How to Translate?

(cs224n: Translation, Seq2Seq, Attention)

자연어 심화세미나

TOBIG'S 19기 이영아

Contents

Unit 01

What is question answering?

Unit 01 | What is question answering?

QA이란 무엇인가?

- 인간의 언어로 제시된 질문에 자동으로 답하는 시스템

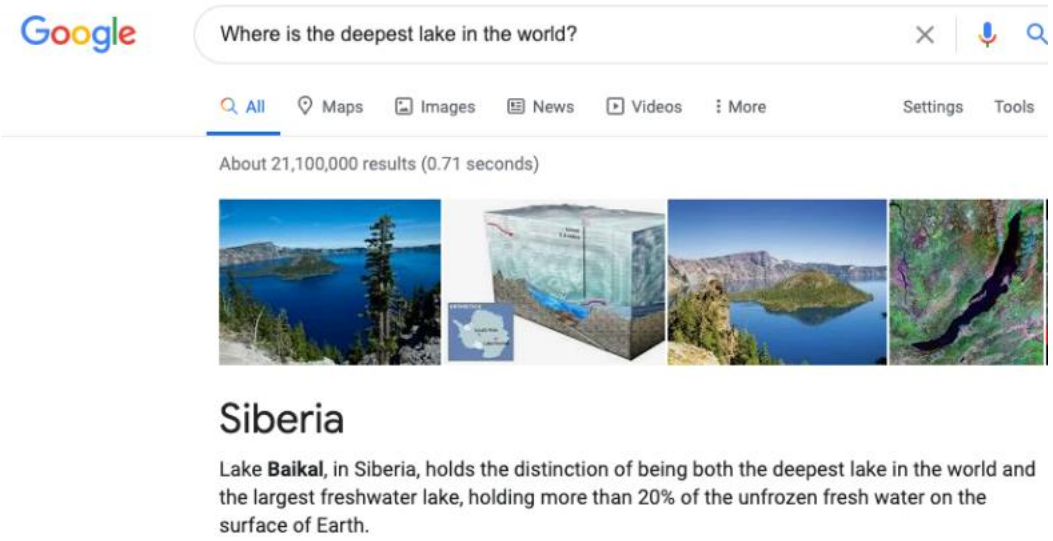
QA의 여러 종류

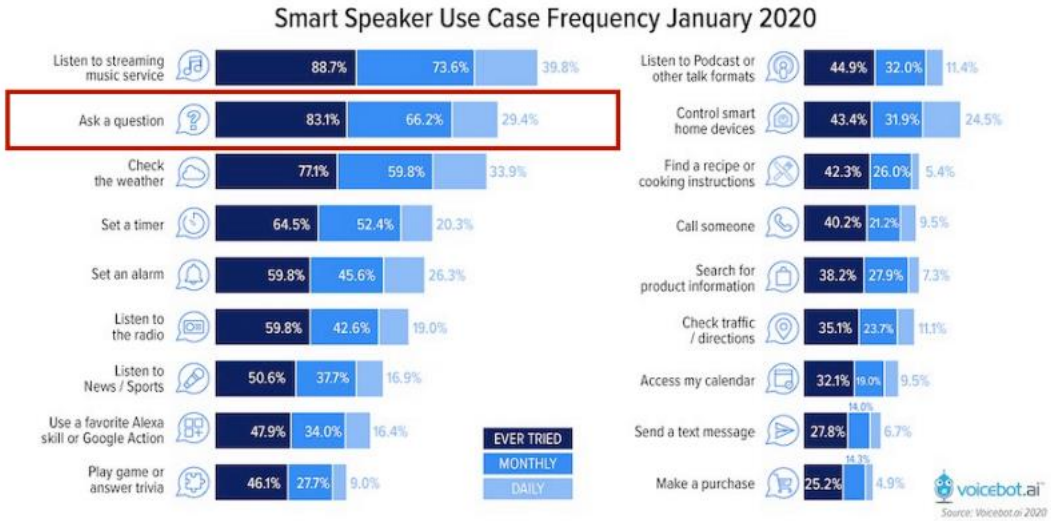
- 시스템 형성의 기반 (text passage, web documents, knowledge bases, tables, images)
- 질문 타입 (factoid vs non-factoid, open-domain vs closed-domain, simple vs compositional)
- 답변 타입 (short segment of text, paragraph, list, yes/no)

Unit 01 | What is question answering?

19기 자연어 심화세미나
TOBIG'S 19기 이영아

QA 쓰임





Unit 01 | What is question answering?

