Question Answering

FUNCUP

Chi-Liang Liu

2019/11/16

National Taiwan University

Outline

1. Introduction
2. Machine Comprehension by Deep Learning
3. Conclusion

Introduction

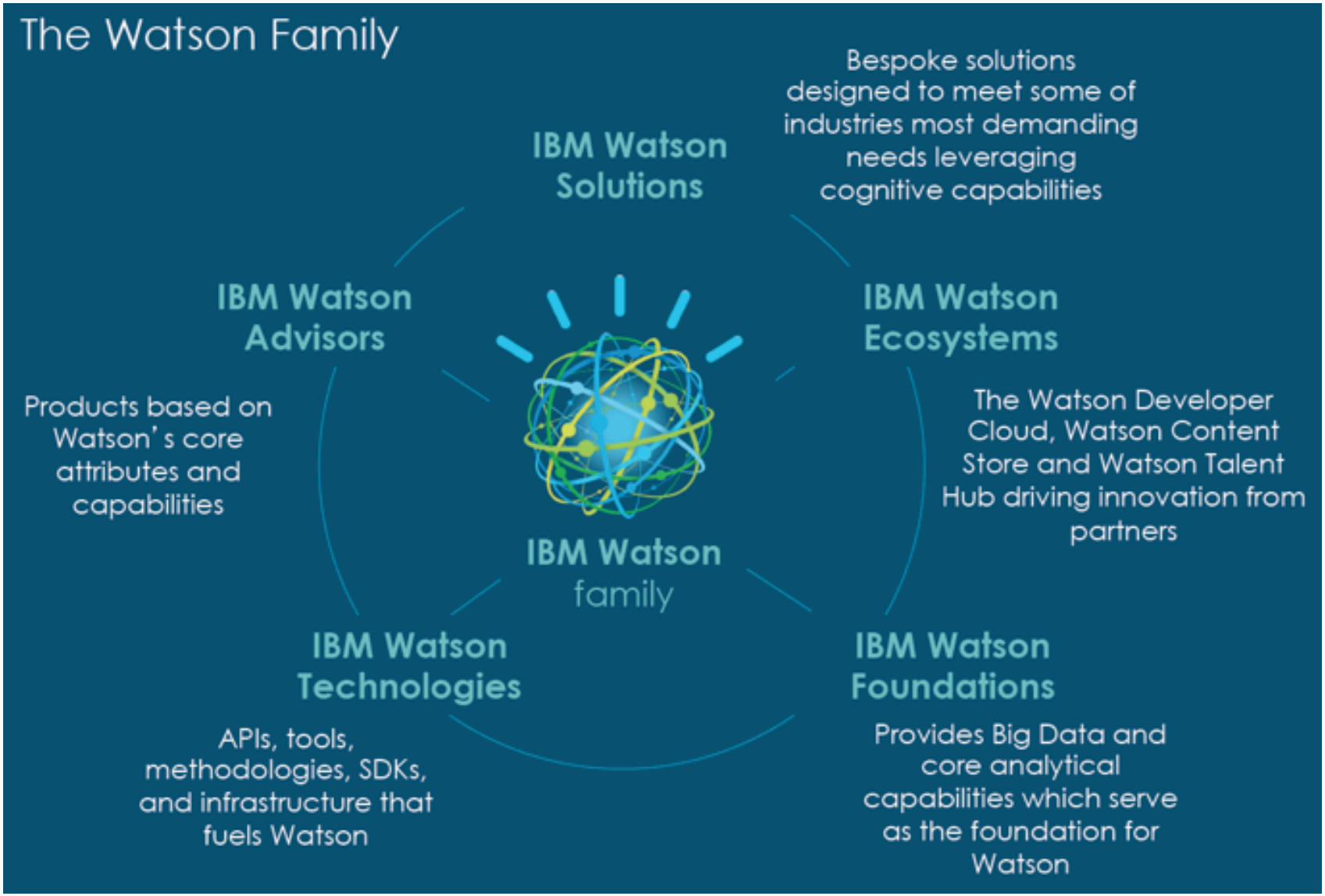
What is Question Answering?

- Answer particular questions
- Colloquial approach
- One of the oldest NLP problem
- A way to evaluate **Reading Comprehension**
  - ”Since questions can be devised to query **any aspect** of text comprehension, the ability to answer questions is **the strongest possible demonstration of understanding**.” Wendy Lehnert, 1977

Example - Google

Example - IBM Waston

Won Jeopardy on February 16, 2011!



Example - Siri

Look Good?

人工智慧?

or

工人智慧?

Type

- **Single** vs Multiple
- **Simple** vs Complex
- **Text** vs Visual
- **Open-domain** vs Closed-domain
- **IR-based** vs Knowledge-based

Introduction

Single vs Multiple

Simple

- A single document Q/A task involves questions associated with one particular document.
- In most cases, the assumption is that **the answer appears somewhere in the document** and probably once.
- Applications involve searching an individual resource, such as a book, encyclopedia, or manual.
- Reading comprehension tests are also a form of single document question answering.
- ex. SQuAD

Multiple

- A multiple document Q/A task involves questions posed against a collection of documents.
- The answer may appear in the collection multiple times or may not appear at all!
- Applications include WWW search engines, and searching text repositories such as news archives, medical literature, or scientific articles.
- ex. MS MARCO

Introduction

Simple vs Complex

Question Type

- Simple (factoid) questions (most commercial systems)
  - Who wrote the Declaration of Independence?
  - What is the average age of the onset of autism?
  - Where is Apple Computer based?
  - ex. SQuAD
- Complex (narrative) questions
  - What do scholars think about Jefferson's position on dealing with pirates?
  - What is a Hajj?
  - In children with an acute febrile illness, what is the efficacy of single medication therapy with acetaminophen or ibuprofen in reducing fever?
  - ex. Narrative QA
- Complex (opinion) questions
  - Was the Gore/Bush election fair?

Introduction

Text vs Visual

Text

Input Document and Question.

Output Answer.

Example

- University of Washington
- Allennlp

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Visual

Input Picture or Video and Question.

Output Answer.

- What is in the image?
- Are there any humans?
- What sport is being played?
- Who has the ball?
- How many players are in the image?
- Who are the teams?
- Is it raining?

Example

- http://vqa.cloudcv.org

Introduction

IR-based vs Knowledge-based

Information Retrieval based

- Information Retrieval:
  QA can be viewed as short passage retrieval.
- Information Extraction:
  QA can be viewed as open-domain information extraction.

Knowledge-based

- Build a semantic representation of the query
  - Times, dates, locations, entities, numeric quantities
- Map from this semantics to query structured data or resources
  - Geospatial databases
  - Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
  - Restaurant review sources and reservation services
  - Scientific databases
- Examples: Siri

Hybrid based

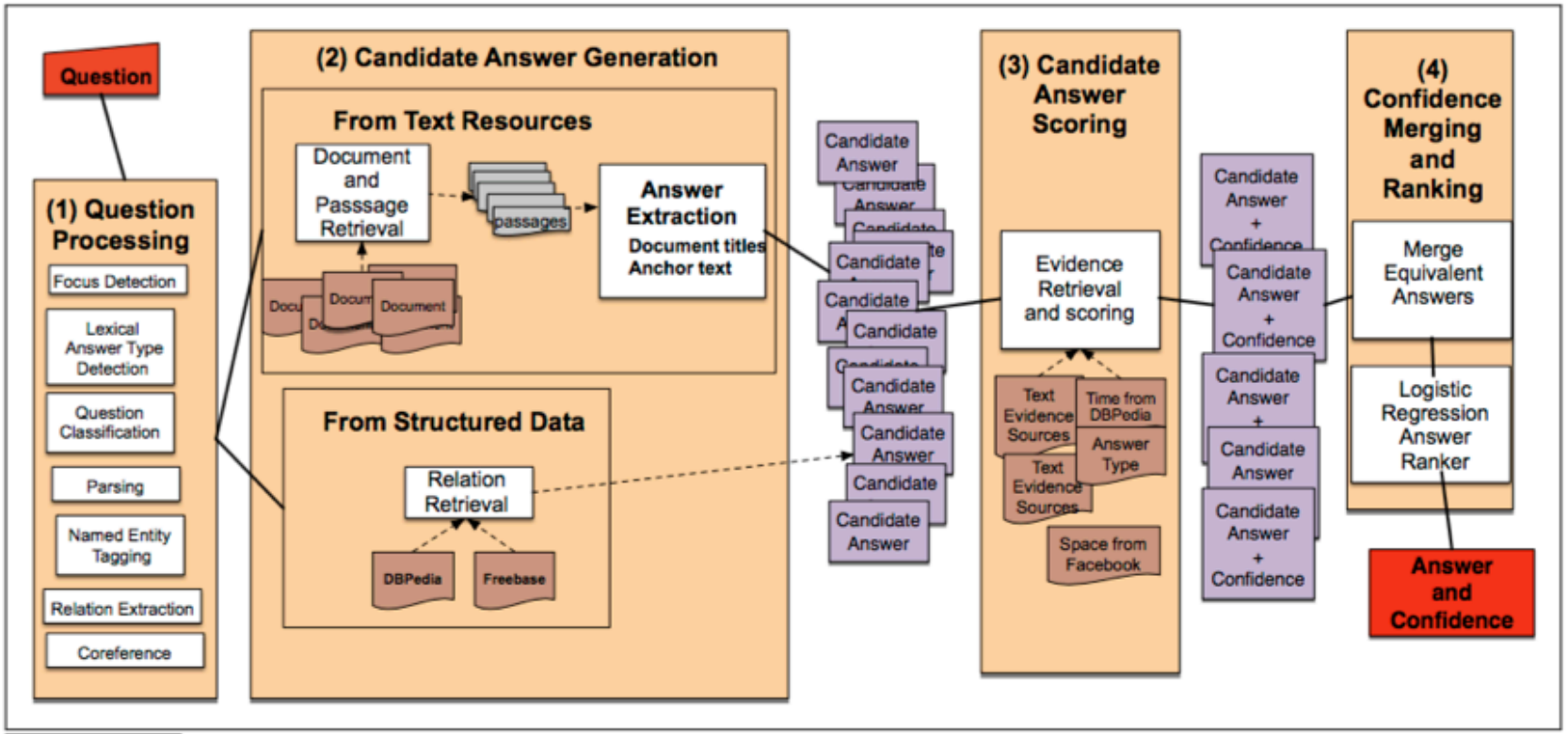
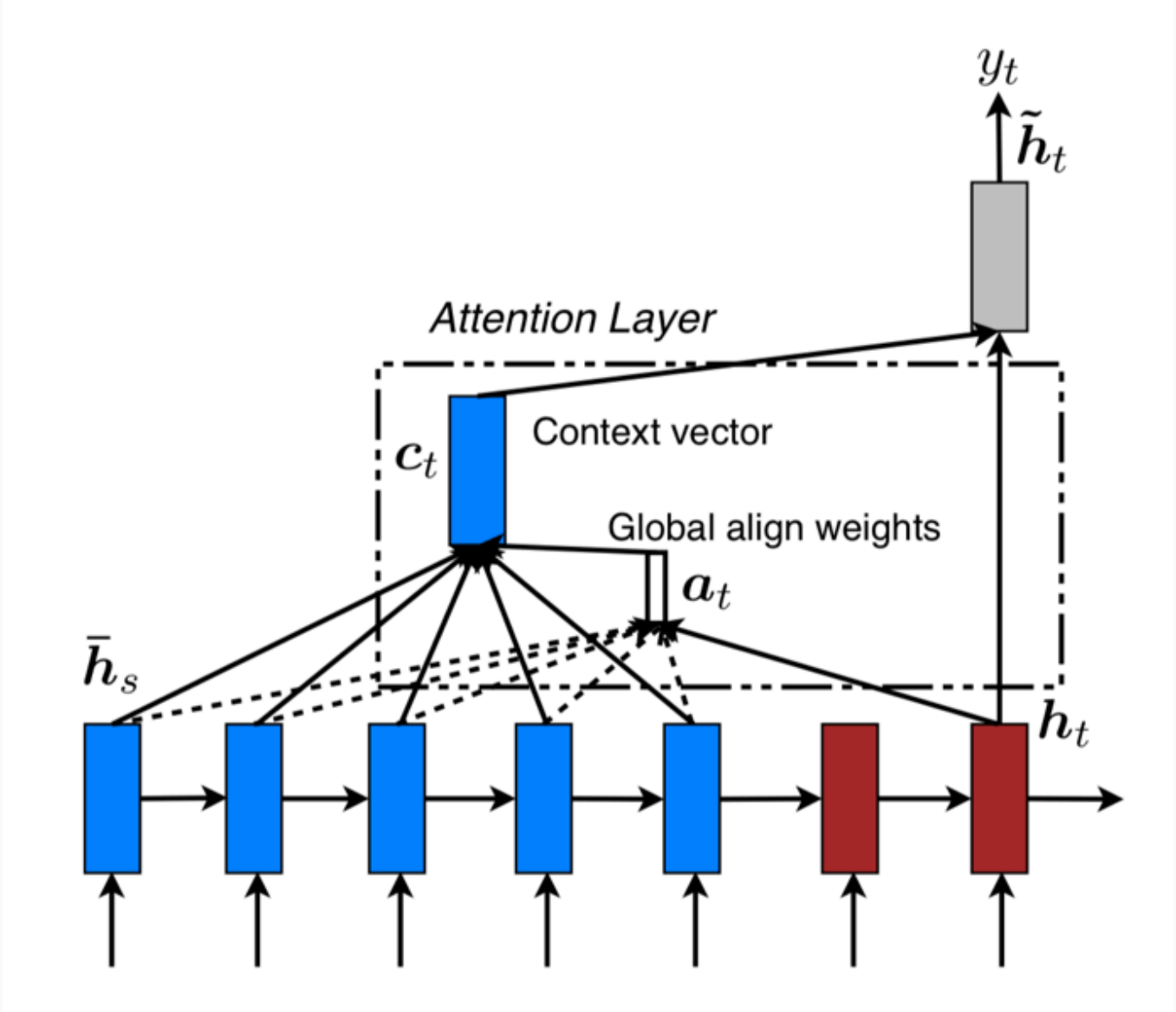


Figure 28.9 The 4 broad stages of Watson QA: (1) Question Processing, (2) Candidate Answer Generation, (3) Candidate Answer Scoring, and (4) Answer Merging and Confidence Scoring.

Machine Comprehension by Deep Learning

Recap: Attention



Machine Comprehension by Deep Learning

bAbi

bAbi

- Whether a system is able to answer questions via chaining facts, simple induction, deduction and many more.
- 20 Types of question:
  Single Supporting Fact, Two or Three Supporting Facts, Two or Three Argument Relations, Yes/No Questions, Counting and Lists/Sets, Simple Negation and Indefinite Knowledge, Basic Coreference, Conjunctions and Compound Coreference, Time Reasoning, Basic Deduction and Induction, Positional and Size Reasoning, Path Finding, Agent's Motivations

bAbi (cont.)

