한국어 임베딩 스터디

Deep contextualized word representations

INDEX

1. Introduction

2. ELMo: Embeddings from Language Models

3. Evaluation

4. Conclusion

Introduction

- Pre-trained word representations은 많은 자연어이해 모델에서 중요한 요소
- ELMo는 쉽게 다른 모델에 적용가능, Deep contextualized word representation을 보여줌
- 이 표상은 (문장 내) 각 token이 전체 입력 sequence의 함수인 representation를 할당
- concatenated language model로 학습된 bidirectional LSTM(biLM)로부터 얻은 vector를 사용
- LSTM 각 layer 결과를 가중합하여 얻어지며 이런 방식이 성능 우수
- LSTM의 낮은 단계의 layer는 품사 등 문법 정보를, 높은 단계의 layer는 문맥 정보를 학습
- 많은 실험에서 최신 논문보다 에러율을 약 20% 감소 시킴

ELMo: Embeddings from Language Models

- Character-based convolution layer
- Bidirectional language models
- ELMo
- Using biLMs for supervised NLP tasks
- Pre-trained bidirectional language model architecture

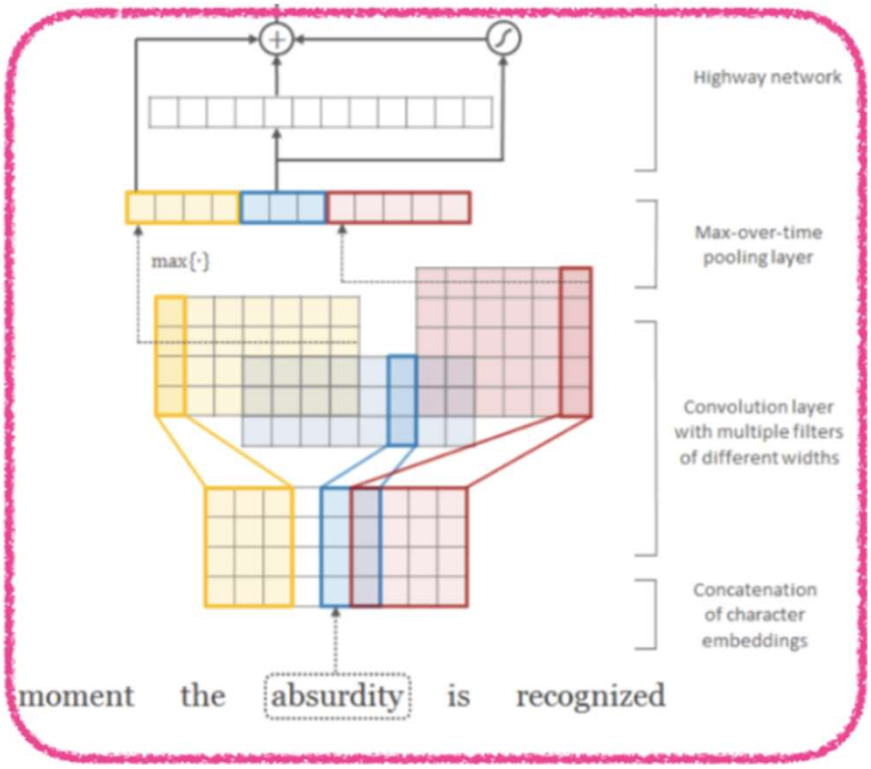
Character-based convolution layer

- 단어를 문자 단위로 변경하여 사용(OOV 문제 발생 안함)
- 다양한 필터 사이즈로 컨볼루션 필터 생성
- 컨볼루션 필터에서 **Max pooling**을 통해 풀링 벡터 생성
- 풀링 벡터를 이어 붙인 후, 하이웨이 네트워크로 전송

- 하이웨이 네트워크

$$y = H(x, W_H) * T(x, W_T) + x * C(x, W_c)$$
$$= H(x, W_H) * T(x, W_T) + x * (1 - T(x, W_T))$$

- 값을 얼마난 변경(**T**)해서 보낼지 결정
- 레이어가 많을 때 학습이 잘 되지 않는 경우에 대한 우회 경로



Bidirectional language models

forward language model $$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_1, t_2, \ldots, t_{k-1})$$

backward language model $$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_{k+1}, t_{k+2}, \ldots, t_N)$$

