HIT-SCIR

FOURTEENTH CIPS SUMMER SCHOOL &
ADVANCED TECHNOLOGY TUTORIAL

基于深度学习的机器阅读理解

DEEP LEARNING BASED MACHINE READING COMPREHENSION

YIMING CUI
ME@YMCUI.COM
JOINT LABORATORY OF HIT AND iFLYTEK RESEARCH (HFL)

2019-07-14

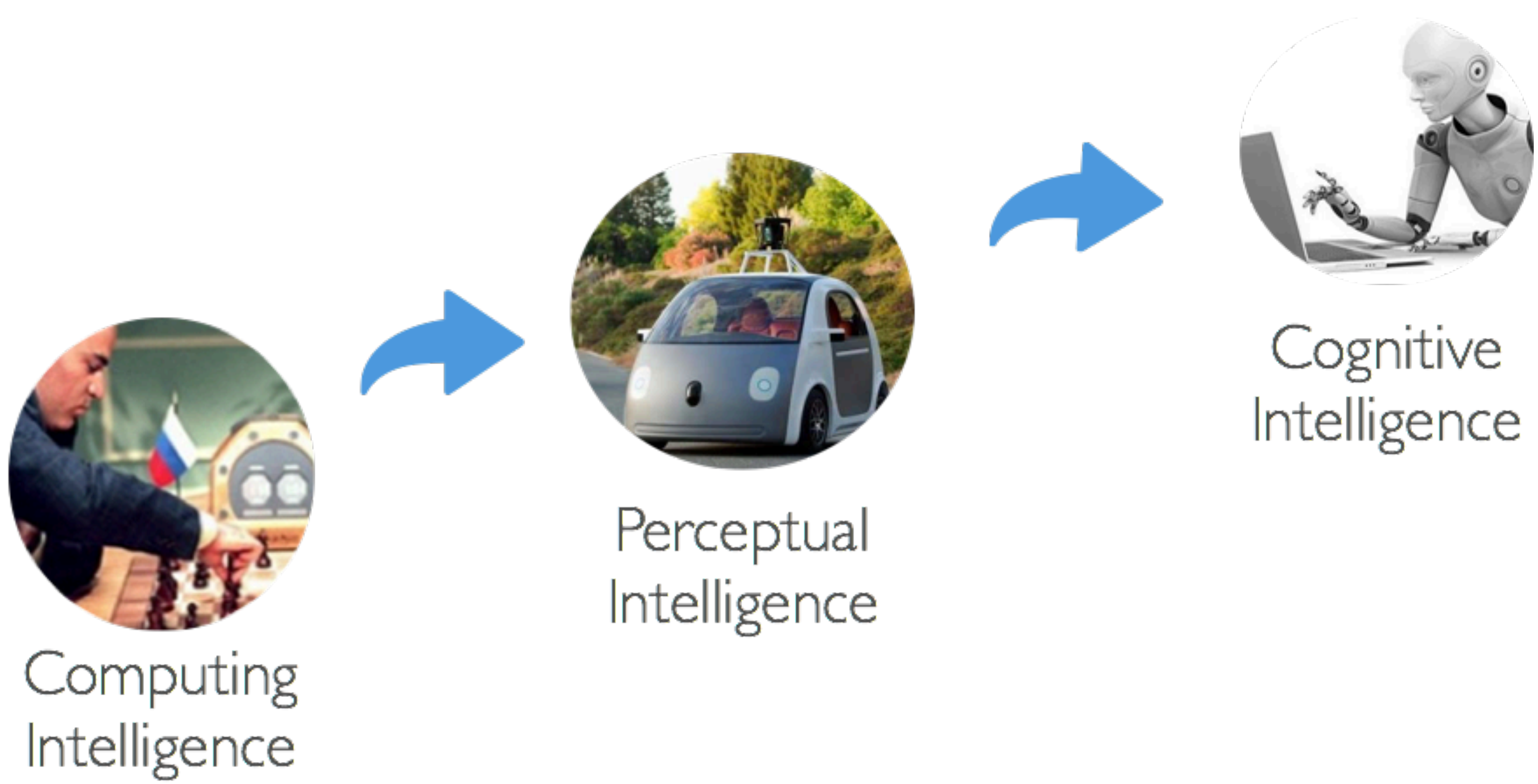
Outlines

- General Introductions to Machine Reading Comprehension (MRC)
- Machine Reading Comprehension in Deep Learning
  - Cloze-Style MRC
  - Span-Extraction MRC
  - MRC with Multiple-Choices
  - BERT-based MRC
- Chinese Machine Reading Comprehension
  - Chinese MRC Datasets
  - Chinese Pre-trained Models
- Conclusions

Introductions to MRC

Introduction

- To comprehend human language is essential in AI
- Machine Reading Comprehension (MRC) has attracted lots of attention from the NLP field



Introduction

- **Reading Comprehension**
- **Macro-view**
  - To learn and do reasoning with world knowledge and common knowledge while we are growing up

- **Micro-view**
  - Read an article/several articles, and answer the questions based on it

Introduction

- **Four key components in RC**
  - **→Document**
  - Question
  - Candidates
  - Answer

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

*Example is chosen from the MCTest dataset (Richardson et al., 2013)

Introduction

- **Four key components in RC**
  - Document
  - **→Question**
  - Candidates
  - Answer

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

*Example is chosen from the MCTest dataset (Richardson et al., 2013)

Introduction

- **Four key components in RC**
  - Document
  - Question
  - **→Candidates**
  - Answer

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

*Example is chosen from the MCTest dataset (Richardson et al., 2013)

Introduction

- **Four key components in RC**
  - Document
  - Question
  - Candidates
  - **→Answer**

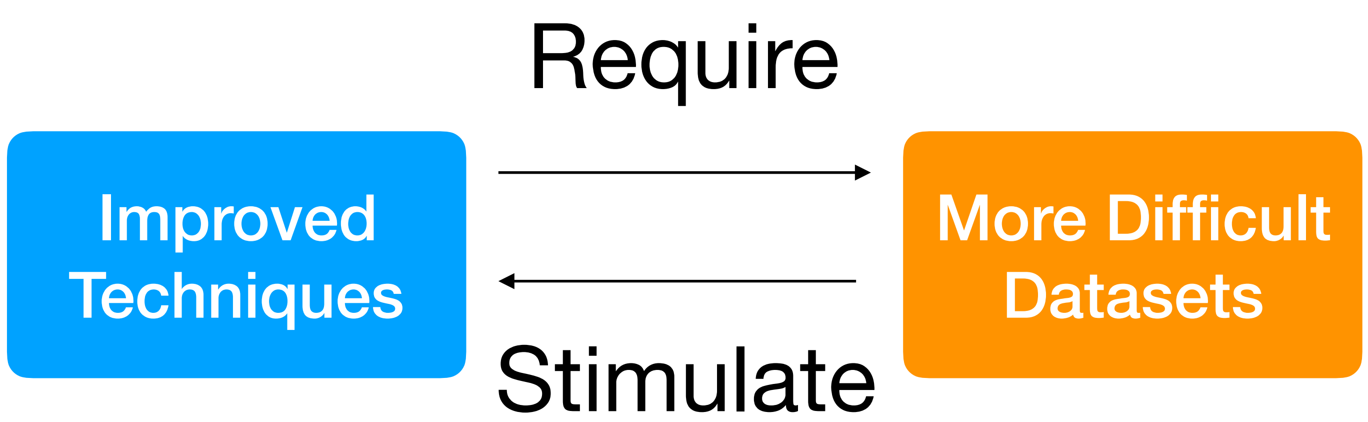
James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

*Example is chosen from the MCTest dataset (Richardson et al., 2013)

Introduction

- Why MRC became enormously popular in recent years?
- Mutual effect by
  - A growing interest in DL techniques
  - Availability of large-scale MRC data

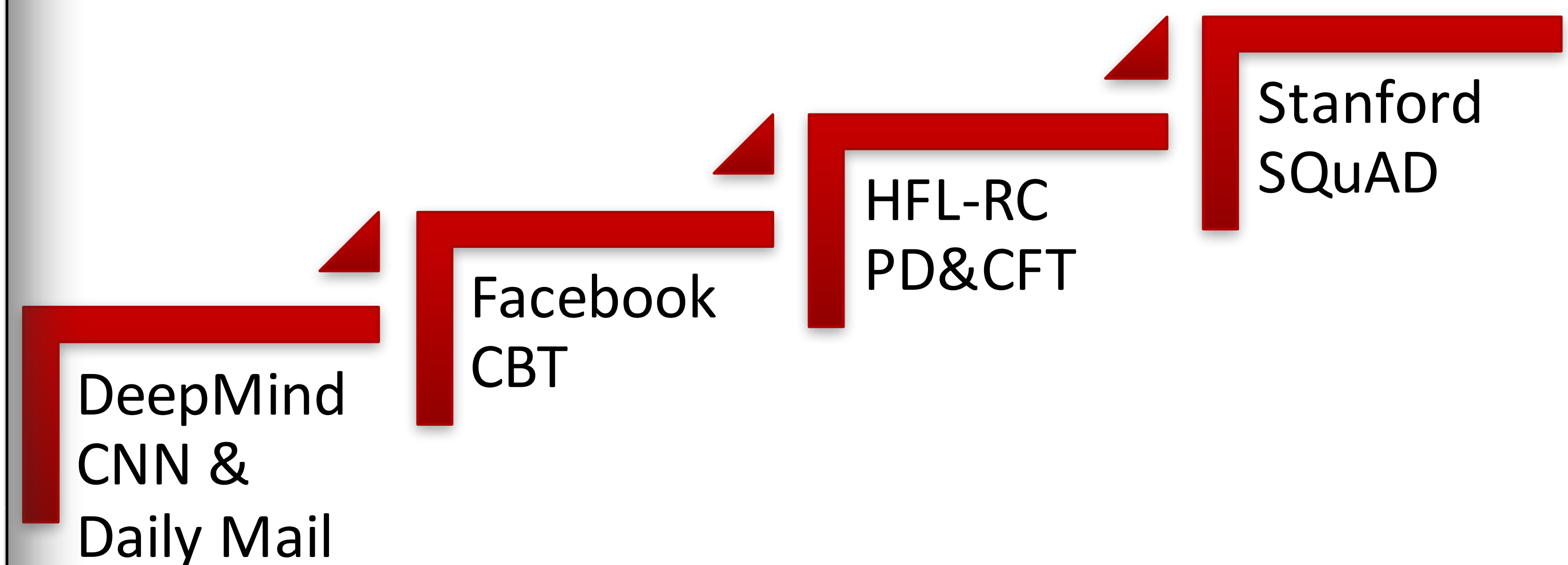


Introduction

- **MCTest (Richardson et al., EMNLP 2013)**

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane



Introduction

- **DeepMind CNN/DailyMail (Hermann et al., NIPS 2015)**

	Original Version	Anonymised Version
Context	The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broad-caster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." . . .	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " . . .
Query	Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer	Oisin Tymon	*ent193*

Stanford
SQuAD

Introduction

- **Facebook CBT (Hill et al., ICLR 2016)**

"Well, Miss Maxwell, I think it only fair to tell you that you may have trouble with those boys when they do come. Forewarned is forearmed, you know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and it is hard to corner him."

"Are the boys big ?" queried Esther anxiously.

"Yes. Thirteen and fourteen and big for their age. You can't whip 'em -- that is the trouble. A man might, but they'd twist you around their fingers. You'll have your hands full, I'm afraid. But maybe they'll behave all right after all."

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that Mr. Baxter had exaggerated matters a little.

```
S: 1 Mr. Cropper was opposed to our hiring you .
   2 Not , of course , that he had any personal objection to you , but he is set
   against female teachers , and when a Cropper is set there is nothing on earth can
   change him .
   3 He says female teachers ca n't keep order .
   4 He 's started in with a spite at you on general principles , and the boys know
   it .
   5 They know he 'll back them up in secret , no matter what they do , just to prove
   his opinions .
   6 Cropper is sly and slippery , and it is hard to corner him . ''
   7 `` Are the boys big ? ''
   8 queried Esther anxiously .
   9 `` Yes .
   10 Thirteen and fourteen and big for their age .
   11 You ca n't whip 'em -- that is the trouble .
   12 A man might , but they 'd twist you around their fingers .
   13 You 'll have your hands full , I 'm afraid .
   14 But maybe they 'll behave all right after all . ''
   15 Mr. Baxter privately had no hope that they would , but Esther hoped for the
   best.
   16 She could not believe that Mr. Cropper would carry his prejudices into a
   personal application .
   17 This conviction was strengthened when he overtook her walking from school the
   next day and drove her home .
   18 He was a big , handsome man with a very suave , polite manner .
   19 He asked interestedly about her school and her work , hoped she was getting on
   well , and said he had two young rascals of his own to send soon .
   20 Esther felt relieved .
q: She thought that Mr. _______ had exaggerated matters a little .
C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.
a: Baxter
```

Introduction

- **PD&CFT (Cui et al., COLING 2016)**

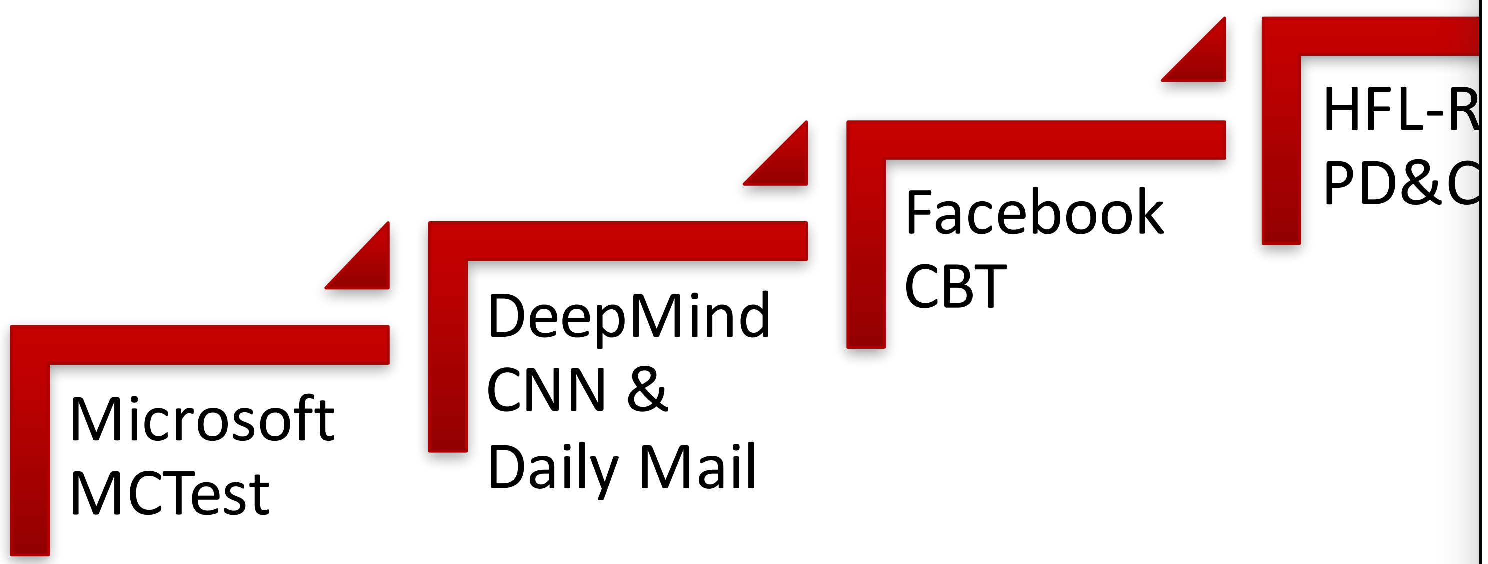
Microsoft MCTest

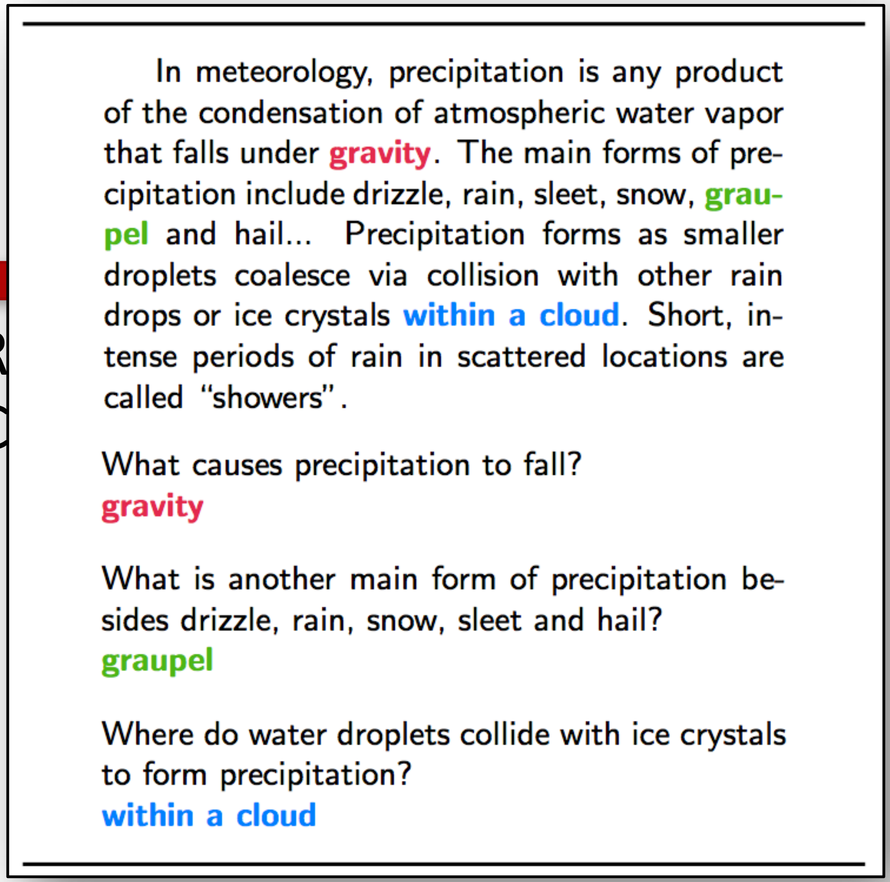
Deep CNN & Daily Mail

<table>
<tr><td>Document</td><td>1 ||| 人民网 1月 1日 讯 据 《 纽约 时报 》 报道 ， 美国 华尔街 股市 在 2013年 的 最后 一 天 继续 上涨 ， 和 全球 股市 一样 ， 都 以 最高 纪录 或 接近 最高 纪录 结束 本年 的 交易 。
2 ||| 《 纽约 时报 》 报道 说 ， 标普 500 指数 今年 上升 29.6% ， 为 1997年 以来 的 最 大 涨幅 ；
3 ||| 道琼斯 工业 平均 指数 上升 26.5% ， 为 1996年 以来 的 最 大 涨幅 ；
4 ||| 纳斯达克 上涨 38.3% 。
5 ||| 就 12月 31日 来说 ， 由于 就业 前景 看好 和 经济 增长 明年 可能 加速 ， 消费者 信心 上升 。
6 ||| 工商 协进会 报告 ， 12月 消费者 信心 上升 到 78.1 ， 明显 高于 11月 的 72 。
7 ||| 另 据 《 华尔街 日报 》 报道 ， 2013年 是 1995年 以来 美国 股市 表现 最 好 的 一 年 。
8 ||| 这 一 年 里 ， 投资 美国 股市 的 明智 做法 是 追 着 " 傻钱 " 跑 。
9 ||| 所谓 的 " 傻钱 " **X** ， 其实 就 是 买 入 并 持有 美国 股票 这样 的 普通 组合 。
10 ||| 这个 策略 要 比 对冲 基金 和 其它 专业 投资者 使用 的 更为 复杂 的 投资 方法 效果 好 得 多 。</td></tr>
<tr><td>Query</td><td>所谓 的 " 傻钱 " X ， 其实 就 是 买 入 并 持有 美国 股票 这样 的 普通 组合 。</td></tr>
<tr><td>Answer</td><td>策略</td></tr>
</table>

Introduction

- **SQuAD (Rajpurkar et al., EMNLP 2016)**





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

Cloze-Style Machine Reading Comprehension

Cloze-style RC

- **Definition of cloze-style RC**
  - Document: the same as the general RC
  - Query: a sentence with a blank
  - Candidate (optional): several candidates to fill in
  - Answer: a single word that exactly matches the query
    - The answer word should appear in the document

Original Version
Context The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." …
Query Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.
Answer Oisin Tymon

*Example is chosen from the CNN dataset (Hermann et al., 2015)

Cloze-style RC

- **CBT Dataset (Hill et al., ICLR 2016)**

Step1: Choose 21 consecutive sentences

"Well, Miss Maxwell, ... ve trouble with those boys when they do come. Forewarned is forearmed, y... know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and it is hard to corner him."

"Are the boys big ?" queried Esther anxiously.

"Yes. Thirteen and fourteen and big for their age. You can't whip 'em -- that is the trouble. A man might, but they'd twist you around their fingers. You'll have your hands full, I'm afraid. But maybe they'll behave all right after all."

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that Mr. Baxter had exaggerated matters a little.

S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''
8 queried Esther anxiously .
9 `` Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly about her school and her work , hoped she was getting on well , and said he had two young rascals of his own to send soon .
20 Esther felt relieved .

q: She thought that Mr. ________ had exaggerated matters a little .

C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

a: Baxter

Cloze-style RC

- **CBT Dataset (Hill et al., ICLR 2016)**

Step1: Choose 21 consecutive sentences

Step2: Choose first 20 sentences as Context

"Well, Miss Maxwell, ... ve trouble with those boys when they do come. Forewarned is forearmed, y... know. Mr. Cropper was opposed to our hiring you. Not, of course, that ...e had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and it is hard to corner him."

"Are the boys big ?" queried Esther anxiously.

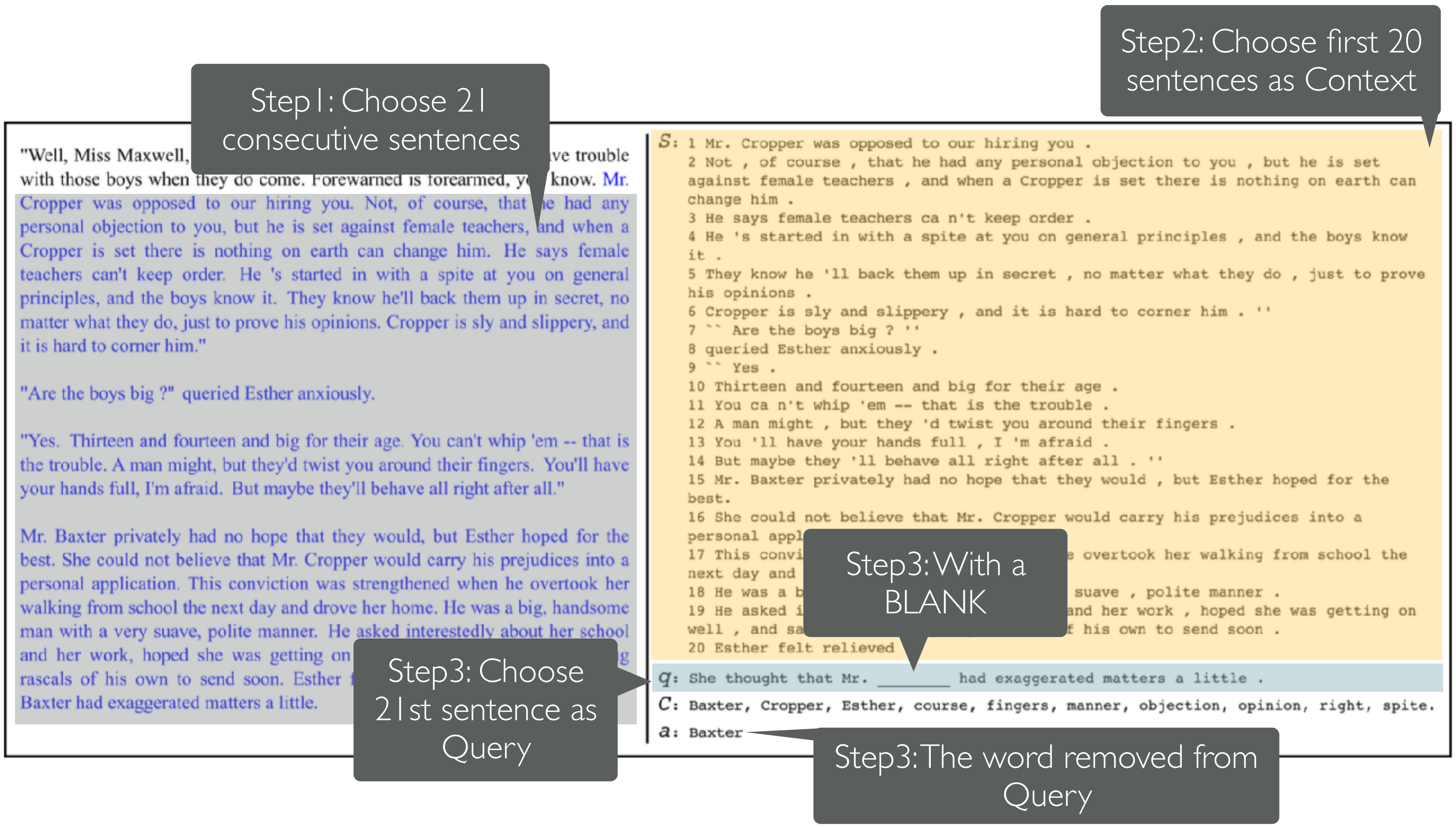
"Yes. Thirteen and fourteen and big for their age. You can't whip 'em -- that is the trouble. A man might, but they'd twist you around their fingers. You'll have your hands full, I'm afraid. But maybe they'll behave all right after all."

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that Mr. Baxter had exaggerated matters a little.

```
S: 1 Mr. Cropper was opposed to our hiring you .
   2 Not , of course , that he had any personal objection to you , but he is set
   against female teachers , and when a Cropper is set there is nothing on earth can
   change him .
   3 He says female teachers ca n't keep order .
   4 He 's started in with a spite at you on general principles , and the boys know
   it .
   5 They know he 'll back them up in secret , no matter what they do , just to prove
   his opinions .
   6 Cropper is sly and slippery , and it is hard to corner him . ''
   7 `` Are the boys big ? ''
   8 queried Esther anxiously .
   9 `` Yes .
   10 Thirteen and fourteen and big for their age .
   11 You ca n't whip 'em -- that is the trouble .
   12 A man might , but they 'd twist you around their fingers .
   13 You 'll have your hands full , I 'm afraid .
   14 But maybe they 'll behave all right after all . ''
   15 Mr. Baxter privately had no hope that they would , but Esther hoped for the
   best.
   16 She could not believe that Mr. Cropper would carry his prejudices into a
   personal application .
   17 This conviction was strengthened when he overtook her walking from school the
   next day and drove her home .
   18 He was a big , handsome man with a very suave , polite manner .
   19 He asked interestedly about her school and her work , hoped she was getting on
   well , and said he had two young rascals of his own to send soon .
   20 Esther felt relieved .
q: She thought that Mr. ________ had exaggerated matters a little .
C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.
a: Baxter
```

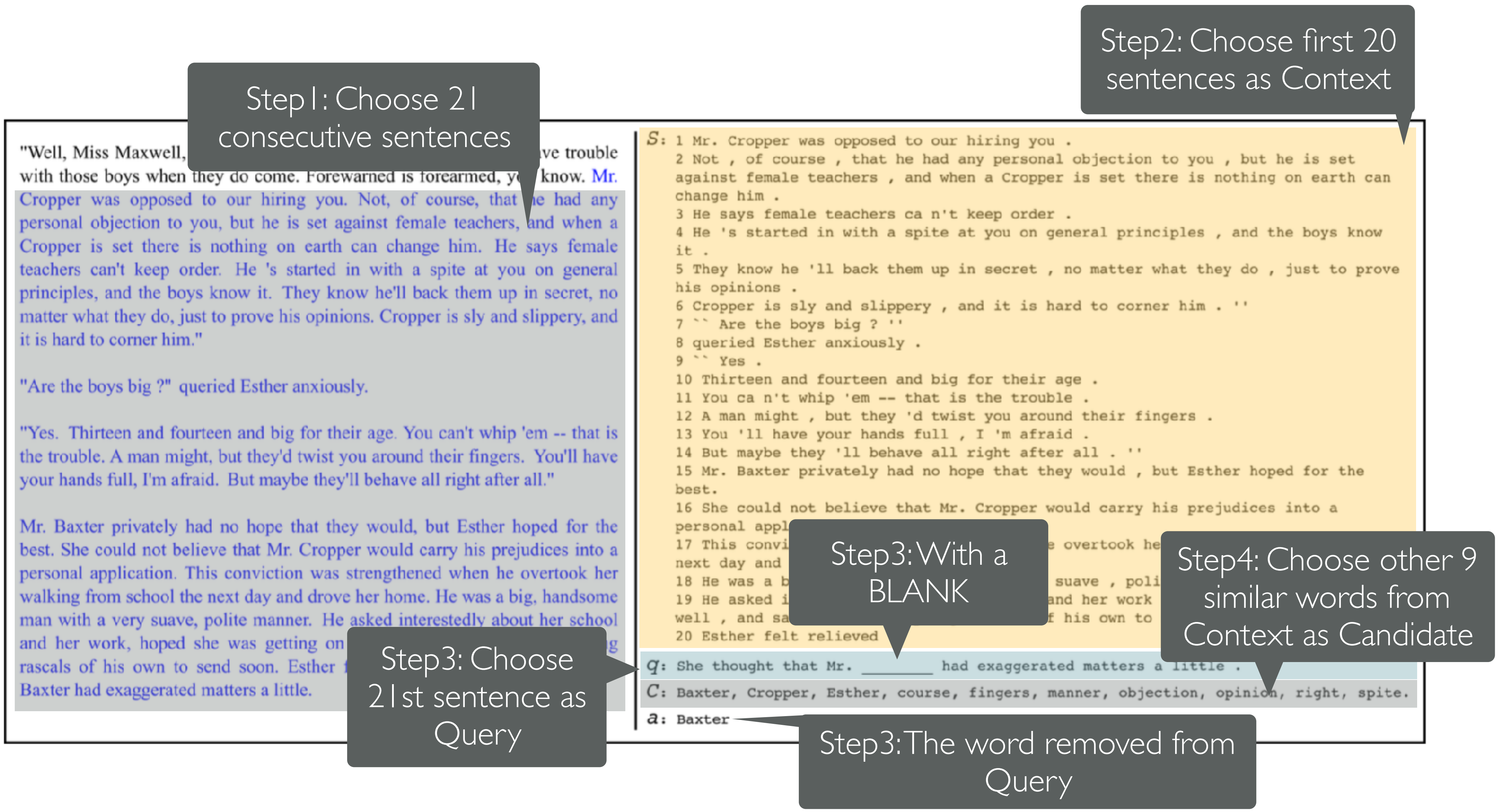
Cloze-style RC

- **CBT Dataset (Hill et al., ICLR 2016)**



Cloze-style RC

- **CBT Dataset (Hill et al., ICLR 2016)**



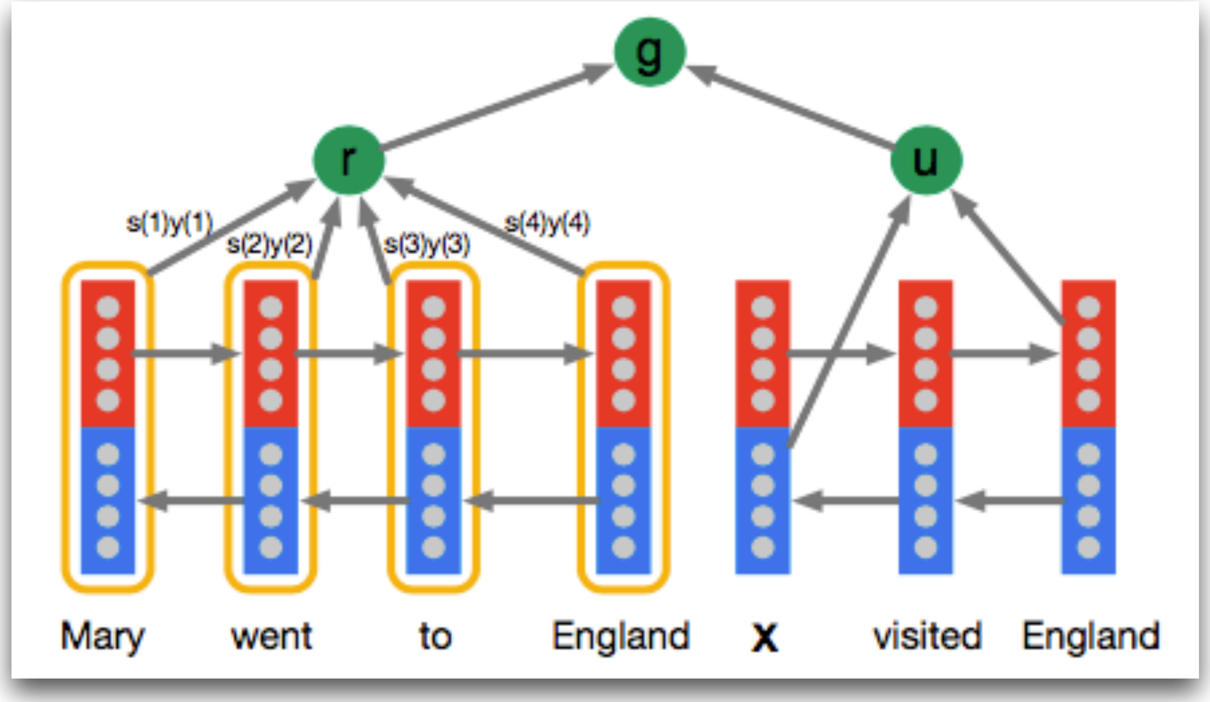
Related Works

- **Predictions on full vocabulary**
  - Attentive Reader (Hermann et al., 2015)
  - Stanford AR (Chen et al., 2016)
- **Pointer-wise predictions (Vinyals et al., 2015)**
  - Attention Sum Reader (Kadlec et al., 2016)
  - Gated-attention Reader (Dhingra et al., 2017)
  - Consensus Attention Reader (Cui et al., 2016)
  - Attention-over-Attention Reader (Cui et al., 2017)

