한국어 단어 표상 내의 문법성 발견을 위한 구조적 탐사

민진우○[1], 나승훈[1], 신종훈[2], 김영길[2]

[1]전북대학교 인지컴퓨팅 연구실, [2]ETRI

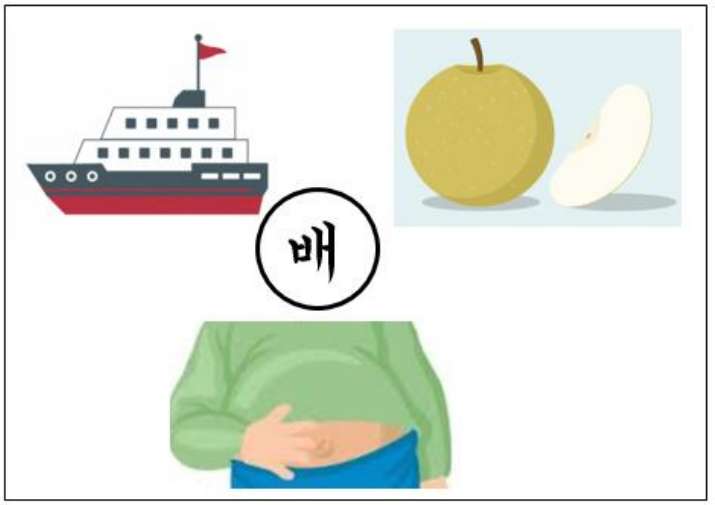
목차

- Contextual Word Representations
- 관련연구
- 구조 입증
- 실험 세팅 & 실험 결과

Contextual Word Representations

[문맥 단어 표현]

- 기존 Word Embedding의 단점
  - 중의적인 단어를 표현할 수 없음
    - 예) 단어 “배“
  - 문맥에 관계 없이 고정된 단어 표현 사용
    - 단어 이불이 언급된 문장



- 문맥 단어 표현(Contextual Word Representations)
  - 같은 단어더라도 문맥에 따라 그 표현 방법(벡터)가 바뀔 수 있는 개념의 Word Embedding
  - ELMo, BERT, GPT 계열의 다양한 모델 연구됨

관점	Word Embedding	문맥 단어 표현
Input	단어 단위	문장 단위(단어의 시퀀스)
Layer	(일반적으로) 단층	(일반적으로) 다계층
Output	해당 단어에 대한 Embedding	문장을 구성하는 각 단어에 대한 Embedding들

