北京航空航天大學
BEIHANG UNIVERSITY

自然语言处理

人工智能研究院

主讲教师 沙磊

问答系统

Contents

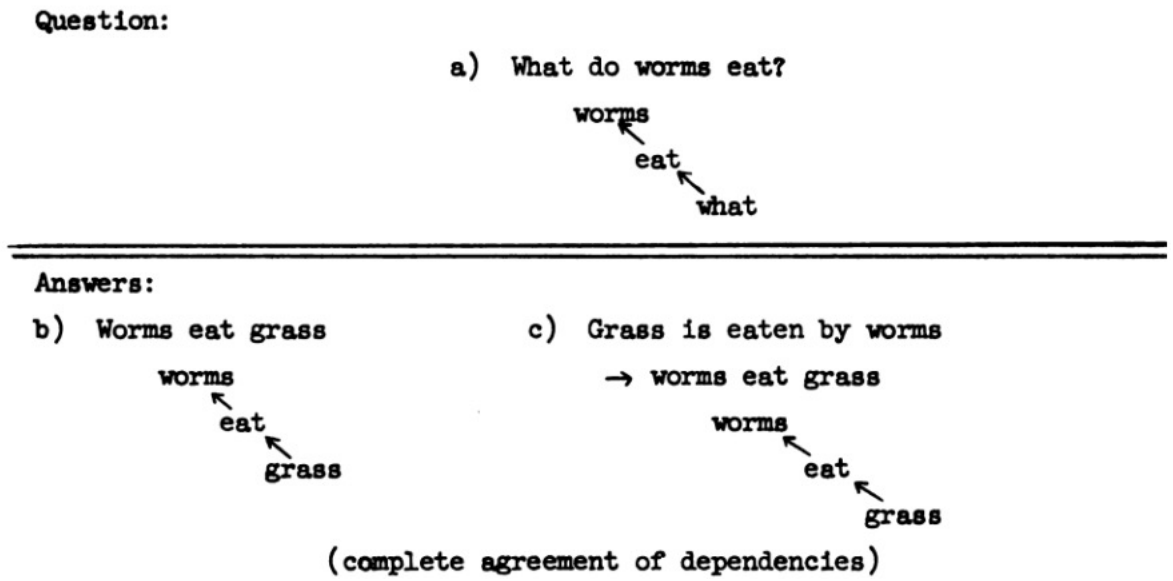
- 今天内容
  - 介绍新的NLP任务：问答系统
  - 机器阅读理解
    - 如何根据一篇单独的文章来回答问题
  - 开域问答Open-domain Question Answering
    - 如何根据大量的文本来回答问题

What is question answering?

- 问答系统目的是自动回答人类提出的自然语言形式的问题

The earliest QA systems dated back to 1960s!

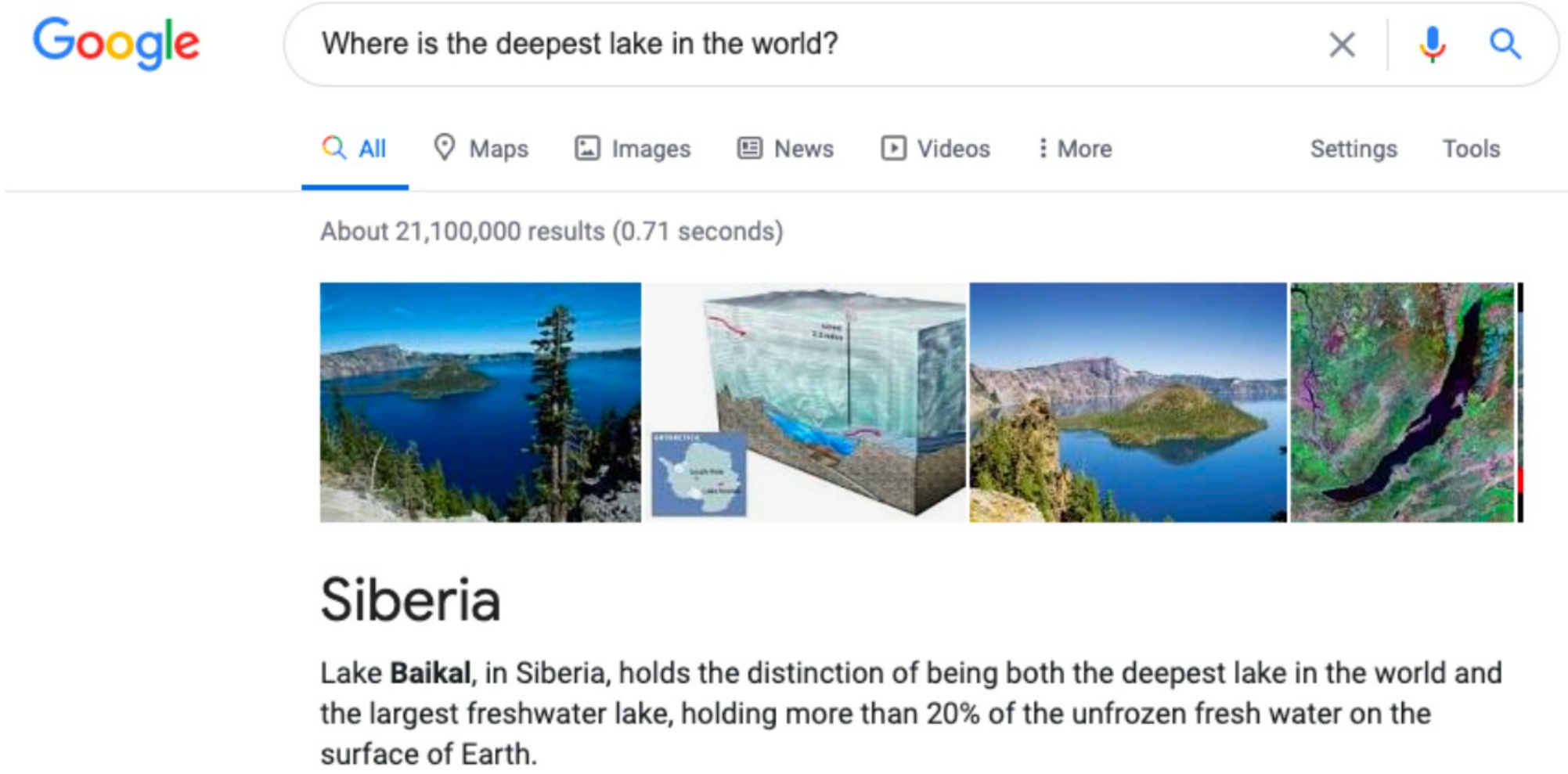
(Simmons et al., 1964)



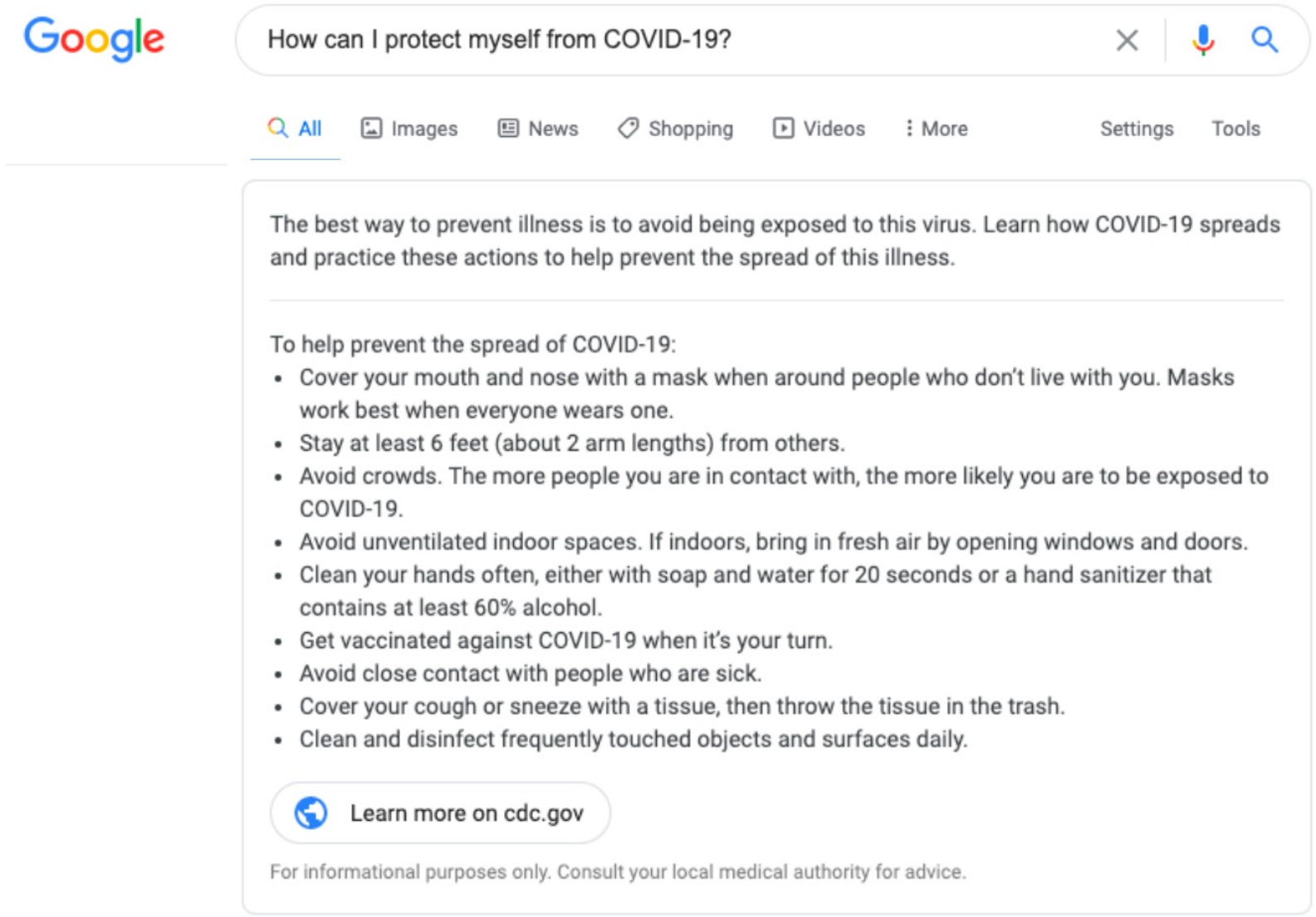
Question answering: a taxonomy

- 系统建立在哪种信息源上
  - 单篇文档，网络文本，知识库，表格，图像。。。
- 问题类型
  - 实施性/非事实性，开域/闭域，简单性/复合型
- 答案类型
  - 文档中的一个短的片段，一段文本，一个列表，yes/no