19기 자연어 심화세미나
TOBIG'S 19기 이영아

여러 타입의 QA

Knowledge based QA

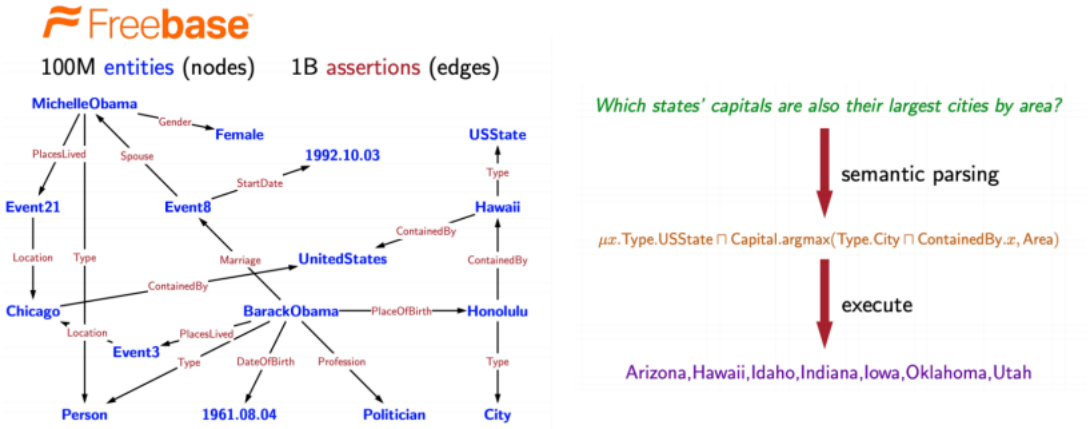


Image credit: Percy Liang

Visual QA

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

Unit 02

Reading Comprehension

Unit 02 | Reading Comprehension

Reading comprehension

- 텍스트를 이해하고 질문에 답하기
- (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

Unit 02 | Reading Comprehension

QA를 다루는 이유

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

UCD ***finished*** the 2006 championship as Dublin champions ,
by ***beating*** St Vincents in the final .

finished	Who finished something? - UCD What did someone finish? - the 2006 championship What did someone finish something as? - Dublin champions How did someone finish something? - by beating St Vincents in the final
beating	Who beat someone? - UCD When did someone beat someone? - in the final Who did someone beat? - St Vincents

(He et al., 2015)

Unit 02 | Reading Comprehension

19기 자연어 심화세미나
TOBIG'S 19기 이영아

SQuAD

- 100k annotated (passage, question, answer)
- passages : English Wikipedia (100~150)
- questions : crowd sourced
- answers : short segment of text in the passage

- 평가 방법
- exact match (0 or 1) / F1

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

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

Unit 02 | Reading Comprehension

How can we build a model to solve SQuAD?

- Problem formulation
  - Input: $C = (c_1, c_2, \ldots, c_N), Q = (q_1, q_2, \ldots, q_M), c_i, q_i \in V$
  - Output: $1 \leq \text{start} \leq \text{end} \leq N$

- (answer은 input passage의 일부일 것이므로)

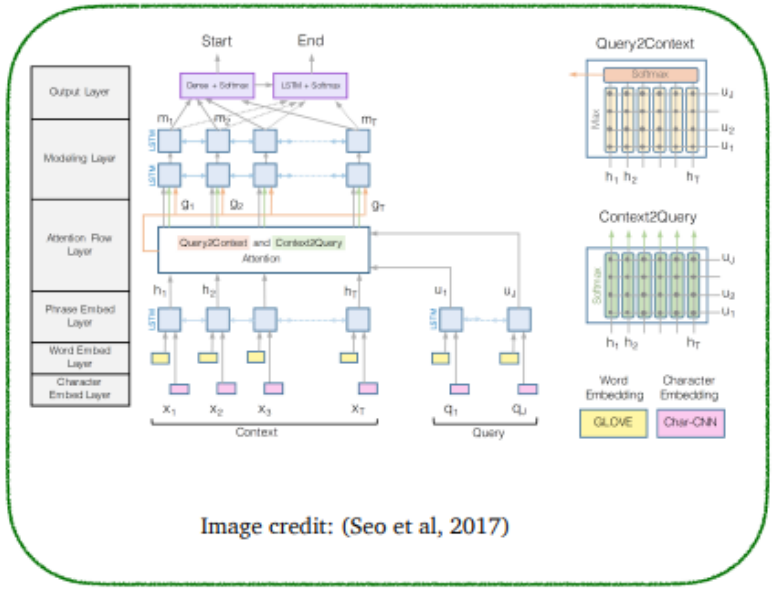


Image credit: (Seo et al, 2017)

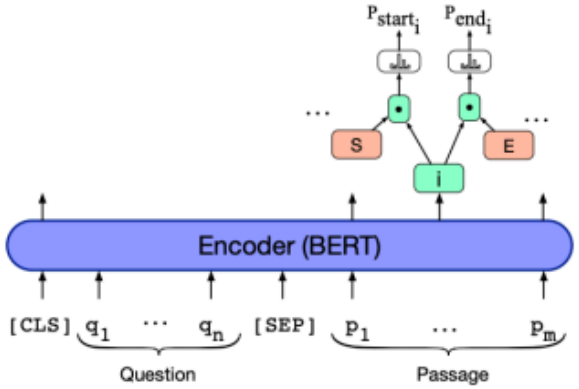
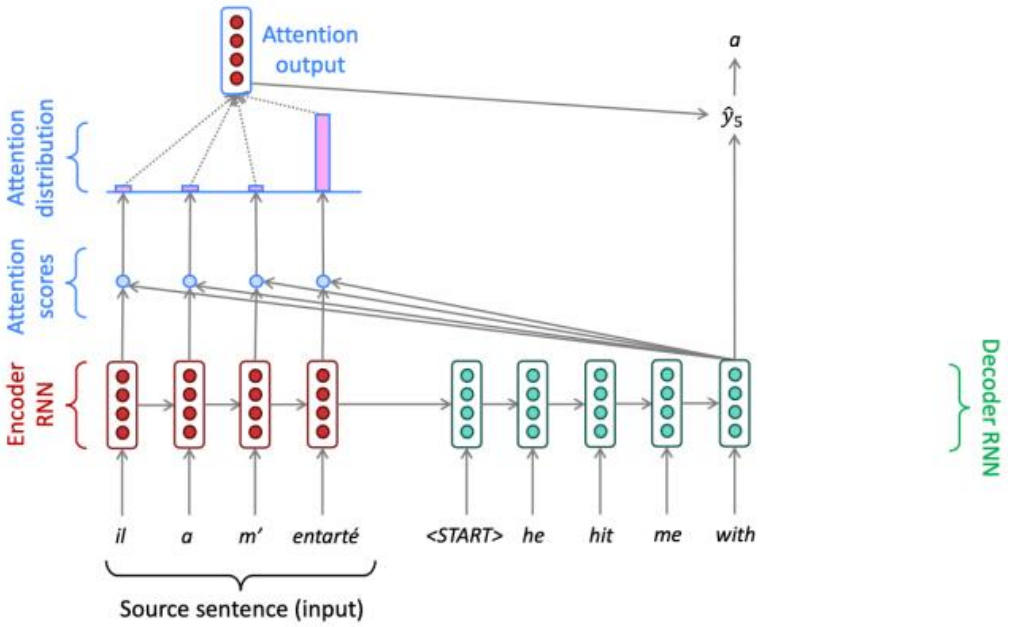