Task 1: Single Supporting Fact

Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts

John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 3: Three Supporting Facts

John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

Task 4: Two Argument Relations

The office is north of the bedroom.
The bedroom is north of the bathroom.
The kitchen is west of the garden.
What is north of the bedroom? A: office
What is the bedroom north of? A: bathroom

Task 5: Three Argument Relations

Mary gave the cake to Fred.
Fred gave the cake to Bill.
Jeff was given the milk by Bill.
Who gave the cake to Fred? A: Mary
Who did Fred give the cake to? A: Bill

Task 6: Yes/No Questions

John moved to the playground.
Daniel went to the bathroom.
John went back to the hallway.
Is John in the playground? A:no
Is Daniel in the bathroom? A:yes

Task 7: Counting

Daniel picked up the football.
Daniel dropped the football.
Daniel got the milk.
Daniel took the apple.
How many objects is Daniel holding? A: two

Task 8: Lists/Sets

Daniel picks up the football.
Daniel drops the newspaper.
Daniel picks up the milk.
John took the apple.
What is Daniel holding? milk, football

Task 9: Simple Negation

Sandra travelled to the office.
Fred is no longer in the office.
Is Fred in the office? A:no
Is Sandra in the office? A:yes

Task 10: Indefinite Knowledge

John is either in the classroom or the playground.
Sandra is in the garden.
Is John in the classroom? A:maybe
Is John in the office? A:no

bAbi (cont.)

Task 11: Basic Coreference

Daniel was in the kitchen.
Then he went to the studio.
Sandra was in the office.
Where is Daniel? A:studio

Task 12: Conjunction

Mary and Jeff went to the kitchen.
Then Jeff went to the park.
Where is Mary? A: kitchen
Where is Jeff? A: park

Task 13: Compound Coreference

Daniel and Sandra journeyed to the office.
Then they went to the garden.
Sandra and John travelled to the kitchen.
After that they moved to the hallway.
Where is Daniel? A: garden

Task 14: Time Reasoning

In the afternoon Julie went to the park.
Yesterday Julie was at school.
Julie went to the cinema this evening.
Where did Julie go after the park? A:cinema
Where was Julie before the park? A:school

Task 15: Basic Deduction

Sheep are afraid of wolves.
Cats are afraid of dogs.
Mice are afraid of cats.
Gertrude is a sheep.
What is Gertrude afraid of? A:wolves

Task 16: Basic Induction

Lily is a swan.
Lily is white.
Bernhard is green.
Greg is a swan.
What color is Greg? A:white

Task 17: Positional Reasoning

The triangle is to the right of the blue square.
The red square is on top of the blue square.
The red sphere is to the right of the blue square.
Is the red sphere to the right of the blue square? A:yes
Is the red square to the left of the triangle? A:yes

Task 18: Size Reasoning

The football fits in the suitcase.
The suitcase fits in the cupboard.
The box is smaller than the football.
Will the box fit in the suitcase? A:yes
Will the cupboard fit in the box? A:no

Task 19: Path Finding

The kitchen is north of the hallway.
The bathroom is west of the bedroom.
The den is east of the hallway.
The office is south of the bedroom.
How do you go from den to kitchen? A: west, north
How do you go from office to bathroom? A: north, west

Task 20: Agent's Motivations

John is hungry.
John goes to the kitchen.
John grabbed the apple there.
Daniel is hungry.
Where does Daniel go? A:kitchen
Why did John go to the kitchen? A:hungry

End-To-End Memory Networks

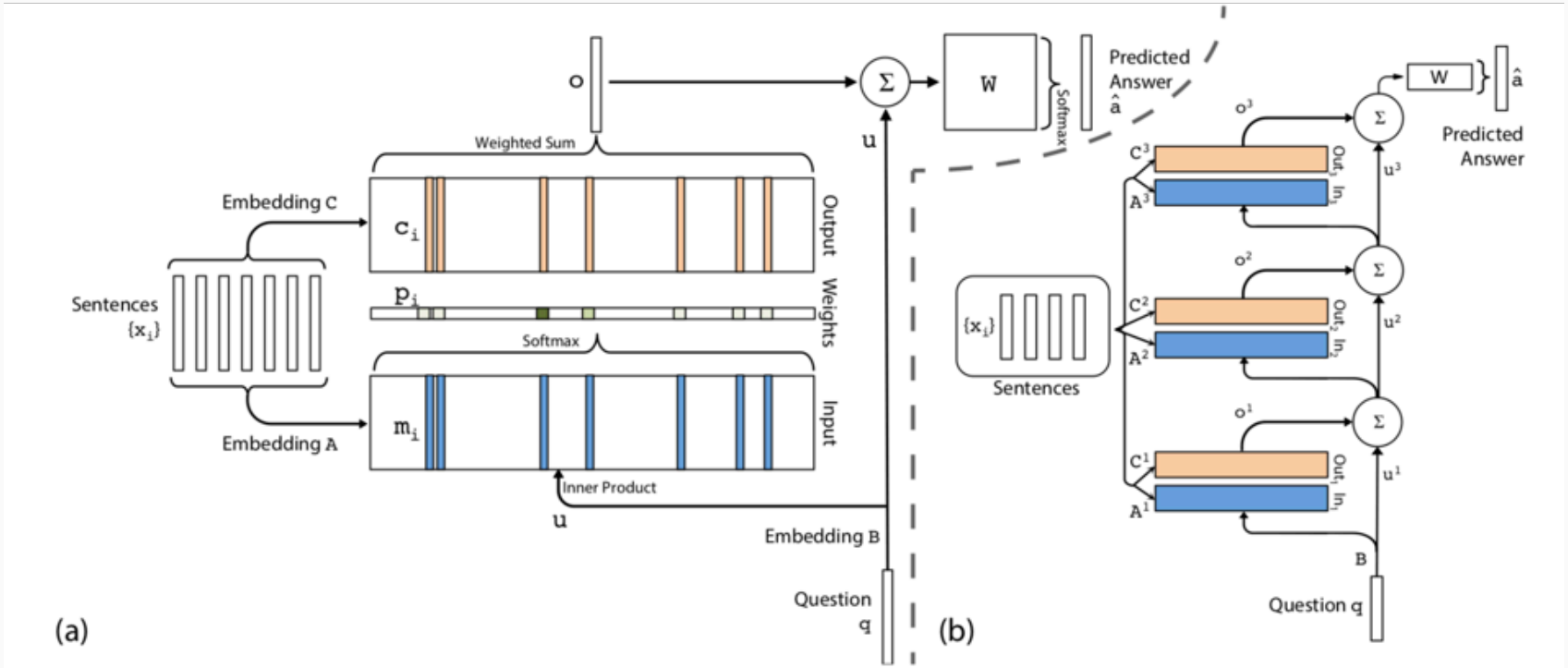
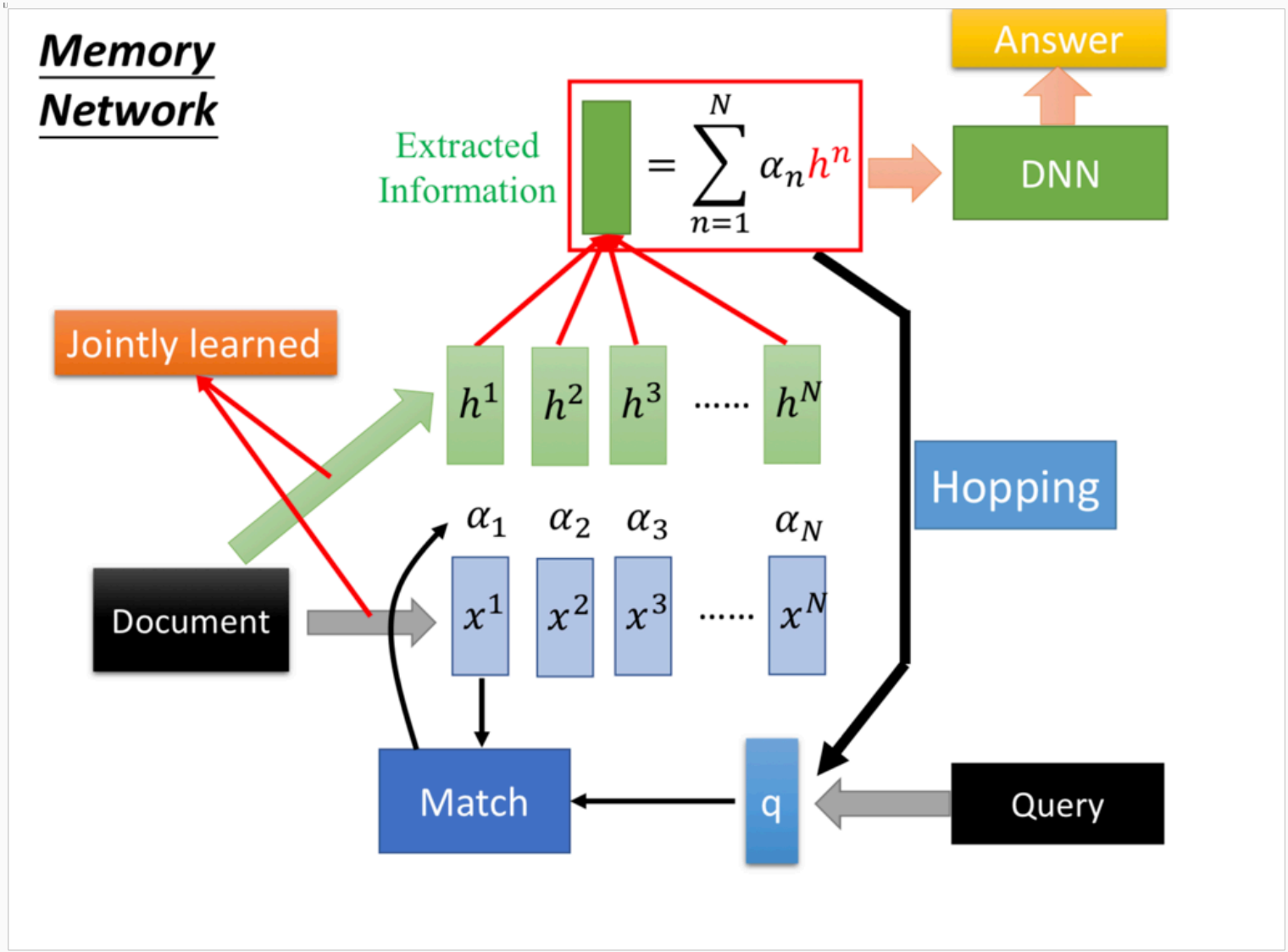


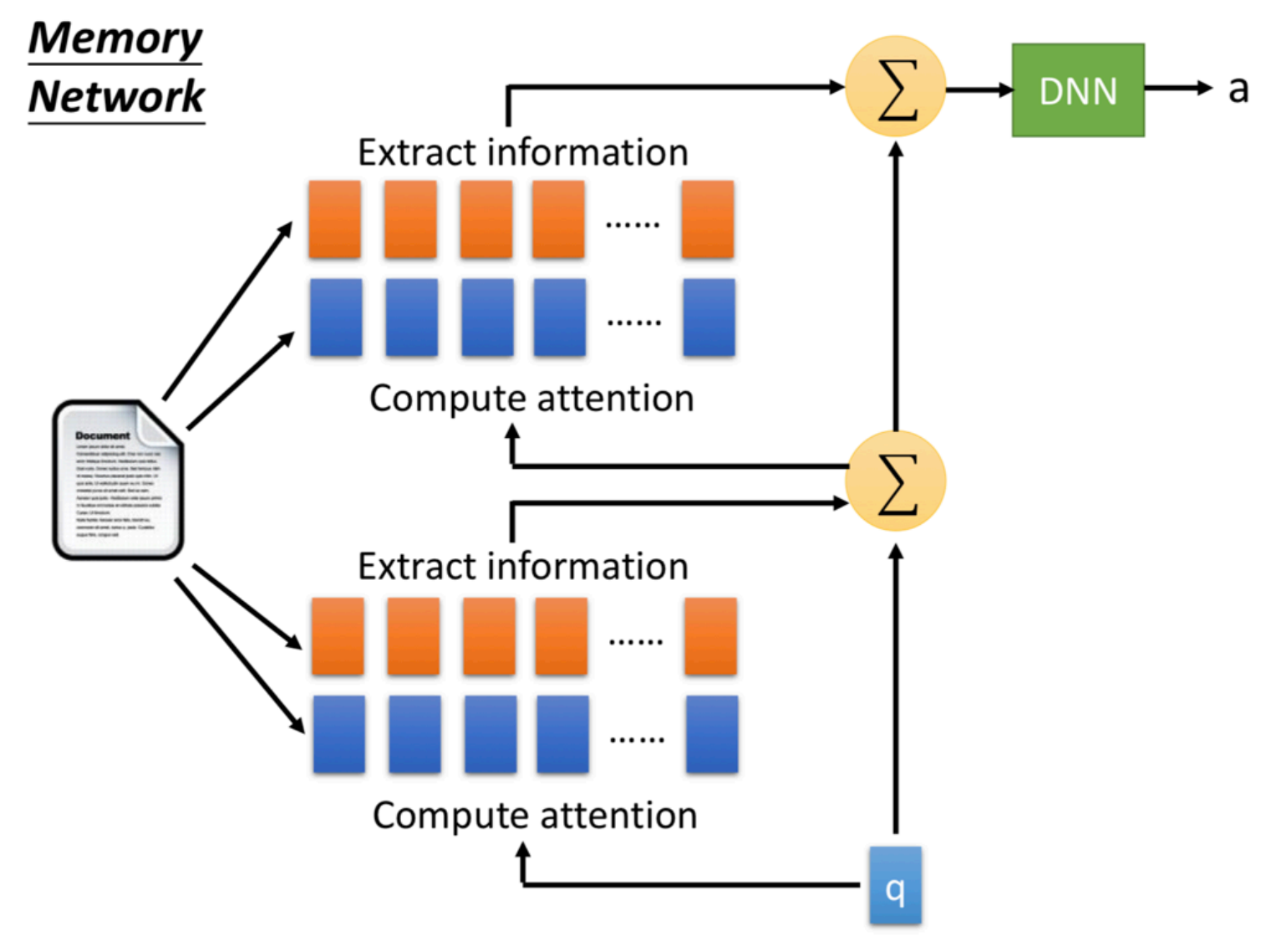
Figure 1: (a): A single layer version of our model. (b): A three layer version of our model. In practice, we can constrain several of the embedding matrices to be the same (see Section 2.2).

End-To-End Memory Networks (cont.)



