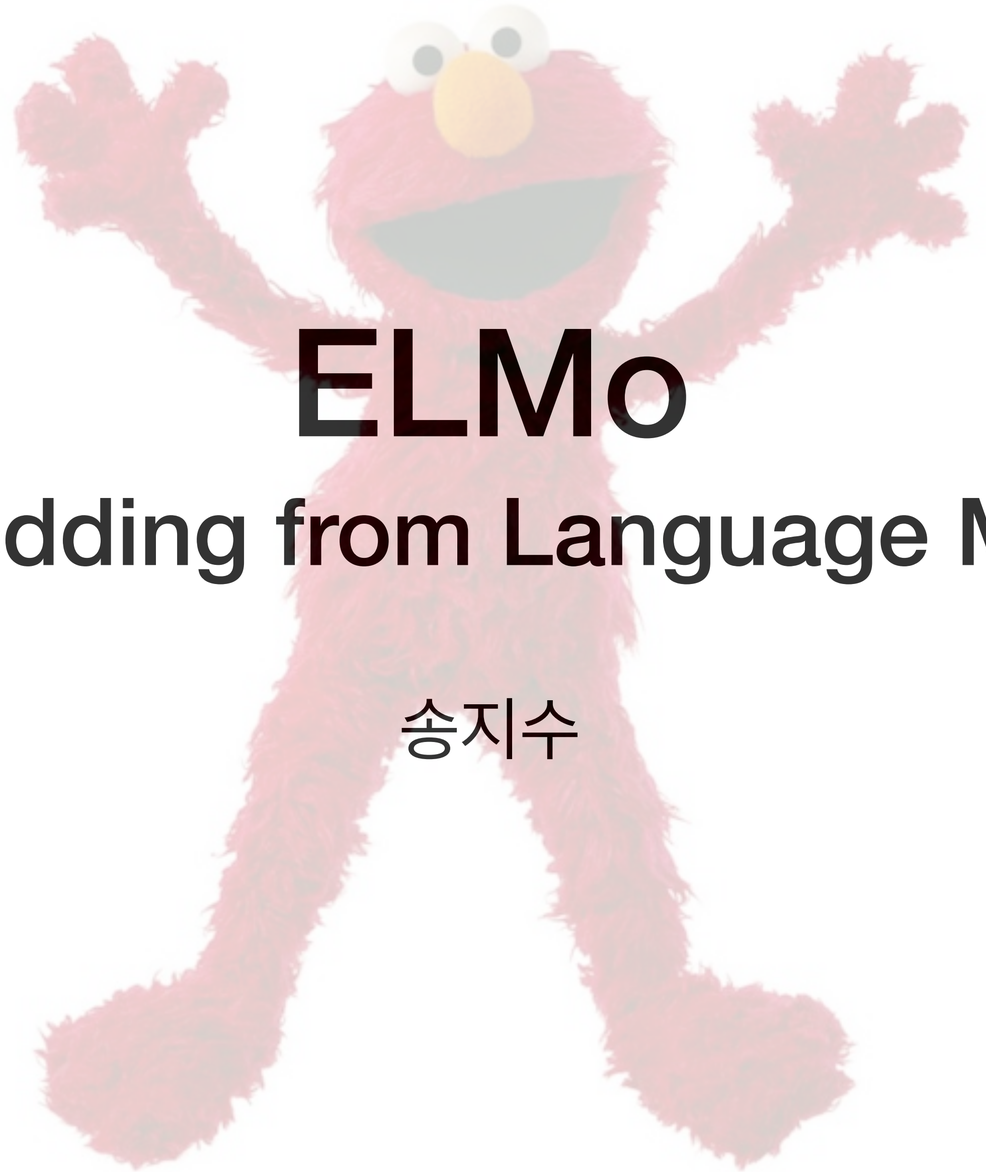


ELMo

(Embedding from Language Model)