Attentive Reader

- **Teaching Machines to Read and Comprehend (Hermann et al., NIPS 2015)**

- Propose **attention-based neural networks** for reading comprehension

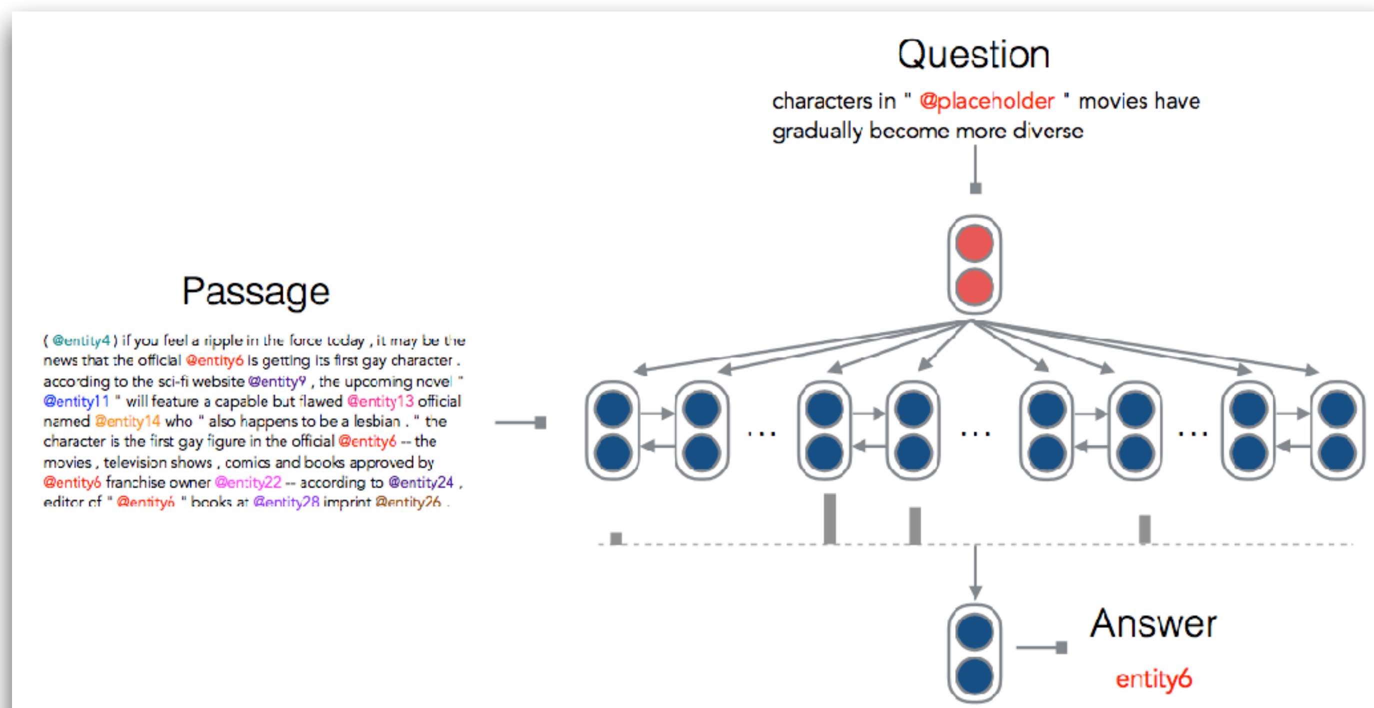


$$m(t) = \tanh\left(W_{ym} y_d(t) + W_{um} u\right),$$
$$s(t) \propto \exp\left(\mathrm{w}_{ms}^{\intercal} m(t)\right),$$
$$r = y_d s,$$

$$g^{\mathrm{AR}}(d, q) = \tanh\left(W_{rg} r + W_{ug} u\right).$$

Stanford AR

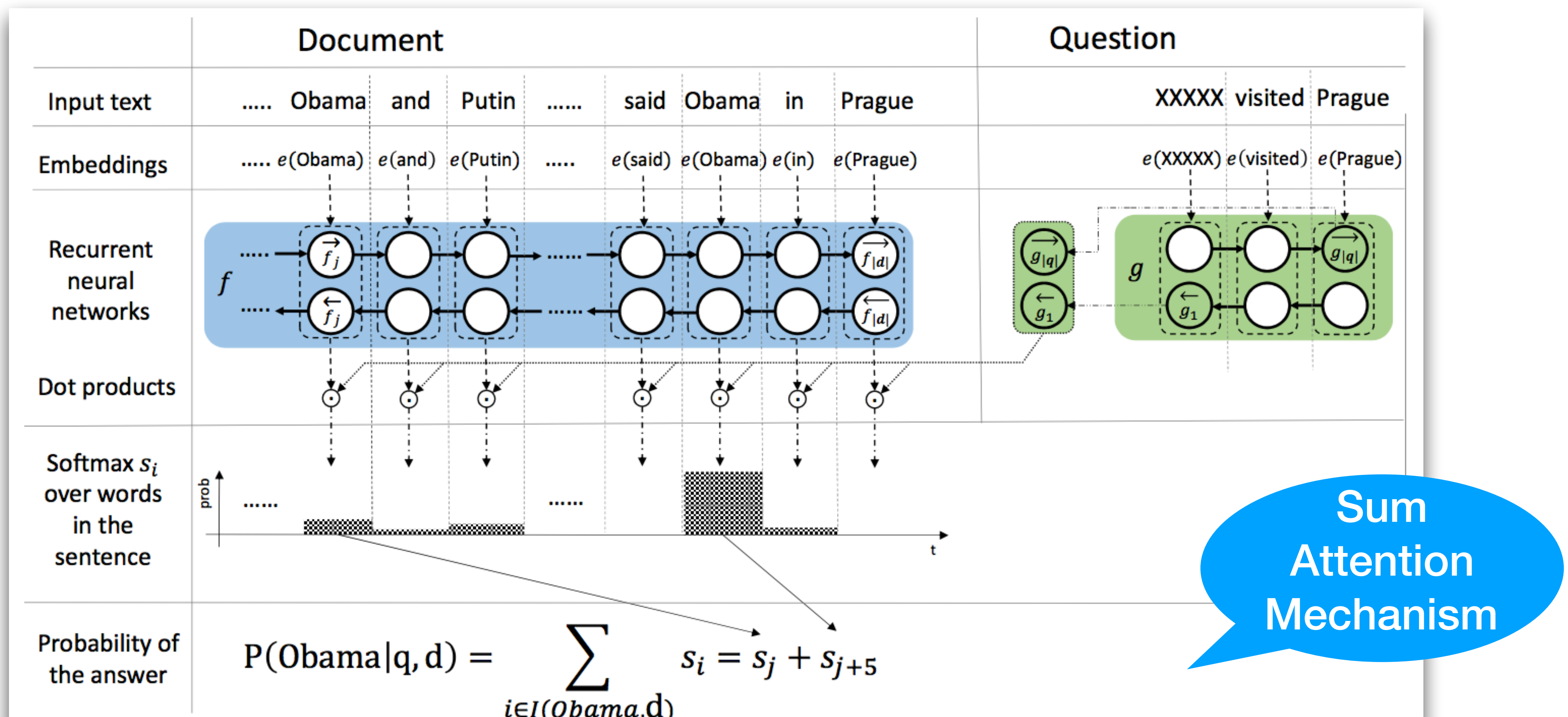
- **A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task (Chen et al., ACL 2016)**
- Nothing special in NN model, but provides **valuable insights on the CNN/ DailyMail datasets**



1) CNN/DailyMail dataset is noisy
2) Current NN models have almost reached CEILING performance
3) Requires less reasoning and inference

Attention Sum Reader

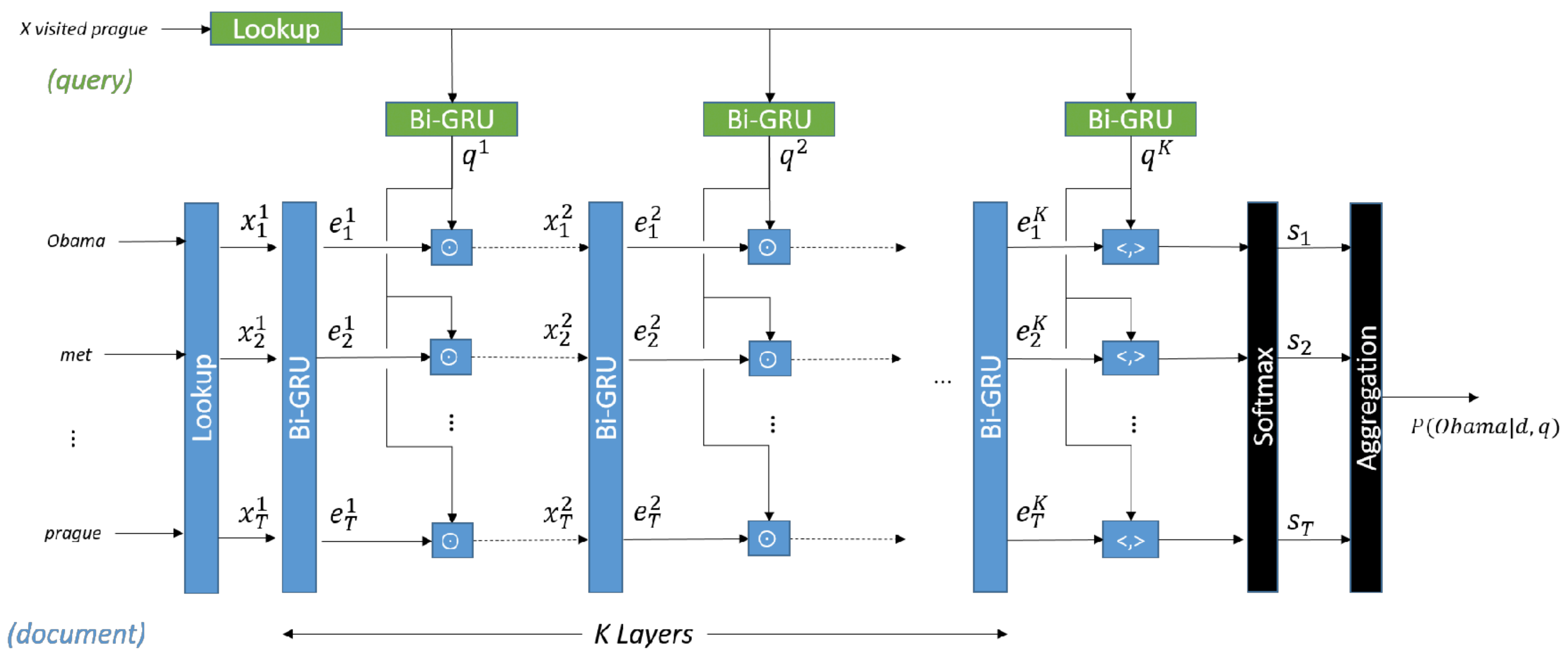
- **Text Understanding with the Attention Sum Reader Network (Kadlec et al., ACL 2016)**
- Propose to **utilize and improve Pointer Network** (Vinyals et al., 2015) in RC



Gated-Attention Reader

- **Gated-Attention Reader for Text Comprehension (Dhingra et al., ACL 2017)**

- Propose to use **multiple hops** for refining attended representations



CAS Reader

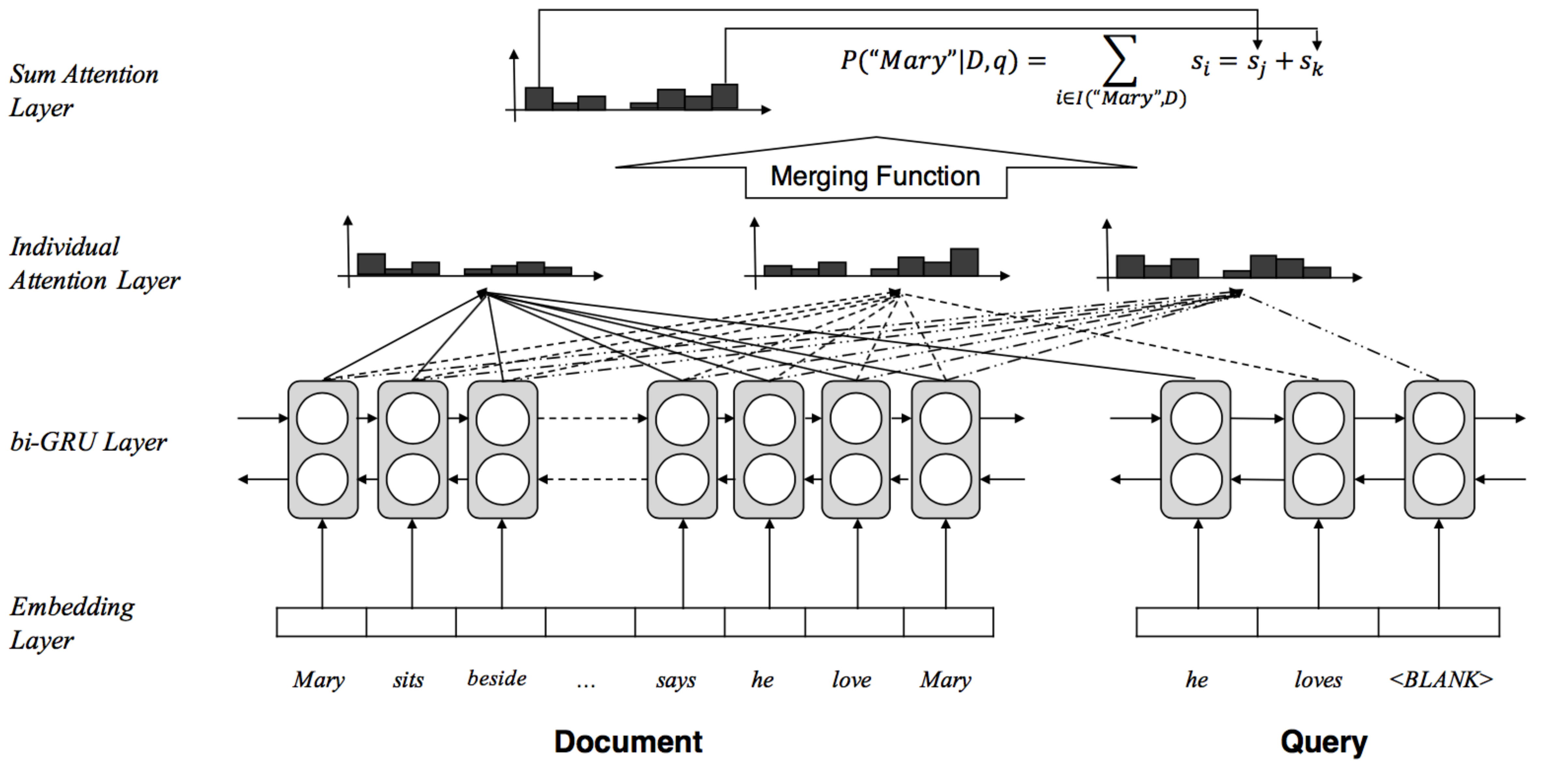
- **Consensus Attention-based Neural Networks for Chinese Reading Comprehension**
  - We propose an extension to AS Reader (Kadlec et al., 2016), which is a popular framework on the cloze-style reading comprehension task
  - Instead of blending query representations into one, we can take EVERY individual query words to generate document-level attention respectively

Consensus Attention-based Neural Networks for Chinese Reading Comprehension

Yiming Cui[†*], **Ting Liu**[‡], **Zhipeng Chen**[†], **Shijin Wang**[†] and **Guoping Hu**[†]
[†]iFLYTEK Research, Beijing, China
[‡]Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin, China
[†]{ymcui,zpchen,sjwang3,gphu}@iflytek.com
[‡]tliu@ir.hit.edu.cn

CAS Reader

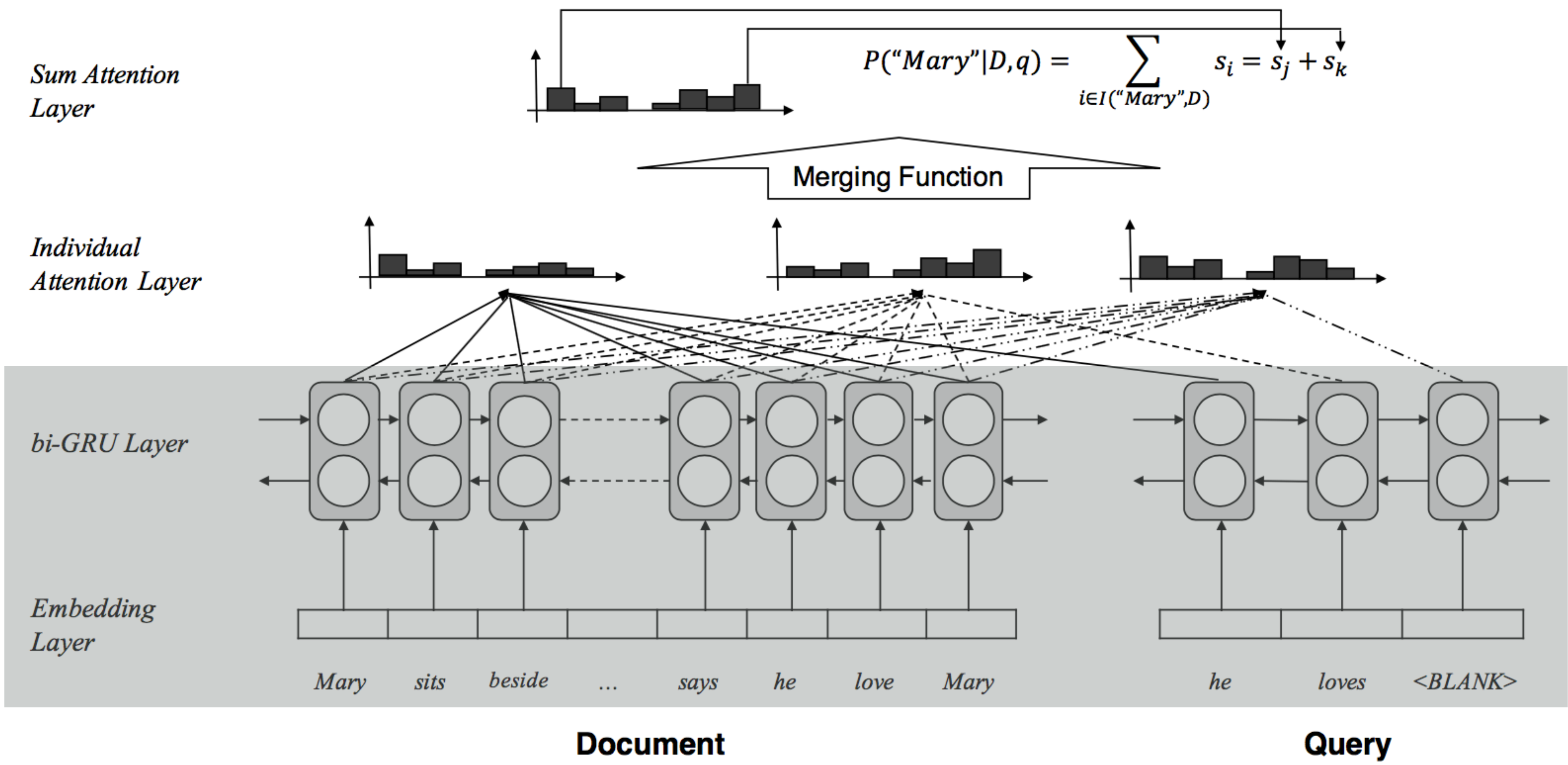
- **Neural Architecture**



Cui et al., COLING 2016. Consensus Attention-based Neural Networks for Chinese Reading Comprehension

CAS Reader

- **Step 1**: Transform document and query into contextual representations using GRU (Cho et al., 2014)



$$e(x) = W_e * x, \ where \ \ x \in \mathcal{D}, \mathcal{Q} \qquad (1)$$

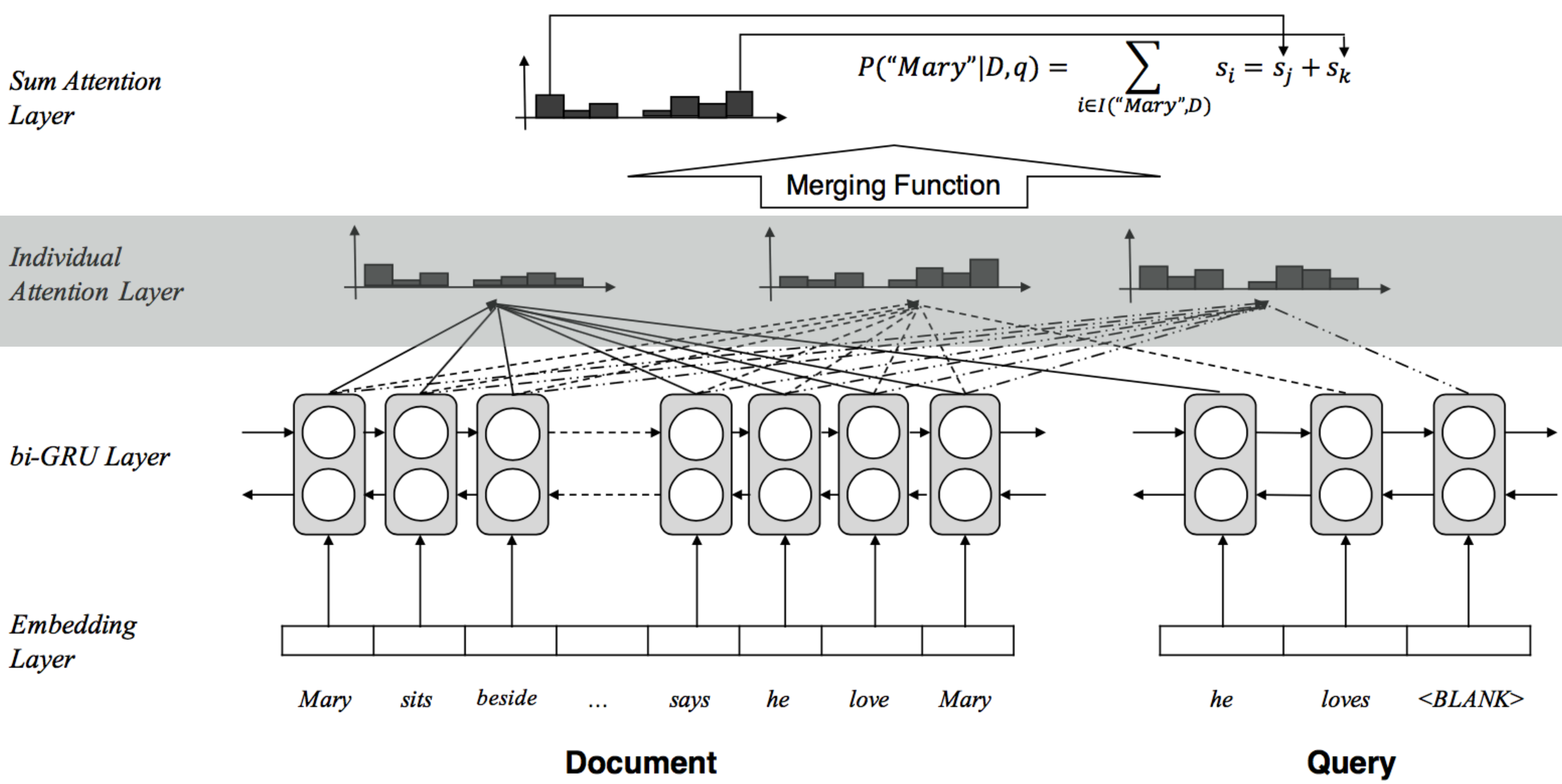
$$\overrightarrow{h_s(x)} = \overrightarrow{GRU}(e(x)) \ ; \ \overleftarrow{h_s(x)} = \overleftarrow{GRU}(e(x)) \qquad (2)$$

$$h_s(x) = [\overrightarrow{h_s(x)}; \overleftarrow{h_s(x)}] \qquad (3)$$

Cui et al., COLING 2016. Consensus Attention-based Neural Networks for Chinese Reading Comprehension

CAS Reader

- **Step 2**: Generate several document-level attentions in terms of every word in the query

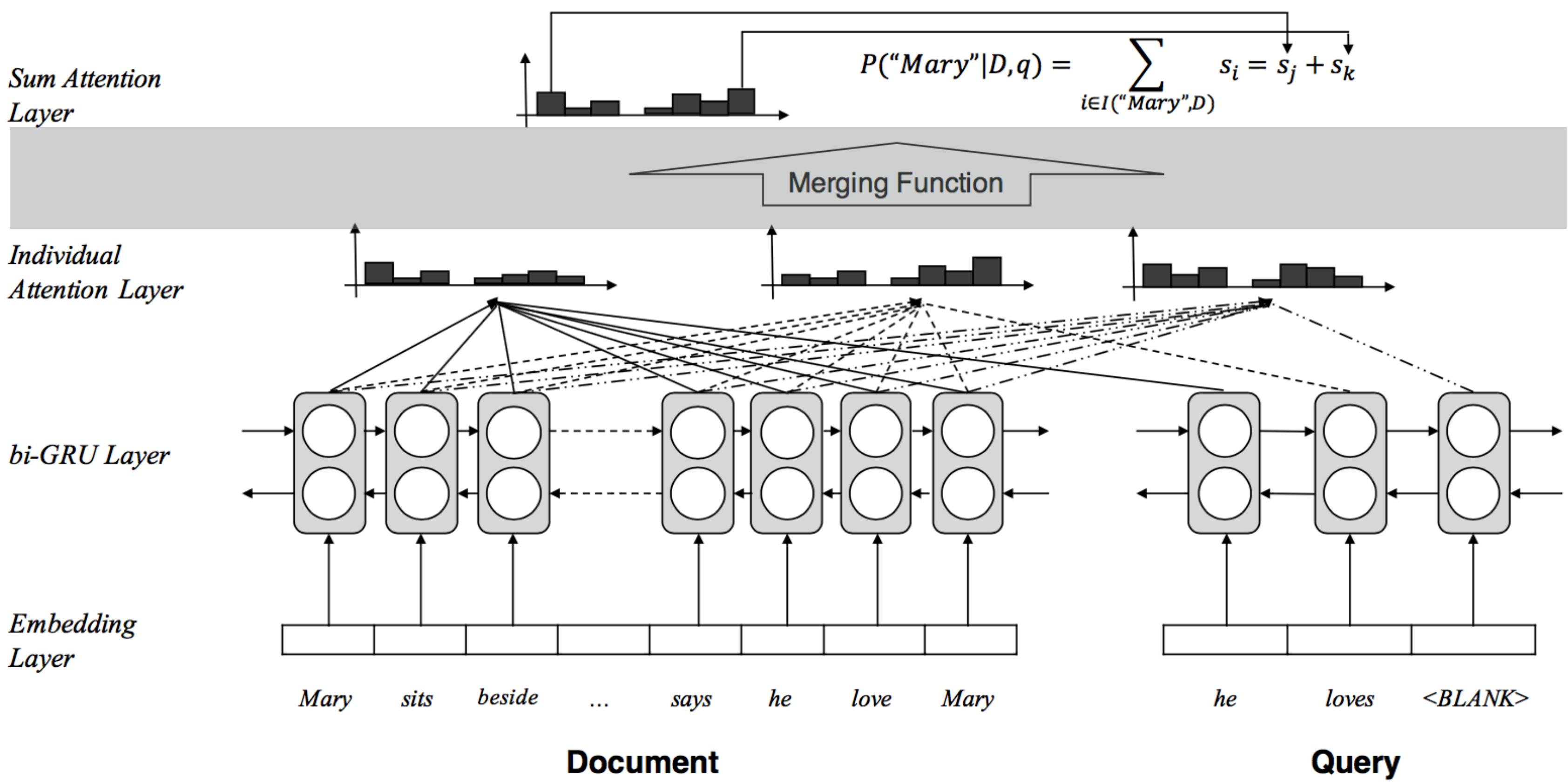


$$\alpha(t) = softmax(h_{doc}^T \cdot h_{query}(t))$$

Cui et al., COLING 2016. Consensus Attention-based Neural Networks for Chinese Reading Comprehension

CAS Reader

- **Step 3**: Induce a consensus attention over these individual attentions with heuristic functions

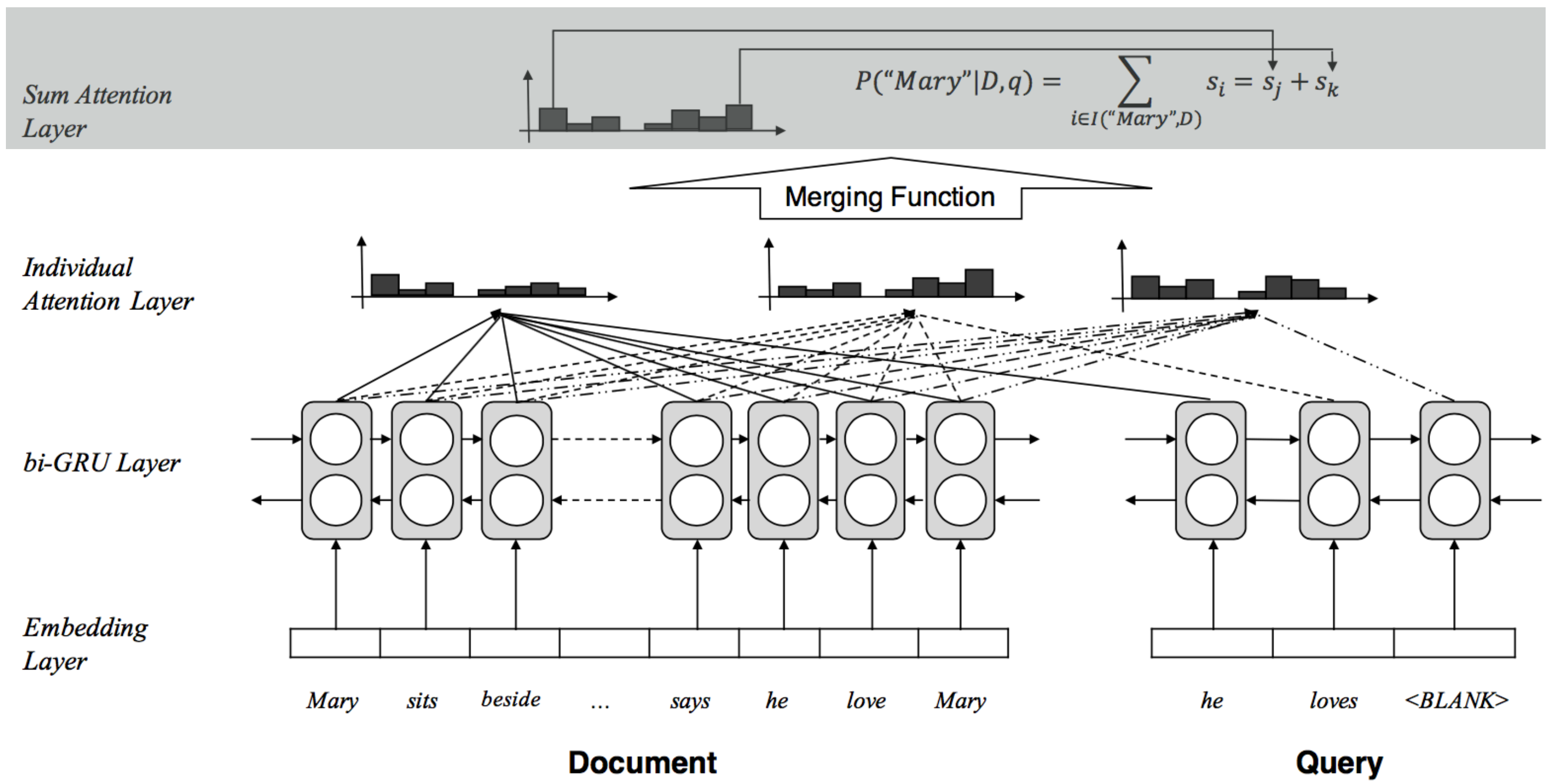


$$
s \propto \begin{cases} softmax(\sum_{t=1}^{m} \alpha(t)), & \text{if } mode = sum; \\ softmax(\frac{1}{m}\sum_{t=1}^{m} \alpha(t)), & \text{if } mode = avg; \\ softmax(\max_{t=1\ldots m} \alpha(t)), & \text{if } mode = max. \end{cases}
$$

Cui et al., COLING 2016. Consensus Attention-based Neural Networks for Chinese Reading Comprehension

CAS Reader

- **Step 4**: Applying sum-attention mechanism (Kadlec et al., 2016) to get the final probability of the answer



$$P(w|\mathcal{D},\mathcal{Q}) = \sum_{i \in I(w,\mathcal{D})} s_i, \quad w \in V$$

Cui et al., COLING 2016. Consensus Attention-based Neural Networks for Chinese Reading Comprehension

Question

- **Any better solutions for choosing the heuristic function?**
  - Though CAS Reader solves the problem of regarding the query as a whole (such as in Attentive Reader), it relies on the heuristic functions to merge final predictions.
  - These heuristic functions regard each prediction EQUALLY.
  - However, it neglects the importance of predictions from different sources.

