Text Understanding with the Attention Sum Reader Network

NLP 스터디

김승유

목차

- 1. cloze test
- 2. 데이터셋
- 3. Attention Sum Reader 구조
- 4. 결과

Text Understanding with the Attention Sum Reader Network

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar & Jan Kleindienst
IBM Watson
V Parku 4, Prague, Czech Republic
{rudolf_kadlec,martin.schmid,obajgar,jankle}@cz.ibm.com

우리가 아는 Attention과 조금 다르다

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In Advances in Neural Information Processing Systems, pages 2674-2682.

Abstract

Several large cloze-style context-question-answer datasets have been introduced recently: the CNN and Daily Mail news data and the Children's Book Test. Thanks to the size of these datasets, the associated text comprehension task is well suited for deep-learning techniques that currently seem to outperform all alternative approaches. We present a new, simple model that uses attention to directly pick the answer from the context as opposed to computing the answer using a blended representation of words in the document as is usual in similar models. This makes the model particularly suitable for question-answering problems where the answer is a single word from the document. Ensemble of our models sets new state of the art on all evaluated datasets.

Document: What was supposed to be a fantasy sports car ride at Walt Disney World Speedway turned deadly when a Lamborghini crashed into a guardrail. The crash took place Sunday at the Exotic Driving Experience, which bills itself as a chance to drive your dream car on a racetrack. The Lamborghini's passenger, 36-year-old Gary Terry of Davenport, Florida, died at the scene, Florida Highway Patrol said. The driver of the Lamborghini, 24-year-old Tavon Watson of Kissimmee, Florida, lost control of the vehicle, the Highway Patrol said. (...)
Question: Officials say the driver, 24-year-old Tavon Watson, lost control of a _______
Answer candidates: Tavon Watson, Walt Disney World Speedway, Highway Patrol, Lamborghini, Florida, (...)
Answer: Lamborghini

Figure 1: Each example consists of a context document, question, answer cadidates and, in the training data, the correct answer. This example was taken from the CNN dataset (Hermann et al., 2015). Anonymization of this example that makes the task harder is shown in Figure 3.

cloze test란

CLOZE TEST

For ***Lawrence*** ***United*** ***Washington*** ***cut*** ***father*** ***he*** ***his***
of ***really*** ***slaves*** ***very***

George Washington was the first President of the United States. He was also the commander in chief of all American forces during the American Revolutionary War. For his central role in the beginning of the United States, he is often called the father of ______ country. His mother was Mary Ball and his ______ was Augustine Washington. They owned a plantation with ______ in Virginia. George studied at local schools. George's ______ died when he was eleven. Then his brother ______ helped train him. There is a story that ______ cut down his father's cherry tree. When asked, ______ did not lie and said that he did ______ down the tree. The story means he was ______ honest. We do not know if the story ______ happened.

SCORE:
3/14

?

빈칸 채우기 문제

왜 cloze test냐?

- 지금까지의 정보는 텍스트로 저장되었고, 이를 인공지능한테 이해시켜야 함.
- 텍스트 이해 수준을 평가하는 방법 중 하나는 텍스트에서 답을 추론할 수 있는 질문을 하는 것이며 따라서 cloze test를 활용한다.
- 데이터 구하기가 쉬우며 난이도 조절도 쉽다. (전치사 ↔ 고유명사)

데이터셋 설명

- 질문, 질문의 본문, 정답, 보기 → (q, d, a, A)

- 이 때, a는 보기 A와 본문 d의 원소

- 고유명사(named entities)들은 익명 토큰으로 대체 ex) @entity5

→ 특정 개체에 대한 실제 세계의 지식을 축적하지 않고,
전적으로 문맥에 의존시키기 위함

데이터셋

- 1. CNN과 Daily Mail의 기사들
 - 본문: 기사 본문
 - 질문: 하이라이트 문장(제목, 소제목)을 cloze 식으로 변형

- 2. Children's book test
 - 본문: 어린이 책 이야기에서의 20개의 연속된 문장
 - 질문: 요약이 없기 때문에, 21번째 문장 이용
 - 표준 LSTM 언어 모델이 동사와 전치사의 경우 사람과 비슷한 성능을 내지만 고유명사과 일반명사는 그렇지 않기 때문에 이 두가지에 대해서만 진행