Image credit: J & M, edition 3

Unit 02 | Reading Comprehension

seq2seq model with attention

- seq2seq 모델의 machine translation task와 유사성
- 입력 : source and target sentence ~ passage and question
- 어텐션 : source word와 target word 사이의 관련성 찾기
  ~ passag의 word와 questio과의 관련성 찾기

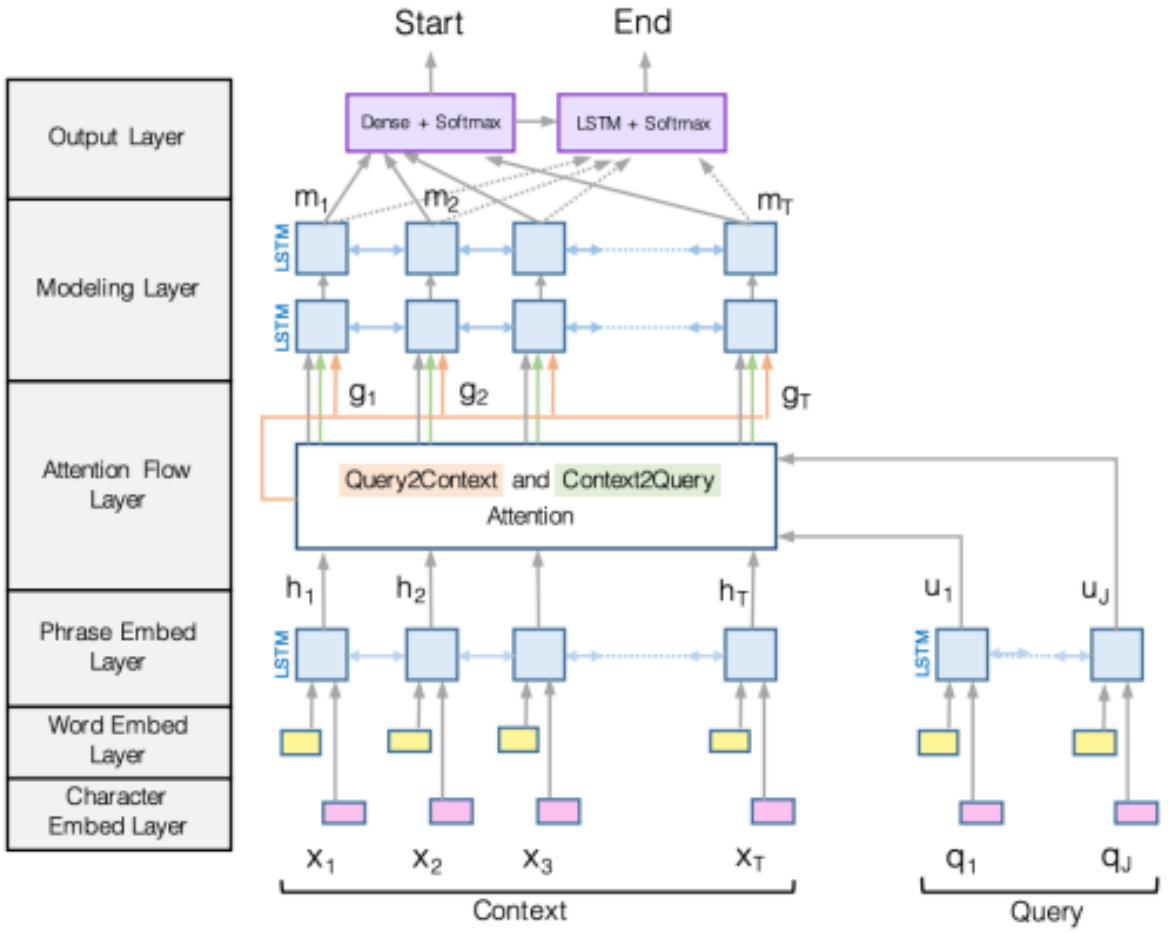


- 차이점
- machine translation은 생성을 위해 디코더 사용
- reading comprehension은 생성할 필요 없음
  -> start, end position classifier

