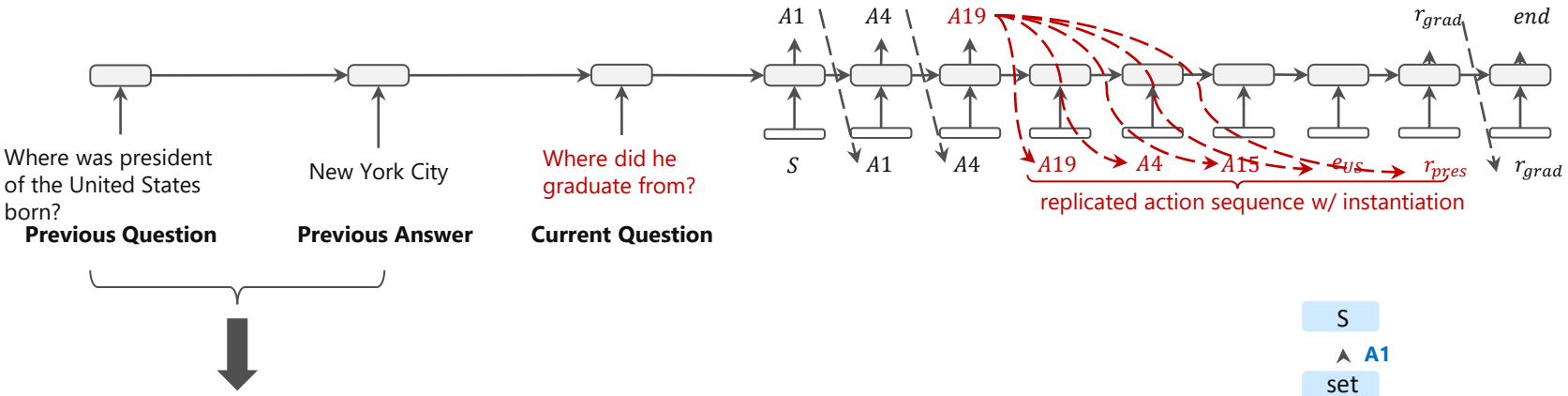
A16: $e \rightarrow United\ States$

A17: $r2 \rightarrow isPresidentOf$

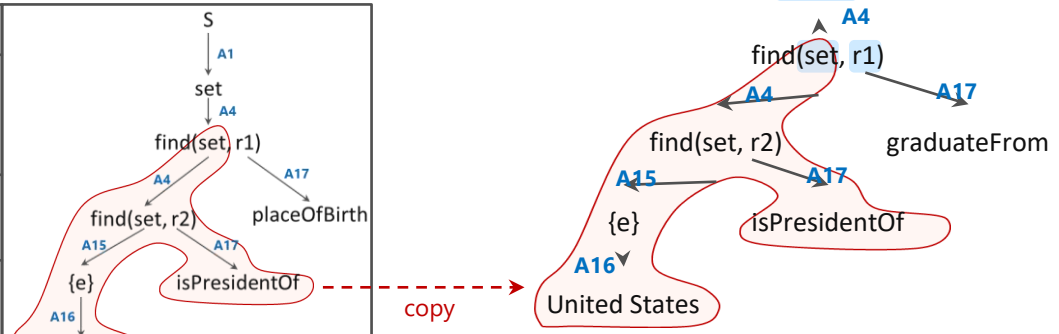
A17: $r1 \rightarrow placeOfBirth$

Action	Operation	Description
A1-A3	S—>Set | Num | Bool	S is start symbol
A4	Set—>Find(R, E)	Set of entities with a r edge to e
A5	Num—>Count(Set)	Total number of set
A6	Bool—>(∈, E, Set)	Whether E∈Set
A7	Set—>Set ∪ Set	Union of Sets
A8	Set—>Set ∩ Set	Intersection of Sets
A9	Set—>Set —Set	Difference of Sets
A10	Set->larger(set, r, num)	Entity from set linking to more than num entities with relation r
A11	Set->less(set, r, num)	Entity from set linking to less than num entities with relation r
A12	Set->equal(set, r, num)	Entity from set linking to num entities with relation r
A13	Set->arxmax(set, r, num)	Entity from set linking to most entities with relation r
A14	Set->argmin(set, r, num)	Entity from set linking to least entities with relation r
A15	Set->{e}	
A16-A18	e | r | num -> constant	instantiation for entity e, predicate r or number num
A19-A21	Set | Num | Bool ->action(i-1)	Replicate previous operation sequence

KBQA with Semantic Parsing (multi-turn)

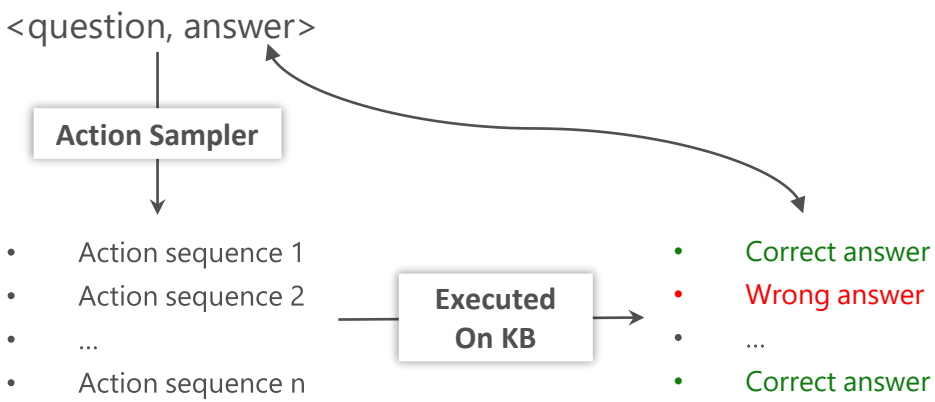


Dialog Memory	
Entity	{United States, tag=utterance} {New York City, tag=answer}
Predicate	{isPresidentOf} {placeOfBirth}
Action Subsequence	$set \rightarrow A4\ A15\ e_{US}\ r_{pres}$ $set \rightarrow A4\ A15$ $set \rightarrow A4\ A4\ A15\ e_{US}$ $r_{pres}\ r_{bth}$ $set \rightarrow A4\ A4\ A15$



Training Data Collection

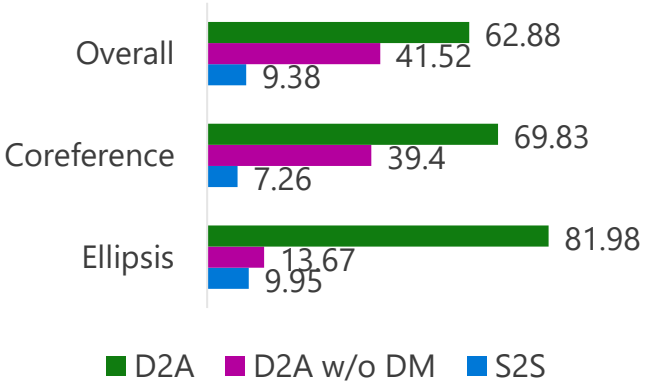
- Input: <question, answer> pairs and a Knowledge Graph (e.g. Freebase/Sartori)
- Output: <question, LFs, answer>



Evaluation on CSQA Dataset (IBM Research, 2018)

CSQA Dataset Statistics

Dialogs	200,000
Turns	1.6M
Entities in KB	12.8M
Unique relations	330
KB Tuples	21.2M
Entity Types	642



id	question type	current question + previous turn	predicted logical form
1	Simple Question (Direct)	**Q1**: N/A **R1**: N/A **Q2**: Who was the dad of Jorgen Ottesen Brahe?	$find(\{Jorgen\ Ottesen\ Brahe\}, father)$
2	Simple Question (Coreferenced)	**Q1**: Who was the dad of Jorgen Ottesen Brahe? **R1**: Otte Brahe **Q2**: Who is the spouse of that one?	$find(\{Otte\ Brehe\}, spouse)$
3	Simple Question (Ellipsis)	**Q1**: What is the profession of Mkihail Beliaiev? **R1**: Military personnel **Q2**: And also tell me about Brett MacLean	$find(\{Brett\ MacLean\}, occupation)$
4	Logical Reasoning (All)	**Q1**: N/A **R1**: N/A **Q2**: Which administrative territories have diplomatic relations with Italy and are not Derikha present in?	$and(\ diff(\ find(\{Italy\}, reverse(diplomatic\ relation)),\ find(\{Derikha\}, country),\ find(\{administrative\ territories\}, isA)\ ))$
5	Quantitative Reasoning	**Q1**: N/A **R1**: N/A **Q2**: Which works did min number of people do the dubbing for?	$argmin(find(\{voice\ actor\}, isa), reverse(work)\ )$
6	Comparative Reasoning	**Q1**: N/A **R1**: N/A **Q2**: Which musical instruments are played by more number of people than electronic keyboard?	$larger(find(\{musical\ instruments\}, isA), reverse(instrument),\ count(and(find(\{electronic\ keyboard\}, reverse(instrument)),\ find(\{people\}, isA))\ ))$
7	Verification (Boolean)	**Q1**: N/A **R1**: N/A **Q2**: Is Arizona Coyotes present in United States of America?	$in(Arizona\ Coyotes\ ,\ find(\{United\ States\ of\ America\}, reverse(country)\ ))$
8	Quantitative Reasoning (Count)	**Q1**: How many people have birthplace at Provence? **R1**: 15 **Q2**: And how about Peterborough?	$copy(\ count(\ find(\{Peterborough\}, reverse(place\ of\ birth)\ )))$
9	Comparative Reasoning (Count)	**Q1**: How many musical instruments are played by greater number of people than Body percussion ? **R1**: 30 **Q2**: And also tell me about timpani?	$copy(count(\ larger(\ find(\{musical\ instrument\}, isA),\ reverse(instrument)\ \ ,\ count(\ find(\ \{timpani\}\ , reverse(instrument))\ )\ )\ ))$

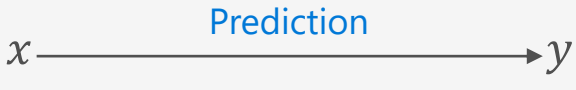
Daya Guo, Duyu Tang, Nan Duan, Ming Zhou and Jian Yin. "*Dialog-to-Action: Conversational Question Answering Over a Large-Scale Knowledge Base.*" NeurIPS-2018 full paper.

Outline

- Table-based QA
  - Retrieval
  - Semantic Parsing
  - Question Generation
  - Conversational Question Answering
  - Table-to-Text Generation
- Conversational KBQA
  - Dialog-to-Action
  - Coupling Retrieval and Meta-Learning
- Image-based QA

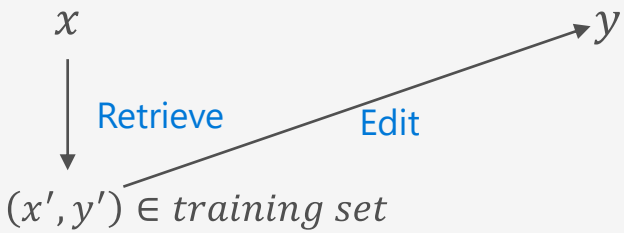
Retrieval-based Semantic Parsing

Standard



➡ $p_{model}(y|x)$

Retrieve-and-Edit



➡ $p_{retrieve}((x', y')|x)$

➡ $p_{edit}(y|x, (x', y'))$

Retrieve-and-Edit

- Retrieve-and-edit

$$p_{model}(y|x) = \sum_{(x',y')} p_{edit}\,(y|x,(x',y'))p_{ret}((x',y')|x)$$

maximize $\mathcal{L}(p_{edit}, p_{ret}) = E[log p_{model}(y|x)]$

- Task-dependent similarity: two inputs x and x′ should be considered similar only if the editor has a high likelihood of editing y′ into y.

See the proof from [Hashimoto+ 2018]

1. Train an encoder-decoder to embed x to v then reconstruct y:

$$(\hat{\theta}, \hat{\phi}) := \arg\max_{\theta,\phi} \mathbb{E}_{(x,y)\sim p_{\text{data}}}[\mathbb{E}_{v|x\sim p_\theta}[\log p_\phi(y \mid v)]].$$

2. Set the retriever to be a nearest neighbor index:

$$\hat{p_{\text{ret}}}(x', y' \mid x) := \mathbf{1}[(x', y') = \arg\min_{(x',y')\in\mathcal{D}} \|\mu_{\hat{\theta}}(x) - \mu_{\hat{\theta}}(x')\|_2^2].$$

3. Train the editor

$$\arg\max_{p_{\text{edit}}} \mathbb{E}_{(x,y)\sim p_{\text{data}}}[\mathbb{E}_{(x',y')\sim \hat{p_{\text{ret}}}}[\log p_{\text{edit}}(y \mid x, (x', y'))]].$$

Retrieval-based Context-Dependent Semantic Parsing

- Context-Dependent Semantic Parsing
  - Conversational QA conditioned on conversational history
  - Code generation conditioned on environment

- Motivation for meta-learning
  - The pattern of a structural output may come from different retrieved examples.

Task I: Conversational Question Answering over KB

- **Q1:** Where was the president of the United States born?
- **A1:** New York City

- **Q2:** Where did he graduate from?

Task II: Code Generation

Class environment:

```
public class SimpleVector implements Serializable {
    double[] vecElements;
    double[] weights;

    NL: Adds a scalar to this vector in place.
    public void add(double arg0)
}
```

NL: Increment this vector in place

(a)

```
public void inc() { this.add(1);}
```

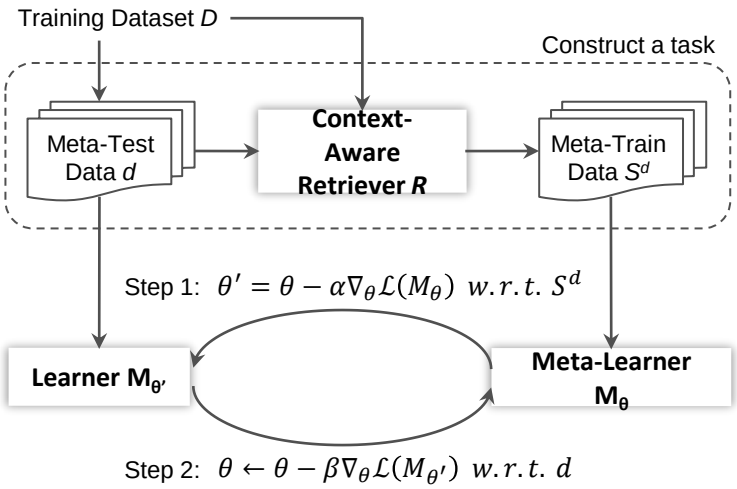
(b)

```
public void inc() {
  for (int i = 0; i< vecElements.length; i++){
      vecElements[i]+=1; } }
```

Figure 1: Code generation based on the class environment and a natural language documentation (NL). (a) shows a example of code generation by applying the class function $add()$, while (b) iterates the $vecElements$ array to increment each element.

Coupling Retrieval and Model-Agnostic Meta-Learning

- Considers retrieved datapoints as a pseudo task for fast adaptation



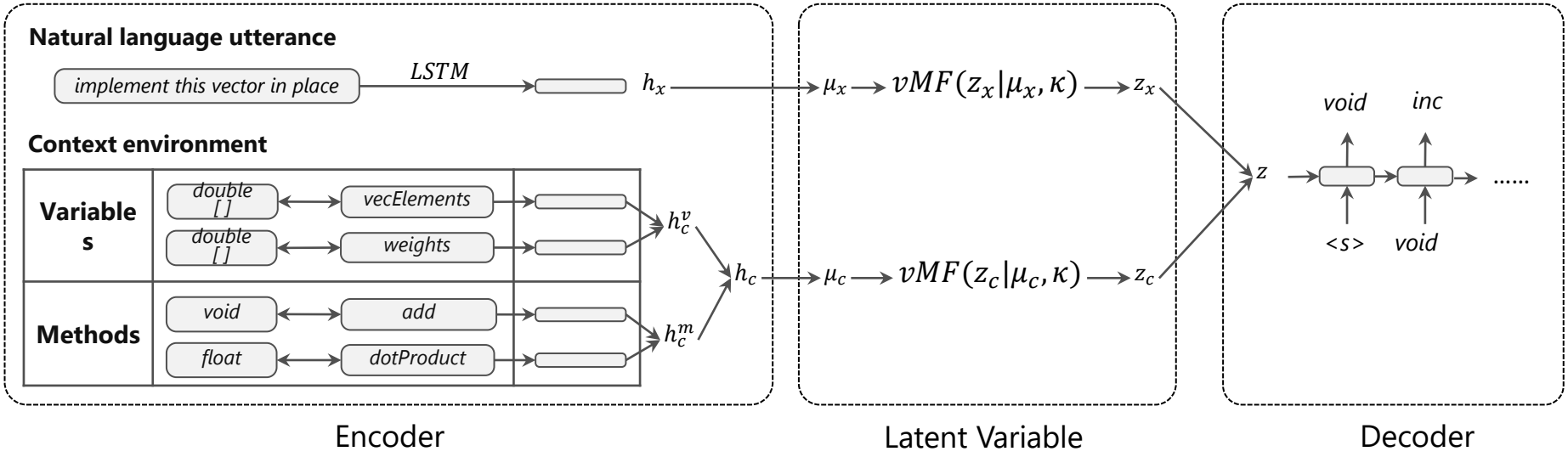
Algorithm 1 Retrieval-MAML

Input: Training dataset $D = (x^{(j)}, c^{(j)}, y^{(j)})$, step size α and β

Output: Meta-learner M

1: Training a context-aware retriever R using D.
2: For each example d, we obtain a support set S^d retrieved by R
3: Randomly initialize θ for M
4: **while** not done **do**
5: Sample a batch of examples D' from D as test examples, and $S' = \bigcup_{d \in D'} S^d$ are viewed as training examples
6: Evaluate $\nabla_\theta\mathcal{L}(M_\theta)$ using S', and compute adapted parameters with gradient descent: $\theta' = \theta - \alpha\nabla_\theta\mathcal{L}(M_\theta)$
7: Update $\theta \leftarrow \theta - \beta\nabla_\theta\mathcal{L}(M_{\theta'})$ using D' for meta-update
8: **end while**

Context-Aware Retrieval Model



Coupling Retrieval and Meta-Learning

- Retrieve datapoints for **each example** to construct a *pseudo* task for fast adaptation

Task I: Conversational Question Answering over KB

Method	Simple Question	Logical Reasoning	Quantitative Reasoning	Comparative Reasoning
Sequence-to-Sequence	13.64	8.33	0.96	2.96
Dialog2Action (MSRA-NLC@NeurIPS2018)	92.01	42.00	45.37	41.51
Dialog2Action + MAML (ours)	**92.66**	**44.34**	**50.30**	**48.13**

Task II: Code Generation

Method	Exact	BLEU
Sequence-to-Sequence	3.20	23.51
Yin+@ACL-2017	6.65	21.29
Iyer+@EMNLP-2018	8.60	22.21
Dialog2Action (MSRA-NLC@NeurIPS2018)	9.15	23.24
Dialog2Action+MAML (ours)	**10.50**	**24.40**

- **New state-of-the-art on both tasks.**

Daya Guo, Duyu Tang, Nan Duan, Ming Zhou and Jian Yin. "*Coupling Retrieval and Meta-Learning for Context-Dependent Semantic Parsing.*" ACL-2019 full paper.

More results on CONCODE

Methods	Dev		Test	
	Exact	BLEU	Exact	BLEU
Retrieval ONLY				
TFIDF	1.25	17.78	1.50	19.73
Context-independent Retrieval	0.85	19.63	0.80	21.98
Context-dependent Retrieval	1.30	21.21	1.00	24.94
Parsing-based methods without retrieved examples				
Seq2Seq	2.90	21.00	3.20	23.51
Seq2Prod (Yin and Neubig, 2017)	5.55	21.00	6.65	21.29
Iyer et al. (2018)	7.05	21.28	8.60	22.11
Seq2Action	**7.75**	**22.47**	**9.15**	**23.34**
Parsing-based methods with retrieved examples				
Seq2Action+Edit vector (Context-independent Retrieval)	6.6	21.27	7.90	22.51
Seq2Action+Edit vector (Context-aware Retrieval)	7.75	20.69	9.20	22.68
Seq2Action+Retrieve-and-edit (Context-independent Retrieval)	5.55	21.27	7.05	22.74
Seq2Action+Retrieve-and-edit (Context-aware Retrieval)	7.55	22.20	9.30	23.95
Seq2Action+MAML (Context-independent Retrieval)	9.15	21.48	9.85	23.22
Seq2Action+MAML (Context-aware Retrieval, w/o finetune)	8.30	21.27	10.30	24.12
Seq2Action+MAML (Context-aware Retrieval)	**8.45**	**21.32**	**10.50**	**24.40**

Retrieve-and-edit

Retrieve-and-MAML

Table 1: Performance of different approaches on the CONCODE dataset.

Model Analysis

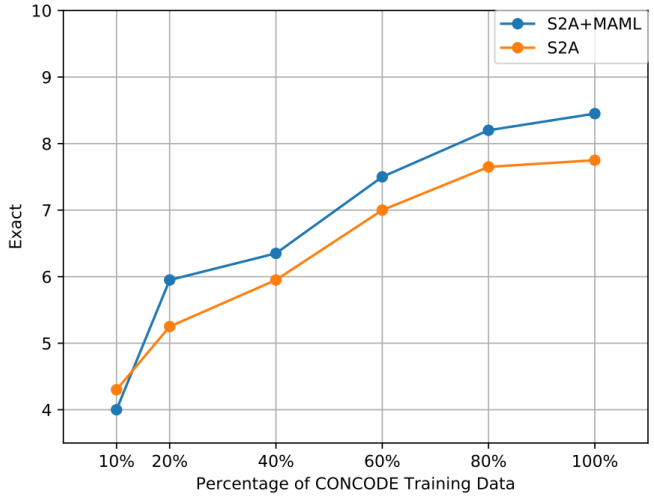


Figure 4: Comparison between S2A and S2A+MAML with different portions of supervised data.

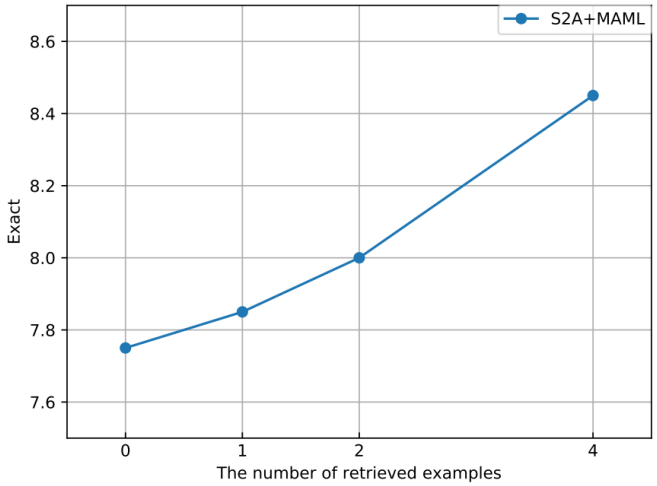


Figure 5: S2A+MAML with different number of retrieved examples on the CONCODE devset.

Retrieved Examples

Input	Context-Aware Retriever	Context-Independent Retriever

Class environment:
HashMap<lalr_item, lalr_item> _all;
NL:
Does the set contain a particular item
Code:
boolean function(lalr_item arg0){
return _all.constainsKey(arg0); }
	Class environment:
	Map<Point, RailwayNode> _nodeMap;
	NL:
	Check if a node at a specific position exits.
	Code:
	boolean function(Point arg0){
	return _nodeMap.constainsKey(arg0);}
		Class environment:
		Node root;
		Node get(Node x, String key, int d);
		NL: Does the set contain the given key
		Code:
		boolean function(String arg0){
		Node loc0==get(root,arg0,0);
		if (loc0==null) return false;
		return loc0.isString; }

Q1: who is the dad of jorgen ottesen brahe? **A1**: otte brahe **Q2**: who is the spouse of that one?	**Q1**: whose child are gio batta bellotti? **A1**: matteo bellotti, paola cresipi guzzo **Q2**: which person is married to that one?	**Q1**: which abstract beings have marge simpson as an offspring? **A1**: clancy bouvier, jacqueline bouvier **Q2**: who is the spouse of that one?