感謝李宏毅教授提供

End-To-End Memory Networks (cont.)



感謝李宏毅教授提供

End-To-End Memory Networks (cont.)

Story (1: 1 supporting fact)	Support	Hop 1	Hop 2	Hop 3
Daniel went to the bathroom.		0.00	0.00	0.03
Mary travelled to the hallway.		0.00	0.00	0.00
John went to the bedroom.		0.37	0.02	0.00
John travelled to the bathroom.	yes	0.60	0.98	0.96
Mary went to the office.		0.01	0.00	0.00
Where is John? Answer: bathroom Prediction: bathroom				

Story (2: 2 supporting facts)	Support	Hop 1	Hop 2	Hop 3
John dropped the milk.		0.06	0.00	0.00
John took the milk there.	yes	0.88	1.00	0.00
Sandra went back to the bathroom.		0.00	0.00	0.00
John moved to the hallway.	yes	0.00	0.00	1.00
Mary went back to the bedroom.		0.00	0.00	0.00
Where is the milk? Answer: hallway Prediction: hallway				

Story (16: basic induction)	Support	Hop 1	Hop 2	Hop 3
Brian is a frog.	yes	0.00	0.98	0.00
Lily is gray.		0.07	0.00	0.00
Brian is yellow.	yes	0.07	0.00	1.00
Julius is green.		0.06	0.00	0.00
Greg is a frog.	yes	0.76	0.02	0.00
What color is Greg? Answer: yellow Prediction: yellow				

Story (18: size reasoning)	Support	Hop 1	Hop 2	Hop 3
The suitcase is bigger than the chest.	yes	0.00	0.88	0.00
The box is bigger than the chocolate.		0.04	0.05	0.10
The chest is bigger than the chocolate.	yes	0.17	0.07	0.90
The chest fits inside the container.		0.00	0.00	0.00
The chest fits inside the box.		0.00	0.00	0.00
Does the suitcase fit in the chocolate? Answer: no Prediction: no				

	Baseline			MemN2N								
Task	Strongly Supervised MemNN [22]	LSTM [22]	MemNN WSH	BoW	PE	PE LS	PE LS RN	1 hop PE LS joint	2 hops PE LS joint	3 hops PE LS joint	PE LS RN joint	PE LS LW joint
1: 1 supporting fact	0.0	50.0	0.1	0.6	0.1	0.2	0.0	0.8	0.0	0.1	0.0	0.1
2: 2 supporting facts	0.0	80.0	42.8	17.6	21.6	12.8	8.3	62.0	15.6	14.0	11.4	18.8
3: 3 supporting facts	0.0	80.0	76.4	71.0	64.2	58.8	40.3	76.9	31.6	33.1	21.9	31.7
4: 2 argument relations	0.0	39.0	40.3	32.0	3.8	11.6	2.8	22.8	2.2	5.7	13.4	17.5
5: 3 argument relations	2.0	30.0	16.3	18.3	14.1	15.7	13.1	11.0	13.4	14.8	14.4	12.9
6: yes/no questions	0.0	52.0	51.0	8.7	7.9	8.7	7.6	7.2	2.3	3.3	2.8	2.0
7: counting	15.0	51.0	36.1	23.5	21.6	20.3	17.3	15.9	25.4	17.9	18.3	10.1
8: lists/sets	9.0	55.0	37.8	11.4	12.6	12.7	10.0	13.2	11.7	10.1	9.3	6.1
9: simple negation	0.0	36.0	35.9	21.1	23.3	17.0	13.2	5.1	2.0	3.1	1.9	1.5
10: indefinite knowledge	2.0	56.0	68.7	22.8	17.4	18.6	15.1	10.6	5.0	6.6	6.5	2.6
11: basic coreference	0.0	38.0	30.0	4.1	4.3	0.0	0.9	8.4	1.2	0.9	0.3	3.3
12: conjunction	0.0	26.0	10.1	0.3	0.3	0.1	0.2	0.4	0.0	0.3	0.1	0.0
13: compound coreference	0.0	6.0	19.7	10.5	9.9	0.3	0.4	6.3	0.2	1.4	0.2	0.5
14: time reasoning	1.0	73.0	18.3	1.3	1.8	2.0	1.7	36.9	8.1	8.2	6.9	2.0
15: basic deduction	0.0	79.0	64.8	24.3	0.0	0.0	0.0	46.4	0.5	0.0	0.0	1.8
16: basic induction	0.0	77.0	50.5	52.0	52.1	1.6	1.3	47.4	51.3	3.5	2.7	51.0
17: positional reasoning	35.0	49.0	50.9	45.4	50.1	49.0	51.0	44.4	41.2	44.5	40.4	42.6
18: size reasoning	5.0	48.0	51.3	48.1	13.6	10.1	11.1	9.6	10.3	9.2	9.4	9.2
19: path finding	64.0	92.0	100.0	89.7	87.4	85.6	82.8	90.7	89.9	90.2	88.0	90.6
20: agent's motivation	0.0	9.0	3.6	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.2
Mean error (%)	6.7	51.3	40.2	25.1	20.3	16.3	13.9	25.8	15.6	13.3	12.4	15.2
Failed tasks (err. $>$ 5%)	4	20	18	15	13	12	11	17	11	11	11	10
On 10k training data												
Mean error (%)	3.2	36.4	39.2	15.4	9.4	7.2	6.6	24.5	10.9	7.9	7.5	11.0
Failed tasks (err. $>$ 5%)	2	16	17	9	6	4	4	16	7	6	6	6

Machine Comprehension by Deep Learning

SQuAD

SQuAD

SQuAD1.1 Leaderboard

Since the release of SQuAD1.0, the community has made rapid progress, with the best models now rivaling human performance on the task. Here are the ExactMatch (EM) and F1 scores evaluated on the test set of v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147

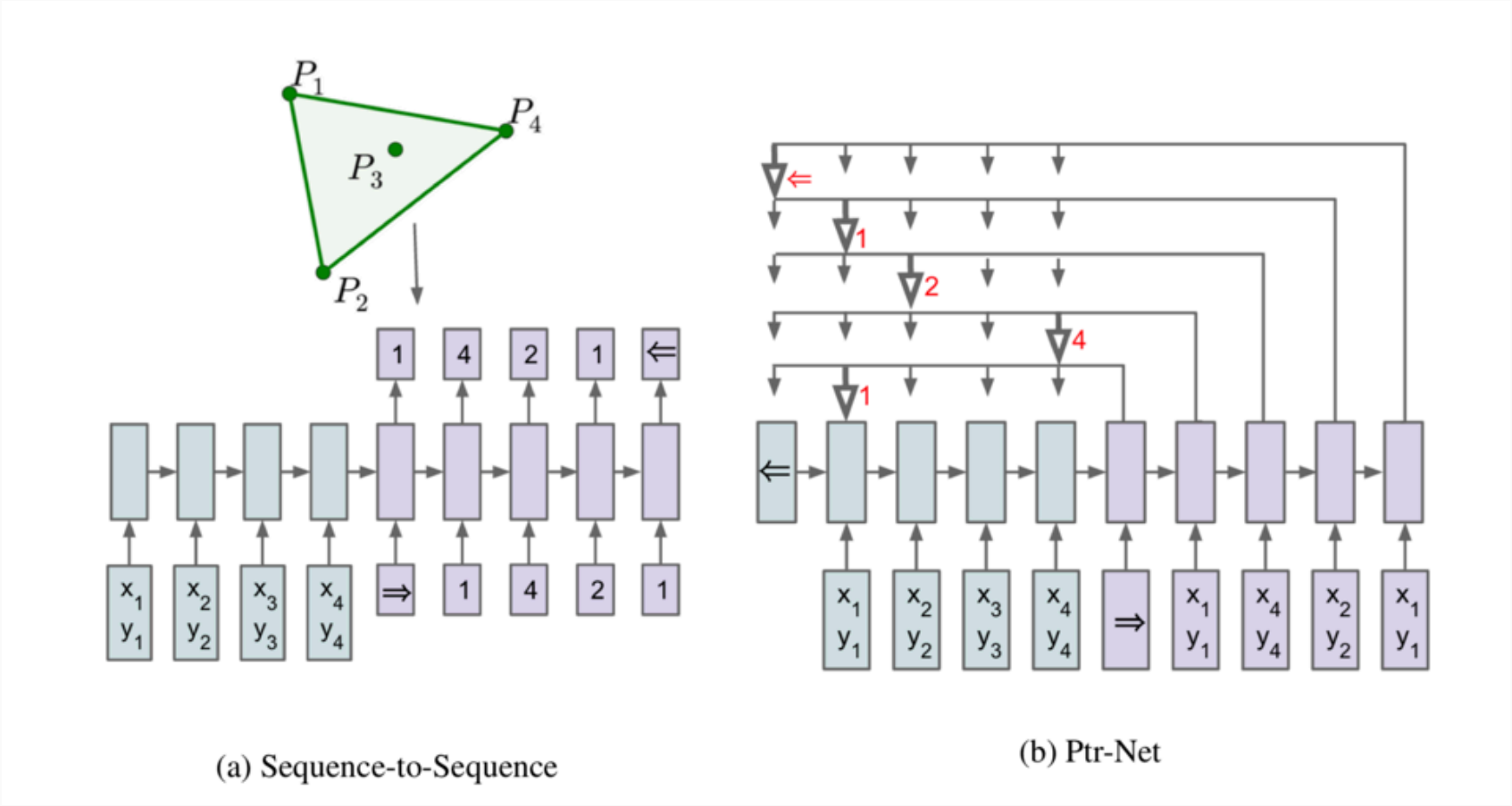
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

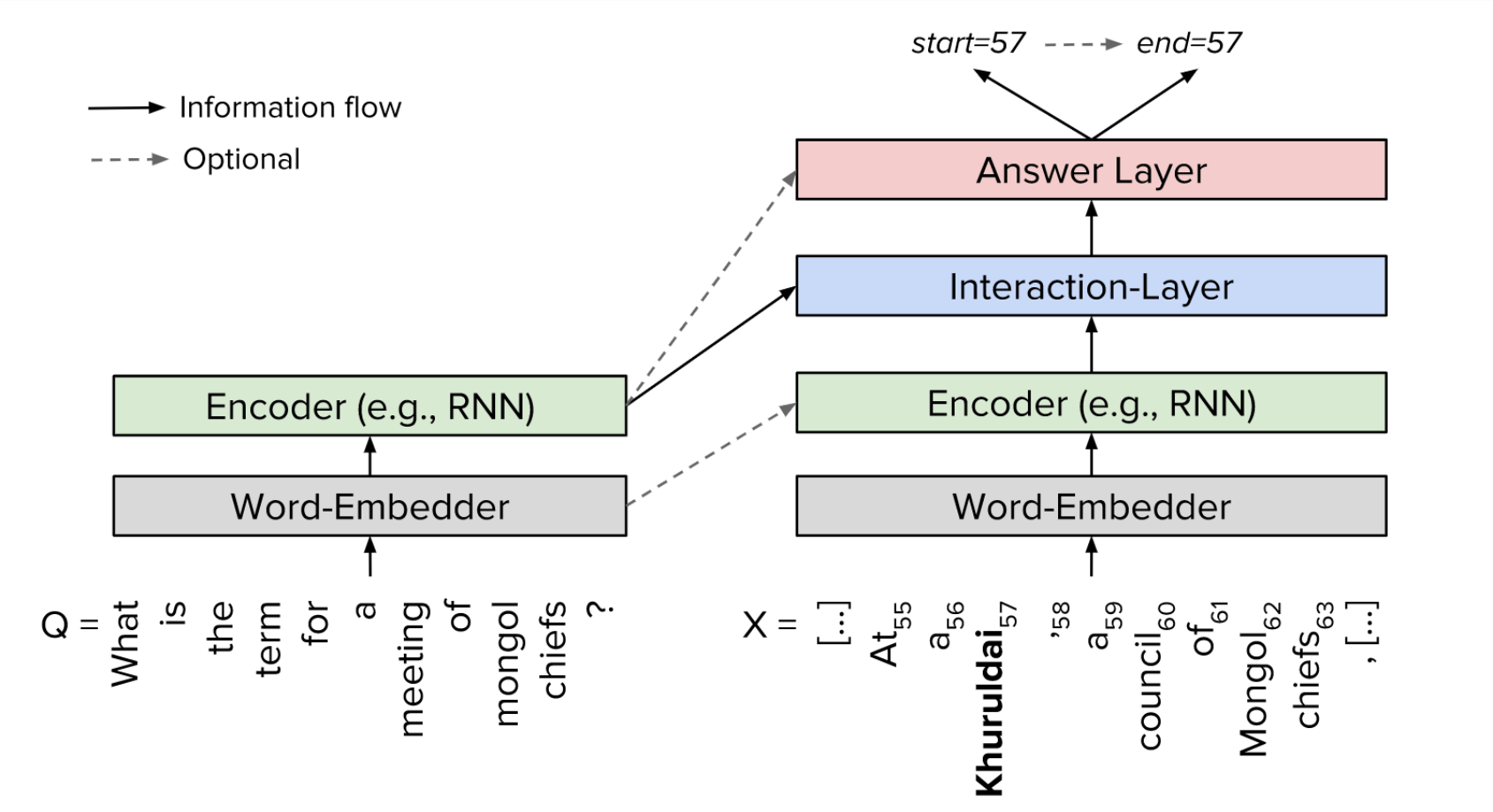
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