Unit 02 | Reading Comprehension

19기 자연어 심화세미나
TOBIG'S 19기 이영아

Bidirectional Attention Flow model

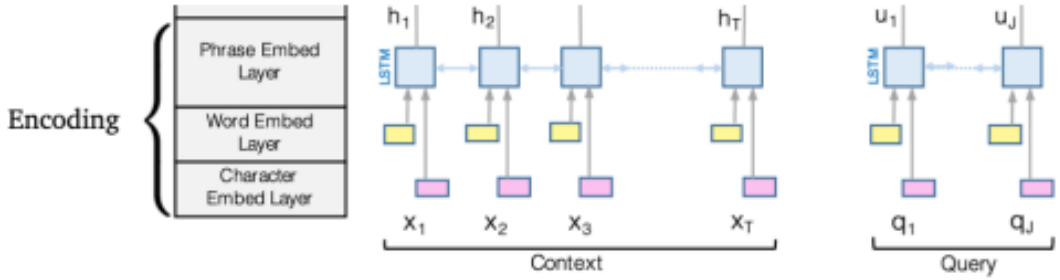


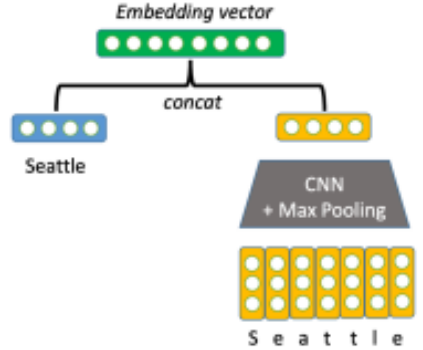
(Seo et al., 2017): Bidirectional Attention Flow for Machine Comprehension

Unit 02 | Reading Comprehension

19기 자연어 심화세미나
TOBIG'S 19기 이영아

BiDAF : Encoding





- character embedding : CNNs
- word embedding : GloVe

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

f: high-way networks omitted here

- contextual embedding : bidirectional LSTM

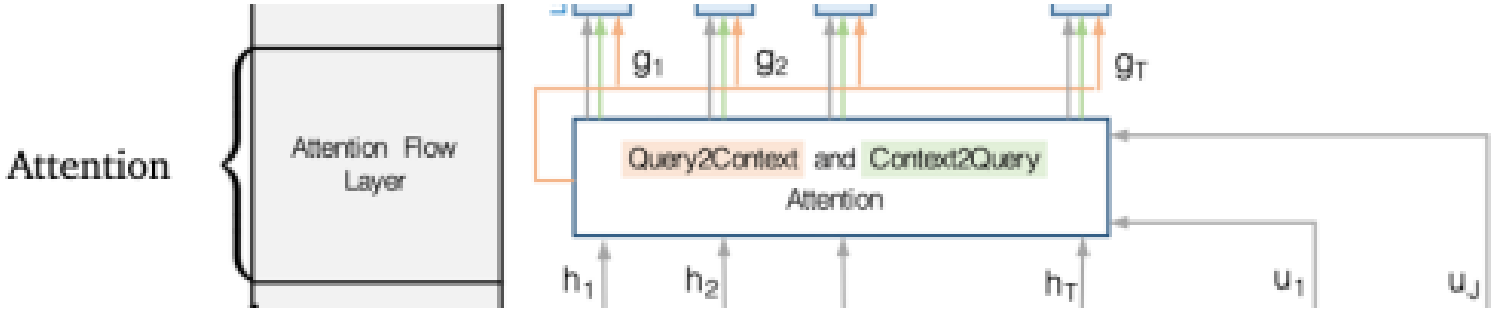
$$\overrightarrow{\mathbf{c}}_i = \text{LSTM}(\overrightarrow{\mathbf{c}}_{i-1}, e(c_i)) \in \mathbb{R}^H$$
$$\overleftarrow{\mathbf{c}}_i = \text{LSTM}(\overleftarrow{\mathbf{c}}_{i+1}, e(c_i)) \in \mathbb{R}^H$$
$$\mathbf{c}_i = [\overrightarrow{\mathbf{c}}_i; \overleftarrow{\mathbf{c}}_i] \in \mathbb{R}^{2H}$$

$$\overrightarrow{\mathbf{q}}_i = \text{LSTM}(\overrightarrow{\mathbf{q}}_{i-1}, e(q_i)) \in \mathbb{R}^H$$
$$\overleftarrow{\mathbf{q}}_i = \text{LSTM}(\overleftarrow{\mathbf{q}}_{i+1}, e(q_i)) \in \mathbb{R}^H$$
$$\mathbf{q}_i = [\overrightarrow{\mathbf{q}}_i; \overleftarrow{\mathbf{q}}_i] \in \mathbb{R}^{2H}$$

- 통상적으로 decoder로 unidirectional LSTM -> generate
- context을 표현하고자 biLSTM사용

Unit 02 | Reading Comprehension

BiDAF : Attention



- attention idea : query와 context 사이의 interaction 포착
- context to query attention : context word에 가장 관련있는 query word 찾기