Outline

- Table-based QA
  - Retrieval
  - Semantic Parsing
  - Question Generation
  - Conversational Question Answering
  - Table-to-Text Generation
- KBQA
  - Dialog-to-Action
  - Coupling Retrieval and Meta-Learning
- Image-based QA

Image-based QA

- Visual Question Answering
  - Input: NL question + Image
  - Output: NL Answer

Q: What color is the drink?

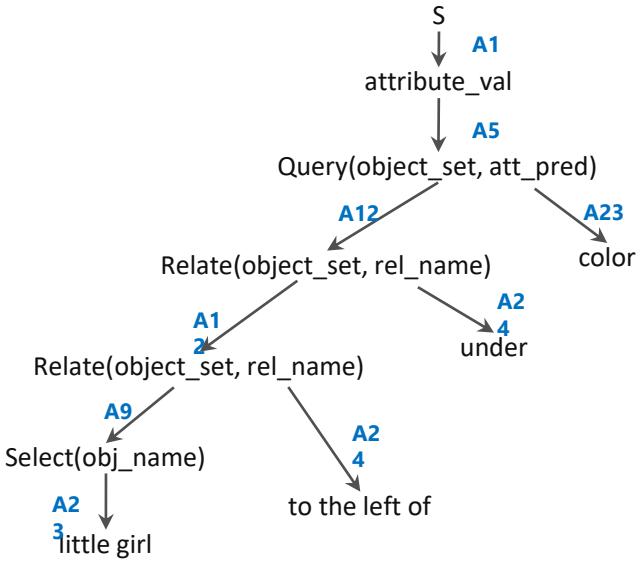
A: Pink

Q: What is on the tray?

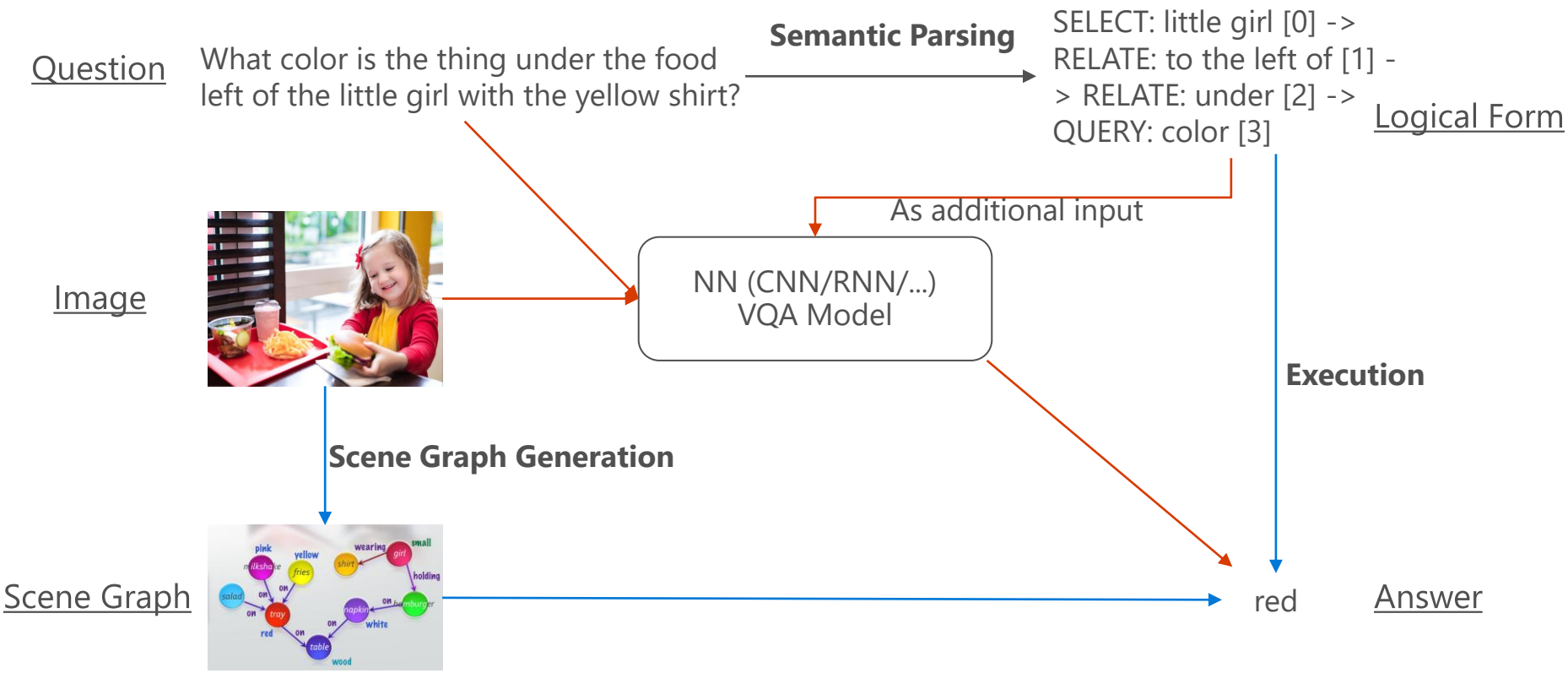
A: Food

Q: What color is the thing under the food left of the little girl?

A: Red

NL Question	What color is the thing under the food left of the little girl?
Image	
Program	SELECT: little girl [0] -> RELATE: to the left of [1] -> RELATE: under [2] -> QUERY: color [3]
Grammar-Guided Structural Representation	

Pipeline

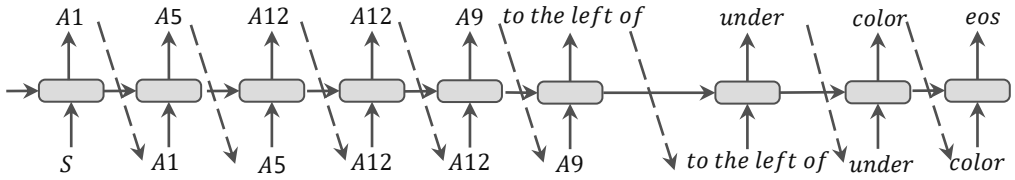

Question
What color is the thing under the food left of the little girl with the yellow shirt?
Semantic Parsing
SELECT: little girl [0] -> RELATE: to the left of [1] -> RELATE: under [2] -> QUERY: color [3]
Logical Form
As additional input
Image
NN (CNN/RNN/...) VQA Model
Execution
Scene Graph Generation
Scene Graph
red
Answer

Visual Semantic Parsing

	LF Exact Match
Seq2Seq	77.4%
Seq2Action	85.6%

What color is the thing under the food left of the little girl?

SELECT: little girl [0] ->
RELATE: to the left of [1] ->
RELATE: under [2] ->
QUERY: color [3]



Action	Operation	Description
A1-A4	S->attribute_val | attribute_pred | relation_name | bool	Possible types of answer
A5	attribute_val->Query(object_set, att_pred)	Get attribute of first object from the 'object_set' with the predicate 'att_pred'
A6	attribute_val->Chooseatt(object_set, att_pred, att_val, att_val)	Choose between attribute values
A7	attribute_pred -> Common(object_set set)	Get the predicate that the all objects in the set have same value
A8	relation_name -> Chooserel(object_set, obj_name, rel_name, rel_name)	Choose between relation names
A9	object_set -> Select(obj_name)	Get objects with name of 'name' from the scene graph
A10	object_set -> Filter(object_set, att_pred?, att_val)	Subset of objects that have attribute 'pred' equals 'val'
A11	object_set -> Group(object_set, object_set)	Union of two object sets
A12	object_set -> Relate(object_set, rel_name)	Get object that is linked to/from the first object of 'set' with relation 'rel'
A13	boolean -> Logic(lg_op, boolean, boolean)	Logical operation
A14	boolean ->Exist(object_set)	True if set is not empty
A15	boolean ->Verifyatt(object_set, att_pred?, att_val)	True if objects have attribute 'pred' equals 'val'
A16	boolean ->Verifyrel(object_set, obj_name, rel_name)	True if the first object of 'set' have the relation with object
A17	boolean ->Same(object_set, att_pred)	True if objects have same value for the attribute 'pred'
A18	boolean ->Different(object_set, att_pred)	True if objects do not have same value for the attribute 'pred'
A19-20	lg_op -> And | Or	
A21-A24	att_pred, att_val, obj_name, rel_name -> instantiation	instantiation

Modular Network

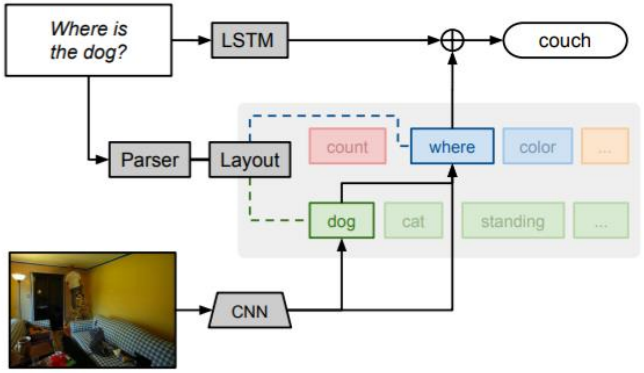
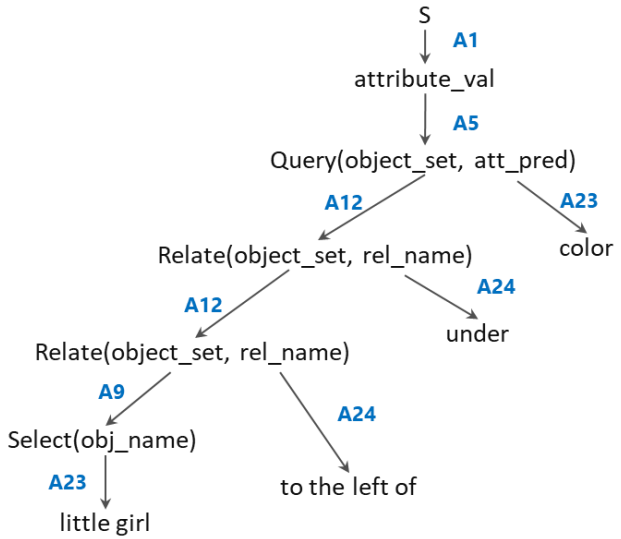
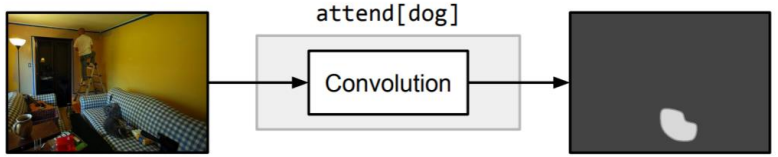


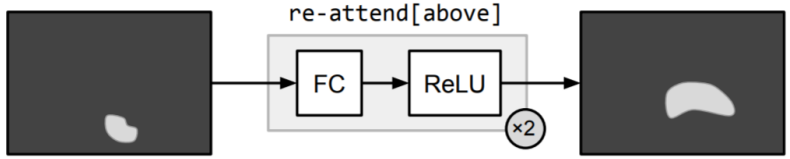
Image courtesy [Andreas et al. 2016]



Select



Relate



Intermediate Reasoning Result

Select(man)

Select(sky)

Intermediate Reasoning Result

Select(bus) Filter(Color, red)

Select(bus) Filter(Color, white)

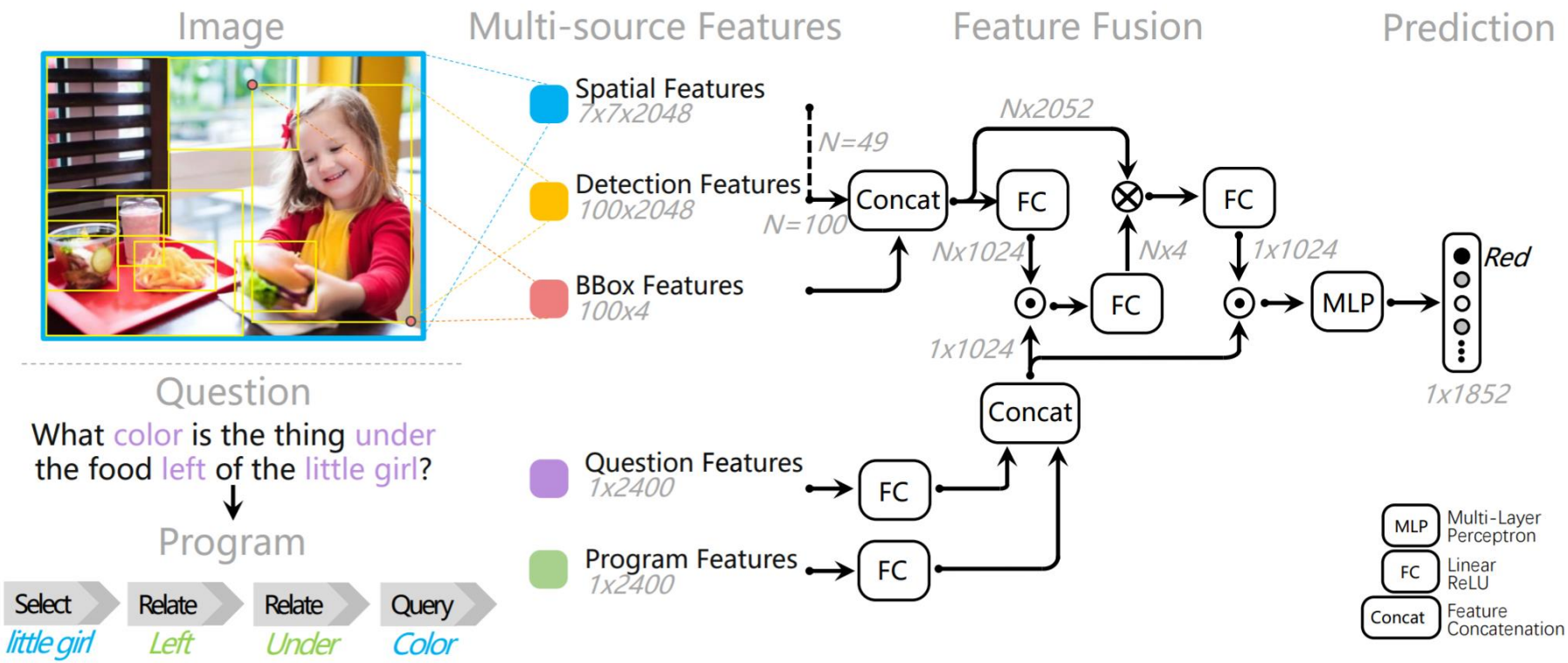
Results

Methods	Simple Questions (program length <3)	Complex Questions (program length >=3)
MLB [20]	71.73	56.25
Mutan [4]	73.89	57.22
CoR [38]	75.07	58.40
MSP(ours)	**76.44**	**58.80**

Improvements over baselines

Still have a big room

System Integration



References

- Retrieval
  - Yibo Sun, Duyu Tang, Nan Duan, Tao Qin, Shujie Liu, Zhao Yan, Ming Zhou, Yuanhua Lv, Wenpeng Yin, Xiaocheng Feng, Bing Qin, Ting Liu. Joint Learning of Question Answering and Question Generation. 2019. TKDE.
  - Duyu Tang, Nan Duan, Zhao Yan, Zhirui Zhang, Yibo Sun, Shujie Liu, Yuanhua Lv and Ming Zhou. Learning to Collaborate for Question Answering and Asking. 2018. NAACL.
- Semantic Parsing
  - Yibo Sun, Duyu Tang, Nan Duan, Jianshu Ji, Guihong Cao, Xiaocheng Feng, Bing Qin, Ting Liu and Ming Zhou. Semantic Parsing with Syntax- and Table-Aware SQL Generation. 2018. ACL.
- Question Generation
  - Daya Guo, Yibo Sun, Duyu Tang, Nan Duan, Jian Yin, Hong Chi, James Cao, Peng Chen and Ming Zhou. Question Generation from SQL Queries Improves Neural Semantic Parsing. 2018. EMNLP.
- Conversation
  - Yibo Sun, Duyu Tang, Nan Duan, Jingjing Xu, Xiaocheng Feng, Bing Qin. Knowledge-Aware Conversational Semantic Parsing Over Web Tables. 2018. arxiv
- Table-to-Text Generation
  - Junwei Bao, Duyu Tang, Nan Duan, Zhao Yan, Yuanhua Lv, Ming Zhou, Tiejun Zhao. Table-to-Text: Describing Table Region with Natural Language. 2018. AAAI.

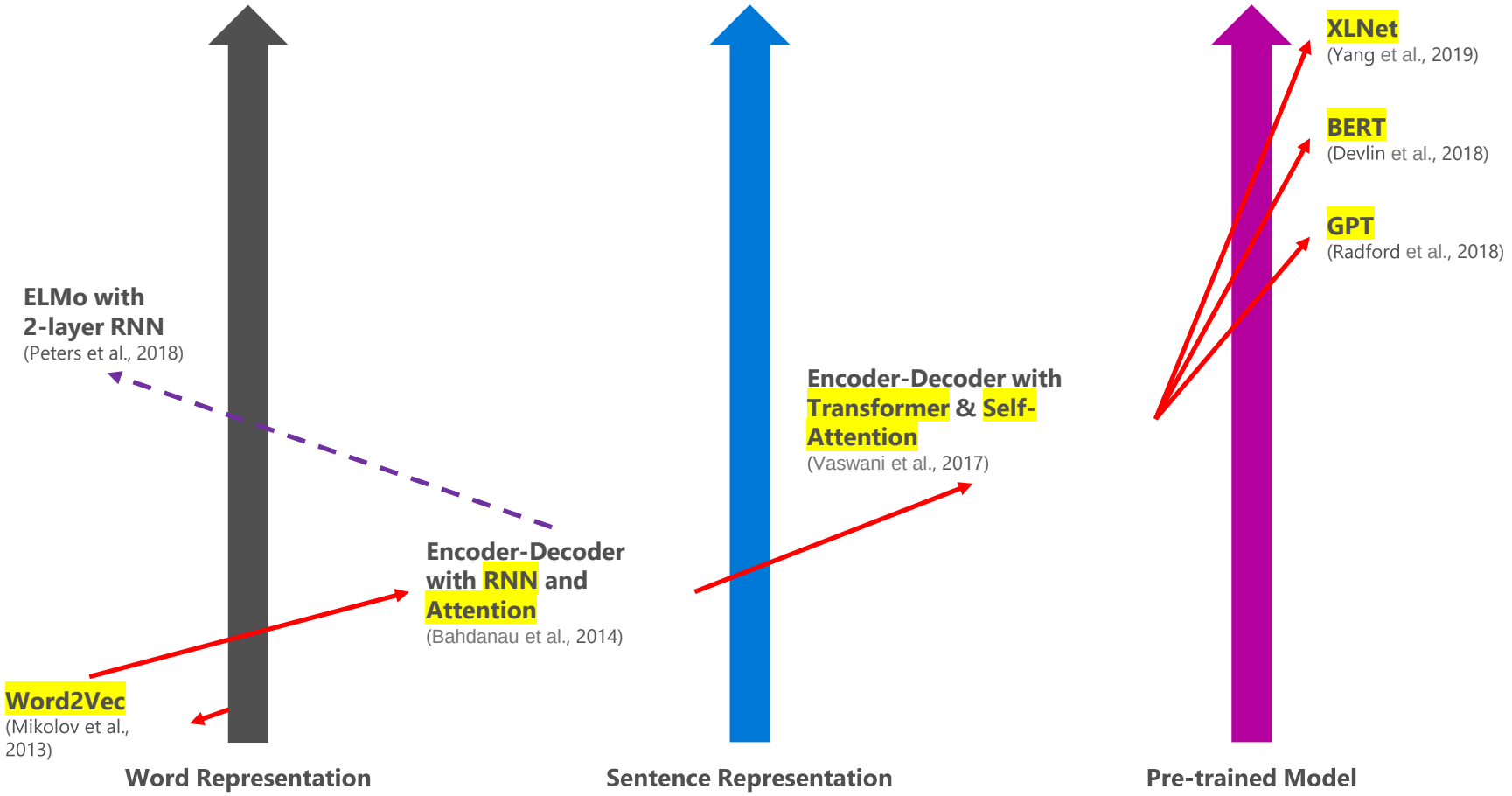
References

- Conversational Knowledge based Question Answering
  - Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, Jian Yin. Dialog-to-Action: Conversational Question Answering Over a Large-Scale Knowledge Base. 2018. NeurIPS.
  - Daya Guo, Duyu Tang, Nan Duan, Ming Zhou and Jian Yin, Coupling Retrieval and Meta-Learning for Context-Dependent Semantic Parsing, ACL, 2019.
- Image Question Answering
  - Chenfei Wu, Yanzhao Zhou, Gen Li, Nan Duan, Duyu Tang, Ming Zhou. Deep Reason: A Strong Baseline for Real-World Visual Reasoning, 2019. arxiv.

Pre-training

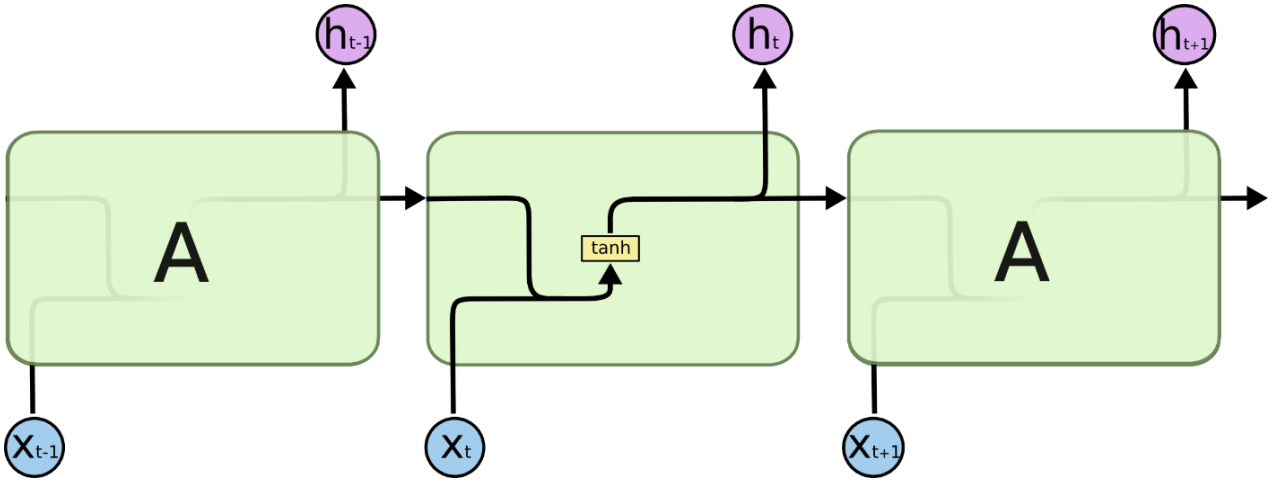
Nan Duan

What happened in NLP recently?



