Question Answering

Wongyu Kim (김원규)

Internet Computing Laboratory
Department of Computer Science
Yonsei University

2020.09.10

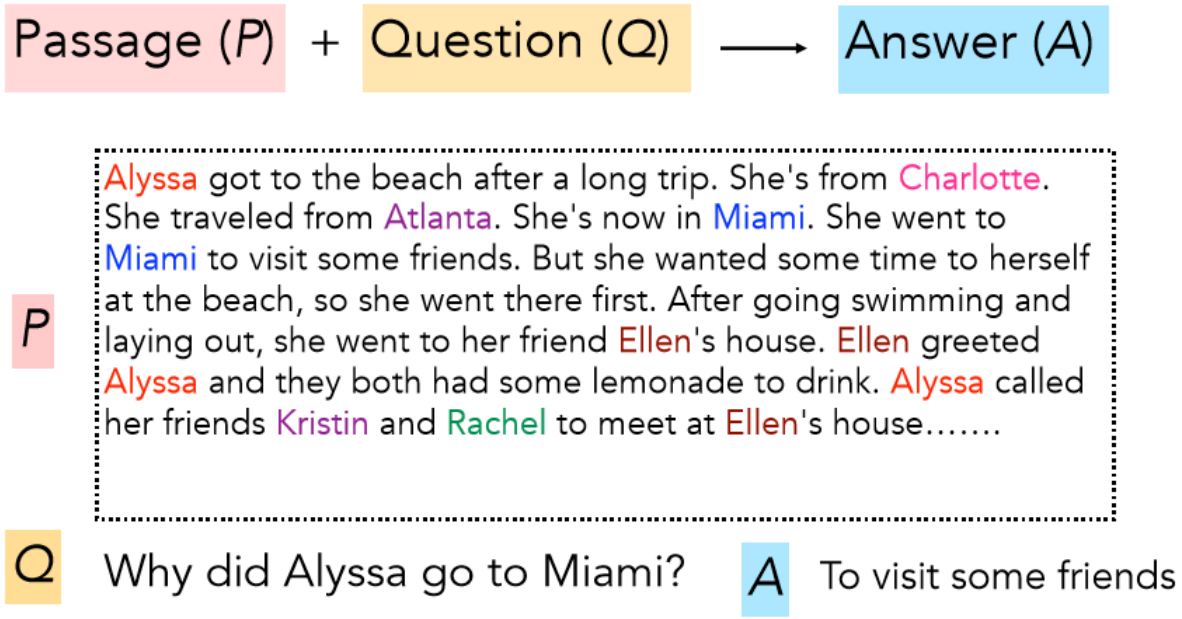
Question Answering

Technical note: This is a "featured snippet" answer extracted from a web page, not a question answered using the (structured) Google Knowledge Graph (formerly known as Freebase).

✓ 위 그림처럼 질문에 대한 답변을 하는 Task로
NLP vs Google KG 방법이 있다.
혹은 최근에 두 방법을 융합하여 사용하기도 한다.