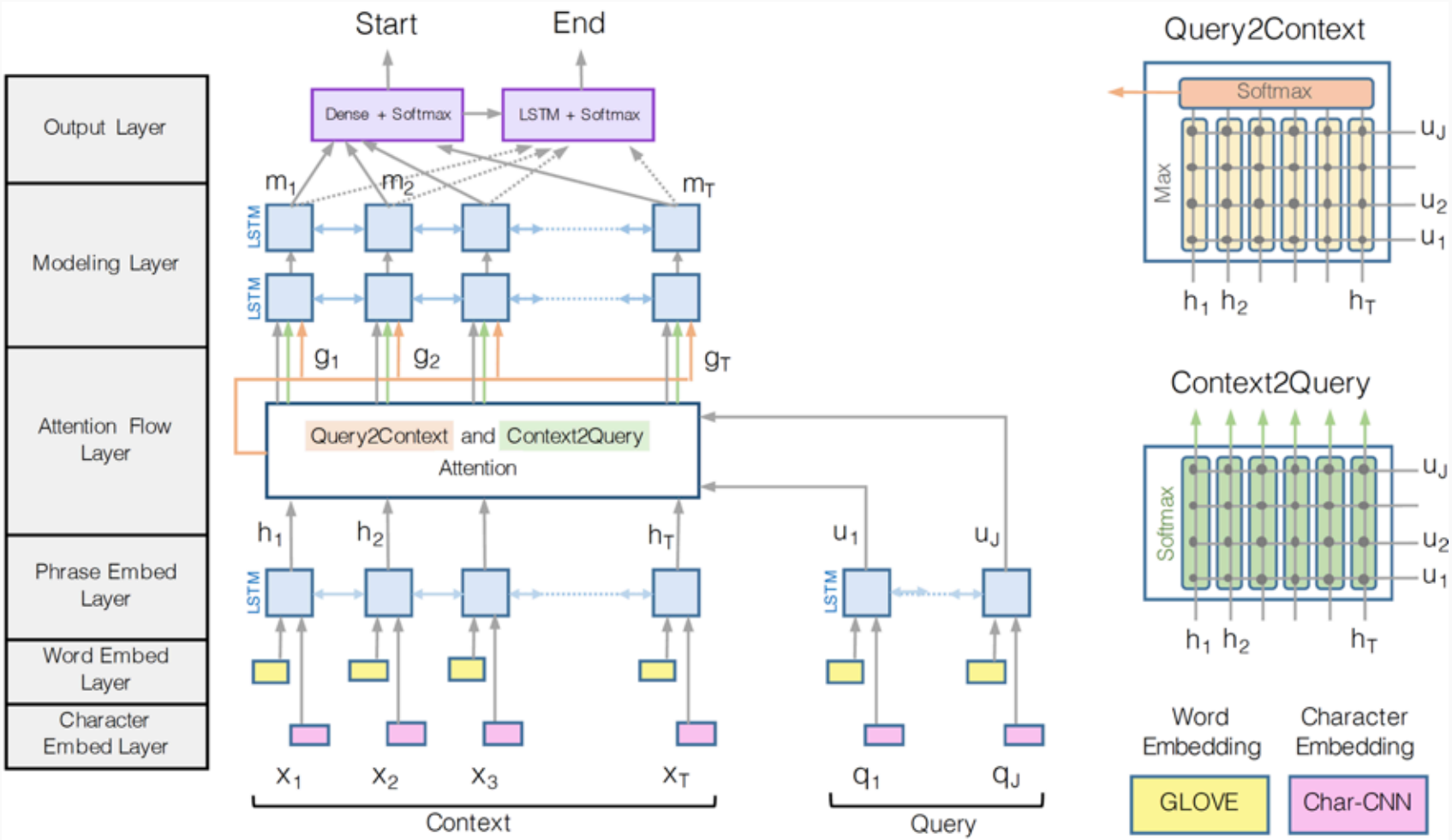
Pointer Network



Model Architecture



BiDAF


Output Layer
Modeling Layer
Attention Flow Layer
Phrase Embed Layer
Word Embed Layer
Character Embed Layer
Start
End
Dense + Softmax
LSTM + Softmax
LSTM
Query2Context and Context2Query Attention
Context
Query
Query2Context
Softmax
Max
Context2Query
Word Embedding
Character Embedding
GLOVE
Char-CNN

DRQA

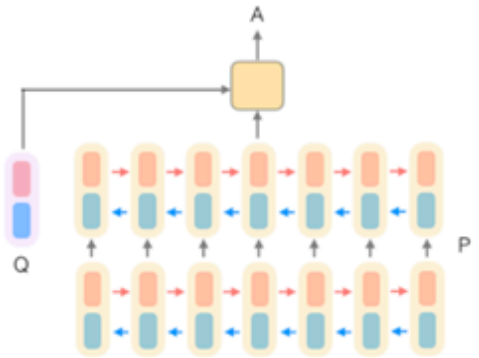
Open-domain QA
SQuAD, TREC, WebQuestions, WikiMovies

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

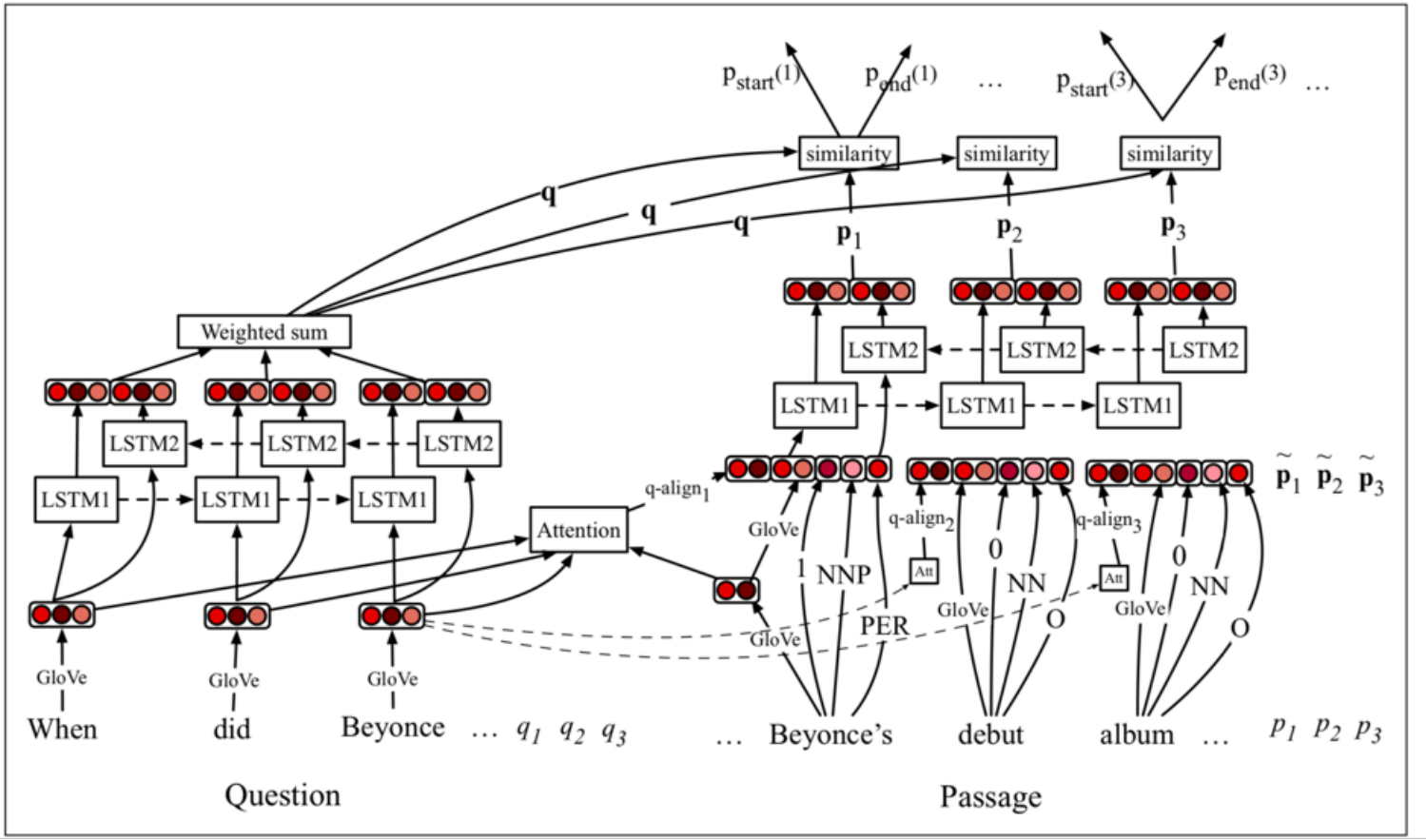
Document Retriever

Document Reader

833,500

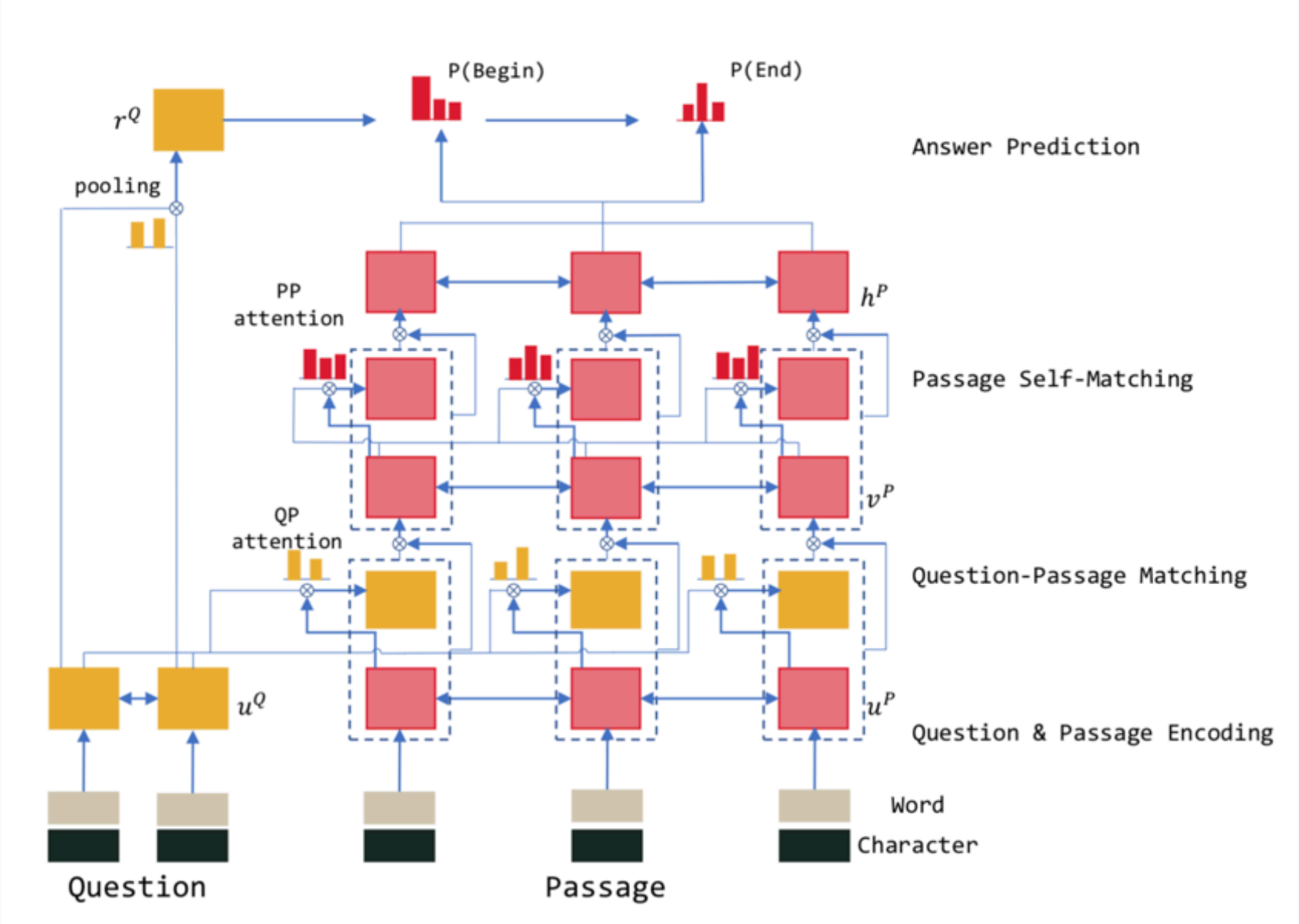


DRQA (cont.)



Method	Dev		Test	
	EM	F1	EM	F1
Dynamic Coattention Networks (Xiong et al., 2016)	65.4	75.6	66.2	75.9
Multi-Perspective Matching (Wang et al., 2016)†	66.1	75.8	65.5	75.1
BiDAF (Seo et al., 2016)	67.7	77.3	68.0	77.3
R-net†	n/a	n/a	71.3	79.7
DrQA (Our model, Document Reader Only)	**69.5**	**78.8**	70.0	79.0

R-net


P(Begin)
P(End)
r^Q
Answer Prediction
pooling
PP
attention
h^P
Passage Self-Matching
v^P
QP
attention
Question-Passage Matching
u^Q
u^P
Question & Passage Encoding
Word
Character
Question
Passage

Fusion Net

Architectures	(1)	(2)	(2')	(3)	(3')
Match-LSTM (Wang & Jiang, 2016)		✓			
DCN (Xiong et al., 2017)		✓			✓
FastQA (Weissenborn et al., 2017)	✓				
FastQAExt (Weissenborn et al., 2017)	✓	✓		✓	
BiDAF (Seo et al., 2017)		✓			✓
RaSoR (Lee et al., 2016)	✓		✓		
DrQA (Chen et al., 2017a)	✓				
MPCM (Wang et al., 2016)	✓	✓			
Mnemonic Reader (Hu et al., 2017)	✓	✓		✓	
R-net (Wang et al., 2017)		✓		✓	

Table 1: A summarized view on the fusion processes used in several state-of-the-art architectures.

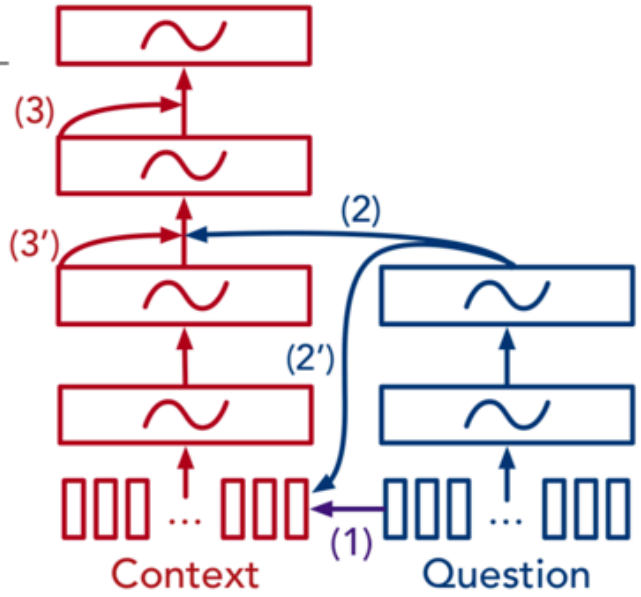


Figure 2: A conceptual architecture illustrating recent advances in MRC.

Context: The Alpine Rhine is part of the Rhine, a famous European river. The **Alpine Rhine** begins in the most western part of the Swiss canton of Graubünden, and later **forms the border** between Switzerland to the West and **Liechtenstein** and later Austria to the East. On the other hand, the **Danube** separates Romania and Bulgaria.