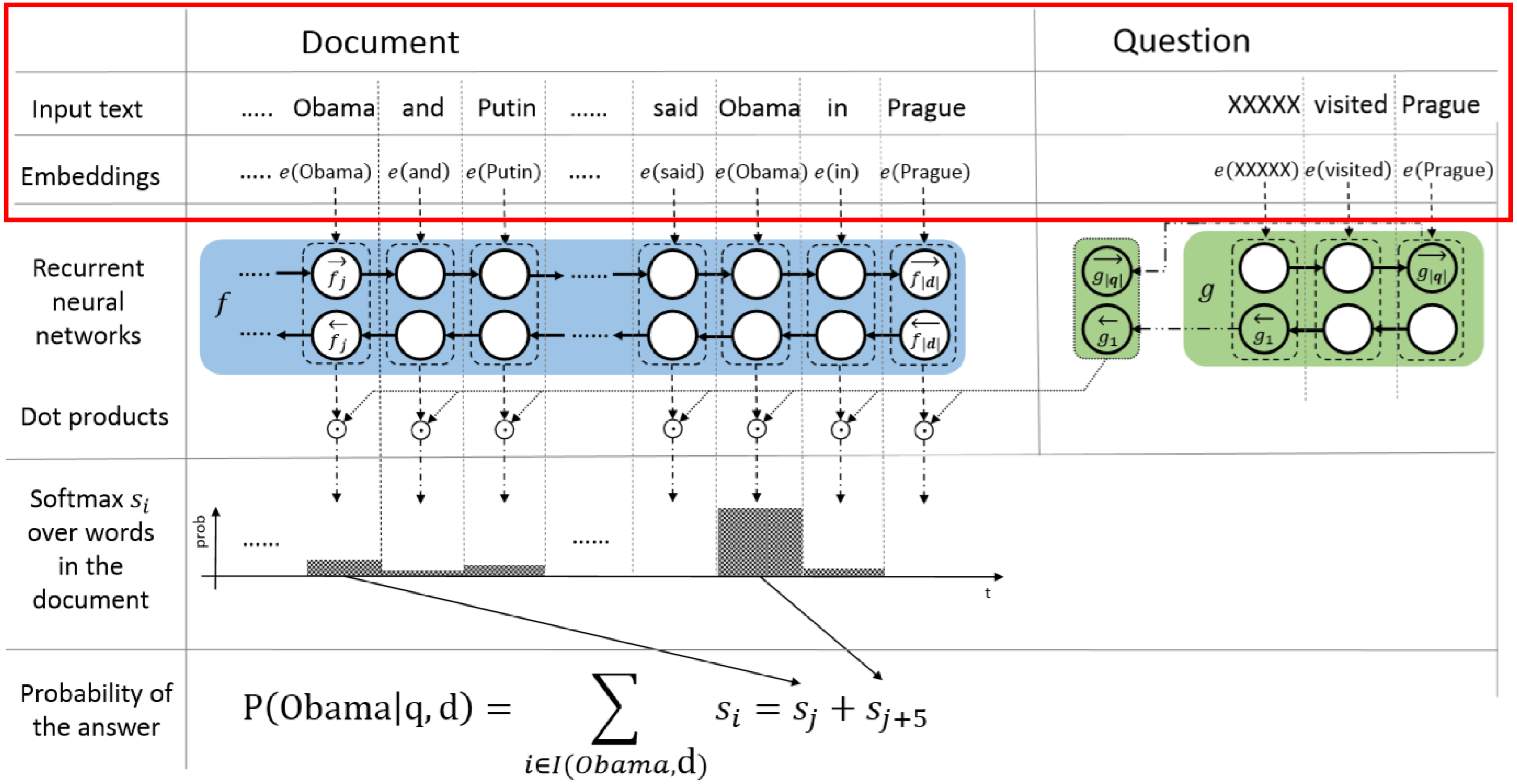
모델 설명

- Attention Sum Reader가 모델 이름
- 정답이 본문에서 나온 단어라는 가정을 전제로 설계됨
  (즉, 본문에 포함되지 않는 답변은 만들 수 없음)
- 가정이 성립하는 경우에는 좋은 성능을 낸다.
- 구조 요약: (1개의 임베딩 레이어 e, 2개의 인코더 g, f)