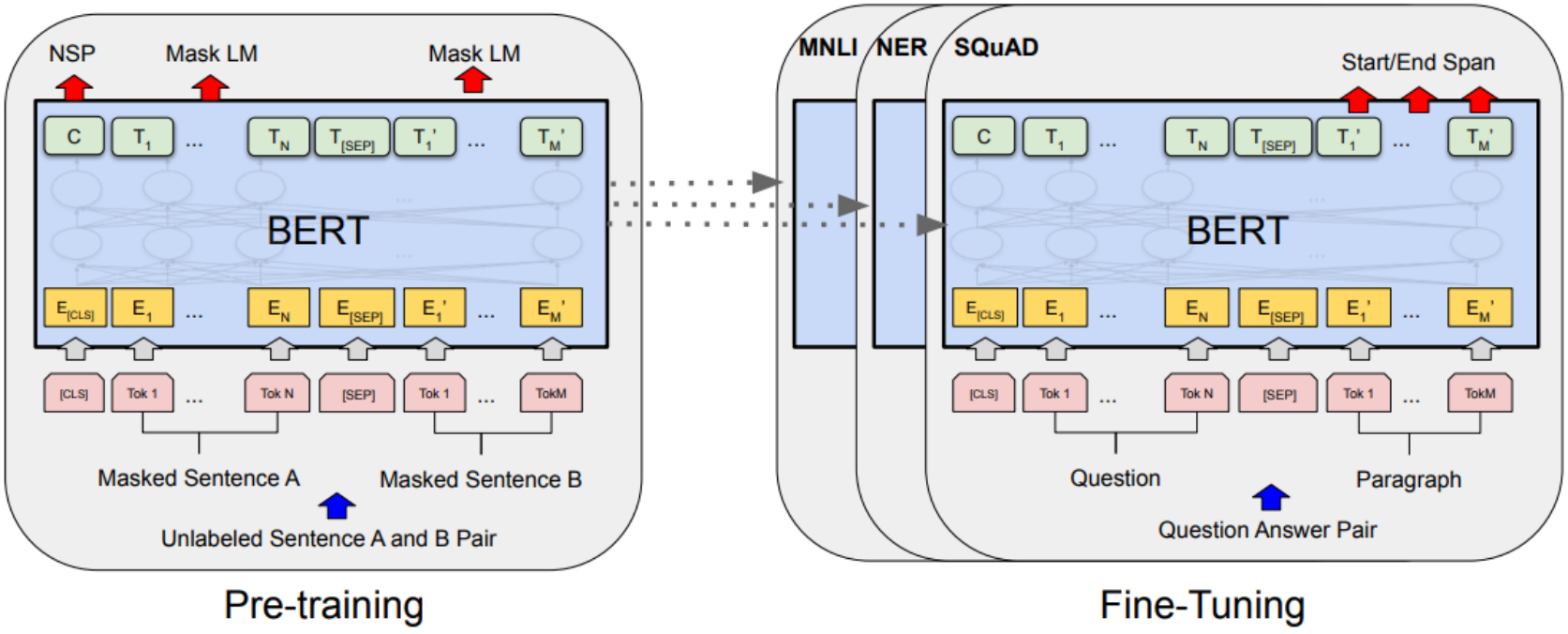
Deep contextualized word representations (Peters et al., '18)

$$\sum_{k=1}^{N} \Big(\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \Big)$$

TASK	PREVIOUS SOTA		OUR BASELINE	ELMO + BASELINE	INCREASE (ABSOLUTE/ RELATIVE)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

- 문맥 정보를 반영하는 단어 임베딩의 Pre-training 방법 제시
  - Char CNN 기반의 word embedding
  - 양방향 LSTM을 이용하여 forward LSTM을 이용하여 다음 단어 backward LSTM을 이용하여 이전 단어를 예측하는 Bi-LM 구조
  - 기존 language Model과 동일한 학습 방법

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding(Yinhan Liu et al., '19)



- BERT : Bidirectional Encoder Representation from Transformer
  - 양방향의 Transformer를 이용하여 문장 내 임의의 단어를 예측하고 다음 문장을 예측하는 두 가지 Task로 언어 모델 학습
  - 응용 Task에 Fine-Tuning하는 방식으로 성능 향상

RoBERTa: A Robustly Optimized BERT Pretraining Approach(Yinhan Liu et al., '19)

- BERT의 최적화 모델
  - BERT의 학습 과정에서 최적화 되지 않은 부분을 최적화
    - **Dynamic Mask LM** 기존의 BERT 모델과 달리 RoBERTa에서는 BERT와 달리 매 학습마다 마스킹하는 단어를 다르게 하는 Dynamic Masking 방식을 사용
    - **NSP 테스크 제외** 다음 문장을 예측하는 NSP 태스크는 학습하지 않고 최대 토큰 길이 512에 가깝도록 다른 문서의 문장으로 채워 넣어 문서 길이가 미리 설정한 하여 학습 효율을 높임
  - 기존 BERT에서의 성능 향상