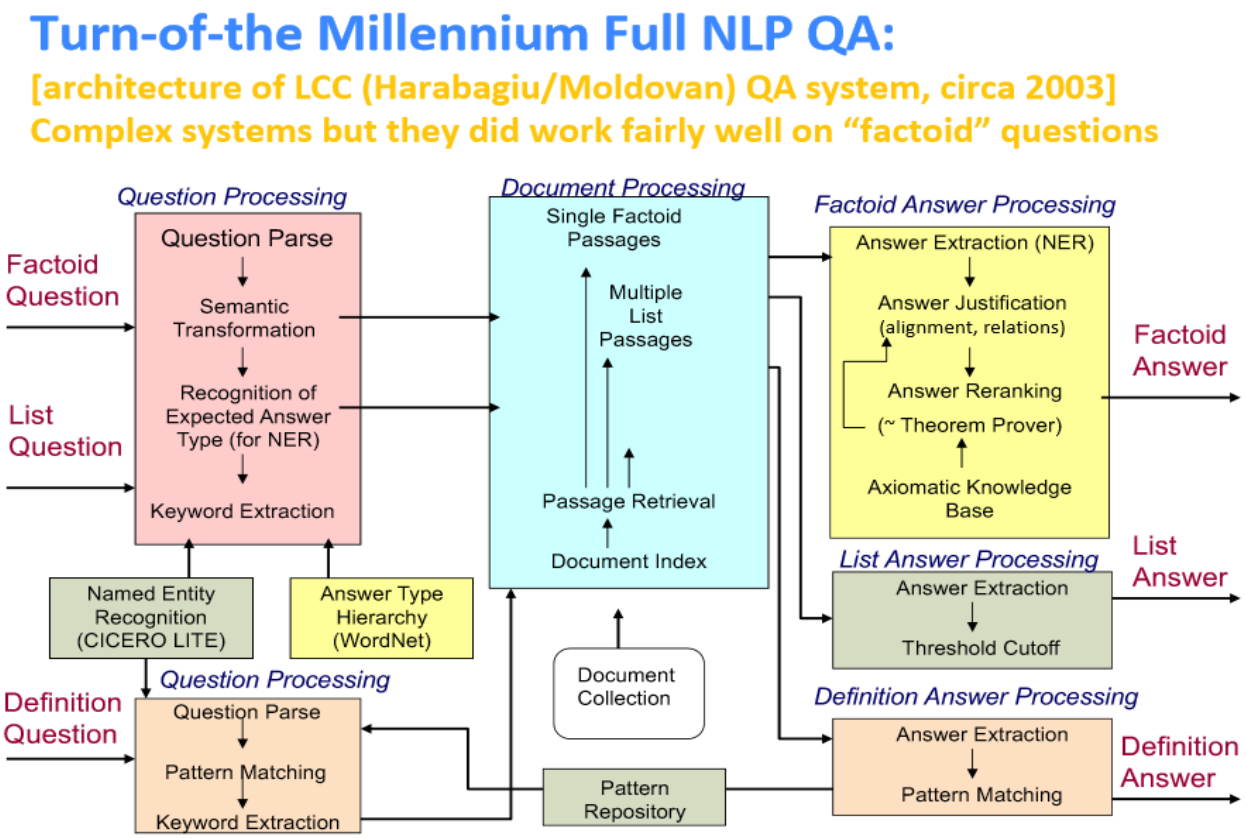
Question Answering

- ✓ QA에 대한 데이터셋은 아래와 같은 구조와 대개 많이 유사하다.
- ✓ 입력으로 Passage와 Question, 출력으로 Answer를 도출하는 구조이다.
- ✓ 아래의 데이터는 MCTest란 대회에서 사용하는 데이터이다.



Question Answering

✓ 아래 그림은 LCC의 QA 아키텍쳐(딥러닝 이전, 2003).
✓ 매우 복잡, 인간이 직접 설계한 많은 components의 결합으로 이루어짐.
✓ IBM의 DeepQA(2011)도 여러 components 를 ensemble, 아래보다 복잡.



SQuAD 1.0 ~ 1.1

- ✓ Stanford에서 만든, 현재까지 대표적인 QA 데이터셋.
- ✓ 각 Question은 Passage 한 개씩 가진다.
- ✓ Passage는 Wikipedia에 존재한다.
- ✓ 목표는 역시 answer를 도출하는 모델을 만드는 것.

- ✓ 100k esamples (대개 1 paragraph - 5 questions).
- ✓ Answer은 항상 paragraph의 span으로 존재 (a.k.a extractive QA)
  -> model can't answer yes/no, counting, implicit why questions
  (paragraph에 없는 말을 모델은 할 수가 없다.)

SQuAD 1.0 ~ 1.1

- ✓ 하지만 아래 그림처럼 정답은 어떻게 해석하는지에 따라 여러 개가 될 수 있다.
- ✓ 만약 gold answer가 1개라면 모델의 성능을 평가하기가 애매할 것이다.
- ✓ 따라서 답이 될 수 있는 경우를 3개로 두어 answer의 variation을 감안한다.
- ✓ 이를 통해 모델의 성능을 평가하기에 유용해진다.

Stanford Question Answering Dataset (SQuAD)

Private schools, also known as independent schools, non-governmental, or nonstate schools, are not administered by local, state or national governments; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, which makes the cost cheaper, depending on a talent the student may have (e.g. sport scholarship, art scholarship, academic scholarship), financial need, or tax credit scholarships that might be available.

Along with non-governmental and nonstate schools, what is another name for private schools?

Gold answers: ① independent ② independent schools ③ independent schools

Along with sport and art, what is a type of talent scholarship?

Gold answers: ① academic ② academic ③ academic

Rather than taxation, what are private schools largely funded by?

Gold answers: ① tuition ② charging their students tuition ③ tuition

Evaluation of SQuAD 1.0 ~ 1.1

- ✓ Exact Match: 모델이 예측한 span과 gold answers 중에 정확히 같은 것이 존재한다면 +1 / 아니라면 +0 을 한다.

- ✓ F1 Score: 위의 방법보다 더 continuous하고 융통적이라고 볼 수 있다.

  1. 예측 span과 gold answers를 BoW로 바꾼다.

  2. 그 후 (gold answer - span)의 3 pairs에 대해

  precision(예측 span가 실제로 gold answer에 있을 확률) &
  recall(gold answer가 실제로 예측 span에 있을 확률) -> F1 = 2PR/(P+R)을
  구하고 max 값 만을 선택한다.

  3. 모든 테스트 데이터에 대해 실행한 후 평균하면 final F1 Score를 구할 수 있다.

- ✓ 두 metric 모두 punctuation and articles(구두점과 정관사)는 무시한다.