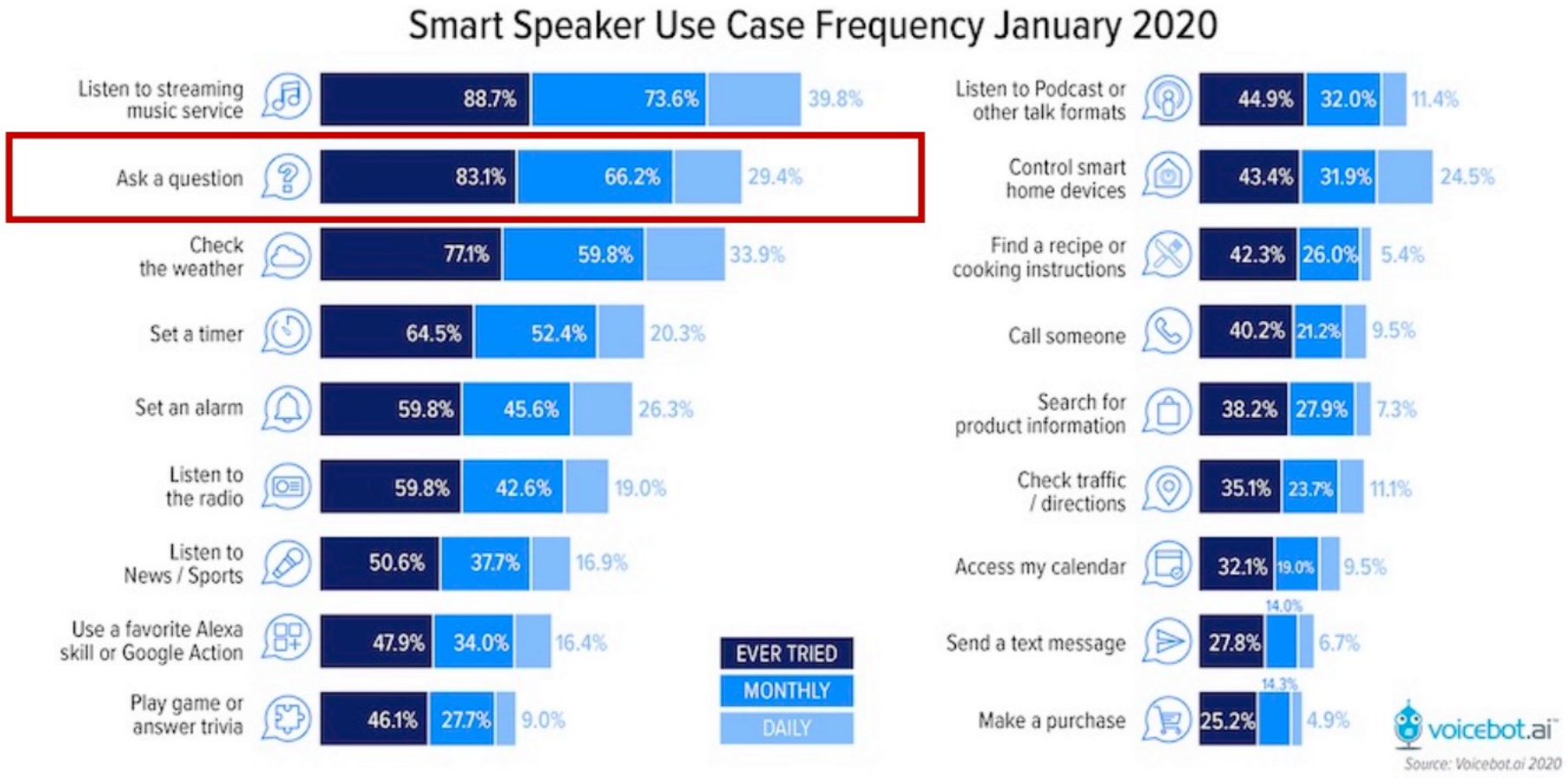
Lots of practical applications



Lots of practical applications



Lots of practical applications



2011: IBM Watson beat Jeopardy champions

IBM Watson defeated two of Jeopardy's greatest champions in 2011

IBM Watson beat Jeopardy champions

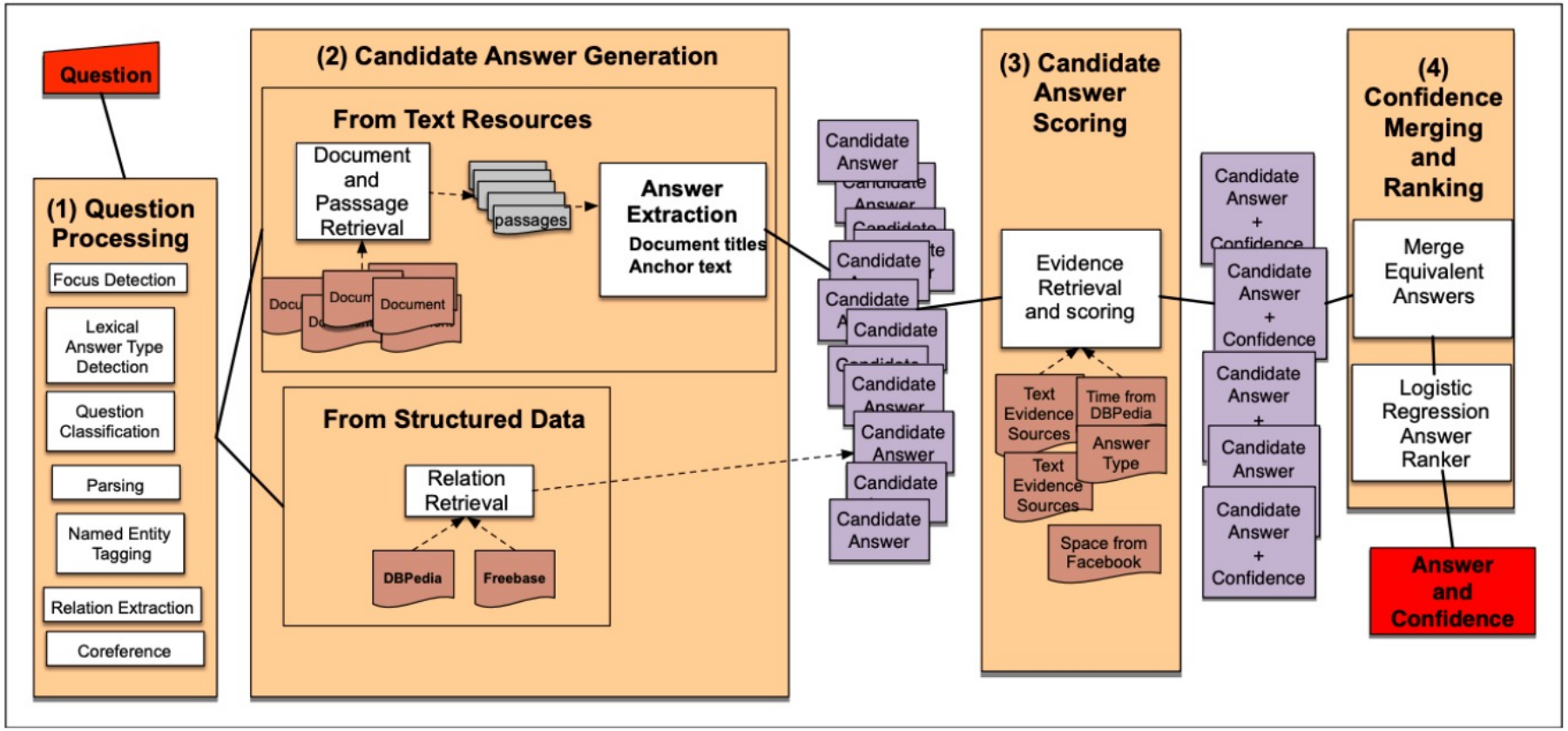


Image credit: J & M, edition 3

(1) Question processing, (2) Candidate answer generation, (3) Candidate answer scoring, and (4) Confidence merging and ranking.

Question answering in deep learning era

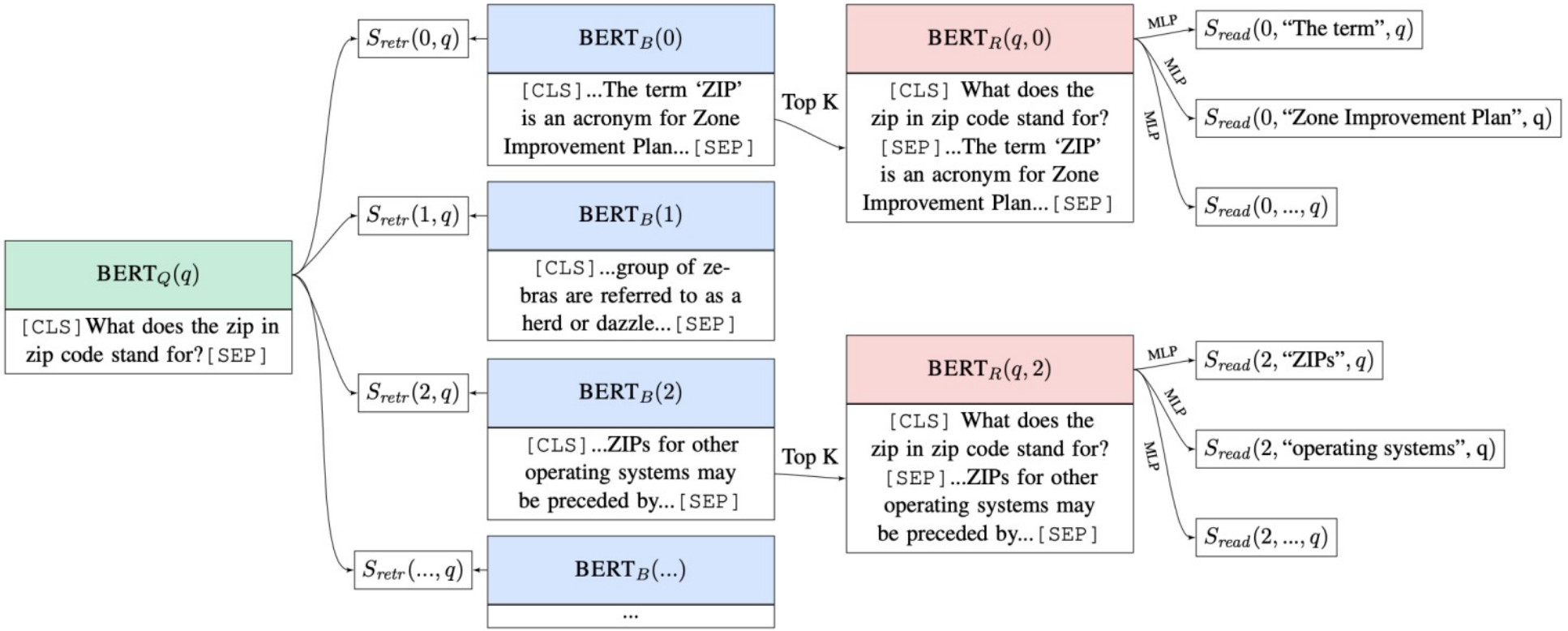


Image credit: (Lee et al., 2019)

Almost all the state-of-the-art question answering systems are built on top of end-to-end training and pre-trained language models (e.g., BERT)!

Beyond textual QA problems

- 今天，我们主要介绍无结构文本上的问答系统

Knowledge based QA

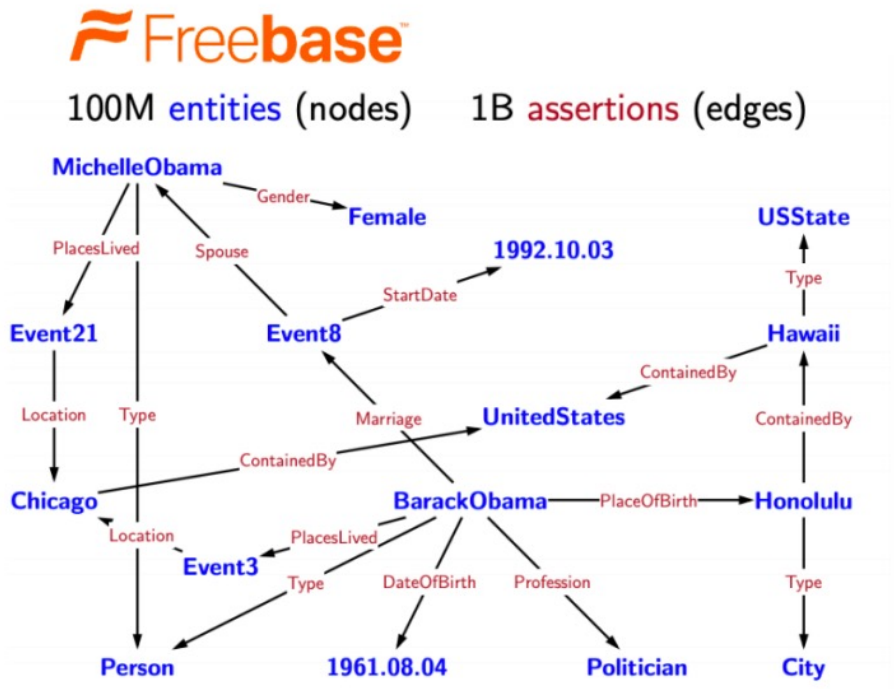


Image credit: Percy Liang

Beyond textual QA problems

Visual QA

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

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

(Antol et al., 2015): Visual Question Answering

Reading comprehension

- **Reading comprehension** = comprehend a passage of text and answer questions about its content (P, Q) ⟶ A

Tesla was the fourth of five children. He had an older brother named Dane and three sisters, Milka, Angelina and Marica. Dane was killed in a horse-riding accident when Nikola was five. In 1861, Tesla attended the "Lower" or "Primary" School in Smiljan where he studied German, arithmetic, and religion. In 1862, the Tesla family moved to Gospić, Austrian Empire, where Tesla's father worked as a pastor. Nikola completed "Lower" or "Primary" School, followed by the "Lower Real Gymnasium" or "Normal School."