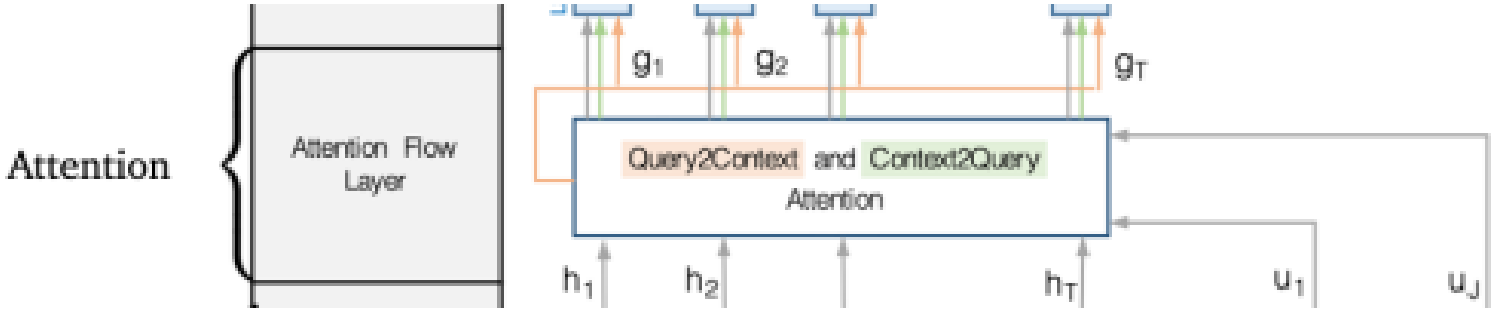
Q: *Who leads the United States?*

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

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

Unit 02 | Reading Comprehension

BiDAF : Attention



- attention idea : query와 context 사이의 interaction 포착
- query to context attention : query word에 가장 관련있는 context word 찾기

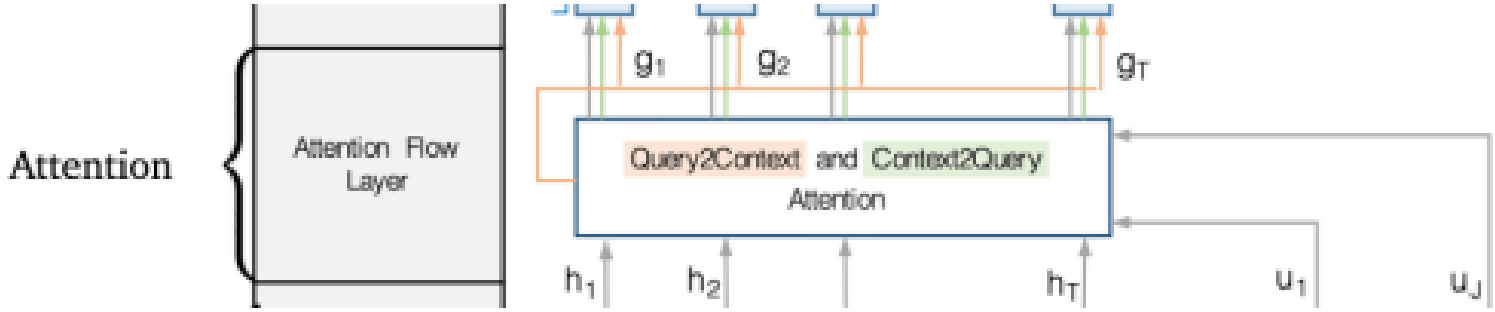
While Seattle's weather is very nice in summer, its weather is very rainy in winter, making it one of the most gloomy cities in the U.S. LA is ...

Q: Which city is gloomy in winter?

Unit 02 | Reading Comprehension

19기 자연어 심화세미나
TOBIG'S 19기 이영아

BiDAF : Attention



- 각 c, q 페어의 유사도 계산

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

- context-to-query attention (ci와 가장 관련있는 question word 찾기)

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

- query-to-context attention (question words와 가장 관련있는 context words 찾기)

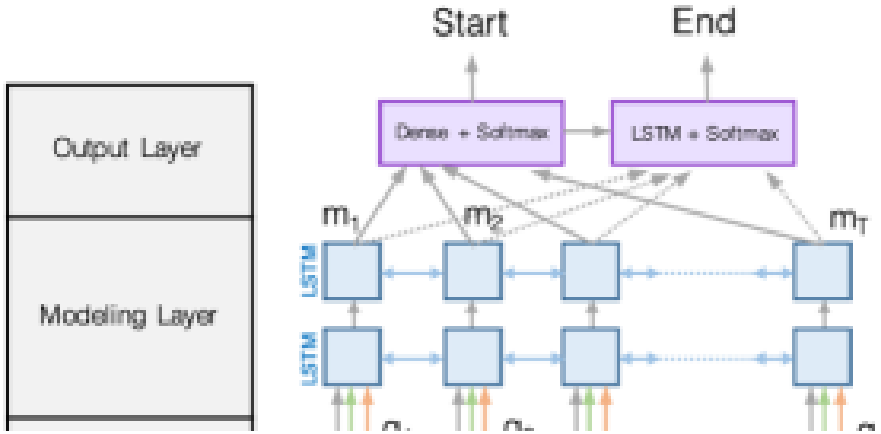
$$\beta_i = \text{softmax}_i(\max_{j=1}^{M}(S_{i,j})) \in \mathbb{R}^N \qquad \mathbf{b} = \sum_{i=1}^{N} \beta_i \mathbf{c}_i \in \mathbb{R}^{2H}$$

The final output is

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

Unit 02 | Reading Comprehension

BiDAF : Modeling and output layers



- Modeling $\mathbf{m}_i = \mathrm{BiLSTM}(\mathbf{g}_i) \in \mathbb{R}^{2H}$
- context word 사이 interaction

- output layers $p_{\text{start}} = \mathrm{softmax}(\mathbf{w}_{\text{start}}^{\intercal}[\mathbf{g}_i; \mathbf{m}_i])$ $p_{\text{end}} = \mathrm{softmax}(\mathbf{w}_{\text{end}}^{\intercal}[\mathbf{g}_i; \mathbf{m}'_i])$
- 2개의 classifier $\mathbf{m}'_i = \mathrm{BiLSTM}(\mathbf{m}_i) \in \mathbb{R}^{2H}$ $\mathbf{w}_{\text{start}}, \mathbf{w}_{\text{end}} \in \mathbb{R}^{10H}$

The final training loss is
$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$

Unit 02 | Reading Comprehension

19기 자연어 심화세미나
TOBIG'S 19기 이영아

BiDAF

This model achieved 77.3 F1 on SQuAD v1.1.