AoA Reader

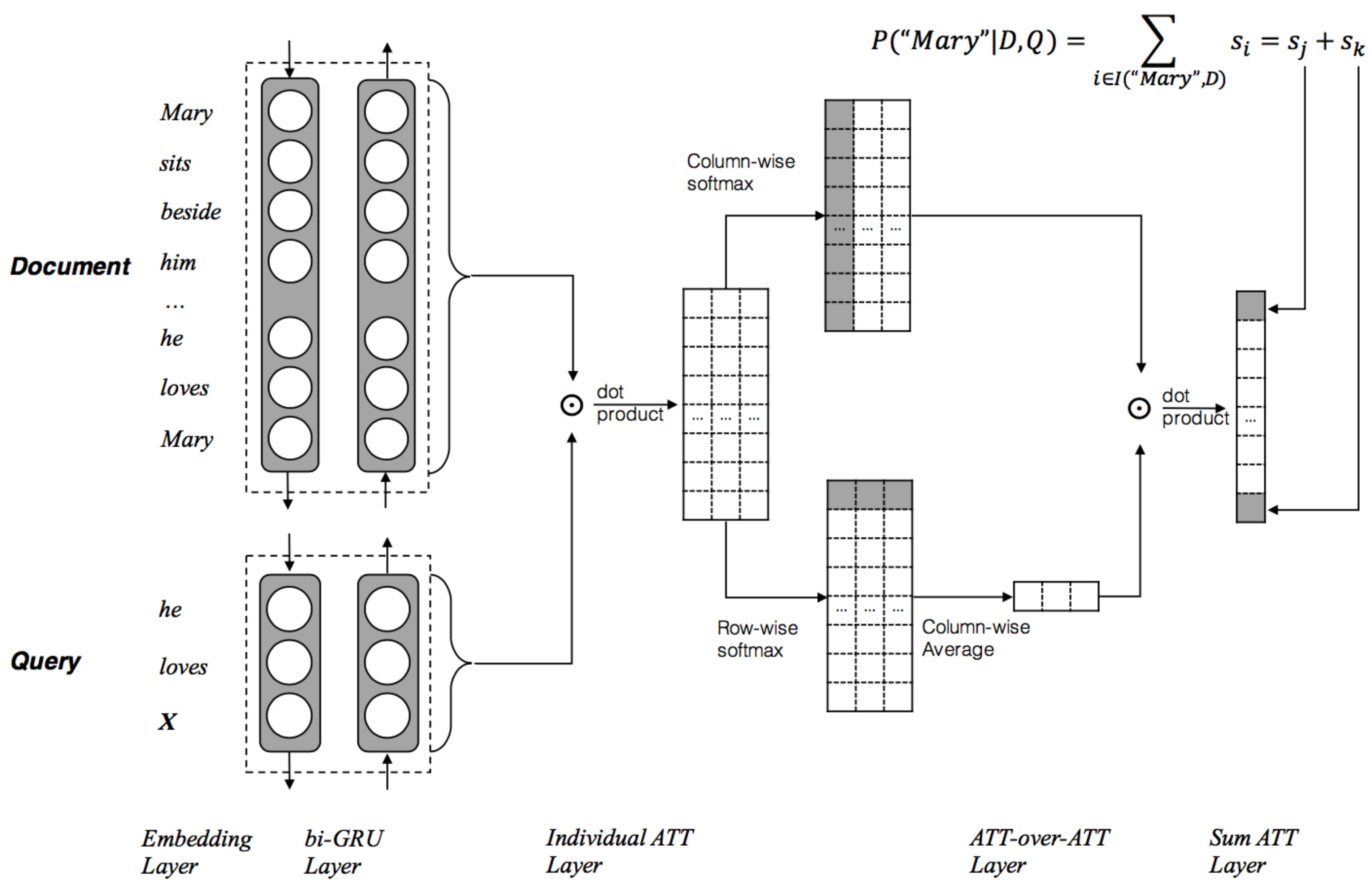
- Primarily motivated by AS Reader (Kadlec et al., 2016) and CAS Reader (Cui et al., 2016)
  - Introduce matching matrix for indicating doc-query relationships
  - Mutual attention: doc-to-query and query-to-doc
  - Instead of using heuristics to combine individual attentions, we place another attention to dynamically assign weights to the individual ones
- Some of the ideas in our work has already been adopted in the follow-up works not only in cloze-style RC but also other types of RC (such as SQuAD).

Attention-over-Attention Neural Networks for Reading Comprehension

Yiming Cui[†], **Zhipeng Chen**[†], **Si Wei**[†], **Shijin Wang**[†], **Ting Liu**[‡] and **Guoping Hu**[†]
[†]Joint Laboratory of HIT and iFLYTEK, iFLYTEK Research, Beijing, China
[‡]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China
[†]{ymcui,zpchen,siwei,sjwang3,gphu}@iflytek.com
[‡]tliu@ir.hit.edu.cn

AoA Reader

- **Model Architecture**



Cui et al., ACL 2017. Attention-over-Attention Neural Networks for Reading Comprehension

AoA Reader

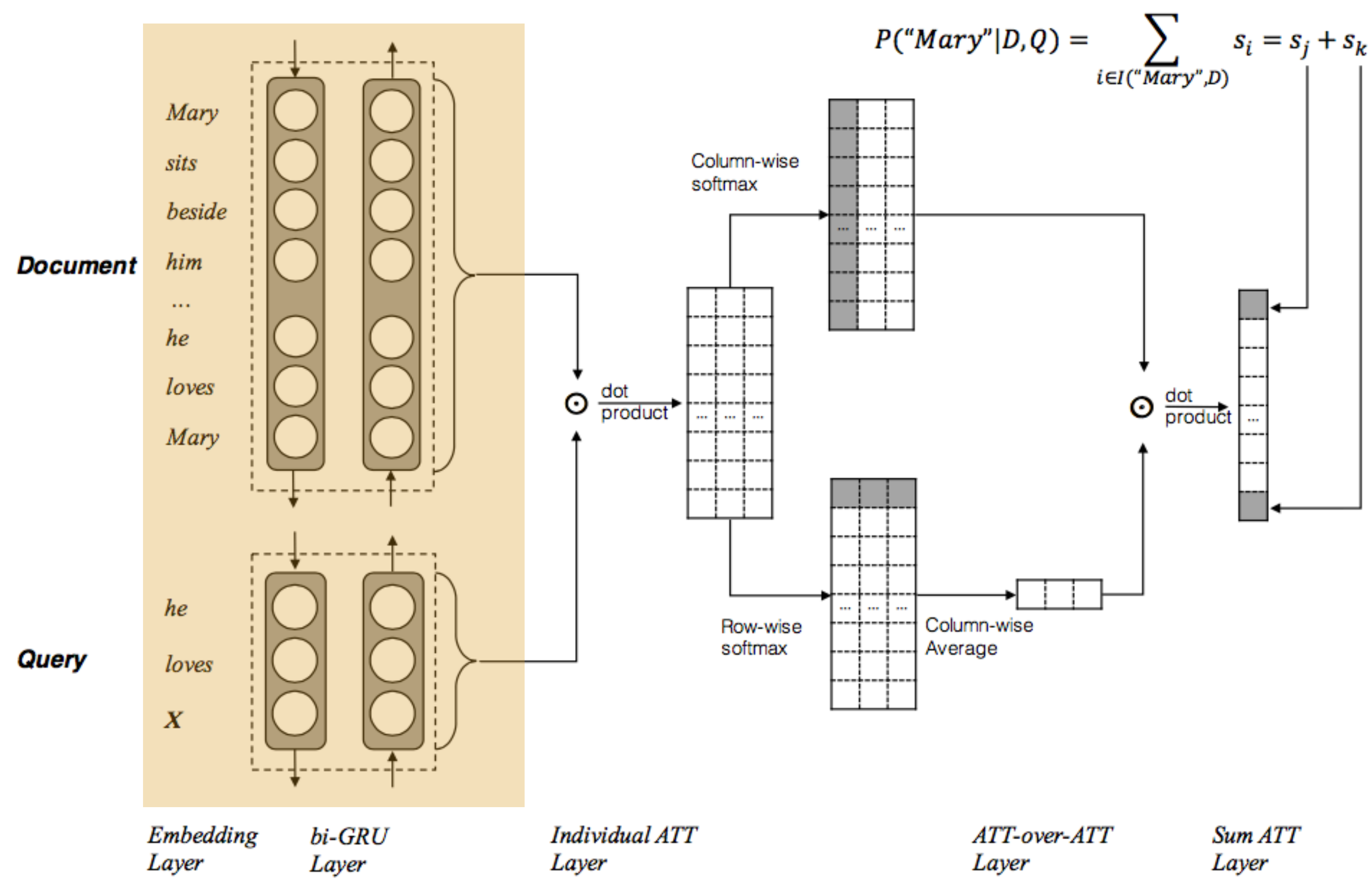
- **Contextual Embedding**
  - Transform document and query into contextual representations using word-embeddings and bi-GRU units

$$e(x) = W_e \cdot x, \ where \ x \in \mathcal{D}, \mathcal{Q} \tag{1}$$

$$\overrightarrow{h_s(x)} = \overrightarrow{GRU}(e(x)) \tag{2}$$

$$\overleftarrow{h_s(x)} = \overleftarrow{GRU}(e(x)) \tag{3}$$

$$h_s(x) = [\overrightarrow{h_s(x)}; \overleftarrow{h_s(x)}] \tag{4}$$

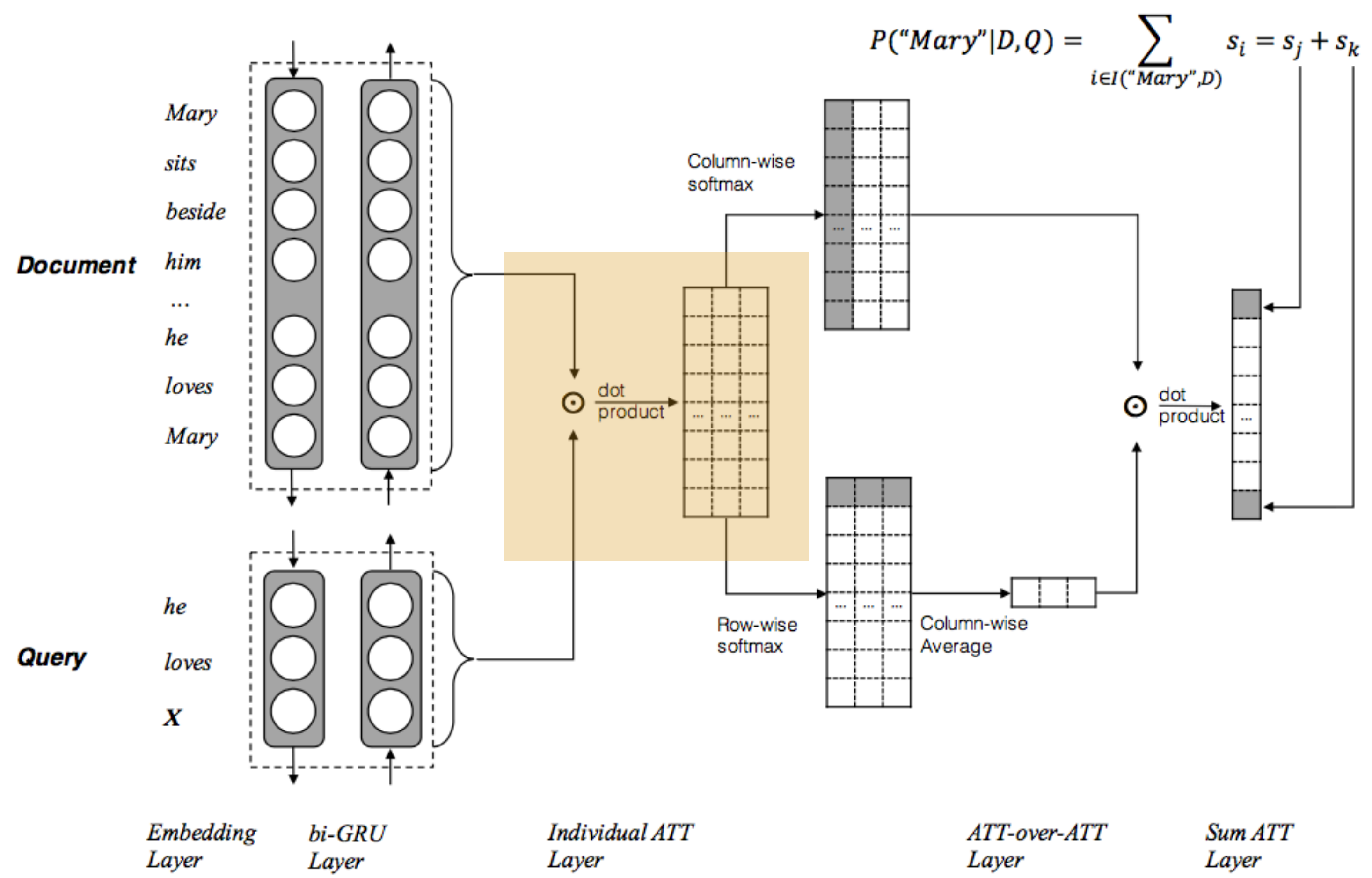


Cui et al., ACL 2017. Attention-over-Attention Neural Networks for Reading Comprehension

AoA Reader

- **Pair-wise Matching Score**
  - Calculate similarity between document and query word
  - Use dot product for attention calculation

$$M(i,j) = h_{doc}(i)^T \cdot h_{query}(j) \tag{5}$$

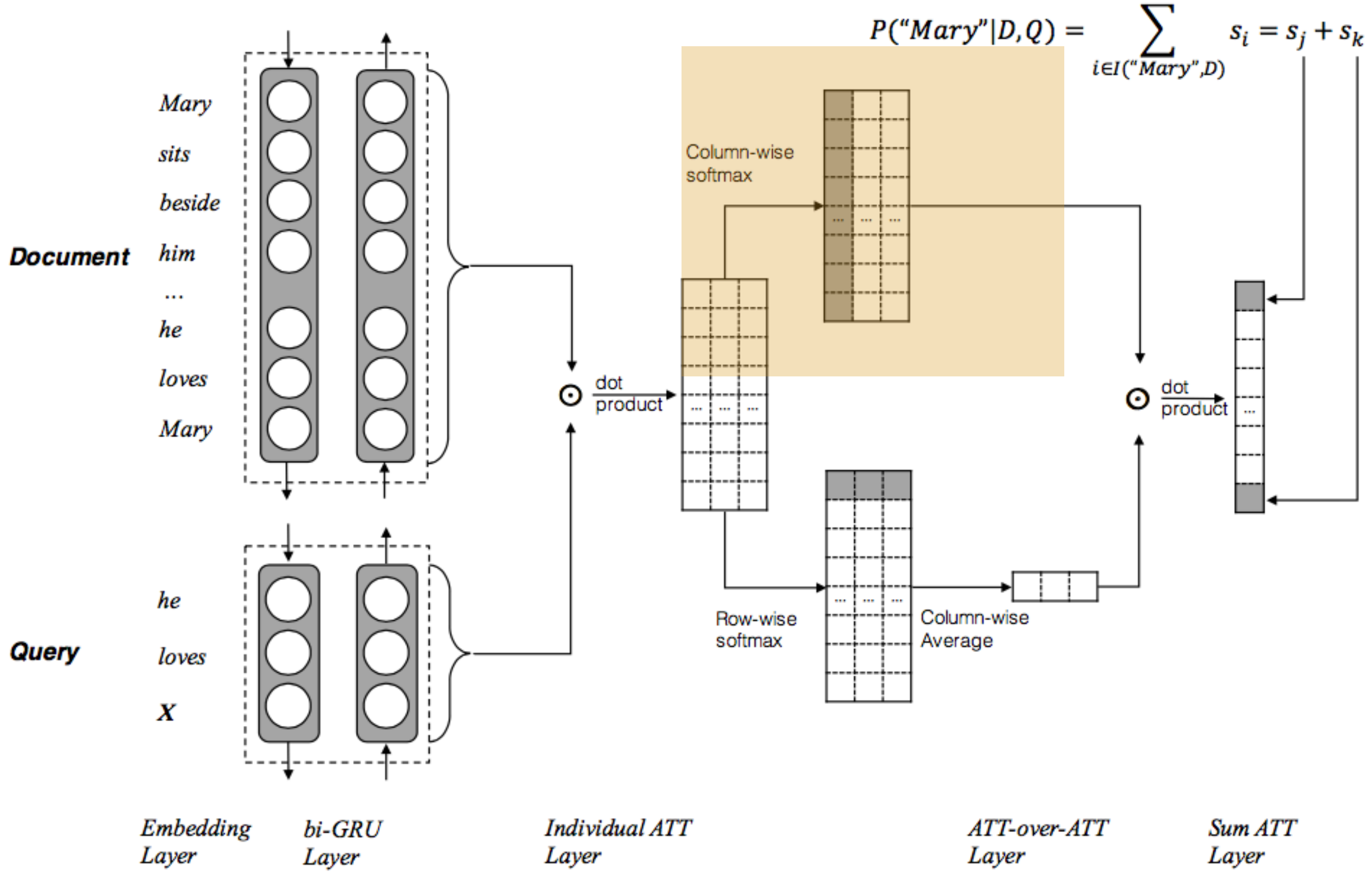


Cui et al., ACL 2017. Attention-over-Attention Neural Networks for Reading Comprehension

AoA Reader

- **Individual Attentions**
  - Calculate doc-level attention w.r.t. each query word

$$\alpha(t) = softmax(M(1,t), ..., M(|\mathcal{D}|,t)) \quad (6)$$
$$\alpha = [\alpha(1), \alpha(2), ..., \alpha(|\mathcal{Q}|)] \quad (7)$$



Cui et al., ACL 2017. Attention-over-Attention Neural Networks for Reading Comprehension

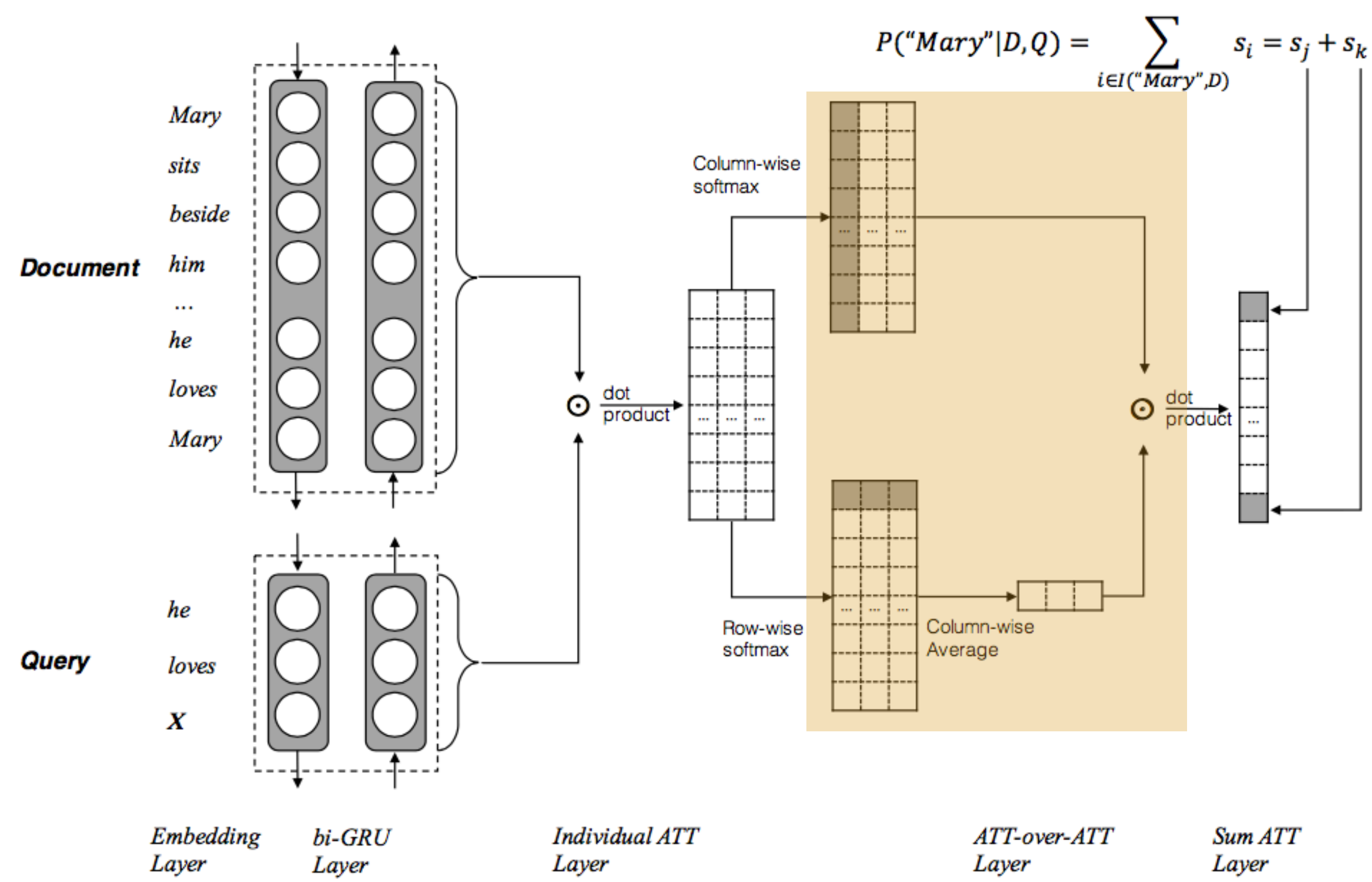
AoA Reader

- **Attention-over-Attention**
  - Dynamically assign weights to individual attentions
  - Get "attended attention"

$$\beta(t) = softmax(M(t,1), ..., M(t,|\mathcal{Q}|)) \quad (8)$$

$$\beta = \frac{1}{n}\sum_{t=1}^{|\mathcal{D}|}\beta(t) \quad (9)$$

$$s = \alpha^T\beta \quad (10)$$



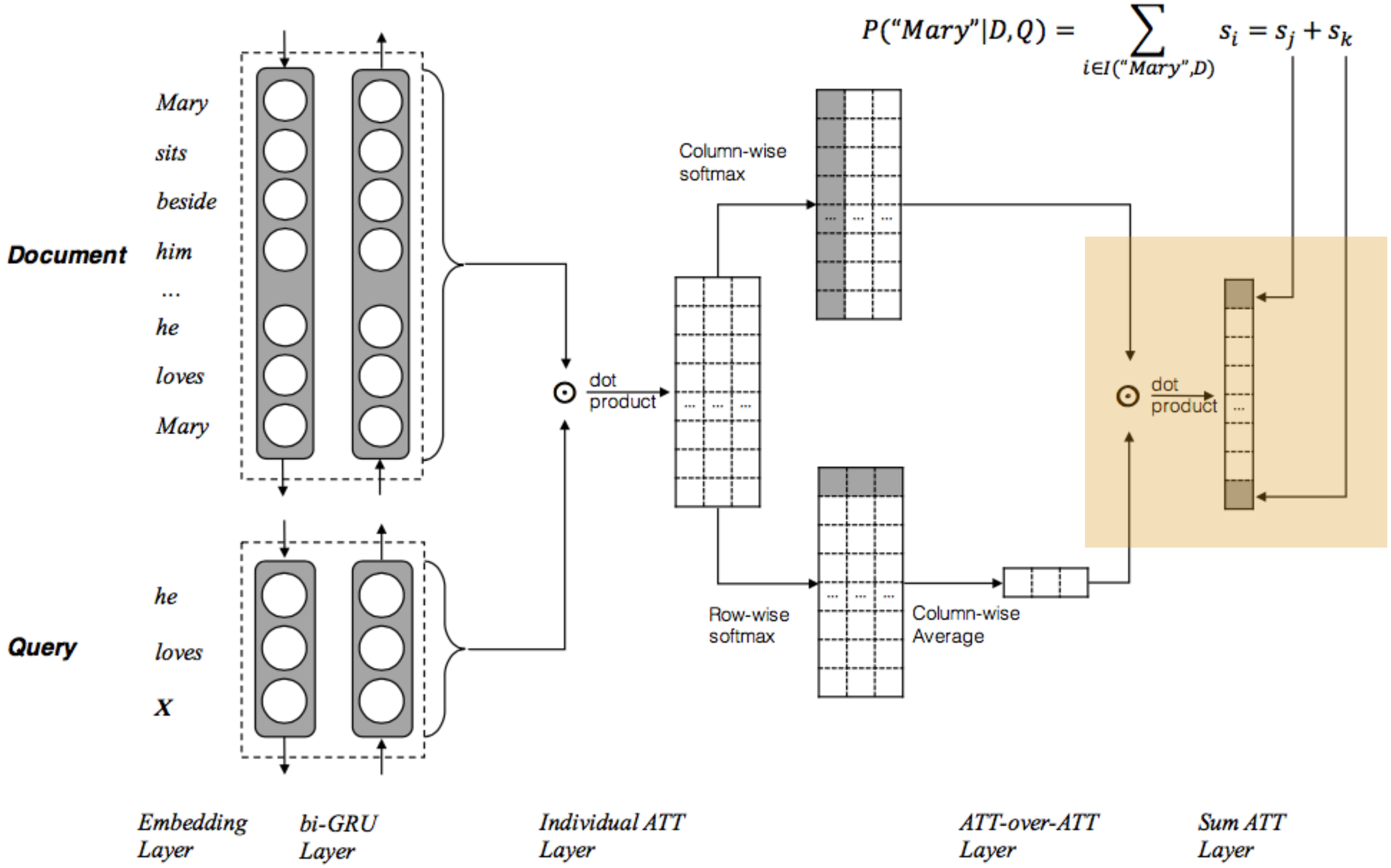
Cui et al., ACL 2017. Attention-over-Attention Neural Networks for Reading Comprehension

AoA Reader

- **Final Predictions**
  - Adopt Pointer Network (Vinyals et al., 2015) for predictions
  - Apply **sum-attention mechanism** (Kadlec et al., 2016) to get the final probability of the answer

$$P(w|\mathcal{D},\mathcal{Q}) = \sum_{i \in I(w,\mathcal{D})} s_i, \quad w \in V \qquad (11)$$

$$\mathcal{L} = \sum_i \log(p(x)) \ , x \in \mathcal{A} \qquad (12)$$



Cui et al., ACL 2017. Attention-over-Attention Neural Networks for Reading Comprehension

AoA Reader

- **An intuitive example**

	Tom	loves	<blank>	.
Query-level Attention	0.5	0.3	0.15	0.05
Candidate Answers	Mary = 0.6 diamond = 0.3 beside = 0.1	Mary = 0.3 diamond = 0.5 beside = 0.2	Mary = 0.4 diamond = 0.4 beside = 0.2	Mary = 0.2 diamond = 0.4 beside = 0.4
Average Score (CAS Reader)	Mary = (0.6+0.3+0.4+0.2) / 4 = 0.375 **diamond = (0.3+0.5+0.4+0.4) / 4 = 0.400** beside = (0.1+0.2+0.2+0.4) / 4 = 0.225			
Weighted Score (AoA Reader)	**Mary = 0.6*0.5+0.3*0.3+0.4*0.15+0.2*0.05 = 0.460** diamond = 0.3*0.5+0.5*0.3+0.4*0.15+0.4*0.05 = 0.380 beside = 0.1*0.5+0.2*0.3+0.2*0.15+0.4*0.05 = 0.160			

Cui et al., ACL 2017. Attention-over-Attention Neural Networks for Reading Comprehension

AoA Reader

- **N-best re-ranking strategy for cloze-style RC**
  - Mimic the process of double-checking, in terms of fluency, grammatical correctness, etc.
  - Main idea: Re-fill the candidate answer into the blank of the query to form a complete sentence and using additional features to score the sentences

Generate candidate answers
N-best decoding

↓

Re-fill the candidate into query

↓

Scoring with additional features (such as N-gram LM)

↓

Feature weight tuning
KBMIRA (Cherry and Foster, 2012)

↓

Re-scoring and Re-ranking

Cui et al., ACL 2017. Attention-over-Attention Neural Networks for Reading Comprehension

AoA Reader

- **Single model performance**
  - Significantly outperform previous works
  - Re-ranking strategy substantially improve performance
  - Introducing attention-over-attention mechanism instead of using heuristic merging function (Cui et al., 2016) may bring significant improvements
- **Ensemble performance**
  - We use a greedy ensemble approach of 4 models trained on different random seeds
  - Significant improvements over various state-of-the-art systems

	CNN News		CBTest NE		CBTest CN	
	Valid	Test	Valid	Test	Valid	Test
Deep LSTM Reader (Hermann et al., 2015)	55.0	57.0	-	-	-	-
Attentive Reader (Hermann et al., 2015)	61.6	63.0	-	-	-	-
Human (context+query) (Hill et al., 2015)	-	-	-	*81.6*	-	*81.6*
MemNN (window + self-sup.) (Hill et al., 2015)	63.4	66.8	70.4	66.6	64.2	63.0
AS Reader (Kadlec et al., 2016)	68.6	69.5	73.8	68.6	68.8	63.4
CAS Reader (Cui et al., 2016)	68.2	70.0	74.2	69.2	68.2	65.7
Stanford AR (Chen et al., 2016)	72.4	72.4	-	-	-	-
GA Reader (Dhingra et al., 2016)	73.0	73.8	74.9	69.0	69.0	63.9
Iterative Attention (Sordoni et al., 2016)	72.6	73.3	75.2	68.6	72.1	69.2
EpiReader (Trischler et al., 2016)	**73.4**	74.0	75.3	69.7	71.5	67.4
AoA Reader	73.1	**74.4**	**77.8**	**72.0**	**72.2**	**69.4**
AoA Reader + Reranking	-	-	**79.6**	**74.0**	**75.7**	**73.1**
MemNN (Ensemble)	66.2	69.4	-	-	-	-
AS Reader (Ensemble)	73.9	75.4	74.5	70.6	71.1	68.9
GA Reader (Ensemble)	76.4	77.4	75.5	71.9	72.1	69.4
EpiReader (Ensemble)	-	-	76.6	71.8	73.6	70.6
Iterative Attention (Ensemble)	74.5	75.7	76.9	72.0	74.1	**71.0**
AoA Reader (Ensemble)	-	-	**78.9**	**74.5**	**74.7**	70.8
AoA Reader (Ensemble + Reranking)	-	-	**80.3**	**75.6**	**77.0**	**74.1**

Cui et al., ACL 2017. Attention-over-Attention Neural Networks for Reading Comprehension

Summary

- **What are the good things in cloze-style RC?**

  - **Pointer Network** is especially useful in this task, as the answer is assumed to be existed in the document, just directly PICK the right answer from document

  - A simple **DOT product** is capable of attention calculation

  - **Mutual attention mechanism** could bring additional information, using both doc-to-query and query-to-doc attentions

  - Re-ranking strategy with **traditional N-gram LMs** could substantially improve cloze-style RC performance due to its nature

Span-Extraction MRC

SQuAD

- **SQuAD: 100,000+ Questions for Machine Comprehension of Text (Rajpurkar et al., EMNLP 2016)**
- **Dataset Features**
  - More Difficult: word-level answers → words, phrases or even sentences
  - High Quality: automatically generated data → human-annotated data
  - Much Bigger: 100K+ questions, bigger than previous human-annotated RC datasets

SQuAD
The Stanford Question Answering Dataset

SQuAD

- **Sample of SQuAD**
  - **Document**: Passages from Wikipedia pages, segment into several small paragraphs
  - **Query**: Human-annotated, including various query types (what/when/where/who/how/why, etc.)
  - **Answer**: Continuous segments (text spans) in the passage, which has a larger search space, and much harder to answer than cloze-style RC

Oxygen

The Stanford Question Answering Dataset

In the meantime, on August 1, 1774, an experiment conducted by the British clergyman Joseph Priestley focused sunlight on mercuric oxide (HgO) inside a glass tube, which liberated a gas he named "dephlogisticated air". He noted that candles burned brighter in the gas and that a mouse was more active and lived longer while breathing it. After breathing the gas himself, he wrote: "The feeling of it to my lungs was not sensibly different from that of common air, but I fancied that my breast felt peculiarly light and easy for some time afterwards." Priestley published his findings in 1775 in a paper titled "An Account of Further Discoveries in Air" which was included in the second volume of his book titled Experiments and Observations on Different Kinds of Air. Because he published his findings first, Priestley is usually given priority in the discovery.