송지수

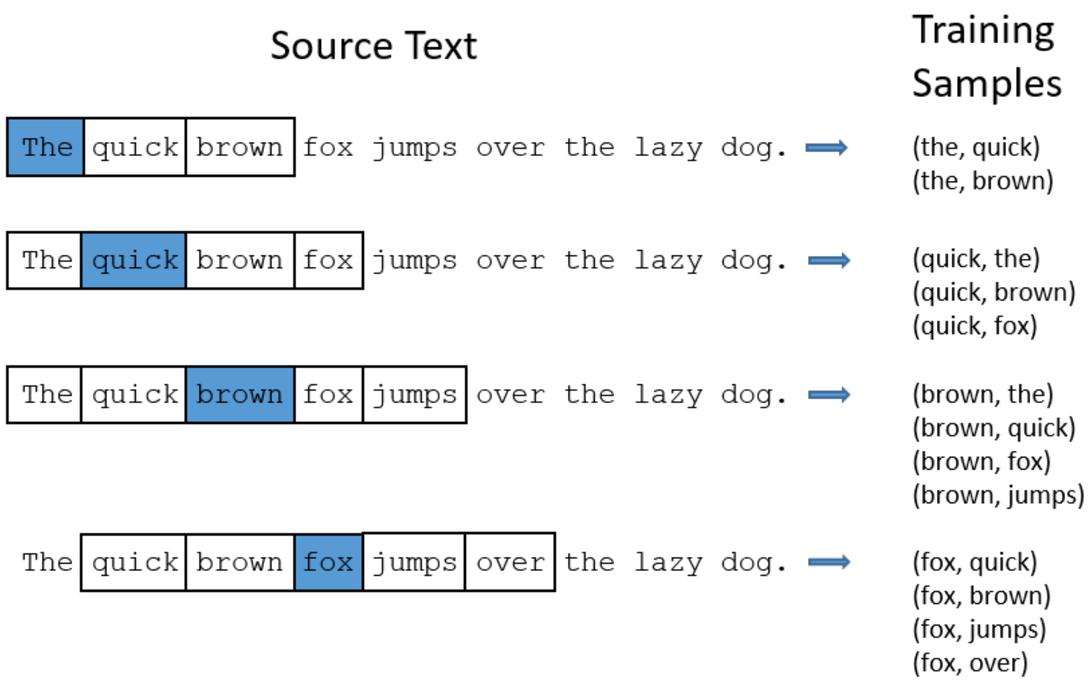
목차

ELMo Introduction

엘모?

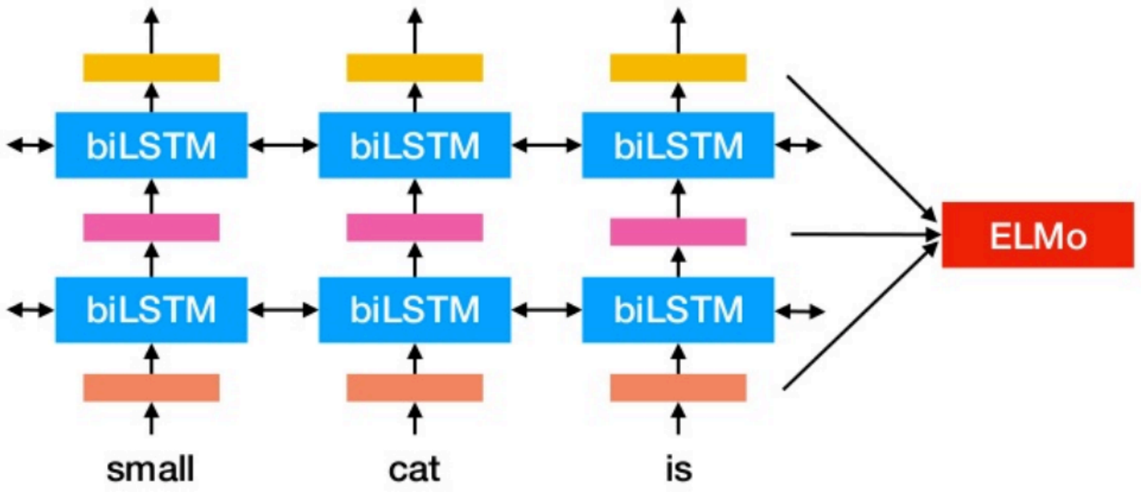
- Embedding from Language Model
- 기존의 Word Embedding 기법인 Word2Vec, FastText 등은 문맥에 관계 없이 단어 당 벡터 하나로 Embedding 됨
- ELMo는 단어를 Vectorization 하는 과정에서 주변 단어를 고려, 같은 단어라도 문맥에 따라 다른 Vector를 가지게 됨

Word2Vec, FastText, …



- Skip-gram 혹은 CBOW (주로 Skip-gram)
- 주변 단어의 분포를 통해 중심 단어의 출현 확률을 학습
- 함께 출현한 단어들 간의 유사도는 높이고,
  함께 출현하지 않은 단어들 간의 유사도는 낮추는 방향으로 학습이 진행
- 결과물은 “임베딩 단어 벡터”