Ranking of SQuAD 1.0 ~ 1.1

- ✓ 아래 그림은 SQuAD 1.1에 대한 모델들의 성능이다.
- ✓ Ensemble은 같은 방법으로 학습한 같은 구조의 여러 모델의 결과를 종합한 것이라 보면 된다.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835

SQuAD 2.0

- ✓ SQuAD 1.1은 답을 항상 찾으려는 경향이 있었다.
- ✓ 하지만 현실 세계에서는 답이 없는 경우가 많다.
- ✓ 따라서 Questions의 반은 Answer가 있고, 반은 없도록 설정했다.(valid & test)
  (training questions -> 1/3만 Answer 존재)

- ✓ 따라서 metric 계산 시 Exact Match든 F1 Score든 이전과 방법이 같지만
  No answer가 Gold일 경우 모델의 예측이 같은 No answer라면 +1, 아니라면 +0.

SQuAD 2.0 Example

Genghis Khan united the Mongol and Turkic tribes of the steppes and became Great Khan in 1206. He and his successors expanded the Mongol empire across Asia. Under the reign of Genghis' third son, Ögedei Khan, the Mongols destroyed the weakened Jin dynasty in 1234, conquering most of northern China. Ögedei offered his nephew Kublai a position in Xingzhou, Hebei. Kublai was unable to read Chinese but had several Han Chinese teachers attached to him since his early years by his mother Sorghaghtani. He sought the counsel of Chinese Buddhist and Confucian advisers. Möngke Khan succeeded Ögedei's son, Güyük, as Great Khan in 1251. He

When did Genghis Khan kill Great Khan?

Gold Answers: <No Answer>

Prediction: 1234 [from Microsoft nlnet]

SQuAD 2.0

- ✓ 당연히 1.1 버전보다는 어려우며 아직 인간을 넘지 못했다.
- ✓ 또한 아직 좋은 Model이라도 가끔은 이상한 결과를 보이고 있다. = Natural Language Understanding Errors

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	**85.082**	**87.615**
2 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967

Good systems are great, but still basic NLU errors

The Yuan dynasty is considered both a successor to the Mongol Empire and an imperial Chinese dynasty. It was the khanate ruled by the successors of Möngke Khan after the division of the Mongol Empire. In official Chinese histories, the Yuan dynasty bore the Mandate of Heaven, following the Song dynasty and preceding the Ming dynasty. The dynasty was established by Kublai Khan, yet he placed his grandfather Genghis Khan on the imperial records as the official founder of the

What dynasty came before the Yuan?

Gold Answers: ① Song dynasty ② Mongol Empire ③ the Song dynasty

Prediction: Ming dynasty [BERT (single model) (Google AI)]

Limitations of SQuAD

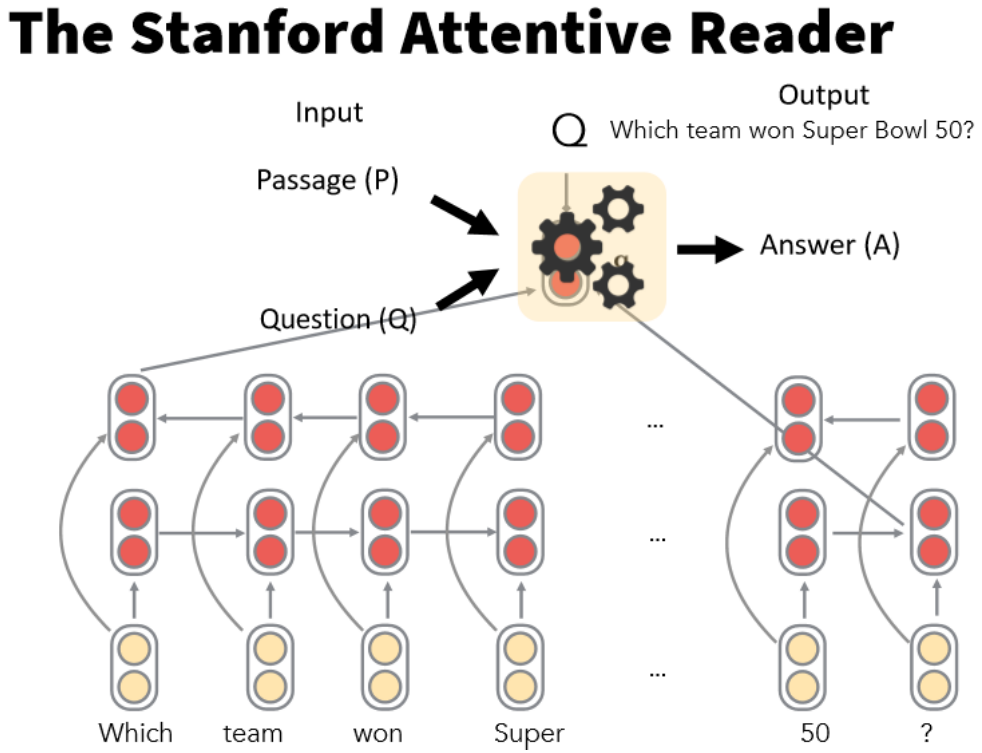
- ✓ 오직 span-based answers
- ✓ Question은 오직 그에 상응하는 Passage에서 파생되어야 한다. = 현실 세계에서는 필요 없는 질문일 수도 있다.
- ✓ 또한 현실 세계에서 얻는 것 보다 형식적, 언어적, 문법적이다.(Wikipedia의 영향)