(1) Recurrent Neural Network (RNN)

(Elman, 1990)



$$h_t = \tanh(W_h \cdot h_{t-1} + W_x \cdot x_t + b) = \tanh(W \cdot [h_{t-1}; x_t] + b)$$

Using RNN in Language Modeling (LM)

An NLP task about computing $p(w_k|w_1 \dots w_{k-1})$

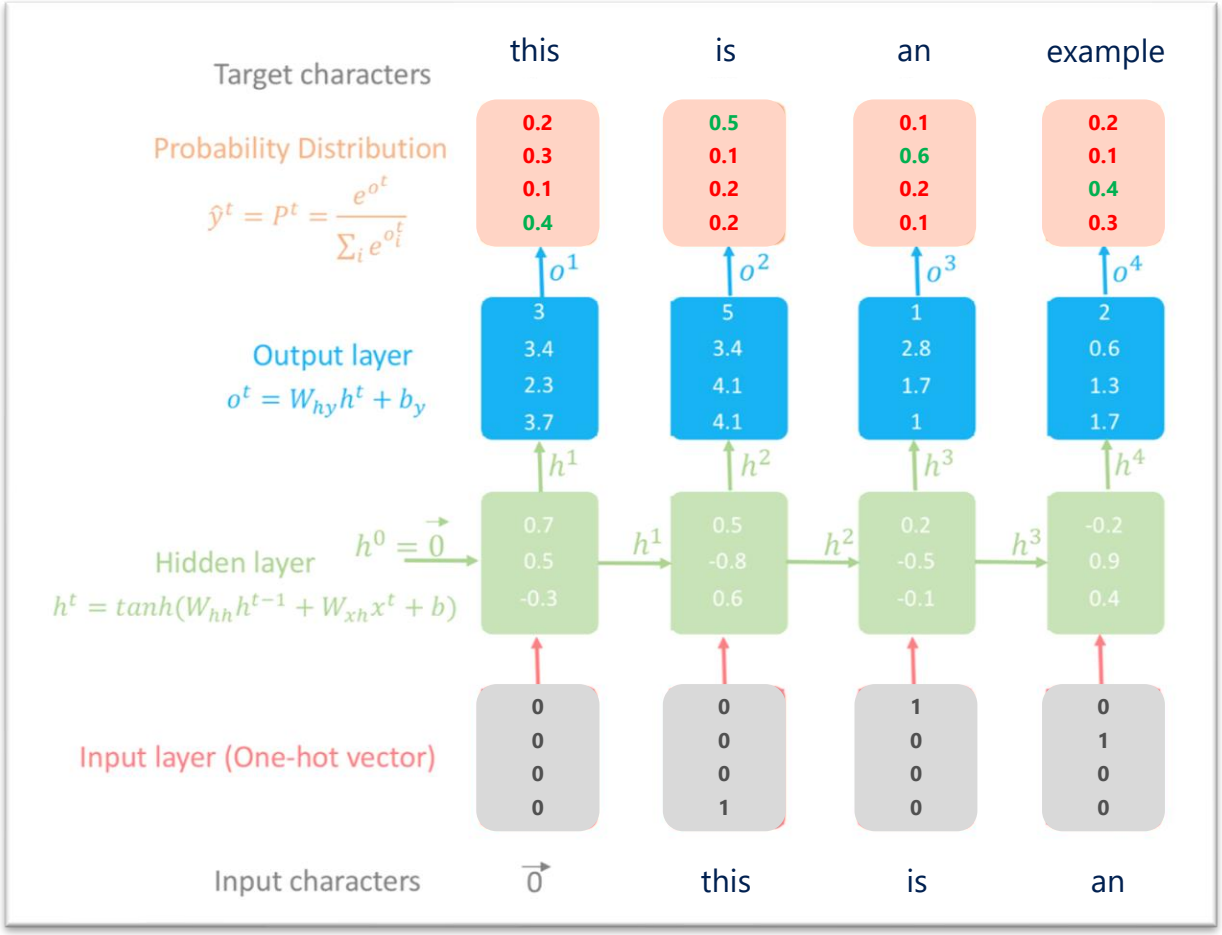
Statistical LM: $p(w_k|w_1 \dots w_{k-1}) = \frac{Count(w_1 \dots w_{k-1} w_k)}{\sum_{w'_k} Count(w_1 \dots w_{k-1} w'_k)}$

$$\approx \frac{Count(w_{k-n+1} \dots w_{k-1} w_k)}{\sum_{w'_k} Count(w_{k-n+1} \dots w_{k-1} w'_k)}$$

Neural LM: $p(w_k|w_1 \dots w_{k-1}) = \frac{e^{f(w_1 \dots w_{k-1} w_k)}}{\sum_{w'_k} e^{f(w_1 \dots w_{k-1} w'_k)}}$

where $e^{f(w_1 \dots w_{k-1} w_k)}$ comes from a neural network (such as RNN)

An Example of RNN-LM

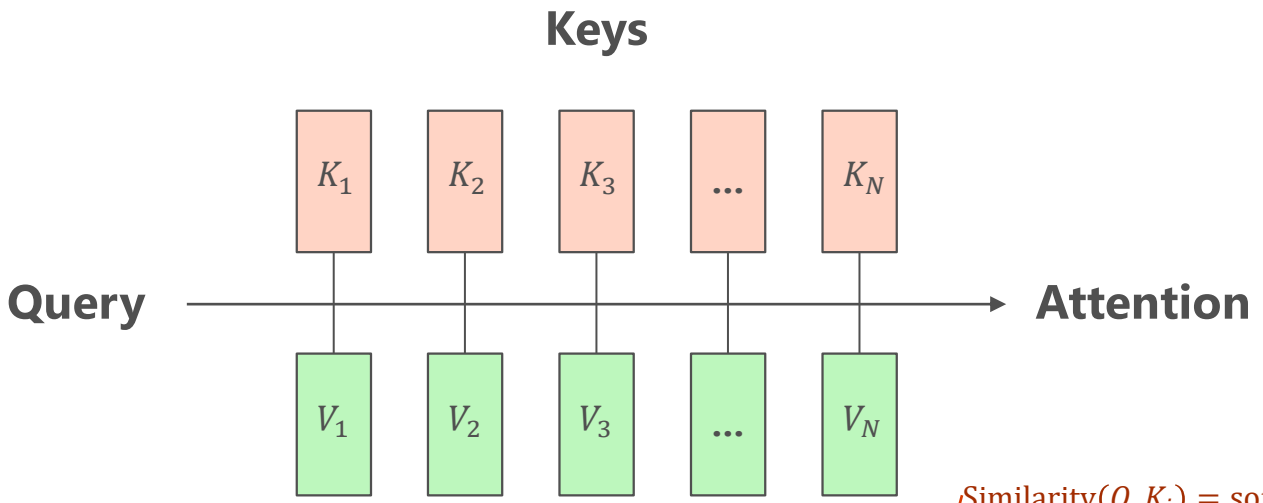


vocab

is
an
example
this

Attention

(An attention function maps a query and a set of key-value pairs to an attention score/vector.)

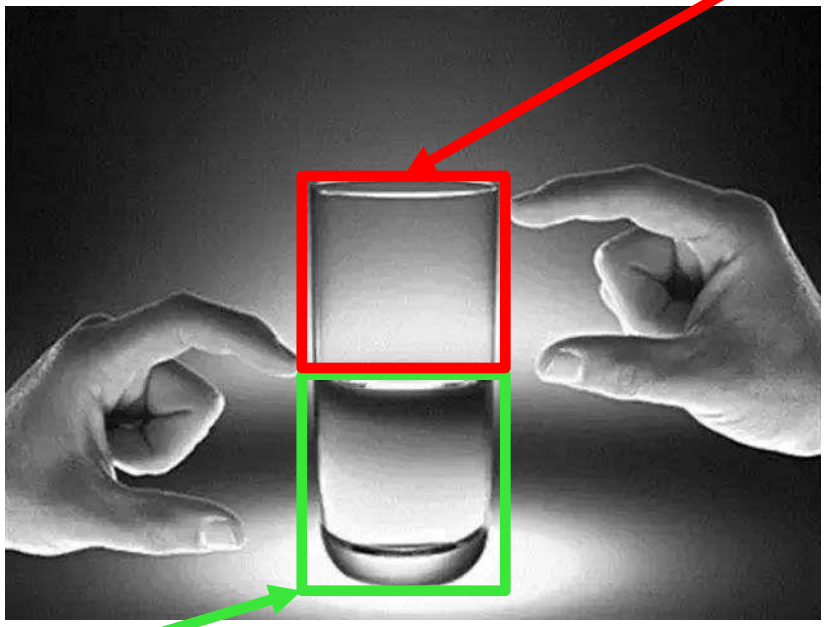


$$\text{Similarity}(Q, K_i) = \text{softmax}(f(Q, K_i))$$

- $f(Q, K_i) = Q^{\mathrm{T}} K_i$ dot-product
- $f(Q, K_i) = W_a[Q; K_i]$ concatenation
- $f(Q, K_i) = v_a^{\mathrm{T}} \tanh(W_a Q + U_a K_i)$ perceptron
- ...

$$\text{Attention}(Q, K_1, \dots, K_N, V_1, \dots, V_N) = \sum_{i=1}^{N} \text{Similarity}(Q, K_i) \cdot V_i$$

How to Understand Attention Intuitively



Score(,) > Score(,)

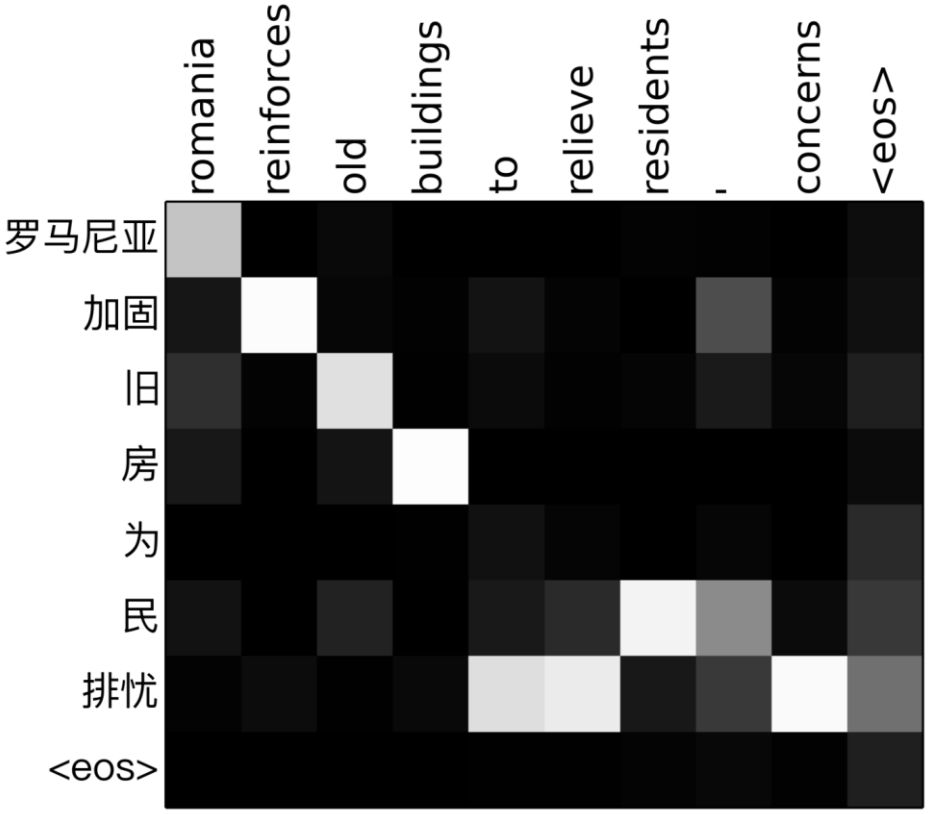
Score(,) > Score(,)

How to Understand Attention Intuitively

Q: where is the dog laying?
A: sidewalk

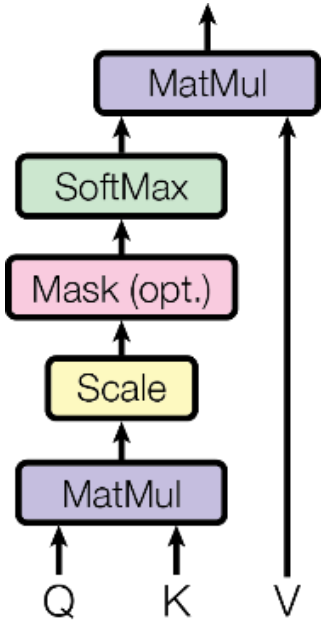
Q: what is around the man's neck? A: tie

Visual Question Answering



Neural Machine Translation

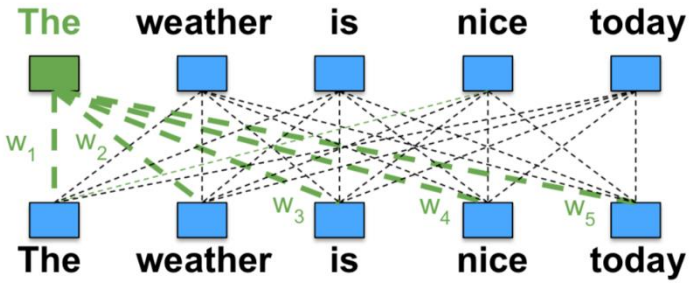
Scaled Dot-Product Attention



$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V$$

For large values of d_k, the dot products grow large in magnitude, pushing the softmax function into regions where it has extremely small gradients. To counteract this effect, the dot products are scaled by $\frac{1}{\sqrt{d_k}}$.

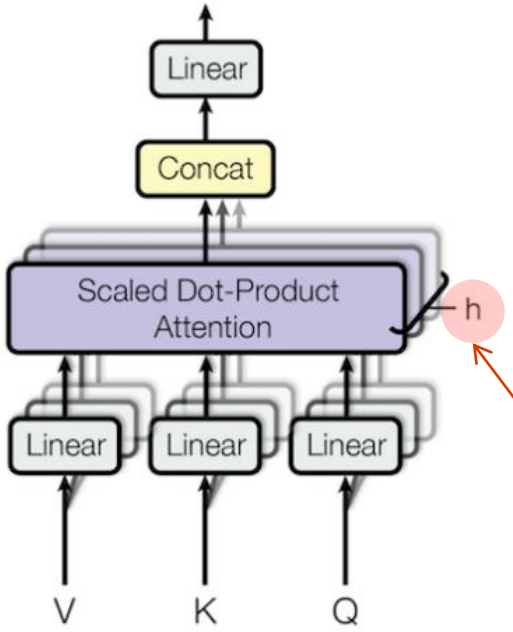
An Example



$$w_1, w_2, w_3, w_4, w_5 = \text{softmax}\left(\begin{bmatrix} 0.6 & 0.2 & 0.8 \end{bmatrix}_{\text{The}} \times \begin{bmatrix} 0.6 & 0.2 & 0.9 & 0.4 & 0.4 \\ 0.2 & 0.3 & 0.1 & 0.1 & 0.1 \\ 0.8 & 0.1 & 0.8 & 0.4 & 0.6 \end{bmatrix}_{\text{The weather is nice today}} \right)$$

$$\begin{bmatrix} 1.8 \\ 2.3 \\ 0.4 \end{bmatrix}_{\text{The}} = w_1 \times \begin{bmatrix} 0.6 \\ 0.2 \\ 0.8 \end{bmatrix}_{\text{The}} + w_2 \times \begin{bmatrix} 0.2 \\ 0.3 \\ 0.1 \end{bmatrix}_{\text{weather}} + w_3 \times \begin{bmatrix} 0.9 \\ 0.1 \\ 0.8 \end{bmatrix}_{\text{is}} + w_4 \times \begin{bmatrix} 0.4 \\ 0.1 \\ 0.4 \end{bmatrix}_{\text{nice}} + w_5 \times \begin{bmatrix} 0.4 \\ 0.1 \\ 0.6 \end{bmatrix}_{\text{today}}$$

Multi-Head Attention



Compute attentions from different representation spaces.

$$head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, \dots, head_h)W^O$$

(2) Transformer

(Vaswani et al., 2017)

Layer Normalization adjusts parameters in each layer to a more stable distribution and accelerates the training of the network.

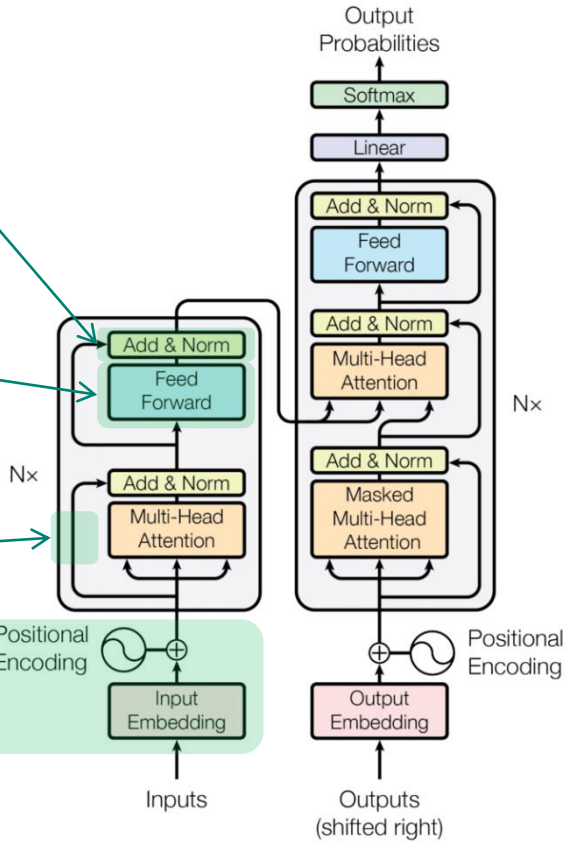
Position-wide Feed-Forward Networks is applied to each position:

$$FFN(x) = \max(0, xW_1 + b_1)\,W_2 + b_2$$

While the linear transformations are the same across different positions, they use different parameters from layer to layer.

Residual connections allow to have deeper models.

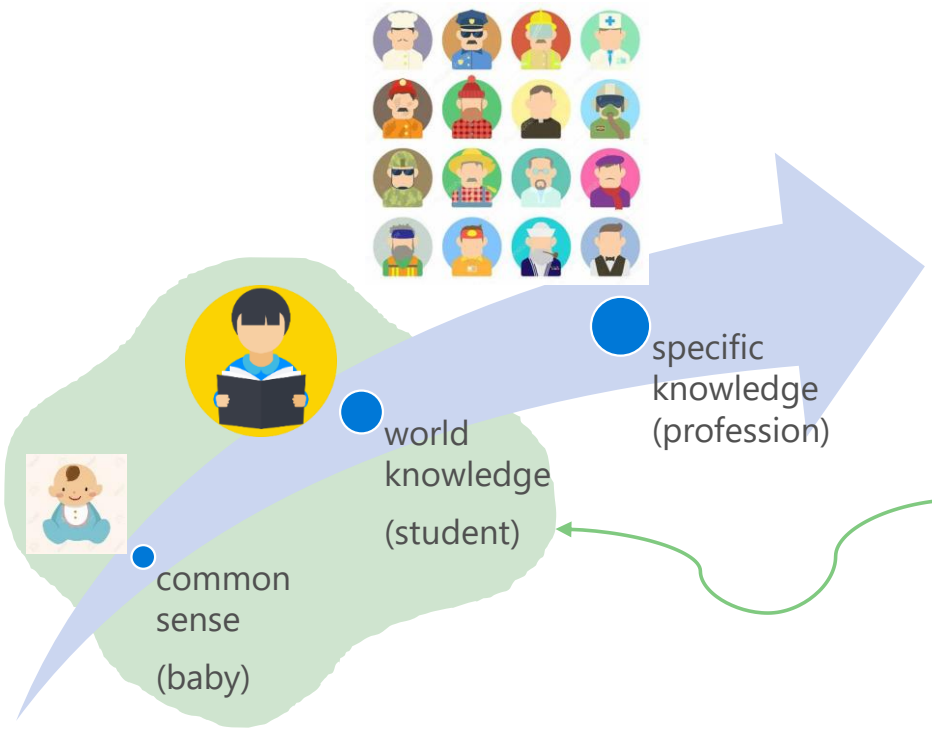
Position encodings make use of the order of the sequence and they have the same dimension as input embeddings. The two can be summed to tell the neural net that there's an underlying sequential structure.



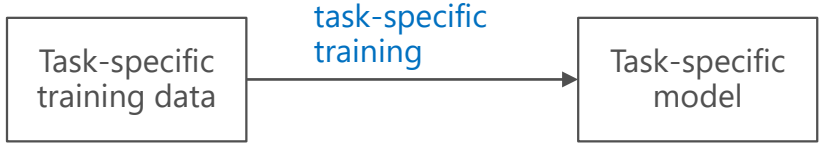
Outline

- Pre-training in NLP
- Pre-training in Language + Vision

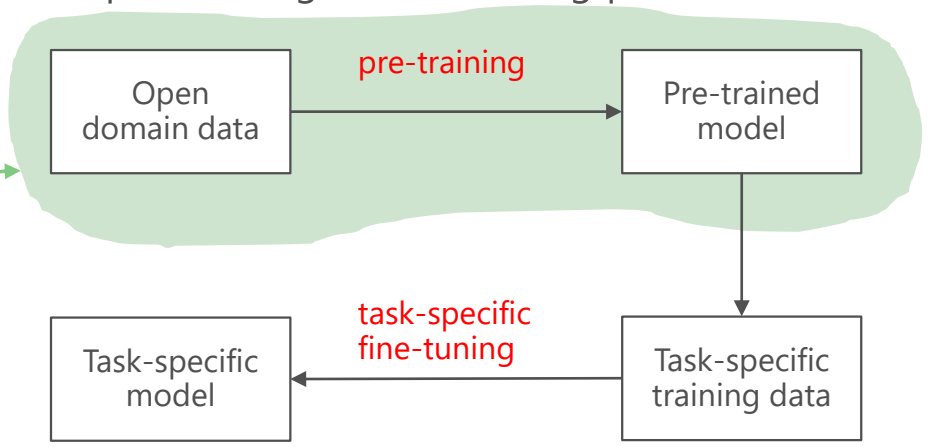
What is Pre-training & Why is it Important



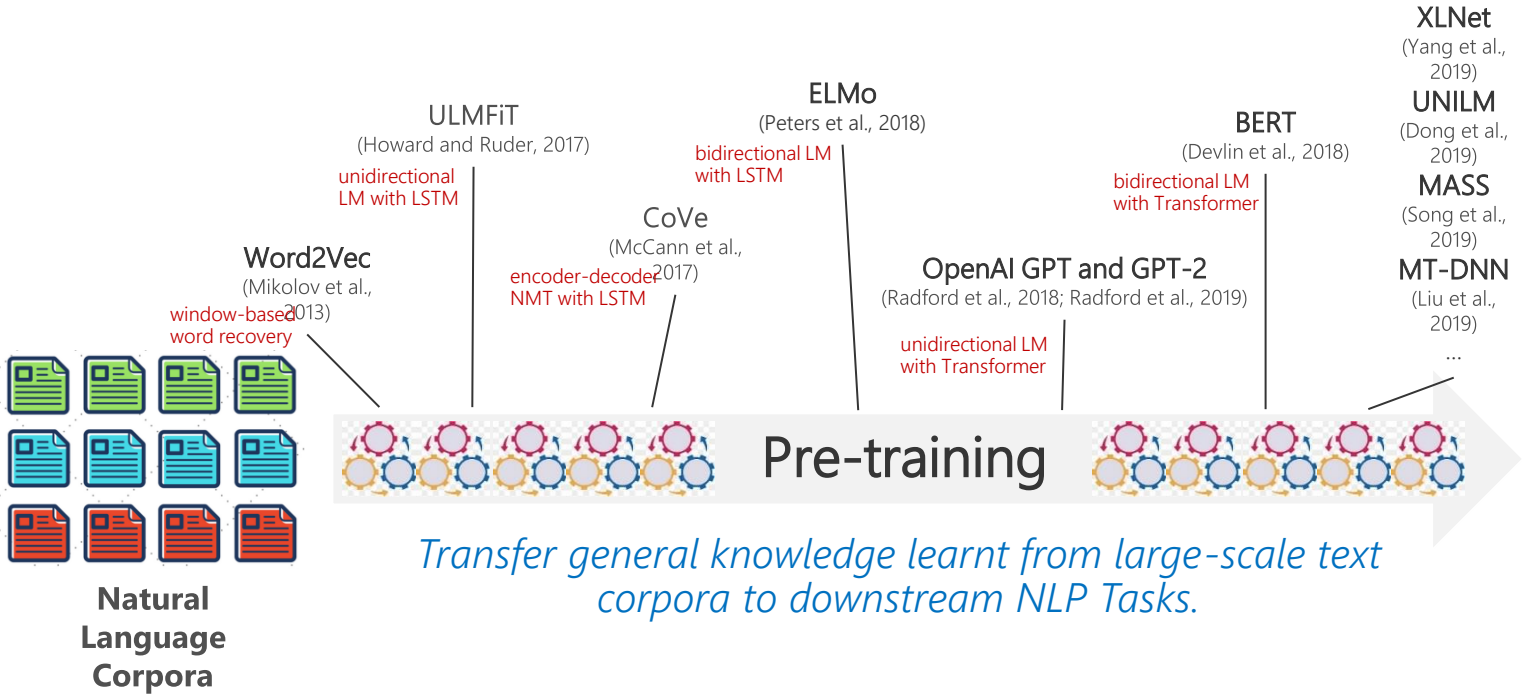
- A typical training procedure



- A pre-training-based training procedure



Pre-training in NLP



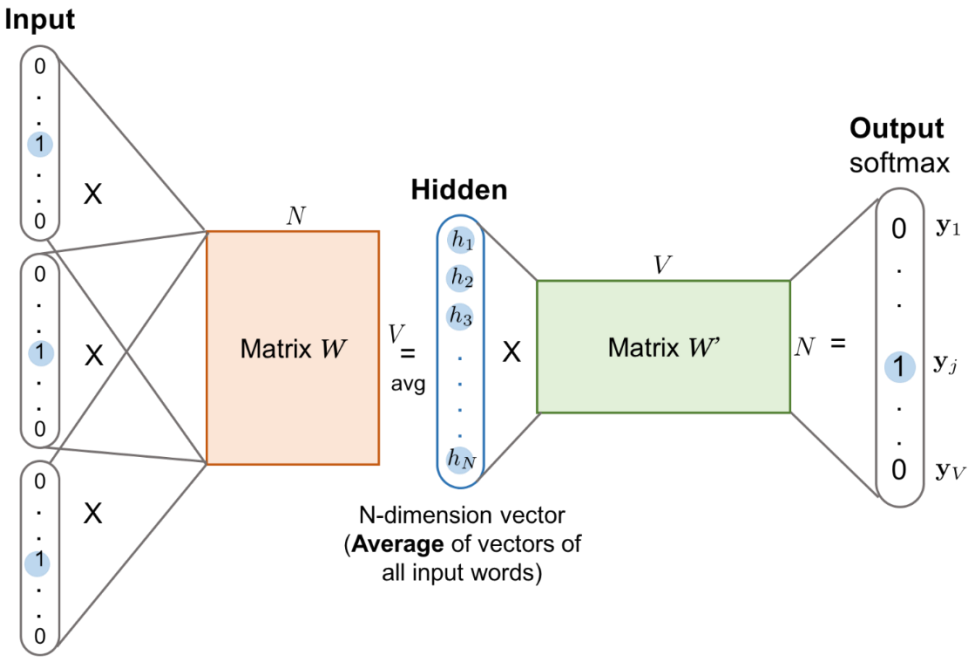
Transfer general knowledge learnt from large-scale text corpora to downstream NLP Tasks.