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

- Θ_x는 token representation
- Θ_s는 softmax layer

- 이 둘의 LSTM의 파라미터는 다르게 고정 됨

ELMo

ELMo는 biLM의 중간 layer representation을 task-specific하게 결합

biLM의 L개의 layer는 각 token x_k당 2L+1개의 representation을 계산

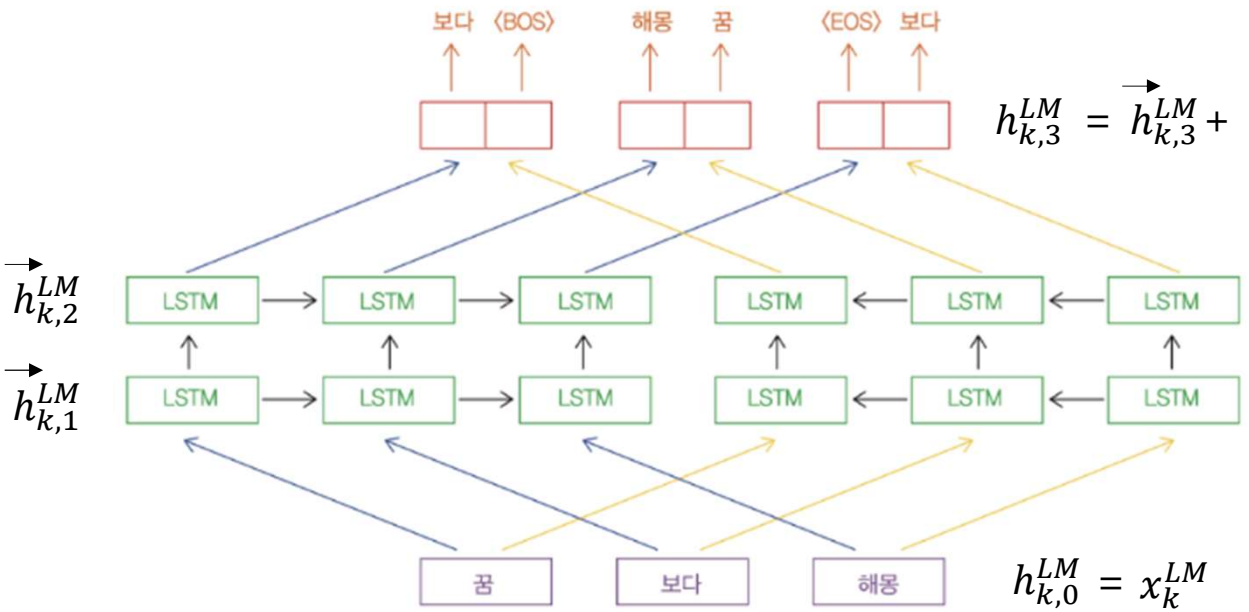
$$\begin{aligned} R_k &= \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} \mid j = 1, \ldots, L\} \\ &= \{\mathbf{h}_{k,j}^{LM} \mid j = 0, \ldots, L\}, \end{aligned}$$

다운스트림 모델에 포함시키기 위해, ELMo는 R의 모든 레이어를 하나 벡터로 분해

$$\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}$$

ELMo

$$R_k = \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} \mid j = 1, \ldots, L\}$$
$$= \{\mathbf{h}_{k,j}^{LM} \mid j = 0, \ldots, L\},$$



$$h_{k,3}^{LM} = \overrightarrow{h}_{k,3}^{LM} + \overleftarrow{h}_{k,3}^{LM} + x_k^{LM}$$

$$\overrightarrow{h}_{k,2}^{LM} \quad \overleftarrow{h}_{k,2}^{LM} \quad h_{k,2}^{LM} = \overrightarrow{h}_{k,2}^{LM} + \overleftarrow{h}_{k,2}^{LM} + x_k^{LM}$$

$$\overrightarrow{h}_{k,1}^{LM} \quad \overleftarrow{h}_{k,1}^{LM} \quad h_{k,1}^{LM} = \overrightarrow{h}_{k,1}^{LM} + \overleftarrow{h}_{k,1}^{LM} + x_k^{LM}$$

$$h_{k,0}^{LM} = x_k^{LM}$$

그림 5-18 ELMo 양방향 LSTM 및 출력 레이어

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

Using biLMs for supervised NLP tasks

- biLM으로 representation들을 얻고 선형 결합을 하여 학습
- 먼저 가장 하위 계층에 biLM이 없는 지도 모델을 고려
  - $(t_1, t_2, \dots, t_N)$이 주어지면 기학습된 단어 embedding(+글자기반 representation)을 사용하여 각 token 위치마다 문맥-독립적 token representation x_k를 만든다.
  - 그러면 모델은 biRNN이든 CNN이든 FFN이든 사용하여 문맥-의존적 representation h_k를 생성한다.