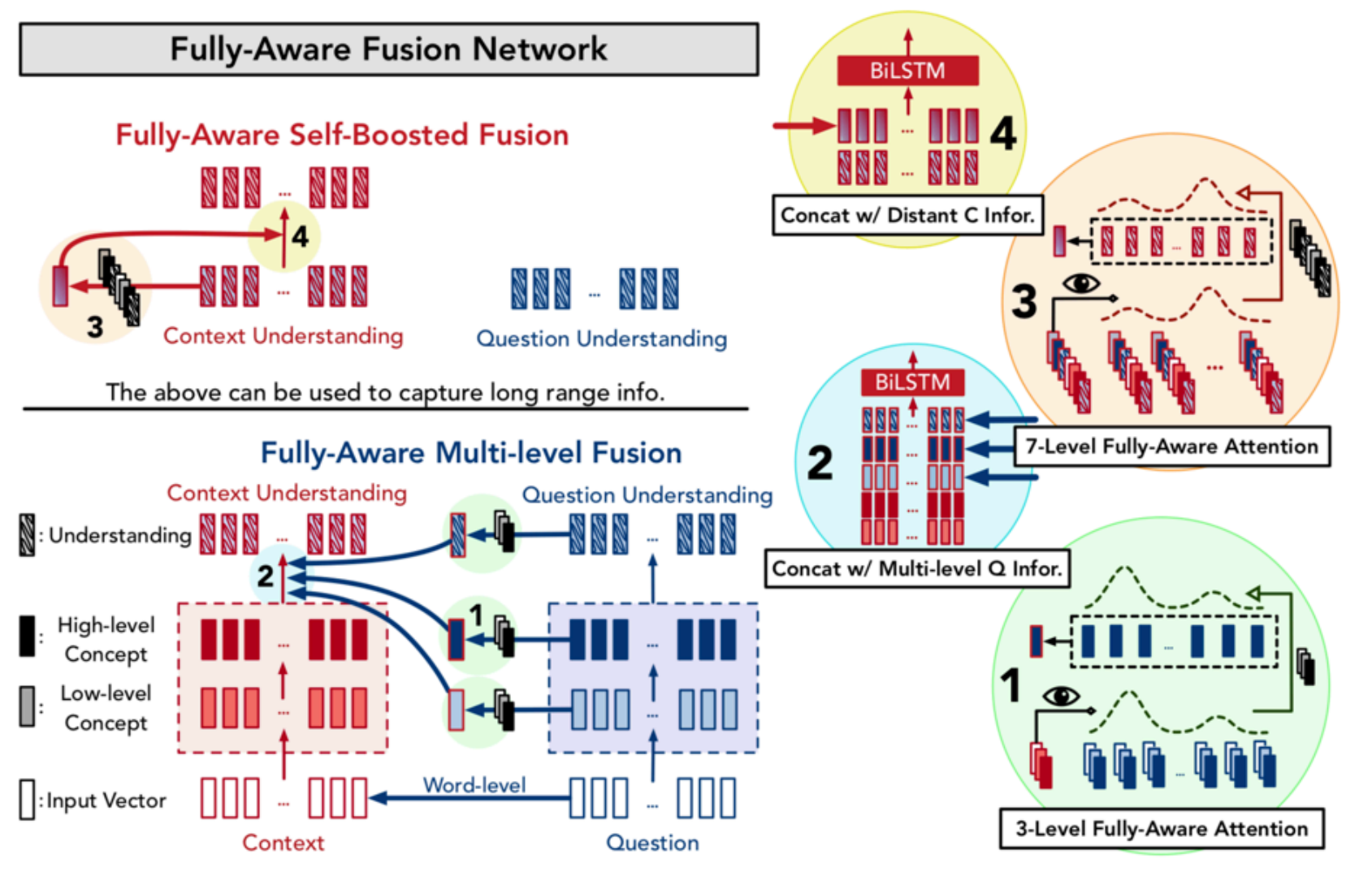
Question: What is the other country the Rhine separates Switzerland to?

Answer: Liechtenstein

"History of Word" Concept

Input Word	Low level	High level
Alpine Rhine	European river	Separating River
Forms the border	Border countries	Separates
Liechtenstein	Country	Country, separate
Danube	European river	Separating River

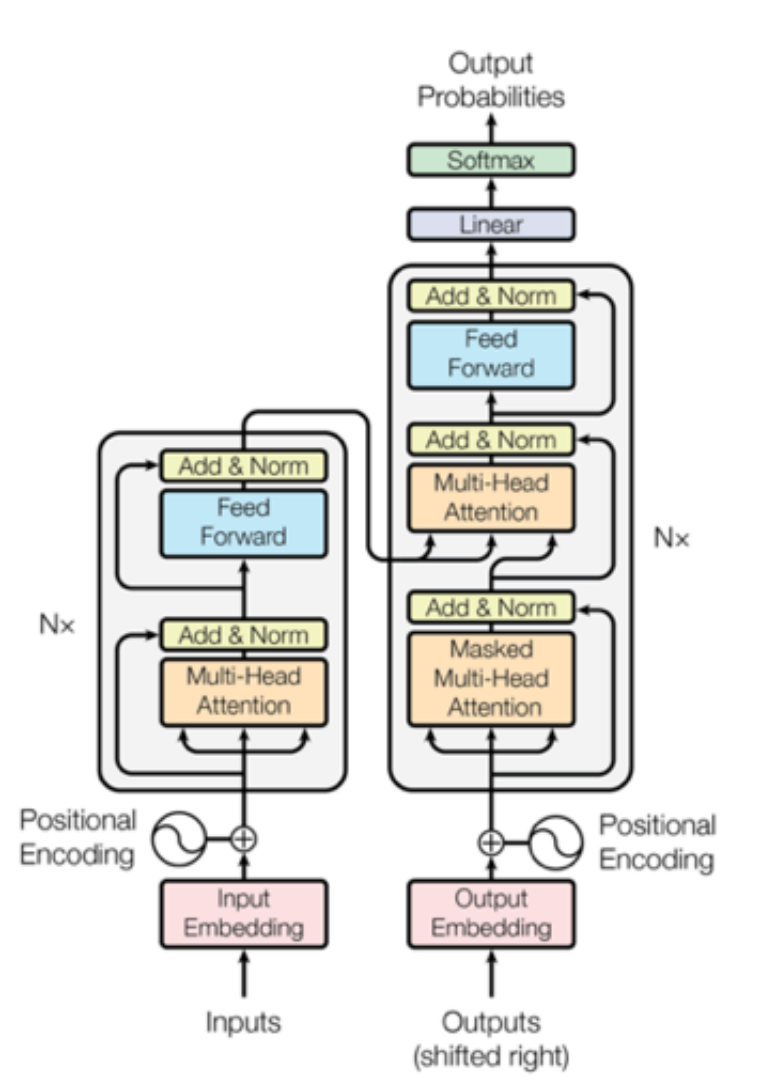
Fusion Net (cont.)

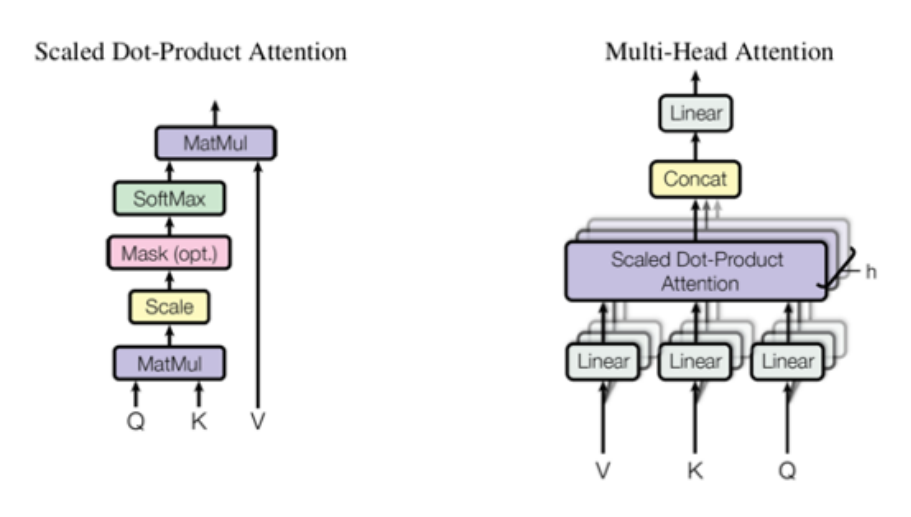


Fusion Net (cont.)

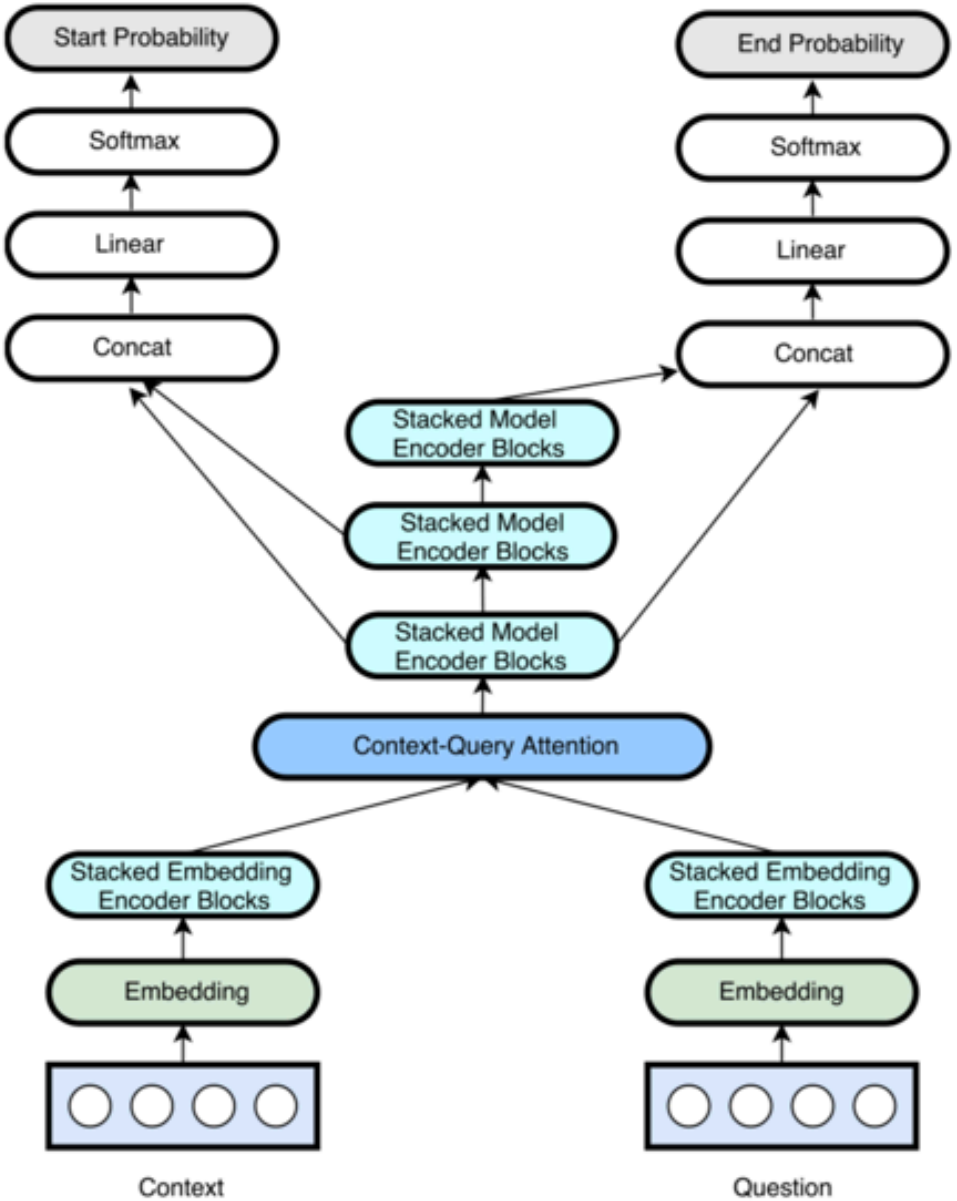
	Test Set
Single Model	**EM / F1**
LR Baseline (Rajpurkar et al., 2016)	40.4 / 51.0
Match-LSTM (Wang & Jiang, 2016)	64.7 / 73.7
BiDAF (Seo et al., 2017)	68.0 / 77.3
SEDT (Liu et al., 2017)	68.2 / 77.5
RaSoR (Lee et al., 2016)	70.8 / 78.7
DrQA (Chen et al., 2017a)	70.7 / 79.4
ReasoNet (Shen et al., 2017)	70.6 / 79.4
R. Mnemonic Reader (Hu et al., 2017)	73.2 / 81.8
DCN+	74.9 / 82.8
R-net†	75.7 / 83.5
FusionNet	**76.0 / 83.9**
Ensemble Model	
ReasoNet (Shen et al., 2017)	75.0 / 82.3
MEMEN (Pan et al., 2017)	75.4 / 82.7
R. Mnemonic Reader (Hu et al., 2017)	77.7 / 84.9
R-net†	78.2 / 85.2
DCN+	78.7 / 85.6
FusionNet	**78.8 / 85.9**
Human (Rajpurkar et al., 2016)	82.3 / 91.2