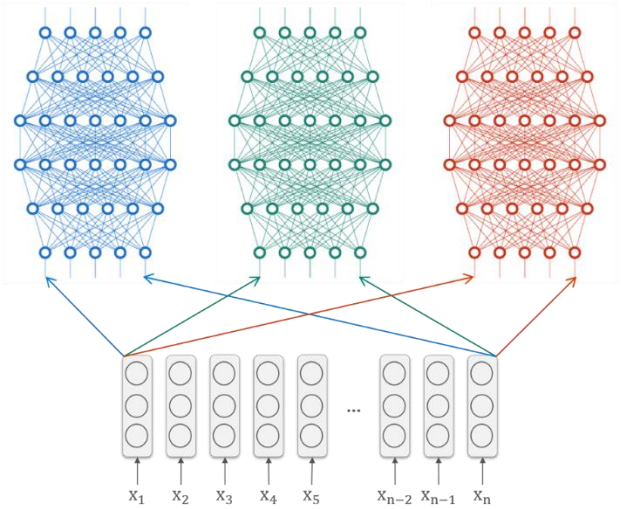
NLP Tasks

- Machine Translation
- Search Engine
- Semantic Parsing
- Question Answering
- Chatbot & Dialogue
- Paraphrase Classification
- Text Entailment
- Sentiment Analysis
- ...

Word2Vec

(Mikolov et al., 2013)





Initialize the first layer of a neural network

objective function:

minimize $-\frac{1}{T}\sum_{t=1}^{T}\log p^{CBOW}(x_t|x_{t-n},\dots,x_{t-1},x_{t+1},\dots,x_{t+n})$

Key Issue

Cannot deal with ***unknown or OOV words***.

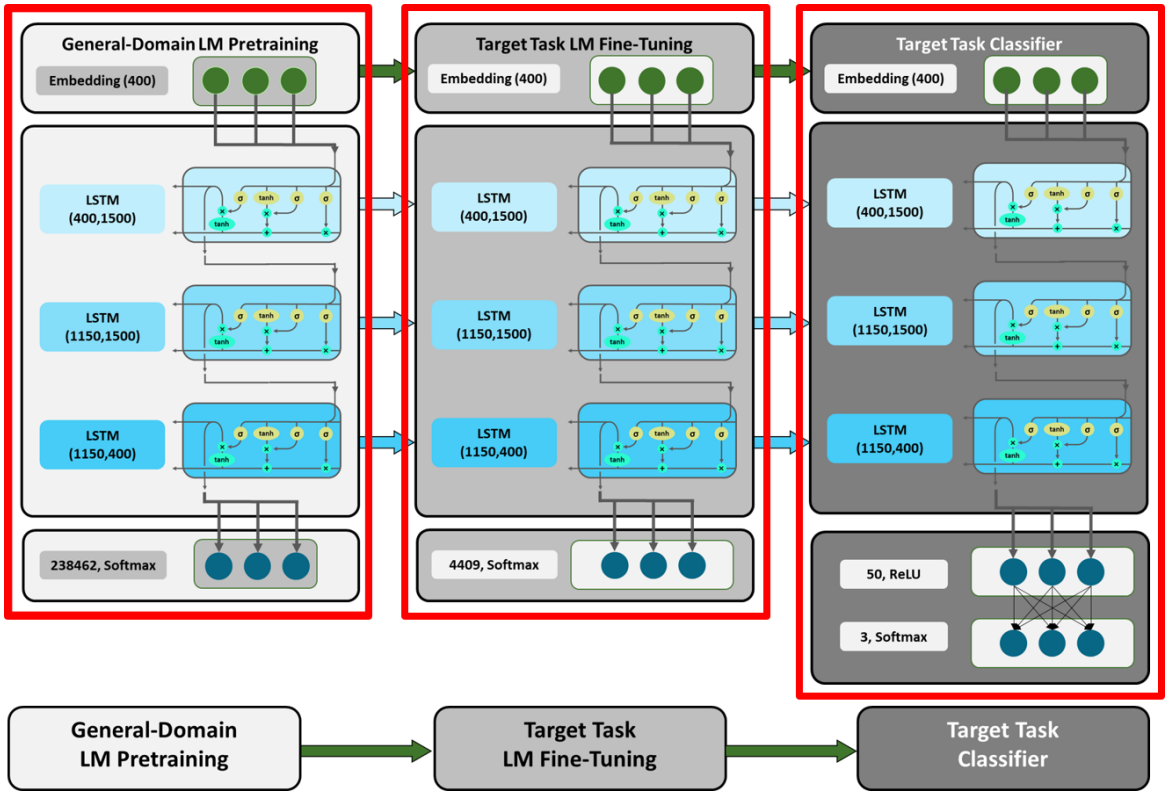
"Yesterday, I watched a new released movie called ***asdjadaklsxa***."

Cannot deal with the ***ambiguity (or context-aware) phenomenon***.

*"He walked along the **bank** of the river."* vs. *"He borrowed money for the **bank**."*

ULMFiT

(Howard and Ruder, 2017)



- **General-Domain LM Pretraining:** In a first step, a LM is pretrained on a large general-domain corpus (the WikiText-103 dataset).
- **Target Task LM Fine-Tuning:** The LM is consequently fine-tuned on the data of the target task (the Twitter US Airline Sentiment dataset), which is likely from a different distribution than the source task dataset.
- **Target Task Classifier:** The pretrained LM is expanded by two linear blocks so that the final output is a probability distribution over the sentiment labels (i.e. positive, negative and neutral)

ULMFiT should win more respect.

Discriminative fine-tuning

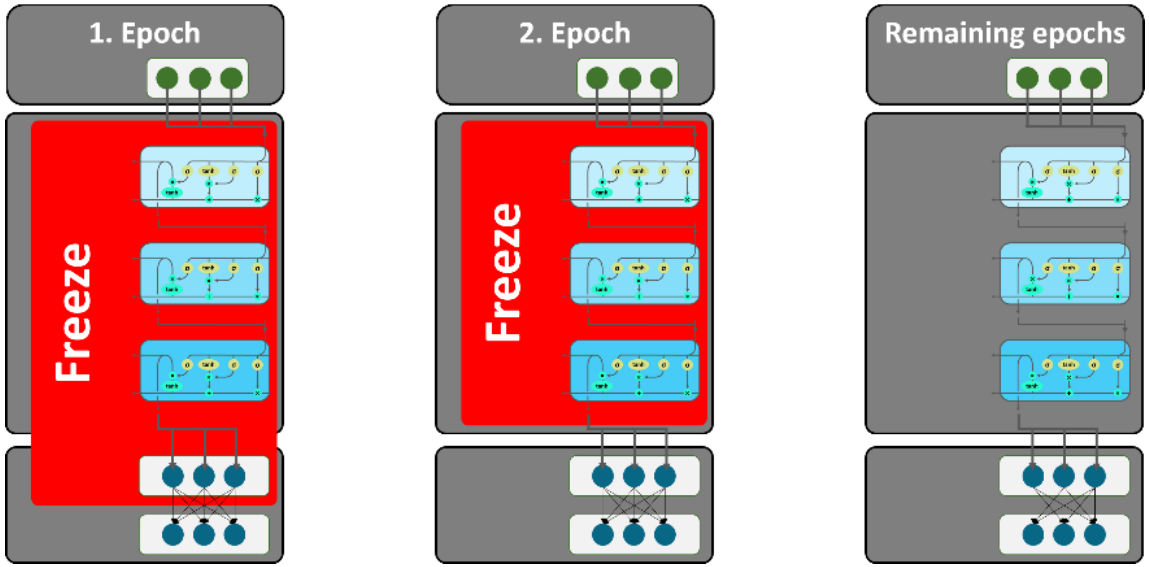
It is likely the first effective approach to fine-tune LM for NLP tasks.

Concat Pooling

The last hidden state is concatenated with a max-pooled and a mean-pooled representation of all hidden states.

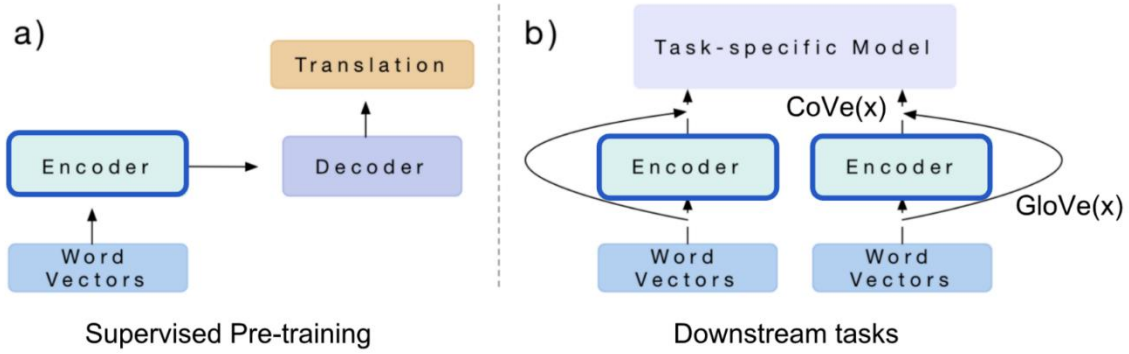
Gradual Unfreezing

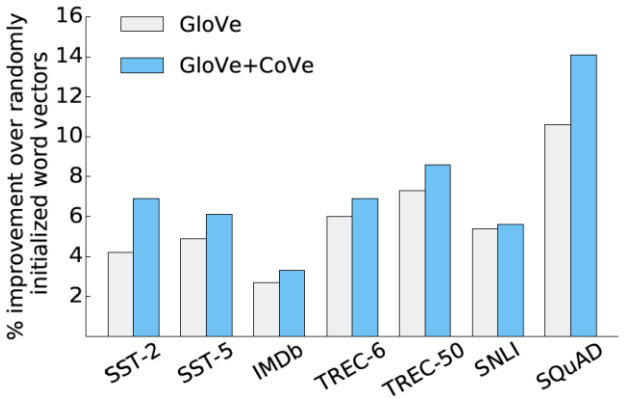
First, all layers but the softmax output layer are frozen;
Next, the ReLU layer is unfrozen and fine-tuned;
Last, the entire model is fine-tuned.



CoVe

(McCann et al., 2017)





$$\text{CoVe}(w) = \text{MTLSTM}(\text{GloVe}(w))$$

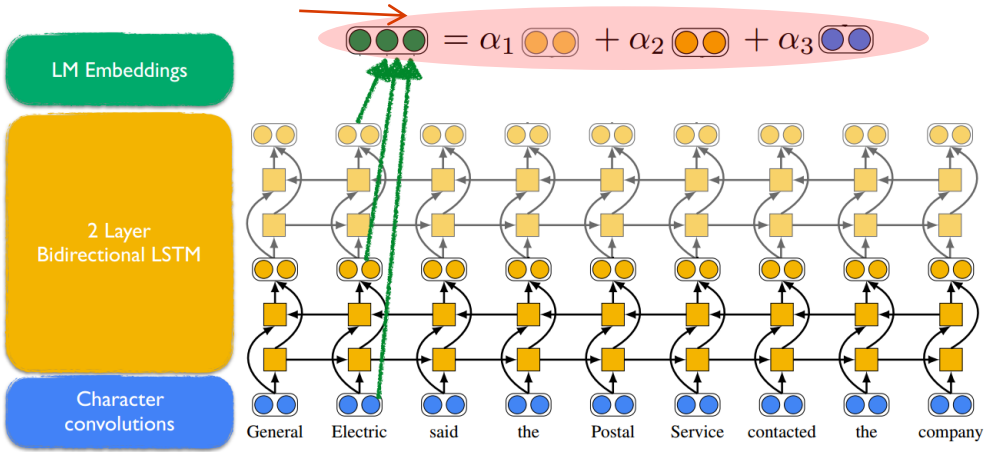
$$v_w = [\text{GloVe}(w); \text{CoVe}(w)]$$

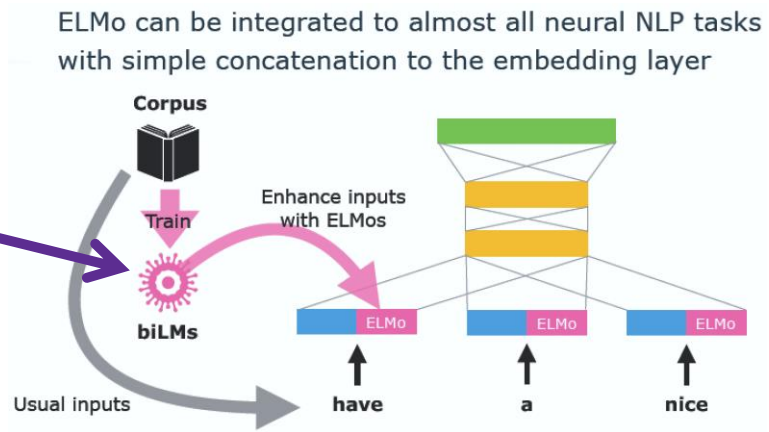
Issue: Pre-training is bounded by ***available datasets on the supervised translation task***.

ELMo

(Peters et al., 2018)

Let the end task model learn a linear combination of these representations.



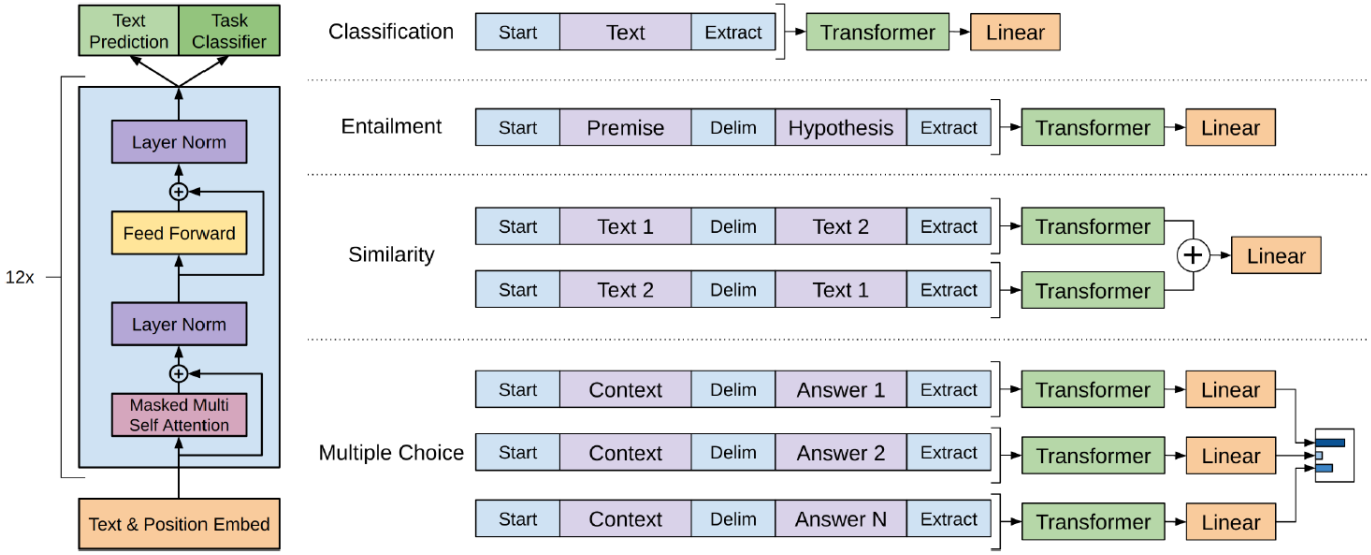


objective function:

maximize $\sum_{t=1}^{T}(\log p(t_k|t_1,\ldots,t_{k-1};\Theta_x,\overrightarrow{\Theta_{LSTM}},\Theta_{softmax}) + \log p(t_k|t_{k+1},\ldots,t_N;\Theta_x,\overleftarrow{\Theta_{LSTM}},\Theta_{softmax}))$

OpenAI GPT and GPT-2

(Radford et al., 2018; Radford et al., 2019)



$h_0 = UW_e + W_p$

$h_l = \text{transformer}_{\text{block}(h_{l-1})}$

$p(u) = \text{softmax}(h_n W_e^{\text{T}})$

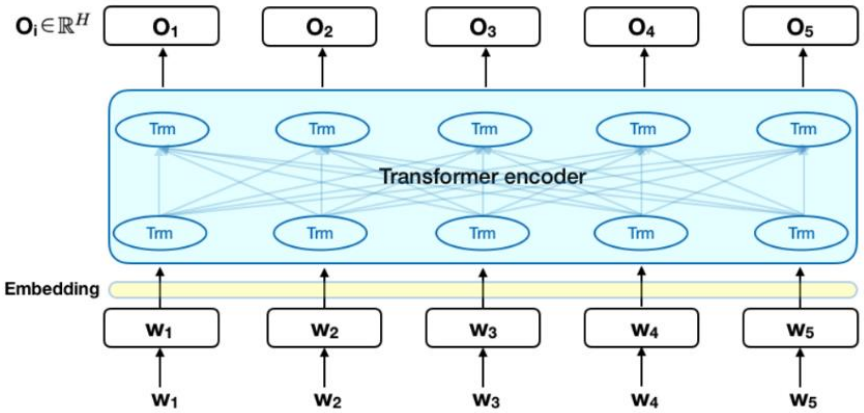
pre-training objective function:
maximize $\sum_i \log p(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$

$p(y|x^1, \ldots, x^m) = \text{softmax}(h_l^m W_y)$

fine-tuning objective function:
maximize $\lambda \cdot \sum_i \log p(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta) + \sum_{(x,y)} \log p(y|x^1, \ldots, x^m)$

BERT

(Devlin et al., 2018)



Use the output of the masked word's position to predict the masked word

Possible classes: All English words

0.1% Aardvark
... ...
10% Improvisation
... ...
0% Zyzzyva

FFNN + Softmax

BERT

Randomly mask 15% of tokens

[CLS] Let's stick to [MASK] in this skit

Input

[CLS] Let's stick to improvisation in this skit

Predict likelihood that sentence B belongs after sentence A

1% IsNext
99% NotNext

FFNN + Softmax

BERT

Tokenized Input

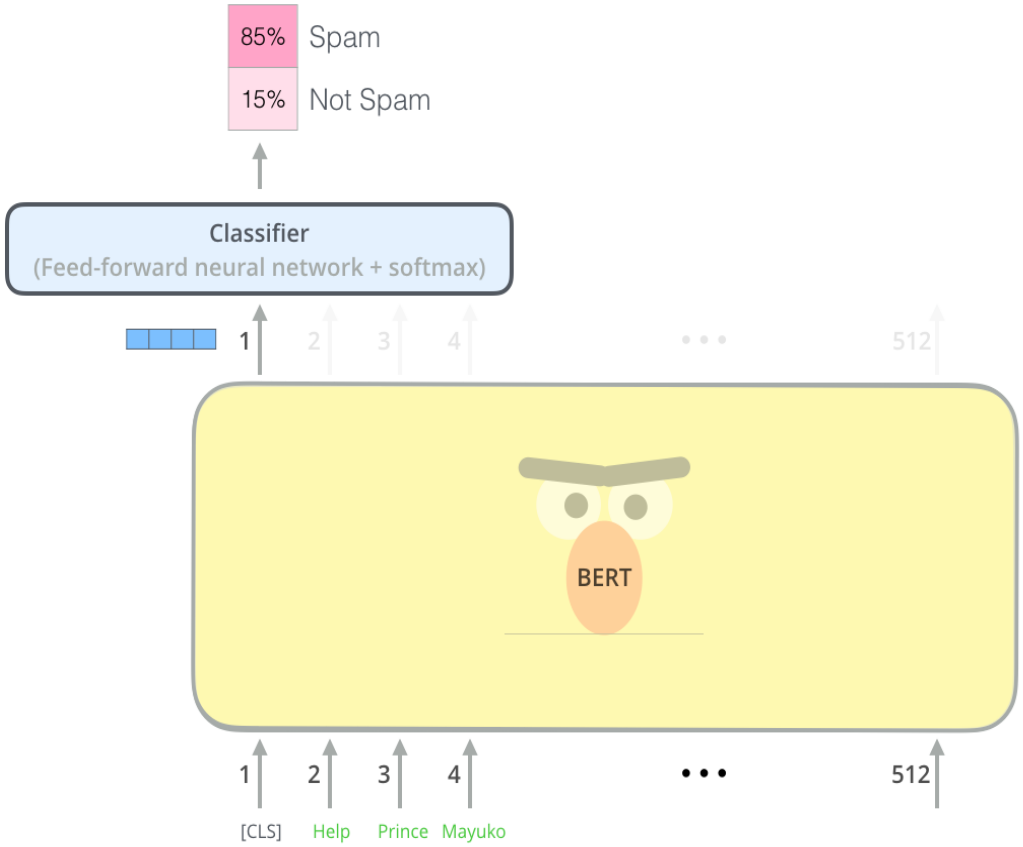
[CLS] the man [MASK] to the store [SEP]

Input

[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds [SEP]