Why is Priestley usually given credit for being first to discover oxygen?

Ground Truth Answers: published his findings first | he published his findings first | he published his findings first | he published his findings first | Because he published his findings first

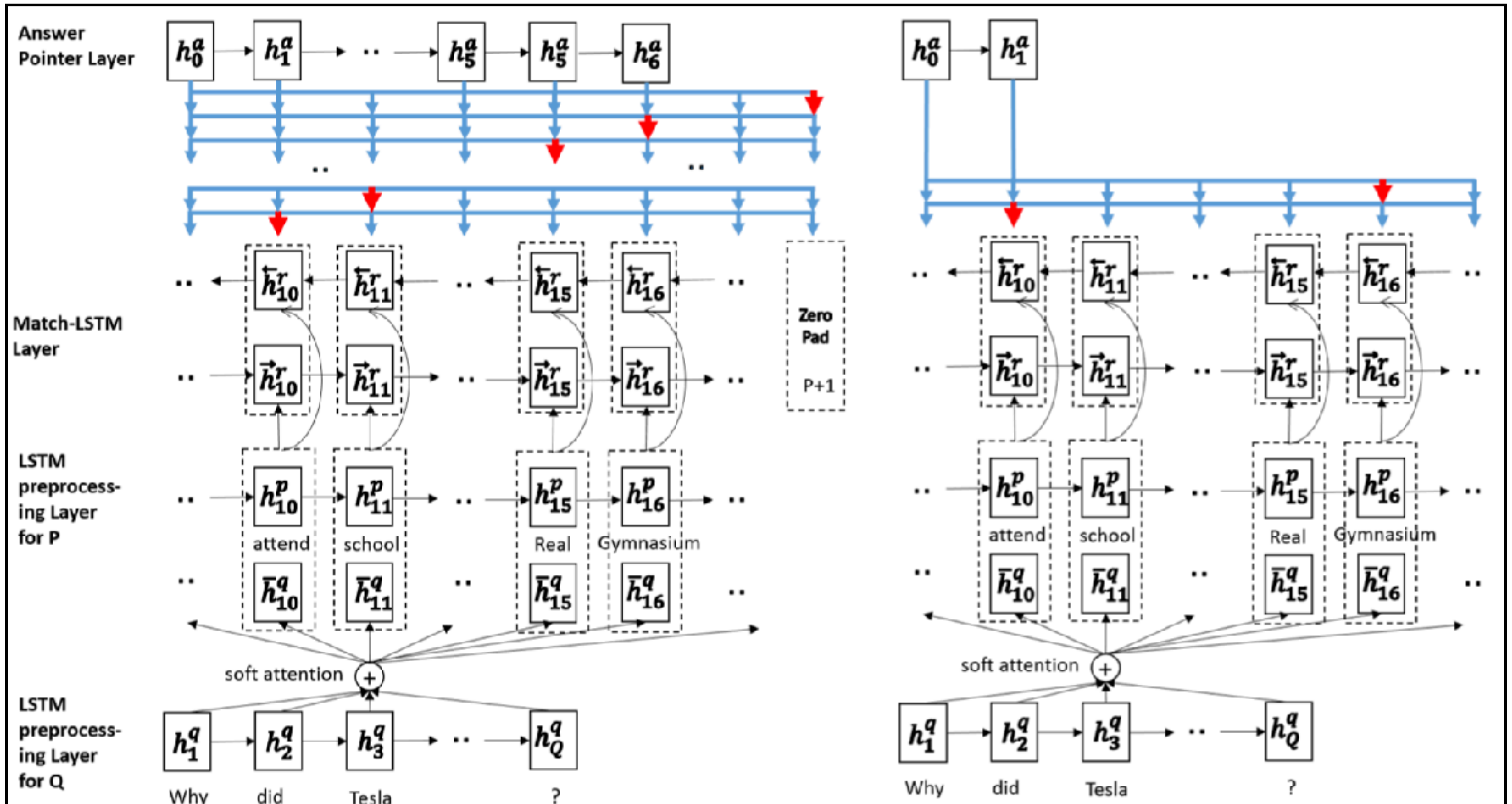
Rajpurkar et al., EMNLP 2016. 100,000+ Questions for Machine Comprehension of Text

Related Works

- A large number of researchers are investigating SQuAD after its release. Tons of models are proposed.
- Representative Works
  - Match-LSTM (Wang and Jiang, 2016)
  - Bi-directional Attention Flow (BiDAF) (Seo et al., 2016)
  - Dynamic Coattention Network (DCN) (Xiong et al., 2017)
  - r-net (Wang et al., 2017)

Match-LSTM

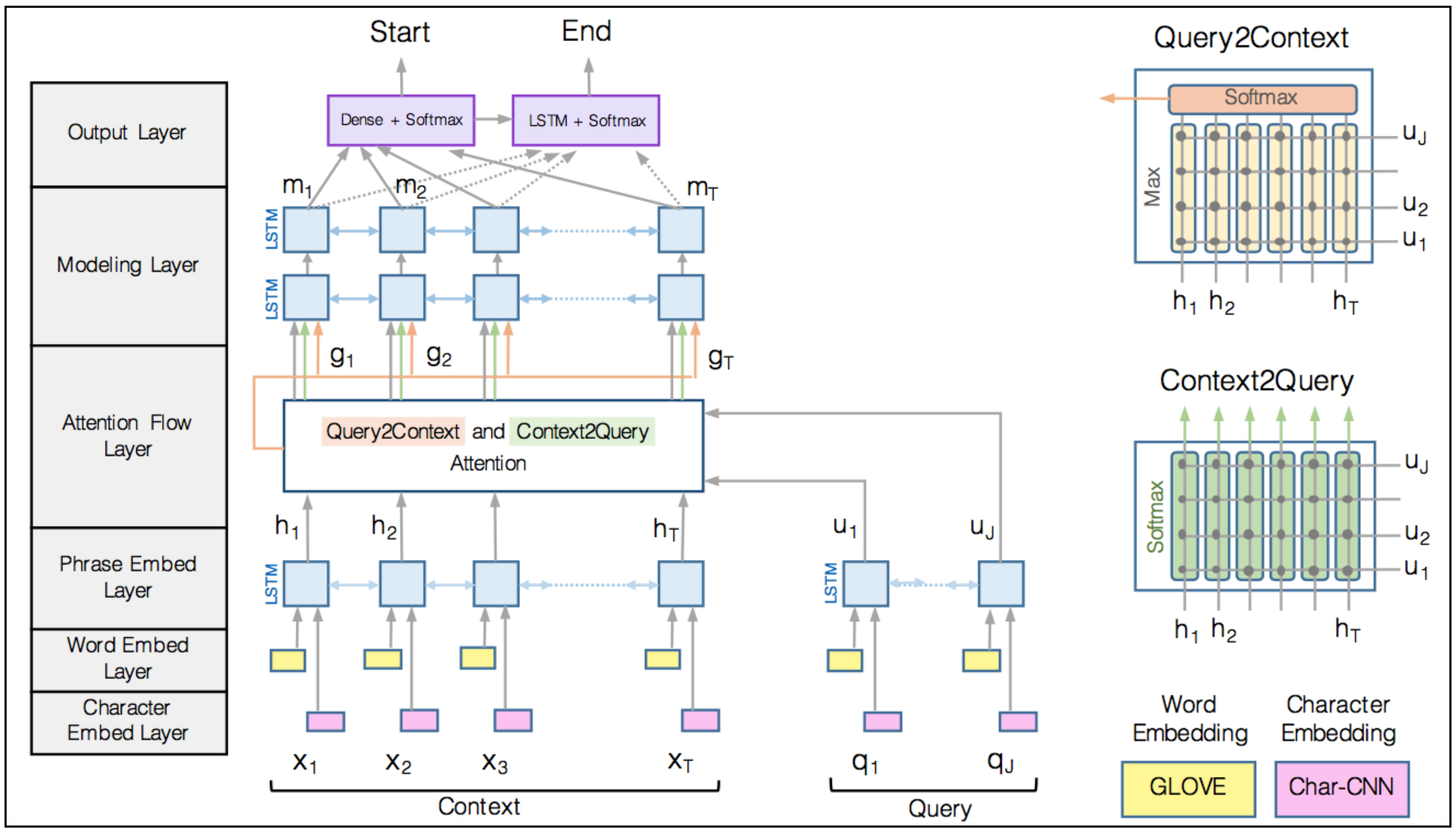
- **Machine Comprehension using Match-LSTM and Answer Pointer (Wang and Jiang, 2016)**
- Propose to use Pointer Network to directly output start and end position in the document



(a) Sequence Model (b) Boundary Model

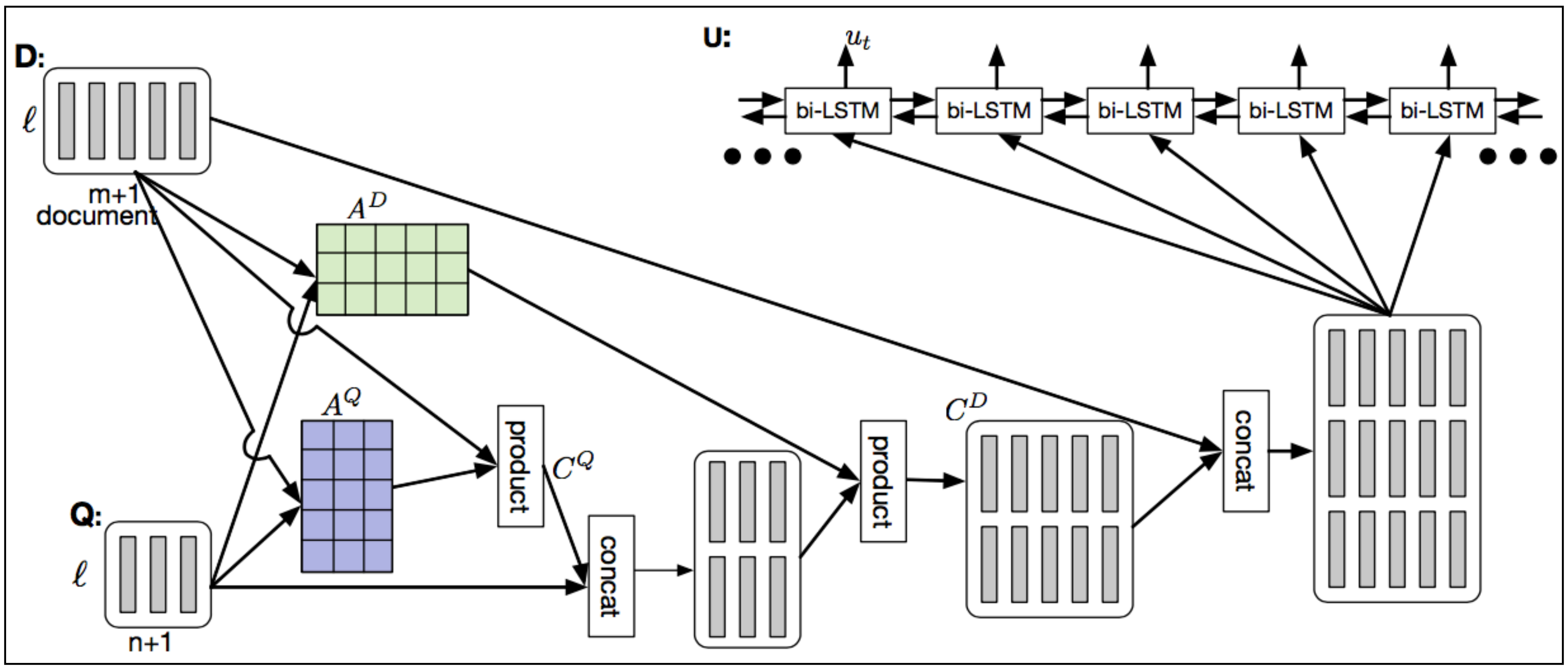
BiDAF

- **Bi-Directional Attention Flow for Machine Comprehension (Seo et al., 2016)**
- Propose bi-directional attention, which has become a **stereotype** in SQuAD task



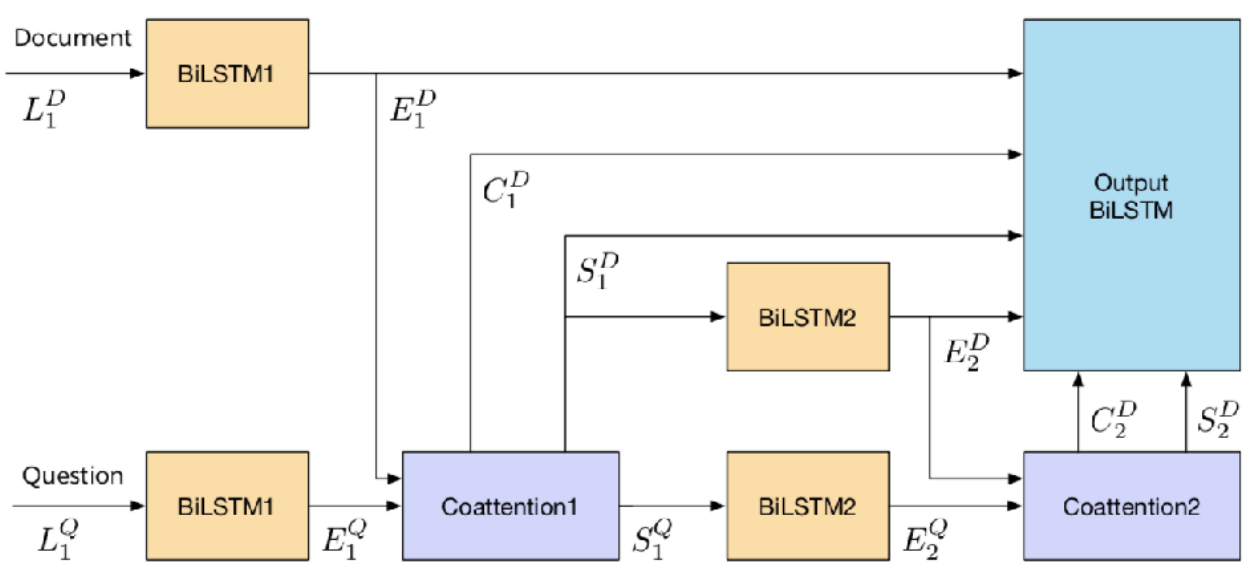
DCN

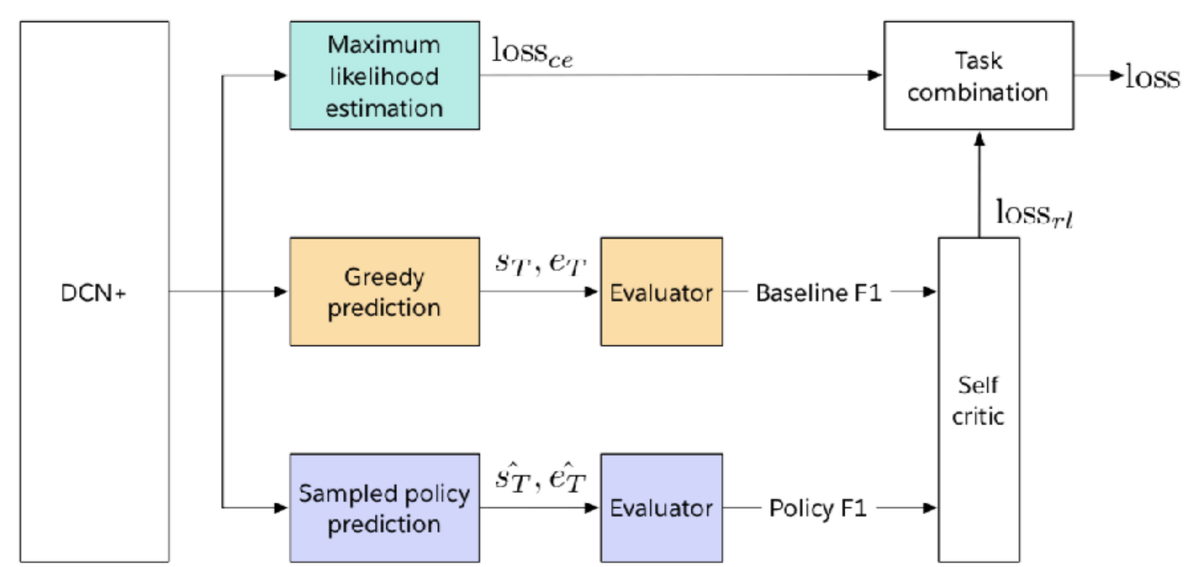
- **Dynamic Coattention Networks for Question Answering (Xiong et al., 2016)**
- Propose dynamic co-attention model, iterative pointer mechanism



DCN+

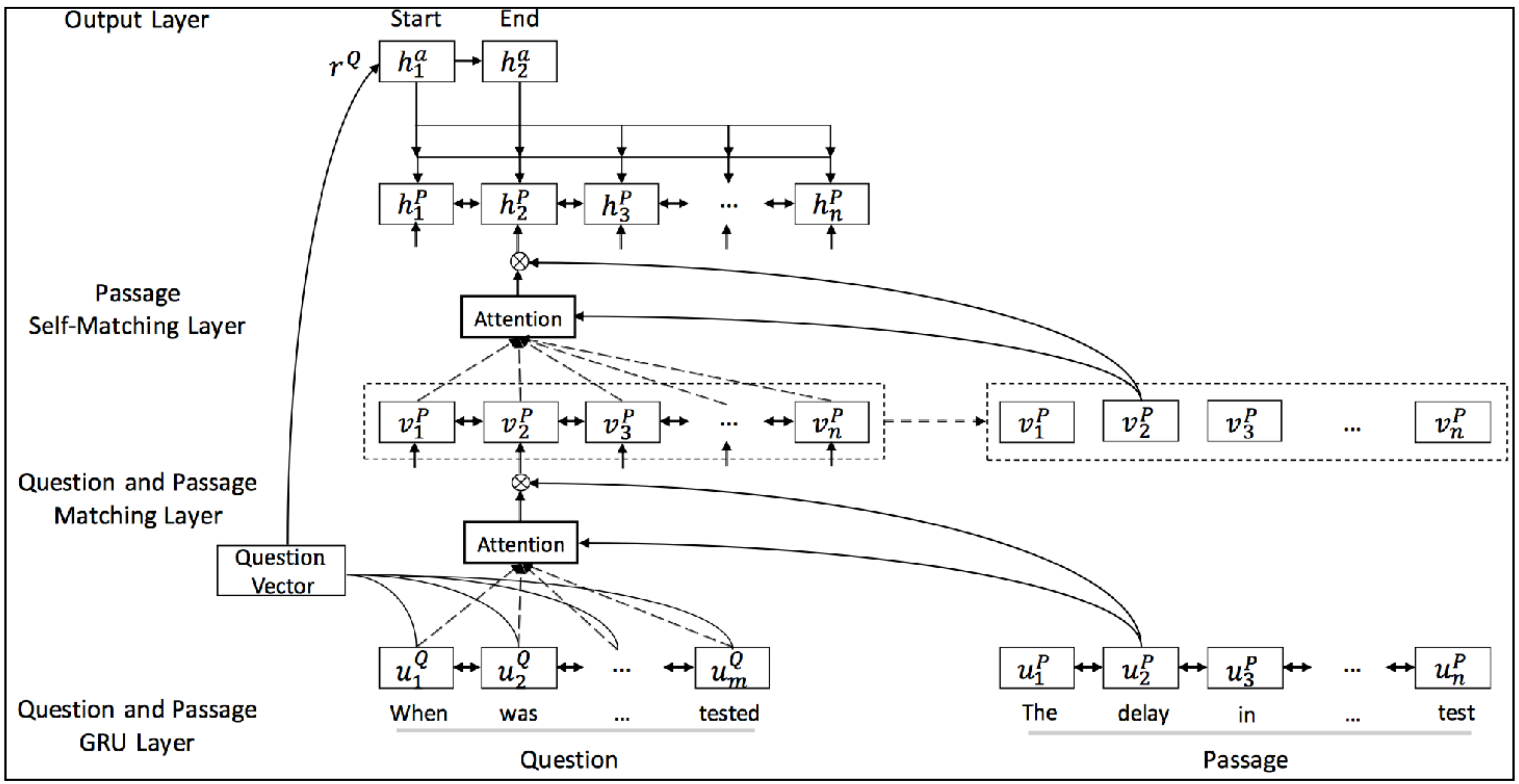
- **DCN+: Mixed Objective and Deep Residual Coattention for Question Answering (Xiong et al., 2017)**
- Utilize deep self-attention and residual networks
- A mixed objective that combines cross entropy loss with self-critical policy learning





R-NET

- **Gated Self-Matching Networks for Reading Comprehension and Question Answering (Wang et al., 2017)**
- Propose to use self-matching and gated attention



Our Work

- **Interactive AoA Reader: an improved version of AoA Reader (Cui et al., 2017)**
- We have been working on SQuAD task for months, and get on the first place in late July, 2017

Rank	Model	EM	F1
1 Jul 2017	Interactive AoA Reader (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	77.845	85.297
2 Jun 2017	r-net (ensemble) *Microsoft Research Asia* http://aka.ms/rnet	77.688	84.666
3 Jul 2017	r-net (single model) *Microsoft Research Asia* http://aka.ms/rnet	75.705	83.496
3 Jul 2017	smarnet (ensemble) *Eigen Technology & Zhejiang University*	75.989	83.475
4 Jul 2017	DCN+ (single model) *Salesforce Research*	74.866	82.806

*As of August 1, 2017. http://stanford-qa.com

Rank	Model	EM	F1
1 Oct 17, 2017	Interactive AoA Reader+ (ensemble) *Joint Laboratory of HIT and iFLYTEK*	79.083	86.450
2 Oct 24, 2017	FusionNet (ensemble) *Microsoft Business AI Solutions Team*	78.978	86.016
3 Nov 03, 2017	BiDAF + Self Attention + ELMo (single model) *Allen Institute for Artificial Intelligence*	78.580	85.833
3 Oct 12, 2017	r-net (ensemble) *Microsoft Research Asia* http://aka.ms/rnet	78.926	85.722
3 Oct 22, 2017	DCN+ (ensemble) *Salesforce Research*	78.852	85.996

*As of November 13, 2017. http://stanford-qa.com

Our Work

- As our work is not published, we cannot reveal the detailed architecture and algorithms

- But…we can tell you a little bit of the techniques that adopted (published techniques with modifications)

    - Char+Word level embeddings

    - Multiple hops for representation refining

    - Incorporating historical attentions

Summary

- **Old things still work**
  - Pointer Network for directly predict start/end position in the document
  - Mutual attention mechanism
- **What's new?**
  - Word-level + Char-level embeddings
  - More complex attention calculation with multiple attended representations

Multiple-Choice MRC

RACE

- **RACE: Large-scale ReAding Comprehension Dataset From Examinations (Lai et al., EMNLP 2017)**
- **Features**
  - Needs a more comprehensive understanding of the context
  - The answer is no longer a span in document
  - Misleading choices among candidates
  - SOTA model in SQuAD failed to give an excellent performance (70%+ → 40%)

Passage:
Is it important to have breakfast every day? A short time ago, a test was given in the United States. People of different ages, from 12 to 83, were asked to have a test. During the test, these people were given all kinds of breakfast, and sometimes they got no breakfast at all. Scientists wanted to see how well their bodies worked after eating different kinds of breakfast.
The results show that if a person eats a right breakfast, he or she will work better than if he or she has no breakfast. If a student has fruit, eggs, bread and milk before going to school, he or she will learn more quickly and listen more carefully in class. Some people think it will help you lose weight if you have no breakfast. But the result is opposite to what they think. This is because people become so hungry at noon that they eat too much for lunch. They will gain weight instead of losing it.

Question: What do the results show?
A) They show that breakfast has affected on work and studies.
B) The results show that breakfast has little to do with a person's work.
C) The results show that a person will work better if he only has fruit and milk.
D) They show that girl students should have less for breakfast.

CSA Model

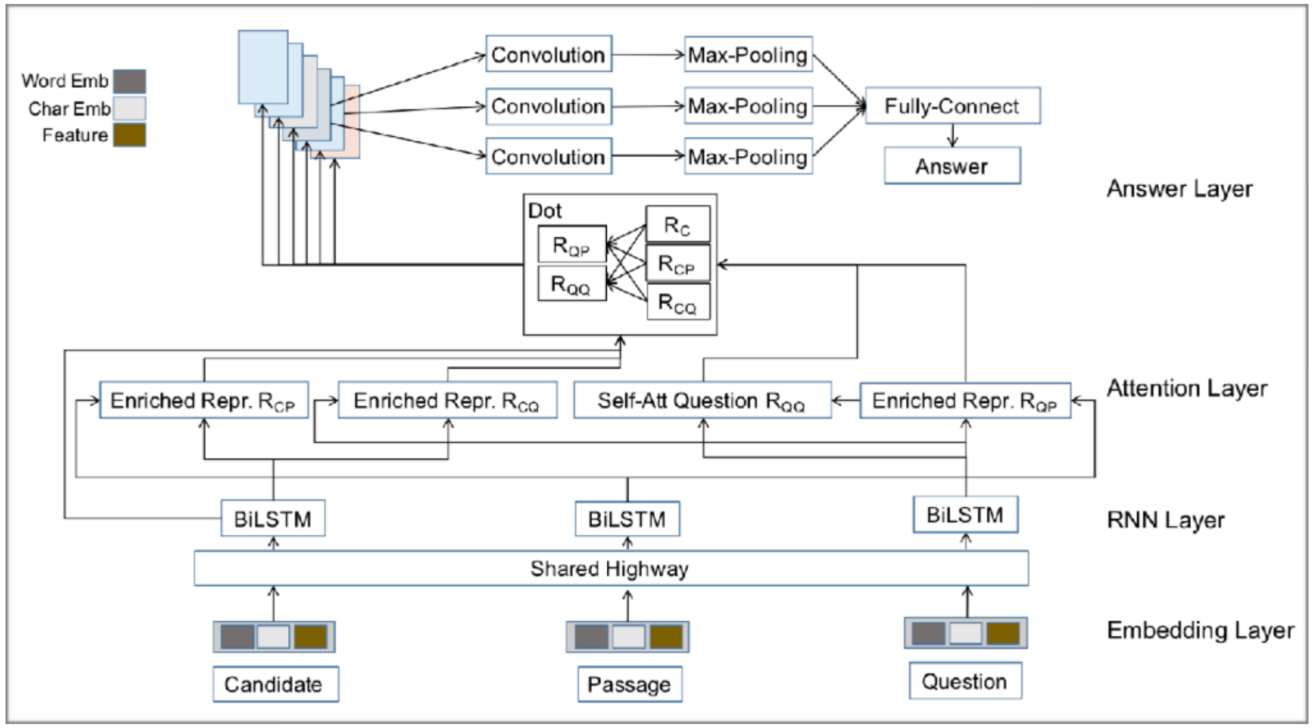
- **Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Question**
- **Contributions**
  - Focus on modeling different semantic aspects of **candidate answers**
  - Propose **C**onvolutional **S**patial **A**ttention (CSA) to simultaneously extract the attentions between various representations
  - Experimental results on RACE and SemEval 2018 Task 11 show that the proposed model achieves state-of-the-art performance.

Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

Zhipeng Chen[†], Yiming Cui[†‡*], Wentao Ma[†], Shijin Wang[†], Guoping Hu[†]
[†]Joint Laboratory of HIT and iFLYTEK (HFL), iFLYTEK Research, Beijing, China
[‡]Research Center for Social Computing and Information Retrieval (SCIR),
Harbin Institute of Technology, Harbin, China
{zpchen,ymcui,wtma,sjwang3,gphu}@iflytek.com

CSA Model

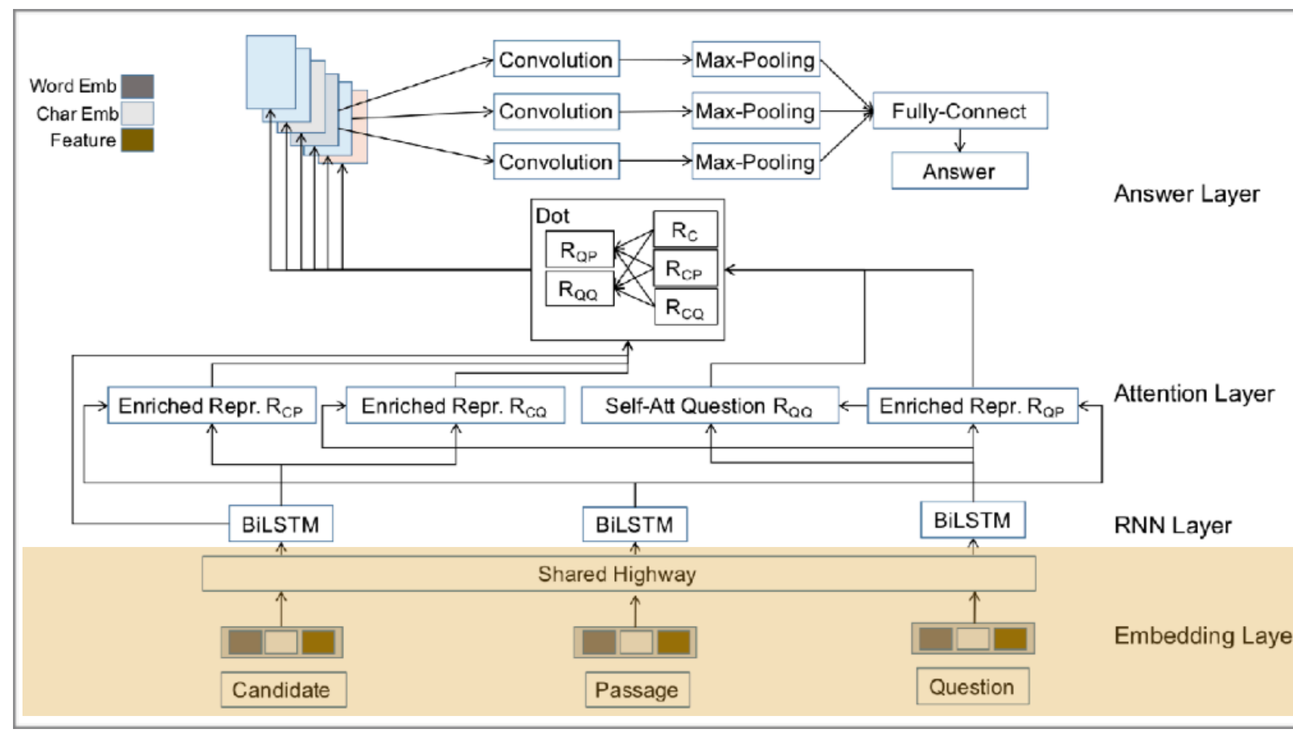
- **Formal Definition of the Task**
  - Inputs: Document, Question, Candidate
  - Output: Candidate score of being the answer
- **Basic Components**
  - Embedding Layer
  - LSTM Layer
  - Enriched Representation Layer
  - Convolutional Spatial Attention Layer
  - Answer Layer



Chen and Cui et al., AAAI 2019. Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

CSA Model

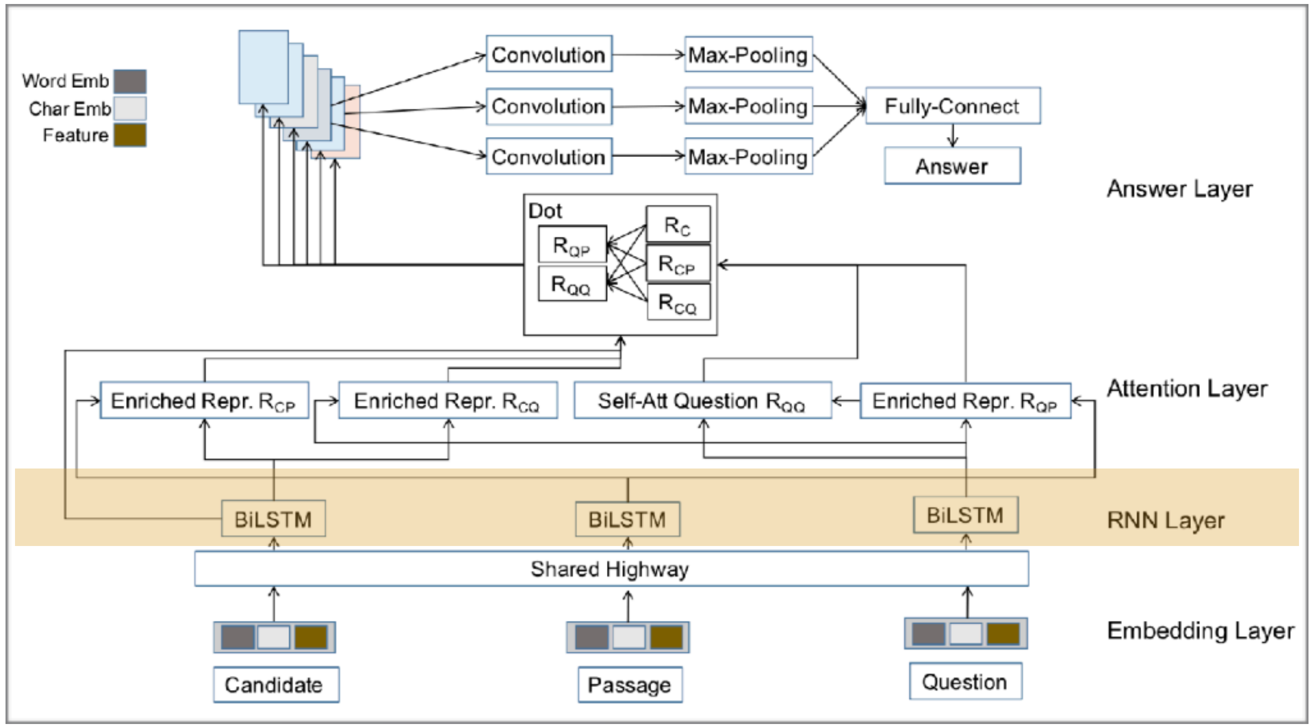
- **Embedding Layer**
  - GloVe Word Embedding [Pennington et al., 2013]
  - ELMo [Peters et al., 2018]
  - POS-tag Embedding
  - Exact Word Matching
  - Fuzzy Word Matching
- Concatenate all the features above to form final embeddings



Chen and Cui et al., AAAI 2019. Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

CSA Model

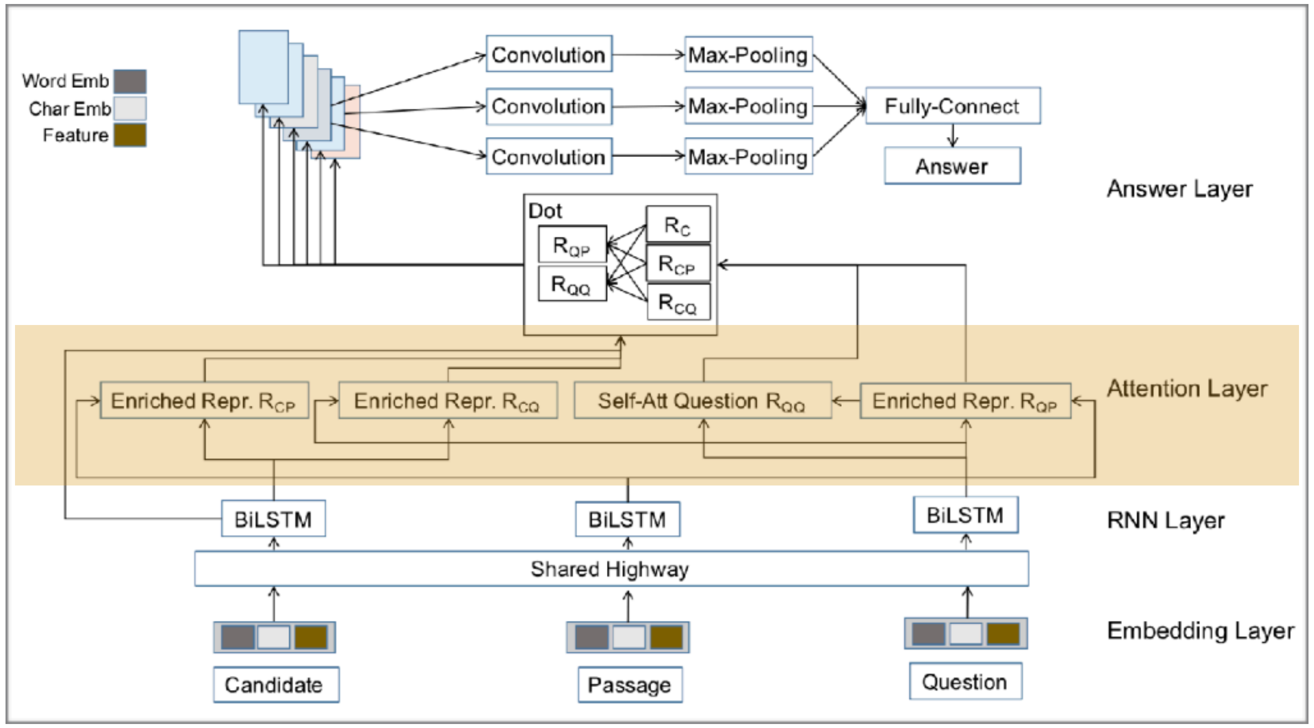
- **LSTM Layer**
  - Apply highway layer to better mix various types of embeddings
  - Place an ordinary Bi-LSTM layer after embedding to obtain contextual representation



Chen and Cui et al., AAAI 2019. Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

CSA Model

- **Enriched Representation Layer**
  - Using 'enriched representation algorithm' to get various attention-guided representations.
  - $\mathbf{R}_{CQ}$: question-aware candidate representation
  - $\mathbf{R}_{CP}$: passage-aware candidate representation
  - $\mathbf{R}_{QP}$: passage-aware question representation
  - $\mathbf{R}_{QQ}$: self-attended question representation



Chen and Cui et al., AAAI 2019. Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

CSA Model

- **Algorithm for Enriched Representation**

- **Two Key Points**
  - Adopt a symmetric attention mechanism [Huang et al., 2017]
  - Apply element-wise weight to the attention matrix

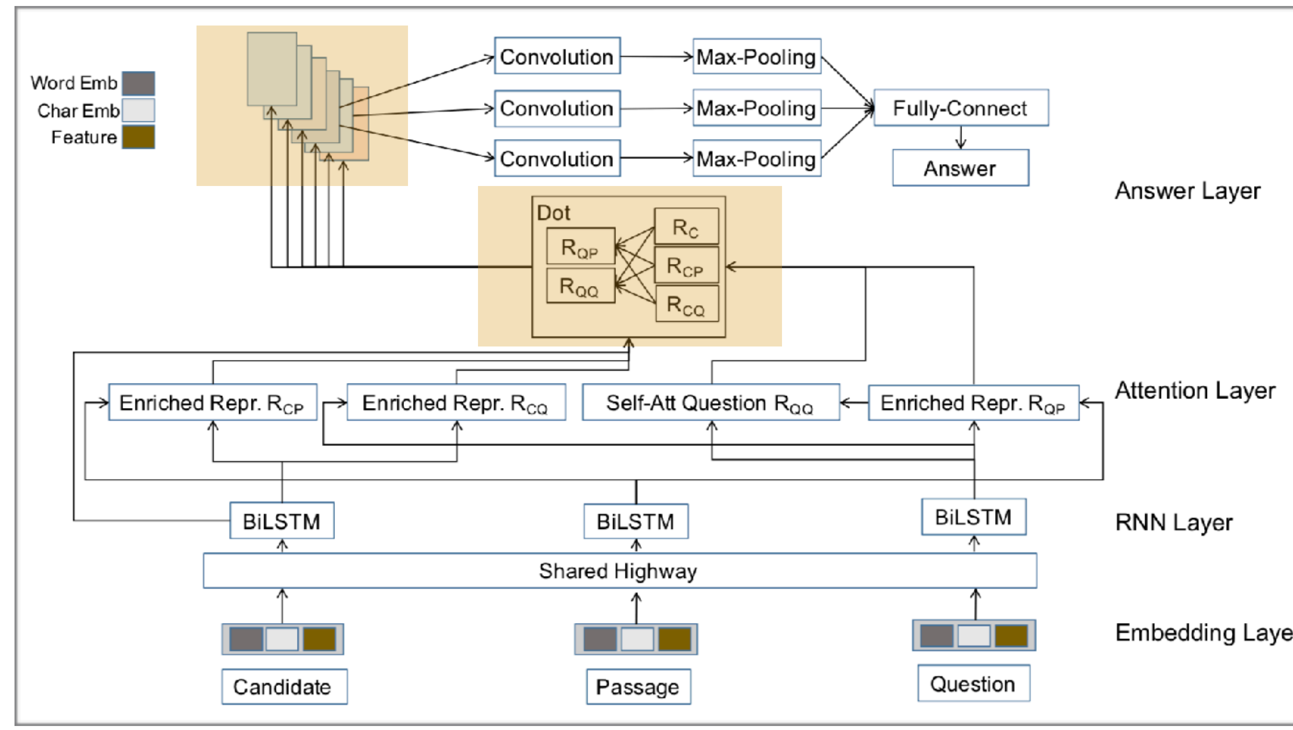
Algorithm 1 Enriched Representation.

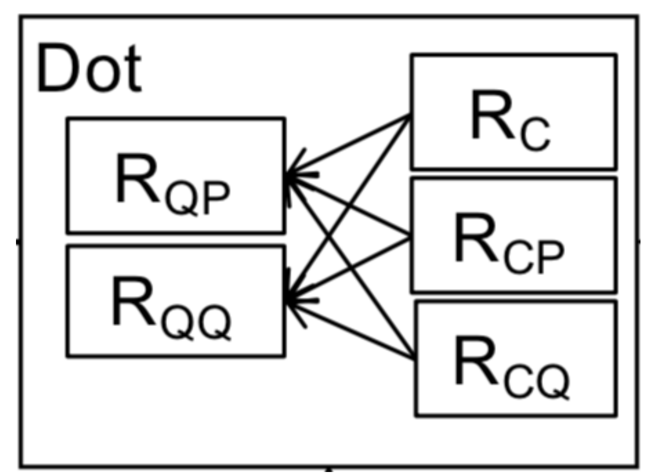
Input:
  Time-Distributed representation X_1
  Time-Distributed representation X_2
Initialize:
  Random weight matrix $W_1 \in \mathbb{R}^{h \times h_{att}}$
  Random weight matrix $W_2 \in \mathbb{R}^{h \times h_{att}}$
  Diagonal weight matrix $D \in \mathbb{R}^{h_{att} \times h_{att}}$
  All-one weight matrix $W \in \mathbb{R}^{|X_1| \times |X_2|}$
Output: X_2-aware X_1 representation Y
1: Calculate attention matrix $M' \in \mathbb{R}^{|X_1| \times |X_2|}$:
   $M' = f(W_1 X^1)^T \cdot D \cdot f(W_2 X^2)$
2: Apply element-wise weight: $M = M' \odot W$
3: Apply softmax function to the last dimension of M:
   $M_{att} = softmax(M)$
4: Calculate raw representation $Y' \in \mathbb{R}^{|X_2| \times h}$:
   $Y' = {M_{att}}^T \cdot X_1$
5: Concatenate raw representation Y' and raw input X_1, then apply Bi-LSTM:
   $Y = \text{Bi-LSTM}([X_1; Y'])$
6: **return** Y

Chen and Cui et al., AAAI 2019. Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

CSA Model

- **Convolutional Spatial Attention Layer**
  - Candidate information is important
  - We calculate dot attentions between three candidate representations and two question representations
  - Concatenate 2*3=6 attention matrices, forming an attention cuboid **M** with shape [6, candidate_len, question_len]





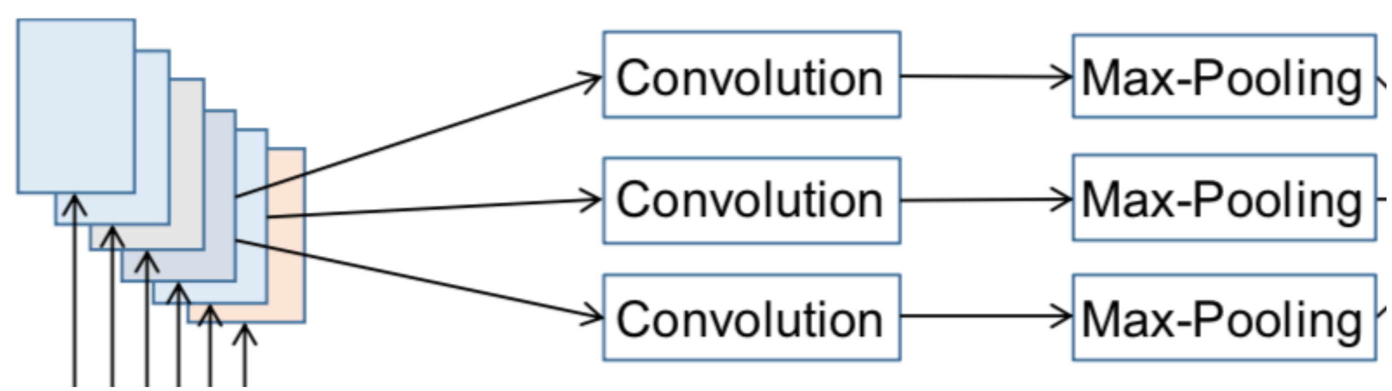
Chen and Cui et al., AAAI 2019. Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

科大讯飞 iFLYTEK

CSA Model

- **Convolutional Spatial Attention Layer**
  - The resulting matching cuboid **M** can be seen as a 2D-image with 6-channels
  - We use Convolution-MaxPooling operation to dynamically extract high-level features with kernel size 5, 10, 15

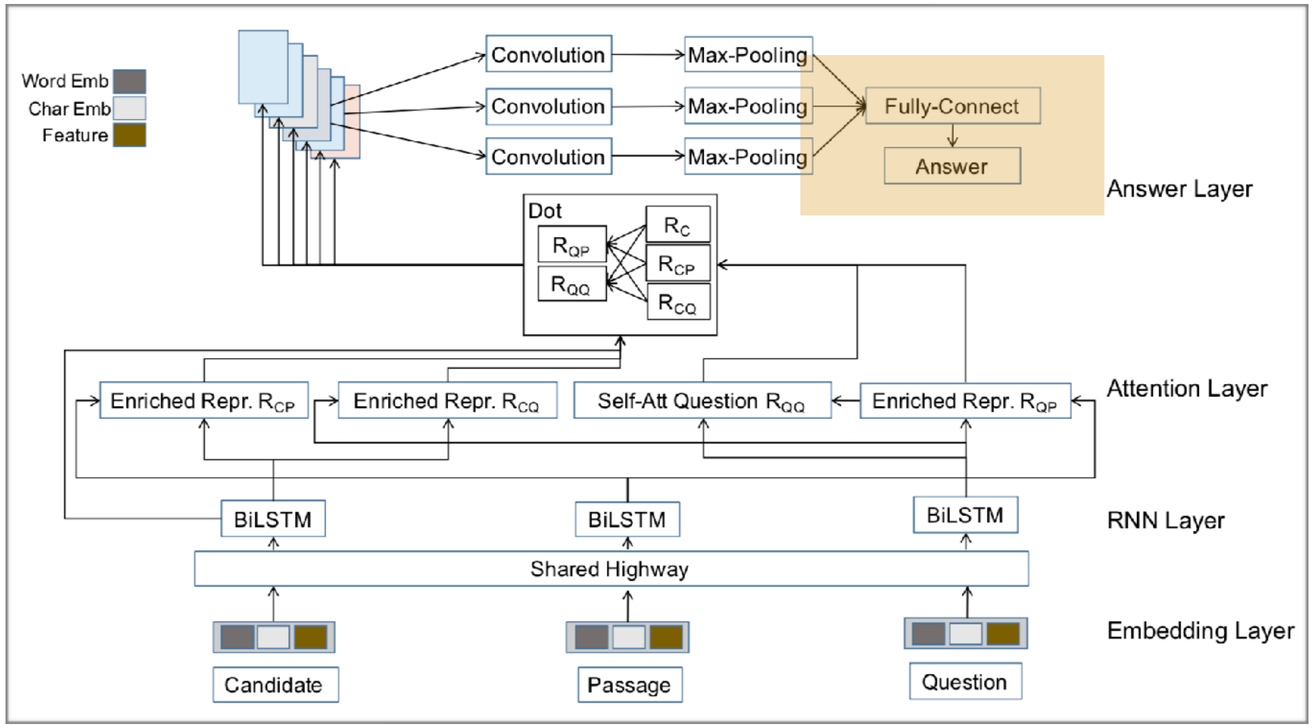
$$O_1 = \text{Max-Pooling}_{1\times 3}\{CNN_{1\times 5}(M)\}$$
$$O_2 = \text{Max-Pooling}_{1\times 2}\{CNN_{1\times 10}(M)\}$$
$$O_3 = \text{Max-Pooling}_{1\times 1}\{CNN_{1\times 15}(M)\}$$



Chen and Cui et al., AAAI 2019. Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

CSA Model

- **Answer Layer**
  - Concatenate all three feature vectors
  - Pass through a fully-connected layer to get a scalar score
- **Prediction**
  - Choose the candidate that has the largest score as the answer



Chen and Cui et al., AAAI 2019. Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

Experiments

- **Dataset**
  - RACE: English examinations of Chinese middle and high school students. (4 candidate selections)
  - SemEval 2018 Task 11: Machine Comprehension using Commonsense Knowledge (2 candidate selections)
- **Hyper-parameters**
  - Passage/Question/Candidate max length: 300 / 20 / 10
  - Word Embedding: 200-dim
  - Bi-LSTM hidden size: 250-dim
  - ELMo: 1024-dim
- **Implementation**
  - Keras + TensorFlow

Experiments

- **Results on RACE**
  - State-of-the-art performance, especially on RACE-H
  - Incorporating ELMo yields another significant improvement

Model	**RACE-M**	**RACE-H**	**RACE**
Sliding Window (Lai et al. 2017)	37.3	30.4	32.2
Stanford AR (Lai et al. 2017)	44.2	43.0	43.3
GA Reader (Lai et al. 2017)	43.7	44.2	44.1
ElimiNet (Parikh et al. 2018)	N/A	N/A	44.5
Hierarchical Attention Flow (Zhu et al. 2018)	45.0	*46.4*	46.0
Dynamic Fusion Network (Xu et al. 2017)	*51.5*	45.7	*47.4*
CSA Model (single model)	51.0	47.3	48.4
CSA Model + ELMo (single model)	**52.2**	**50.3**	**50.9**
GA Reader (6-ensemble)	-	-	45.9
ElimiNet (6-ensemble)	-	-	46.5
GA + ElimiNet (12-ensemble)	-	-	47.2
Dynamic Fusion Network (9-ensemble)	*55.6*	*49.4*	*51.2*
CSA Model (7-ensemble)	55.2	52.4	53.2
CSA Model + ELMo (9-ensemble)	**56.8**	**54.8**	**55.0**

Chen and Cui et al., AAAI 2019. Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

Experiments

- **Results on SemEval 2018**
  - Baselines are the top teams in SemEval 2018 Task 11.
  - CSA model shows marginal but consistent improvements on single/ensemble settings.
  - With the help of ELMo, there is another boost in performance.

Model	Dev	Test
HMA (Chen et al. 2018)	**84.48**	80.94
TriAN (Wang 2018)	83.84	81.94
CSA Model (single model)	83.63	82.20
CSA Model + ELMo (single model)	83.84	**83.27**
TriAN (ensemble)	85.27	83.95
HMA (ensemble)	**86.46**	84.13
CSA Model (ensemble)	84.05	84.34
CSA Model + ELMo (ensemble)	85.05	**85.23**

Chen and Cui et al., AAAI 2019. Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

Experiments

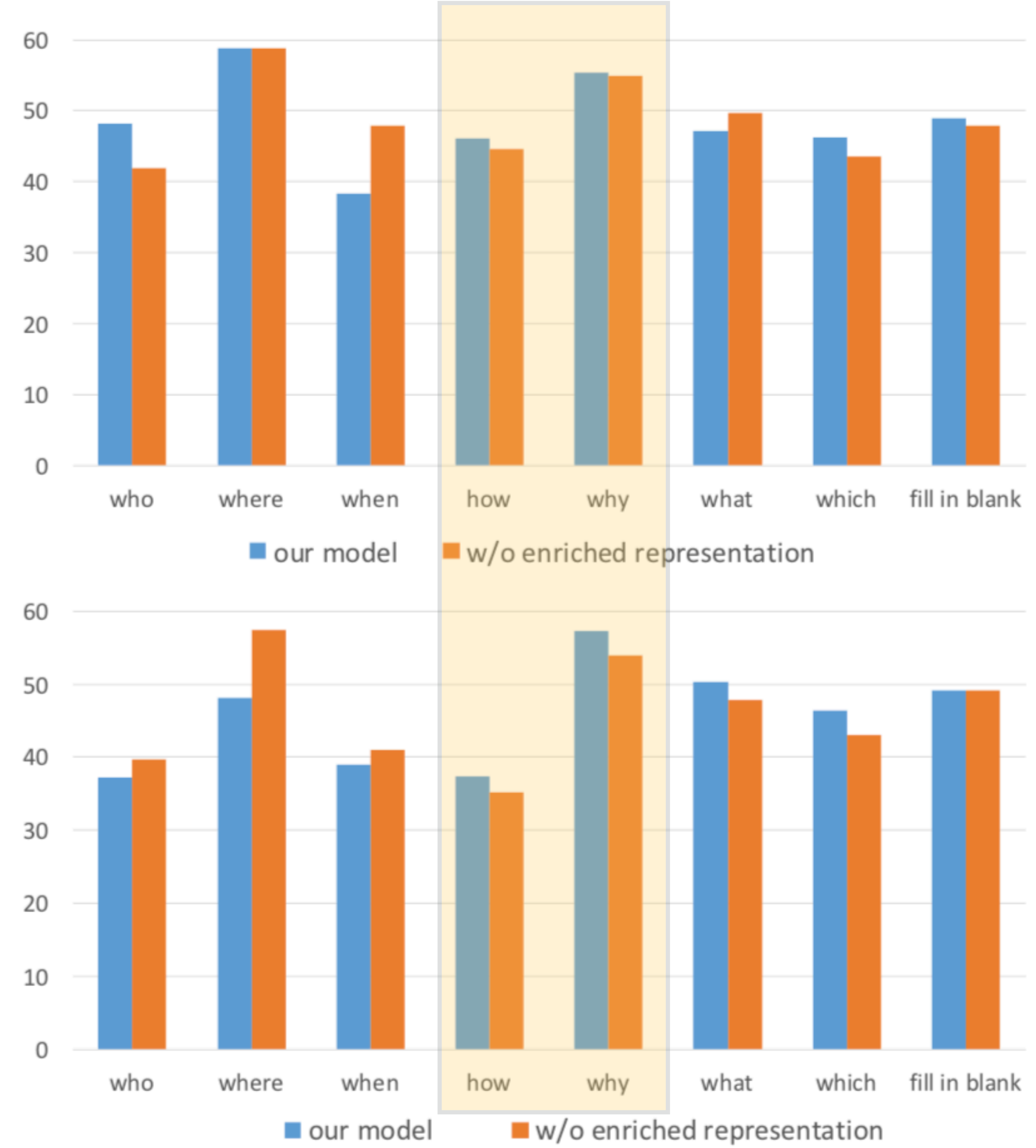
- **Ablation Results on RACE**
  - w/o attention weight: do not apply element-wise weight
  - w/o enriched repr: only use LSTM outputs
  - w/o CSA: using two fully connected layers to achieve dimensionality reduction of the 3D-attention
- **Importance: CSA > enriched repr > att weight**

Model	RACE
CSA Model	48.52
w/o attention weight	48.18
w/o enriched representation	47.52
w/o convolutional spatial attention	47.30
CSA Model + ELMo	50.89
w/o attention weight	49.49
w/o enriched representation	49.78
w/o convolutional spatial attention	48.47

Chen and Cui et al., AAAI 2019. Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

Analysis

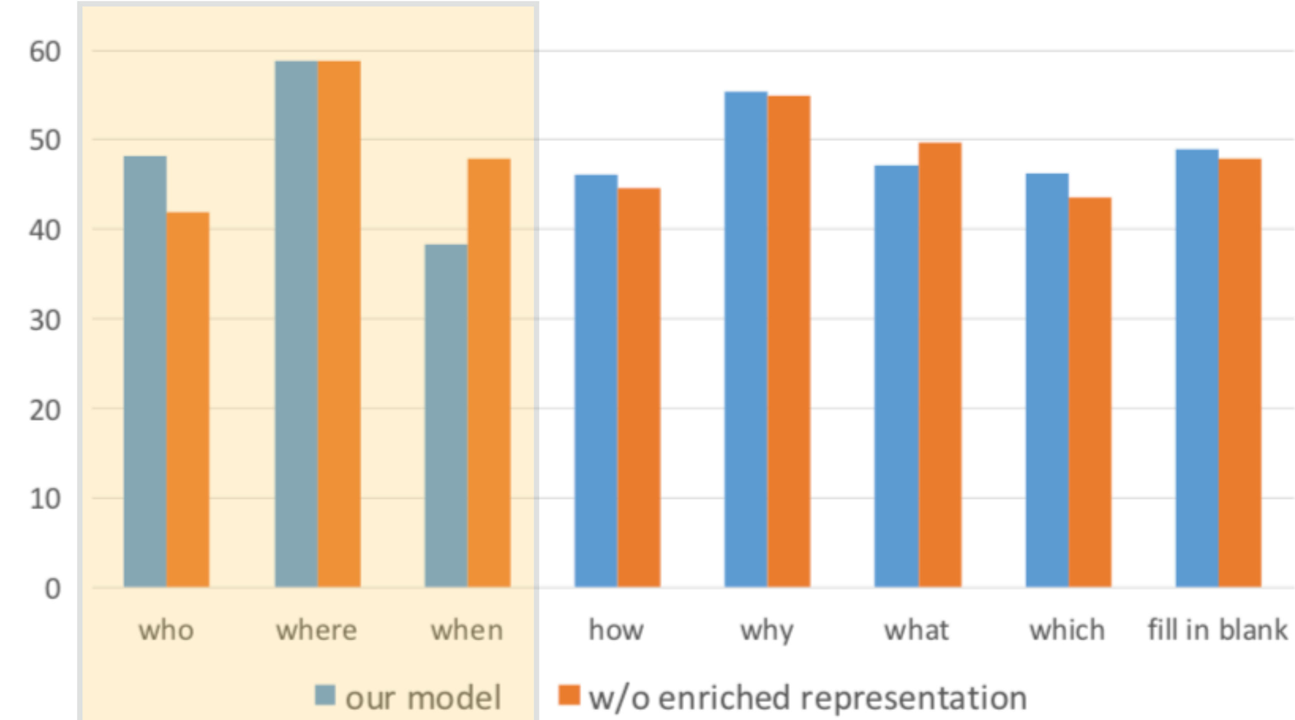
- **Quantitative Analysis on Different Type of Questions (on RACE data)**
  - [+] CSA model is good at handling 'how' and 'why' questions, which needs comprehensive reasoning on the document