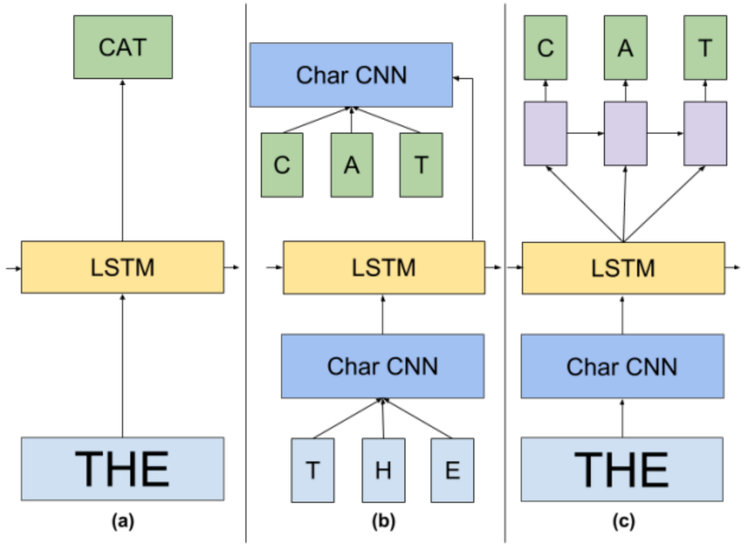
Then How About ELMo

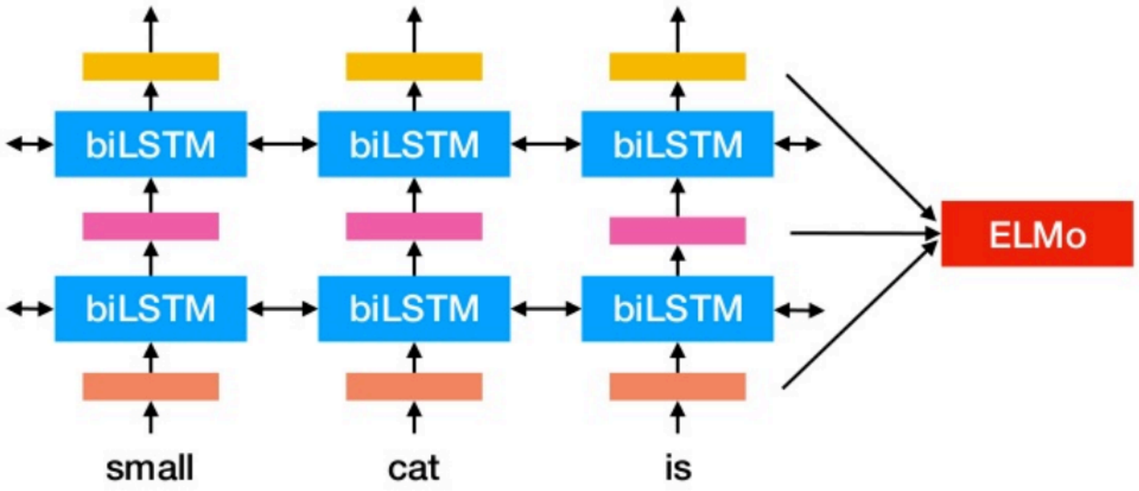


- biLSTM Layer의 Layer Representations의 Combination이 결과물
- Inference를 위한 Input이 Word가 아닌 ‘Sentence’
- 따라서 모델에 들어가는 Sequence length 만큼의 맥락을 고려하여 Vectorization 가능

ELMo 소개

Training





- ELMo의 Input을 Character로 넣을 건지, Word로 넣을 것인지는 선택 가능
  (TF 구현체 참고)
- Character로 Input을 준다면 token layer에서 Character CNN 사용
- Word로 Input을 준다면, 미리 Embedding 된 자료 필요