Model	SQuAD 1.1 EM	SQuAD 1.1 F1	SQuAD 2.0 EM	SQuAD 2.0 F1
Single models on dev, w/o data augmentation				
$\text{BERT}_{\text{LARGE}}$	84.1	90.9	79.0	81.8
$\text{XLNet}_{\text{LARGE}}$	**89.0**	94.5	86.1	88.8
RoBERTa	88.9	**94.6**	**86.5**	**89.4**
Single models on test (as of July 25, 2019)				
$\text{XLNet}_{\text{LARGE}}$			86.3†	89.1†
RoBERTa			86.8	89.8
XLNet + SG-Net Verifier			**87.0**†	**89.9**†

한국어 Contextual word representations 연구

- ETRI BERT
  - 한국어의 특성을 반영한 두 가지 단위의 BERT 언어 모델 제공
    - 형태소 분석 기반 모델 : 입력 문장을 형태소 분석기를 이용해 형태소 단위로 분리하여 입력 토큰으로 사용.

    - 어절 기반 언어 모델 : 문자의 어절에서 고빈도로 발생하는 문자(음절)들을 결합하여 토큰을 구성한 BPE(Byte Pair Encoding)방식.

  - 높은 성능 향상
    - BERT를 이용한 한국어 의존 구문 분석[박천음, KCC '2019]

Dependency parsing	UAS	LAS
이창기[16] with MI	90.37	88.17
나승훈[6]: deep biaffine attention	91.78	89.76
박천음[5]: 포인터 네트워크	92.16	89.88
안휘진[10]: deep biaffine + 스택 포인터 네트워크	92.17	90.08
박성식[11]: ELMo + 멀티헤드 어텐션	92.85	90.65
BERT + LSTM deep bilinear	93.85	91.78
BERT + LSTM deep biaffine	**94.06**	**92.00**

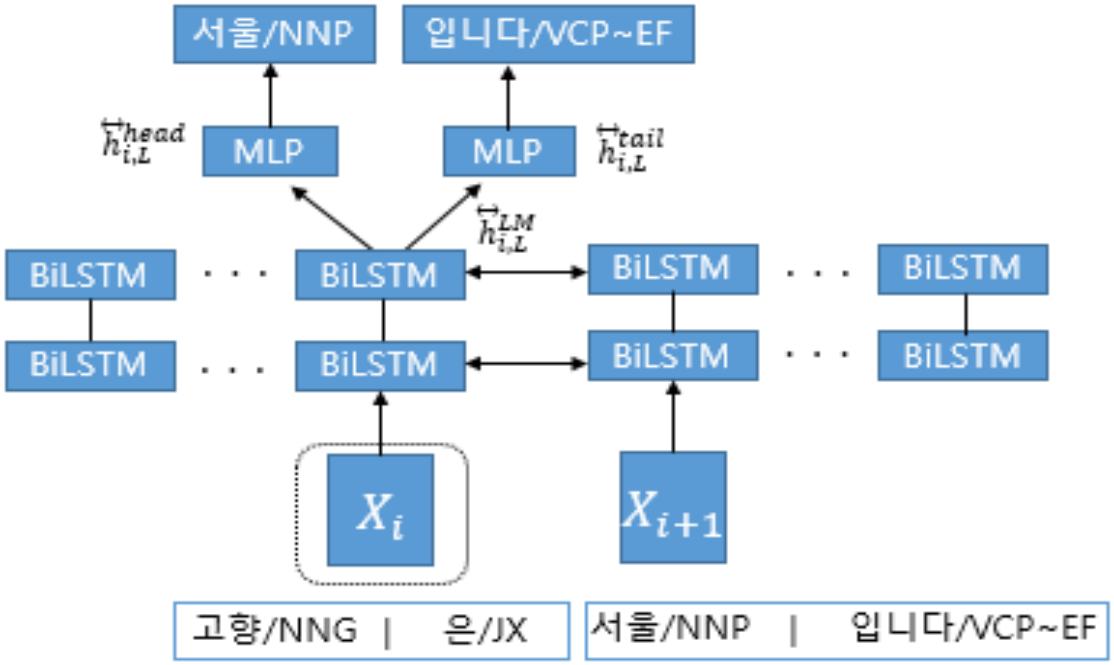
한국어 Contextual word representations 연구

- 한국어 RoBERTa 모델(KCC ‘19)
  - RoBERTa를 한국어 코퍼스 상에서 학습
  - Hybrid Tokenizer 제안
    - 형태소 토큰 5만개와 BPE 토큰 2만개를 단어장으로 구성
    - 분석된 형태소를 형태소 단위를 우선적으로 단어장에서 매칭한 후 해당 형태소가 미등록어일 경우 형태소를 BPE 단위로 토크나이징 하는 하이브리드 방식을 사용