Q: What language did Tesla study while in school?

A: German

Reading comprehension

- **Reading comprehension** = comprehend a passage of text and answer questions about its content (P, Q) ⟶ A

Kannada language is the official language of Karnataka and spoken as a native language by about 66.54% of the people as of 2011. Other linguistic minorities in the state were Urdu (10.83%), Telugu language (5.84%), Tamil language (3.45%), Marathi language (3.38%), Hindi (3.3%), Tulu language (2.61%), Konkani language (1.29%), Malayalam (1.27%) and Kodava Takk (0.18%). In 2007 the state had a birth rate of 2.2%, a death rate of 0.7%, an infant mortality rate of 5.5% and a maternal mortality rate of 0.2%. The total fertility rate was 2.2.

Q: Which linguistic minority is larger, Hindi or Malayalam?

A: Hindi

Why do we care about this problem?

- 对很多实际的应用均有用
- 可以作为一个评价计算机对于人类语言理解程度的测试任务
  - Wendy Lehnert 1977: “由于问题可以被设计为对文本理解的任何方面提出疑问，因此回答问题的能力是理解的最有力证明.”
- 许多其他的NLP任务都可以被归结为阅读理解任务:

Information extraction

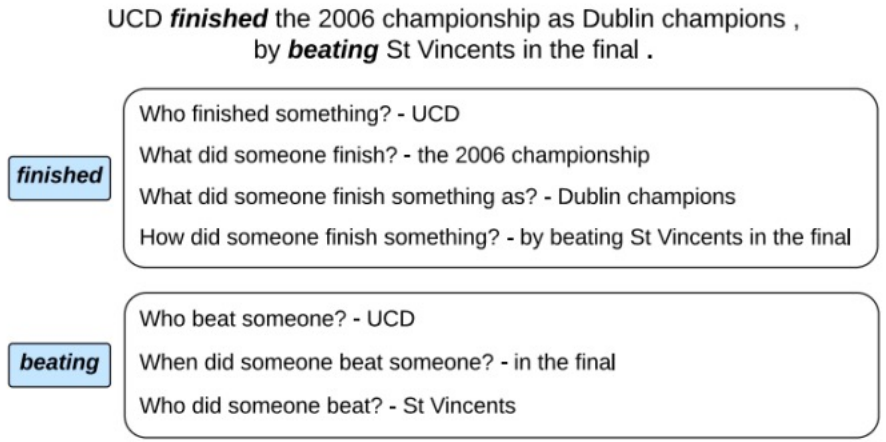
(Barack Obama, educated_at, ?)

Question: Where did Barack Obama graduate from?

Passage: Obama was born in Honolulu, Hawaii. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago.

(Levy et al., 2017)

Semantic role labeling



(He et al., 2015)

Stanford question answering dataset (SQuAD)

- 100k annotated (passage, question, answer) triples
  - 大规模有监督的数据集也是训练阅读理解有效神经模型的关键因素！
- Passages are selected from English Wikipedia, usually 100~150 words.
- Questions are crowd-sourced.
- Each answer is an short segment of text (or span) in the passage.
  - 这是一个局限的点，事实上不是所有问题都能这么回答的
- SQuAD至今都是机器阅读任务最广泛认可的dataset，目前几乎已经解决，并超越人类水平

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

Stanford question answering dataset (SQuAD)

- 评价方法： 精确匹配(0 or 1) and F1 (partial credit).
- 对于开发集和测试集，每个问题给出三个候选答案
- 我们将预测答案与每个候选答案进行比较（删除a、an、the等），并取最大分数。最后，我们取所有case的精确匹配和F1的平均值。
- 人类水平： EM = 82.3, F1 = 91.2

Q: What did Tesla do in December 1878?

A: {left Graz, left Graz, left Graz and severed all relations with his family}

Prediction: {left Graz and served}

Exact match: max{0, 0, 0} = 0

F1: max{0.67, 0.67, 0.61} = 0.67

Other question answering datasets

- TriviaQA：琐事爱好者的问答。单独收集似乎讨论了问题并且包含答案的网页段落，但没有人工去验证
- Natural Questions：来自谷歌搜索的常见问题。答案从维基百科段落中收集得到。答案可以是substring, yes, no, 或NOT_PRESENT。经人类标注验证。
- HotpotQA。从整个维基百科构建要回答的问题。本任务涉及从两个页面获取信息去回答一个多步骤的提问：
  - Q：《无敌舰队》作者的哪部小说将被史蒂文·斯皮尔伯格改编成电影
  - A： Ready Player One

Neural models for reading comprehension

- 如何构建一个模型去解决SQuAD?
- 问题形式化：
  - 输入：$C = (c_1, c_2, \ldots, c_N), Q = (q_1, q_2, \ldots, q_M), c_i, q_i \in V$　　N~100，M~15
  - 输出：$1 \leq \text{start} \leq \text{end} \leq N$　　答案是文中的某片段
- LSTM+Attention(2016-2018)
  - Attentive Reader (Hermann et al., 2015), Stanford Attentive Reader (Chen et al., 2016), MatchLSTM (Wang et al., 2017), BiDAF (Seo et al., 2017), Dynamic coattention network (Xiong et al., 2017), DrQA (Chen et al., 2017), R-Net (Wang et al., 2017), ReasoNet (Shen et al., 2017).
- Fine-tuning BERT-like models for reading comprehension (2019+)

LSTM-based vs BERT models

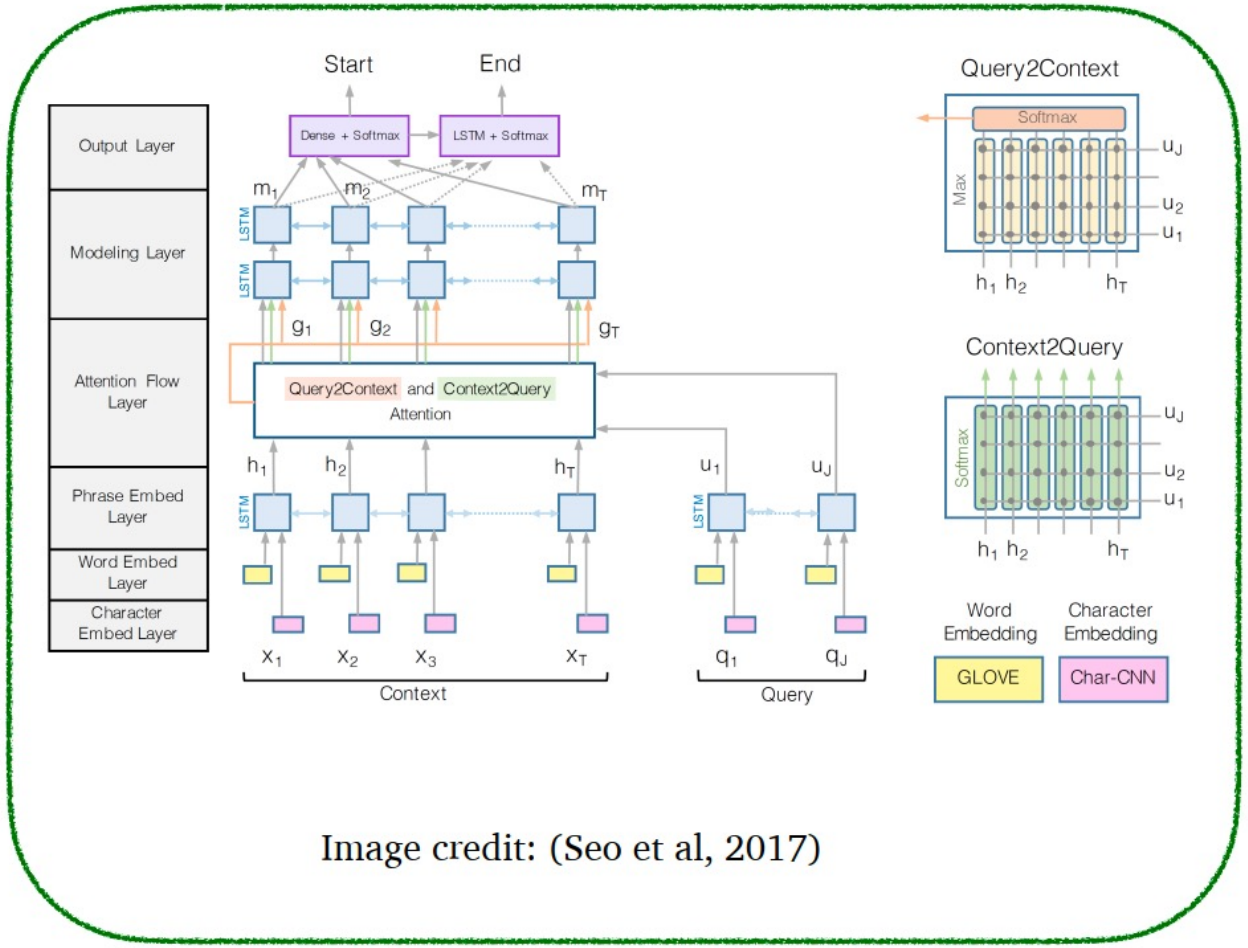


Image credit: (Seo et al, 2017)

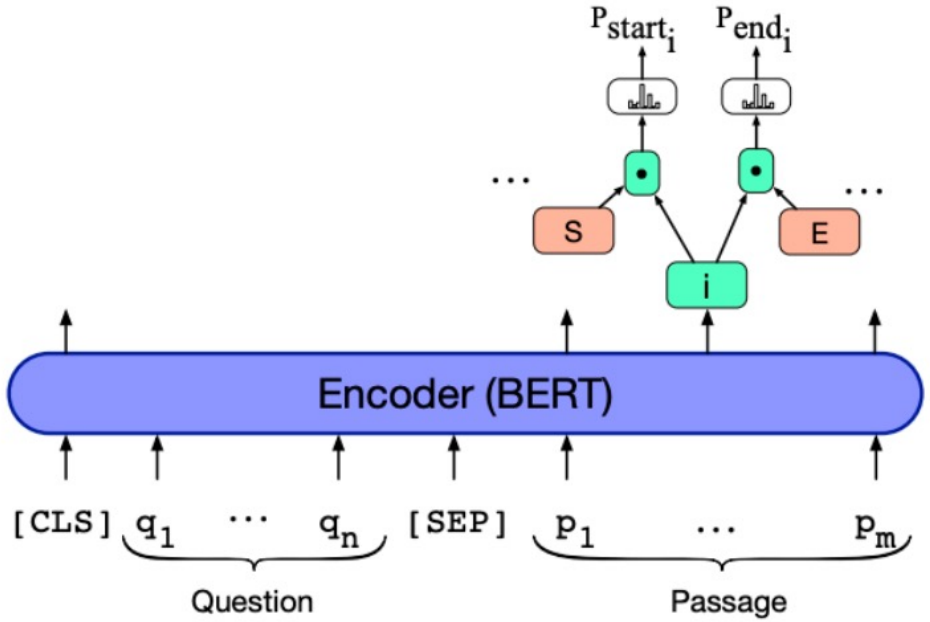
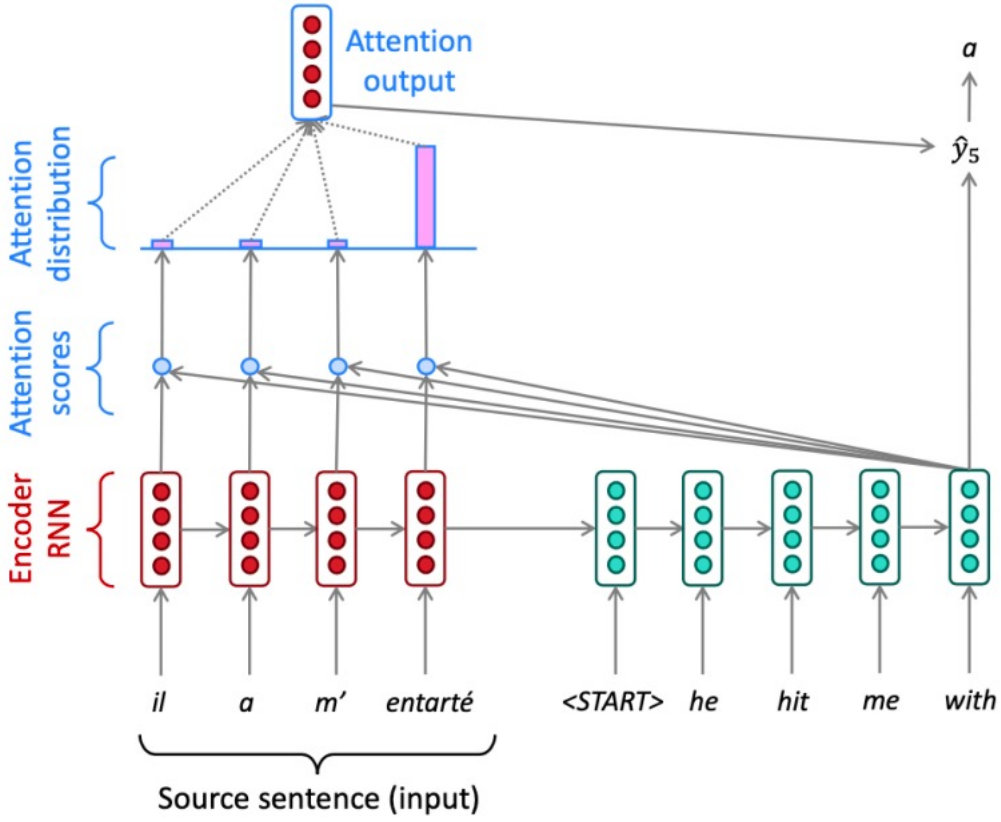


