Ko-SenseBERT: 한국어 어휘지도(UWordMap)에 기반한 의미 강화된 언어 모델

이현민, 이건희, 나승훈, 옥철영
전북대학교
울산대학교 한국어처리연구실
{leehm, dkghszkfhs, nash} @ jbnu.ac.kr
okcy@ulsan.ac.kr

서 론

- Transformer는 self-attention 매커니즘을 이용하여 순차적데이터에서 관계를 추적하여 문맥과 의미를 학습하는 신경망이다.

- BERT 모델은 기본적으로 입력 문장에서 MASK된 단어를 추측하는 학습 방법을 사용한다.

- 좀 더 명확한 의미를 끌어내기 위해서 단어의 의미를 외부 지식으로부터 입력 받아 보강하는 SenseBERT가 제시되었다.

- 한국어 어휘지도(UWordMap)로 부터 단어의 sense를 맵핑한 외부지식을 사용하여 한국어 버전 Ko-SenseBERT를 학습하고 국립국어원 동형이의어 분별 태스크를 미세조정(Fine-tuning)하여 성능을 비교한다.

UWordMap

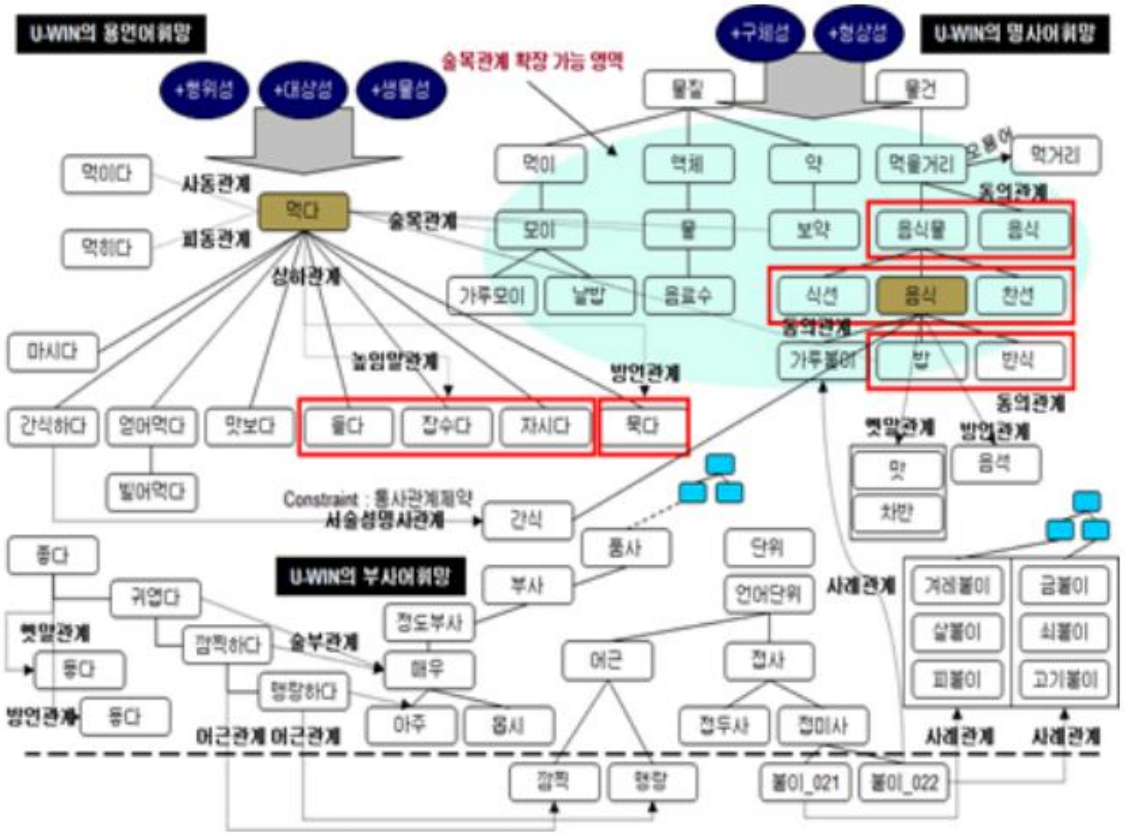


그림 1 한국어 어휘지도(UWordMap) 구조도

- 표준국어대사전(2002년 CD 버전)을 기반으로 명사 의미망(상하위어), 용언 논항의 의미 제약정보, 부사와 의미적으로 결합하는 용언/부사/명사 등을 연결한 어휘의미망이다.
- 명사의미망(상하위어)을 이용하여 단어의 super-sense를 사용하였다.
- 단어의 super-sense는 공간, 과정, 관계, 기호, 단위, 대상, 모양, 물건, 방법, 범위, 생물, 성질, 시간, 요소, 인지, 용언/부사, 작용, 재료, 정도, 존재, 종류, 집단, 행위, 힘으로 총 24개 분야로 나누어져 있다.