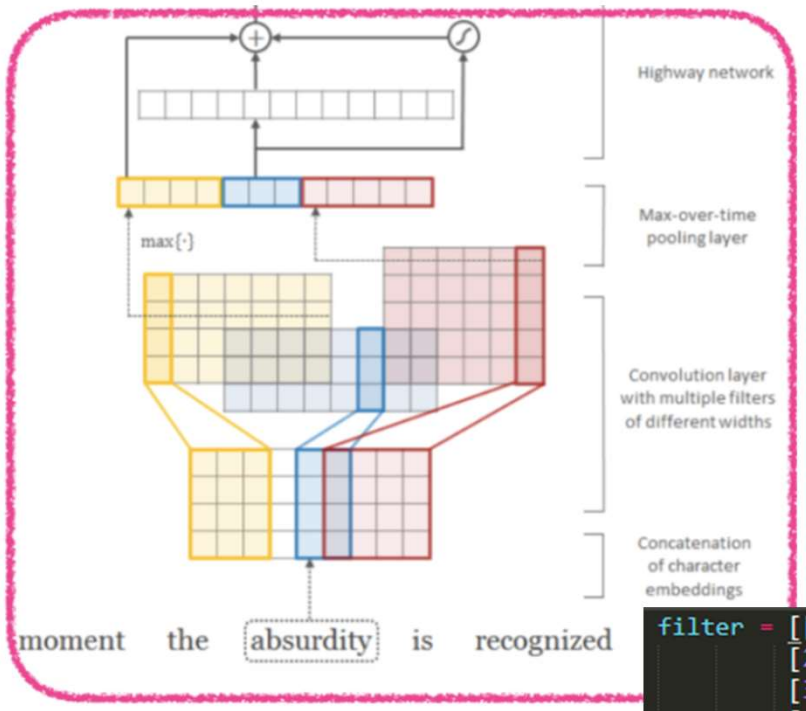
- ELMo를 지도 모델에 추가하려면
  - 먼저 biLM의 weight를 고정시키고
  - ELMo 벡터 $ELMo_k^{task}$와 x_k를 이어붙인 후
  - ELMo enhanced representation $[x_k; ELMo_k^{task}]$를 task RNN에 전달한다.

Pre-trained bidirectional language model architecture

- biLM은 (Józefowicz et al. 2016)의 것과 비슷
- but 양방향 학습의 동시학습을 가능하게 하고 LSTM layer 사이에 residual connection을 추가
- 완전히 문자기반인 입력 representation을 유지하면서도 모델복잡도와 계산요구량의 균형을 맞추기 위해, embedding과 은닉차원을 반으로 줄였다.
- 최종 모델은 4096개의 unit과 512차원의 projection layer, 1-2번 layer 사이 residual connection을 갖는 L=2 biLSTM을 사용한다.
- 그 결과 biLM은 각 입력 token마다 순수 문자기반 입력 때문에 학습셋을 벗어나는 것을 포함한, 3개의 layer of representation을 생성한다.

$$\gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}$$

Pre-trained bidirectional language model architecture




```
filter = [[1, 32],
          [2, 32],
          [3, 64],
          [4, 128],
          [5, 256],
          [6, 512],
          [7, 1024],
          [8, 2048]]
```