Image credit: J & M, edition 3

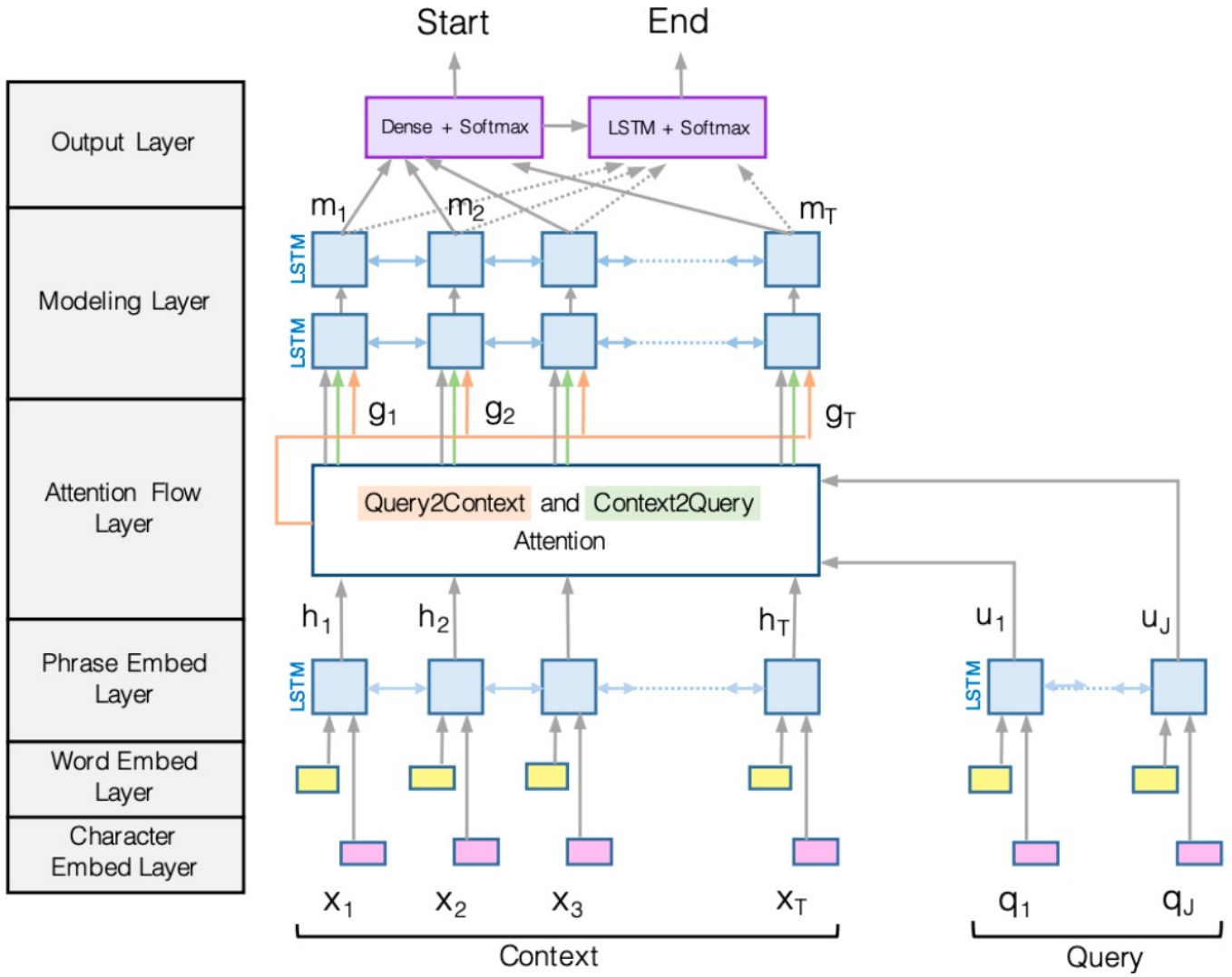
Recap: seq2seq model with attention

- 此处的seq2seq我们处理的不是原句和目标句，而是文章序列和问题序列（长度可能差距很大）
- 我们需要针对文章中的哪些词与问题（问题中的哪些词）最相关进行建模
  - 注意力机制Attention在这里很关键
- 此处我们不需要自回归的解码器去生产目标句的每一个词，只需要用两个分类器找到答案序列的开始和结束的位置

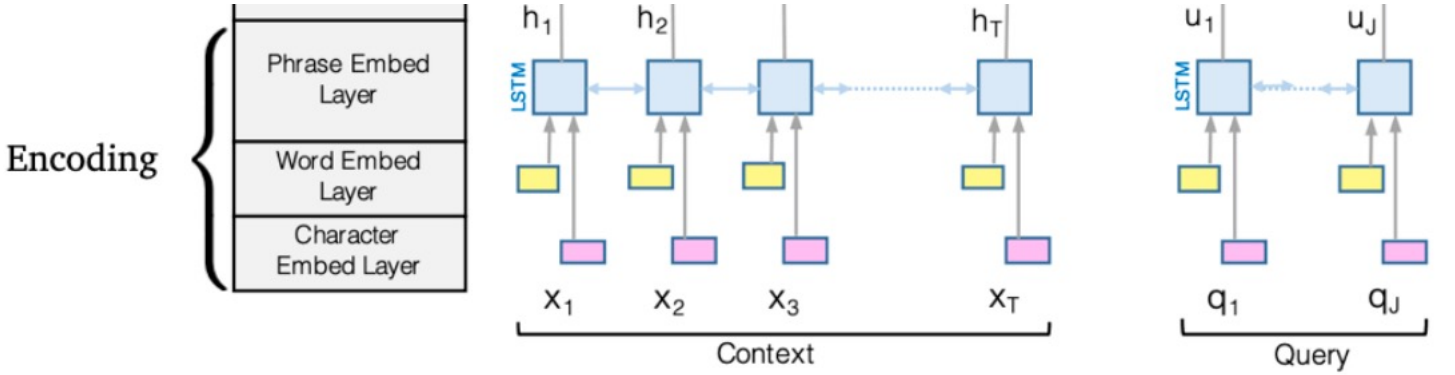


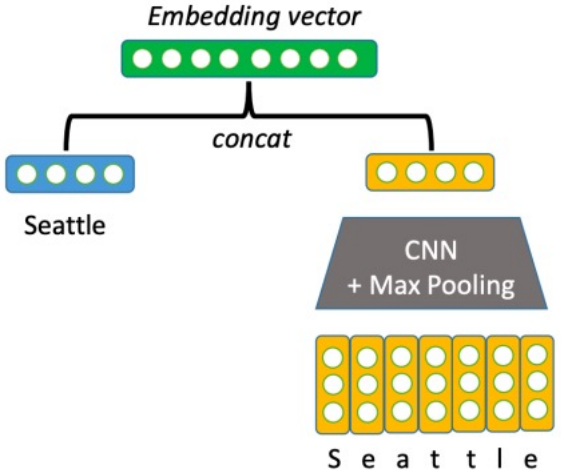
BiDAF: the Bidirectional Attention Flow model

(Seo et al., 2017): Bidirectional Attention Flow for Machine Comprehension



BiDAF: Encoding





- Use a concatenation of word embedding (GloVe) and character embedding (CNNs over character embeddings) for each word in context and query. f is high-way network

$$e(c_i) = f([\text{GloVe}(c_i); \text{charEmb}(c_i)]) \qquad e(q_i) = f([\text{GloVe}(q_i); \text{charEmb}(q_i)])$$

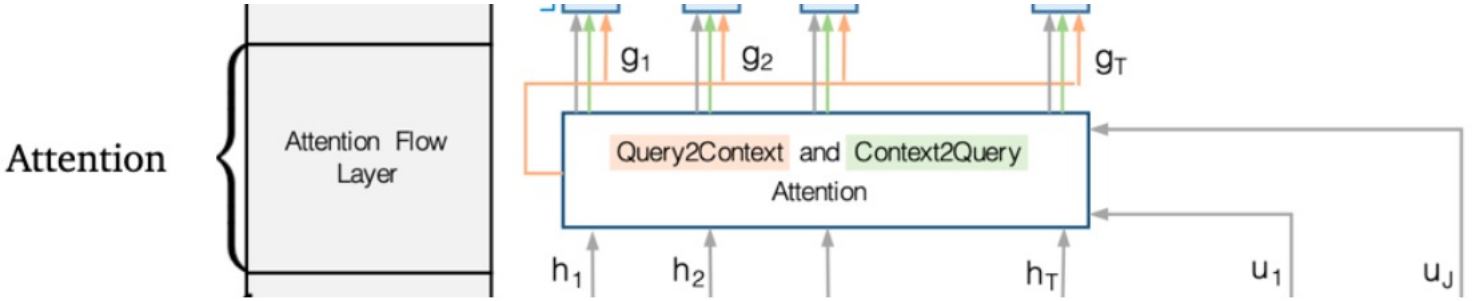
- Then, use two bidirectional LSTMs separately to produce contextual embeddings for both context and query

$$\overrightarrow{\mathbf{c}}_i = \text{LSTM}(\overrightarrow{\mathbf{c}}_{i-1}, e(c_i)) \in \mathbb{R}^H \qquad \overrightarrow{\mathbf{q}}_i = \text{LSTM}(\overrightarrow{\mathbf{q}}_{i-1}, e(q_i)) \in \mathbb{R}^H$$

$$\overleftarrow{\mathbf{c}}_i = \text{LSTM}(\overleftarrow{\mathbf{c}}_{i+1}, e(c_i)) \in \mathbb{R}^H \qquad \overleftarrow{\mathbf{q}}_i = \text{LSTM}(\overleftarrow{\mathbf{q}}_{i+1}, e(q_i)) \in \mathbb{R}^H$$

$$\mathbf{c}_i = [\overrightarrow{\mathbf{c}}_i; \overleftarrow{\mathbf{c}}_i] \in \mathbb{R}^{2H} \qquad \mathbf{q}_i = [\overrightarrow{\mathbf{q}}_i; \overleftarrow{\mathbf{q}}_i] \in \mathbb{R}^{2H}$$

BiDAF: Attention



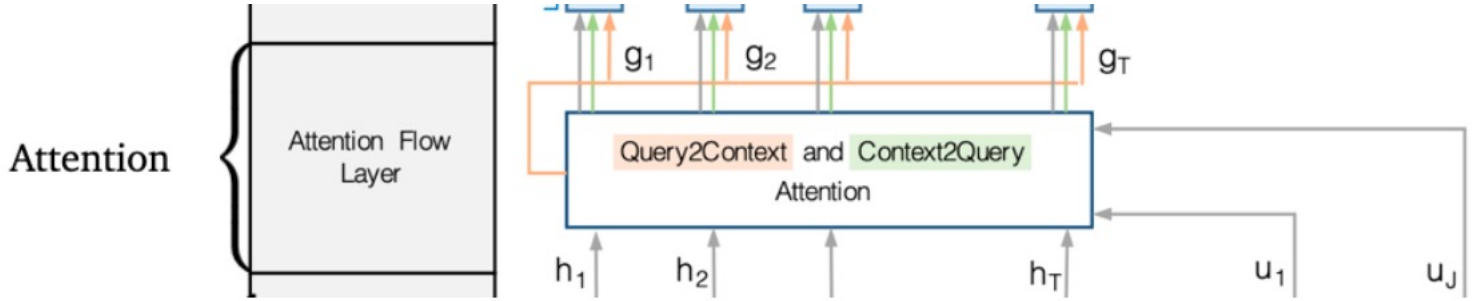
- Context-to-query attention: For each context word, choose the most relevant words from the query words.