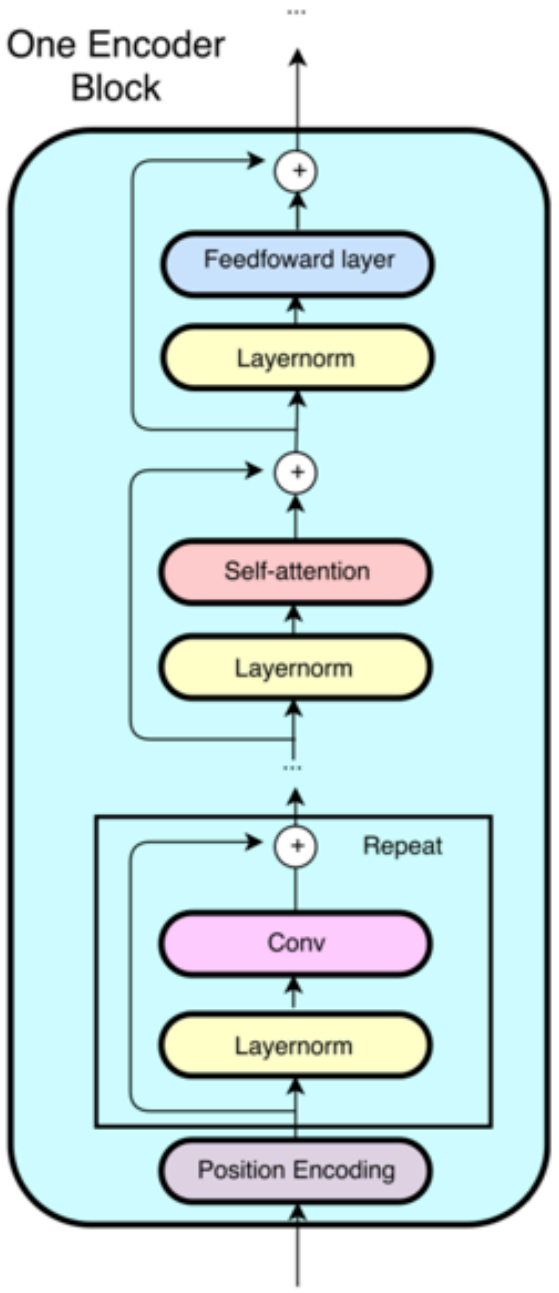
Recap: Transformer



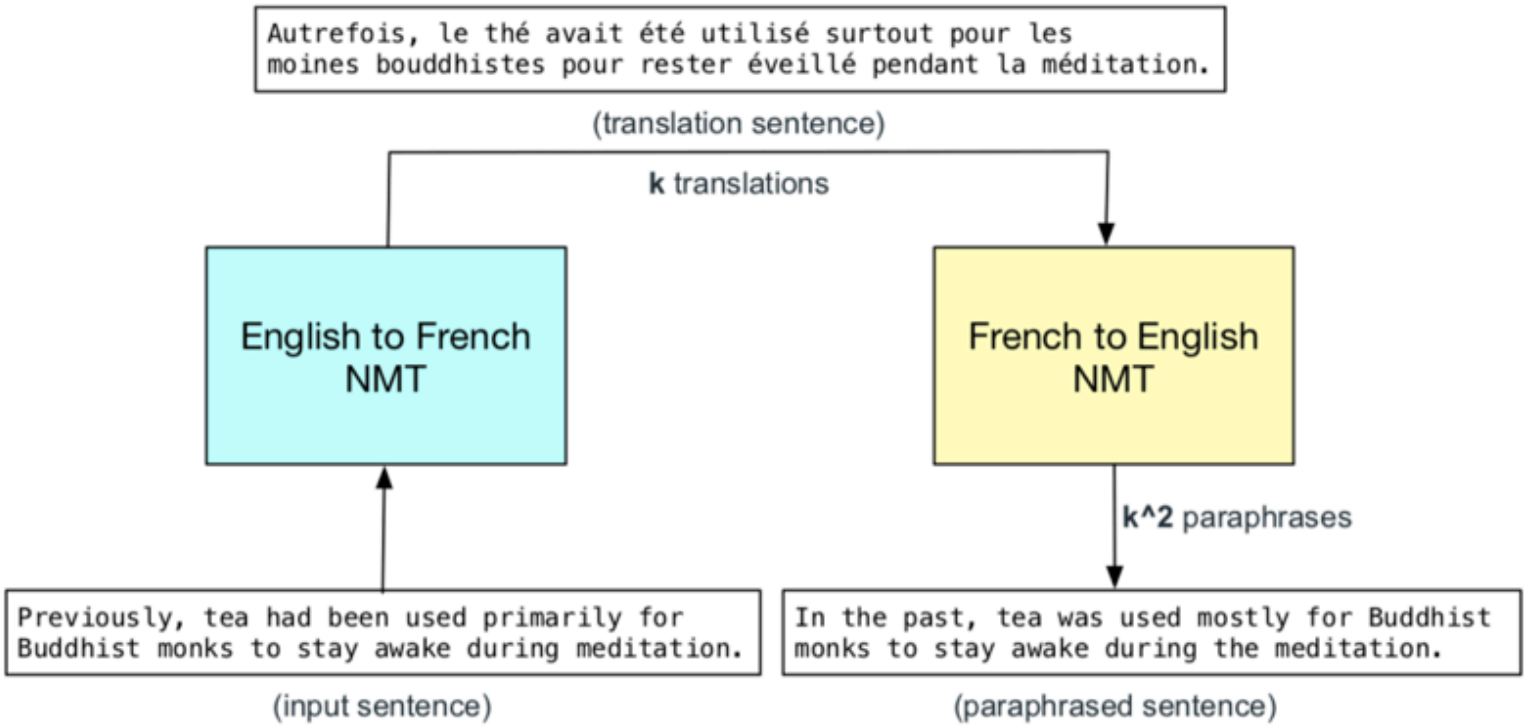


QANET


Model
Start Probability
Softmax
Linear
Concat
End Probability
Softmax
Linear
Concat
Stacked Model Encoder Blocks
Stacked Model Encoder Blocks
Stacked Model Encoder Blocks
Context-Query Attention
Stacked Embedding Encoder Blocks
Embedding
Stacked Embedding Encoder Blocks
Embedding
Context
Question


One Encoder Block
Feedfoward layer
Layernorm
Self-attention
Layernorm
Repeat
Conv
Layernorm
Position Encoding

QANET (cont.)



	Sentence that contains an answer	Answer
Original	All of the departments in the College of Science offer PhD programs, except for the Department of Pre-Professional Studies.	Department of Pre-Professional Studies
Paraphrase	All departments in the College of Science offer PHD programs with the exception of the Department of Preparatory Studies.	Department of Preparatory Studies

Table 1: Comparison between answers in original sentence and paraphrased sentence.

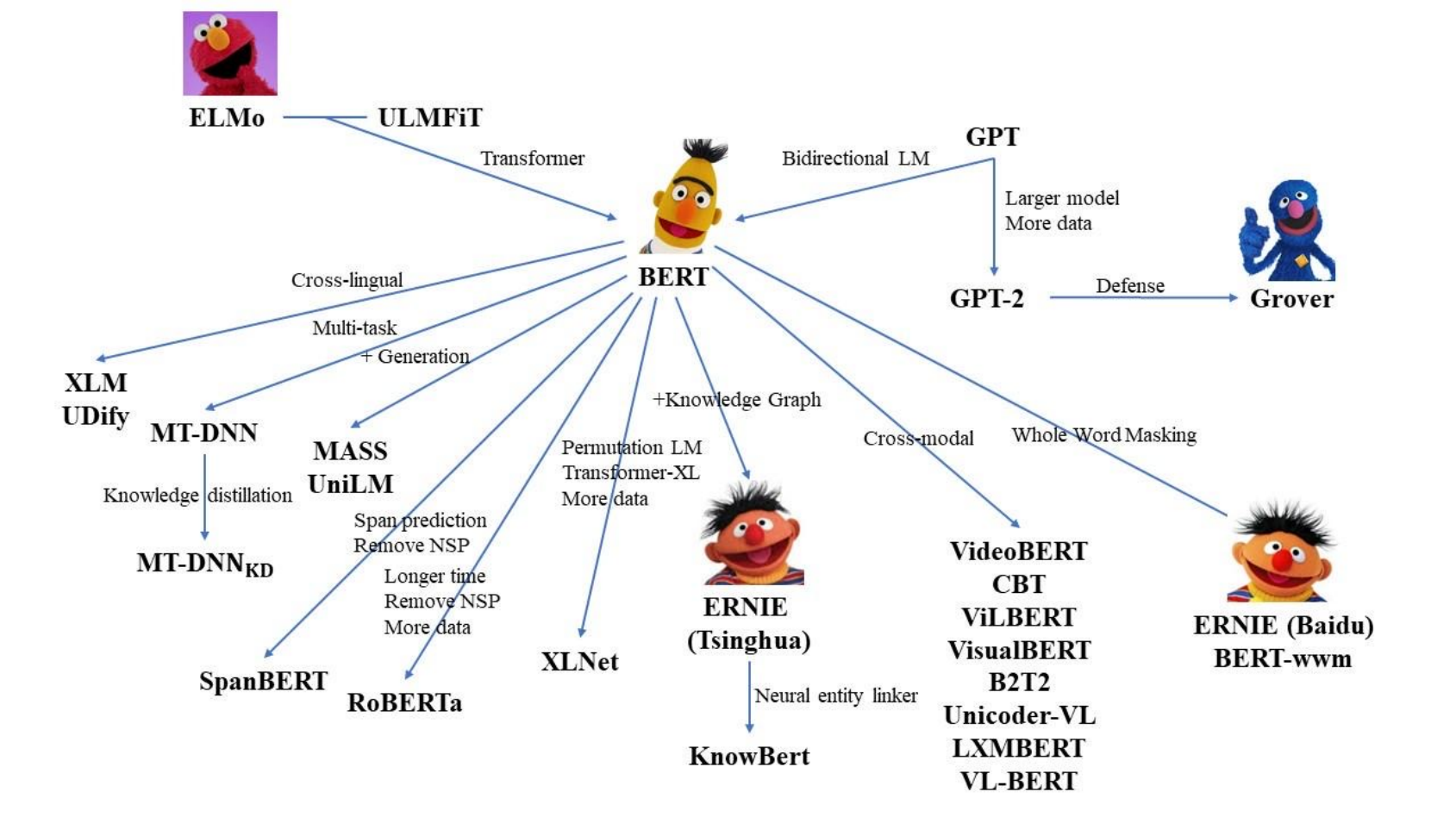
QANET (cont.)

	Published[12]	LeaderBoard[13]
Single Model	EM / F1	EM / F1
LR Baseline (Rajpurkar et al., 2016)	40.4 / 51.0	40.4 / 51.0
Dynamic Chunk Reader (Yu et al., 2016)	62.5 / 71.0	62.5 / 71.0
Match-LSTM with Ans-Ptr (Wang & Jiang, 2016)	64.7 / 73.7	64.7 / 73.7
Multi-Perspective Matching (Wang et al., 2016)	65.5 / 75.1	70.4 / 78.8
Dynamic Coattention Networks (Xiong et al., 2016)	66.2 / 75.9	66.2 / 75.9
FastQA (Weissenborn et al., 2017)	68.4 / 77.1	68.4 / 77.1
BiDAF (Seo et al., 2016)	68.0 / 77.3	68.0 / 77.3
SEDT (Liu et al., 2017a)	68.1 / 77.5	68.5 / 78.0
RaSoR (Lee et al., 2016)	70.8	
FastQAExt (Weissenborn et al., 2017)	70.8	
ReasoNet (Shen et al., 2017b)	69.1	
Document Reader (Chen et al., 2017)	70.0	
Ruminating Reader (Gong & Bowman, 2017)	70.6	
jNet (Zhang et al., 2017)	70.6	
Conductor-net		
Interactive AoA Reader (Cui et al., 2017)		
Reg-RaSoR	N/A	75.8 / 83.3
DCN+	N/A	74.9 / 82.8
AIR-FusionNet	N/A	76.0 / 83.9
R-Net (Wang et al., 2017)	72.3 / 80.7	76.5 /84.3
BiDAF + Self Attention + ELMo	N/A	**77.9/ 85.3**
Reinforced Mnemonic Reader (Hu et al., 2017)	73.2 / 81.8	73.2 / 81.8
Dev set: QANet	**73.6 / 82.7**	N/A
Dev set: QANet + data augmentation ×2	**74.5 / 83.2**	N/A
Dev set: QANet + data augmentation ×3	**75.1 / 83.8**	N/A
Test set: QANet + data augmentation ×3	**76.2 / 84.6**	76.2 / 84.6

	Train time to get 77.0 F1 on Dev set	Train speed	Inference speed
QANet	3 hours	102 samples/s	259 samples/s
BiDAF	15 hours	24 samples/s	37samples/s
Speedup	**5.0x**	**4.3x**	**7.0x**

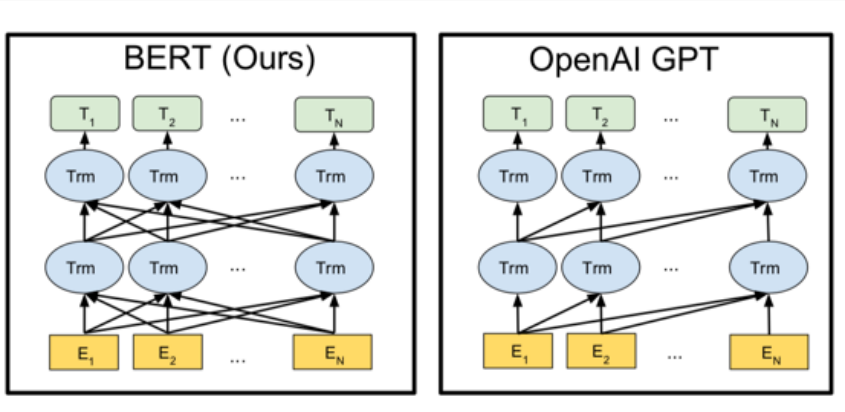
	QANet	RNN-1-128	Speedup	RNN-2-128	Speedup	RNN-3-128	Speedup
Training	**3.2**	1.1	**2.9x**	0.34	**9.4x**	0.24	**13.3x**
Inference	**8.1**	2.2	**3.7x**	1.3	**6.2x**	0.92	**8.8x**

Large Pretrain Model

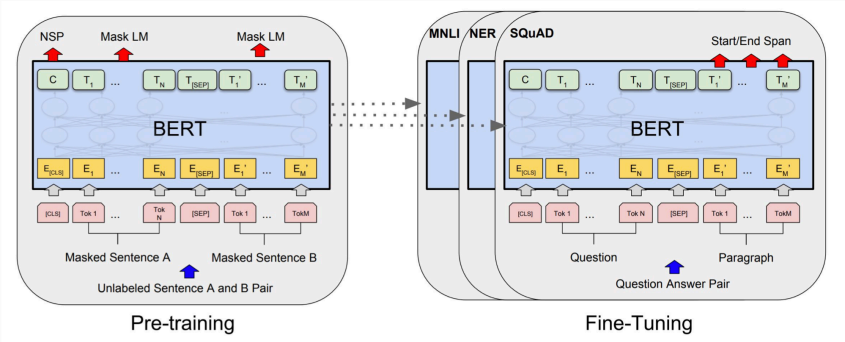


BERT

- LM pre-training on large text corpora + fine-tuning on SQuAD?
  - Improving Language Understanding with Unsupervised Learning, from OpenAI GPT



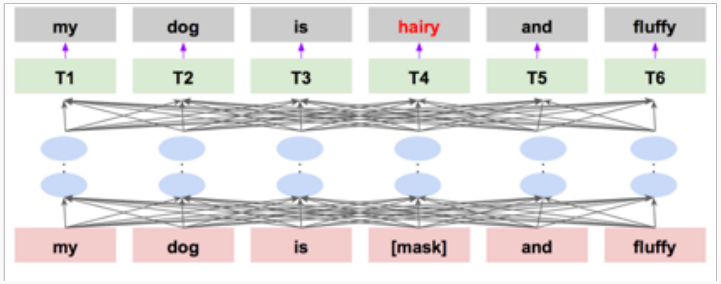
  - minimal task-specific model architecture/parameters?

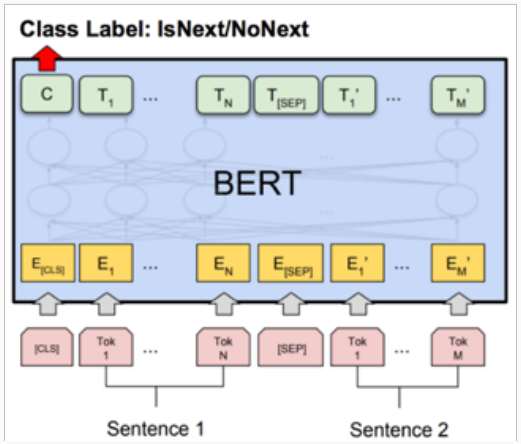


  - Pre-training: 4 days on 64 TPU chips
    Fine-tuning: very fast, only 3 or 4 epochs?
  - ” Massive Compute Is All You Need “??