SenseBERT

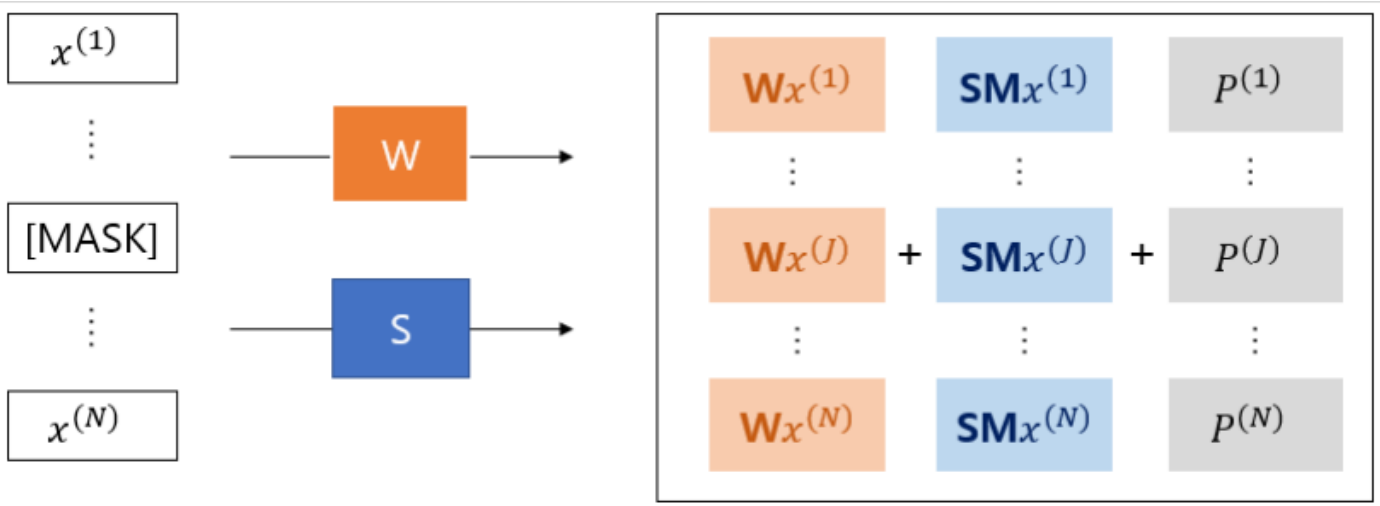


그림 1: SenseBERT Embedding 구조

- SenseBERT는 단어의 Word 임베딩뿐만 아니라 각 word에 대한 sense 임베딩을 추가로 구성하여, 최종 임베딩을 얻는다.
- $S \in \mathbb{R}^{D_s \times D_w}$를 추가한다. 이때, $D_s =$ 45의 사이즈를 가진다.
- 단어의 입력 x에 따른 transformer input값은 다음과 같다.
- $V_{input}^{(j)} = (W + SM)x^{(j)} + p^{(j)}$

SenseBERT

1. “This **bass** is delicious”
(supersenses: noun.food, noun.artifact, etc)

2. “This **chocolate** is delicious”
(supersenses: noun.food, noun.attribute, etc)

3. “This **pickle** is delicious”
(supersenses: noun.food, noun.state, etc)

입력문장의 예시, “This [MASK] is delicious”에서 3문장 모두 supersense 레이블로 noun.food를 가지고 있으므로, 정답 레이블이 강화될 것으로 기대된다.