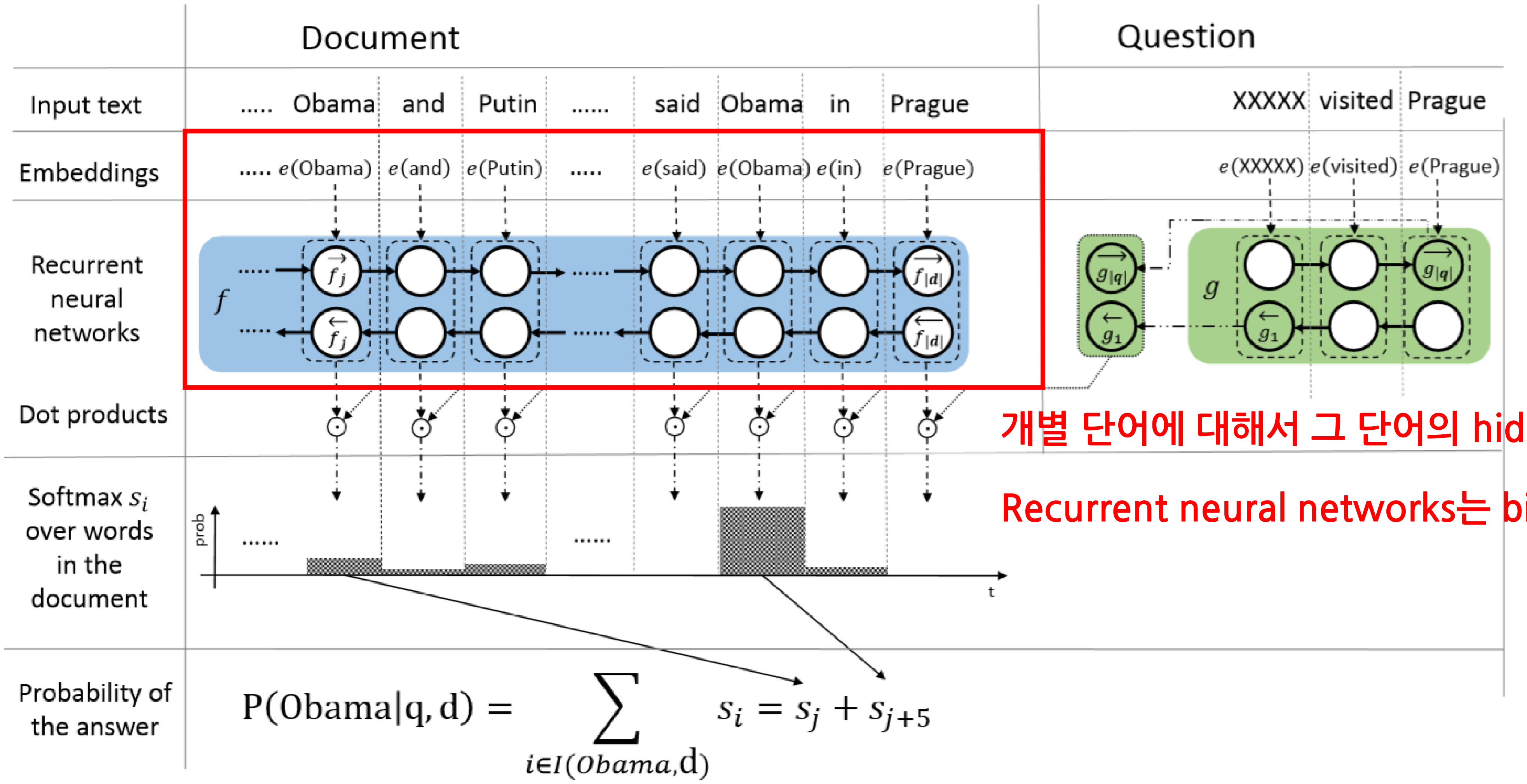
1. 질문을 인코더 g에 통과
2. 본문에서 개별 단어들에 대해 인코더 f 통과
3. 질문과 개별 단어의 내적 후 가장 가능성이 높은 답 선택