Q: *Who leads the United States?*

C: *Barak Obama is the president of the USA.*

For each context word, find the most relevant query word.

BiDAF: Attention



- Query-to-context attention: choose the context words that are most relevant to one of query words.

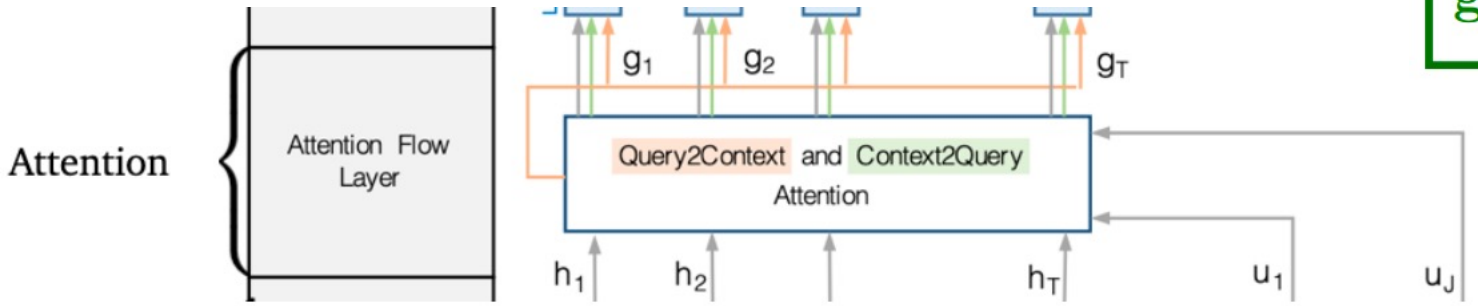
While Seattle's weather is very nice in summer, its weather is very rainy in winter, making it one of the most gloomy cities in the U.S. LA is …

Q: Which city is gloomy in winter?

BiDAF: Attention

The final output is

$$\mathbf{g}_i = [\mathbf{c}_i; \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{b}] \in \mathbb{R}^{8H}$$



- First, compute a similarity score for every pair of $(\mathbf{c}_i, \mathbf{q}_j)$

$$S_{i,j} = \mathbf{w}_{\text{sim}}^{\intercal}[\mathbf{c}_i; \mathbf{q}_j; \mathbf{c}_i \odot \mathbf{q}_j] \in \mathbb{R} \qquad \mathbf{w}_{\text{sim}} \in \mathbb{R}^{6H}$$

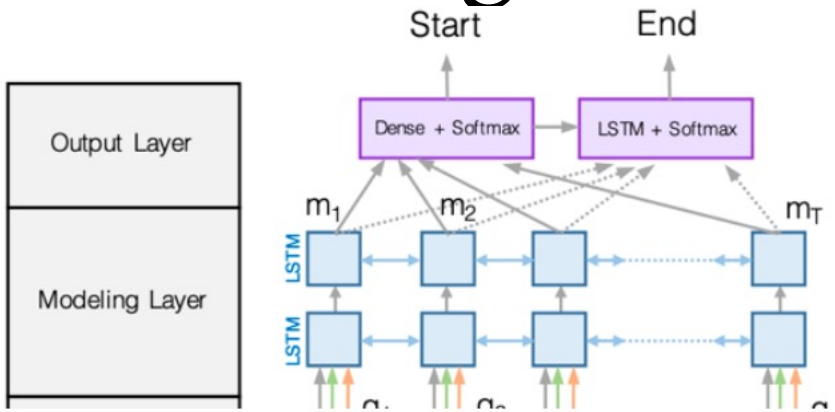
- Context-to-query attention (which question words are more relevant to c_i):

$$\alpha_{i,j} = \text{softmax}_j(S_{i,j}) \in \mathbb{R} \qquad \mathbf{a}_i = \sum_{j=1}^{M} \alpha_{i,j}\mathbf{q}_j \in \mathbb{R}^{2H}$$

- Query-to-context attention (which context words are relevant to some question words):

$$\beta_i = \text{softmax}_i(\max_{j=1}^{M}(S_{i,j})) \in \mathbb{R}^N \qquad \mathbf{b} = \sum_{i=1}^{N} \beta_i \mathbf{c}_i \in \mathbb{R}^{2H}$$

BiDAF: Modeling and output layers



The final training loss is

$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

- Modeling layer: pass $\mathbf{g}_i$ to another two layers of **bi-directional** LSTMs.
  - Attention layer is modeling interactions between query and context
  - Modeling layer is modeling interactions within context words

$$\mathbf{m}_i = \text{BiLSTM}(\mathbf{g}_i) \in \mathbb{R}^{2H}$$

- Output layer: two classifiers predicting the start and end positions:

$$p_{\text{start}} = \text{softmax}(\mathbf{w}_{\text{start}}^\intercal [\mathbf{g}_i; \mathbf{m}_i]) \qquad p_{\text{end}} = \text{softmax}(\mathbf{w}_{\text{end}}^\intercal [\mathbf{g}_i; \mathbf{m}_i'])$$

$$\mathbf{m}_i' = \text{BiLSTM}(\mathbf{m}_i) \in \mathbb{R}^{2H} \quad \mathbf{w}_{\text{start}}, \mathbf{w}_{\text{end}} \in \mathbb{R}^{10H}$$

BiDAF: Performance on SQuAD

- This model achieved 77.3 F1 on SQuAD v1.1.
  - Without context-to-query attention ➔ 67.7 F1
  - Without query-to-context attention ➔73.7 F1
  - Without character embeddings ➔ 75.4 F1

	F1
Logistic regression	51.0
Fine-Grained Gating (Carnegie Mellon U)	73.3
Match-LSTM (Singapore Management U)	73.7
DCN (Salesforce)	75.9
BiDAF (UW & Allen Institute)	77.3
Multi-Perspective Matching (IBM)	78.7
ReasoNet (MSR Redmond)	79.4
DrQA (Chen et al. 2017)	79.4
r-net (MSR Asia) [Wang et al., ACL 2017]	79.7
Human performance	91.2

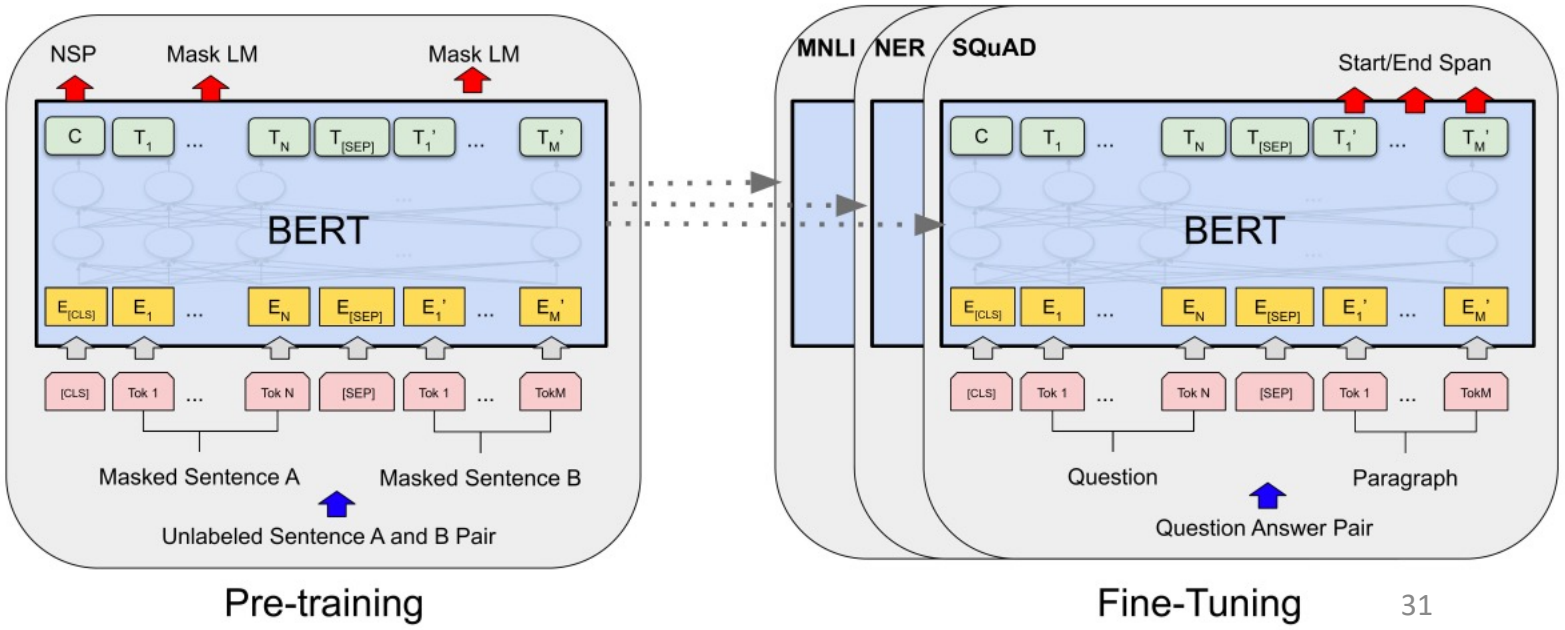
64

Attention visualization

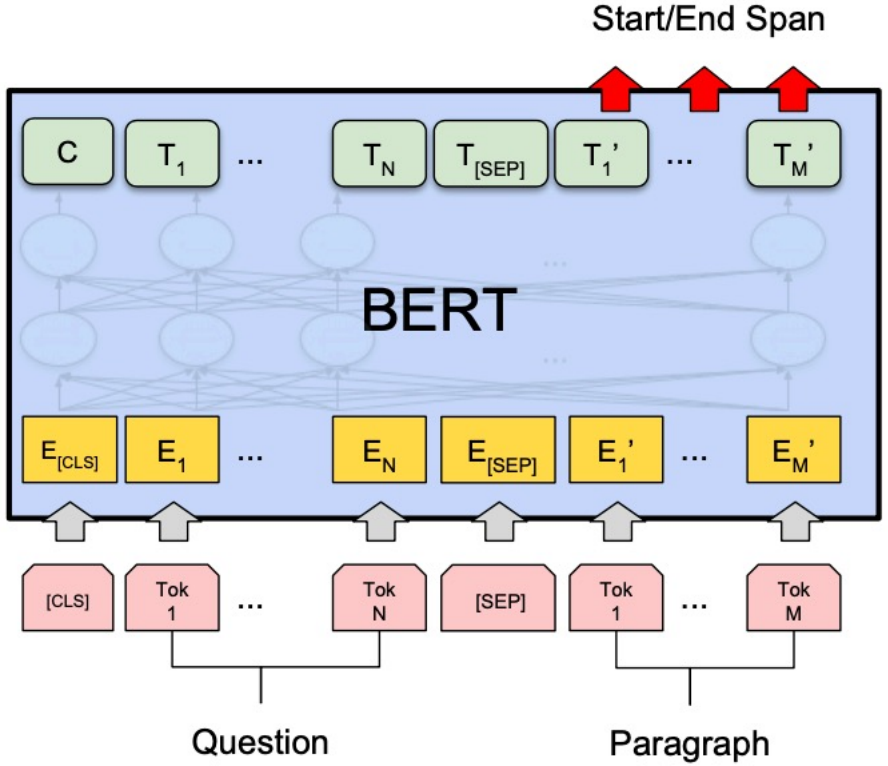
Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season . The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title . The game was played on February 7 , 2016 , at Levi 's Stadium in the San Francisco Bay Area at Santa Clara , California . As this was the 50th Super Bowl , the league emphasized the " golden anniversary " with various gold-themed initiatives , as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as " Super Bowl L ") , so that the logo could prominently feature the Arabic numerals 50 .