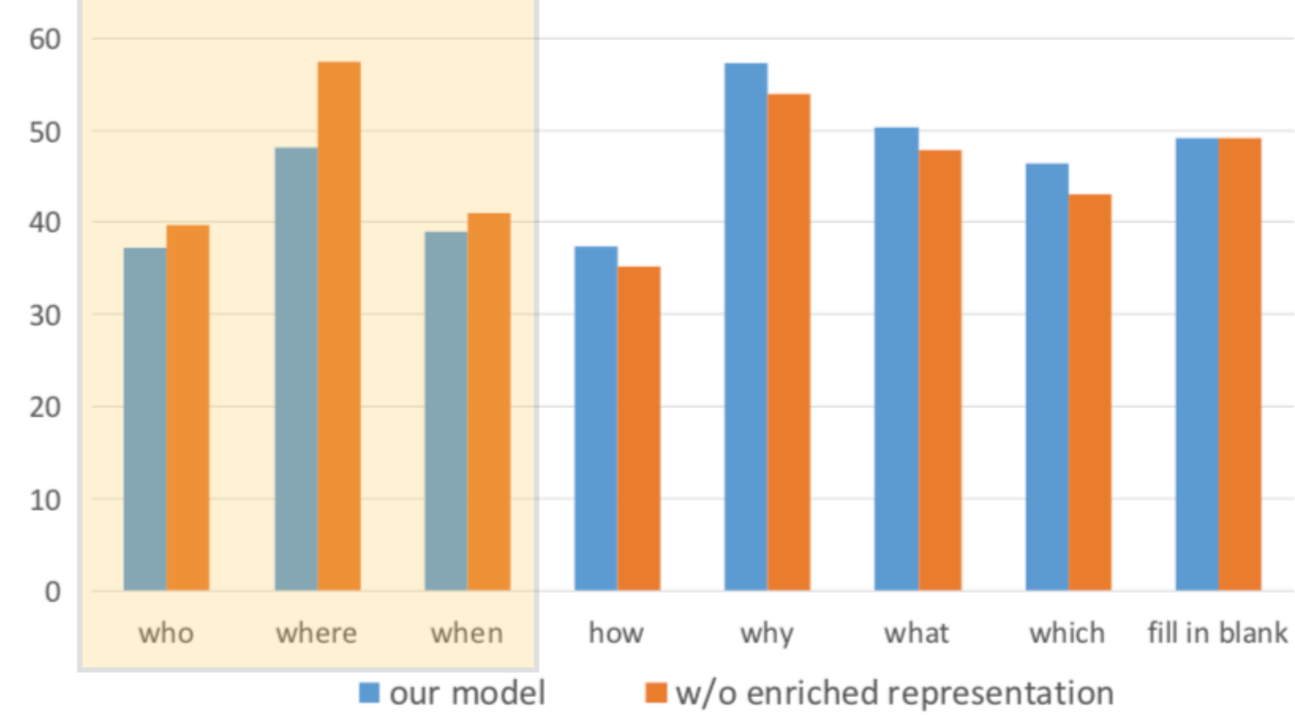


Chen and Cui et al., AAAI 2019. Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

Analysis

- **Quantitative Analysis on Different Type of Questions (on RACE data)**
  – [-] On the contrary, CSA model shows inferior performance on 'who', 'where', 'when' questions

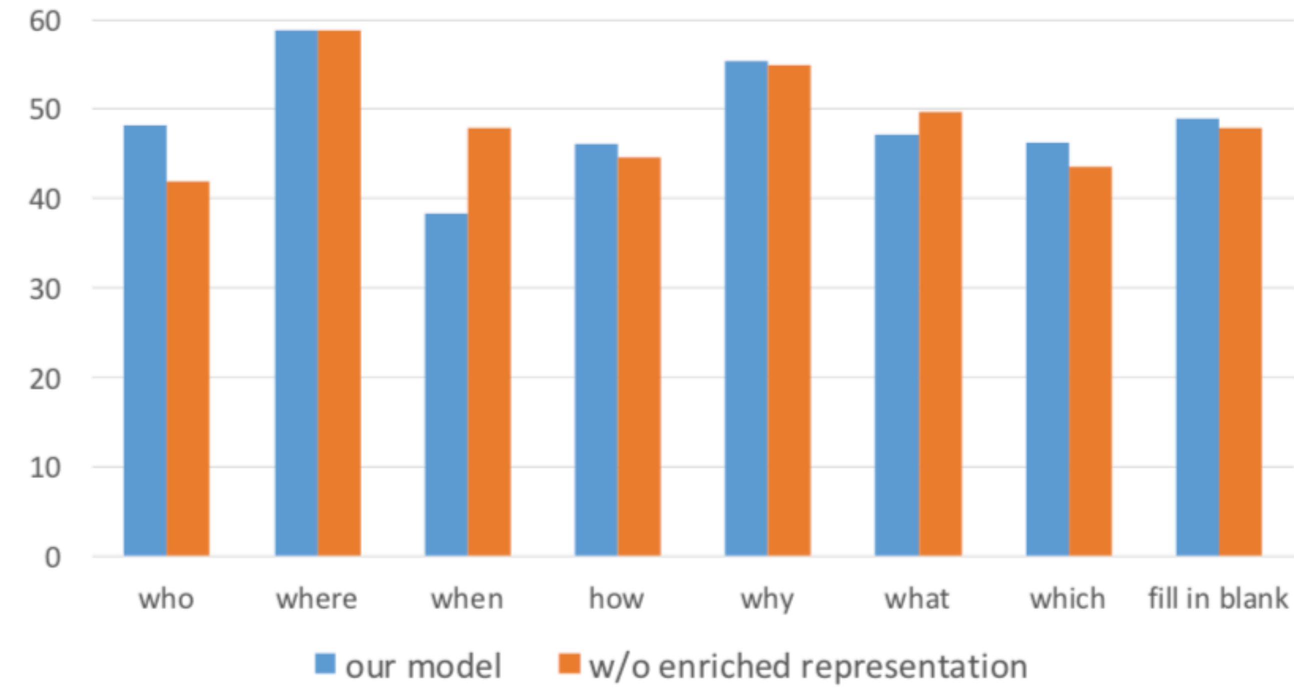


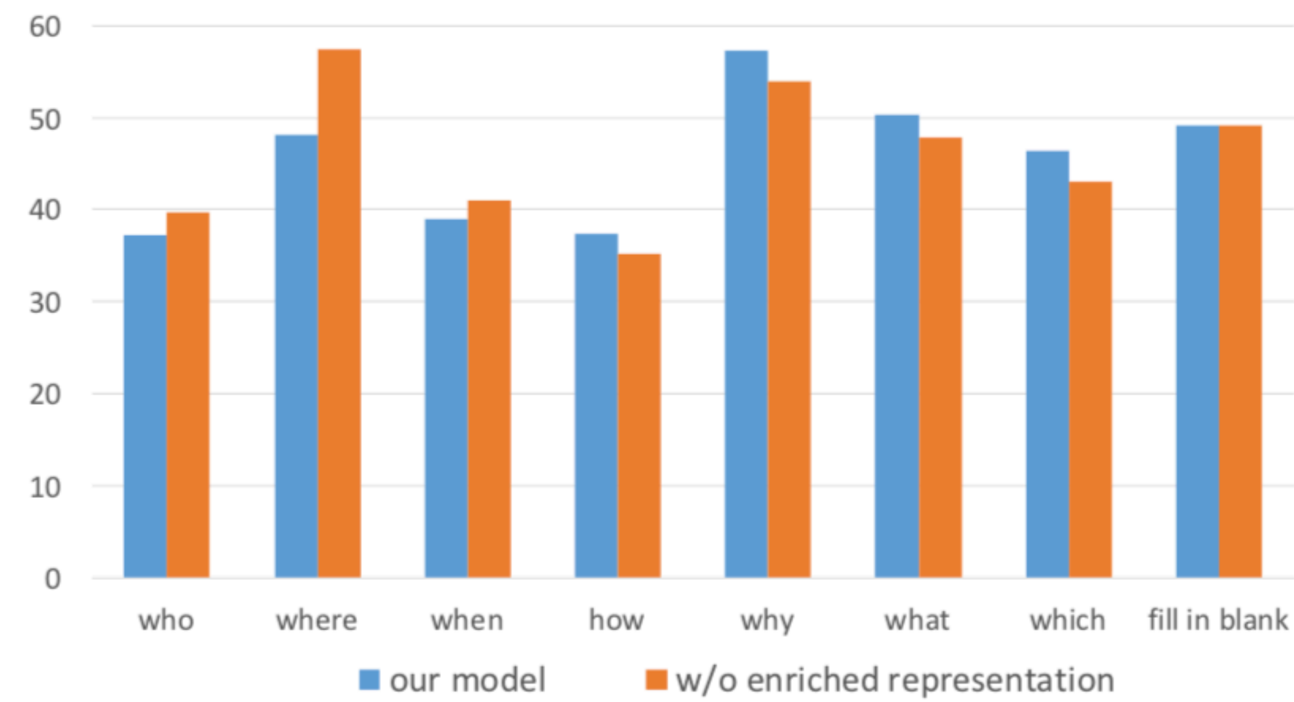


Chen and Cui et al., AAAI 2019. Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

Analysis

- **Quantitative Analysis on Different Type of Questions (on RACE data)**
  - Further efforts should be made on balancing the word-level attention and highly abstracted attention.





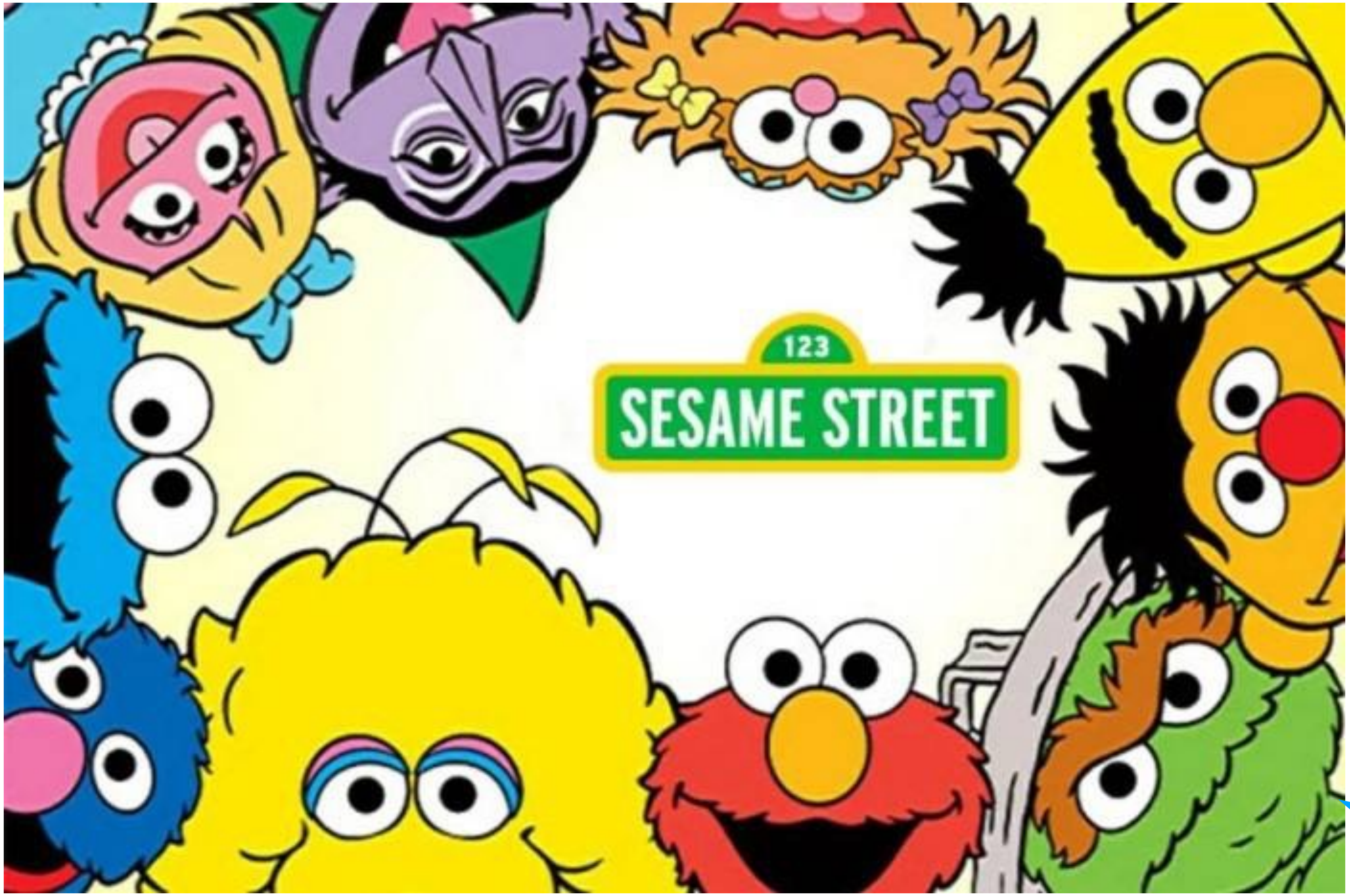
Chen and Cui et al., AAAI 2019. Convolutional Spatial Attention Model for Reading Comprehension with Multiple-Choice Questions

Analysis

- **Conclusions for CSA Model**
  - Propose Convolutional Spatial Attention model for RC with multiple-choice questions
  - The proposed model done well on hard problems types, such as ‘how’ and ‘why’
  - Experimental results show significant improvements on RACE and SemEval 2018 datasets

BERT-based MRC

BERT-based MRC



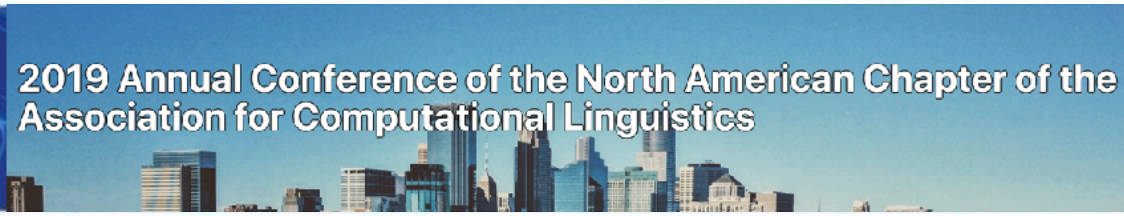
BERT

ERNIE

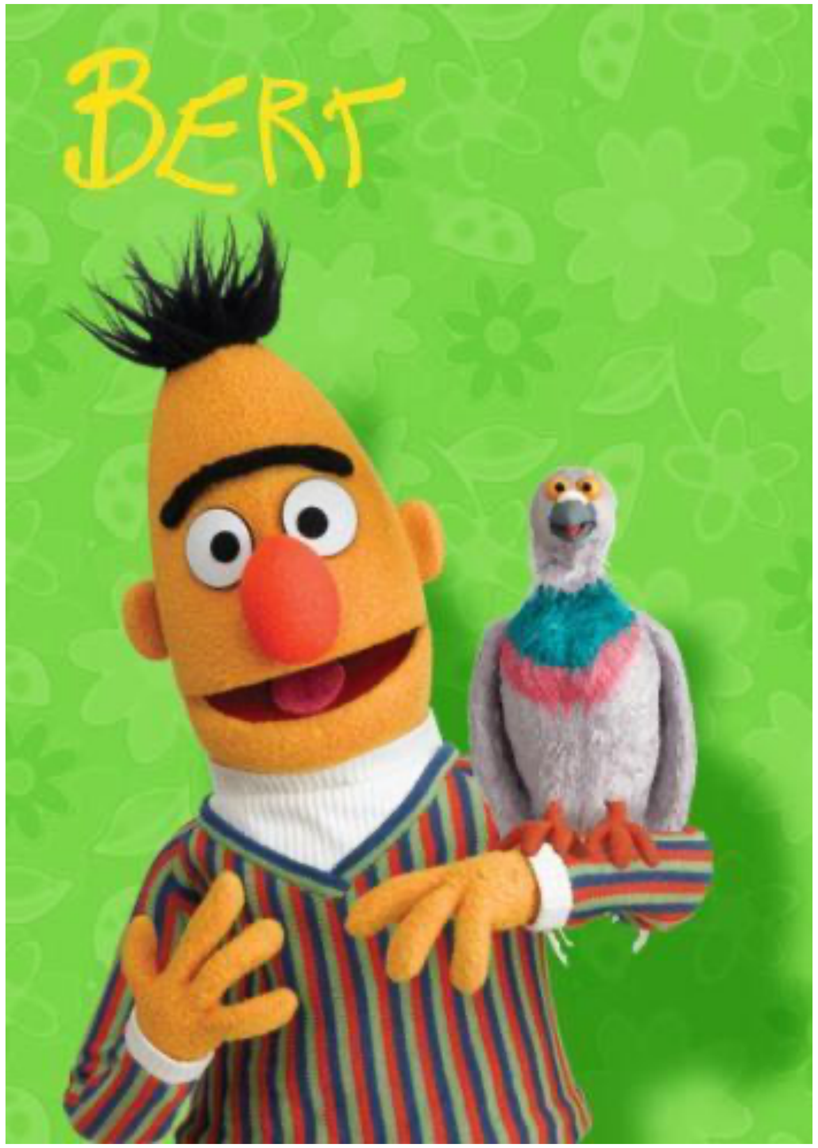
ELMo

Who will be the next?

BERT-based MRC



- **BERT: Bidirectional Encoder Representations from Transformers**
  - **NAACL 2019 Best Paper**
  - 16,000+ stars on GitHub
  - 600+ citations (only half a year)
  - State-of-the-art text representation



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT-based MRC

- **BERT: Bidirectional Encoder Representations from Transformers**
- **Contributions**
  - Demonstrate the importance of bidirectional pre-training for language representations
  - Pre-trained representations eliminate the needs of many heavily-engineered task-specific architectures
  - Pre-trained models are released to the community for future research

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT-based MRC

- **Pre-training Task I: Masked LM (MLM)**
  – Mask out several input words, and then predict the masked words

```
                          store               gallon
                            ↑                    ↑
the man went to the [MASK] to buy a [MASK] of milk
```

  – Too little masking: Easy to pick them out
  – Too much masking: Not enough context
  – In this paper, use a percentage of **15%**

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT-based MRC

- **Pre-training Task I: Masked LM (MLM)**
  - Problem: Mask token never appear at fine-tuning (realistic data)
  - Solution: 15% of the words to predict, but don't replace with `[MASK]` 100% of the time
  - Instead
    - 80% of the time, replace with `[MASK]`
      - `went to the store → went to the [MASK]`
    - 10% of the time, replace random word
      - `went to the store → went to the apple`
    - 10% of the time, keep the same word
      - `went to the store → went to the store`

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT-based MRC

- **Let's see how this implemented in source code**
- File: `create_pretraining_data.py`
- Function: `create_masked_lm_predictions()`
- Arguments
  - Tokens (list): tokenized sequence tokens
  - masked_lm_prob (float): how many words (proportion) should be masked
  - max_predictions_per_seq (int): maximum predictions per sequence
  - vocab_words (list): vocabulary
  - rng: random.Random(seed)

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT-based MRC

- **Generate candidate indexes**
  - Skip `[CLS]` and `[SEP]`
  - Shuffle candidate indexes
  - Determine the prediction number

```
cand_indexes = []
for (i, token) in enumerate(tokens):
  if token == "[CLS]" or token == "[SEP]":
    continue
  cand_indexes.append(i)

rng.shuffle(cand_indexes)

output_tokens = list(tokens)

num_to_predict = min(max_predictions_per_seq,
                     max(1, int(round(len(tokens) * masked_lm_prob))))
```

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT-based MRC

- **Mask out proper tokens**
  - Regular checks for overflow
  - Generate random number to determine the masking action

```
masked_lms = []
covered_indexes = set()
for index in cand_indexes:
  if len(masked_lms) >= num_to_predict:
    break
  if index in covered_indexes:
    continue
  covered_indexes.add(index)

  masked_token = None
  # 80% of the time, replace with [MASK]
  if rng.random() < 0.8:
    masked_token = "[MASK]"
  else:
    # 10% of the time, keep original
    if rng.random() < 0.5:
      masked_token = tokens[index]
    # 10% of the time, replace with random word
    else:
      masked_token = vocab_words[rng.randint(0, len(vocab_words) - 1)]

  output_tokens[index] = masked_token

  masked_lms.append(MaskedLmInstance(index=index, label=tokens[index]))
```

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT-based MRC

- **Pre-training Task II: Next Sentence Prediction (NSP)**
  - Learn the relationships between sentences, i.e., contextual information
  - Predict whether Sentence B is the actual sentence that comes after Sentence A, or a random sentence

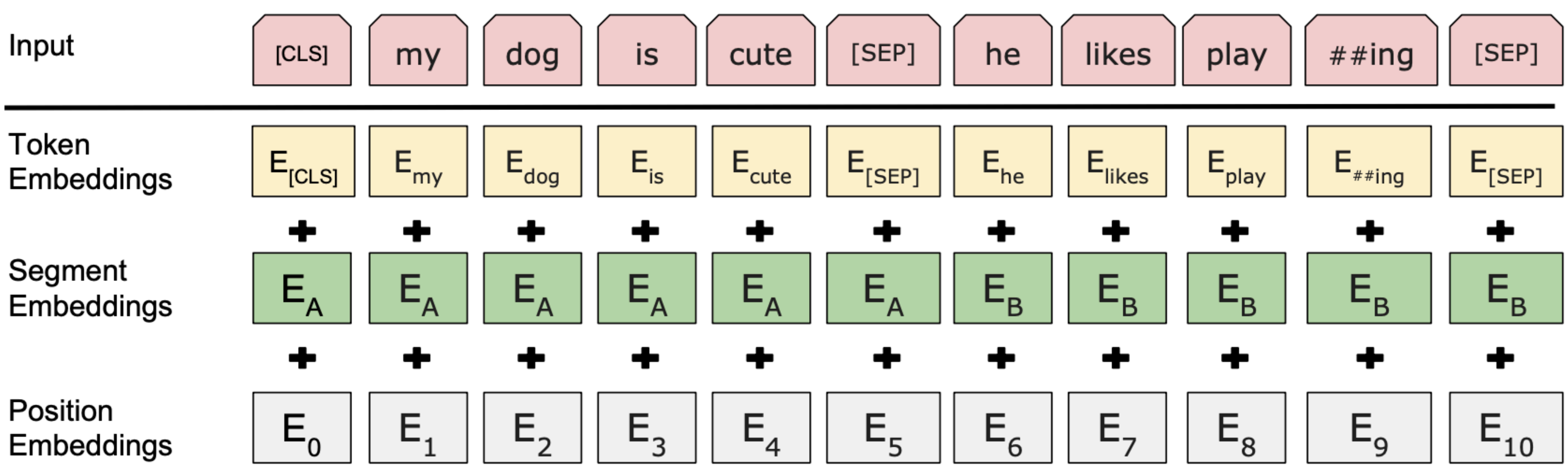
Sentence A = `The man went to the store.`
Sentence B = `He bought a gallon of milk.`
Label = `IsNextSentence`

Sentence A = `The man went to the store.`
Sentence B = `Penguins are flightless.`
Label = `NotNextSentence`

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT-based MRC

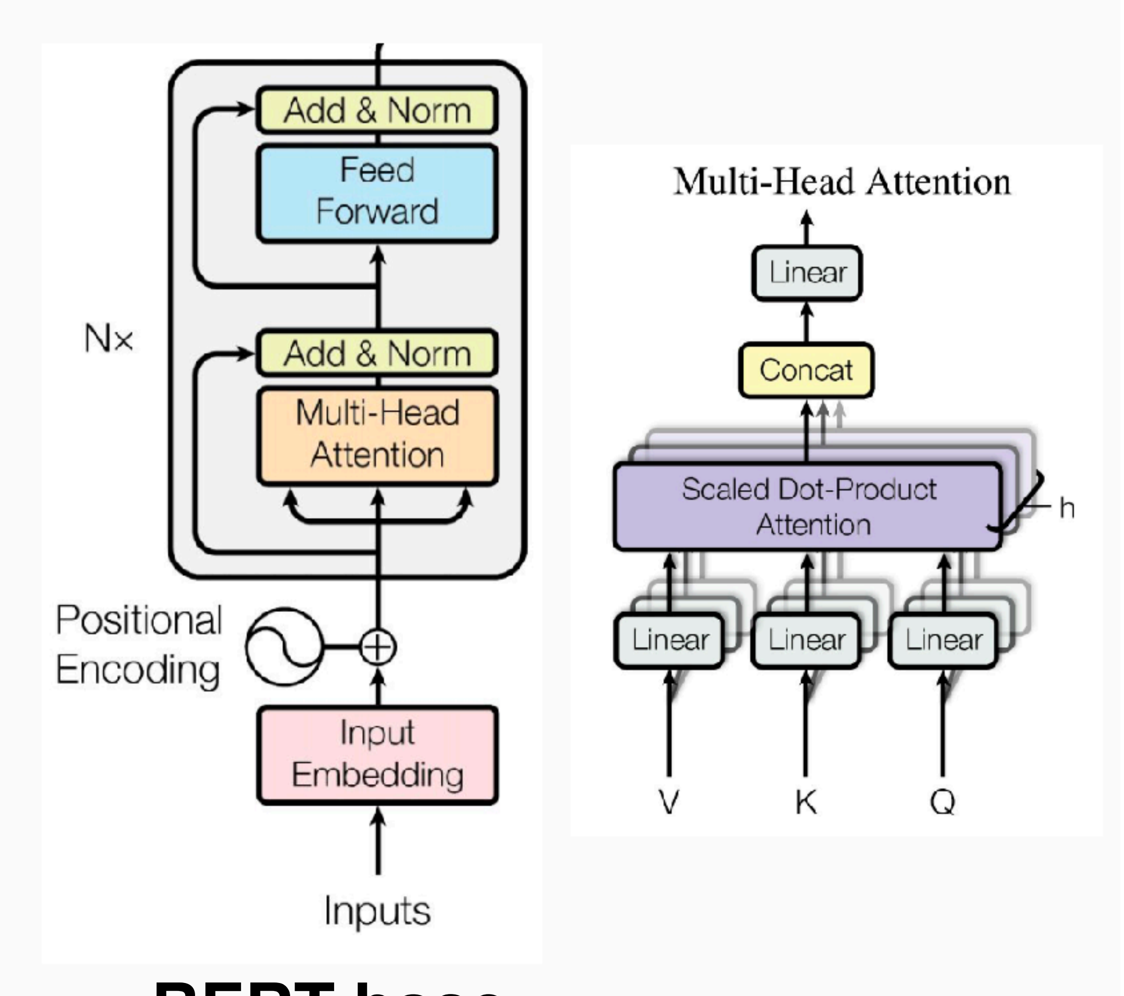
- **Input Representation**
  - Use 30,000 WordPiece vocabulary
  - The final input is the sum of three embeddings
    - Token Embeddings
    - Segment Embeddings
    - Position Embeddings



Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT-based MRC

- **Transformer Encoder**
  - **Multi-head self-attention**
    - Models context
  - **Feed-Forward Layers**
    - Computes non-linear hierarchical features
  - **LayerNorm and Residual Connection**
    - Makes training deep networks healthy
  - **Positional Embeddings**
    - Allows the model to learn relative positioning

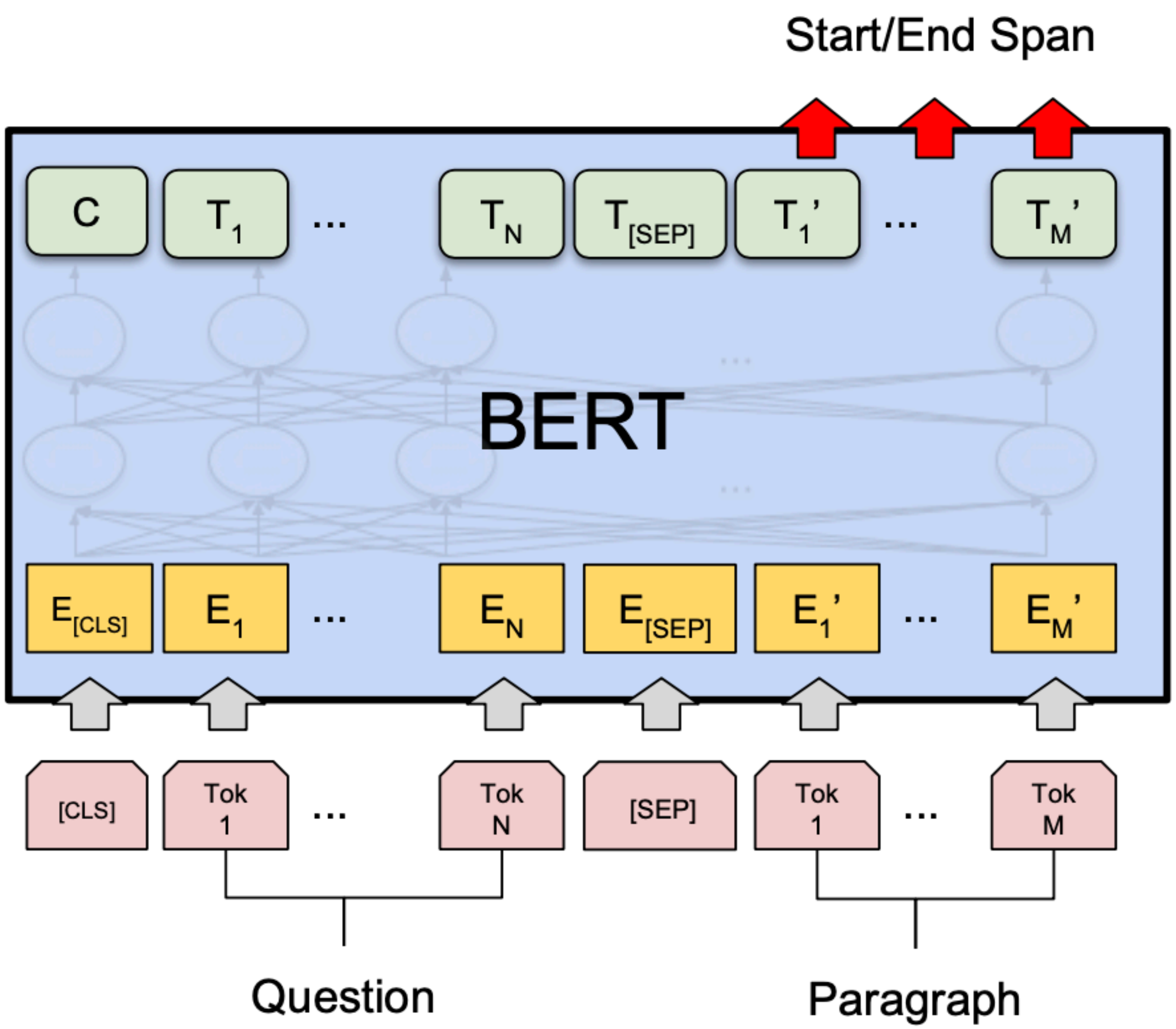


- **BERT-base**
  - 12-layer, 768-hidden, 12-head, 110M params
- **BERT-large**
  - 24-layer, 1024-hidden, 16-head, 340M params

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT-based MRC

- **Fine-tuning BERT on SQuAD Task**



Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

run_squad.py

- **Let's see how this implemented in source code**
- File: `run_squad.py`
- Function: `create_model()`
- Arguments
  - bert_config (json): BERT config file
  - is_training (bool): training mode option
  - input_ids (tensor): input ids for token embeddings
  - input_mask (tensor): input mask for indicating non-padding positions
  - segment_ids (tensor): segment_id tensor
  - use_one_hot_embeddings (bool)

