BERT의 웹 문서 질의 응답 성능 향상을 위한 HTML 태그 스택 및 HTML 임베딩 기법 설계

목진왕 [1], 이현섭 [2‡]
[1] 백석대학교 컴퓨터공학부
[2] 백석대학교 첨단 IT 학부
‡*교신저자*
wlsdhkd8040@bu.ac.kr, hyunseob@bu.ac.kr

A Design of HTML Tag Stack and HTML Embedding Method to Improve Web Document Question Answering Performance of BERT

Jin-Wang Mok[1], Hyun-Seob Lee[2]
[1]Div. of Computer Engineering, Baek-Seok University
[2]Div. of Advanced IT, Baek-Seok University

요 약

최근 기술의 발전으로 인해 자연어 처리 모델의 성능이 증가하고 있다. 그에 따라 평문 지문이 아닌 KorQuAD 2.0 과 같은 웹 문서를 지문으로 하는 기계 독해 과제를 해결하려는 연구가 증가하고 있다. 최근 기계 독해 과제의 대부분의 모델은 트랜스포머를 기반으로 하는 추세를 보인다. 그 중 대표적인 모델인 BERT 는 문자열의 순서에 대한 정보를 임베딩 과정에서 전달받는다. 한편 웹 문서는 태그 구조가 존재하므로 문서를 이해하는데 위치 정보 외에도 태그 정보도 유용하게 사용될 수 있다. 그러나 BERT 의 기존 임베딩은 웹 문서의 태그 정보를 추가적으로 모델에 전달하지 않는다는 문제가 있었다. 본 논문에서는 BERT 에 웹 문서 태그 정보를 효과적으로 전달할 수 있는 HTML 임베딩 기법 및 이를 위한 전처리 기법으로 HTML 태그 스택을 소개한다. HTML 태그 스택은 HTML 태그의 정보들을 추출할 수 있고 HTML 임베딩 기법은 이 정보들을 BERT 의 임베딩 과정에 입력으로 추가함으로써 웹 문서 질의 응답 과제의 성능 향상을 기대할 수 있다.

1. 서론

기계 독해(Machine Reading Comprehension)는 대표적인 자연어 처리 분야 중 하나로 질의와 지문을 입력으로 받아 응답에 해당하는 범위를 모델이 예측하는 과제이다. 최근 기술의 발전에 따라 자연어 처리 모델이 고도화되었고 그에 따라 평문 형식의 지문이 아닌 웹 문서를 지문으로 하여 기계 독해 과제를 해결하려는 연구가 진행되고 있다. 대표적인 웹 문서 기계 독해 데이터셋으로는 SQuAD 2.0[1]과 KorQuAD 2.0[2] 등이 있으며 이 중 KorQuAD 2.0 은 기계 독해 과제 해결을 위한 한국어 웹 문서 데이터셋이다.

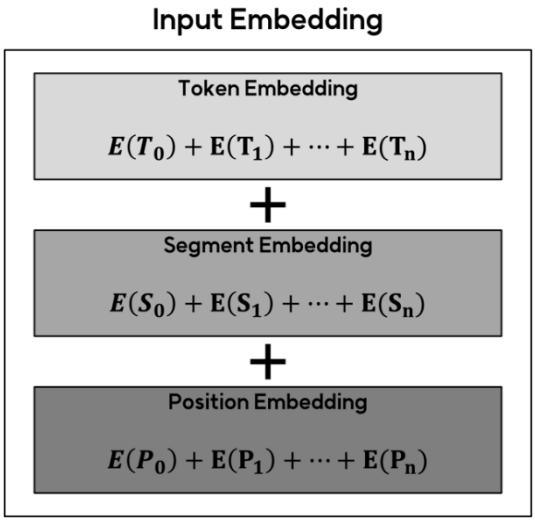
한편 기계 독해 과제에서 좋은 성능을 보이는 자연어 처리 모델들은 대부분 자기주의(Self-Attention)[3] 구조를 활용한 트랜스포머[4] 기반의 모델들이다. 대표적인 트랜스포머 기반의 자연어 처리 모델로는 BERT[5]와 XL-NET[6] 등이 있다. BERT 는 RNN 기반의 모델들과 달리 순차적인 위치 정보를 가질 수 없기 때문에 모델에 위치 정보를 추가하는 구조로 되어 있다. 위치 정보를 모델에 전달하는 방식 중 절대 위치 임베딩은 임베딩 과정에서 주기 함수나 순차적인 정숫값을 이용하여 고정적인 위치 정보를 모델에 전달하는 방식이다. 그러나 평문 지문과 달리 웹 문서는 태그 구조를 가지고 있어 BERT 에 순차적인 위치 정보 뿐만 아니라 태그의 이름이나 태그의 위치와 같은 추가적인 정보를 전달할 수 있다. 따라서 본 논문에서는 웹 문서를 지문으로 하는 기계 독해 과제에서 BERT 에 적용하여 기존 임베딩보다 정확도를 향상시킬 수 있는 HTML 임베딩 기법과 이를 위한 전처리 기법으로 HTML 태그 스택을 소개한다.

