Deep Contextualized Word Representations (ELMo)

https://arxiv.org/pdf/1802.05365

introduction

- Pretrained word representations
  단어의 의미와 문맥 정보를 효과적으로 담아야 함

- ELMo 이전: Word2Vec, Glove embeddings 등
  1. 하나의 단어에 하나의 embedding
  2. 여러가지 의미가 존재하는 단어를 하나의 단어로 취급

-> 대량의 코퍼스로 학습된 한쌍의 LM, bi-LSTM에서 파생된 벡터를 사용(ELMo[Embeddings from Language Models] representation)

ELMo(Embeddings from Language Models)

- Live in present not past. -> 현재
- Here is your b-day present. -> 선물

기존에는 각 word마다의 embedding vector만을 추출함.

과거: present의 임베딩 값은 하나(의미 구분 불가)

따라서, 각 구문에 따른 다양한 의미의 임베딩한 값이 달라야 함

-> biLM

Bidirectional language models(1)

순방향 언어 모델과 역방향 언어 모델을 합친 것을 뜻함.

-> 입력문장의 전체 문맥을 고려하여 토큰에 대한 벡터값 생성

ELMo는 **모든 레이어의 출력값을 활용**하여 임베딩을 생성

- 단순히 최상위 LSTM 레이어만 사용하는 것보다 나음
- 상위 LSTM: 단어 문맥에 따른 **의미 변화**를 포착하는데 유용
- 하위 LSTM: **구문적 특성**을 모델링하는데 유용

Bidirectional language models(2)