모델 구조

One hot 벡터로 표기된 단어들 임베딩



모델 구조

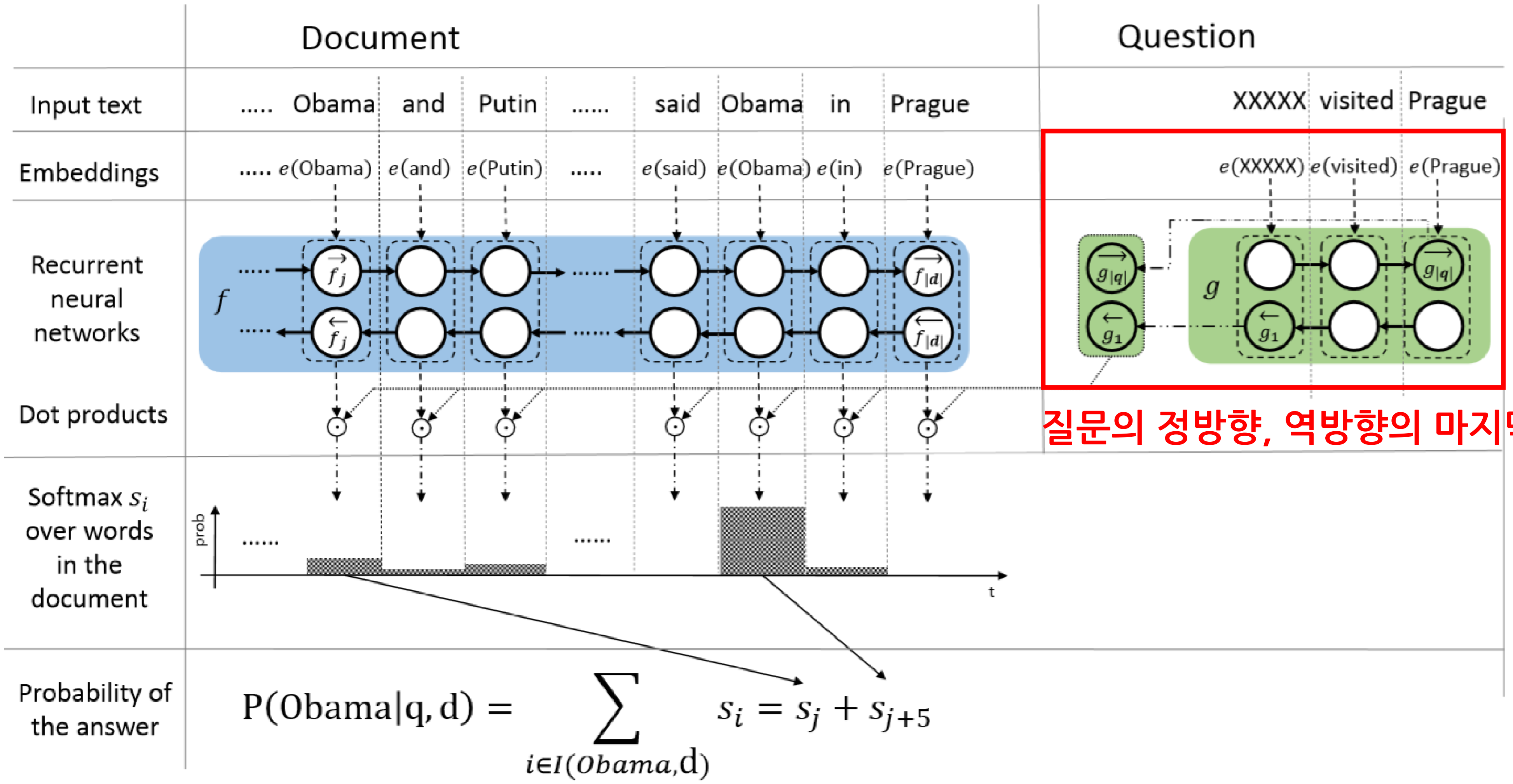


개별 단어에 대해서 그 단어의 hidden state를 concat함

Recurrent neural networks는 bi_GRU이용