- ✓ 그럼에도 불구하고 많이 쓰이고, 이 데이터셋으로 pre-train한 웨이트를 많이 사용하기도 한다.

DrQA(Chen 2016, 2017)

- ✓ The Stanford Attentive Reader라고도 한다.
- ✓ Bidirectional Encoding + Attention이 핵심적인 기술이다.

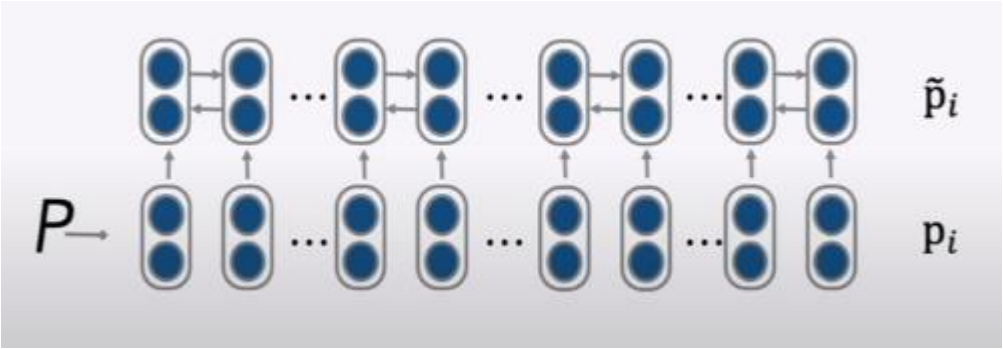
- ✓ 처음에는 Question의 각 단어를 GloVe로 임베딩한다.
- ✓ 그 후 BiLSTM으로 인코딩하여 question representation.



DrQA(Chen 2016, 2017)

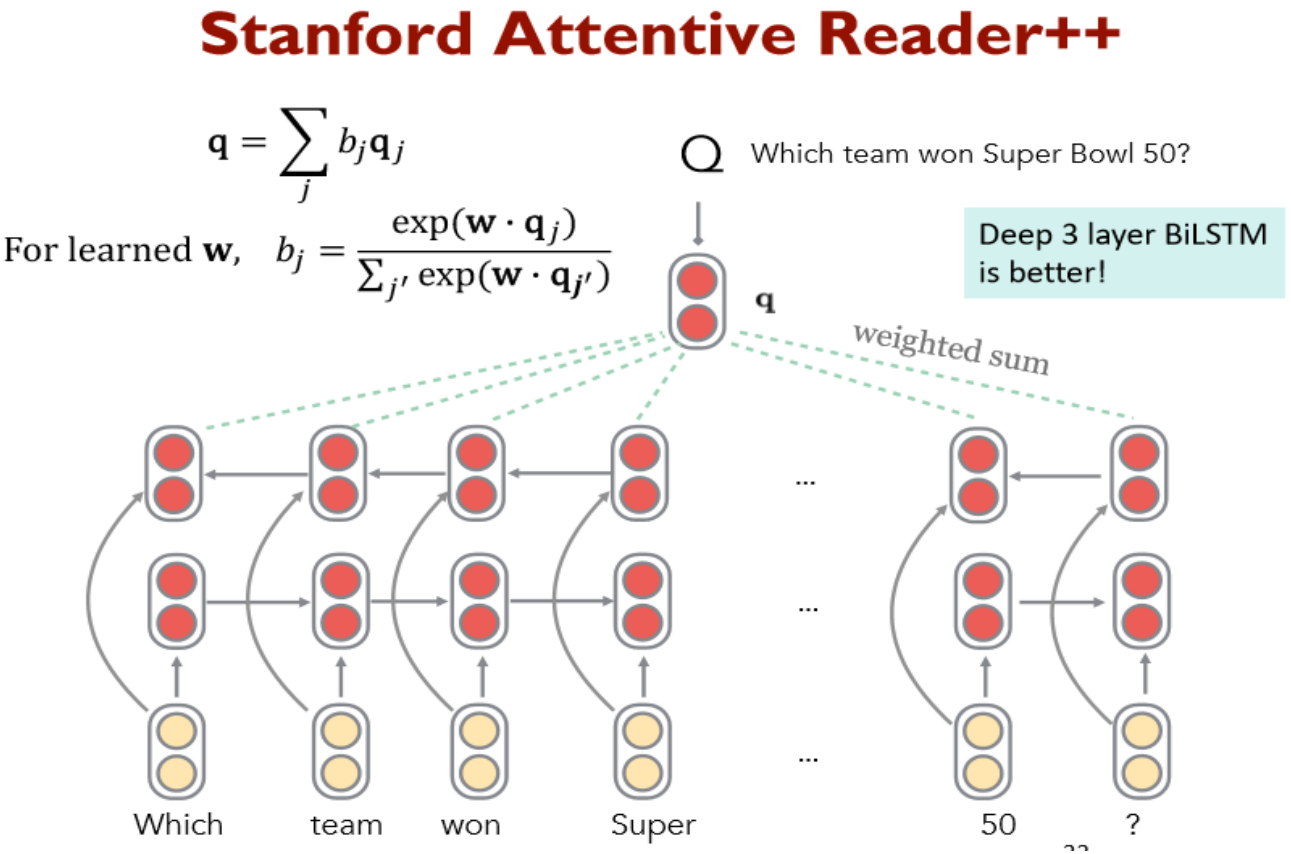
- ✓ 그 다음은 Passage의 인코딩 과정인데 여기서는 비슷하면서 약간 다르다.
- ✓ 역시 GloVe로 임베딩(pi)을 한 뒤 BiLSTM에 넣는다.
- ✓ 하지만 여기서는 양 끝에서만 representation을 가져오는 것이 아니라 각 token에서의 두 output을 concatenate한 결과(pi~)를 가지고 온다.