- Without context-to-query attention $\Longrightarrow$ 67.7 F1
- Without query-to-context attention $\Longrightarrow$ 73.7 F1
- Without character embeddings $\Longrightarrow$ 75.4 F1

	Published[12]	LeaderBoard[13]
Single Model	EM / F1	EM / F1
LR Baseline (Rajpurkar et al., 2016)	40.4 / 51.0	40.4 / 51.0
Dynamic Chunk Reader (Yu et al., 2016)	62.5 / 71.0	62.5 / 71.0
Match-LSTM with Ans-Ptr (Wang & Jiang, 2016)	64.7 / 73.7	64.7 / 73.7
Multi-Perspective Matching (Wang et al., 2016)	65.5 / 75.1	70.4 / 78.8
Dynamic Coattention Networks (Xiong et al., 2016)	66.2 / 75.9	66.2 / 75.9
FastQA (Weissenborn et al., 2017)	68.4 / 77.1	68.4 / 77.1
BiDAF (Seo et al., 2016)	68.0 / 77.3	68.0 / 77.3
SEDT (Liu et al., 2017a)	68.1 / 77.5	68.5 / 78.0
RaSoR (Lee et al., 2016)	70.8 / 78.7	69.6 / 77.7
FastQAExt (Weissenborn et al., 2017)	70.8 / 78.9	70.8 / 78.9
ReasoNet (Shen et al., 2017b)	69.1 / 78.9	70.6 / 79.4
Document Reader (Chen et al., 2017)	70.0 / 79.0	70.7 / 79.4
Ruminating Reader (Gong & Bowman, 2017)	70.6 / 79.5	70.6 / 79.5
jNet (Zhang et al., 2017)	70.6 / 79.8	70.6 / 79.8
Conductor-net	N/A	72.6 / 81.4
Interactive AoA Reader (Cui et al., 2017)	N/A	73.6 / 81.9
Reg-RaSoR	N/A	75.8 / 83.3
DCN+	N/A	74.9 / 82.8
AIR-FusionNet	N/A	76.0 / 83.9
R-Net (Wang et al., 2017)	72.3 / 80.7	76.5 /84.3
BiDAF + Self Attention + ELMo	N/A	**77.9/ 85.3**
Reinforced Mnemonic Reader (Hu et al., 2017)	73.2 / 81.8	73.2 / 81.8