BERT (cont.)

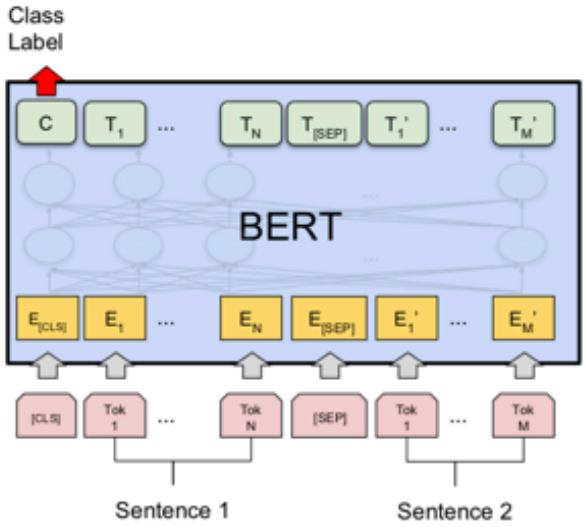
- New pre-training tasks
  1. Masked Language Model
  2. Next Sentence Prediction



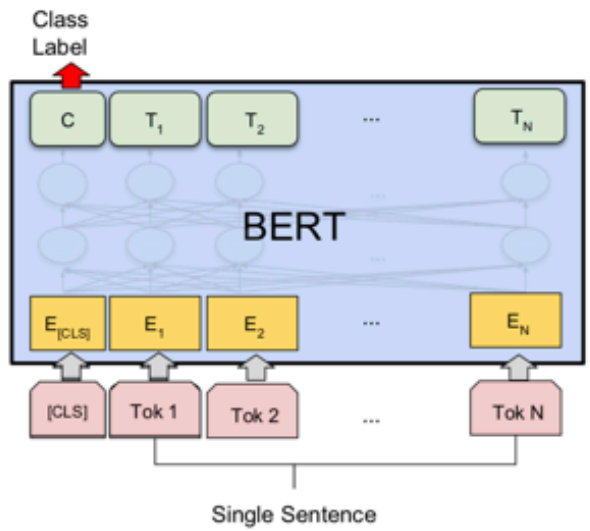


感謝許宗媛學姊提供

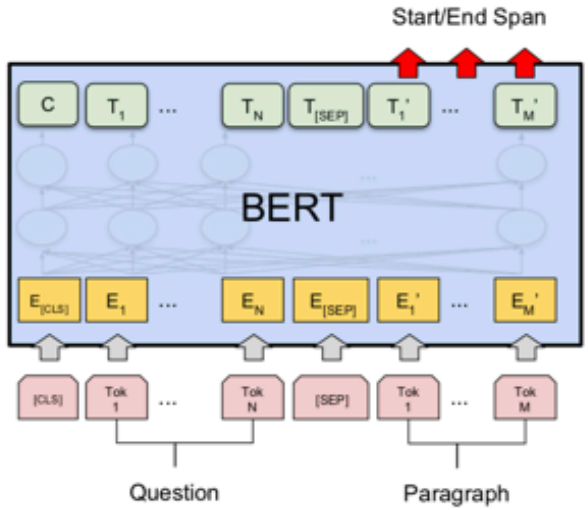
BERT (cont.)



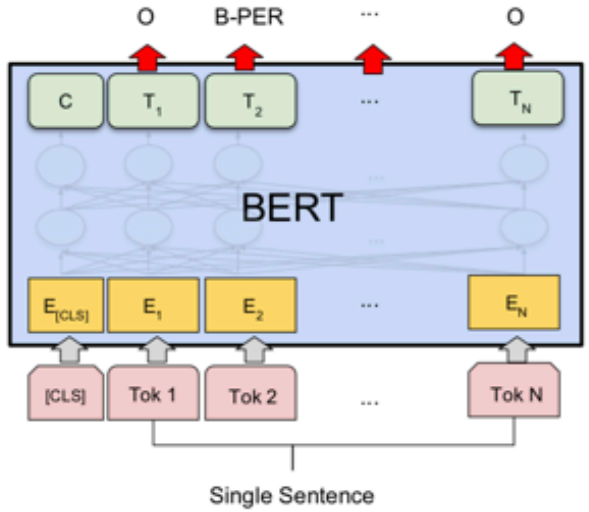
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1

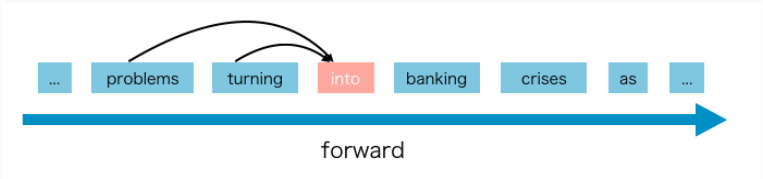


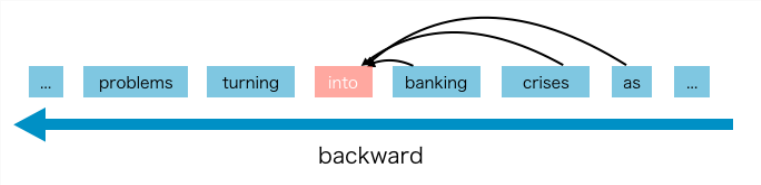
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

XLNet

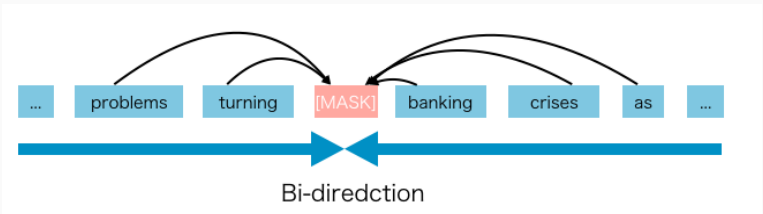
Autoregressive (AR) language modeling and Autoencoding (AE) language modeling

- AR
  - Pro: does not rely on data corruption
  - Con: can only be forward or backward



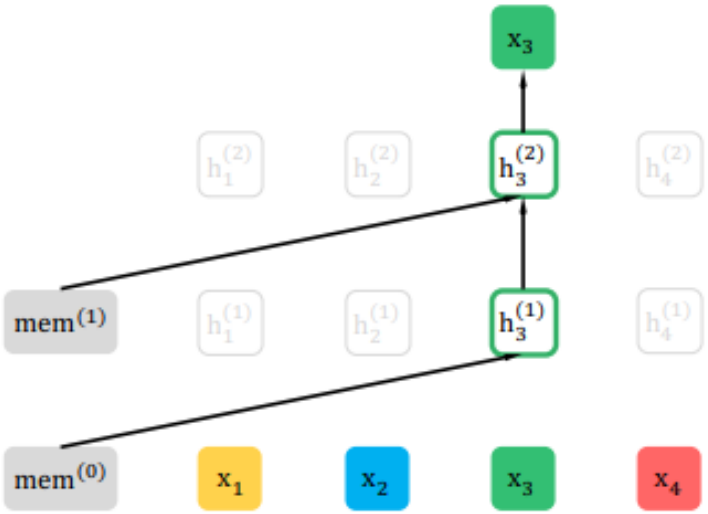


- AE
  - Pro: can use bidirectional information (Context dependency)
  - Con: pretrain-finetune discrepancy (Input noise), Independence Assumption