- ✓ 그 다음은 passage의 각 단어에 대한 attention weight 를 구하기 위해서
  ai = softmax(q * Ws * pi~) -> pi가 start token이 될 확률
  ai` = softmax(q * Ws` * pi~) -> pi가 end token이 될 확률
  최종 output vector -> ai * pi의 합이다. (& ai`*pi의 합)
  이것을 answer vector와 비교해서 학습



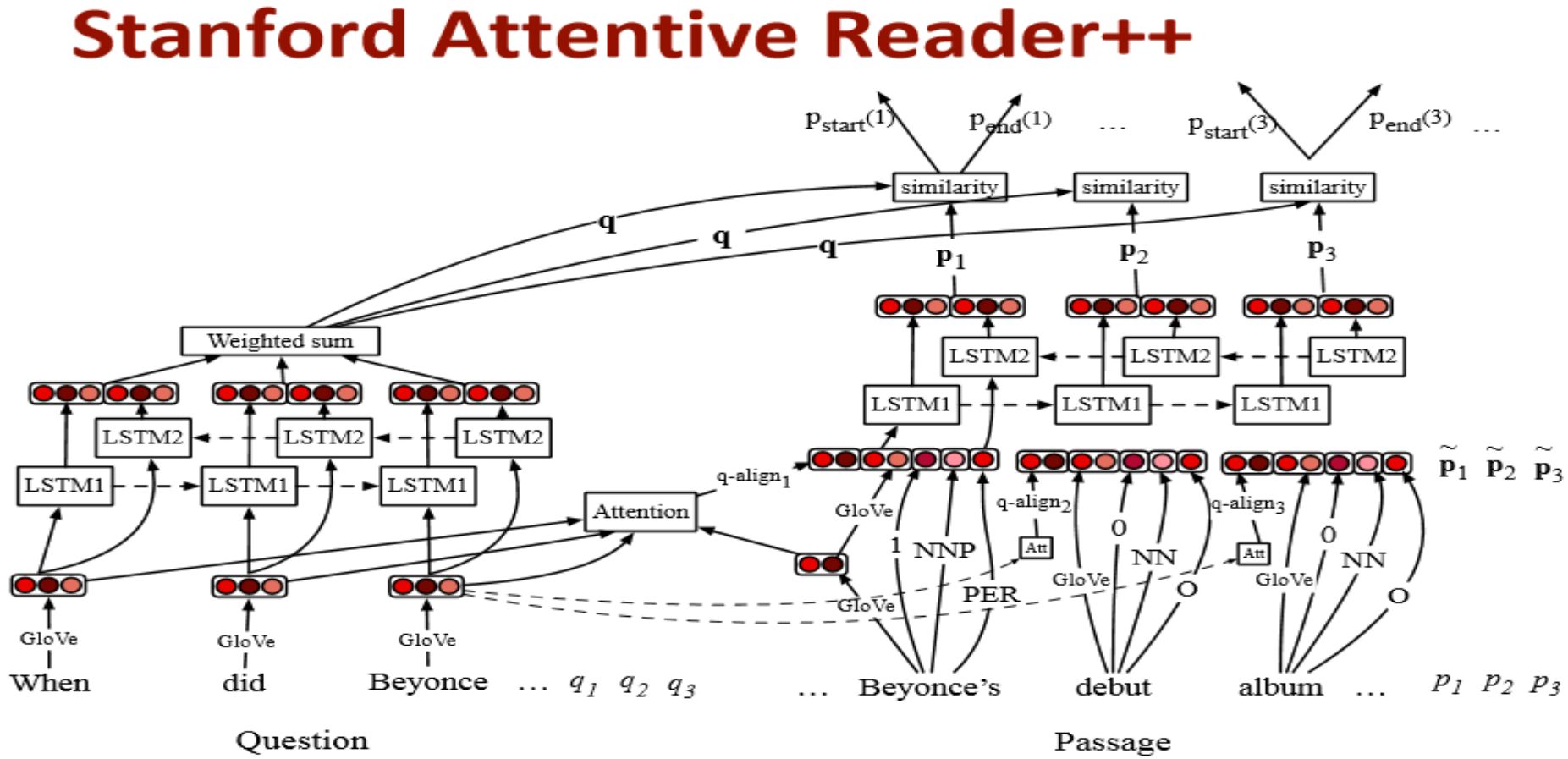
Stanford Attentive Reader++(2018)