원문
고전주의와 바로크는 공통의 면이 있다.
형태소 분석 결과
고전주의/NNG, 와/JC, 바로크/NNG, 는/JX, 공통/NNG, 의/JKG, 면/NNG, 이/JKS, 있/VV, 다/EF, ./SF
토크나이징 결과
_고전, 주의, 와/JC, 바로크/NNG, 는/JX, 공통/NNG, 의/JKG, 면/NNG, 이/JKS, 있/VV, 다/EF, ./SF

한국어 Contextual word representations 연구

- KoELMo: 한국어를 위한 문맥화된 단어 표상(홍승연, HCLT '18)
  - 기존 ELMo를 한국어에 적용할 수 있도록 확장
  - 형태소들의 조합으로 단어 표상을 얻고 다음 단어(어절)의 시작과 끝 형태소를 예측하도록 ELMo 학습



구조 입증

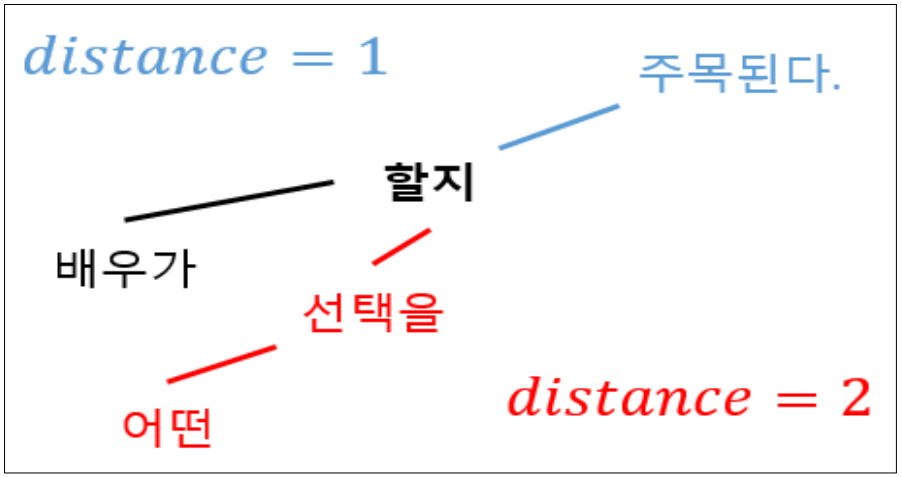
- 사전 학습된 단어 표상이 실제로 어떠한 언어학적 지식을 표현하고 있는지를 분석하는 연구들의 수행

- (Hewitt et al.,NAACL ‘19)의 연구에서는 단어 표현 내에 전체적인 트리 구조가 얼마나 잘 내재되어 있는지 여부를 판별하는 Structural Probe(구조 입증) 방법을 제안

- 본 연구에서는 한국어 ELMo, BERT 등의 문맥 단어 표현 방법이 벡터 공간 상에 전체 문법적 트리 구조가 반영되어 있음을 결과로 제시

구조 입증

- 구조 입증 방법
  - 먼저 트리 T가 주어졌을 때, 바로 이웃한 노드(단어) 사이의 거리를 $d_T(u, v) = 1$로 정의
  - 단어 "할지"를 기준으로 바로 이웃한 노드인 단어 "주목된다"와의 거리는 1이고 단어 단어 "어떤"의 거리는 2