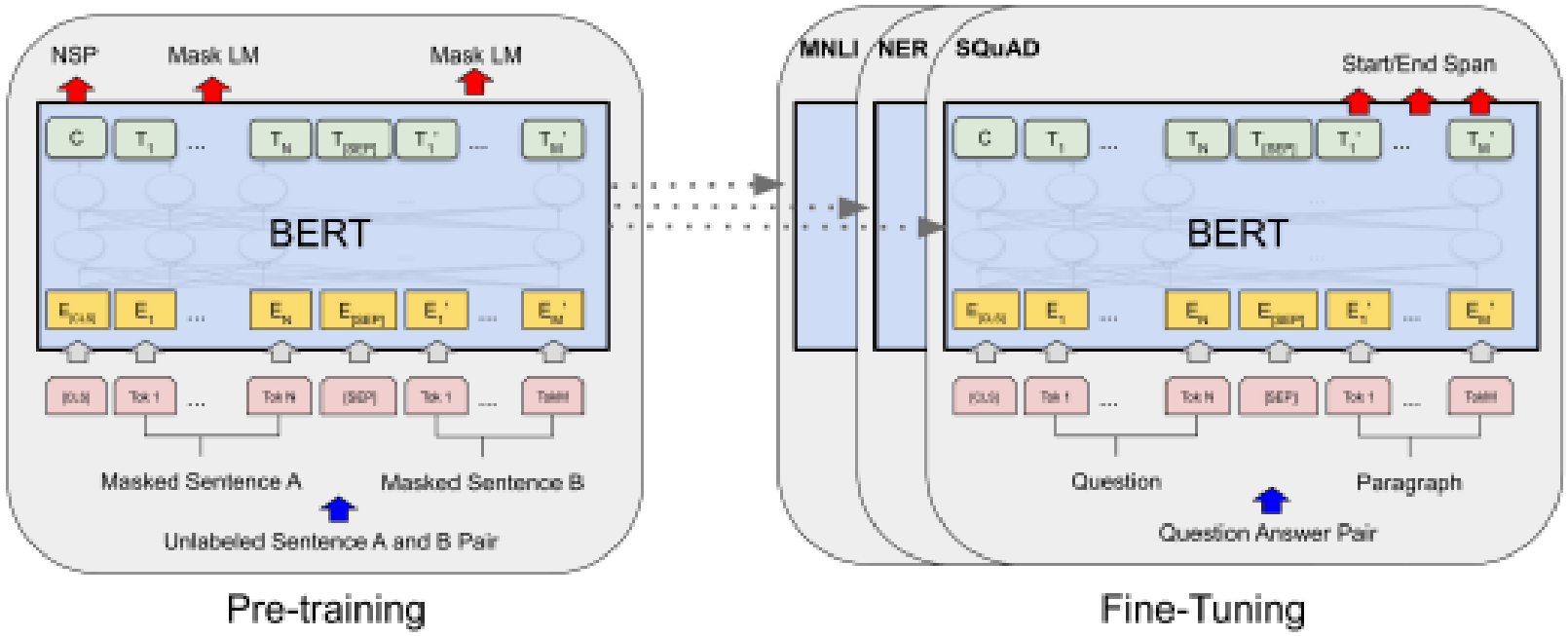
Unit 02 | Reading Comprehension

Attention visualization

Unit 02 | Reading Comprehension

BERT for reading comprehension

- deep bidirectional Transformer encoder

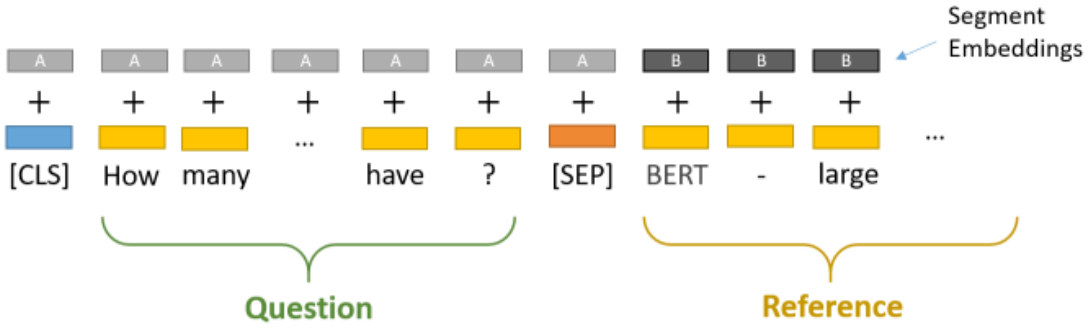


Unit 02 | Reading Comprehension

BERT for reading comprehension

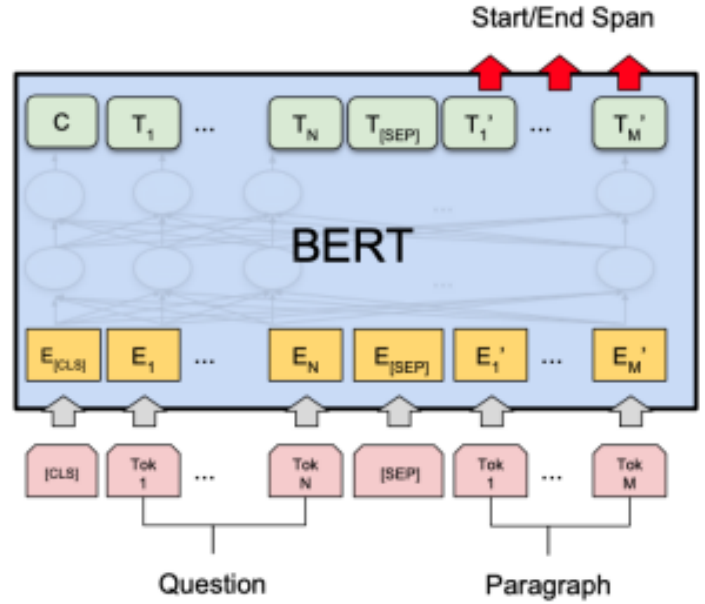
- deep bidirectional Transformer encoder

- segment A : Question
- segment B : Paragraph



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.



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

$$p_{\text{start}}(i) = \text{softmax}_i(\mathbf{w}_{\text{start}}^{\mathsf{T}}\mathbf{H})$$

$$p_{\text{end}}(i) = \text{softmax}_i(\mathbf{w}_{\text{end}}^{\mathsf{T}}\mathbf{H})$$

where $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N]$ are the hidden vectors of the paragraph, returned by BERT

Comparisons between BiDAF and BERT models

- BERT 모델은 BiDAF에 비해 훨씬 많은 파라미터 (110M vs 2.5M)
- BiDAF는 BiLSTM으로 구성된 반면 BERT는 Transformer로 구성 -> recurrence가 없어 병렬화에 용이
- BERT는 pre-trained 된 반면 BiDAF는 GloVe에서부터 쌓음

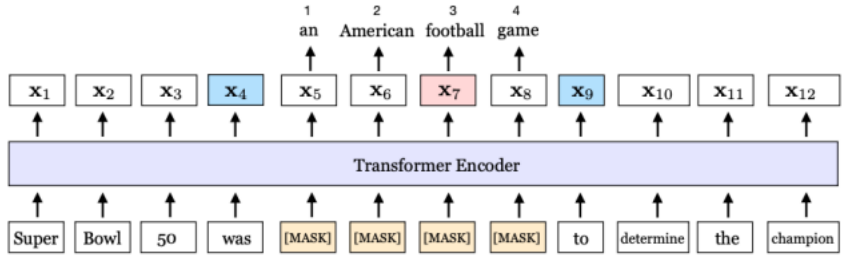
	F1	EM
Human performance	91.2*	82.3*
BiDAF	77.3	67.7
BERT-base	88.5	80.8
BERT-large	90.9	84.1
XLNet	94.5	89.0
RoBERTa	94.6	88.9
ALBERT	94.8	89.3

19기 자연어 심화세미나
TOBIG'S 19기 이영아

Can we design better pre-training objectives?