- ✓ 기존 DrQA에서 몇가지 추가되었다 .
- ✓ 아래 그림처럼 인코딩 시 각 token에 대한 bj를 구하기 위해 w를 가져온다.
- ✓ 그 후 q(question representation)을 생성한다.



Stanford Attentive Reader++(2018)

✓ 모델 아키텍쳐도 바뀐다. 하지만 token 마다 start, end token이 될 확률이 나오는 건 똑같고 그에 따라 학습을 진행하면 된다. => 자세한건 논문



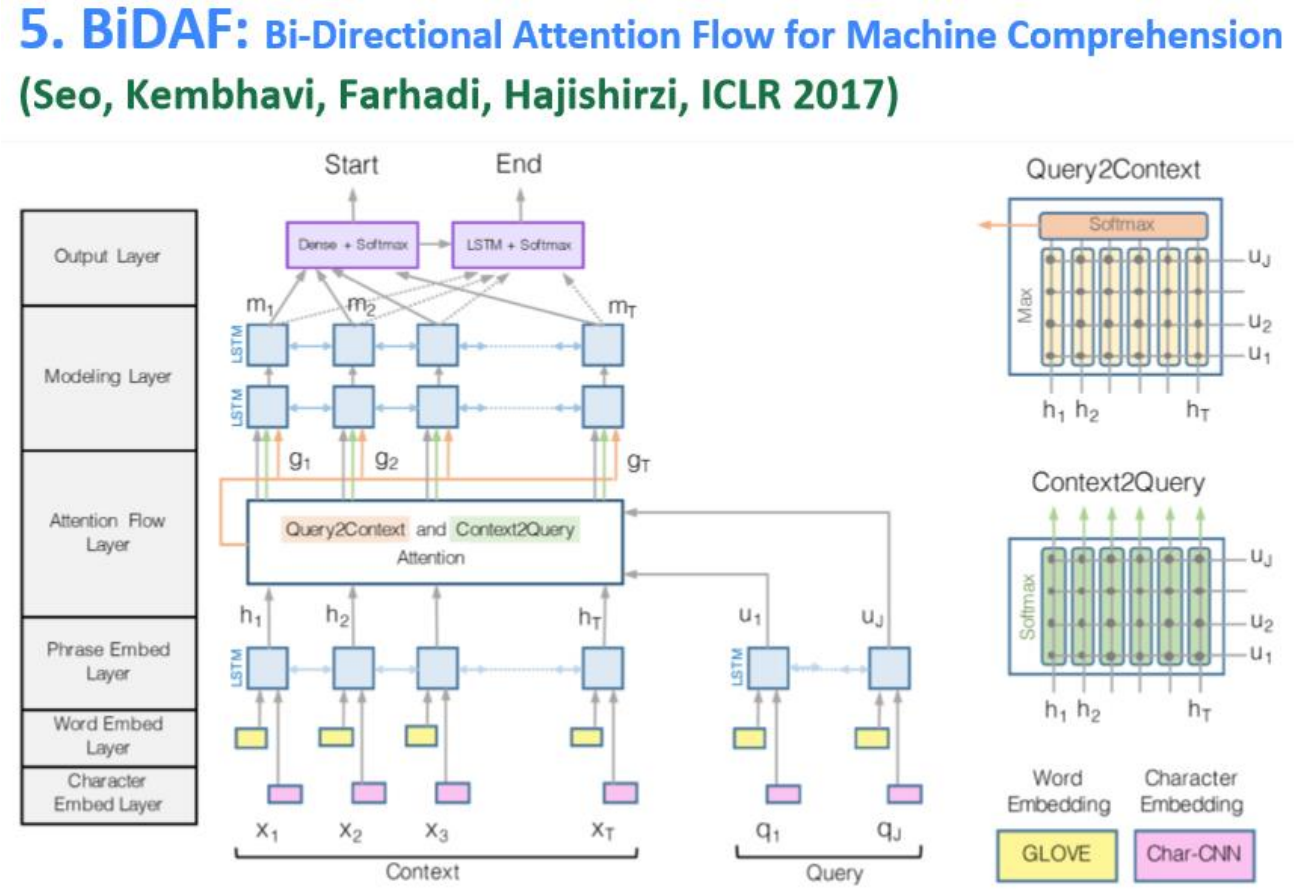
Training objective:

$$\mathcal{L} = -\sum \log P^{(\text{start})}(a_{\text{start}}) - \sum \log P^{(\text{end})}(a_{\text{end}})$$