1. 순방향 언어 모델

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

현재 단어들이 다음 단어 예측하도록 학습

2. 역방향 언어 모델

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

뒤에 오는 단어들을 통해 앞에 있는 단어 예측

-> 두 방향에 대한 로그 가능도(log likelihood)를 공통으로 최대화하는 것을 목표로 함

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

ELMo

1. 각 층의 출력값을 합침

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

2. 각 층의 출력값 별로 가중치 매김

$$E(R_k) = \mathbf{h}_{k,L}^{LM}$$

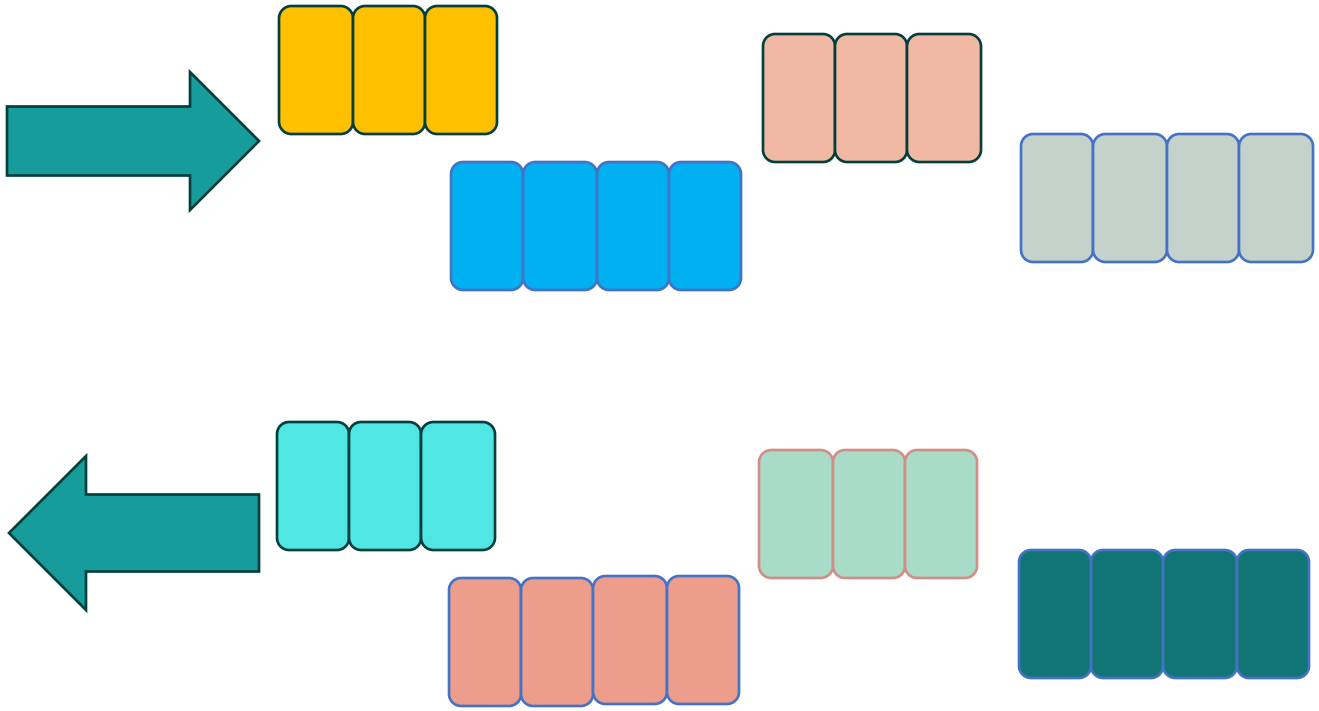
3. 각 층의 출력값을 모두 더함

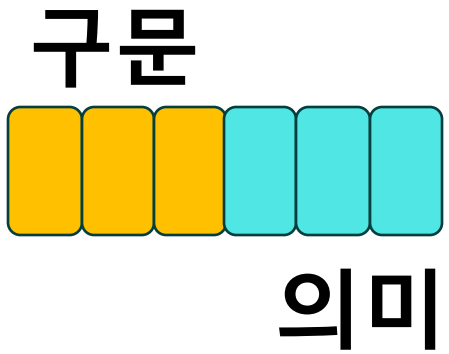
4. 벡터의 크기를 결정하는 스칼라 매개변수 곱

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

ELMo

I want to sleep...





Evaluation

- SQuAD: 질문에 대한 답 찾기
- SNLI: 주어진 전제로 가설이 참인지 판단
- SRL: 문장의 동사와 그 인자들의 관계 식별
- Coref: 서로 참조하는 엔티티들을 연결하는 작업
- NER: 특정 엔티티를 식별하는 작업
- SST-5: 감정 분석

Evaluation(1)

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%

Evaluation(2)

Task	Baseline	Last Only	All layers	
			λ=1	λ=0.001
SQuAD	80.8	84.7	85.0	**85.2**
SNLI	88.1	89.1	89.3	**89.5**
SRL	81.6	84.1	84.6	**84.8**

Task	Input Only	Input & Output	Output Only
SQuAD	85.1	**85.6**	84.8
SNLI	88.9	**89.5**	88.7
SRL	**84.7**	84.3	80.9

Evaluation(3)

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

Table 4: Nearest neighbors to "play" using GloVe and the context embeddings from a biLM.

Model	**F_1**
WordNet 1st Sense Baseline	65.9
Raganato et al. (2017a)	69.9
Iacobacci et al. (2016)	**70.1**
CoVe, First Layer	59.4
CoVe, Second Layer	*64.7*
biLM, First layer	67.4
biLM, Second layer	*69.0*

Model	**Acc.**
Collobert et al. (2011)	97.3
Ma and Hovy (2016)	97.6
Ling et al. (2015)	**97.8**
CoVe, First Layer	*93.3*
CoVe, Second Layer	92.8
biLM, First Layer	*97.3*
biLM, Second Layer	96.8

Conclusion

1. 양방향 언어 모델(biLM)로부터 고품질의 깊은 문맥 의존적 표현(단어의 의미를 문맥에 따라 다르게 반영하는 방법)을 학습하기 위한 접근법, ELMo를 제안함

2. ELMo를 다양한 자연어 처리(NLP) Task에 적용했을 때 큰 성능 개선을 보임

3. layer의 층이 올라갈수록 구문보다 의미 정보를 담아낸다는 사실

4. 모든 layer의 정보를 사용하는 것이 전체 Task 성능을 향상시키는 데 도움이 됨