2. 관련연구

2.1 트랜스포머(Transformer)와 위치 임베딩

트랜스포머는 자기주의(Self-Attention) 구조를 활용한 기계 번역 모델이다. 트랜스포머 이전의 기계 번역 모델은 LSTM[7]과 같은 RNN 기반의 모델을 사용했다. 이에 따라 순차적인 위치 정보를 재귀적인 구조로 인해 모델이 자연스럽게 학습할 수 있었다. 그러나 트랜스포머를 포함한 이후의 모델들은 순차적인 위치 정보를 인위적으로 생성하여 모델에 전달하는 구조로 발전했다. 트랜스포머의 경우 주기함수 기반의 절대 위치 임베딩을 사용했다. 절대 위치 임베딩은 토큰의 순차적인 위치 정보를 모델에 전달하는 것으로 주기함수 또는 0 부터 최대 길이까지 1 만큼씩 증가하는 정숫값 등을 기반으로 한다. 트랜스포머 기반의 모델에 위치 정보를 전달하는 다른 방법으로는 상대 위치 임베딩이 있다. 절대 위치 임베딩과 달리 상대 위치 임베딩은 임의의 토큰에 대해 해당 토큰의 위치와 그 외 토큰의 상대적인 위치를 기반으로 위치 정보를 계산하는 방식으로 임베딩 과정이 아닌 자기주의 구조 내에서 구해야 한다는 차이점이 있다. 본 논문에서는 절대 위치 임베딩을 사용한 임베딩을 기준으로 한다.

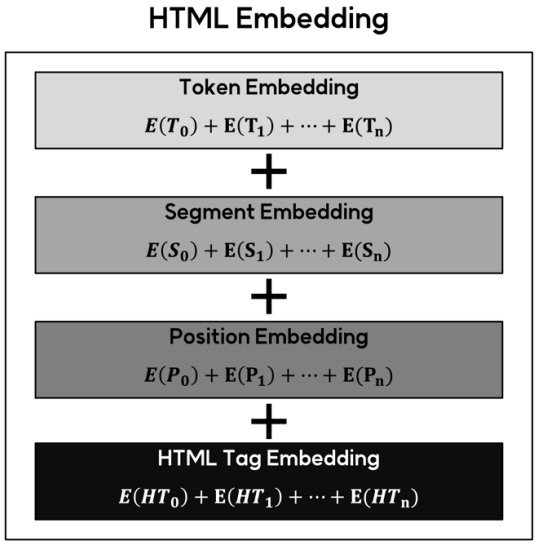
2.2 BERT 의 임베딩



(그림 1) BERT 의 임베딩 구조

대표적인 트랜스포머 기반 모델로 BERT 가 있다. BERT 는 입력에 대해 토큰 임베딩, 세그먼트 임베딩, 위치 임베딩을 각각 구한 다음 더해 입력 임베딩으로 사용한다. 토큰 임베딩은 토크나이저로 토큰화된 입력 시퀀스에 대한 임베딩을 의미하고 세그먼트 임베딩은 BERT 의 입력 시퀀스로 주어지는 질의 부분과 지문 부분을 구분하는 0 과 1 인 값에 대한 임베딩이다. 위치 임베딩은 절대 위치 임베딩을 의미하며 0 부터 최대 입력 길이에 해당하는 순차적인 정숫값에 대한 임베딩이다.

3. HTML 임베딩과 HTML 태그 스택



(그림 2) HTML 임베딩 구조

HTML 임베딩 기법은 BERT 의 임베딩 구조에 HTML 태그 임베딩을 더한 임베딩 기법이다. HTML 문서와 평문의 가장 큰 차이점은 태그의 존재 여부이다. 태그는 “a”, “h1”, “span” 등 고유한 이름을 가지며 대부분 여는 태그와 닫는 태그의 쌍을 가지고 있다. 그러나 “meta”, “link”, “img”와 같이 여닫는 태그 쌍이 아닌 단일 태그로 구성된 태그가 존재하며 “</>”와 같이 의미 없는 태그도 존재한다.