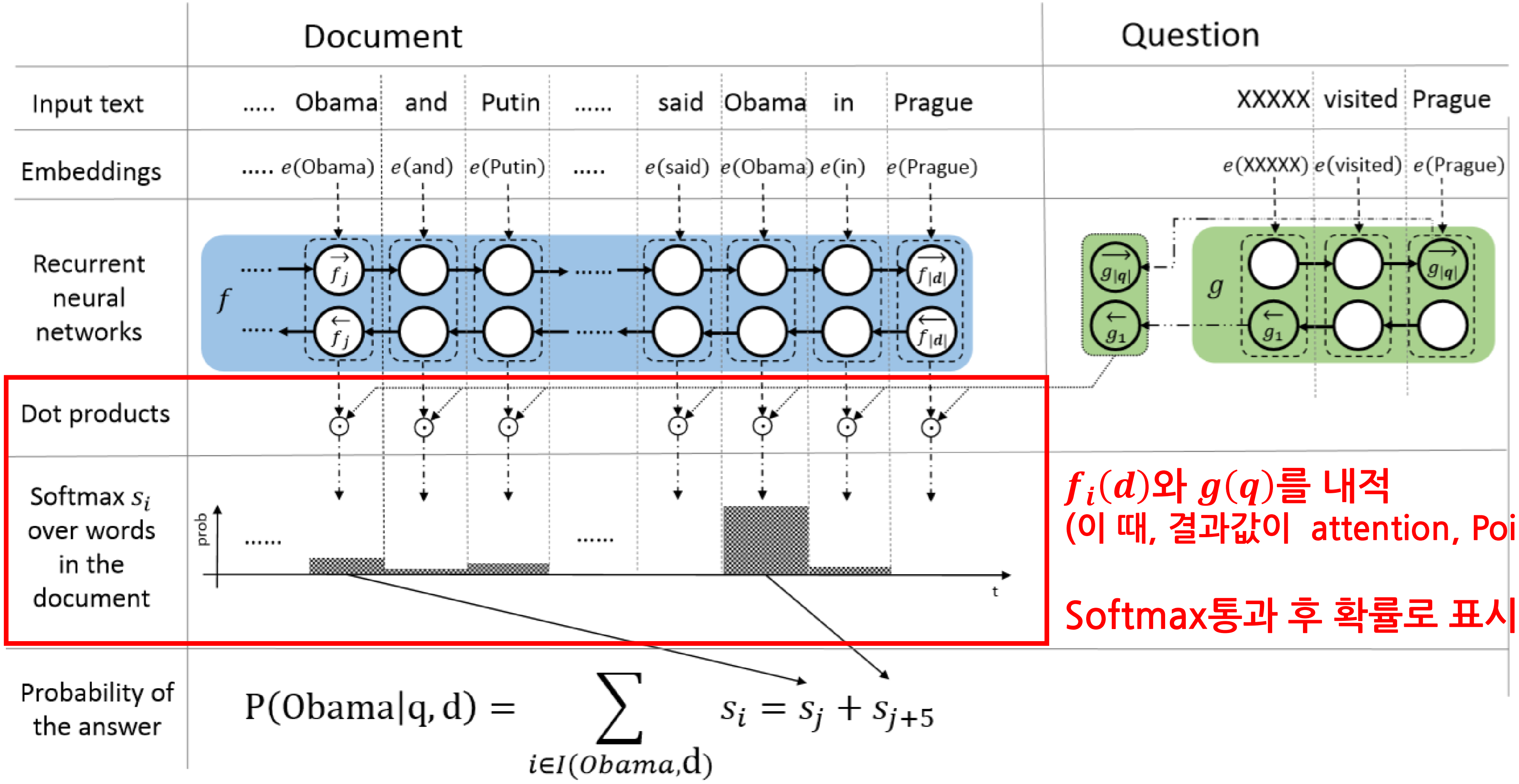
ybigta
연세대학교 빅데이터 학회

모델 구조



질문의 정방향, 역방향의 마지막 hidden state를 concat함

모델 구조



$f_i(d)$와 $g(q)$를 내적

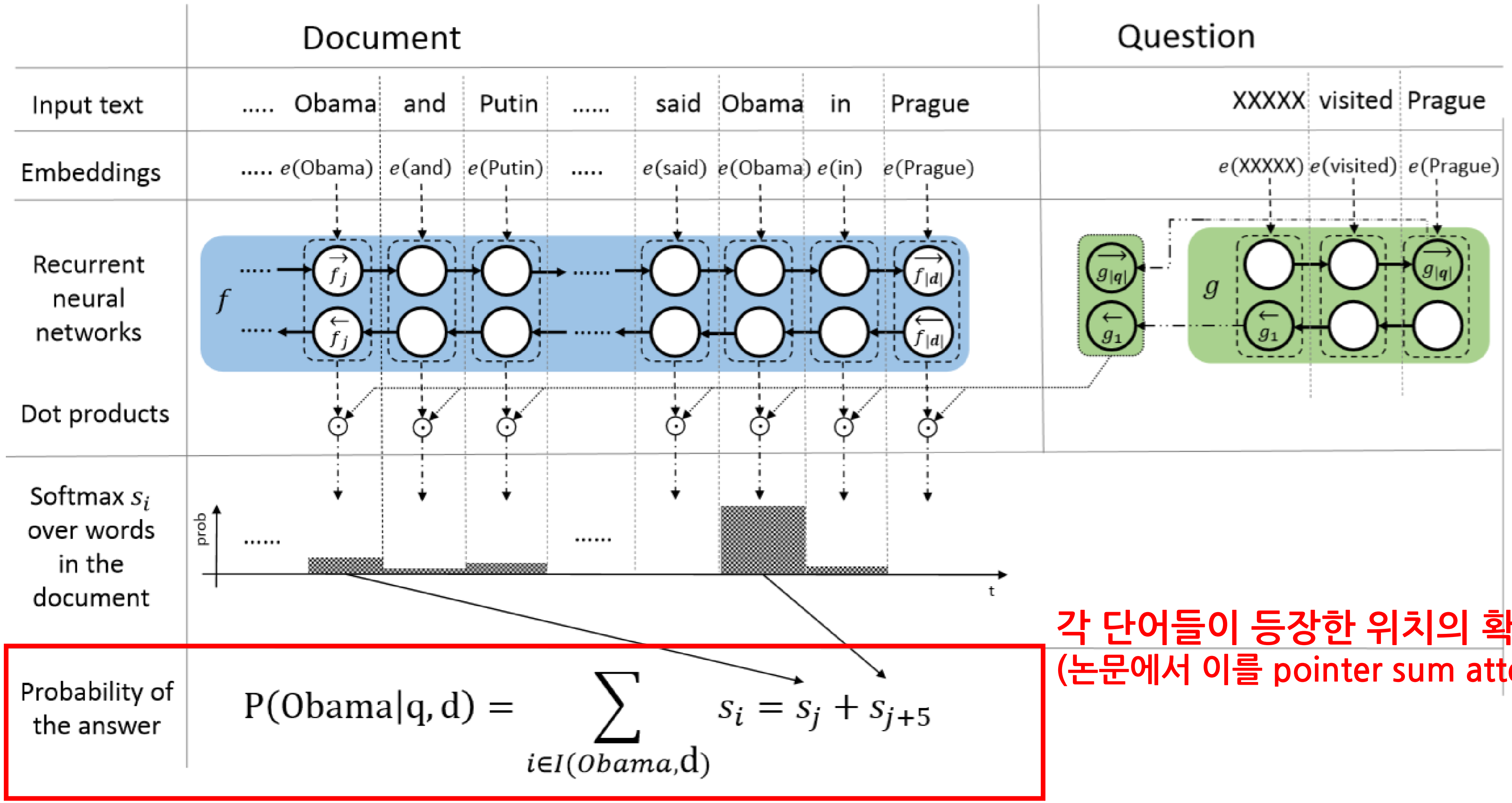
(이 때, 결과값이 attention, Pointer Network의 attentnion)

Softmax통과 후 확률로 표시

ybigta

연세대학교 빅데이터 학회

모델 구조



각 단어들이 등장한 위치의 확률을 모두 더해 줌
(논문에서 이를 pointer sum attention 메커니즘 이라고 명칭)