BERT for reading comprehension

- BERT is a deep bidirectional Transformer encoder pre-trained on large amounts of text (Wikipedia + BooksCorpus)
- BERT is pre-trained on two training objectives:
  - Masked language model (MLM)
  - Next sentence prediction (NSP)
- BERTbase has 12 layers and 110M parameters, BERTlarge has 24 layers and 330M parameters



BERT for reading comprehension



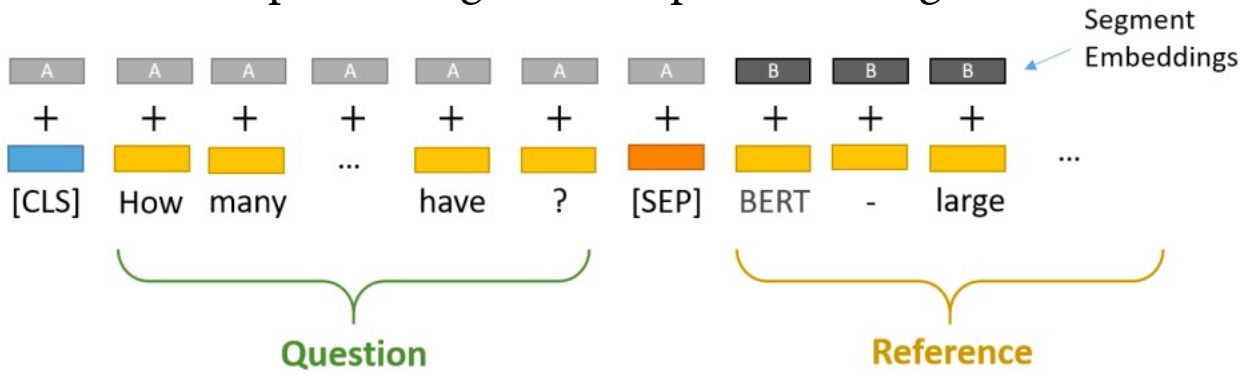
$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

$$p_{\text{start}}(i) = \text{softmax}_i(\mathbf{w}_{\text{start}}^{\intercal}\mathbf{h}_i)$$

$$p_{\text{end}}(i) = \text{softmax}_i(\mathbf{w}_{\text{end}}^{\intercal}\mathbf{h}_i)$$

where $\mathbf{h}_i$ is the hidden vector of c_i, returned by BERT

- Question = Segment A
- Passage = Segment B
- Answer = predicting two endpoints in segment B



Question: How many parameters does BERT-large have?

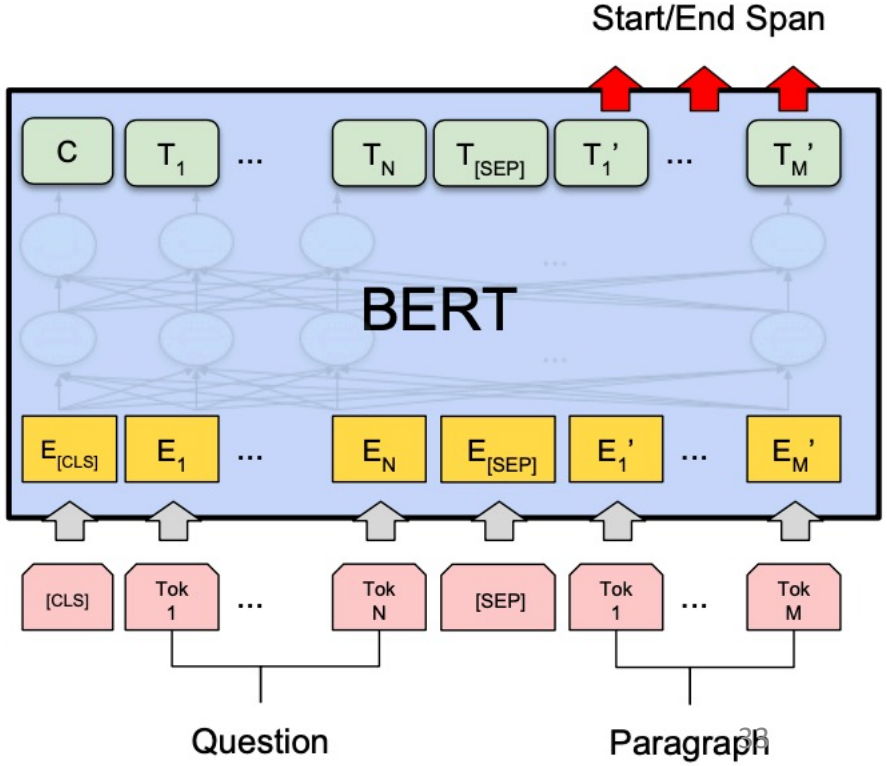
Reference Text: BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

BERT for reading comprehension

- BERT的所有参数（110M）以及新加的参数在优化L的过程中均会被优化。 $\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$
- 效果拔群！越强的预训练模型提供的效果越好。SQuAD从此演变成一个评价预训练模型的标准任务

	F1	EM
Human performance	91.2*	82.3*
BiDAF	77.3	67.7
BERT-base	88.5	80.8
BERT-large	90.9	84.1
XLNet	94.5	89.0
RoBERTa	94.6	88.9
ALBERT	94.8	89.3

(dev set, except for human performance)

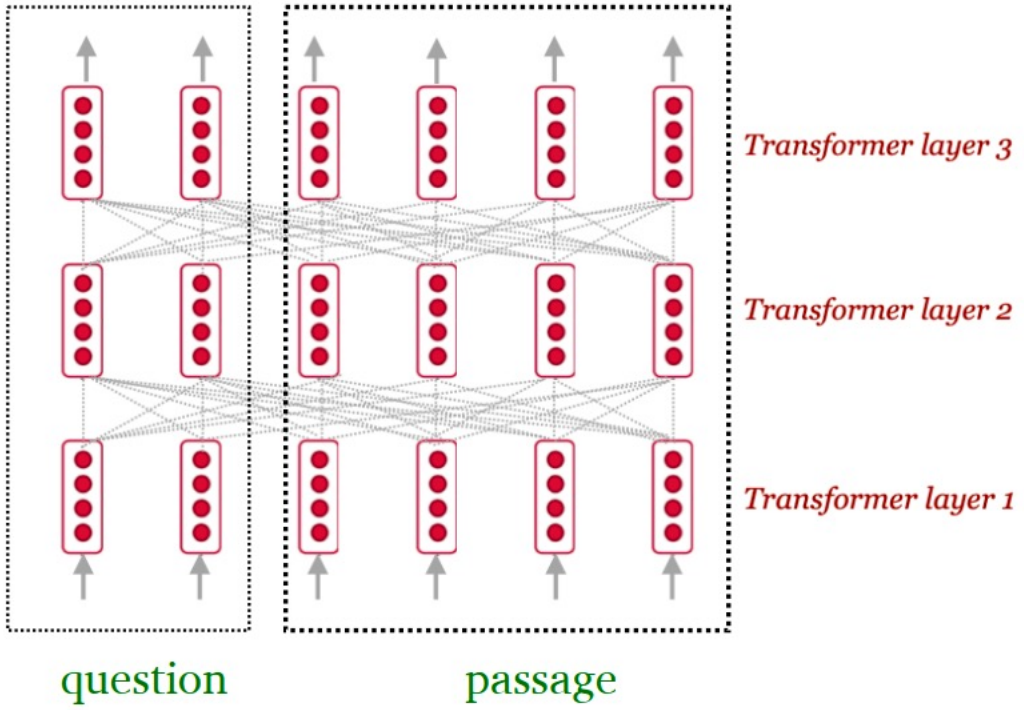


Comparisons between BiDAF and BERT models

- BERT参数更多(110M or 330M)，BiDAF大约2.5M
- BiDAF是建立在多个双向LSTM上的，BERT建立在Transformer上（没有循环结构，容易并行）
- BERT整体是预训练的模型。BiDAF是建立在GloVe上的（所有剩下的参数都需要从训练集中学习）

Comparisons between BiDAF and BERT models

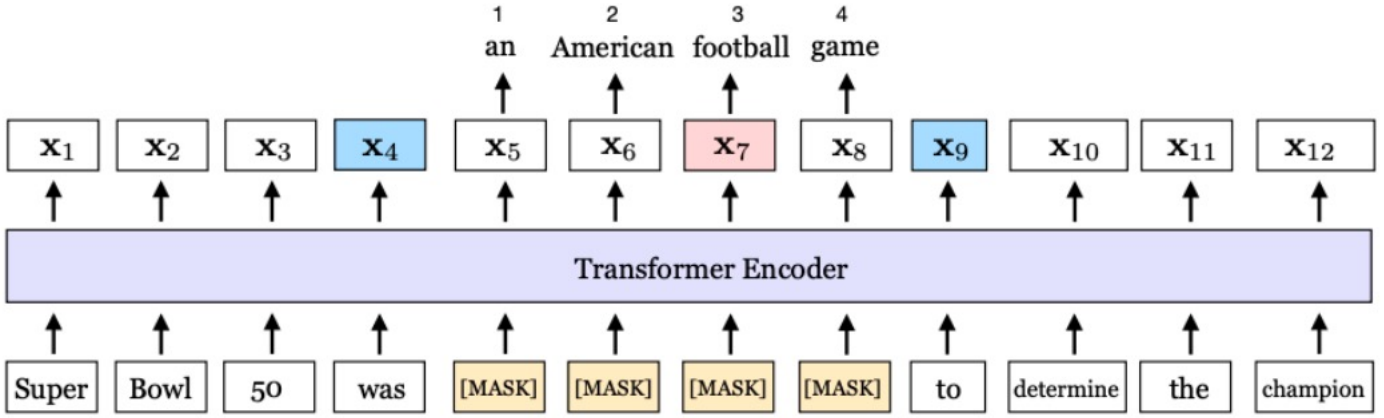
- 它们的本质是否不一样？
  - BiDAF集中建模问题与篇章之间的交互
  - BERT将问题Q与篇章P拼接后使用self-attention
    - attention(P, P) + attention(P, Q) + attention(Q, P) + attention(Q, Q)
  - (Clark and Gardner, 2018) 表明：在BiDAF中加入篇章的self-attention： attention(P, P) 同样可以提高效果



Can we design better pre-training objectives?

The answer is Yes!