Tag Stacks by Input Sequence

Index	$m-1$	m	m
Input Tokens	<h1>	lorem	</h1>
Is valid?	true		true
Operation	push()	None	pop()
Tag Stack	<h1> <body> <html>	<h1> <body> <html>	 <body> <html>

(그림 3) 입력 시퀀스에 따른 태그 스택의 변화

이를 해결하는 방법으로 본 논문은 스택 구조를 통해 단일 태그와 의미 없는 태그를 전처리하는 HTML 태그 스택 기법을 소개한다. HTML 태그 스택은 HTML 문서를 토큰화한 후 각 토큰에 대해 해당 토

큰이 태그이면서 유효한지 확인한다. 이 때 유효성은 태그가 단일 태그이거나 의미 없는 태그일 경우 거짓을 반환한다. 이후 해당 토큰이 여는 태그일 경우 스택에 푸쉬 연산을 수행하여 삽입하고 해당 토큰이 닫는 태그일 경우 팝 연산을 수행하여 스택에서 제거한다. 이를 통해 각각의 토큰 위치에 대하여 현재 태그 정보와 상위 태그의 정보들을 추출하고 HTML 태그 스택에 저장할 수 있다.

Tag to Value Mapping

Key	Value
...	...
<h1>	12
<h2>	13
...	...
	37
...	...

(그림 4) HTML 태그-정수 매핑 테이블

각 토큰은 해당 시점의 HTML 태그 스택을 가질 수 있다. 본 논문에서 제안하는 HTML 태그 임베딩의 입력 시퀀스는 각 토큰에 해당하는 태그 스택의 태그 값을 특정 정숫값으로 매핑한다. 그리고 태그 스택에서 해당 태그의 순서 값을 곱하고 이렇게 곱한 각각의 태그 값을 더해 하나의 스칼라 값을 생성한다. 이 과정을 입력 토큰 시퀀스의 각각의 토큰에 적용하면 동일한 길이의 HTML 태그 시퀀스를 생성할 수 있다. HTML 태그 시퀀스의 원소는 동일한 태그 구조에서 동일한 값을 가지므로 웹 문서에서 태그 구조의 패턴을 모델에 전달할 수 있다는 장점이 있다. 또한 본 논문에서 제안하는 HTML 태그 임베딩을 추가한 HTML 임베딩 기법은 모델이 기존에 가지고 있는 위치 정보에 더해 HTML 문서 구조를 효과적으로 전달할 수 있으므로 웹 문서 기계 독해 과제에 대한 성능 향상을 기대할 수 있다.

4. 결론

최근 자연어 처리 모델의 고도화에 따라 평문보다 높은 난이도인 웹 문서 지문에서 기계 독해 과제를 해결하려는 연구가 증가하고 있다. 본 논문은 기계 독해 과제의 대표적인 모델 중 하나인 BERT 에 적용할 수 있는 HTML 임베딩 기법과 이를 위한 전처리 기법인 HTML 태그 스택을 제안했다. HTML 임베딩 기법은 BERT 의 기존 임베딩에 HTML 태그 임베딩을 더한 임베딩 기법이다. HTML 태그 임베딩의 입력은 입력 토큰 시퀀스와 동일한 길이의 HTML 태그 시퀀스로 BERT 의 입력 토큰 시퀀스 각각에 해당하는 태그 정보를 HTML 태그 스택을 이용하여 추출하고 정숫값으로 매핑한 후 순차 정숫값과 곱해 모두 더한 것이다. HTML 태그 스택은 입력되는 각 토큰에 대해 유효한 태그인지 검사 후 스택에 푸쉬와 팝 연산을 수행함으로써 해당 토큰의 HTML 문서 내의 태그 위치를 저장할 수 있다.

HTML 임베딩 기법은 자연어 처리 모델이 웹 문서의 태그 구조를 학습할 수 있기 때문에 KorQuAD 2.0 과 같은 웹 문서 기계 독해 과제에서의 성능 향상을 기대할 수 있다. 향후에는 소개한 HTML 임베딩 기법을 구현하고 기준 모델과 비교하여 성능을 측정하고자 한다. 또한 웹 문서 기계 독해 성능을 향상시킬 수 있는 임베딩 기법과 모델 아키텍처를 연구할 계획이다.

본 논문은 2022 년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업(2021RIS-004)과 기초연구사업(NRF-2021R1I1A3061020)의 결과 입니다.

참고문헌

[1] Rajpurkar, Pranav, Robin Jia, and Percy Liang. "Know what you don't know: Unanswerable questions for SQuAD." *arXiv preprint arXiv:1806.03822* (2018).

[2] Kim, Youngmin, et al. "KorQuAD 2.0: Korean QA dataset for web document machine comprehension." *Annual Conference on Human and Language Technology*. Human and Language Technology, (2019).

[3] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

[4] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[5] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[6] Yang, Zhilin, et al. "Xlnet: Generalized autoregressive pretraining for language understanding." *Advances in neural information processing systems* 32 (2019).

[7] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997).