Training

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

- t1, t2, … , tk-1 로 tk가 나타날 확률을 예측

< Forward >

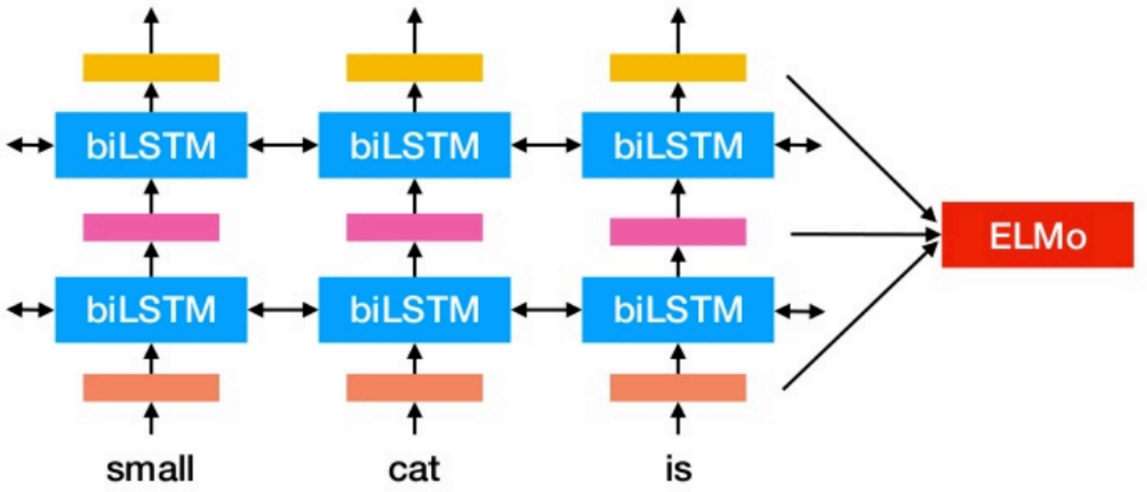
$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_{k+1}, t_{k+2}, \ldots, t_N).$$

- tk+1, tk+2, … , tN 으로 tk가 나타날 확률을 예측

< Backward >

Training

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



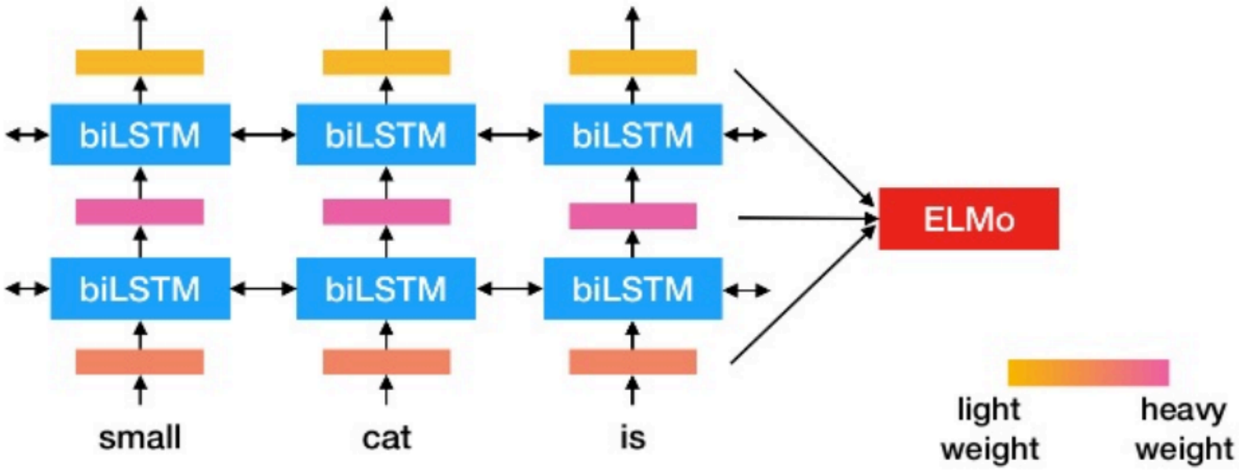
- biLM은 Forward와 Backward의 log-likelihood를 합하고, 그를 maximize 하는 방향으로 학습
- 이는 다음 biLSTM Layer의 Input이 됨

ELMo는 뭘 뱉나요

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

- L: biLM의 레이어 수
- $\mathbf{h}_k^{LM,0}$: token layer
- $\mathbf{h}_k^{LM,j}$: each biLSTM layer
- 각 token마다, 각 biLSTM layer의 Representation을 얻고 그 Representation들을 조합해 최종 벡터를 반환

ELMo는 뭘 뱉나요



- s = Softmax(w)
- γ : Vector Scale을 결정하는 상수

$$E(R_k; \mathbf{w}, \gamma) = \gamma \sum_{j=0}^{L} s_j \mathbf{h}_k^{LM,j}.$$

Evaluation

얼마나 좋은가요

Task	Previous SOTA		Our baseline	ELMo + baseline	Increase (absolute/ relative)
SNLI	McCann et al. (2017)	88.1	88.0	88.7 ± 0.17	0.7 / 5.8%
SQuAD^2	r-net Wang et al. (2017)	84.3	81.1	85.3	4.2 / 22.2%
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%

Question Answering, Semantic Role Labeling, Sentiment Analysis 등의
대부분 NLP Task에서 성능 향상을 보임

얼마나 좋은가요

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

	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 .

문맥에 따라 같은 단어라도 다른 Vector로 Embedding 되는 것을 확인 가능

감사합니다!