32

BiDAF(Bi-Directional Attention Flow for Machine Comprehension, 2017)

- ✓ 이 당시 앞서 말했듯이 Attention을 어떻게 쓰는 것이 트렌드이자 핵심이였음.
- ✓ 여기서는 Query to Context 만이 아닌 Context to Query의 두 방향에 대한 Attention을 고려



BiDAF(Bi-Directional Attention Flow for Machine Comprehension, 2017)

1. Character Embed Layer: Char-CNN으로 d차원 임베딩
2. Word Embed Layer: GloVe로 d차원 임베딩
3. Contextual Embed Layer: 임베딩된 것으로 2d차원 인코딩 (Question과 Context의 모든 tokens에 대해)
4. 어텐션이 양방향으로 진행됨. (BiDAF인 이유)
   - Query2Context = Query가 있을 때 Context의 어느 정보가 관련 있는가 학습
   - Context2Query = Context가 있을 때 Query의 어느 정보가 관련 있는가 학습
   - 유사도 파악하기 위해 Similarity Mat 사용, Context의 t번째 단어, Query의 j번째 단어의 Similarity를 학습하게 된다.

BiDAF(Bi-Directional Attention Flow for Machine Comprehension, 2017)

1

$$\boldsymbol{S}_{ij} = \boldsymbol{w}_{\text{sim}}^T [\boldsymbol{c}_i; \boldsymbol{q}_j; \boldsymbol{c}_i \circ \boldsymbol{q}_j] \in \mathbb{R}$$

2

$$\boldsymbol{m}_i = \max_j \boldsymbol{S}_{ij} \in \mathbb{R} \quad \forall i \in \{1, \ldots, N\}$$

$$\beta = \text{softmax}(\boldsymbol{m}) \in \mathbb{R}^N$$

$$\boldsymbol{c}' = \sum_{i=1}^{N} \beta_i \boldsymbol{c}_i \in \mathbb{R}^{2h}$$

3

$$\alpha^i = \text{softmax}(\boldsymbol{S}_{i,:}) \in \mathbb{R}^M \quad \forall i \in \{1, \ldots, N\}$$

$$\boldsymbol{a}_i = \sum_{j=1}^{M} \alpha_j^i \boldsymbol{q}_j \in \mathbb{R}^{2h} \quad \forall i \in \{1, \ldots, N\}$$

4

$$\boldsymbol{b}_i = [\boldsymbol{c}_i; \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{c}'] \in \mathbb{R}^{8h}$$

BiDAF(Bi-Directional Attention Flow for Machine Comprehension, 2017)

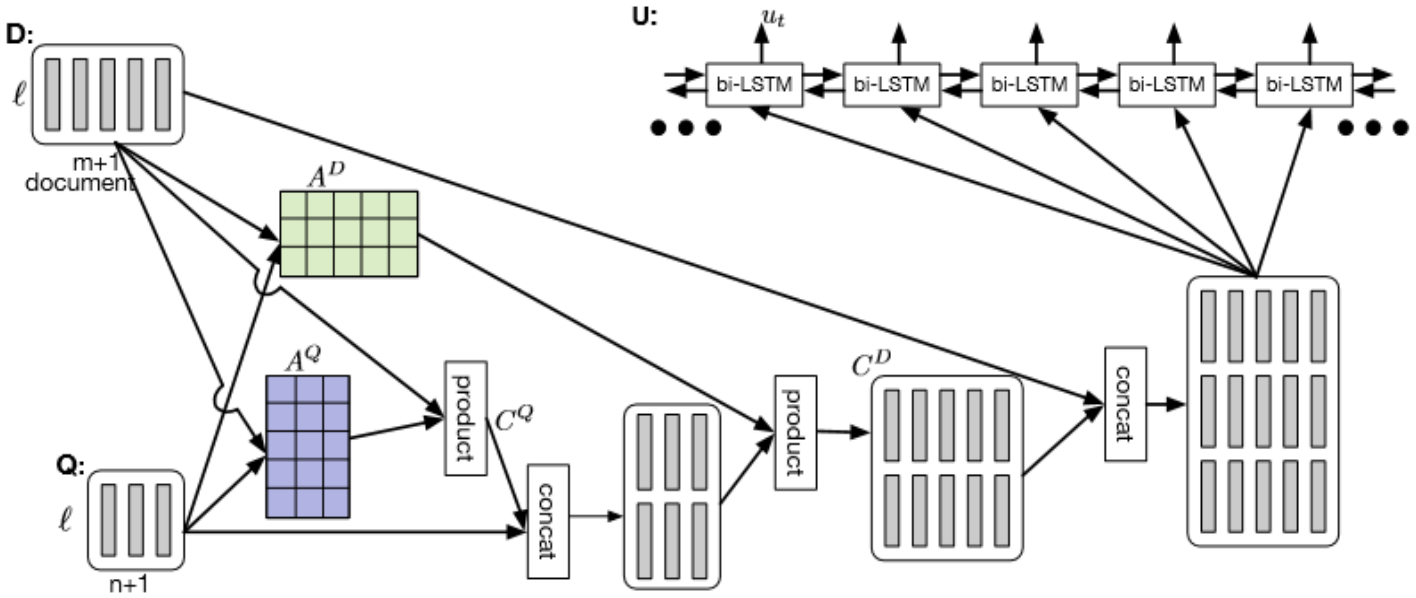
5. Modeling Layer: 정보 정리 단계, 4단계의 출력을 다시 BiLSTM에 넣는다. (GRU)