구조 입증

- 구조 입증 방법
  - 문장 길이 l의 n개의 단어를 갖는 단어의 시퀀스 $w_{1:n}^{l}$이 주어질 때 문맥 단어 표현 모델을 통한 출력된 벡터 표현의 시퀀스를 $h_{1:n}^{l}$
  - 두 단어 사이의 거리를 수식의 제곱 거리(Squared Distance) 함수로 정의

$$d_B(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l)^2 = \left(B(\boldsymbol{h}_i^l - \boldsymbol{h}_j^l)\right)^T \left(B(\boldsymbol{h}_i^l - \boldsymbol{h}_j^l)\right)$$

  - 학습
    - 실제 단어 사이의 거리와 두 벡터 간의 거리의 차가 최소가 되도록 학습
    - $|s^l|$은 문장의 길이를 나타내며 전체 단어 쌍은 $|s^l|^2$이므로 제곱으로 정규화

$$min \sum_l \frac{1}{|s^l|^2} \sum_{i,j} |d_{T^l}(w_i^l, w_j^l) - d_B(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l)^2|$$

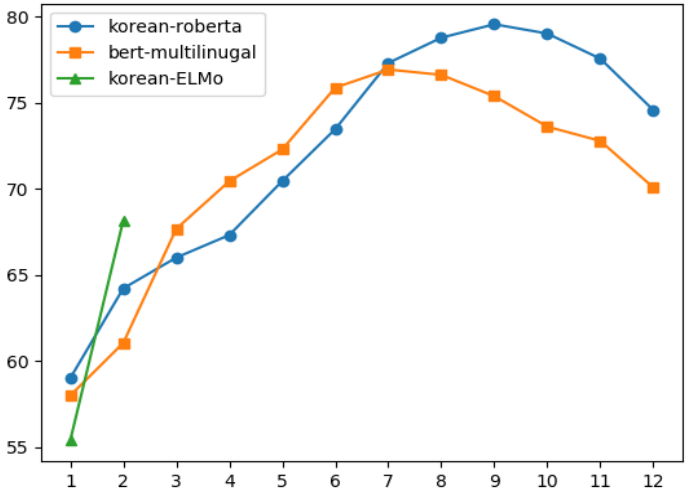
  - 디코딩
    - 거리함수를 통해 계산된 모든 단어 쌍에 대한 거리를 표현하는 무방향 그래프는 MST(minimum spaning tree)알고리즘을 통해 방향성이 없는 구문 트리를 반환

구조 입증

- 실험 세팅
  - 구조 입증을 위해 실험한 문맥 단어 표현 3가지
    - 한국어 ELMo : 2층의 LSTM 인코더로 구성
    - Bert-Multilingual, 한국어 RoBERTa : 기존 BERT-Base와 세팅이 동일하여 12층의 트랜스포머로 구성
  - 실험 집합
    - 실험을 위한 데이터 셋으로 세종 구문 분석 데이터 셋을 사용
  - 평가 지표
    - UUAS(Undirected Unlabeled Attach Score)를 이용

구조 입증

- 실험 결과
  - 각 모델의 Layer 별 UUAS 성능
    - Bert-Multilingual은 7번째 층에서 가장 성능이 높았으며 한국어 RoBERTa은 9번째 층에서 가장 높은 성능을 보여줌
  - 각 모델 별 최종 UUAS 성능
    - 실제로 응용 테스크에서 가장 높은 성능을 보이고 있는 모델인 RoBERTa가 가장 강력한 문법성을 내포하고 있음을 확인



	UUAS
한국어 ELMo Layer2	68.17%
Bert-Multilingual Layer7	76.94%
한국어 RoBERTa Layer9	79.57%

- 결론
  - 구조 입증 방법을 이용하여 사전 학습 한국어 문맥 단어 표현들이 문법 구조를 잘 반영하고 있음을 실험 결과로 얻음

Q&A

감사합니다.