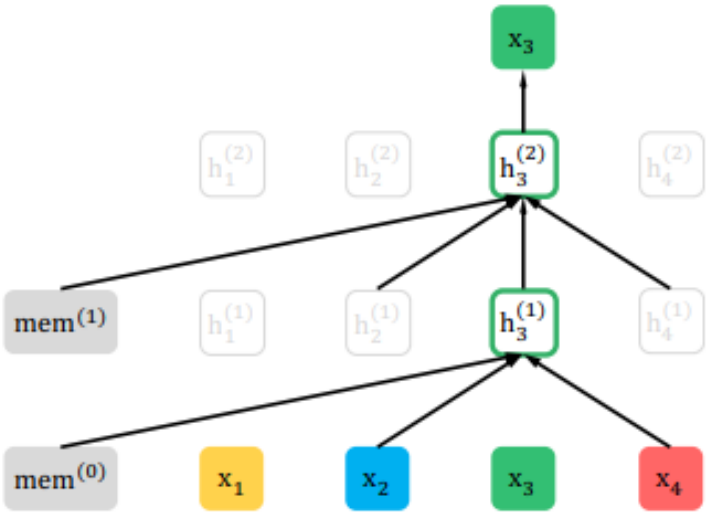


XLNet

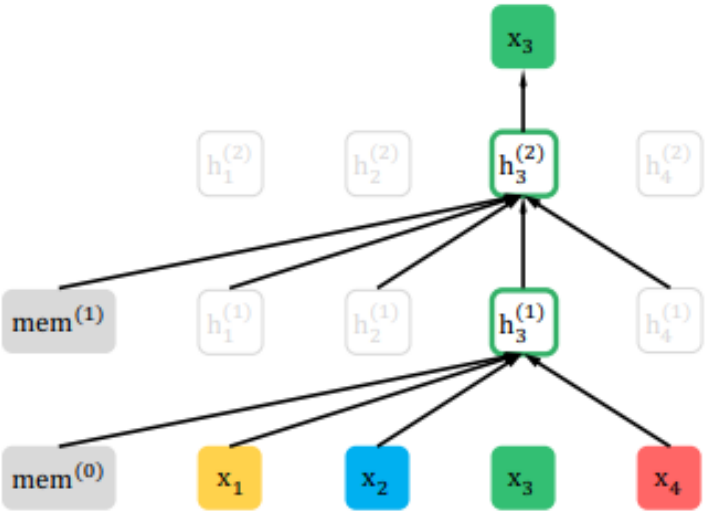
Permutation Language Modeling

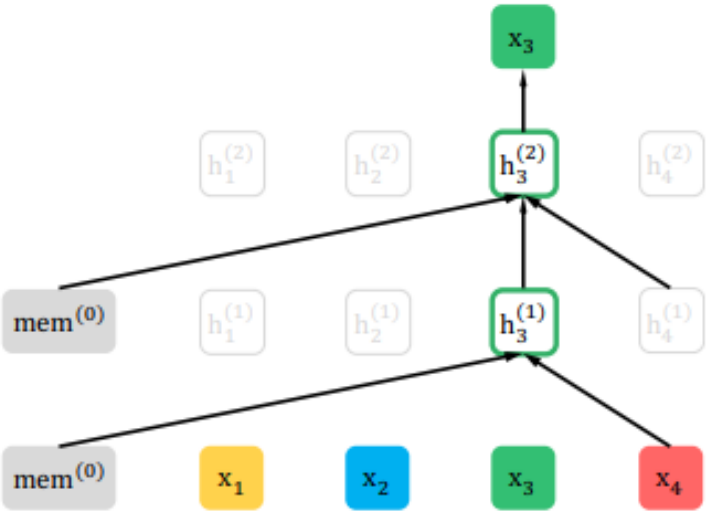


Factorization order: 3 → 2 → 4 → 1

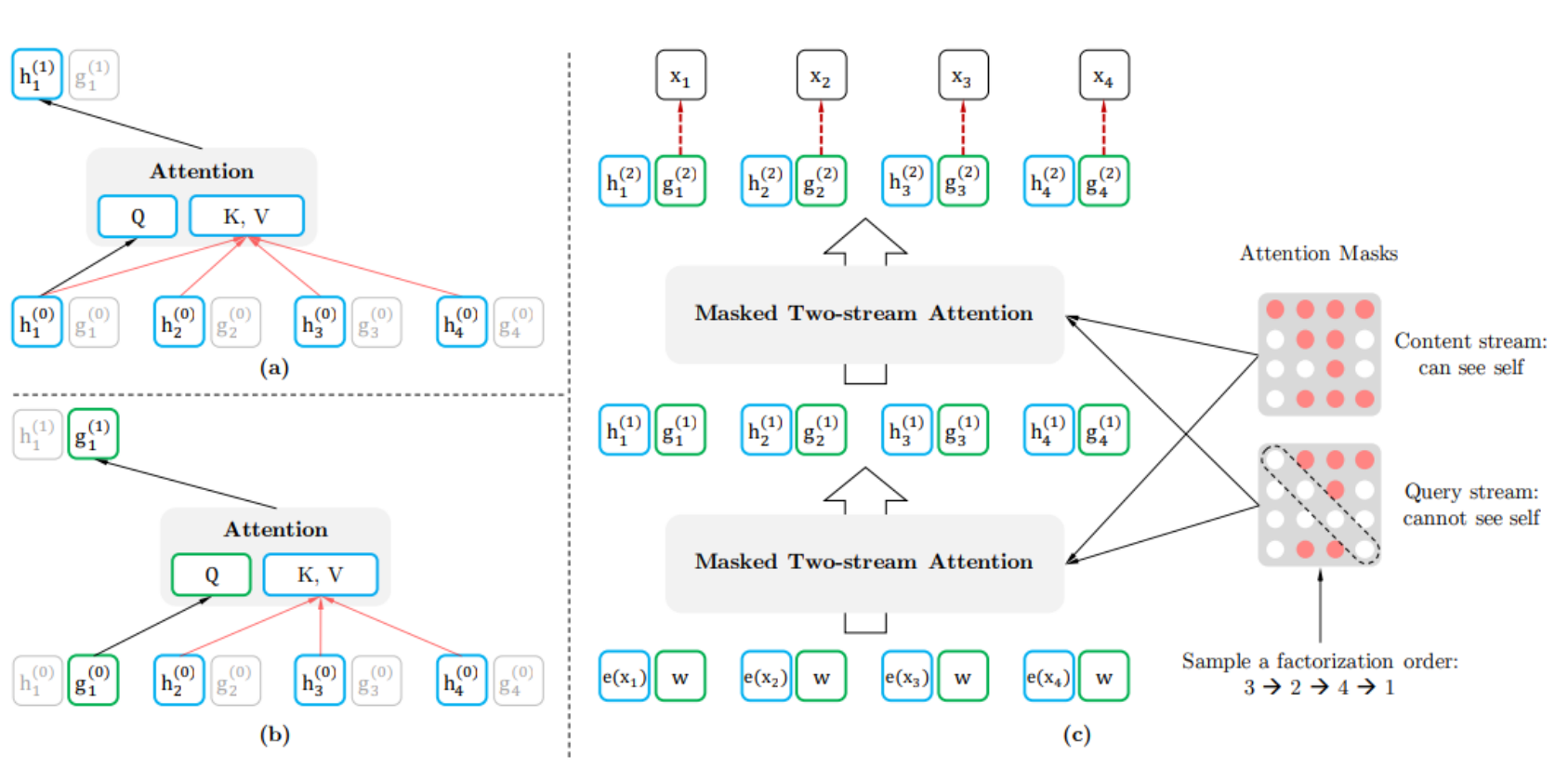


Factorization order: 2 → 4 → 3 → 1





Two-Stream Self-Attention



ERINE 2.0

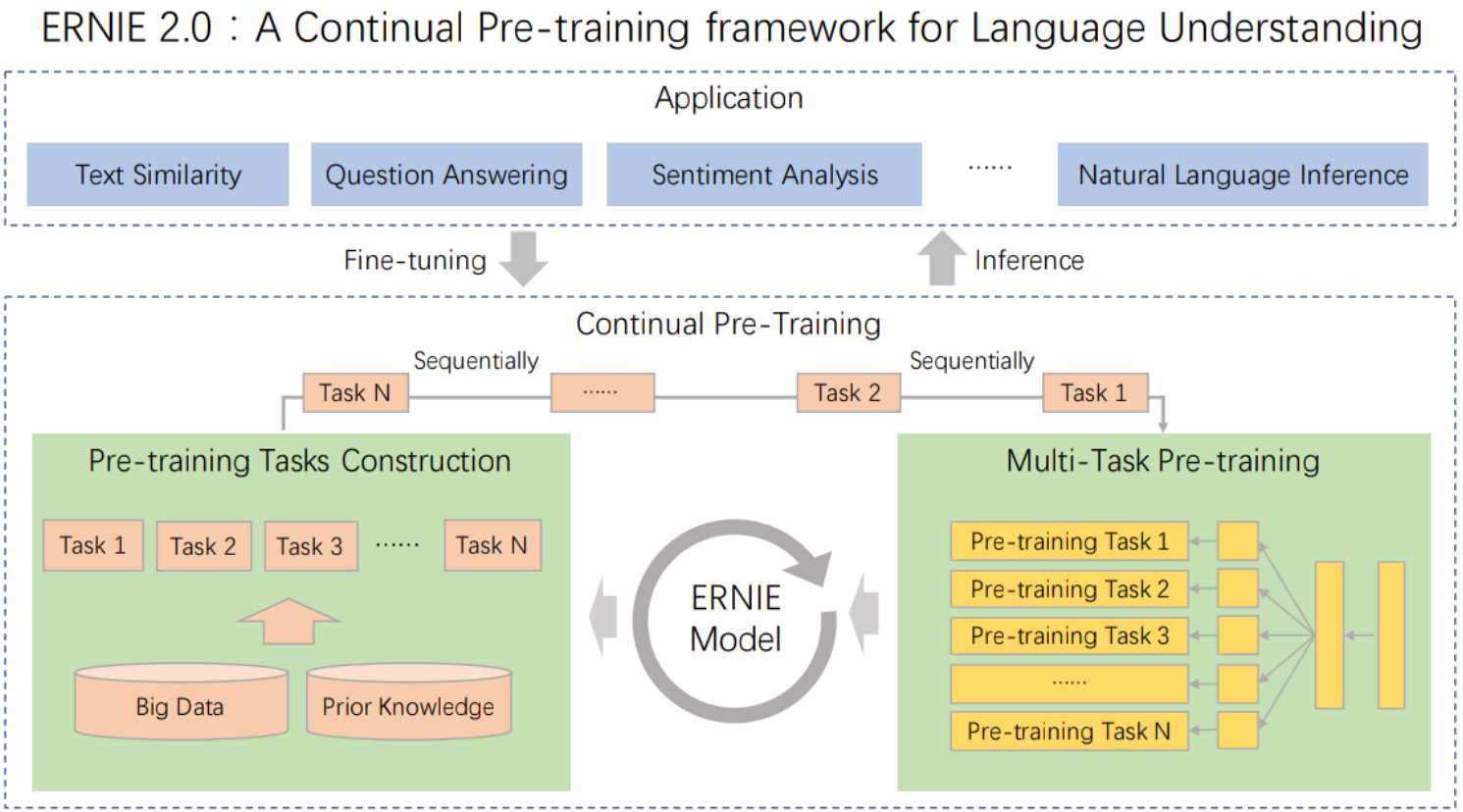


Figure 1: The framework of ERNIE 2.0, where the pre-training tasks can be incrementally constructed, the models are pre-trained through multi-task learning, and the pre-trained model is fine-tuned to adapt to various language understanding tasks.

ALBERT

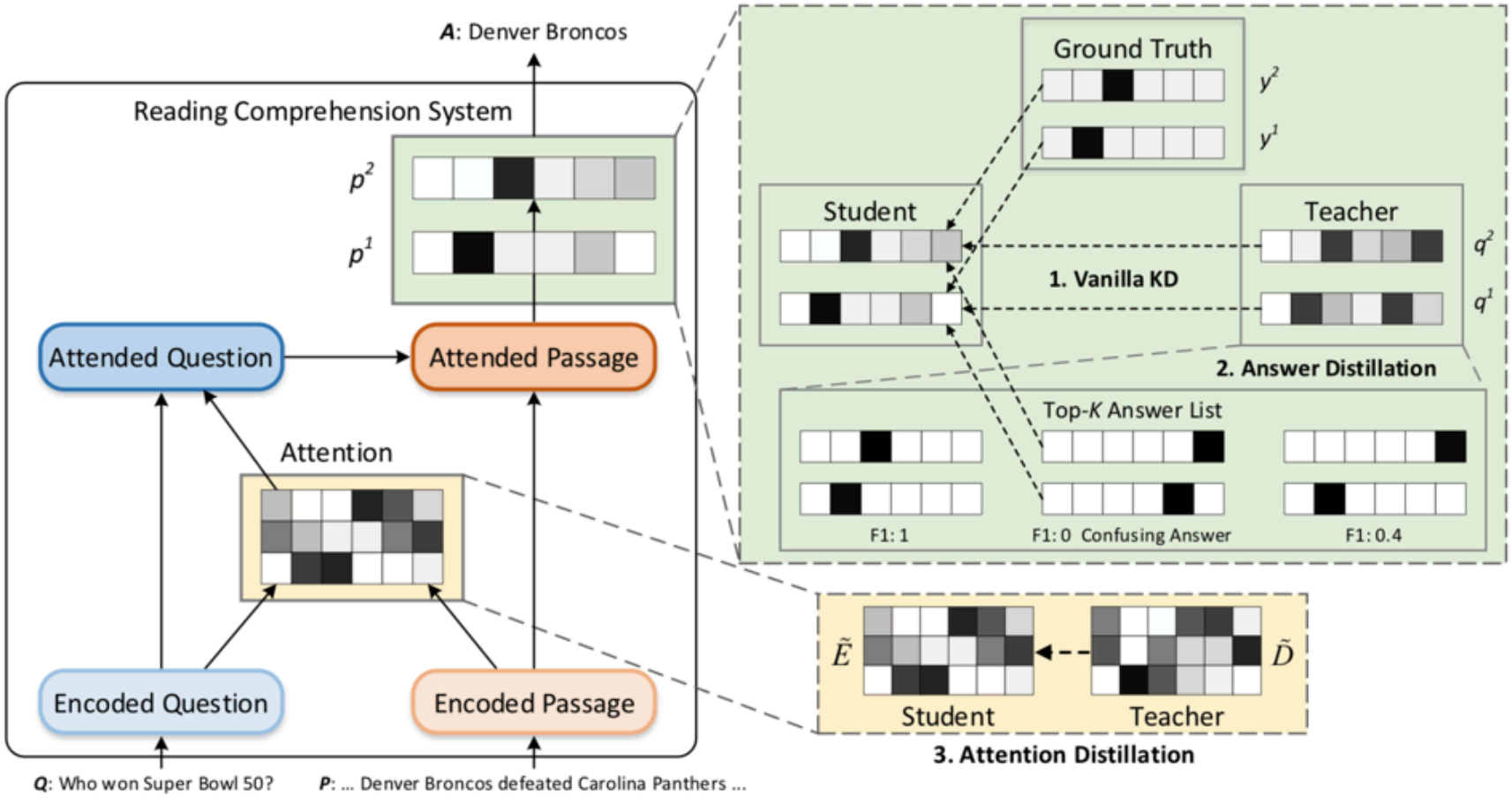
1. Factorized embedding parameterization: $O(VH) \rightarrow O(VE + EH)$
2. Cross-layer parameter sharing
3. Inter-sentence coherence loss

Models	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
BERT-large	86.6	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
XLNet-large	89.8	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa-large	90.2	94.7	**92.2**	86.6	96.4	**90.9**	68.0	92.4	-	-
ALBERT (1M)	90.4	95.2	92.0	88.1	96.8	90.2	68.7	92.7	-	-
ALBERT (1.5M)	**90.8**	**95.3**	**92.2**	**89.2**	**96.9**	**90.9**	**71.4**	**93.0**	-	-
Ensembles on test (from leaderboard as of Sept. 16, 2019)										
ALICE	88.2	95.7	**90.7**	83.5	95.2	92.6	**69.2**	91.1	80.8	87.0
MT-DNN	87.9	96.0	89.9	86.3	96.5	92.7	68.4	91.1	89.0	87.6
XLNet	90.2	98.6	90.3	86.3	96.8	93.0	67.8	91.6	90.4	88.4
RoBERTa	90.8	98.9	90.2	88.2	96.7	92.3	67.8	92.2	89.0	88.5
Adv-RoBERTa	91.1	98.8	90.3	88.7	96.8	93.1	68.0	92.4	89.0	88.8
ALBERT	**91.3**	**99.2**	90.5	**89.2**	**97.1**	**93.4**	69.1	**92.5**	**91.8**	**89.4**

Others

- Megatron-LM
- RoBERTa

How to Compress Model?

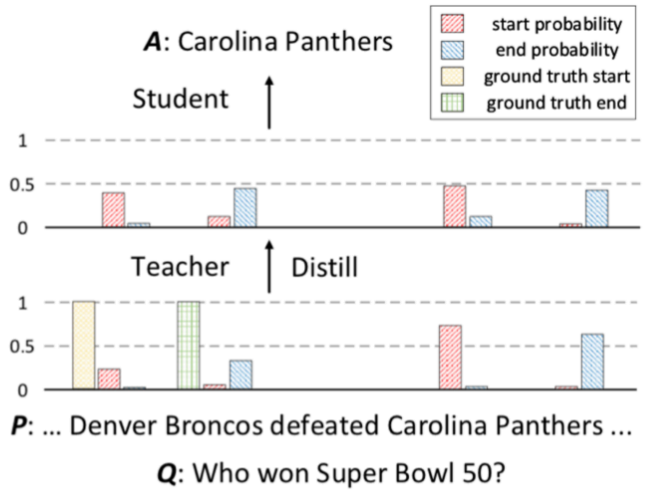


How to Compress Model? (cont.)

Passage: Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. The Champ Bowl 40 took place in Chicago.

Question1: Who won Super Bowl 50?
Question2: Where did Super Bowl 50 take place?

(a) Gold answers (red) versus confusing answers (blue) in the SQuAD and adversarial SQuAD datasets.



(b) The knowledge biased towards the confusing answer is distilled from the teacher to the student.

Figure 1: An illustration of confusing answer and biased distillation in machine reading comprehension.

Model	Dev		Test	
	EM	F1	EM	F1
LR Baseline[1]	40.0	51.0	40.4	51.0
FusionNet[2]	75.3	83.6	76.0	83.9
BiSAE[3]	77.9	85.6	78.6	85.8
R-Net+[4]	-	-	79.9	86.5
SLQA+[5]	80.0	87.0	80.4	87.0
QANet[6]	-	-	80.9	87.8
BiSAE (E)	79.6	86.6	81.0	87.4
R-Net+ (E)	-	-	82.6	88.5
SLQA+ (E)	82.0	88.4	82.4	88.6
QANet (E)	-	-	82.7	89.0
RMR[7]	78.9	86.3	79.5	86.6
RMR (E)	**81.2**	**87.9**	**82.3**	**88.5**
RMR + A2D	80.3	87.5	81.5	88.1

Model	AddSent		AddOneSent	
	EM	F1	EM	F1
LR Baseline	17.0	23.2	22.3	30.4
BiSAE	38.7	44.4	48.0	54.7
SLQA+	-	52.1	-	62.7
DCN+ (MINI)[1]	52.2	59.7	60.1	67.5
FusionNet (E)	46.2	51.4	54.7	60.7
SLQA+ (E)	-	54.8	-	64.2
RMR	53.0	58.5	60.9	67.0
RMR (E)	56.0	61.1	62.7	68.5
RMR + A2D	**56.0**	**61.3**	**63.3**	**69.3**

other compress models

- Distil Bert
- Tiny Bert

Machine Comprehension by Deep Learning

SQuAD2.0

SQuAD2.0

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 Mar 16, 2019	BERT + DAE + AoA (single model) *Joint Laboratory of HIT and iFLYTEK Research*	85.884	88.621
5 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	85.082	87.615
5 Mar 13, 2019	BERT + ConvLSTM + MTL + Verifier (single model) *Layer 6 AI*	84.924	88.204

Article: Endangered Species Act
Paragraph: "*. . . Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

Figure 1: Two unanswerable questions written by crowdworkers, along with plausible (but incorrect) answers. Relevant keywords are shown in blue.

BERT Again

- How to train on SQuAD2.0
  - Use token 0 ([CLS]) to emit logit for ”no answer”.
  - :No answer” directly competes with answer span.
  - Threshold is optimized on dev set.
- Synthetic self-training
  - Use seq2seq model to generate positive questions from context+answer.
  - Heuristically transform positive questions into negatives (i.e., ”no answer”/impossible).
  - Result: +3.0 F1/EM score, new state-of-the-art.

Machine Comprehension by Deep Learning

DataSet

DataSet

- **English DataSet**
  CliCR, CNN / Daily Mail, CoQA, HotpotQA, MS MARCO, MultiRC, NewsQA, QAngaroo, QuAC, RACE, SQuAD, Story Cloze Test, Recipe QA, NarrativeQA, DuoRC, DROP
- **Chinese DataSet**
  DRCD, DuReader
- **Audio DataSet**
  Spoken SQuAD, ODSQA, 科技大擂台-與 AI 對話

Conclusion

Conclusion

- Is modeling ”solved” in NLP?
- Do we still need KB?
- Near-term improvements in NLP will be mostly about making clever use of ”free” data.

Demo

- 簡答題: https://pse.is/LUC26
- 選擇題: https://pse.is/JFYEY

感謝吳宗翰、謝濬丞協助選擇題版本