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$
$$= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)$$



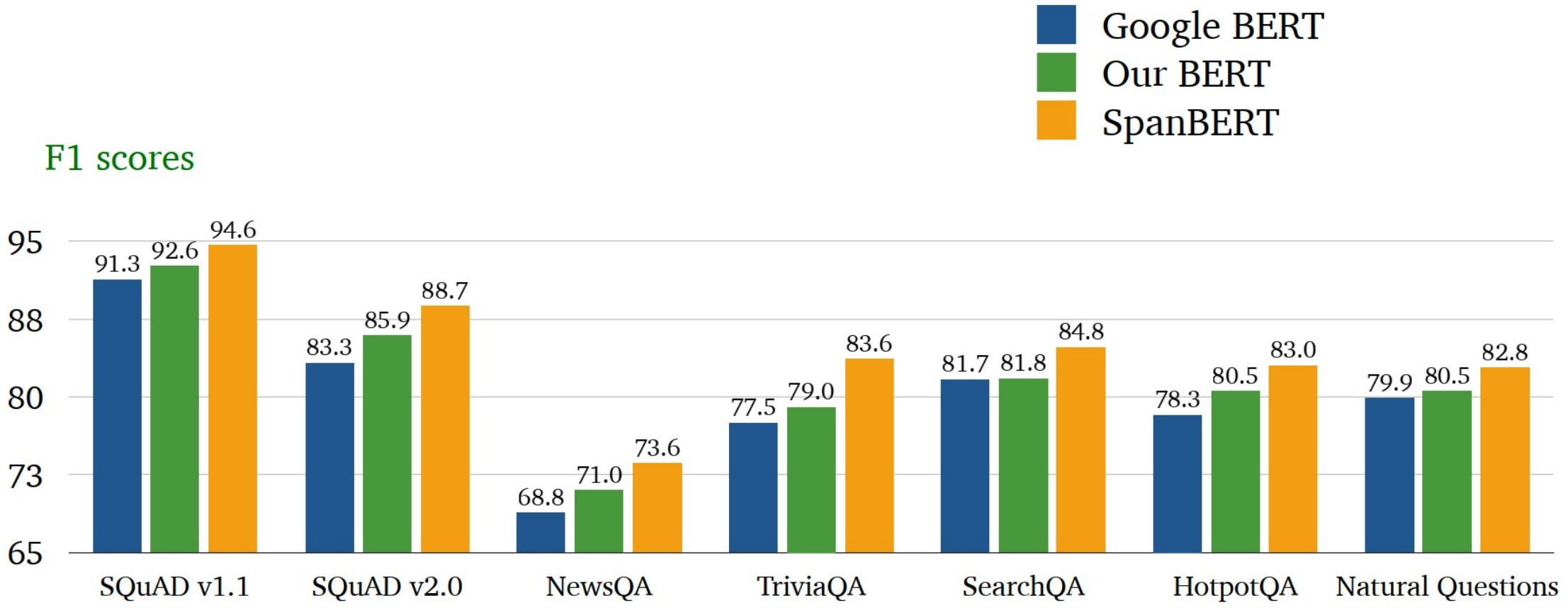
- Two ideas:
  - 1) 将整个答案区间mask掉，而不是随机mask15%的词
  - 2) 利用答案区间的两端来预测所有中间的单词= 把答案区间的所有信息压缩到两端

$$\mathbf{y}_i = f(\mathbf{x}_{s-1}, \mathbf{x}_{e+1}, \mathbf{p}_{i-s+1})$$

Position embedding

(Joshi & Chen et al., 2020): SpanBERT: Improving Pre-training by Representing and Predicting Spans

SpanBERT performance



Is reading comprehension solved?

- 既然SQuAD目前的效果已经超越人类，代表这个问题解决了？
- 目前系统在面临对抗样本以及领域外样本时依然表现不佳

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

	Match Single	Match Ens.	BiDAF Single	BiDAF Ens.
Original	71.4	75.4	75.5	80.0
ADDSENT	27.3	29.4	34.3	34.2
ADDONESENT	39.0	41.8	45.7	46.9
ADDANY	7.6	11.7	4.8	2.7
ADDCOMMON	38.9	51.0	41.7	52.6

(Jia and Liang, 2017): Adversarial Examples for Evaluating Reading Comprehension Systems

Is reading comprehension solved?

- 一个数据集上训练出的模型无法直接迁移到另一个数据集

Fine-tuned on \ Evaluated on	SQuAD	TriviaQA	NQ	QuAC	NewsQA
SQuAD	**75.6**	46.7	48.7	20.2	41.1
TriviaQA	49.8	**58.7**	42.1	20.4	10.5
NQ	53.5	46.3	**73.5**	21.6	24.7
QuAC	39.4	33.1	33.8	**33.3**	13.8
NewsQA	52.1	38.4	41.7	20.4	**60.1**

(Sen and Saffari, 2020): What do Models Learn from Question Answering Datasets?

Is reading comprehension solved?

BERT-large model trained on SQuAD

	Test *TYPE* and Description	Failure Rate (🤖)	Example Test cases (with expected behavior and 🤖 prediction)
Vocab	***MFT:*** comparisons	20.0	**C:** Victoria is younger than Dylan. **Q:** Who is less young? **A:** Dylan 🤖: Victoria
	MFT: intensifiers to superlative: most/least	91.3	**C:** Anna is worried about the project. Matthew is extremely worried about the project. **Q:** Who is least worried about the project? **A:** Anna 🤖: Matthew
Taxonomy	***MFT:*** match properties to categories	82.4	**C:** There is a tiny purple box in the room. **Q:** What size is the box? **A:** tiny 🤖: purple
	MFT: nationality vs job	49.4	**C:** Stephanie is an Indian accountant. **Q:** What is Stephanie's job? **A:** accountant 🤖: Indian accountant
	MFT: animal vs vehicles	26.2	**C:** Jonathan bought a truck. Isabella bought a hamster. **Q:** Who bought an animal? **A:** Isabella 🤖: Jonathan
	MFT: comparison to antonym	67.3	**C:** Jacob is shorter than Kimberly. **Q:** Who is taller? **A:** Kimberly 🤖: Jacob
	MFT: more/less in context, more/less antonym in question	100.0	**C:** Jeremy is more optimistic than Taylor. **Q:** Who is more pessimistic? **A:** Taylor 🤖: Jeremy
Robust.	***INV:*** Swap adjacent characters in **Q** (typo)	11.6	**C:** ...Newcomen designs had a duty of about 7 million, but most were closer to 5 million.... **Q:** What was the ideal duty → udty of a Newcomen engine? **A:** INV 🤖: 7 million → 5 million
	INV: add irrelevant sentence to **C**	9.8	(no example)

(Ribeiro et al., 2020): Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Is reading comprehension solved?

BERT-large model trained on SQuAD

Temporal	***MFT:*** change in one person only	41.5	**C:** Both Luke and Abigail were writers, but there was a change in Abigail, who is now a model. **Q:** Who is a model? **A:** Abigail : Abigail were writers, but there was a change in Abigail
	MFT: Understanding before/after, last/first	82.9	**C:** Logan became a farmer before Danielle did. **Q:** Who became a farmer last? **A:** Danielle : Logan
Neg.	***MFT:*** Context has negation	67.5	**C:** Aaron is not a writer. Rebecca is. **Q:** Who is a writer? **A:** Rebecca : Aaron
	MFT: **Q** has negation, **C** does not	100.0	**C:** Aaron is an editor. Mark is an actor. **Q:** Who is not an actor? **A:** Aaron : Mark
Coref.	***MFT:*** Simple coreference, he/she.	100.0	**C:** Melissa and Antonio are friends. He is a journalist, and she is an adviser. **Q:** Who is a journalist? **A:** Antonio : Melissa
	MFT: Simple coreference, his/her.	100.0	**C:** Victoria and Alex are friends. Her mom is an agent **Q:** Whose mom is an agent? **A:** Victoria : Alex
	MFT: former/latter	100.0	**C:** Kimberly and Jennifer are friends. The former is a teacher **Q:** Who is a teacher? **A:** Kimberly : Jennifer
SRL	***MFT:*** subject/object distinction	60.8	**C:** Richard bothers Elizabeth. **Q:** Who is bothered? **A:** Elizabeth : Richard
	MFT: subj/obj distinction with 3 agents	95.7	**C:** Jose hates Lisa. Kevin is hated by Lisa. **Q:** Who hates Kevin? **A:** Lisa : Jose

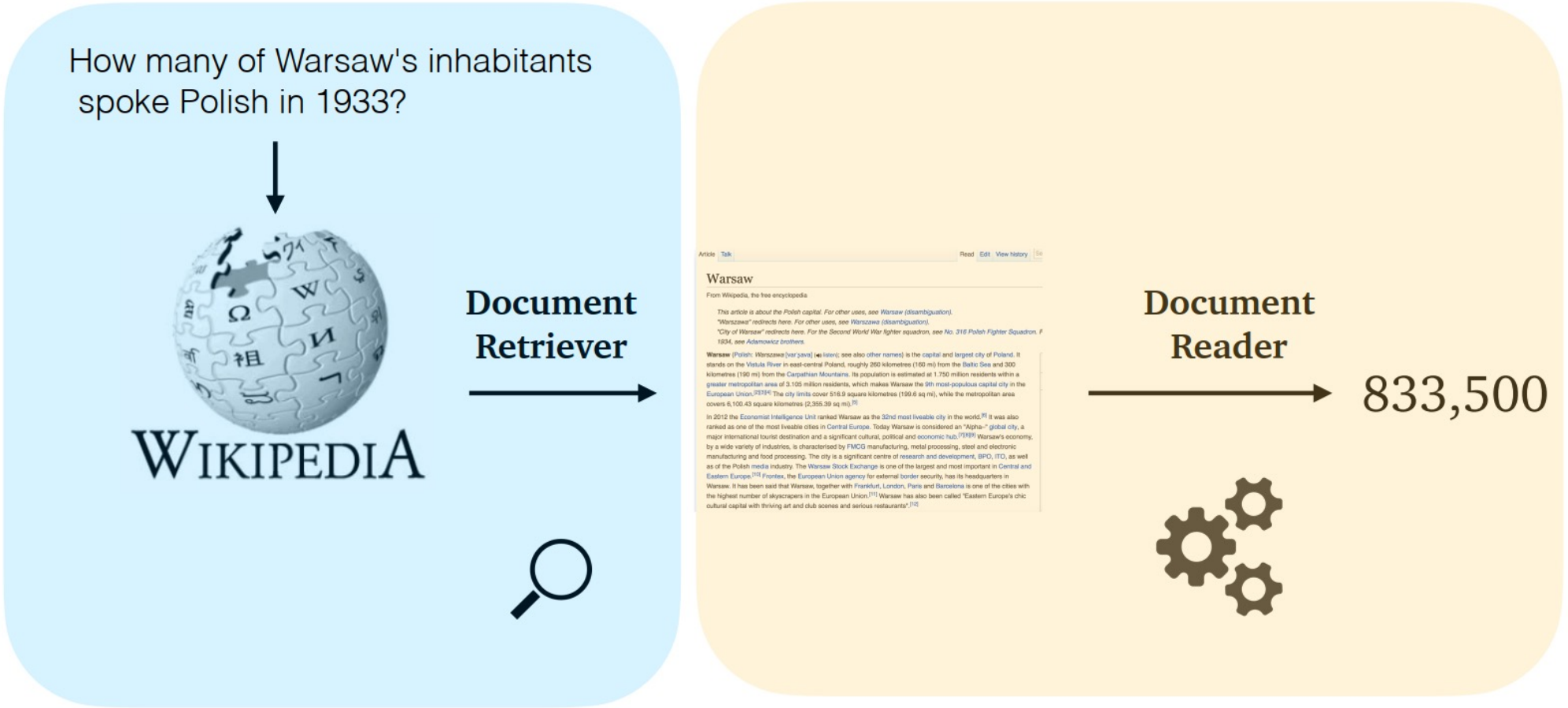
(Ribeiro et al., 2020): Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Open-domain question answering

- 与阅读理解不同，开域QA不给定一篇文章
- 开域QA允许访问一大组文档，答案可能在其中一篇中。目标是为任何给定问题返回一个答案
- 更加有挑战性！更加贴近实际！