연세대학교 빅데이터 학회

모델 학습

- Adam optimizer이용
- $-\log P_\theta(a|q,d)$를 최소화 하도록 학습
- 다른 모델과 비교하기 위해 pre-train된 word embedding 사용하지 않음
- 이미 tokenized된 데이터셋이기 때문에 특별한 전처리하지 않음 등…

ybigta
연세대학교 빅데이터 학회

결과

	CNN		Daily Mail	
	valid	test	valid	test
Attentive Reader †	61.6	63.0	70.5	69.0
Impatient Reader †	61.8	63.8	69.0	68.0
MemNNs (single model) ‡	63.4	66.8	NA	NA
MemNNs (ensemble) ‡	66.2	69.4	NA	NA
Dynamic Entity Repres. (max-pool) ♯	71.2	70.7	NA	NA
Dynamic Entity Repres. (max-pool + byway)♯	70.8	72.0	NA	NA
Dynamic Entity Repres. + w2v ♯	71.3	72.9	NA	NA
Chen et al. (2016) (single model)	72.4	72.4	76.9	75.8
AS Reader (single model)	68.6	69.5	75.0	73.9
AS Reader (avg for top 20%)	68.4	69.9	74.5	73.5
AS Reader (avg ensemble)	73.9	**75.4**	78.1	77.1
AS Reader (greedy ensemble)	74.5	74.8	78.7	**77.7**

Table 2: Results of our AS Reader on the CNN and Daily Mail datasets. Results for models marked with † are taken from (Hermann et al., 2015), results of models marked with ‡ are taken from (Hill et al., 2015) and results marked with ♯ are taken from (Kobayashi et al., 2016). Performance of ‡ and ♯ models was evaluated only on CNN dataset.

	Named entity		Common noun	
	valid	test	valid	test
Humans (query) (∗)	NA	52.0	NA	64.4
Humans (context+query) (∗)	NA	***81.6***	NA	***81.6***
LSTMs (context+query) ‡	51.2	41.8	62.6	56.0
MemNNs (window memory + self-sup.) ‡	70.4	66.6	64.2	63.0
AS Reader (single model)	73.8	68.6	68.8	63.4
AS Reader (avg for top 20%)	73.3	68.4	67.7	63.2
AS Reader (avg ensemble)	74.5	70.6	71.1	**68.9**
AS Reader (greedy ensemble)	76.2	**71.0**	72.4	67.5

Table 3: Results of our AS Reader on the CBT datasets. Results marked with ‡ are taken from (Hill et al., 2015). (∗)Human results were collected on 10% of the test set.

연세대학교 빅데이터 학회

예시

...
what was supposed to be a fantasy sports car ride at @entity3 turned deadly when a @entity4 crashed into a guardrail . the crash took place sunday at the @entity8 , which bills itself as a chance to drive your dream car on a racetrack . the @entity4 's passenger , 36 - year - old @entity14 of @entity15 , @entity16 , died at the scene , @entity13 said . the driver of the @entity4 , 24 - year - old @entity18 of @entity19 , @entity16 , lost control of the vehicle , the @entity13 said .
...

0 1

officials say the driver , 24 - year - old @entity18 , lost control of a _____

Figure 3: Attention in an example with anonymized entities where our system selected the correct answer. Note that the attention is focused only on named entities.

...
@entity11 film critic @entity29 writes in his review that "anyone nostalgic for childhood dreams of transformation will find something to enjoy in an uplifting movie that invests warm sentiment in universal themes of loss and resilience , experience and maturity . " more : the best and worst adaptations of "@entity" @entity43, @entity44 and @entity46 star in director @entity48's crime film about a hit man trying to save his estranged son from a revenge plot. @entity11 chief film critic @entity52 writes in his review that the film
...

0 1

_____ stars in crime film about hit man trying to save his estranged son

Figure 4: Attention over an example where our system failed to select the correct answer (entity43). The system was probably mislead by the co-occurring word 'film'. Namely, entity11 occurs 7 times in the whole document and 6 times it is together with the word 'film'. On the other hand, the correct answer occurs only 3 times in total and only once together with 'film'.

결론

- 자연어 이해를 위한 새로운 신경망 구조 제시
- 모델은 단순하지만, 성능은 가장 좋음(state-of-the-art)
- (Chen et al., 2016)에서 CNN과 Daily Mail 데이터셋이 인간에게도 모호하거나 어렵다고 분석하고 있기 때문에 데이터셋에서 최대 성능에 근접했다고 할 수 있음.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A Thorough Examination of the CNN / Daily Mail Reading Comprehension Task. In Association for Computational Linguistics (ACL).

참고 문헌

- Kadlec, R., Schmid, M., Bajgar, O., & Kleindienst, J. (2016). Text understanding with the attention sum reader network. *arXiv preprint arXiv:1603.01547*.

ybigta
연세대학교 빅데이터 학회

감사합니다