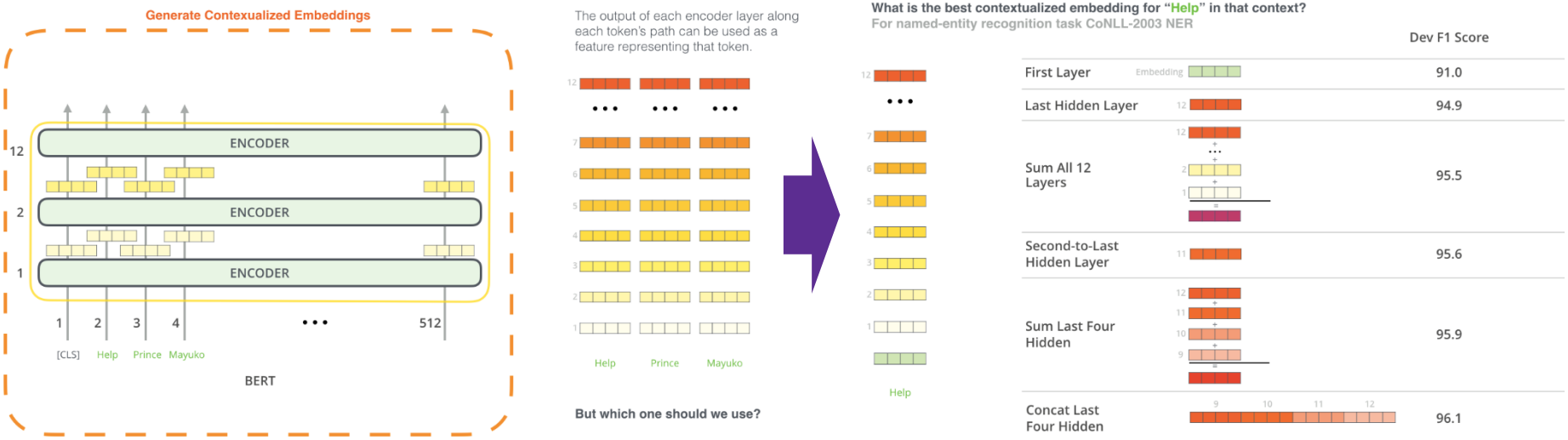
Sentence A | Sentence B

How to use BERT: Discriminative Fine-tuning



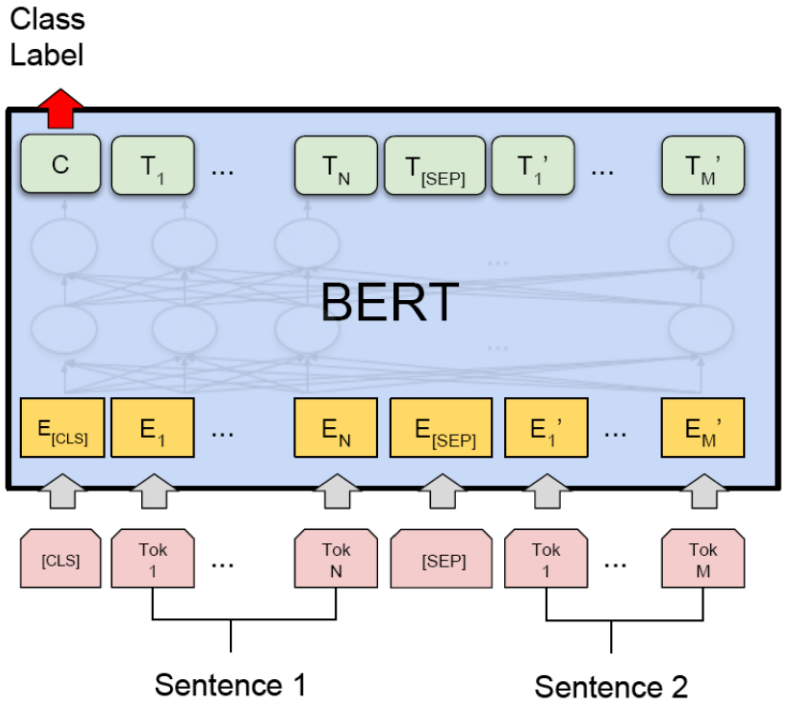
BERT for Spam Email Detection

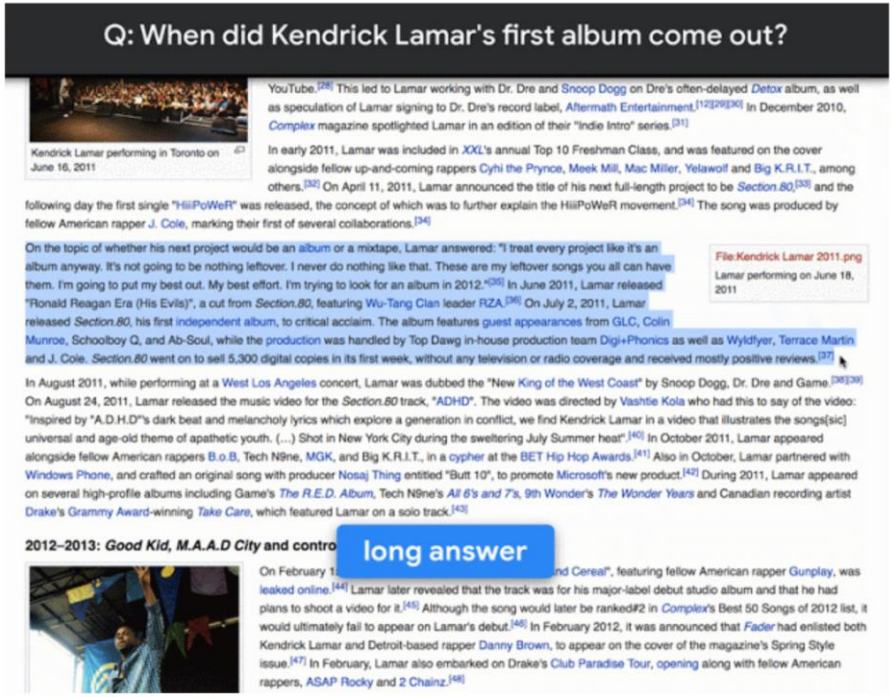
How to use BERT: Feature Extraction



The fine-tuning approach isn't the only way to use BERT. Just like ELMo, you can ***use the pre-trained BERT to create contextualized word embeddings***. Then you can feed these embeddings to your existing model – a process the paper shows yield results not far behind fine-tuning BERT on a task such as named-entity recognition.

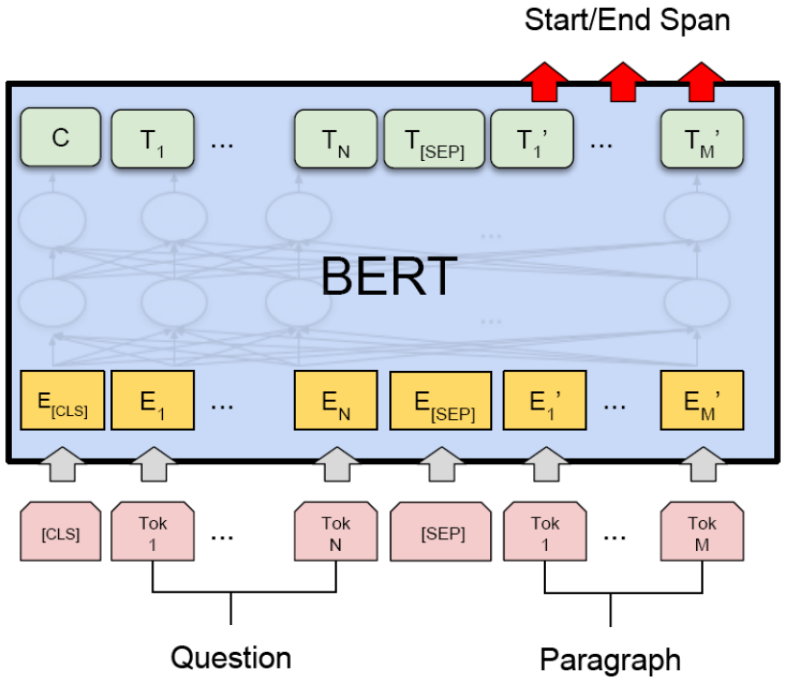
BERT for Matching-based QA Tasks





NQ (Kwiatkowski et al. '19)

BERT for Extraction-based QA Tasks



Passage

S_1 : Pharmacists are healthcare professionals with specialized education and training who perform various roles to ensure optimal health outcomes for their patients through the quality use of medicines.

S_2 : Pharmacists may also be **small-business proprietors**, owning the pharmacy in which they practice.

S_3 : Since pharmacists know about the mode of action of a particular drug, and its metabolism and physiological effects on the human body in great detail, they play an important role in optimization of a drug treatment for an individual.

Question: What other role do many pharmacists play?

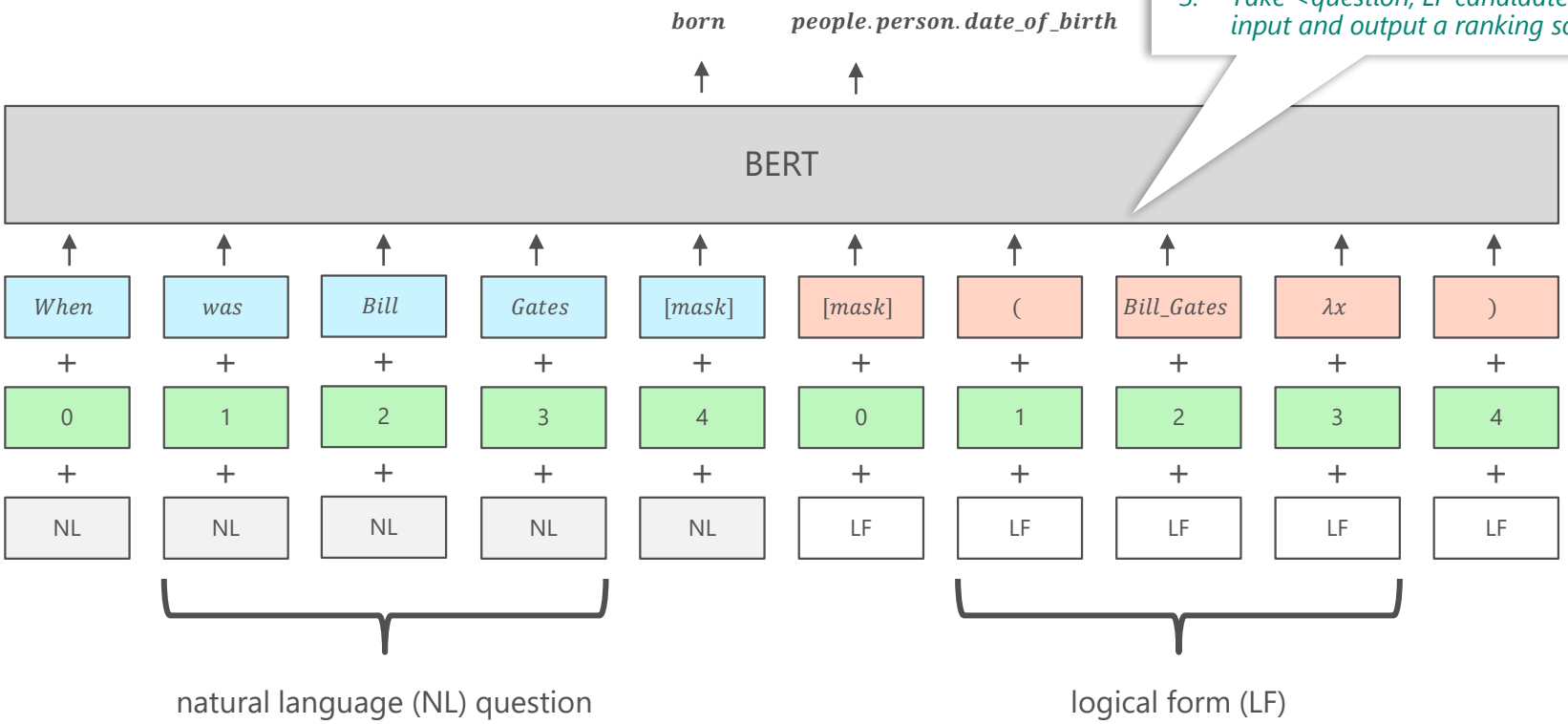
Answer: **small-business proprietors**

SQuAD (Rajpurkar et al. '16)

BERT for Knowledge-based QA Tasks

BERT for semantic parsing re-ranking

1. *Fine-tuning BERT on NL-LF pairs;*
2. *Given a question, generate LF n-bests with any semantic parser;*
3. *Take <question, LF candidate> as input and output a ranking score.*



微软亚洲研究院语义分析数据集: MSParS
(a Multi-perspective Semantic ParSing Dataset)

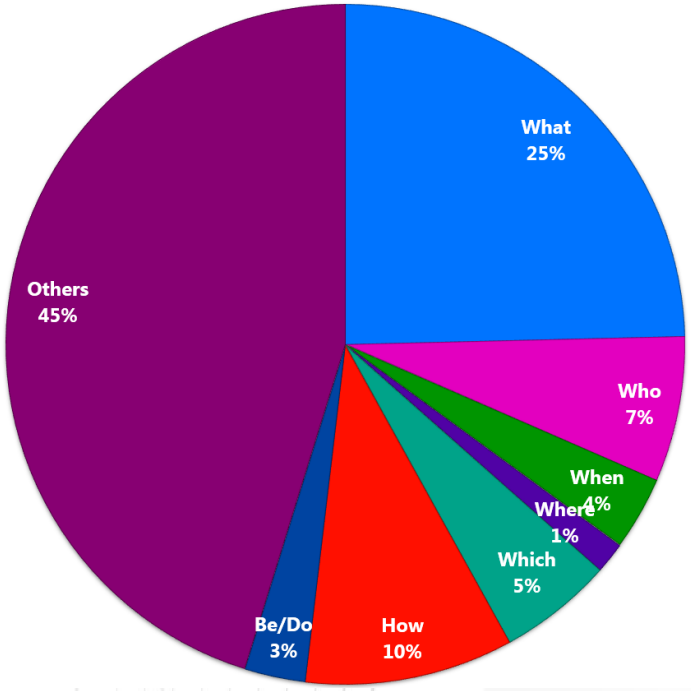
Introduction

MSParS is a large-scale dataset for Knowledge-based Semantic Parsing including single and multi-turn Question Answering. The whole dataset consists of 81,826 samples annotated by human English-speakers. We randomly shuffle these samples and use 80% of them (63,826) as training set, 10% as validation set (9,000), and the remaining 10% as test set (9,000). Note that for test set we only publish the questions without annotations, as this dataset is supporting an open evaluation now.

Each sample is a quadruple consists of:

- a question (or multiple questions for multi-turn QA)
- the logical form(s) representing the question(s)
- the parameters (entity/type/value) extracted from the question(s)
- the question type(s)

	TOTAL		TRAIN		DEV		TEST	
single-relation	34,316	0.419	26,955	0.422	3,727	0.414	3,634	0.404
multi-turn-entity	9,617	0.118	7,362	0.115	1,091	0.121	1,164	0.129
superlative	8,429	0.103	6,623	0.104	898	0.1	908	0.101
aggregation	7,710	0.094	5,871	0.092	906	0.101	933	0.104
multi-hop	7,452	0.091	5,938	0.093	780	0.087	734	0.082
cvt	5,115	0.063	3,849	0.06	619	0.069	647	0.072
yesno	2,688	0.033	2,086	0.033	300	0.033	302	0.034
multi-constraint	2,601	0.032	2,029	0.032	293	0.033	279	0.031
multi-choice	1,344	0.016	1,071	0.017	134	0.015	139	0.015
multi-turn-answer	1,304	0.016	1,068	0.017	106	0.012	130	0.014
multi-turn-predicate	893	0.011	706	0.011	100	0.011	87	0.01
comparative	357	0.004	268	0.004	46	0.005	43	0.005
	81,826		63,826		9,000		9,000	



The dataset contains **2,071** different knowledge graph predicates and **121** different entity types.

数据样例

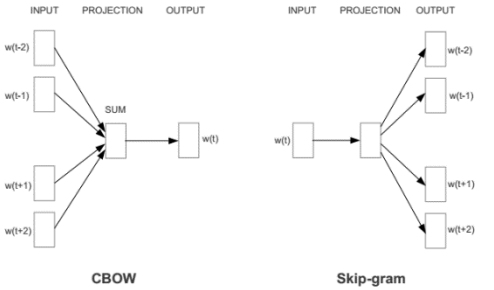
```
==================================================↓
<question id=8657>      who directed movie the disaster artist ( 2017 )↓
<logical form id=8657>  ( lambda ?x ( mso:film.film.director the_disaster_artist_(_2017_) ?x ) )↓
<parameters id=8657>    the_disaster_artist_(_2017_) (entity) [3,8]↓
<question type id=8657> single-relation↓
==========================================================================================↓
<question id=1094>      when was director of call me by your name born↓
<logical form id=1094>  ( lambda ?x exist ?y ( and ( mso:film.film.director call_me_by_your_name ?y ) ( mso:people.person.date_of_birth ?y ?x ) ) )↓
<parameters id=1094>    call_me_by_your_name (entity) [4,8]↓
<question type id=1094> multi-hop↓
==================================================↓
<question id=36502>     when is lionel morton birthday ||| in what films was this actor in?↓
<logical form id=36502> ( lambda ?x ( mso:people.person.date_of_birth lionel_morton ?x ) ) ||| ( lambda ?x ( mso:film.actor.film lionel_morton ?x ) )↓
<parameters id=36502>   lionel_morton (entity) [2,3] @Q1 ||| lionel_morton (entity) [2,3] @Q1↓
<question type id=36502>        multi-turn-entity↓
==================================================↓
<question id=40596>     is it true that the top elevation of madonna di campiglio is 3220.0 meters ?↓
<logical form id=40596> ( mso:skiing.ski_area.top_elevation madonna_di_campiglio 3220.0 )↓
<parameters id=40596>   madonna_di_campiglio (entity) [8,10] ||| 3220.0 (value) [12,12]↓
<question type id=40596>        yesno↓
==================================================↓
```

Experiments

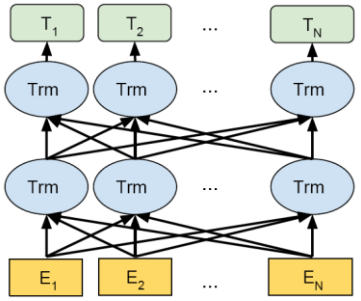
Results on test set	Seq2Seq (top-1)		Seq2Seq + Re-ranking (top-5)	
Question Type	**Number**	**ACC**	**Number**	**ACC**
single-relation	3,033/3,634	0.835	3199/3,634	0.880
multi-hop	631/734	0.860	611/734	0.832
superlative	635/908	0.699	710/908	0.782
multi-turn-entity	712/1,164	0.612	815/1,164	0.700
multi-turn-predicate	61/87	0.701	65/87	0.747
yesno	253/302	0.838	240/302	0.795
aggregation	614/933	0.658	453/647	0.700
cvt	416/647	0.643	655/933	0.702
multi-constraint	184/279	0.659	194/279	0.695
comparative	40/43	0.930	37/43	0.860
multi-turn-answer	46/130	0.354	56/130	0.431
multi-choice	46/139	0.331	62/139	0.446
TOTAL	6,671/9,000	0.741	7,097/9,000	**0.789 (+0.048)**

Comparison

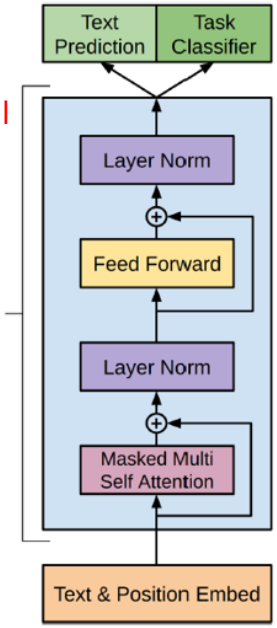
- Bi-directional language model pre-training
- More suitable to *matching tasks*
- Less parameters (BERT_BASE: 110M; BERT_LARGE: 340M)
- Achieve SOTA results on non-generative NLP tasks

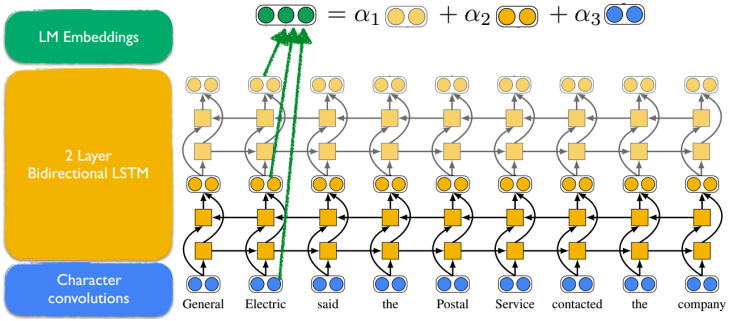


same idea, but more complicated network structure



from single-directional to bi-directional

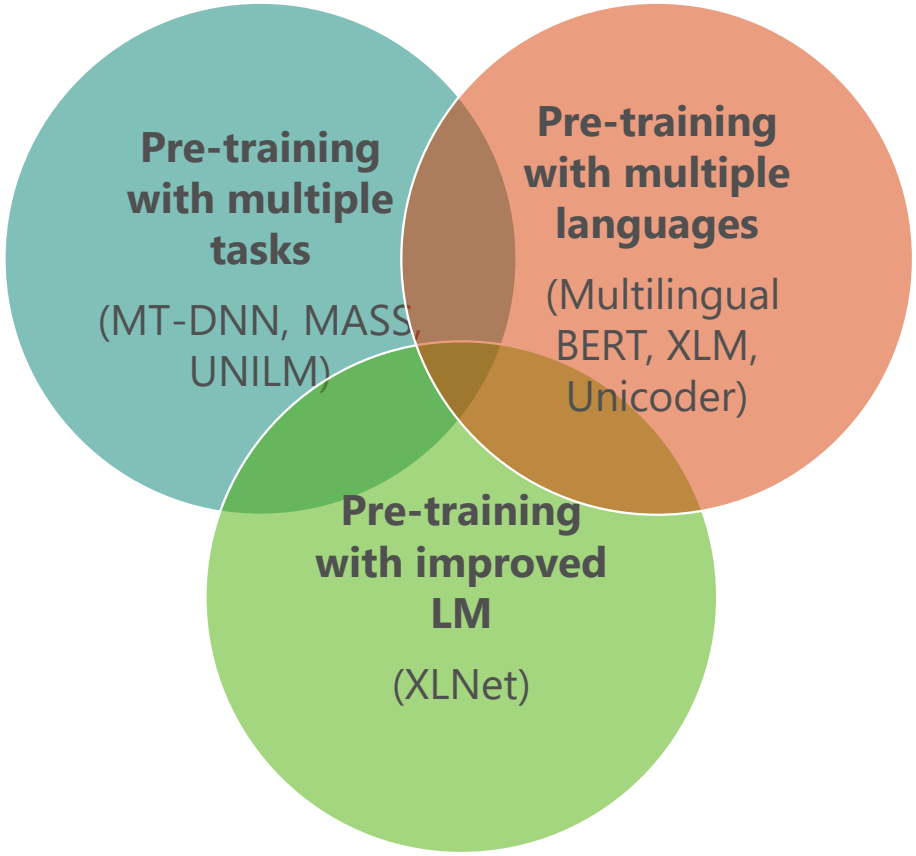




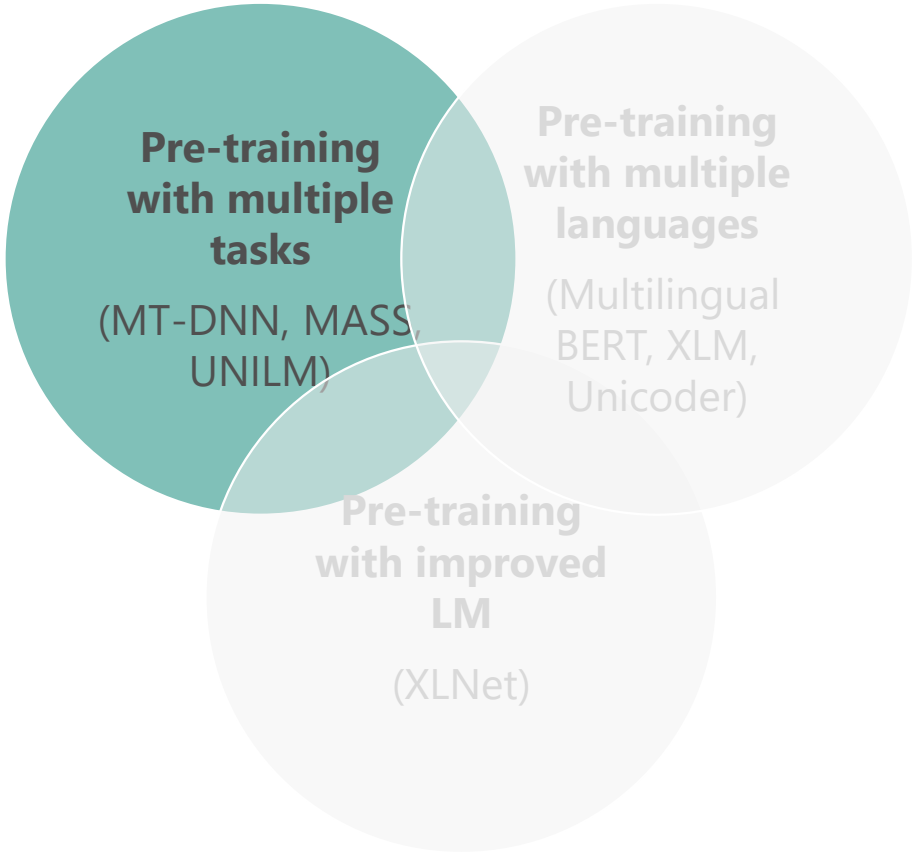
from LSTM to Transformer

- Uni-directional language model pre-training
- More suitable to *generation tasks*
- More parameters (GPT-2: 1.5B)
- Achieve SOTA results on zero/few-shot NLP tasks