6. Output Layer: 각 token이 start & end일 확률을 구하는 Layer

Others

- ✓ Coattention Encoder (ICLR, 2017)
- ✓ Two-way attention - 자세한건 논문

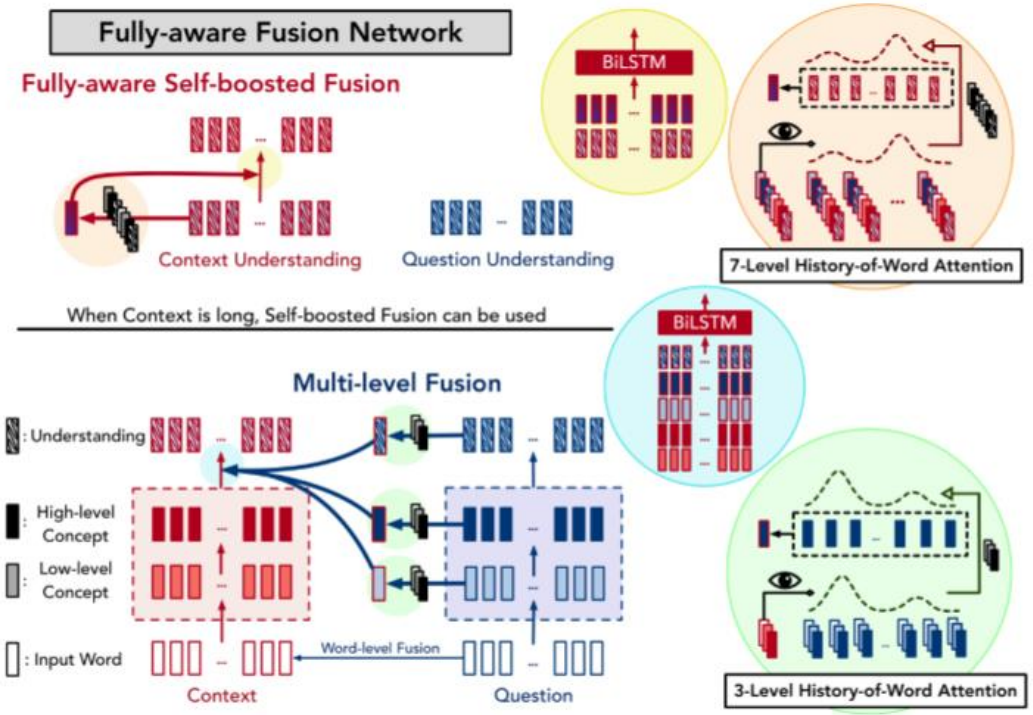
Coattention Encoder



Others

- ✓ FusionNet
- ✓ 다양한 어텐션을 시도하고 결합함.
- ✓ Multi-level inter-attention 비스무리 이용

FusionNet tries to combine many forms of attention



Others

- ✓ 후에 Transformer 등 여러 방법 이용해서 Contextual Word Representations 더 발전시켜 QA 문제 해결에 도움 주었다.
- ✓ 또한 Document Retriever를 이용해 Question이 주어졌을 때 그에 관련된 Document를 찾는 task 도 있다.

참고 문헌

한글 블로그 자료

- https://sumniya.tistory.com/26
- https://ratsgo.github.io/natural%20language%20processing/2017/10/22/manning/Cs224n-winter 2019-lecture 7

https://www.youtube.com/watch?v=QEw0qEa0E50&list=PLoROMvodv4rOhcuXMZkNm7j3fVwBBY42z&index=10

CS224n-winter 2019-syllabus-lecture 7

https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/

Questions & Answers

Thank you