- Sense loss를 구하는 allowed-sense loss 식
- $\mathcal{L}_{SLM}^{allowed} = -logp(\ s\ \in A(w)|context)$

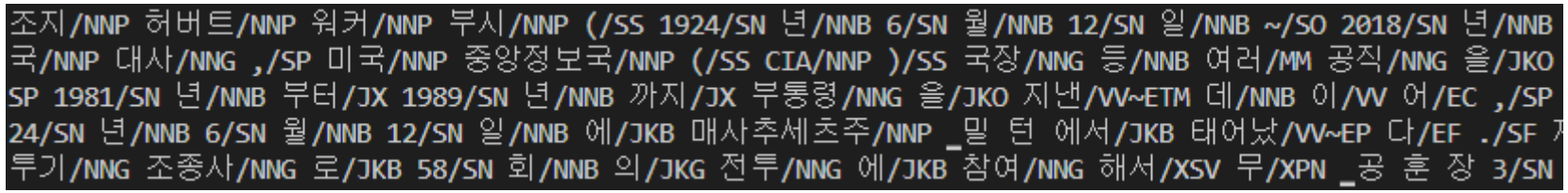
$$= -log \sum_{s \in A(w)} P(s|context)$$

- $P(s|context) = \frac{\exp(y_s^{senses})}{\sum_s \exp(y_s^{senses})}$

- Sense loss를 구하는 reg loss 식
- $\mathcal{L}_{SLM}^{reg} = -\sum_{s \in A(w)} \frac{1}{|A(w)|} logP(s|context)$

- *최종 sense loss, 두 loss의 합을 사용한다.*
- $\mathcal{L}_{SLM} = \mathcal{L}_{SLM}^{allowed} + \mathcal{L}_{SLM}^{reg}$

Ko-SenseBERT
토크나이저



Tokenizer 예시

- UWordMap의 경우 명사에 한정하여 상하위어를 제공한다.
- 명사 단어만 sense를 가지고 있으며, 입력된 문장에서 명사를 분리해내야 한다.
- 형태소 기반 하이브리드 토크나이저는 형태소 토큰들과 BPE(byte pair encoding)토큰들로 이루어져 있다.
- 한국어 버전 M 매트릭스를 구축하기 위해서는 온전한 단어 토큰이 필요하다.
- 형태소 토큰이 포함되어 있는 하이브리드 토크나이저를 사용하였다.

Ko-SenseBERT
M 매트릭스

- 영어와 다른 한국어의 상황을 고려하여 두 가지 방법으로 구축
- 1. 한 글자 Token을 '제외한' 매트릭스 M 구축
- 2. 한 글자 Token을 '포함한' 매트릭스 M 구축
- 예를 들어 단어 "달"의 경우 물건, 생물, 모양, 시간, 단위 총 5개의 supersense를 가지고 있다.
- 1번 매트릭스 M에서 단어 "달"은 아무런 sense를 갖지 못한다.
- 2번 매트릭스 M에서 단어 "달"은 5개의 sense를 가지게 된다.
- 최종적으로 Ko-SenseBERT의 단어 집합은 $M \in \mathbb{R}^{D_s \times D_w}$의 고정 매트릭스를 가지게 된다.

Ko-SenseBERT

실험 세팅 및 실험 결과

Dataset	*Accuracy*(%)
BERT-based-multilingaul-cased(google)	72.8
BERT-based	70.2
Ko-SenseBERT(M_2)	71.2
Ko-SenseBERT(M_1)	69.7

국립국어원 동형이의어 분별 Task

- TITAN Xp 8G, 10 Epoch
- Learning-rate: 1e -4, Warm up: 0.06
- Batch size: 4, Seq length: 512

국립국어원 동형이의어 분별 데이터 예시
Target: 단정
Sentence1: 그의 죽음은 타살로 단정이 되었다.
Sentence2: 단정이 된 교실은 정돈되어 있었다.
Label: False
Start s1, end s1: 11, 13
Start s2, end s2: 0, 2

Ko-SenseBERT
결론

- 본 논문에서는 한국어 어휘지도(UWordMap)을 이용하여 단어의미에 강건한 언어모델 Ko-SenseBERT를 적용해 보았다.
- 한국어 적용을 위해 외부 지식 매트릭스 M을 두 가지 버전으로 실험을 진행하였다.
- M1의 모델의 경우 한 글자 단어를 포함하지 않아 단어의 sense가 오히려 단어 학습을 방해하는 효과가 일어났다.
- 그 결과 BERT-based 대비 0.5% 낮은 스코어를 기록하였다.
- M2 모델의 경우 한 글자 단어를 포함하여 단어의 sense를 충분히 학습할 수 있었다.
- 그 결과 BERT-based 대비 1% 높은 스코어를 기록하였다.
- 차후 실험에서 사전 학습 코퍼스 양을 증가하여 더 효과적은 sense 학습을 진행해볼 예정이다.