More Pre-trained Models after BERT/GPT

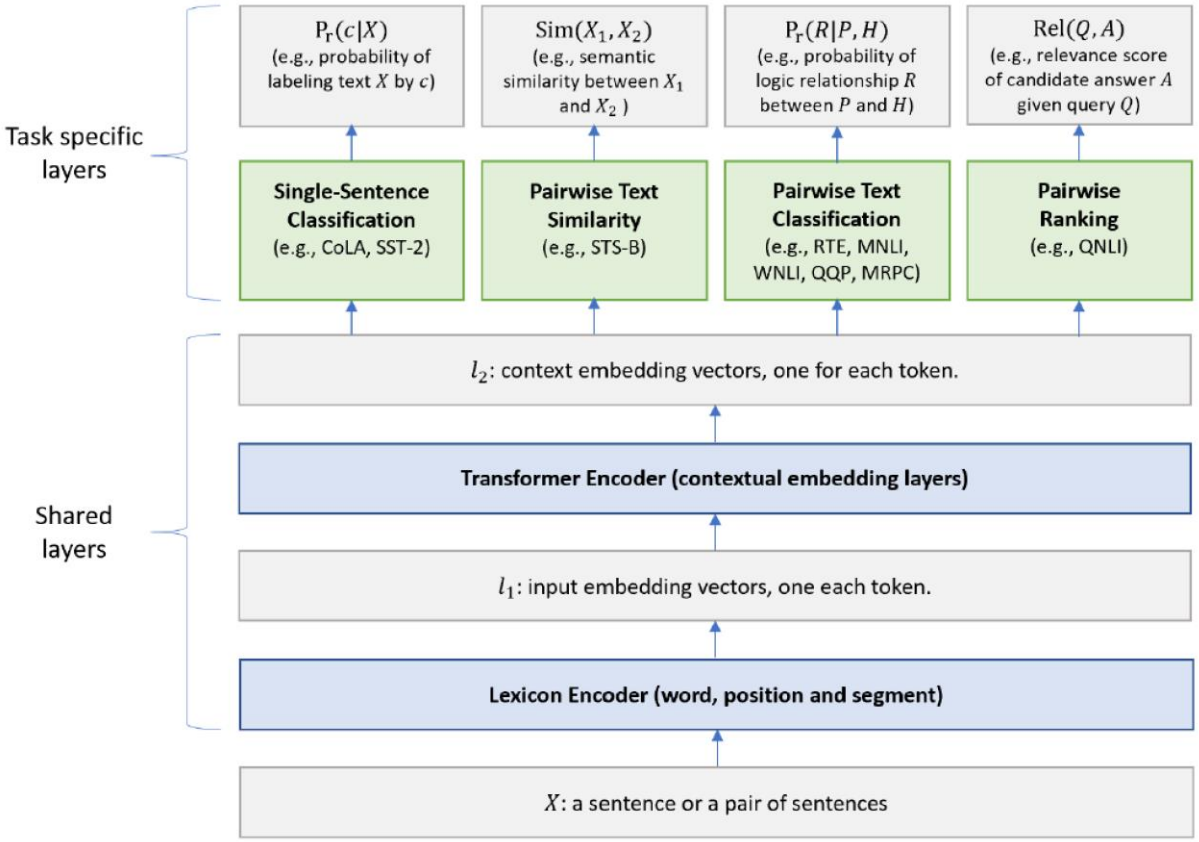


More Pre-trained Models after BERT/GPT



Pre-training with Multiple Tasks: MT-DNN

(Liu et al., 2019)

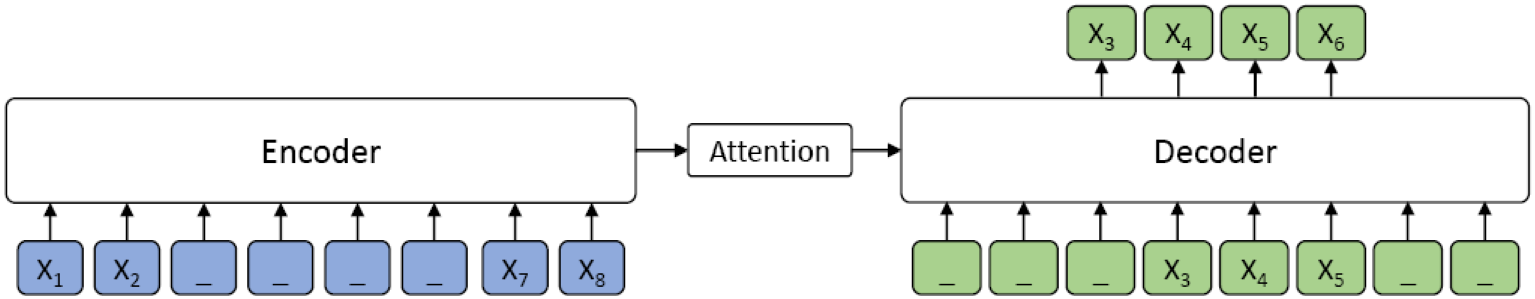


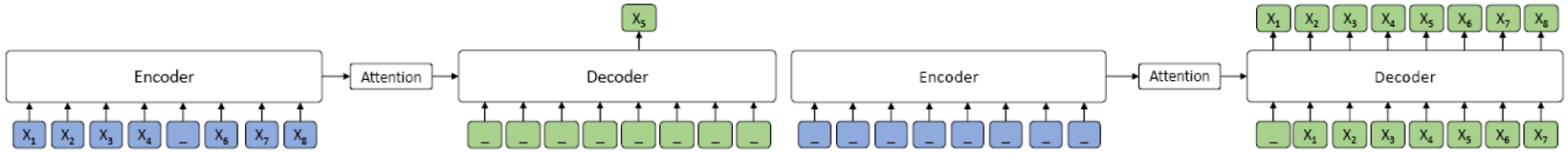
Algorithm 1: Training a MT-DNN model.

Initialize model parameters Θ randomly.
Pre-train the shared layers (i.e., the lexicon encoder and the transformer encoder).
Set the max number of epoch: $epoch_{max}$.
//Prepare the data for T tasks.
for t *in* $1, 2, ..., T$ **do**
  Pack the dataset t into mini-batch: D_t.
end
for $epoch$ *in* $1, 2, ..., epoch_{max}$ **do**
  1. Merge all the datasets:
    $D = D_1 \cup D_2 ... \cup D_T$
  2. Shuffle D
  for b_t *in* D **do**
    //b_t is a mini-batch of task t.
    3. Compute loss : $L(\Theta)$
      $L(\Theta)$ = Eq. 6 for classification
      $L(\Theta)$ = Eq. 7 for regression
      $L(\Theta)$ = Eq. 8 for ranking
    4. Compute gradient: $\nabla(\Theta)$
    5. Update model: $\Theta = \Theta - \epsilon\nabla(\Theta)$
  end
end

Pre-training with Multiple Tasks: MASS

(Song et al., 2019)



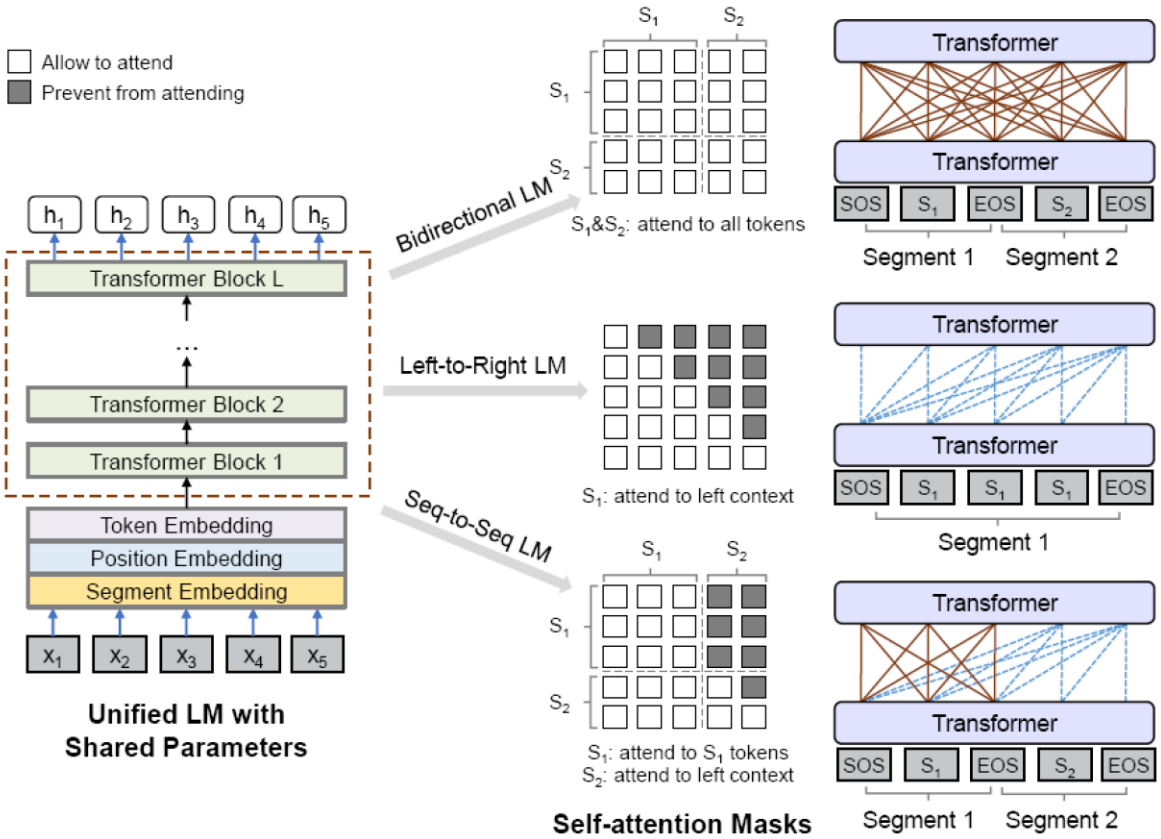


(a) Masked language modeling in BERT ($k = 1$)

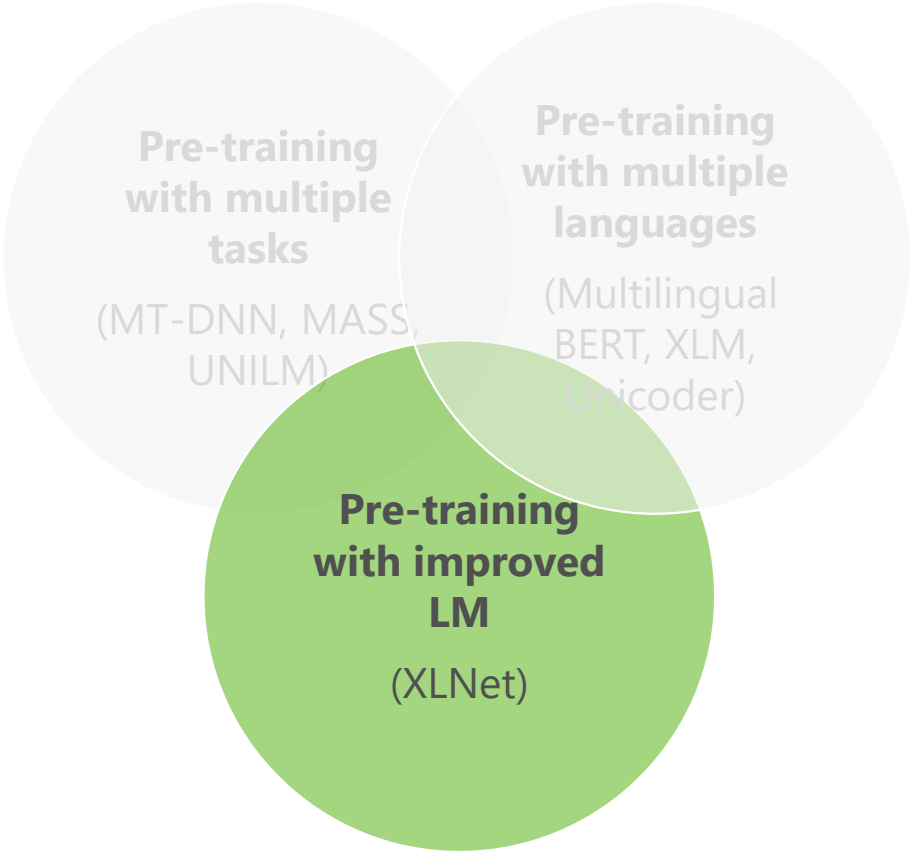
(b) Standard language modeling ($k = m$)

Pre-training with Multiple Tasks: UNILM

(Dong et al., 2019)

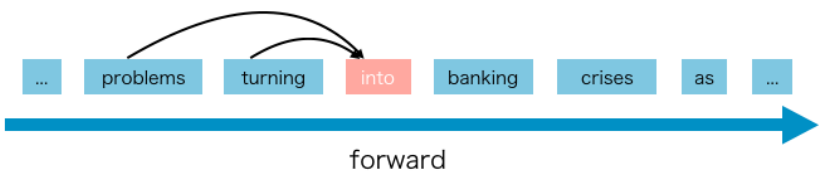


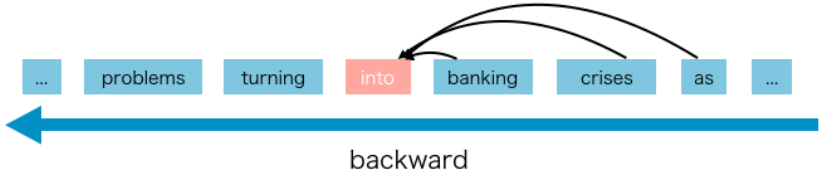
More Pre-trained Models after BERT/GPT



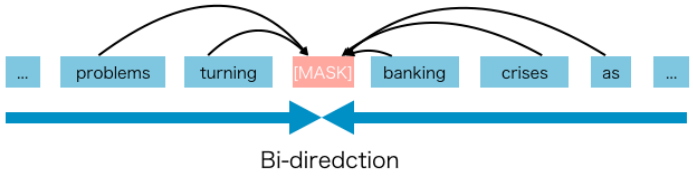
AutoRegressive (AR) LM & AutoEncoder (AE) LM

- **AR LM** aims to predict the next word using its forward or backward context
  - **Pros**: be good at generative NLP tasks
  - **Cons**: can't use forward and backward context at the same time





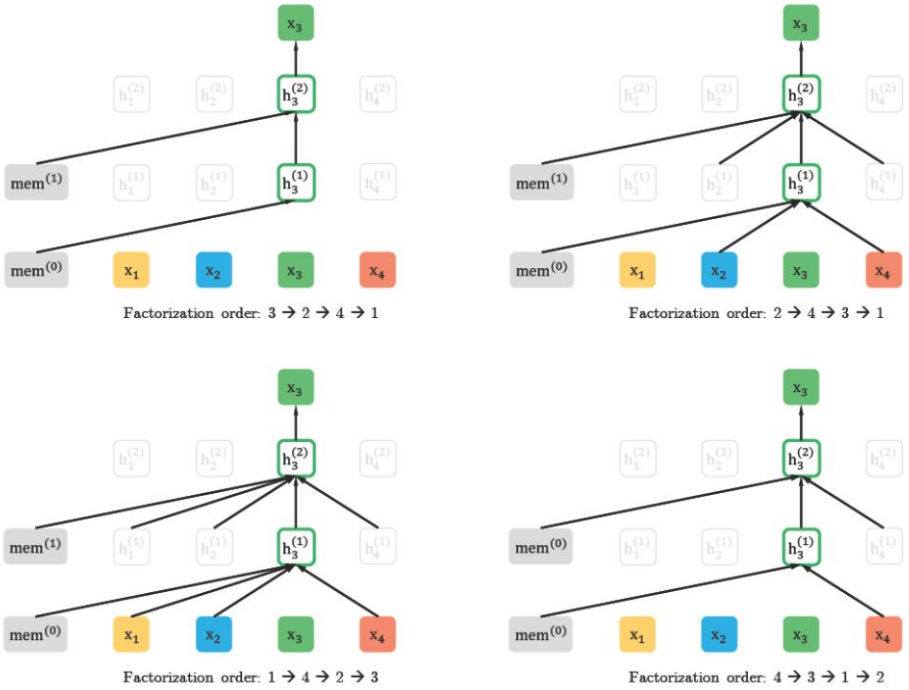
- **AE LM** aims to reconstruct the original data from corrupted input
  - **Pros**: can see the context on both forward and backward direction
  - **Cons**: use MASK in the pretraining, which are unseen in fine-tuning, resulting in a pretrain-finetune discrepancy; assume the predicted (masked) tokens are independent of each other given the unmasked tokens



Pre-training with Improved LM: XLNet

(Yang et al., 2019)

- **XLNet** is a generalized AE LM, which learns from bi-directional context to avoid disadvantages brought by the MASK method in the original AE language model (i.e. BERT).



$$\log p(\text{New York} \mid \text{is a city})$$

Input of BERT: [mask] [mask] is a city

$$\mathcal{J}_{\text{BERT}} = \log p(\text{New} \mid \text{is a city}) + \log p(\text{York} \mid \text{is a city}),$$

$$\mathcal{J}_{\text{XLNet}} = \log p(\text{New} \mid \text{is a city}) + \log p(\text{York} \mid \text{New}, \text{is a city})$$

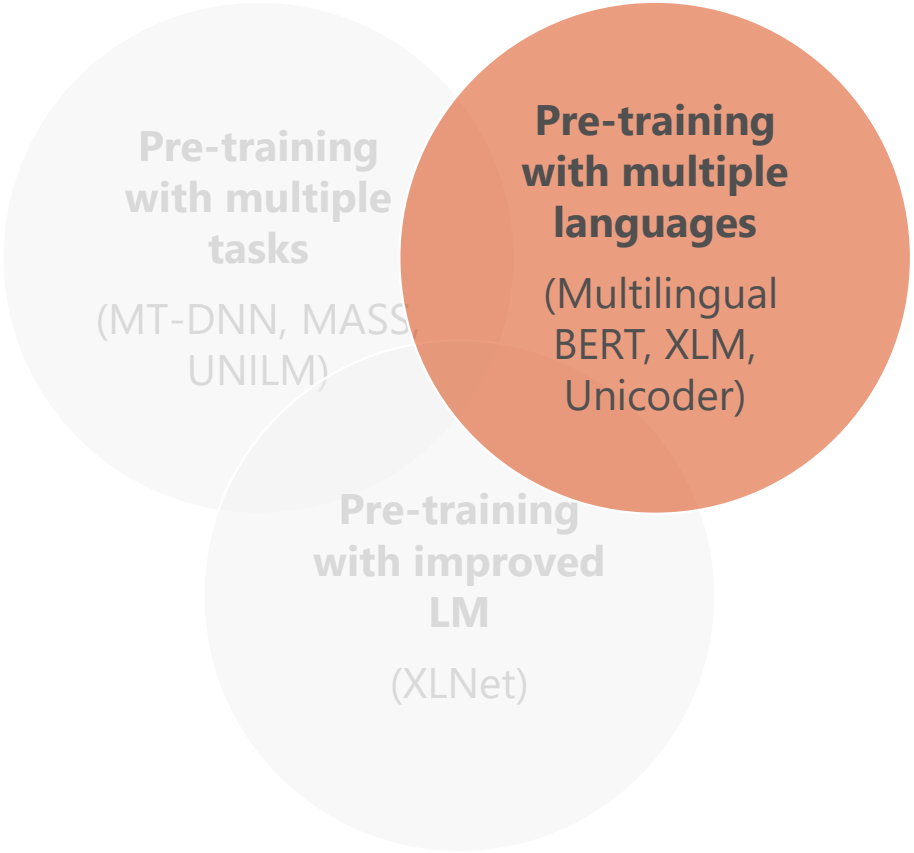
Input of XLNet: is a city New York

Evaluation

SQuAD1.1	EM	F1	SQuAD2.0	EM	F1
Dev set results without data augmentation					
BERT [10]	84.1	90.9	BERT† [10]	78.98	81.77
XLNet	**88.95**	**94.52**	XLNet	**86.12**	**88.79**
Test set results on leaderboard, with data augmentation (as of June 19, 2019)					
Human [27]	82.30	91.22	BERT+N-Gram+Self-Training [10]	85.15	87.72
ATB	86.94	92.64	SG-Net	85.23	87.93
BERT* [10]	87.43	93.16	BERT+DAE+AoA	85.88	88.62
XLNet	**89.90**	**95.08**	XLNet	**86.35**	**89.13**

Model	MNLI	QNLI	QQP	RTE	SST-2	MRPC	CoLA	STS-B	WNLI
Single-task single models on dev									
BERT [2]	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-
XLNet	**89.8/-**	**93.9**	**91.8**	**83.8**	**95.6**	**89.2**	**63.6**	**91.8**	-
Single-task single models on test									
BERT [10]	86.7/85.9	91.1	89.3	70.1	94.9	89.3	60.5	87.6	65.1
Multi-task ensembles on test (from leaderboard as of June 19, 2019)									
Snorkel* [29]	87.6/87.2	93.9	89.9	80.9	96.2	91.5	63.8	90.1	65.1
ALICE*	88.2/87.9	95.7	**90.7**	83.5	95.2	92.6	**68.6**	91.1	80.8
MT-DNN* [18]	87.9/87.4	96.0	89.9	**86.3**	96.5	92.7	68.4	91.1	89.0
XLNet*	**90.2/89.7**†	**98.6**†	90.3†	**86.3**	**96.8**†	**93.0**	67.8	**91.6**	**90.4**

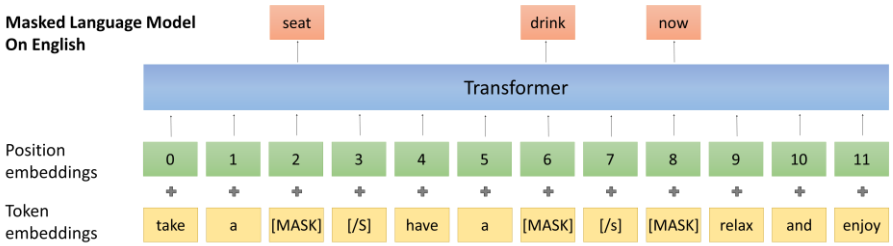
More Pre-trained Models after BERT/GPT

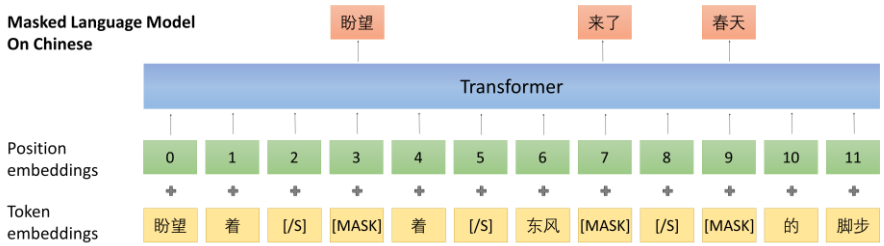


Pre-training with Multiple Languages: Multilingual BERT

(Devlin et al., 2018)

- Training objective
  1. Masked language model



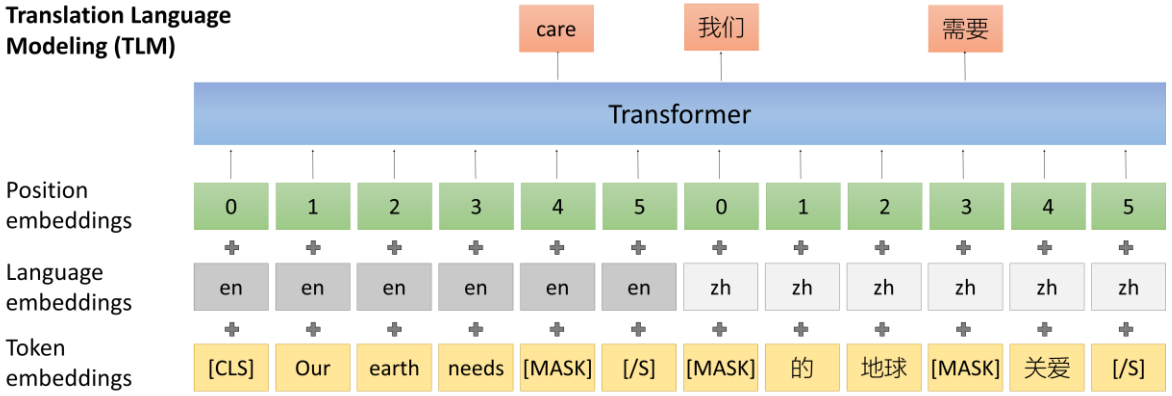


Share vocabulary and parameters for all **104 languages**.