Close-domain：在某一个特定领域内回答问题

Retriever-reader framework



https://github.com/facebookresearch/DrQA

Retriever-reader framework

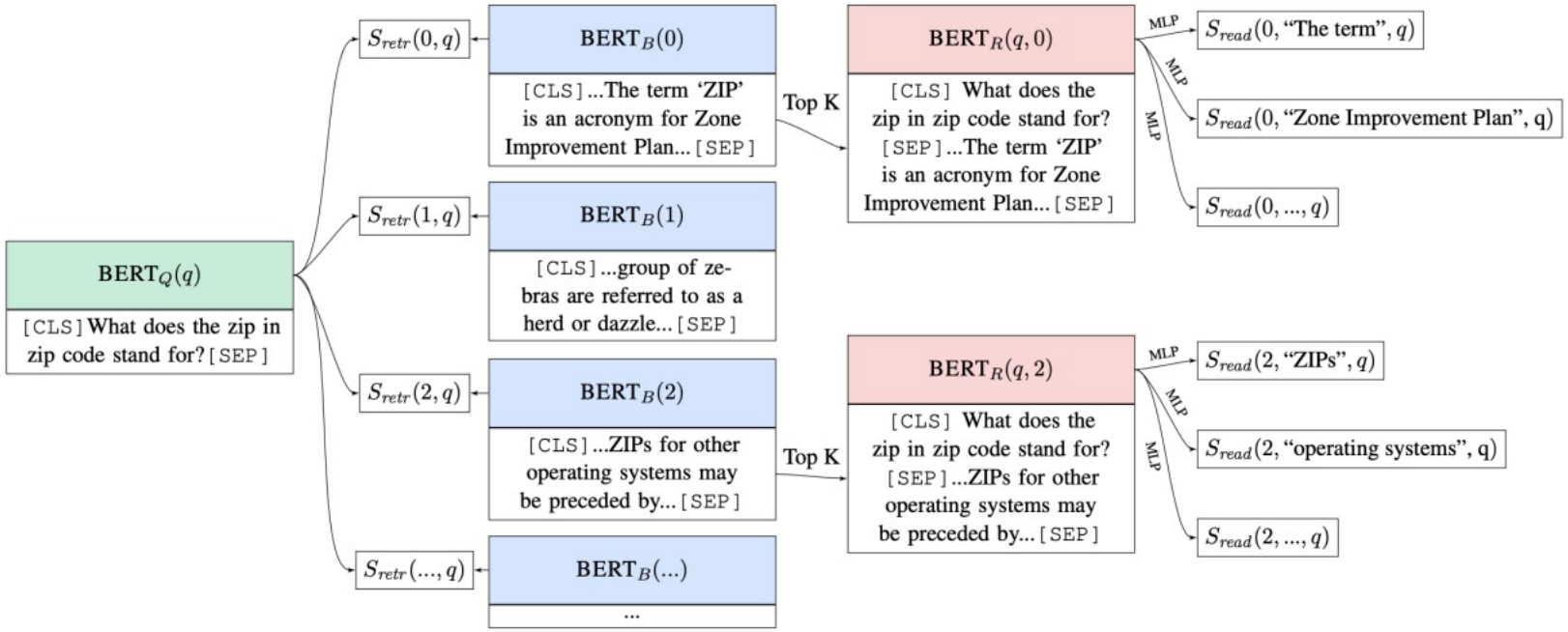
- 输入：一组文档 $\mathscr{D} = D_1, D_2, \ldots, D_N$ 和问题Q
- 输出：答案字符串A

- Retriever: $f(\mathscr{D}, Q) \longrightarrow P_1, \ldots, P_K$　K是预定义的（比如100）
- Answer: $g(Q, \{P_1, \ldots, P_K\}) \longrightarrow A$　阅读理解问题

- In DrQA,
  - Retriever = 一个标准的基于TF-IDF 信息检索的系数模型(固定模块)
  - Reader = 神经网络阅读理解模型
  - 用SQuAD 以及其他的远程指导的QA数据集训练
  - Distantly-supervised examples: (Q, A) ⟶ (P, Q, A)

We can train the retriever too

- Joint training of retriever and reader



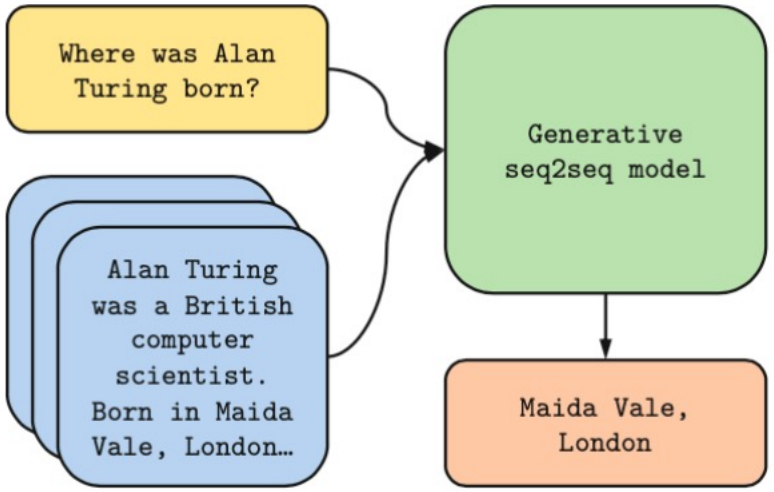
- 每篇文章都利用BERT编码成向量，检索分数就可以是问题的向量与篇章向量的点积
- 然而文章数量太多，很难建模(e.g., 21M in English Wikipedia)

[1] Lee et al., 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering
[2] Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

Dense retrieval + generative models

- 直接生成答案

Fusion-in-decoder (FID) = DPR + T5

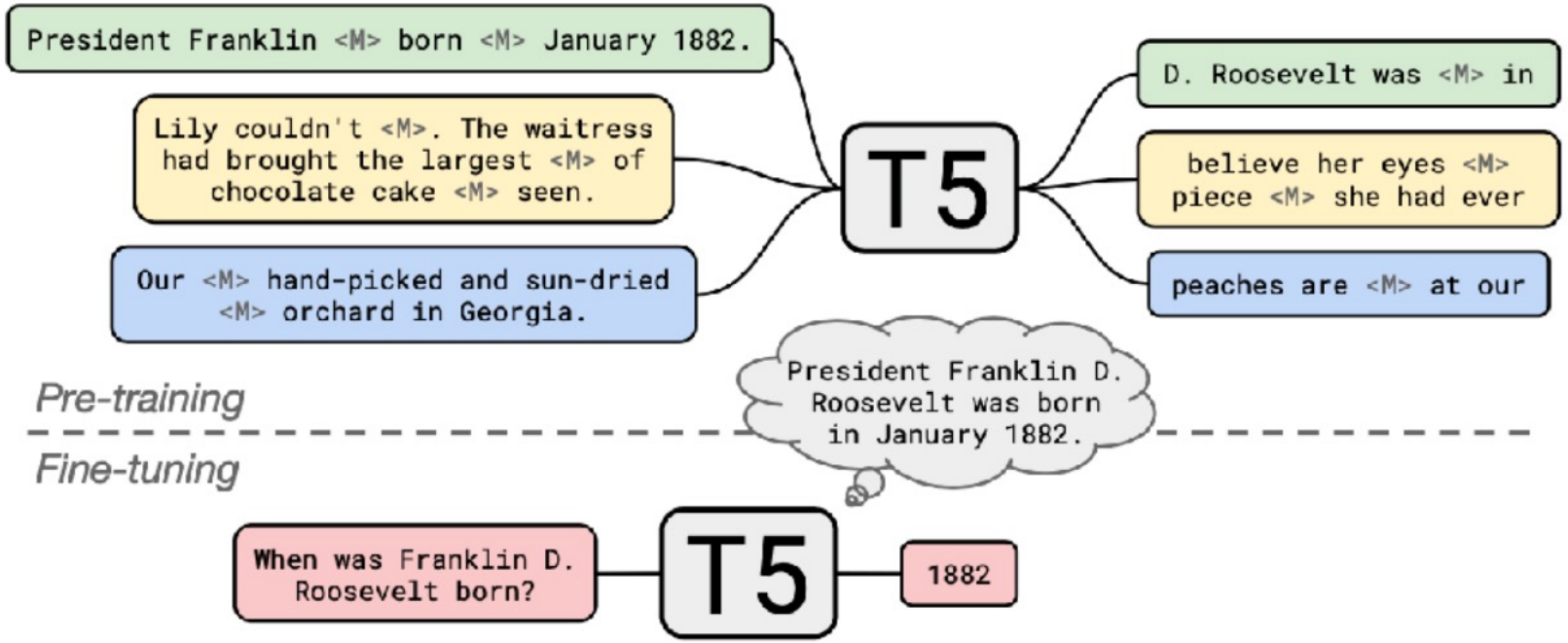


Model	NaturalQuestions	TriviaQA	
ORQA (Lee et al., 2019)	31.3	45.1	-
REALM (Guu et al., 2020)	38.2	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-
SpanSeqGen (Min et al., 2020)	42.5	-	-
RAG (Lewis et al., 2020)	44.5	56.1	68.0
T5 (Roberts et al., 2020)	36.6	-	60.5
GPT-3 few shot (Brown et al., 2020)	29.9	-	71.2
Fusion-in-Decoder (base)	48.2	65.0	77.1
Fusion-in-Decoder (large)	**51.4**	**67.6**	**80.1**

Izacard and Grave 2020. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering

Large language models can do open-domain QA well

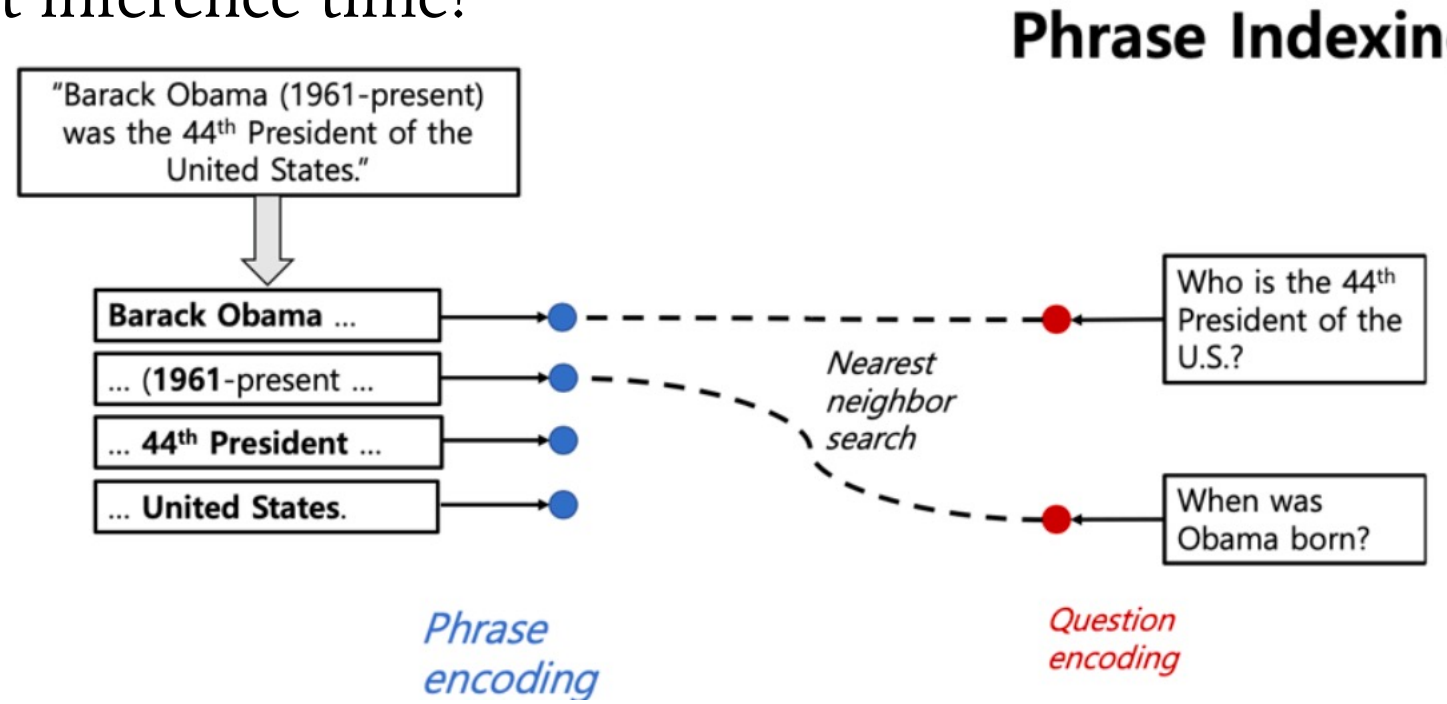
- 取消明确的检索步骤



Roberts et al., 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?

Maybe the reader model is not necessary too!

- It is possible to encode all the phrases (60 billion phrases in Wikipedia) using dense vectors and only do nearest neighbor search without a BERT model at inference time!



[1] Seo et al., 2019. Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index
[2] Lee et al., 2020. Learning Dense Representations of Phrases at Scale

Thank you