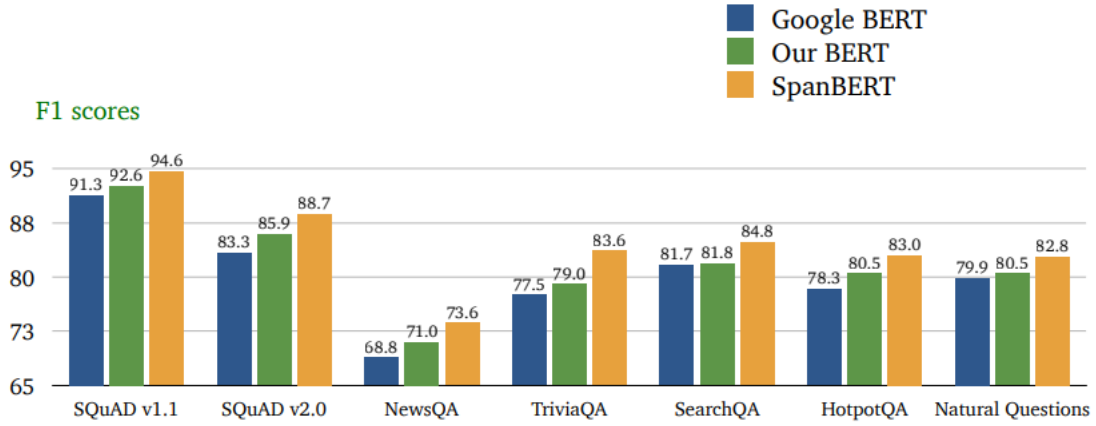
- SpanBERT에서 더 좋은 objective 제안

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



- 단어 각각이 아닌 contiguous span 단위로 마스킹
- 2개의 end point를 이용하여 그 사이의 단어들 마스킹

- 더 좋은 pre-training objective를 이용해 모델 사이즈를 키우지 않고도 성능 향상



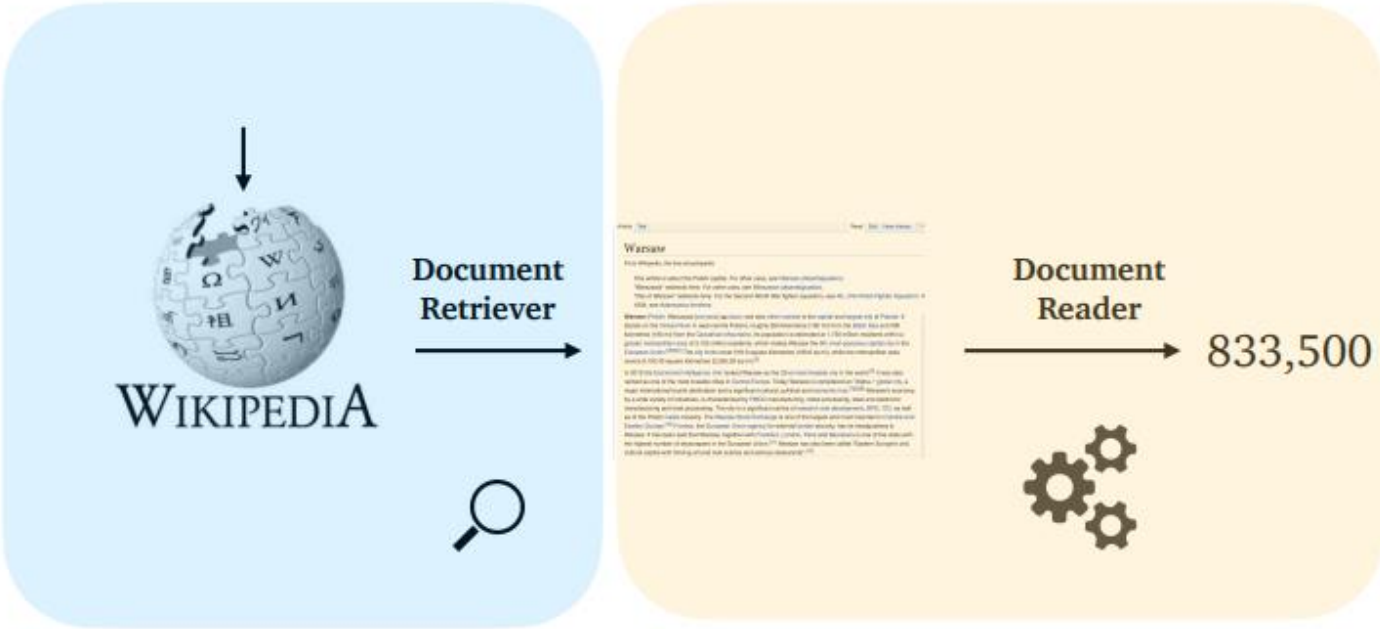
Unit 03

Open-domain question answering

open-domain QA

- 기존의 QA는 passage를 기반으로 답변
- open-domain QA는 large collection of document(Wikipedia)를 이용하여 답변

- retrieval text in question
- -> 질문과 관련있는 적은 수의 documents
- -> reader model
- -> answer



https://github.com/facebookresearch/DrQA

Retriever-reader framework

- Input: a large collection of documents $\mathscr{D} = D_1, D_2, \ldots, D_N$ and Q
- Output: an answer string A

- Retriever: $f(\mathscr{D}, Q) \longrightarrow P_1, \ldots, P_K$ K is pre-defined (e.g., 100)
- Reader: $g(Q, \{P_1, \ldots, P_K\}) \longrightarrow A$ A reading comprehension problem!

- (P : passage)
- retriever : TF-IDF information-retrieval sparse model
- reader : neural reading comprehensive model (trained on SQuAD and other QA datasets)

Reference

Stanford CS224N NLP with Deep Learning | Winter 2021 | Lecture 11 – Question Answering

https://www.youtube.com/watch?v=NcqfHa0_YmU&list=PLoROMvodv4rOSH4v6133s9LFPRHjEmbmJ

***All Images without clarified source are retrieved on the above reference.**

감사합니다