Pre-training with Multiple Languages: XLM

(Lample and Conneau, 2019)

- Training objective
  1. Masked language model
  2. Translation language model



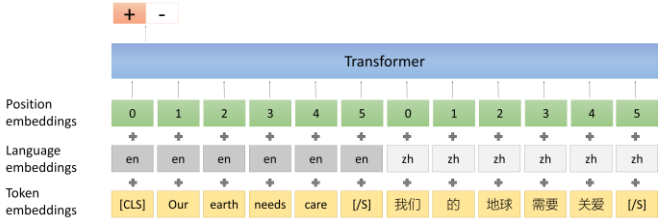
English	French	Spanish
German	Greek	Bulgarian
Russian	Turkish	Arabic
Vietnamese	Thai	Chinese
Hindi	Swahili	Urdu

Pre-training with Multiple Languages: Unicoder

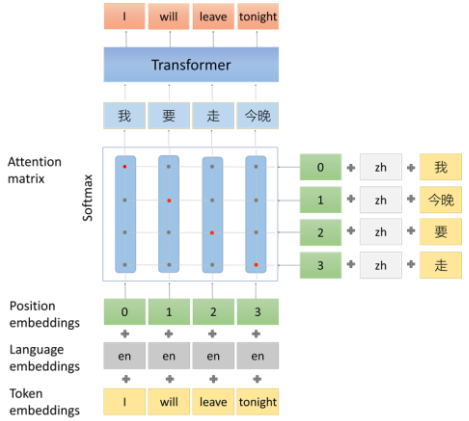
(Huang et al., 2019)

- Training objective
  1. Masked language model
  2. Translation language model
  3. Cross-lingual masked language model
  4. Cross-lingual paraphrase classification
  5. Cross-lingual word recovery

(a) cross-lingual masked language model



(b) cross-lingual paraphrase classification



(c) cross-lingual word recovery

Evaluation

- **Task**: predict the entailment relation between two sentences
- **Dataset**: XNLI by Conneau et al. (2018) (https://www.nyu.edu/projects/bowman/xnli/)

Language	Premise / Hypothesis	Label
English	You don't have to stay there. You can leave.	Entailment
French	La figure 4 montre la courbe d'offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Entailment
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.	Entailment
German	Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod. Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an.	Neutral
Swahili	Ni silaha ya plastiki ya moja kwa moja inayopiga risasi. Inadumu zaidi kuliko silaha ya chuma.	Neutral
Russian	И мы занимаемся этим уже на протяжении 85 лет. Мы только начали этим заниматься.	Contradiction
Chinese	让我告诉你，美国人最终如何看待你作为独立顾问的表现。 美国人完全不知道您是独立律师。	Contradiction
Arabic	تحتاج الوكالات لأن تكون قادرة على قياس مستويات النجاح. لا يمكنالوكالات أ اتعرف ما إذا كانت ناجحة أم لا	Contradiction

English	French	Spanish
German	Greek	Bulgarian
Russian	Turkish	Arabic
Vietnamese	Thai	Chinese
Hindi	Swahili	Urdu

Language	Train	Validation	Test
English	393K	2,500	5,000
14 languages	0	2,500	5,000

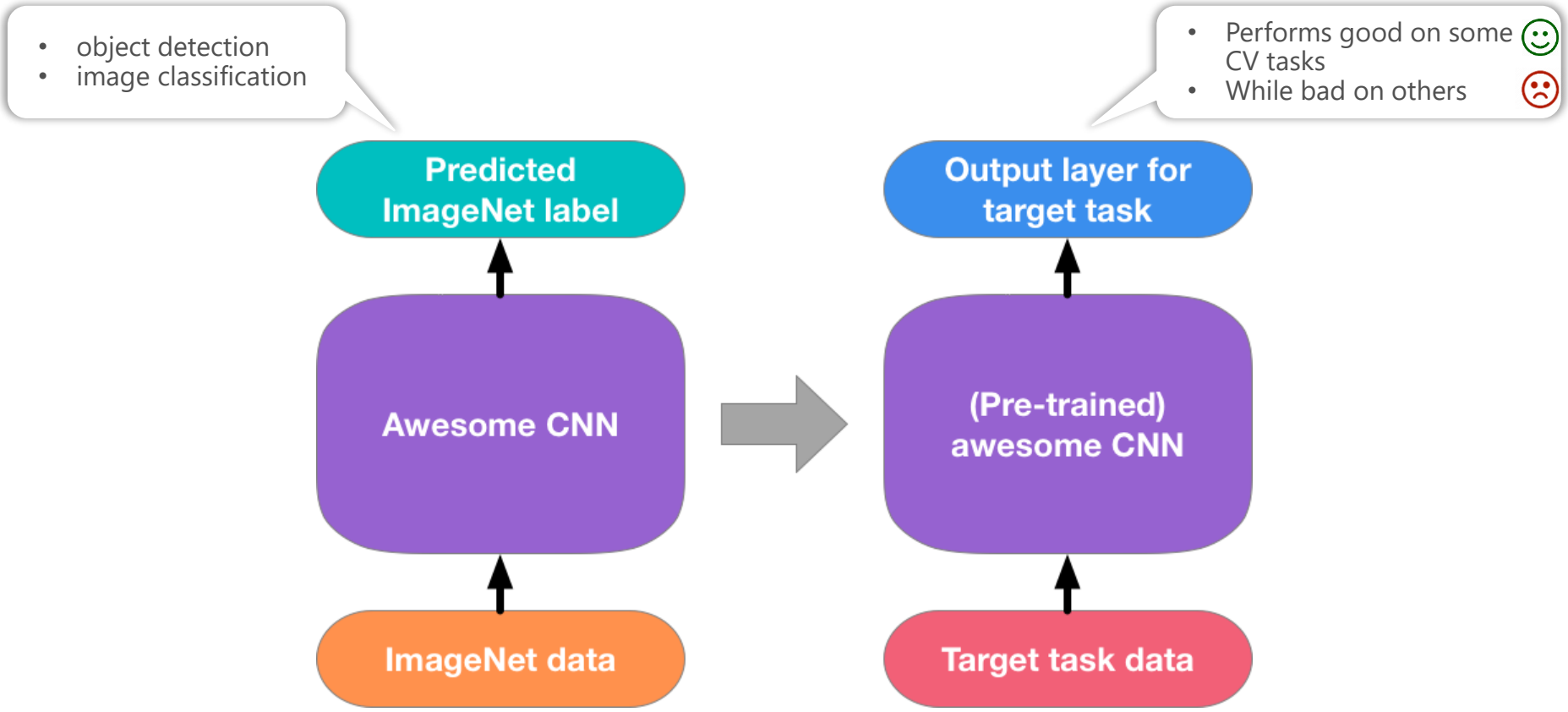
- **Result**: compare to Multilingual BERT (Devlin et al., 2018) and XLM (Lample and Conneau, 2019)

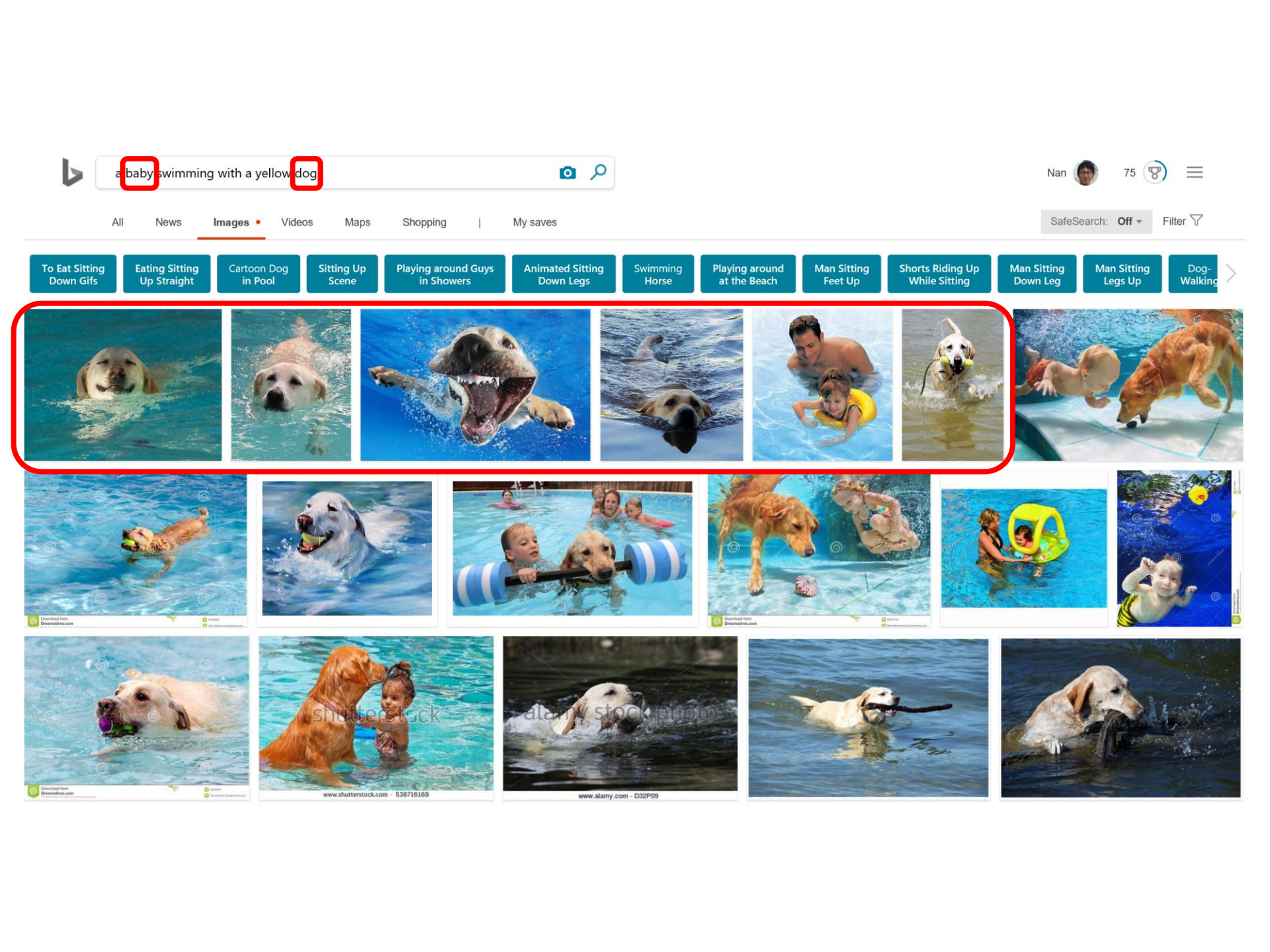
Setting	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	avg
Multilingual BERT	82.1	76.9	78.5	74.8	72.1	75.4	74.3	70.6	70.8	67.8	63.2	76.2	65.3	65.3	60.6	71.6
XLM	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Unicoder	**85.6**	**81.1**	**82.3**	**80.9**	**79.5**	**81.4**	**79.7**	**76.8**	**78.2**	**77.9**	**77.1**	**80.5**	**73.4**	**73.8**	**69.6**	**78.5**

Outline

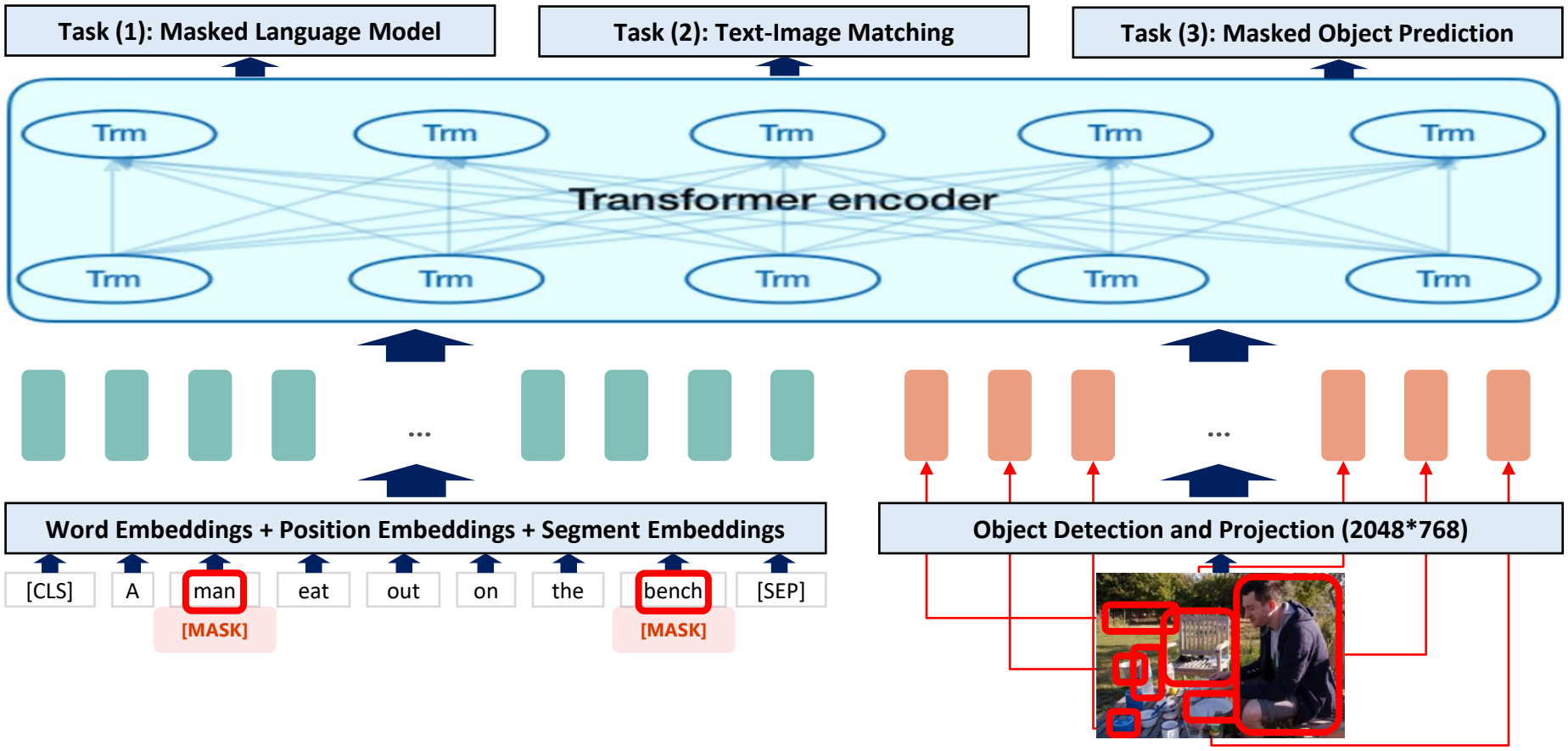
- Pre-training in NLP
- Pre-training in Language + Vision

Pre-training with ImageNet




a baby swimming with a yellow dog
Nan
75
All
News
Images
Videos
Maps
Shopping
My saves
SafeSearch: Off
Filter
To Eat Sitting Down Gifs
Eating Sitting Up Straight
Cartoon Dog in Pool
Sitting Up Scene
Playing around Guys in Showers
Animated Sitting Down Legs
Swimming Horse
Playing around at the Beach
Man Sitting Feet Up
Shorts Riding Up While Sitting
Man Sitting Down Leg
Man Sitting Legs Up
Dog-Walking
shutterstock
www.shutterstock.com · 538716169
alamy stock photo
www.alamy.com · D32P09

ImageBERT (version 1.0)



Experiment

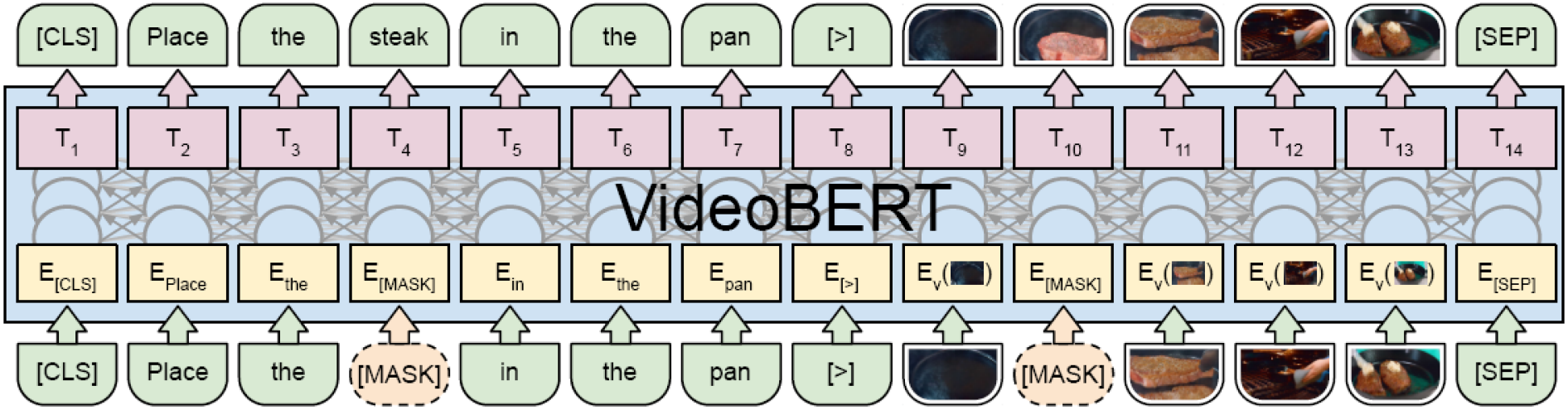
MS-COCO 1K	Text-to-Image Retrieval			Image-to-Text Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Two Branch Network (Wang et al., 2018)	43.3	76.8	87.6	54.0	84.0	91.2
SCAN (Lee et al., 2018)	58.8	88.4	94.8	72.7	94.8	98.4
Scene Concept Graph (Shi et al., 2019)	61.4	88.9	95.1	76.6	96.3	**99.2**
Cross-lingual Pre-training (Ours)	**68.5**	**92.7**	**96.9**	**82.6**	**96.8**	98.2

Flickr 1K	Text-to-Image Retrieval			Image-to-Text Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Two Branch Network (Wang et al., 2018)	28.4	56.3	67.4	37.5	64.7	75.0
SCAN (Lee et al., 2018)	48.6	77.7	85.2	67.4	90.3	95.8
Scene Concept Graph (Shi et al., 2019)	49.3	76.4	85.6	71.8	90.8	94.8
Cross-lingual Pre-training (Ours)	**64.1**	**88.2**	**93.5**	**79.7**	**94.8**	**97.4**

VideoBERT

(Sun et al., 2019)

A joint visual-linguistic model to learn high-level features without any explicit supervision.



Language Preprocessing

- Break each ASR sequence into sentences by adding punctuations
- Tokenize each sentence into WordPieces

Video Preprocessing

- Sample frames at 20 fps
- Create clips from 30-frame (1.5 seconds) over the video
- Apply S3D to extract features from each video clip
- Generate “visual words” based on the extracted features

Experiment (on YouCookII dataset)

Method	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Zhou *et al.* [39]	-	1.42	11.20	-	-
S3D [34]	6.12	3.24	10.00	26.05	0.35
VideoBERT	6.80	4.07	10.99	27.51	0.50
VideoBERT + S3D	**7.81**	**4.52**	**11.85**	**28.78**	**0.55**

Table 3: Video captioning performance on YouCook II. We follow the setup from [39] and report captioning performance on the validation set, given ground truth video segments. Higher numbers are better.

GT: add some chopped basil leaves into it

VideoBERT: chop the basil and add to the bowl

S3D: cut the tomatoes into thin slices

GT: cut the top off of a french loaf

VideoBERT: cut the bread into thin slices

S3D: place the bread on the pan

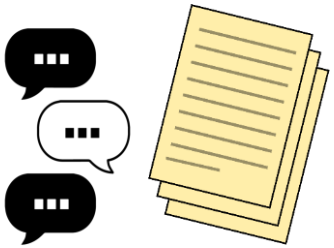
Outline

- Pre-training in NLP
- Pre-training in Language + Vision

Summary (Question Answering)

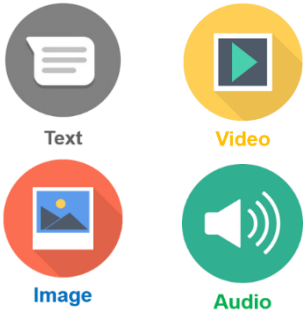
- Low-resource
- Reasoning
- Commonsense
- Multi-turn
- Multi-modal

Summary (Pre-training)

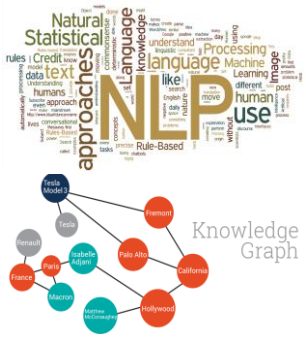


Cross-(longer) Context

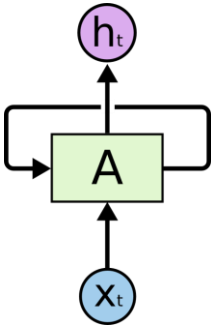
Cross-Language



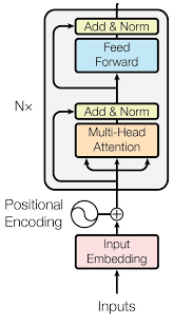
Cross-Modality



Cross-Heterogeneous Data



RNN



Transformer

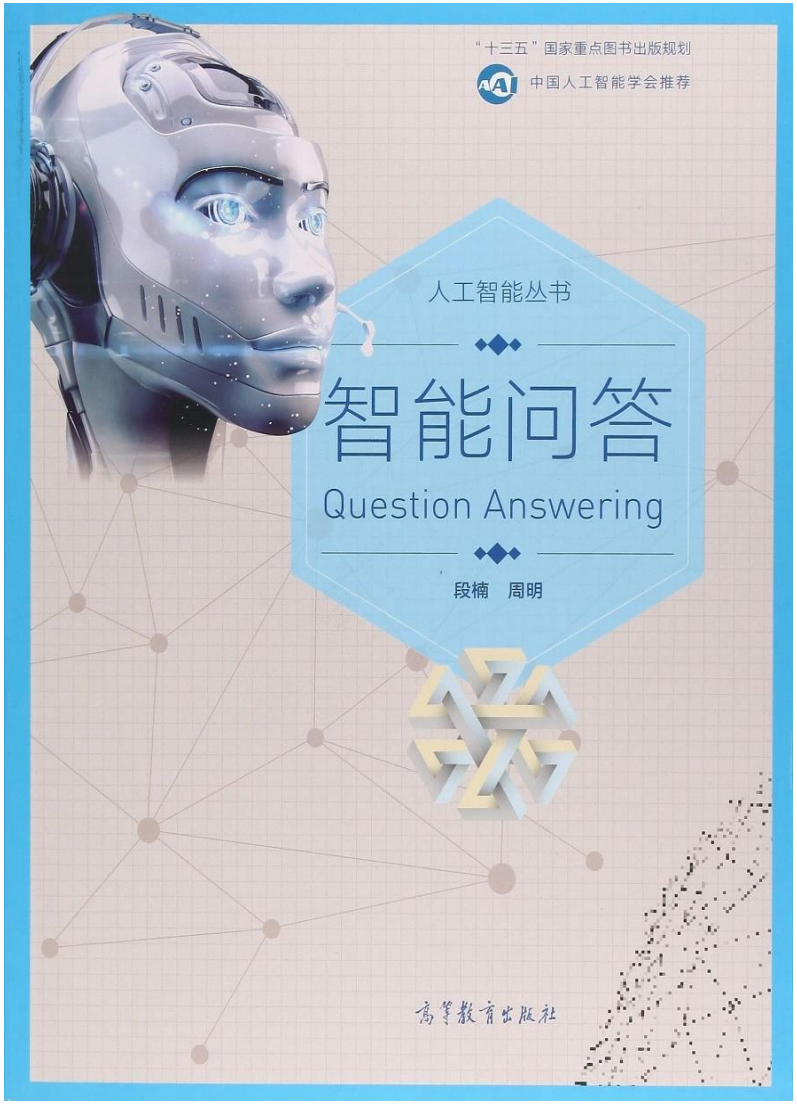
What's next?