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

run_squad.py

- **Generate BERT representation**
  - Define a BERT model
  - Generate sequence output (3D-tensor)

```
model = modeling.BertModel(
    config=bert_config,
    is_training=is_training,
    input_ids=input_ids,
    input_mask=input_mask,
    token_type_ids=segment_ids,
    use_one_hot_embeddings=use_one_hot_embeddings)

final_hidden = model.get_sequence_output()

final_hidden_shape = modeling.get_shape_list(final_hidden, expected_rank=3)
batch_size = final_hidden_shape[0]
seq_length = final_hidden_shape[1]
hidden_size = final_hidden_shape[2]
```

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

run_squad.py

- **Simple Output Layer for Span Prediction**
  - Define a fully-connected (dense) layer
  - Squeeze the vector to a scalar to get raw span output

```
output_weights = tf.get_variable(
    "cls/squad/output_weights", [2, hidden_size],
    initializer=tf.truncated_normal_initializer(stddev=0.02))

output_bias = tf.get_variable(
    "cls/squad/output_bias", [2], initializer=tf.zeros_initializer())

final_hidden_matrix = tf.reshape(final_hidden,
                                 [batch_size * seq_length, hidden_size])
logits = tf.matmul(final_hidden_matrix, output_weights, transpose_b=True)
logits = tf.nn.bias_add(logits, output_bias)

logits = tf.reshape(logits, [batch_size, seq_length, 2])
logits = tf.transpose(logits, [2, 0, 1])

unstacked_logits = tf.unstack(logits, axis=0)

(start_logits, end_logits) = (unstacked_logits[0], unstacked_logits[1])

return (start_logits, end_logits)
```

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

run_squad.py

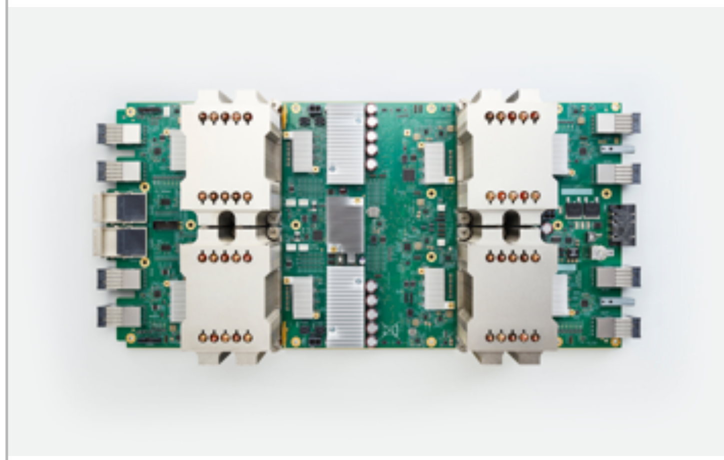
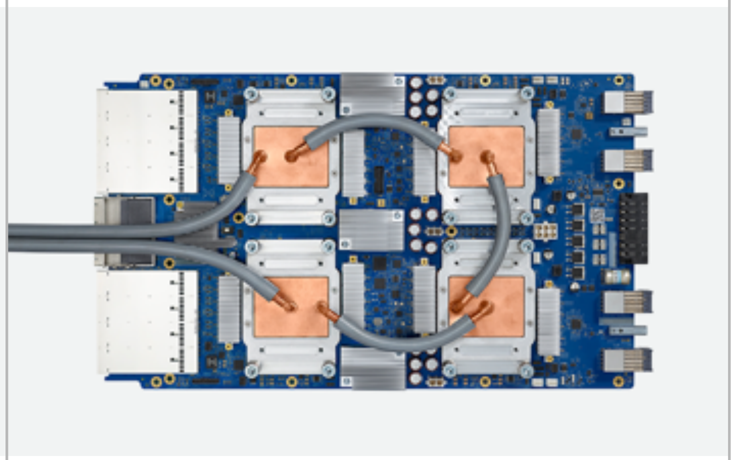
- **Create Loss for Span**
  - Function: `model_fn_builder()` → `compute_loss()`
  - Compute regular cross-entropy loss for start and end positions

```
def compute_loss(logits, positions):
  one_hot_positions = tf.one_hot(
      positions, depth=seq_length, dtype=tf.float32)
  log_probs = tf.nn.log_softmax(logits, axis=-1)
  loss = -tf.reduce_mean(
      tf.reduce_sum(one_hot_positions * log_probs, axis=-1))
  return loss
```

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

TPU

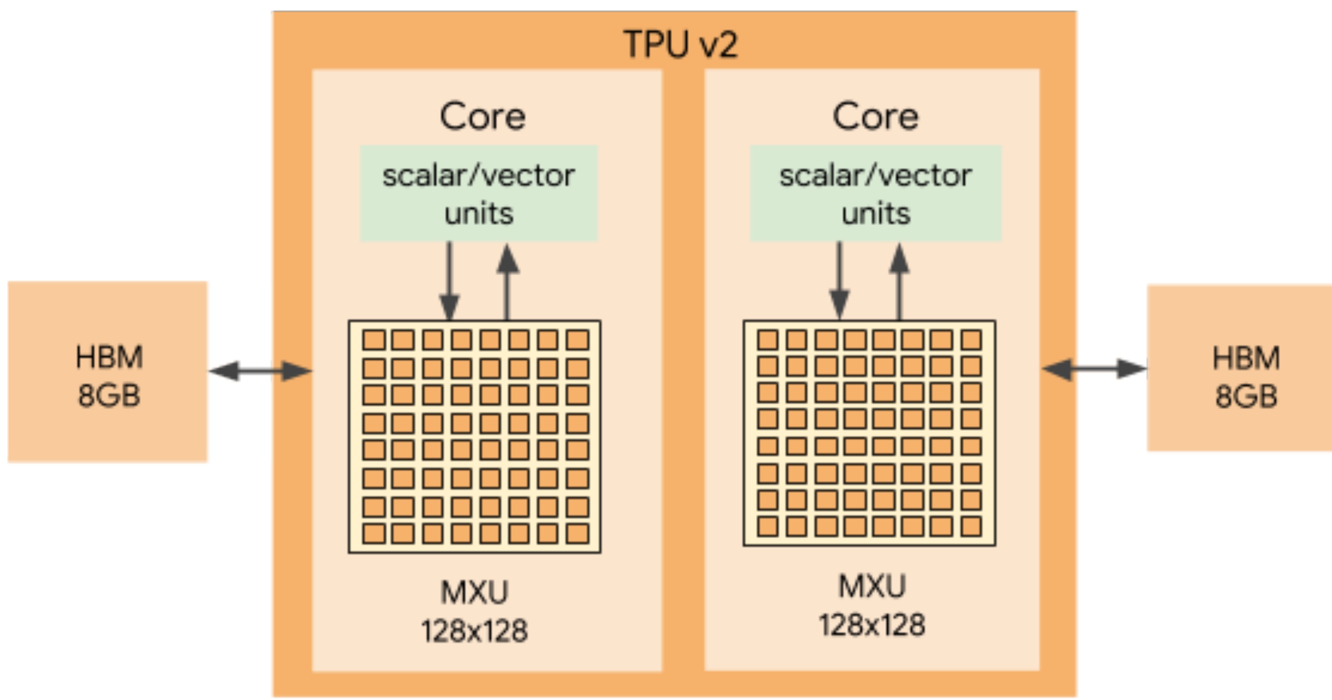
- **Before experiments, let's see what TPU is.**
- **https://cloud.google.com/tpu/**

	NVIDIA V100	TPU v2	TPU v3
			
ARCH	NVIDIA Volta GPU	Google Cloud TPU	Google Cloud TPU
MEM	16GB / 32GB	64GB	128GB
FLOPS	Double: 7 TFLOPS Single: 14 TFLOPS DL: 112 TFLOPS	180 TFLOPS	420 TFLOPS

Google Cloud TPU. https://cloud.google.com/tpu/

TPU

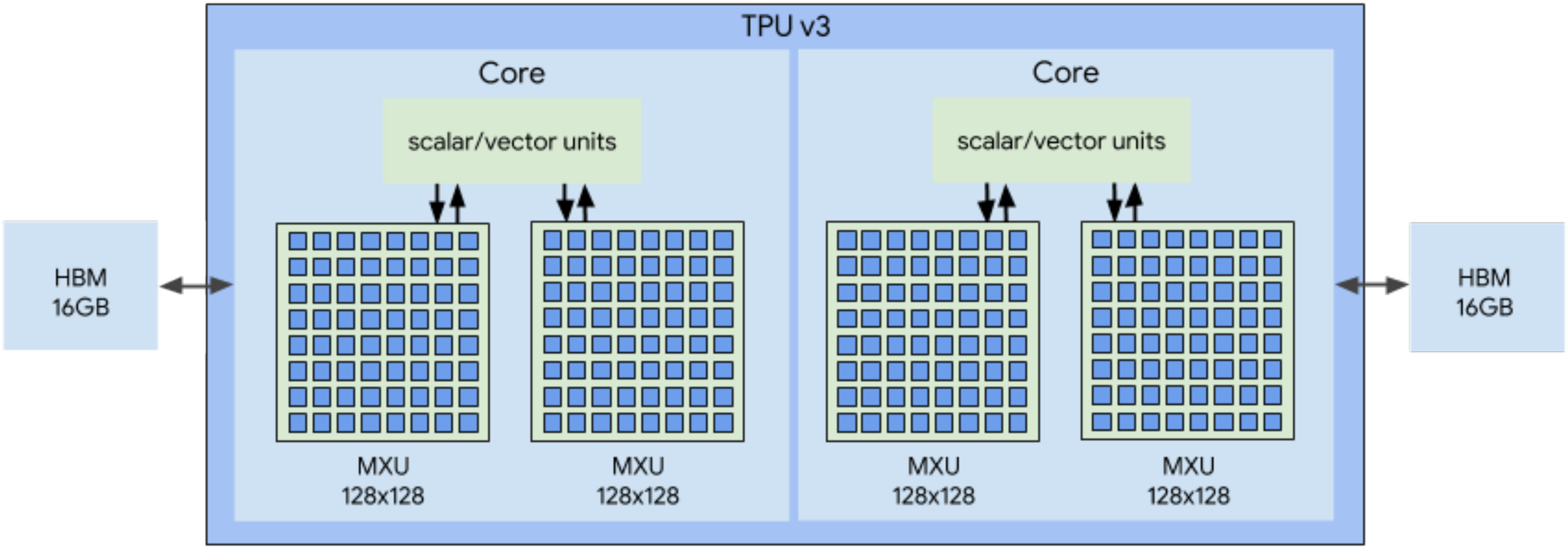
- **TPU v2 (64 GB HBM)**
  - 1 hardware: 4 chips
    - 1 chip: 2 cores, each core: 8GB HBM
  - 64 GB HBM = 4 chips * 2 cores * 8 GB
  - Price (per hour): 4.5 USD or 1.35 USD (preemptible)



Google Cloud TPU. https://cloud.google.com/tpu/

TPU

- **TPU v3 (128 GB HBM)**
  – 1 hardware: 4 chips
    - 1 chip: 2 cores, each core: 16GB HBM
  – 128 GB HBM = 4 chips * 2 cores * 16 GB
  – Price (per hour): 8.0 USD or 2.4 USD (preemptible)



Google Cloud TPU. https://cloud.google.com/tpu/

Experiments

- **Pre-training Setups**
  - Data: Wikipedia + BookCorpus (33B words in total)
  - Training: 256 batch * 512 max_token_length, 1M steps
  - Warmup: 10K steps
  - Time: 4 days
  - Computing Device
    - BERT-base: 4 Cloud TPUs in Pod config (16 chips)
    - BERT-large: 16 Cloud TPUs (64 chips)

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

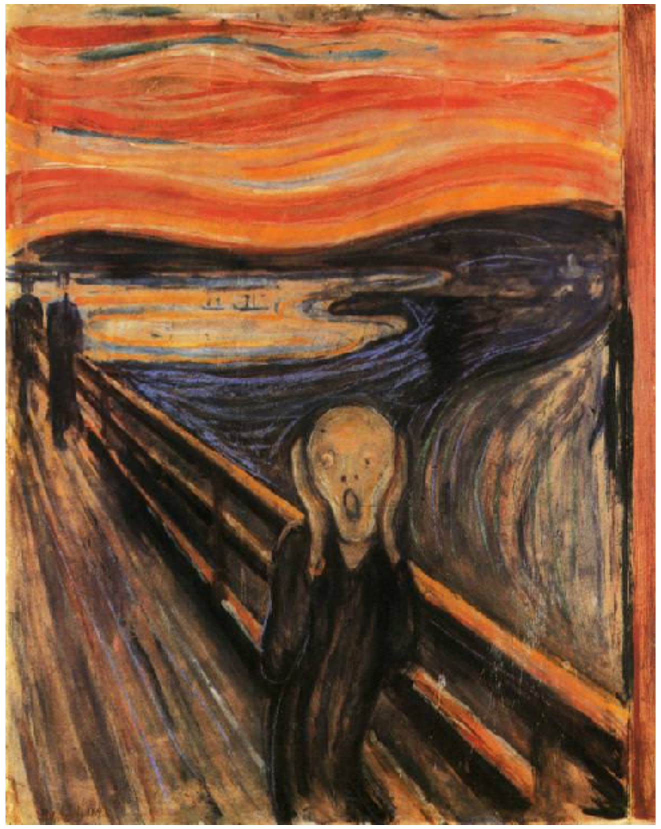
Experiments

- **Question: How much does it cost to train such a model?**
- Take BERT-large as an example,

16 Cloud TPUs = 16 * 4.5 = 72 USD / hour

One-day cost = 72 * 24 = 1,728 USD

Four-days cost = 1,728 USD * 4 = 6,912 USD



6,912 USD ≈ 47,715 CNY

**Actually, it costs way more,
as you won't be able to train
a model only once!**

Experiments

- **Results on SQuAD 1.1**
  - Substantially outperform all previous models, even ensemble models
  - BERT-large yields another significant gain over BERT-base
  - With data augmentation with TriviaQA data, another moderate gain could be obtained

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\mathrm{BERT}_{\mathrm{BASE}}$ (Single)	80.8	88.5	-	-
$\mathrm{BERT}_{\mathrm{LARGE}}$ (Single)	84.1	90.9	-	-
$\mathrm{BERT}_{\mathrm{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\mathrm{BERT}_{\mathrm{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\mathrm{BERT}_{\mathrm{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Experiments

- **Results on SQuAD 2.0**
  - Not surprisingly, BERT-large also got the best performance over previous works
  - After the release of BERT, almost all top-ranked system adopt BERT as a default manner.
  - With the powerful BERT, we managed to be the first one that surpassed average human performance

System	Dev EM	Dev F1	Test EM	Test F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	86.3	89.0	86.9	89.5
#1 Single - MIR-MRC (F-Net)	-	-	74.8	78.0
#2 Single - nlnet	-	-	74.2	77.1
Published				
unet (Ensemble)	-	-	71.4	74.9
SLQA+ (Single)	-		71.4	74.4
Ours				
$\text{BERT}_{\text{LARGE}}$ (Single)	78.7	81.9	80.0	83.1

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Nov 16, 2018	AoA + DA + BERT (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**82.374**	**85.310**
2 Nov 16, 2018	AoA + DA + BERT (single model) *Joint Laboratory of HIT and iFLYTEK Research*	81.178	84.251

Devlin et al., NAACL 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Chinese MRC Datasets

Chinese MRC

- **Cloze-Style (word / entity)**
  - PD & CFT (Cui et al., COLING 2016), WebQA (Li et al., 2016), CMRC 2017 (Cui et al., LREC 2018)
- **Span-Extraction**
  - CMRC 2018 (Cui et al., 2018), DuReader (He et al., MRQA 2018), DRCD (Shao et al., 2018)
- **Multiple-Choice**
  - C^3 (Sun et al., 2019)
- **Sentence Cloze-Style**
  - CMRC 2019 (Cui et al., 2019)
- Note that, we only list the dataset that has public access with proper technical report or paper.

PD & CFT

- **PD&CFT: People Daily and Children's Fairy Tale**
  - First Chinese cloze-style RC datasets, which add diversity in the community
  - Along with the traditional news datasets (People Daily), we also provide an out-of-domain dataset (Children's Fairy Tale)

Consensus Attention-based Neural Networks for Chinese Reading Comprehension

Yiming Cui[†*], **Ting Liu**[‡], **Zhipeng Chen**[†], **Shijin Wang**[†] and **Guoping Hu**[†]
[†]iFLYTEK Research, Beijing, China
[‡]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China
[†]{ymcui,zpchen,sjwang3,gphu}@iflytek.com
[‡]tliu@ir.hit.edu.cn

Cui et al., COLING 2016. Consensus Attention-based Neural Networks for Chinese Reading Comprehension

PD & CFT

- **Step 1: Select one sentence in the (truncated) document**

1 ||| People Daily (Jan 1). According to report of "New York Times", the Wall Street stock market continued to rise as the global stock market in the last day of 2013, ending with the highest record or near record of this year.
2 ||| "New York times" reported that the S&P 500 index rose 29.6% this year, which is the largest increase since 1997.
3 ||| Dow Jones industrial average index rose 26.5%, which is the largest increase since 1996.
4 ||| NASDAQ rose 38.3%.
5 ||| In terms of December 31, due to the prospects in employment and possible acceleration of economy next year, there is a rising confidence in consumers.
6 ||| As reported by Business Association report, consumer confidence rose to 78.1 in December, significantly higher than 72 in November.
7 ||| Also as "Wall Street journal" reported that 2013 is the best U.S. stock market since 1995.
8 ||| In this year, to chase the "silly money" is the most wise way to invest in U.S. stock.
9 ||| The so-called "silly money" strategy is that, to buy and hold the common combination of U.S. stock.
10 ||| This strategy is better than other complex investment methods, such as hedge funds and the methods adopted by other professional investors.

Cui et al., COLING 2016. Consensus Attention-based Neural Networks for Chinese Reading Comprehension

PD & CFT

- **Step 2: Choose one word in this sentence**
  - Only named entity and common noun is considered

1 ||| People Daily (Jan 1). According to report of "New York Times", the Wall Street stock market continued to rise as the global stock market in the last day of 2013, ending with the highest record or near record of this year.
2 ||| "New York times" reported that the S&P 500 index rose 29.6% this year, which is the largest increase since 1997.
3 ||| Dow Jones industrial average index rose 26.5%, which is the largest increase since 1996.
4 ||| NASDAQ rose 38.3%.
5 ||| In terms of December 31, due to the prospects in employment and possible acceleration of economy next year, there is a rising confidence in consumers.
6 ||| As reported by Business Association report, consumer confidence rose to 78.1 in December, significantly higher than 72 in November.
7 ||| Also as "Wall Street journal" reported that 2013 is the best U.S. stock market since 1995.
8 ||| In this year, to chase the "silly money" is the most wise way to invest in U.S. stock.
9 ||| The so-called "silly money" XXXXX is that, to buy and hold the common combination of U.S. stock.
10 ||| This strategy is better than other complex investment methods, such as hedge funds and the methods adopted by other professional investors.

Cui et al., COLING 2016. Consensus Attention-based Neural Networks for Chinese Reading Comprehension

PD & CFT

Colings 2016

- **Step 3: Leave out that word, and the sentence will become the query**

1 ||| People Daily (Jan 1). According to report of "New York Times", the Wall Street stock market continued to rise as the global stock market in the last day of 2013, ending with the highest record or near record of this year.
2 ||| "New York times" reported that the S&P 500 index rose 29.6% this year, which is the largest increase since 1997.
3 ||| Dow Jones industrial average index rose 26.5%, which is the largest increase since 1996.
4 ||| NASDAQ rose 38.3%.
5 ||| In terms of December 31, due to the prospects in employment and possible acceleration of economy next year, there is a rising confidence in consumers.
6 ||| As reported by Business Association report, consumer confidence rose to 78.1 in December, significantly higher than 72 in November.
7 ||| Also as "Wall Street journal" reported that 2013 is the best U.S. stock market since 1995.
8 ||| In this year, to chase the "silly money" is the most wise way to invest in U.S. stock.
9 ||| The so-called "silly money" XXXXX is that, to buy and hold the common combination of U.S. stock.
10 ||| This strategy is better than other complex investment methods, such as hedge funds and the methods adopted by other professional investors.

Document

Query

The so-called "silly money" XXXXX is that, to buy and hold the common combination of U.S. stock.

Cui et al., COLING 2016. Consensus Attention-based Neural Networks for Chinese Reading Comprehension

科大讯飞 iFLYTEK

PD & CFT

Colins 2016

- **Step 4: The removed word becomes the answer**

Document

1 ||| People Daily (Jan 1). According to report of "New York Times", the Wall Street stock market continued to rise as the global stock market in the last day of 2013, ending with the highest record or near record of this year.
2 ||| "New York times" reported that the S&P 500 index rose 29.6% this year, which is the largest increase since 1997.
3 ||| Dow Jones industrial average index rose 26.5%, which is the largest increase since 1996.
4 ||| NASDAQ rose 38.3%.
5 ||| In terms of December 31, due to the prospects in employment and possible acceleration of economy next year, there is a rising confidence in consumers.
6 ||| As reported by Business Association report, consumer confidence rose to 78.1 in December, significantly higher than 72 in November.
7 ||| Also as "Wall Street journal" reported that 2013 is the best U.S. stock market since 1995.
8 ||| In this year, to chase the "silly money" is the most wise way to invest in U.S. stock.
9 ||| The so-called "silly money" XXXXX is that, to buy and hold the common combination of U.S. stock.
10 ||| This strategy is better than other complex investment methods, such as hedge funds and the methods adopted by other professional investors.

Query

The so-called "silly money" XXXXX is that, to buy and hold the common combination of U.S. stock.

strategy Answer

Cui et al., COLING 2016. Consensus Attention-based Neural Networks for Chinese Reading Comprehension

PD & CFT

- **Comparison of cloze-style RC datasets**

Dataset	Language	Genre	Blank Type	Doc	Query
CNN/DM	English	News	NE	News Article	Summary w/ a blank
CBT	English	Story	NE,CN,V,P	20 consecutive sentences	21th sentence w/ a blank
PD&CFT	Chinese	News, story	NE,CN	Doc w/ a blank	the sentence that blank belongs to

Cui et al., COLING 2016. Consensus Attention-based Neural Networks for Chinese Reading Comprehension

WebQA

- **WebQA: Large-scale real-world factoid QA dataset**
  - WebQA is significantly larger (42K questions) than previous datasets
  - All questions are asked (natural annotation) by real-world users in daily life
  - http://paddlepaddle.bj.bcebos.com/dataset/webqa/WebQA.v1.0.zip

Dataset and Neural Recurrent Sequence Labeling Model for Open-Domain Factoid Question Answering

Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, Wei Xu
Baidu Research - Institute of Deep Learning
{lipeng17,liwei26,hezhengyan,wangxuguang,caoying03,
zhoujie01,wei.xu}@baidu.com

CMRC 2017

LREC 2018 MIYAZAKI

- **CMRC 2017: The First Evaluation Workshop on Chinese Machine Reading Comprehension**
  - A cloze-style reading comprehension dataset
  - Data #: ~364k questions
  - Domain: Children's book
  - https://github.com/ymcui/cmrc2017

Dataset for the First Evaluation on Chinese Machine Reading Comprehension

Yiming Cui[†], Ting Liu[‡], Zhipeng Chen[†], Wentao Ma[†], Shijin Wang[†] and Guoping Hu[†]
[†]Joint Laboratory of HIT and iFLYTEK, iFLYTEK Research, Beijing, China
[‡]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China
[†]{ymcui,zpchen,wtma,sjwang3,gphu}@iflytek.com
[‡]tliu@ir.hit.edu.cn

DuReader

MRQA 2018: Machine Reading for Question Answering

Workshop at ACL 2018

Date: Thursday, July 19, 2018

Room: 210

- **DuReader: A Chinese Machine Reading Comprehension Dataset from Real-world Applications**
  - Open-domain MRC datasets
  - Data #: ~200k questions
  - Domain: Articles from Baidu Search and Zhidao
  - https://github.com/baidu/DuReader

DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, Haifeng Wang

Baidu Inc., Beijing, China

{hewei06, liukai20, liujing46, lvyajuan, zhaoshiqi, xiaoxinyan, liuyuan04, wangyizhong01, wu_hua, sheqiaoqiao, liuxuan, wutian, wanghaifeng}@baidu.com

DuReader

- **Example**

Question	学士服颜色/ What are the colors of academic dresses?
Question Type	*Entity-Fact*
Answer 1	**[绿色, 灰色, 黄色, 粉色]**：农学学士服绿色，理学学士服灰色，工学学士服黄色，管理学学士服灰色，法学学士服粉色，文学学士服粉色，经济学学士服灰色。/ **[green, gray, yellow, pink]** Green for Bachelor of Agriculture, gray for Bachelor of Science, yellow for Bachelor of Engineering, gray for Bachelor of Management, pink for Bachelor of Law, pink for Bachelor of Art, gray for Bachelor of Economics
Document 1	农学学士服绿色，理学学士服灰色，...，确定为文、理、工、农、医、军事六大类，与此相应的饰边颜色为粉、灰、黄、绿、白、红六种颜色。
...	
Document 5	学士服是学士学位获得者在学位授予仪式上穿戴的表示学位的正式礼服，...，男女生都应着深色皮鞋。
Question	智慧牙一定要拔吗/ Do I have to have my wisdom teeth removed
Question Type	*YesNo-Opinion*
Answer 1	**[Yes]**因为智齿很难清洁的原因，比一般的牙齿容易出现口腔问题，所以医生会建议拔掉/ **[Yes]** The wisdom teeth are difficult to clean, and cause more dental problems than normal teeth do, so doctors usually suggest to remove them
Answer 2	**[Depend]**智齿不一定非得拔掉，一般只拔出有症状表现的智齿，比如说经常引起发炎... / **[Depend]** Not always, only the bad wisdom teeth need to be removed, for example, the one often causes inflammation ...
Document 1	为什么要拔智齿？智齿好好的医生为什么要建议我拔掉?主要还是因为智齿很难清洁...
...	
Document 5	根据我多年的临床经验来说,智齿不一定非得拔掉.智齿阻生分好多种...

He et al., MRQA 2018. DuReader: A Chinese Machine Reading Comprehension Dataset from Real-world Applications

CMRC 2018

- **CMRC 2018: The Second Evaluation Workshop on Chinese Machine Reading Comprehension**
  - SQuAD-like dataset in Simplified Chinese
  - Data #: ~18K question
  - Domain: Wikipedia
  - https://github.com/ymcui/cmrc2018

A Span-Extraction Dataset for Chinese Machine Reading Comprehension

Yiming Cui[†‡], **Ting Liu**[‡],
Li Xiao[†], **Zhipeng Chen**[†], **Wentao Ma**[†], **Wanxiang Che**[‡], **Shijin Wang**[†], **Guoping Hu**[†]
[†]Joint Laboratory of HIT and iFLYTEK (HFL), iFLYTEK Research, Beijing, China
[‡]Research Center for Social Computing and Information Retrieval (SCIR),
Harbin Institute of Technology, Harbin, China
[†]{ymcui,lixiao3,zpchen,wtma,sjwang3,gphu}@iflytek.com
[‡]{ymcui,tliu,car}@ir.hit.edu.cn

CMRC 2018

- **Example (Normal)**


```
[
    {
        "title": "傻钱策略"
        "context_id": "TRIAL_0"
        "context_text": "工商协进会报告，12月消费者信心上升到78.1，明显高于11月的72。另据《华尔街日报》报道，2013年是1995年以来美国股市表现最好的一年。这一年里，投资美国股市的明智做法是追着“傻钱”跑。所谓的“傻钱”策略，其实就是买入并持有美国股票这样的普通组合。这个策略要比对冲基金和其它专业投资者使用的更为复杂的投资方法效果好得多。"
        "qas":[
                {
                "query_id": "TRIAL_0_QUERY_0",
                "query_text": "什么是傻钱策略？",
                "answers": [
                     "所谓的“傻钱”策略，其实就是买入并持有美国股票这样的普通组合",
                     "其实就是买入并持有美国股票这样的普通组合",
                     "买入并持有美国股票这样的普通组合"
                    ]
                },
                {
                "query_id": "TRIAL_0_QUERY_1",
                "query_text": "12月的消费者信心指数是多少？",
                "answers": [
                    "78.1",
                    "78.1",
                    "78.1"
                    ]
                }
            ]
    }
]
```


Cui et al., arXiv pre-print. A Span-Extraction Dataset for Chinese Machine Reading Comprehension

CMRC 2018

第二届"讯飞杯"中文机器阅读理解评测 (CMRC 2018)
The 2nd Evaluation Workshop on Chinese Machine Reading Comprehension
2018年10月19日
湖南，长沙

- **Example (Challenge)**

[Document]
《黄色脸孔》是柯南·道尔所著的福尔摩斯探案的56个短篇故事之一，收录于《福尔摩斯回忆录》。孟罗先生素来与妻子恩爱，但自从最近邻居新入伙后，孟罗太太则变得很奇怪，曾经凌晨时份外出，又藉丈夫不在家时偷偷走到邻居家中。于是孟罗先生向福尔摩斯求助，福尔摩斯听毕孟罗先生的故事后，认为孟罗太太被来自美国的前夫勒索，所以不敢向孟罗先生说出真相，所以吩咐孟罗先生，如果太太再次走到邻居家时，即时联络他，他会第一时间赶到。孟罗太太又走到邻居家，福尔摩斯陪同孟罗先生冲入，却发现邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿，所以不敢说出真相。

[Question]
孟罗太太为什么在邻居新入伙后变得很奇怪?

[Answer 1]
邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿

[Answer 2]
邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿，所以不敢说出真相。

[Answer 3]
邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿，所以不敢说出真相。

Cui et al., arXiv pre-print. A Span-Extraction Dataset for Chinese Machine Reading Comprehension

DRCD

- **DRCD: A Span-Extraction MRC Dataset in Traditional Chinese**
  - SQuAD-like dataset in Traditional Chinese
  - Data #: ~30k questions
  - Domain: Wikipedia
  - https://github.com/DRCSolutionService/DRCD

DRCD: a Chinese Machine Reading Comprehension Dataset

Chih Chieh Shao and **Trois Liu** and **Yuting Lai** and **Yiying Tseng** and **Sam Tsai**
{cchieh.shao,trois.liu,yuting.lai,yiying.tz,i-sam.tsai}@deltaww.com
Delta Research Center
Delta Electronics, Inc.