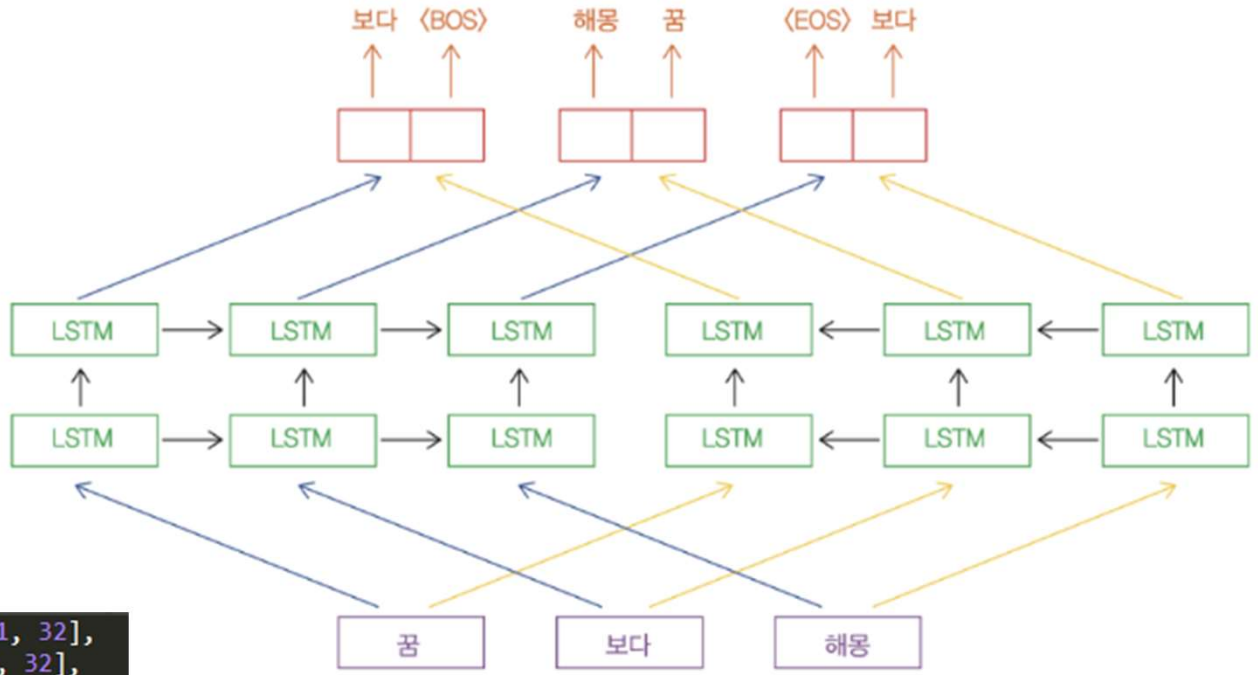


그림 5-18 ELMo 양방향 LSTM 및 출력 레이어

Evaluation

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%

Premise:
Ruth Bader Ginsburg being appointed to the US Supreme Court.

Hypothesis:
A grilled sandwich on a plate.

Label:
Contradiction [different scenes]

- SQuAD(Question Answering)
  - 데이터: The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016)
  - 기준: Bidirectional Attention Flow model(Clark and Gardner, 2017)
  - Self attention 계층 추가, 풀링 단순화, GRU 사용, ELMo 추가
  - 11개의 앙상블 이용
- SNLI (Textual Entailment)
  - 데이터: The Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015)
  - 기준: Chen et al. (2017)
  - biLSTM를 사용하여 premise와 hypothesis를 인코딩, matrix attention 계층, 로컬 추론 계층, 또 다른 biLSTM 추론 구성 계층, 출력 계층 앞에 풀링 작업. 모델에 ELMo 추가
  - 5개의 앙상블 이용

Evaluation

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%

- SRL(Semantic Role Labeling)
  - 데이터: 문장의 술어적 구조를 모델링, e.g.) 누가 누구에게 무슨 짓을 했느냐
  - 기준: He et al. (2017) - BIO 태깅 문제로 모델링, 8층 biLSTM
  - re-implementation에 ELMo를 추가

- Coreferecne resolution
  - 데이터: Stanford Sentiment Treebank (SST-5; Socher et al., 2013)
  - 기준: Lee et al. (2017) - biLSTM와 attention 메커니즘을 사용
  - biLSTM와 주의 메커니즘을 사용한 다음, coreference 체인을 찾기 위해 softmax mention ranking 모델을 이용

Evaluation

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%

- NER
  - 데이터: CoNLL 2003 NER task (Sang and Meulder, 2003)
  - 기준: Lample et al.(2016); Peters et al.(2017)
  - ELMo를 통해 biLSTM-CRF를 강화하여 5회 시행

- SST-5(Sentimental classification)
  - 데이터: Stanford Sentiment Treebank (SST-5; Socher et al., 2013) -영화 리뷰와 별점
  - 기준: McCann et al. (2017) biattentive classification network (BCN)
  - BCN 모델에서 CoVe를 ELMo로

Conclusion

- biLM으로부터 고품질의 깊은 문맥의존 representation을 학습하는 일반적인 방법을 소개
- 많은 NLP 문제들에서 ELMo를 적용했을 때 많은 성능 향상
- ablation을 통해 biLM의 모든 layer들이 각각 효율적으로 문맥 정보를 포착함을 보여줌