- **AR LM**
- **AE LM**

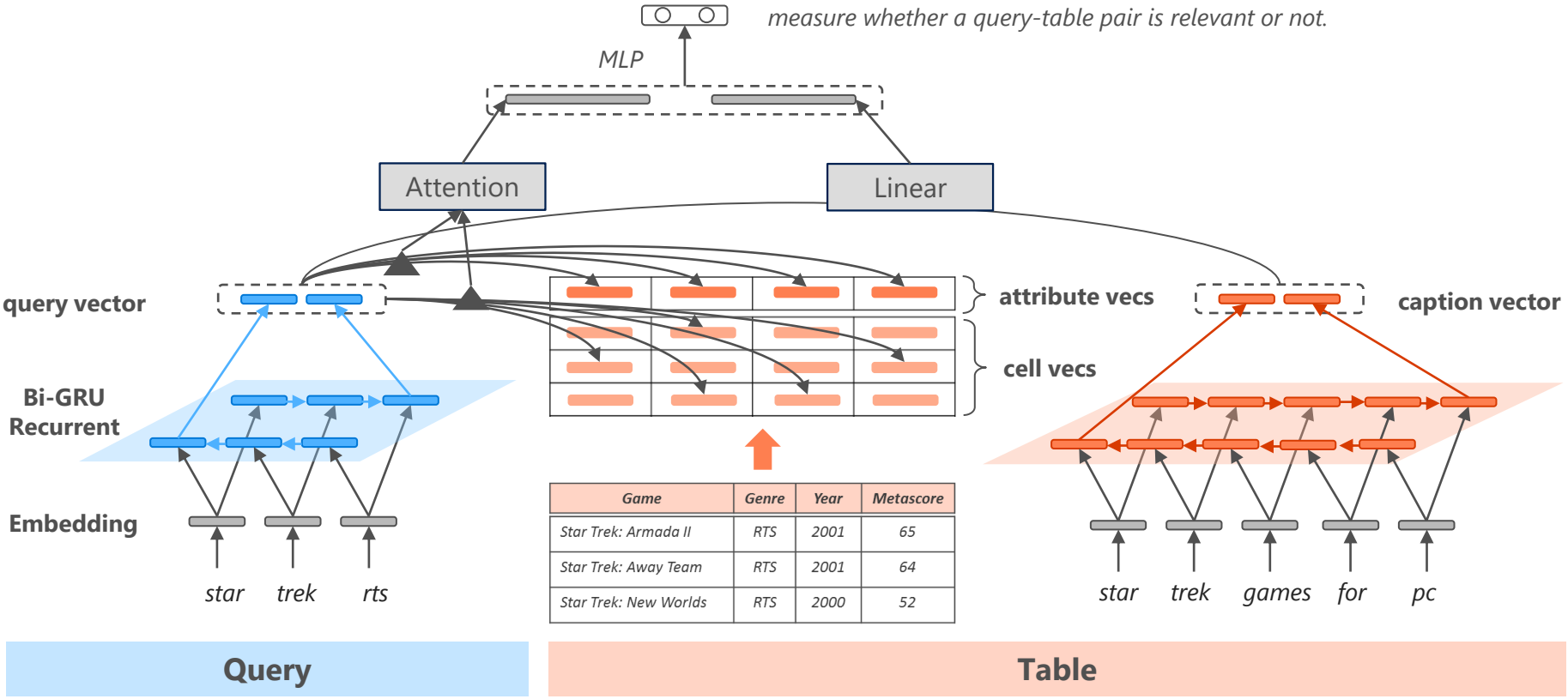
What's next?

Thanks!

And here are two NLP books by MSRA-NLC group ^_^

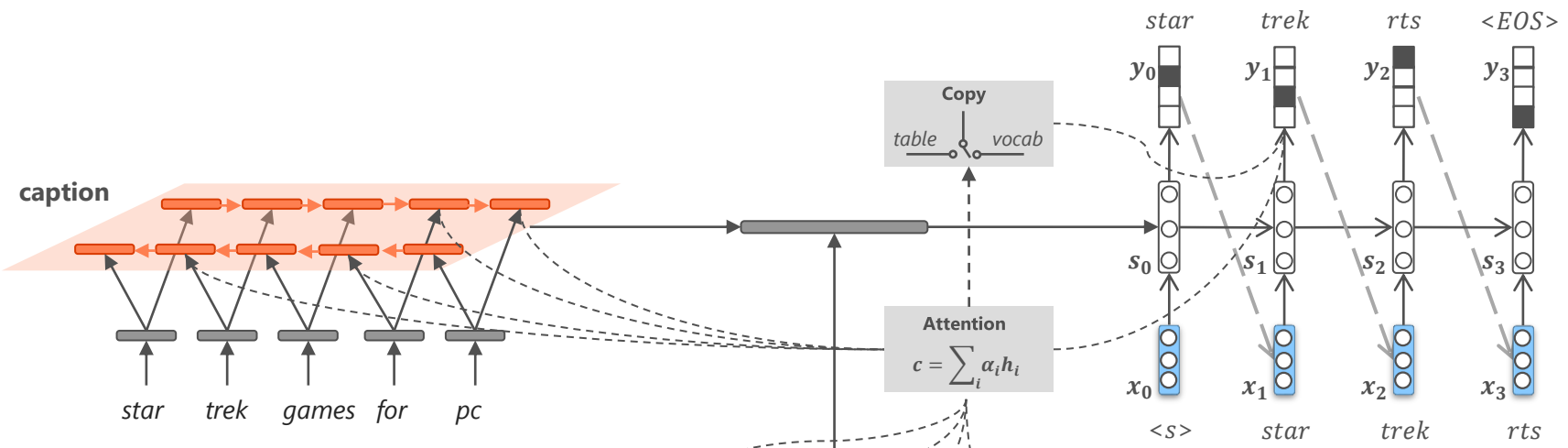


Background: NN based TableQA model



Game	Genre	Year	Metascore
Star Trek: Armada II	RTS	2001	65
Star Trek: Away Team	RTS	2001	64
Star Trek: New Worlds	RTS	2000	52

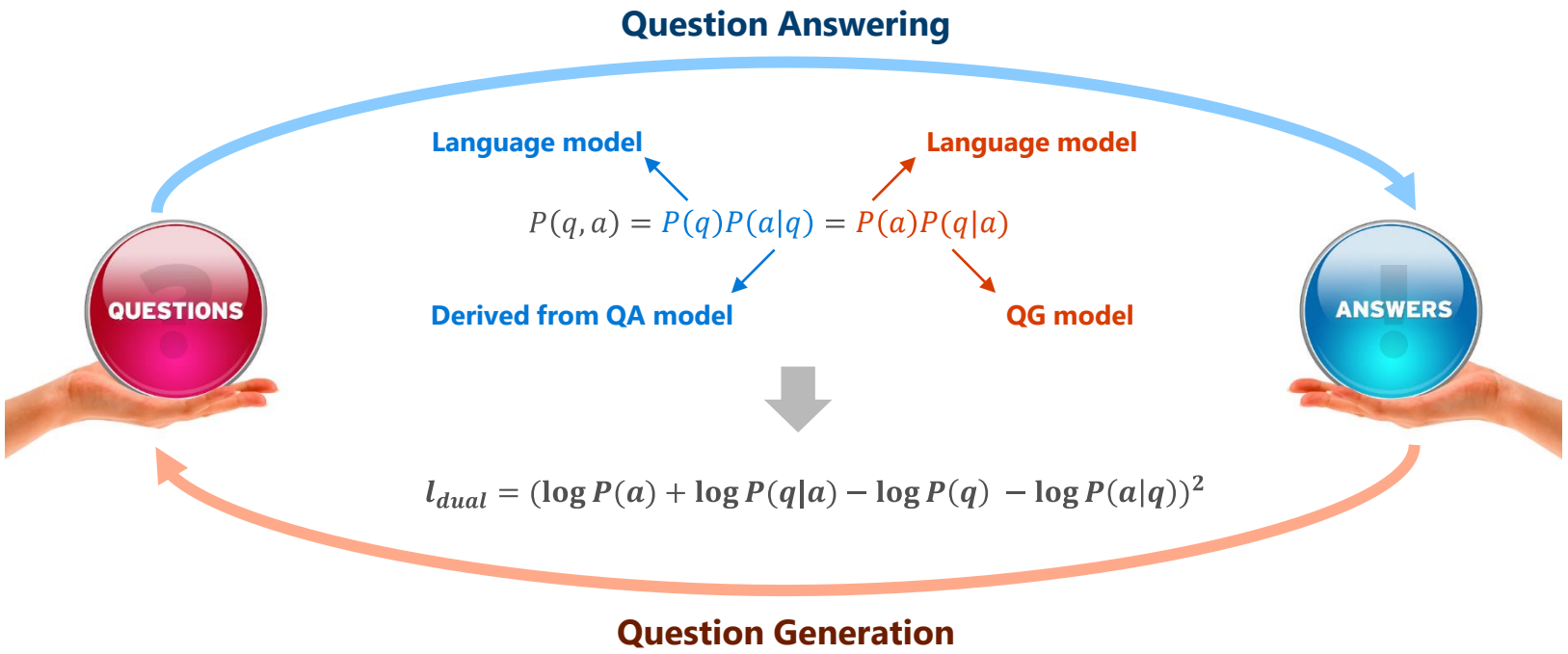
Background: NN based TableQG Model



Game	Genre	Year	Metascore
Star Trek: Armada II	RTS	2001	65
Star Trek: Away Team	RTS	2001	64
Star Trek: New Worlds	RTS	2000	52

Table **Query**

Training QA and QG with Dual Supervised Learning

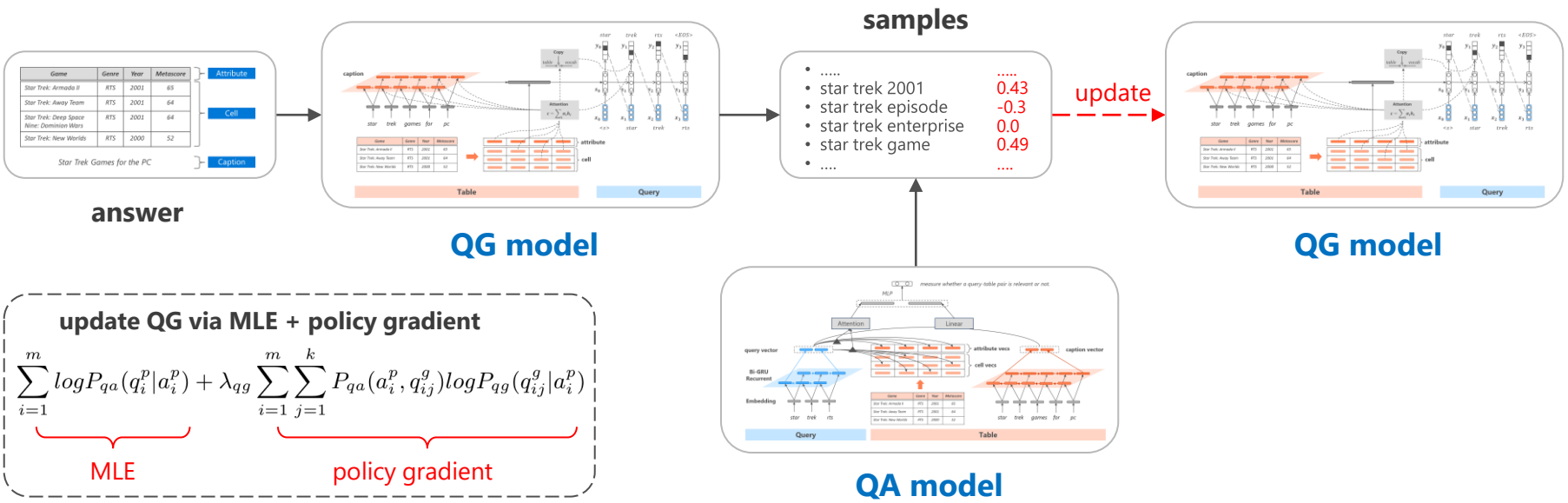


$$P(a|q) = \frac{\exp(f(q,a|\theta_{QA})}{\exp(f(q,a|\theta_{QA}) + \sum_{\tilde{q}} \exp(f(\tilde{q},a|\theta_{QA})}$$

Training QA and QG with Generative Collaborative Nets

(a) QA improves QG

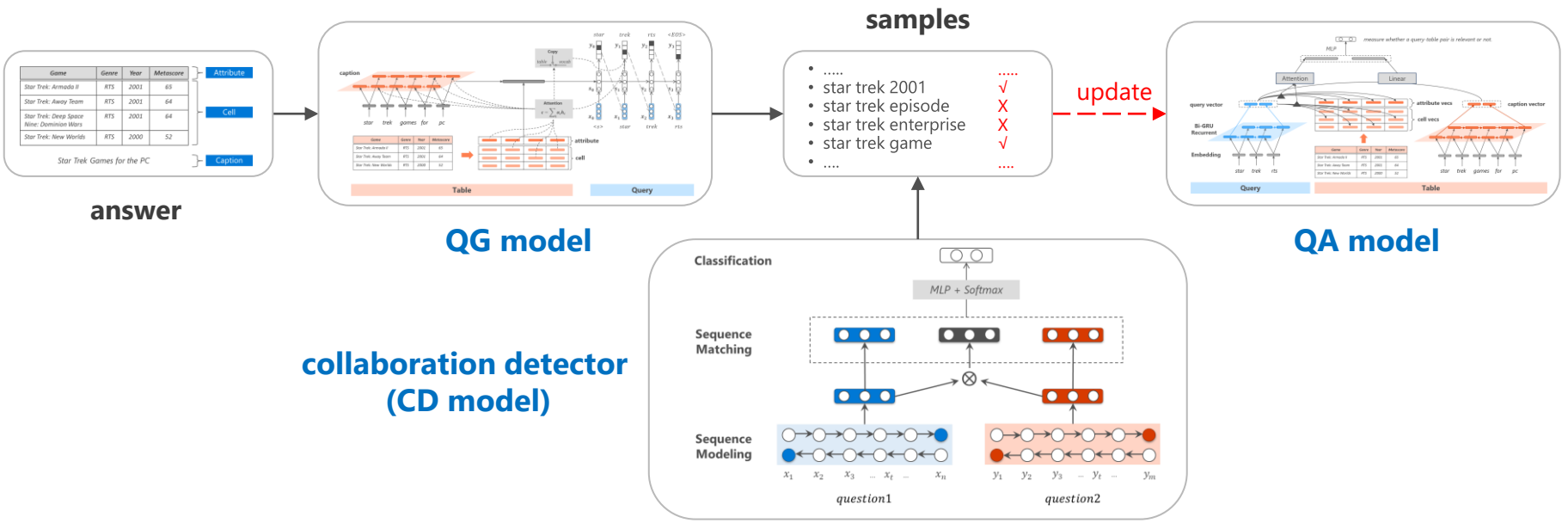
- The intuition is that QA and QG could improve each other
  - **QA improves QG**: incorporating additional QA-specific signal as the loss function for QG



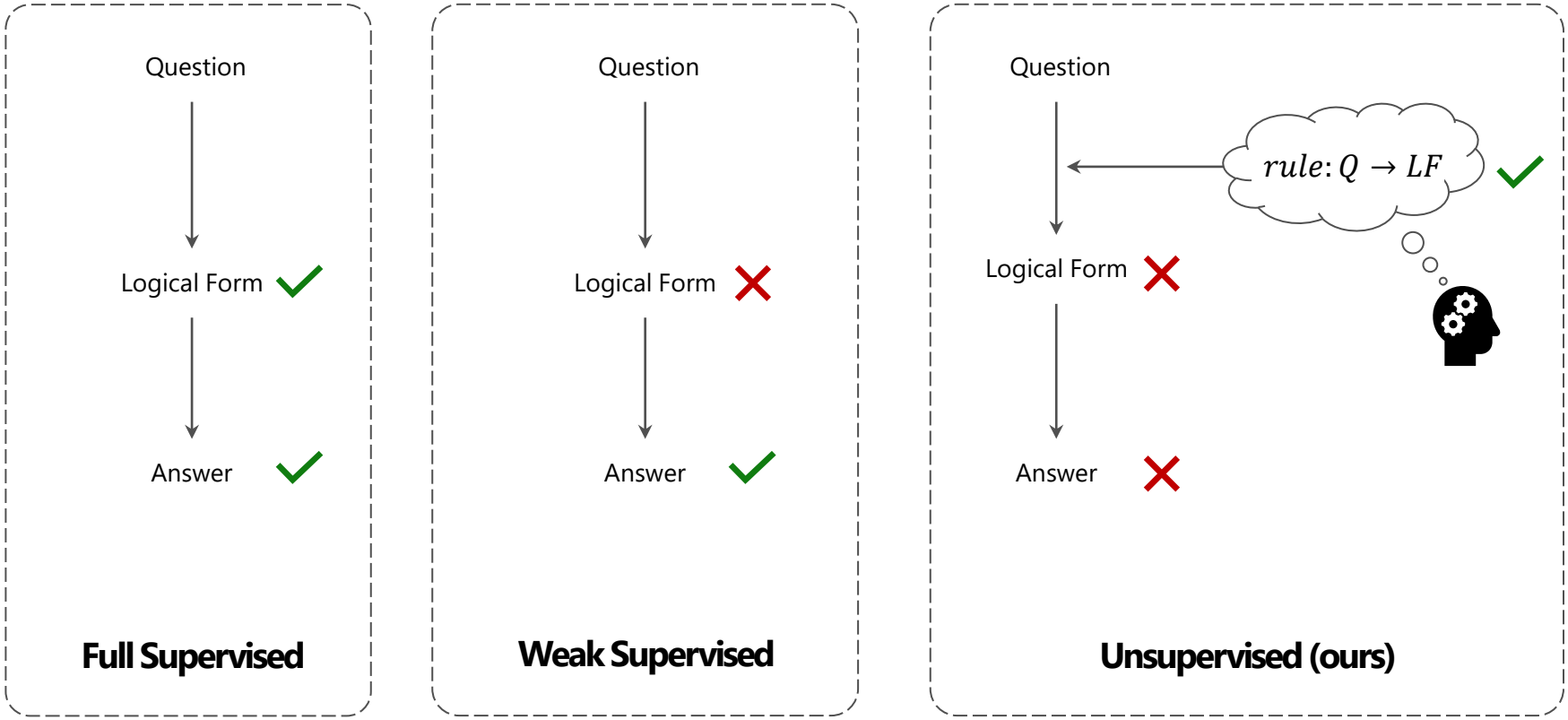
Training QA and QG with Generative Collaborative Nets

- The intuition is that QA and QG could improve each other
  - **QA improves QG**: incorporating additional QA-specific signal as the loss function for QG
  - **QG improves QA**: adding artificially generated training instances for QA

(b) QG improves QA



Learning from Rules



Back-Translation + Meta-Learning

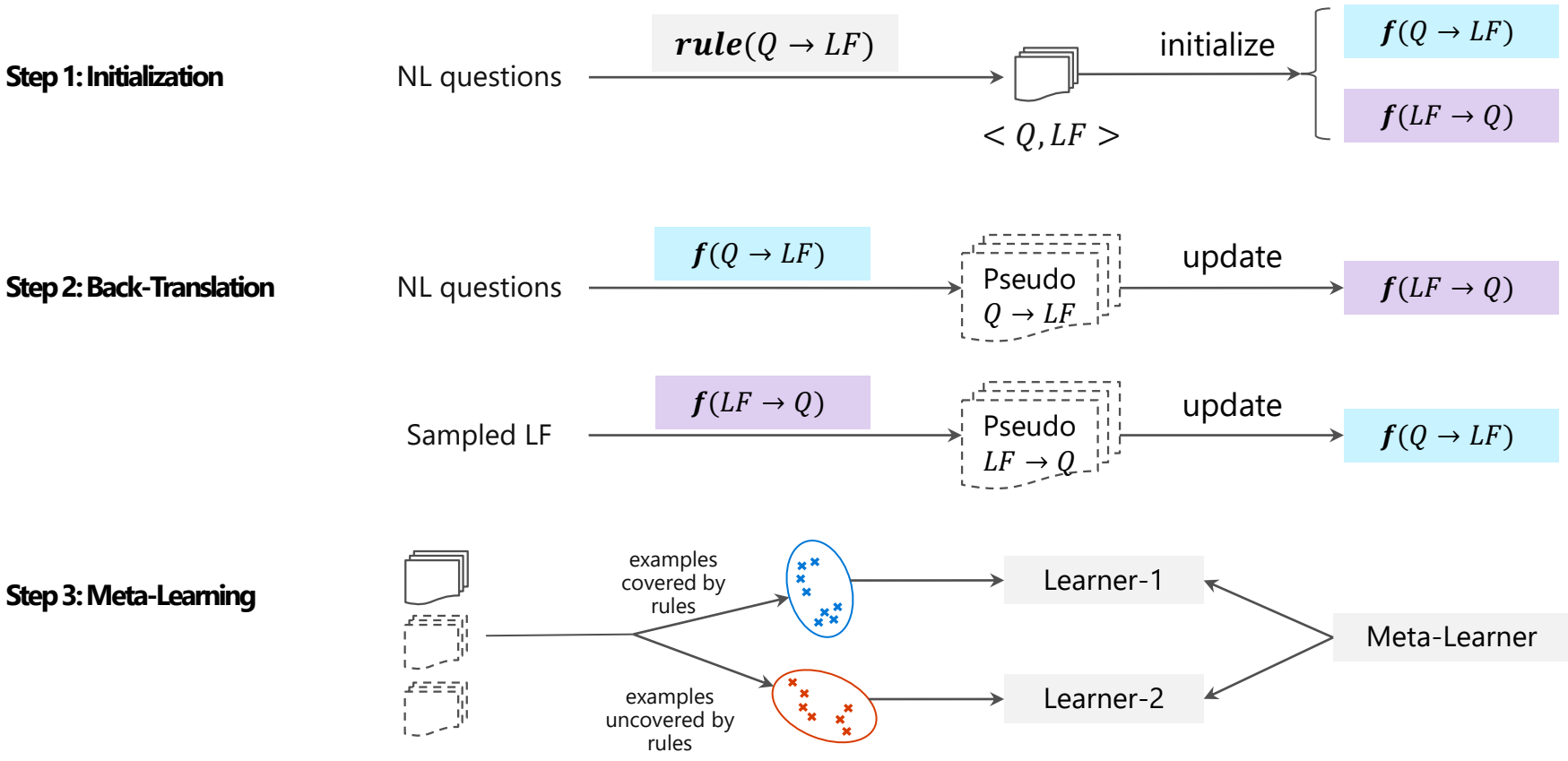


Table-based Semantic Parsing

- Execution accuracy on WikiSQL dataset (Salesforce Research, 2017)

Method	Supervision	Execution ACC
MSRA-NLC @ACL2018	Supervised	74.6%
Dong+ @ACL2018	Supervised	78.5%
Google Brain @NIPS2018	**Weak Supervised**	**72.4%**
Our base model	Supervised	82.3%
Rule	Unsupervised	62.8%
Rule + Data Combination	Unsupervised	70.3%
Rule + Self Training	Unsupervised	70.3%
Rule + Back Translation	Unsupervised	72.3%
Rule + Back Translation+ MAML (ours)	**Unsupervised**	**72.7%**