C³

- **C³: Probing Prior Knowledge Needed in Challenging Chinese Machine Reading Comprehension**
  - MRC with multiple-choice questions
  - Data #: ~24K questions
  - Domain: HSK, MHK
  - https://dataset.org/c3/

Probing Prior Knowledge Needed in Challenging Chinese Machine Reading Comprehension

Kai Sun[1]* Dian Yu[2] Dong Yu[2] Claire Cardie[1]
[1]Cornell University, Ithaca, NY
[2]Tencent AI Lab, Bellevue, WA
ks985@cornell.edu, {yudian, dyu}@tencent.com, cardie@cs.cornell.edu

- # Example

1928年，经徐志摩介绍，时任中国公学校 长的胡适聘用了沈从文做讲师，主讲大学 一年级的现代文学选修课。
当时，沈从文已经在文坛上崭露头角，在 社会上也小有名气，因此还未到上课时 间，教室里就坐满了学生。上课时间到 了，沈从文走进教室，看见下面黑压压 一片，心里陡然一惊，脑子里变得一片空 白，连准备了无数遍的第一句话都堵在嗓 子里说不出来了。 他呆呆地站在那里，面色尴尬至极，双手 拧来拧去无处可放。上课前他自以为成竹 在胸，所以就没带教案和教材。整整10 分 钟，教室里鸦雀无声，所有的学生都好奇 地等着这位新来的老师开口。沈从文深吸 了一口气，慢慢平静了下来，原先准备好 的东西也重新在脑子里聚拢，然后他开始 讲课了。不过由于他依然很紧张，原本预 计一小时的授课内容，竟然用了不到15 分 钟就讲完了。 接下来怎么办?他再次陷入了窘境。无奈 之下，他只好拿起粉笔在黑板上写道:我 第一次上课，见你们人多，怕了。
顿时，教室里爆发出了一阵善意的笑声， 随即一阵鼓励的掌声响起。得知这件事之 后，胡适对沈从文大加赞赏，认为他非常 成功。 有了这次经历，在以后的课堂上，沈从文 都会告诫自己不要紧张，渐渐地，他开始 在课堂上变得从容起来。

Q1 第2段中，“黑压压一片”指的是：
A. 教室很暗
B. 听课的人多★
C. 房间里很吵
D. 学生们发言很积极

Q2 沈从文没拿教材，是因为他觉得：
A. 讲课内容不多
B. 自己准备得很充分★
C. 这样可以减轻压力
D. 教材会限制自己的发挥

Q3 看见沈从文写的那句话，学生们：
A. 急忙安慰他
B. 在心里埋怨他
C. 受到了极大的鼓舞
D. 表示理解并鼓励了他★

Q4 上文主要谈的是：
A. 中国教育制度的发展
B. 紧张时应如何调整自己
C. 沈从文第一次讲课时的情景★
D. 沈从文如何从作家转变为教师的

Sun et al., arXiv pre-print. C^3: Probing Prior Knowledge Needed in Challenging Chinese Machine Reading Comprehension

CMRC 2019

- **CMRC 2019: The Third Evaluation Workshop on Chinese Machine Reading Comprehension**
  - Sentence Cloze-Style MRC Dataset
  - Data #: ~18K
  - Features
    - We propose Sentence Cloze-Style MRC (SCMRC) task to fill in the blank with proper candidate sentence
    - The blank is composed by sentence, which forces the machine to learn longer contextual information
    - Fake candidate sentence (very similar to the real one) is added, which is much more challenging for MRC
  - https://github.com/ymcui/cmrc2019

- **Example**


```
{
    "data": [
        {
            "context": "森林里有一棵大树，树上有一个鸟窝。[BLANK1]，还从来没有看到过鸟宝宝长什么样。
            小松鼠说：“我爬到树上去看过，鸟宝宝光溜溜的，身上一根羽毛也没有。” “我不相信，”小白兔说，“所有的鸟都是有羽毛的。”
            “鸟宝宝没有羽毛。”小松鼠说，“你不信自己去看。”
            小白兔不会爬树，它没有办法去看。小白兔说：“我请蓝狐狸去看一看，我相信蓝狐狸的话。” 小松鼠说：“蓝狐狸跟你一样，也不会爬树。”
            蓝狐狸说：“我有魔法树叶，我能变成一只狐狸鸟。” [BLANK2]，一下子飞到了树顶上。 “蓝狐狸，你看到了吗？”小白兔在树下大声喊。
            “我看到了，鸟窝里有四只小鸟，他们真是光溜溜的，一根羽毛也没有。”蓝狐狸说。 就在这时候，鸟妈妈和鸟爸爸回来了，
            [BLANK3]，立刻大喊大叫： “抓强盗啊！抓强盗啊！强盗闯进了我们家里，想偷我们的孩子！”
            [BLANK4]，全都飞了过来。他们扇着翅膀，朝蓝狐狸冲过来，用尖尖的嘴啄他，用爪子抓他。 蓝狐狸扑扇翅膀，赶紧飞。
            鸟儿们排着队伍，紧紧追上来。[BLANK5]，它飞得不高，也飞得不快。 “救命啊，救命！”蓝狐狸说，“我不是强盗，我是蓝狐狸！”
            小白兔在草丛说：“你不是鸟，你飞不过他们，你赶快变回狐狸吧！” 蓝狐狸急忙落到地上，变回了狐狸，躲进深深的草丛里。
            鸟儿们找不到蓝狐狸，只得飞走了。 蓝狐狸对小白兔说：“谢谢你。”",
            "choices": [
                "蓝狐狸是第一次变成狐狸鸟",
                "森林里所有的鸟听到喊声",
                "他们看到鸟窝里蹲着一只蓝色的大鸟",
                "蓝狐狸真的变成了一只蓝色的大鸟",
                "小动物们只看到过鸟妈妈和鸟爸爸在鸟窝里飞进飞出",
                "小松鼠变成了一只蓝色的大鸟"
                ],
            "context_id": "SAMPLE_00002",
            "answers": [4,3,2,1,0]
        }
    ]
}
```


Fake candidate

科大讯飞 iFLYTEK

Dataset Summary

Dataset	Genre	Query Type	Answer Type	Document #	Query #
PD&CFT [1]	News & Tale	Cloze	Word	28K	100K
WebQA [2]	Web	User log	Entity	-	42K
CMRC 2017 [3]	Tale	Cloze & Query	Word	-	364K
DuReader [4]	Web	User log	Free form	1M	200K
CMRC 2018 [5]	Wiki	Query	Span	-	18K
DRCD繁体 [6]	Wiki	Query	Span	-	34K
C³ [7]	Mixed	Query	Choice	14K	24K
CMRC 2019 [8]	Story	Cloze	Sentence	1K	100K

Chinese Pre-Trained Models

Overview

- **Chinese Pre-trained Models**
  - Chinese BERT (Google)
  - ERNIE (Baidu)
  - Chinese BERT-wwm (HFL)

Chinese BERT

- **As a part of BERT open-source program, Google released pre-trained Chinese BERT**
  - Data: Chinese Wikipedia dump
  - Sample #: 24M sentences
  - Model: BERT-base (12-layer, 768-hidden, 12-heads, 110M parameters)
  - Framework: TensorFlow

System	Chinese
XNLI Baseline	67.0
BERT Multilingual Model	74.2
BERT Chinese-only Model	77.2

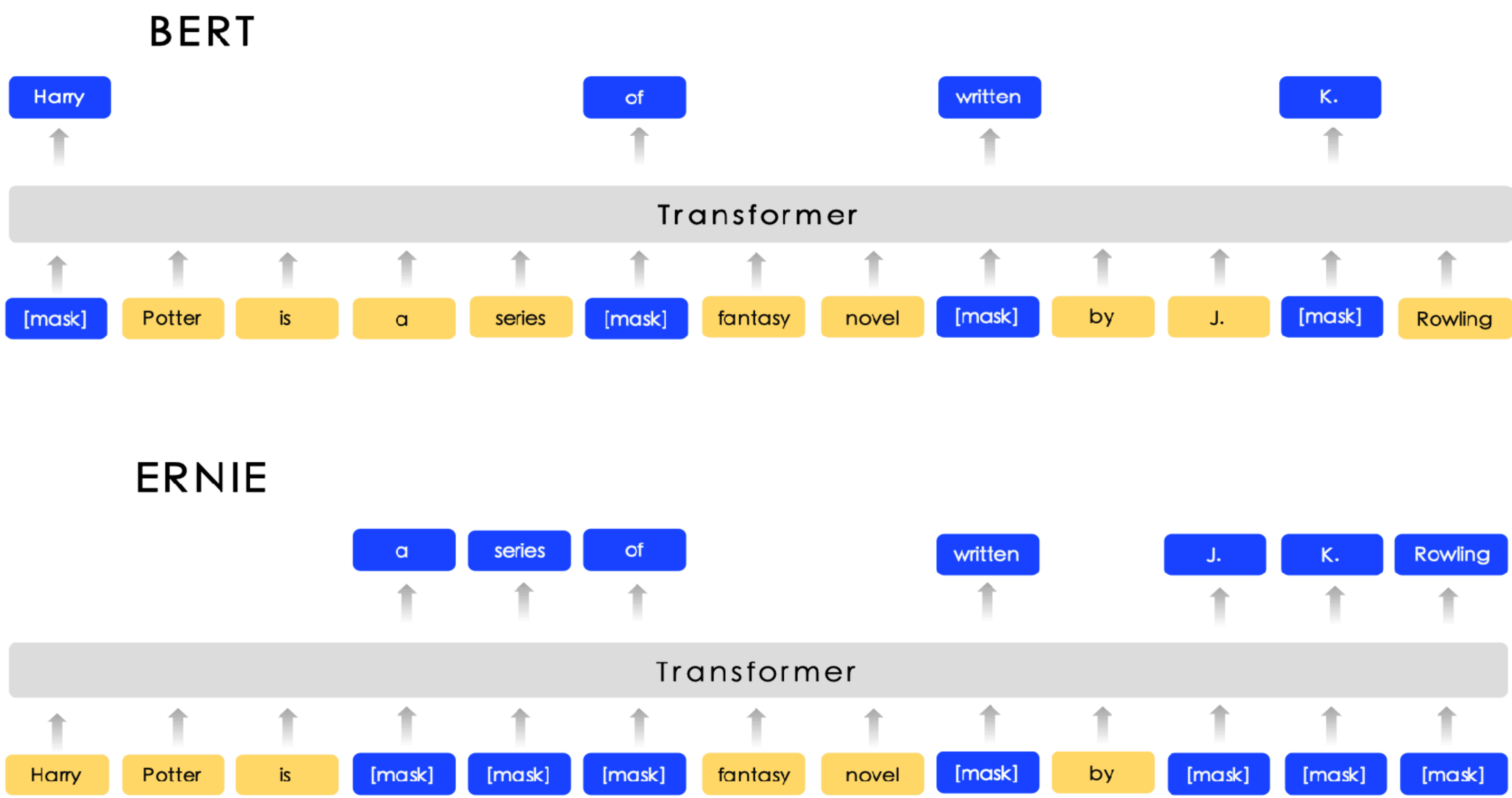
ERNIE

- **ERNIE: Enhanced Representation through kNowledge IntEgration**
  - Released by Baidu on April 2019
  - Data: Chinese Wikipedia, Baidu Baike/News/Tieba
  - Sample #: 21M, 51M, 47M, 54M → 173M
  - Model: BERT-base
  - Framework: PaddlePaddle

ERNIE: Enhanced Representation through Knowledge Integration

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng
Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, Hua Wu
Baidu Inc.
{sunyu02,wangshuohuan,liyukun01,fengshikun01,tianhao,wu_hua}@baidu.com

- **Key Idea**



Sun et al., arXiv pre-print. ERNIE: Enhanced Representation through Knowledge Integration

ERNIE

- **Basic-Level Masking**
  - 15% basic language units are masked.
- **Phrase-Level Masking**
  - Consecutive words are masked. The phrase boundary is identified by lexical analysis and chunking tools.
- **Entity-Level Masking**
  - Mask named entity, such as person names, locations, organizations, etc.

Sentence	Harry	Potter	is	a	series	of	fantasy	novels	written	by	British	author	J.	K.	Rowling
Basic-level Masking	[mask]	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	J.	[mask]	Rowling
Entity-level Masking	Harry	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]
Phrase-level Masking	Harry	Potter	is	[mask]	[mask]	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]

Sun et al., arXiv pre-print. ERNIE: Enhanced Representation through Knowledge Integration

- **Experimental Results**
    - Yields significant improvements over BERT on five Chinese datasets

Data	XNLI		LCQMC		MSRA-NER(SIGHAN 2006)		ChnSentiCorp		nlpcc-dbqa			
Eval	acc		acc		f1-score		acc		mrr		f1-score	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
BERT	78.1	77.2	88.8	87.0	94.0	92.6	94.6	94.3	94.7	94.6	80.7	80.8
ERNIE	79.9 (**+1.8**)	78.4 (**+1.2**)	89.7 (**+0.9**)	87.4 (**+0.4**)	95.0 (**+1.0**)	93.8 (**+1.2**)	95.2 (**+0.6**)	95.4 (**+1.1**)	95.0 (**+0.3**)	95.1 (**+0.5**)	82.3 (**+1.6**)	82.7 (**+1.9**)

Sun et al., arXiv pre-print. ERNIE: Enhanced Representation through Knowledge Integration

Chinese BERT-wwm

- **Background**
  - Google released new pre-trained BERT-large model on Github
  - The modification was called "Whole Word Masking"
  - BERT-large-wwm yields another significant improvement over BERT-large

Model	SQuAD 1.1 F1/EM	Multi NLI Accuracy
BERT-Large, Uncased (Original)	91.0/84.3	86.05
BERT-Large, Uncased (Whole Word Masking)	92.8/86.7	87.07
BERT-Large, Cased (Original)	91.5/84.8	86.09
BERT-Large, Cased (Whole Word Masking)	92.9/86.7	86.46

Chinese BERT-wwm

- **Whole Word Masking**
  - In the original pre-processing code, we randomly select WordPiece tokens to mask.
  - In WWM, we always mask all of the tokens corresponding to a word at once. The overall masking rate remains the same.

	Example
Original Sentence	the man jumped up , put his basket on phil ##am ##mon ' s head
Original Masked Input	[MASK] man [MASK] up , put his [MASK] on phil [MASK] ##mon ' s head
BERT-wwm Input	the man [MASK] up , put his basket on [MASK] [MASK] [MASK] ' s head

Chinese BERT-wwm

- **Important Note on Whole Word Masking**
  - Terminology ‘**Masking**’ does not ONLY represent replacing a word into `[MASK]` token. It could also be in another form, such as ‘keep original word’ or ‘randomly replaced by another word’.

Original Sentence: there is an apple tree nearby.	
Tokenized Sentence: ["there", "is", "an", "ap", "##p", "##le", "tr", "##ee", "nearby", “."]	
w/o wwm	there [MASK] an ap [MASK] ##le tr [RANDOM] nearby . [MASK] [MASK] an ap ##p [MASK] tr ##ee nearby . there is [MASK] ap ##p ##le [MASK] ##ee [MASK] . there is [MASK] ap [MASK] ##le tr ##ee nearby [MASK] . there is an! ap ##p ##le tr [MASK] nearby [MASK] . there is an [MASK] ##p [MASK] tr ##ee nearby [MASK] .
w/ wwm	there is an [MASK] [MASK] [RANDOM] tr ##ee nearby . there is! [MASK] ap ##p ##le tr ##ee nearby [MASK] . there is [MASK] ap ##p ##le [MASK] [MASK] nearby . there [MASK] [MASK] ap ##p ##le tr ##ee [RANDOM] . there is an ap ##p ##le [MASK] [MASK] nearby [MASK] .

Chinese BERT-wwm

- **Chinese BERT with Whole Word Masking**
  - For further accelerating Chinese natural language processing, we provide Chinese pre-trained BERT with Whole Word Masking.
  - We also compare the state-of-the-art Chinese pre-trained models in depth, including BERT, ERNIE, BERT-wwm
  - https://github.com/ymcui/Chinese-BERT-wwm

Pre-Training with Whole Word Masking for Chinese BERT

Yiming Cui[†‡*], Wanxiang Che[†], Ting Liu[†], Bing Qin[†], Ziqing Yang[‡], Shijin Wang[‡], Guoping Hu[‡]
[†]Research Center for Social Computing and Information Retrieval (SCIR), Harbin Institute of Technology, Harbin, China
[‡]Joint Laboratory of HIT and iFLYTEK (HFL), iFLYTEK Research, Beijing, China
[‡]iFLYTEK Hebei AI Research, Hebei, China

Chinese BERT-wwm

- **Chinese BERT with Whole Word Masking**
  - BERT-wwm is similar to ERNIE but different in the following aspects.
    - BERT-wwm is trained on Chinese Wikipedia ONLY.
    - BERT-wwm does not exploit entity-masking or phrase-masking
  - **Example**

[Original Sentence]
使用语言模型来预测下一个词的probability。
[Original Sentence with CWS]
使用 语言 **模型** 来 **预测** 下 一个 词 的 **probability** 。

[Original BERT Input]
使 用 语 言 **[MASK] 型** 来 **[MASK] 测** 下 一 个 词 的 **pro [MASK] ##lity** 。
[Whold Word Masking Input]
使 用 语 言 **[MASK] [MASK]** 来 **[MASK] [MASK]** 下 一 个 词 的 **[MASK] [MASK] [MASK]** 。

Remember: [MASK] could also be 'replace by another word' or 'keep original word'

Cui et al., arXiv pre-print. Pre-Training with Whole Word Masking for Chinese BERT

Chinese BERT-wwm

- **Comparisons of BERT, ERNIE, BERT-wwm**

	BERT	BERT-wwm	ERNIE
Pre-Train Data	Wikipedia	Wikipedia	Wikipedia +Baike+Tieba, etc.
Sentence #	24M		173M
Vocabulary #	21,128		18,000 (17,964)
Hidden Activation	GeLU		ReLU
Hidden Size/Layers	768 & 12		
Attention Head #	12		

Cui et al., arXiv pre-print. Pre-Training with Whole Word Masking for Chinese BERT

Chinese BERT-wwm

- **Experiments**
  - We tested BERT, ERNIE, BERT-wwm on various Chinese datasets covering a wide spectrum of text length (from sentence-level to document-level)

Dataset	Task	MaxLen	Batch	Epoch	Train #	Dev #	Test #	Domain
CMRC 2018	MRC	512	64	2	10K	3.2K	4.9K	Wikipedia
DRCD	MRC	512	64	2	27K	3.5K	3.5K	Wikipedia
CJRC	MRC	512	64	2	10K	3.2K	3.2K	law
People Daily	NER	256	64	3	51K	4.6K	-	news
MSRA-NER‡	NER	256	64	5	45K	-	3.4K	news
XNLI†‡	NLI	128	64	2	392K	2.5K	2.5K	various
ChnSentiCorp‡	SC	256	64	3	9.6K	1.2K	1.2K	various
Sina Weibo	SC	128	64	3	100K	10K	10K	microblogs
LCQMC‡	SPM	128	64	3	240K	8.8K	12.5K	Zhidao
BQ Corpus	SPM	128	64	3	100K	10K	10K	QA
THUCNews	DC	512	64	3	50K	5K	10K	news

† means the data was also tested in the original paper of BERT
‡ means the data was also tested in the original paper of ERNIE

Cui et al., arXiv pre-print. Pre-Training with Whole Word Masking for Chinese BERT

Chinese BERT-wwm

- **Experiments: MRC**
  - BERT-wwm yields significant improvements on CMRC 2018 (Simplified Chinese) and DRCD (Traditional Chinese)
  - ERNIE does not show competitive performance, especially on Traditional Chinese data

CMRC 2018	Dev		Test		Challenge	
	EM	F1	EM	F1	EM	F1
BERT	65.5 (64.4)	84.5 (84.0)	70.0 (68.7)	87.0 (86.3)	18.6 (17.0)	43.3 (41.3)
ERNIE	65.4 (64.3)	84.7 (84.2)	69.4 (68.2)	86.6 (86.1)	19.6 (17.0)	44.3 (42.8)
BERT-wwm	**66.3 (65.0)**	**85.6 (84.7)**	**70.5** (69.1)	**87.4 (86.7)**	**21.0 (19.3)**	**47.0 (43.9)**

Table 3: Results on CMRC 2018 (Simplified Chinese). The average score of 10 independent runs is depicted in brackets. Best LR: BERT (3e-5), BERT-wwm (3e-5), ERNIE (8e-5).

DRCD	Dev		Test	
	EM	F1	EM	F1
BERT	83.1 (82.7)	89.9 (89.6)	82.2 (81.6)	89.2 (88.8)
ERNIE	73.2 (73.0)	83.9 (83.8)	71.9 (71.4)	82.5 (82.3)
BERT-wwm	**84.3 (83.4)**	**90.5 (90.2)**	**82.8 (81.8)**	**89.7 (89.0)**

Table 4: Results on DRCD (Traditional Chinese). Best LR: BERT (3e-5), BERT-wwm (3e-5), ERNIE (8e-5).

Cui et al., arXiv pre-print. Pre-Training with Whole Word Masking for Chinese BERT

Chinese BERT-wwm

- **Experiments: MRC**
  - BERT-wwm only shows moderate improvements over BERT
  - CJRC is composed of the text regarding Chinese laws, which is written in professional ways, which is not friendly to the models in general domains
  - Further fine-tuning should be done on the dataset that is far different from pre-training data

CJRC	Dev		Test	
	EM	F1	EM	F1
BERT	54.6 (**54.0**)	**75.4** (74.5)	**55.1** (**54.1**)	75.2 (74.3)
ERNIE	54.3 (53.9)	75.3 (74.6)	55.0 (53.9)	75.0 (73.9)
BERT-wwm	**54.7** (**54.0**)	75.2 (**74.8**)	**55.1** (**54.1**)	**75.4** (**74.4**)

Table 5: Results on CJRC. Best LR: BERT (4e-5), BERT-wwm (4e-5), ERNIE (8e-5).

Cui et al., arXiv pre-print. Pre-Training with Whole Word Masking for Chinese BERT

Chinese BERT-wwm

- **Experiments: NER**
  - ERNIE has a good performance on NER data, especially for peak performance, while BERT-wwm shows better average performance on these data
  - During training, we encountered training failure in ERNIE over half of ten independent runs, where the results are significantly lower than the average score (say lower than 90). We eliminate these results to ensure fair comparisons

NER	People Daily			MSRA-NER		
	P	R	F	P	R	F
BERT	95.3 (95.0)	95.1 (94.8)	95.2 (94.9)	**95.4** (94.8)	95.3 (95.0)	95.3 (94.9)
ERNIE	**95.8** (94.7)	**95.6** (94.3)	**95.7** (94.5)	95.3 (94.9)	**95.7** (**95.4**)	**95.4** (**95.1**)
BERT-wwm	95.4 (**95.1**)	95.3 (**95.0**)	95.3 (**95.1**)	**95.4** (**95.1**)	95.6 (95.3)	**95.4** (**95.1**)

Table 6: Results on People Daily and MSRA-NER. Best LR for PD: BERT (3e-5), BERT-wwm (3e-5), ERNIE (5e-5). Best LR for MSRA-NER: BERT (3e-5), BERT-wwm (4e-5), ERNIE (5e-5).

Cui et al., arXiv pre-print. Pre-Training with Whole Word Masking for Chinese BERT

Chinese BERT-wwm

- **Experiments: NLI**
  - ERNIE shows the best performance on natural language inference task, compared to BERT and BERT-wwm.

XNLI	Dev	Test
BERT	77.8 (77.4)	77.8 (77.5)
ERNIE	**79.7 (79.4)**	**78.6 (78.2)**
BERT-wwm	79.0 (78.4)	78.2 (78.0)

Table 7: Results on XNLI. Best LR: BERT (3e-5), BERT-wwm (3e-5), ERNIE (5e-5).

Cui et al., arXiv pre-print. Pre-Training with Whole Word Masking for Chinese BERT

Chinese BERT-wwm

- **Experiments: Sentiment Classification (binary)**
  - ERNIE achieves the best performance on ChnSentiCorp
  - Both BERT-wwm and ERNIE show better performance on Weibo data
  - As ERNIE was trained on additional web text, it is beneficial to use ERNIE to process the task in a similar domain

Sentiment Classification	ChnSentiCorp Dev	ChnSentiCorp Test	Sina Weibo (100k) Dev	Sina Weibo (100k) Test
BERT	94.7 (94.3)	95.0 (94.7)	97.49 (97.38)	**97.37** (97.32)
ERNIE	**95.4 (94.8)**	**95.4 (95.3)**	**97.54 (97.41)**	**97.37** (97.29)
BERT-wwm	95.1 (94.5)	**95.4** (95.0)	97.49 (97.40)	**97.37 (97.35)**

Table 8: Results on ChnSentiCorp and Sina Weibo. Best LR for ChnSentiCorp: BERT (2e-5), BERT-wwm (2e-5), ERNIE (5e-5). Best LR for Sina Weibo: BERT (2e-5), BERT-wwm (3e-5), ERNIE (3e-5).

Cui et al., arXiv pre-print. Pre-Training with Whole Word Masking for Chinese BERT

Chinese BERT-wwm

- **Experiments: Sentence Pair Matching**
  - ERNIE shows better performance on LCQMC data
  - While, when it comes to BQ Corpus, BERT-wwm generally outperform ERNIE and BERT, especially the averaged scores

Sentence Pair Matching	**LCQMC** Dev	**LCQMC** Test	**BQ Corpus** Dev	**BQ Corpus** Test
BERT	89.4 (88.4)	86.9 (86.4)	86.0 (85.5)	84.8 (84.6)
ERNIE	**89.8 (89.6)**	**87.2 (87.0)**	**86.3** (85.5)	85.0 (84.6)
BERT-wwm	89.4 (89.2)	87.0 (86.8)	86.1 **(85.6)**	**85.2 (84.9)**

Table 9: Results on LCQMC and BQ Corpus. Best LR for LCQMC: BERT (2e-5), BERT-wwm (2e-5), ERNIE (3e-5). Best LR for BQ Corpus: BERT (3e-5), BERT-wwm (3e-5), ERNIE (5e-5).

Cui et al., arXiv pre-print. Pre-Training with Whole Word Masking for Chinese BERT

Chinese BERT-wwm

- **Experiments: Document Classification**
  - BERT-wwm and BERT generally outperform ERNIE again on long sequence modeling tasks

THUCNews	Dev	Test
BERT	97.7 (97.4)	**97.8 (97.6)**
ERNIE	97.6 (97.3)	97.5 (97.3)
BERT-wwm	**98.0 (97.6)**	**97.8 (97.6)**

Table 10: Results on THUCNews. Best learning rate: BERT (2e-5), BERT-wwm (2e-5), ERNIE (5e-5).

Cui et al., arXiv pre-print. Pre-Training with Whole Word Masking for Chinese BERT

Chinese BERT-wwm

- **Useful Tips**
  - Initial learning rate is the most important hyper-parameters (regardless of BERT or other neural networks), and should ALWAYS be tuned for better performance.
  - BERT and BERT-wwm share almost the same best initial learning rate, so it is straightforward to apply your initial learning rate in BERT to BERT-wwm.
  - However, we find that ERNIE does not share the same characteristics, so it is STRONGLY recommended to tune the learning rate.

Cui et al., arXiv pre-print. Pre-Training with Whole Word Masking for Chinese BERT

Chinese BERT-wwm

- **Useful Tips**
  - As BERT and BERT-wwm were trained on Wikipedia data, they show relatively better performance on the formal text. While, ERNIE was trained on larger data, including web text, which will be useful on casual text, such as Weibo (microblogs).
  - In long-sequence tasks, such as machine reading comprehension and document classification, we suggest using BERT or BERT-wwm.
  - If the task data is extremely different from the pre-training data (Wikipedia for BERT/BERT-wwm), we suggest taking another pre-training steps on the task data, which was also suggested by Devlin et al. (2019).

Cui et al., arXiv pre-print. Pre-Training with Whole Word Masking for Chinese BERT

Chinese BERT-wwm

- **Useful Tips**
  - When dealing with Traditional Chinese text, use BERT or BERT-wwm.
  - As there are so many possibilities in the pre-training stage (such as initial learning rate, global training steps, warm-up steps, etc.), our implementation may not be optimal using the same pre-training data. Readers are advised to train their own model if seeking for another boost in performance. However, if it is unable to do pre-training, choose one of these pre-trained models which were trained on a similar domain to the down-stream task.

Cui et al., arXiv pre-print. Pre-Training with Whole Word Masking for Chinese BERT

Episode 1: Personal (Shallow) Advice for Beginners

Advice for Beginners

- **Begin with pre-trained models, then dive into specific MRC task**
  - You have to admit that pre-trained models become new basic skills for NLP, just like word segmentation/parsing in ‘traditional NLP’
  - You may excuse for not using pre-trained models in your scientific paper, but the reviewer will always ask “why not try/compare your method on BERT?” (at least from my experience)

Advice for Beginners

- **MRC is not <u>ONLY</u> about neural network models, there are many things to do**
  - Data: create much more challenging data for MRC
  - Approach: design more sophisticated models
  - Cross-task: apply MRC to other NLP tasks
  - Multi-lingual: solve MRC other than English
  - Evaluation: does machine really comprehend human language?
  - …
  - Open your mind, embrace new coming techniques

Episode 2: Useful Resources

Useful Resources

- **《机器阅读理解任务综述》**
  - 林鸿宇、韩先培（中科院软件所）
  - http://www.cipsc.org.cn/qngw/?p=930
- **Must-read papers on Machine Reading Comprehension**
  - Yankai Li, Deming Ye, Haozhe Ji
  - https://github.com/thunlp/RCPapers
- **Tracking Progress in Natural Language Processing**
  - Sebastian Ruder
  - https://github.com/sebastianruder/NLP-progress

Useful Resources

- **Neural Machine Reading Comprehension: Methods and Trends**
  - Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang ,Weiming Zhang
  - National University of Defense Technology (NUDT)
  - https://arxiv.org/abs/1907.01118
- **Machine Reading Comprehension: a Literature Review**
  - Xin Zhang, An Yang, Sujian Li, Yizhong Wang
  - Peking University (PKU)
  - https://arxiv.org/abs/1907.01686

Useful Resources

- **Domestic MRC Competitions**
  - **The First Evaluation Workshop on Chinese Machine Reading Comprehension (CMRC 2017)**
    - Host: CIPS-CL, Joint Laboratory of HIT and iFLYTEK Research (HFL), iFLYTEK Co. Ltd
    - Competition Type: Cloze-style RC, User Query RC
    - http://cmrc2017.hfl-rc.com
  - **The Second Evaluation Workshop on Chinese Machine Reading Comprehension (CMRC 2018)**
    - Host: CIPS-CL, Joint Laboratory of HIT and iFLYTEK Research (HFL), iFLYTEK Co. Ltd
    - Competition Type: Span-Extraction RC
    - http://cmrc2018.hfl-rc.com

Useful Resources

- **Domestic MRC Competitions**
  - **2018 NLP Challenge on Machine Reading Comprehension**
    - Host: CCF, CIPSC, Baidu Inc.
    - Competition Type: Open-Domain RC
    - http://mrc2018.cipsc.org.cn
  - **CIPS-SOGOU QA Competition**
    - Host: CIPSC, SOGOU
    - Competition Type: Factoid QA, Non-Factoid QA
    - http://task.www.sogou.com/cips-sogou_qa/
  - **2019 NLP Language and Intelligence Challenge**
    - Host: CCF, CIPSC, Baidu Inc.
    - Competition Type: Open-Domain RC
    - http://lic2019.ccf.org.cn

Useful Resources

- **Domestic MRC Competitions**
  - **The Third Evaluation Workshop on Chinese Machine Reading Comprehension (CMRC 2019)**
    - Host: CIPS-CL, Joint Laboratory of HIT and iFLYTEK Research (HFL), iFLYTEK Co. Ltd
    - Competition Type: Sentence Cloze
    - http://cmrc2019.hfl-rc.com
  - **Chinese AI Law Competitions 2019**
    - Competition Type: Law-related MRC, etc.
    - http://cail.cipsc.org.cn/

Useful Resources

- **Personal Repository (https://github.com/ymcui/)**

Name	Description	Genre	Stars
Chinese-BERT-wwm	Pre-trained Chinese BERT with Whole Word Masking	Data	650+
Chinese-Cloze-RC	A Chinese Cloze-style RC Dataset: People Daily & Children's Fairy Tale (CFT)	Data	111
Eval-on-NN-of-RC	Empirical Evaluation on Current Neural Networks on Cloze-style Reading Comprehension	Text	82
Chinese-RC-Datasets	Collections of Chinese reading comprehension datasets	Data	24
LAMB_Optimizer_TF	LAMB Optimizer for larger batch	Code	20
CMRC2018-DRCD-BERT	BERT baselines for CMRC 2018 & DRCD (Chinese reading comprehension datasets)	Code	18
cmrc2017	The First Evaluation Workshop on Chinese Machine Reading Comprehension (CMRC 2017)	Data	70
cmrc2018	The Second Evaluation Workshop on Chinese Machine Reading Comprehension (CMRC 2018)	Data	52
cmrc2019	The Third Evaluation Workshop on Chinese Machine Reading Comprehension (CMRC 2019)	Data	47

Thank You !

哈工大讯飞联合实验室
微信公众号

Personal Website

E-mail: me@ymcui.com