

# Deep contextualized word representations

김웅희

- Author
  - Peters, M. E.
  - Neumann, M
  - Iyyer, M.
  - Gardner, M.
  - Clark, C.
  - Lee, K., & Zettlemoyer, L.
- Title of Conference(Journal)
  - NAACL 2018

---

# • Introduction

- Existing models like Word2Vec or Glove provides a single context-independent representations
- Rest of models like TagLM, utilizes representations from final layer of the model which essentially leave out information from the lower layers
- Higher-level LSTM states capture context-dependent aspects of word meaning while lower-level states model aspects of syntax

---

# • Language Model

- 문장의 확률을 통해 다음의 단어를 예측하는 것이 목표

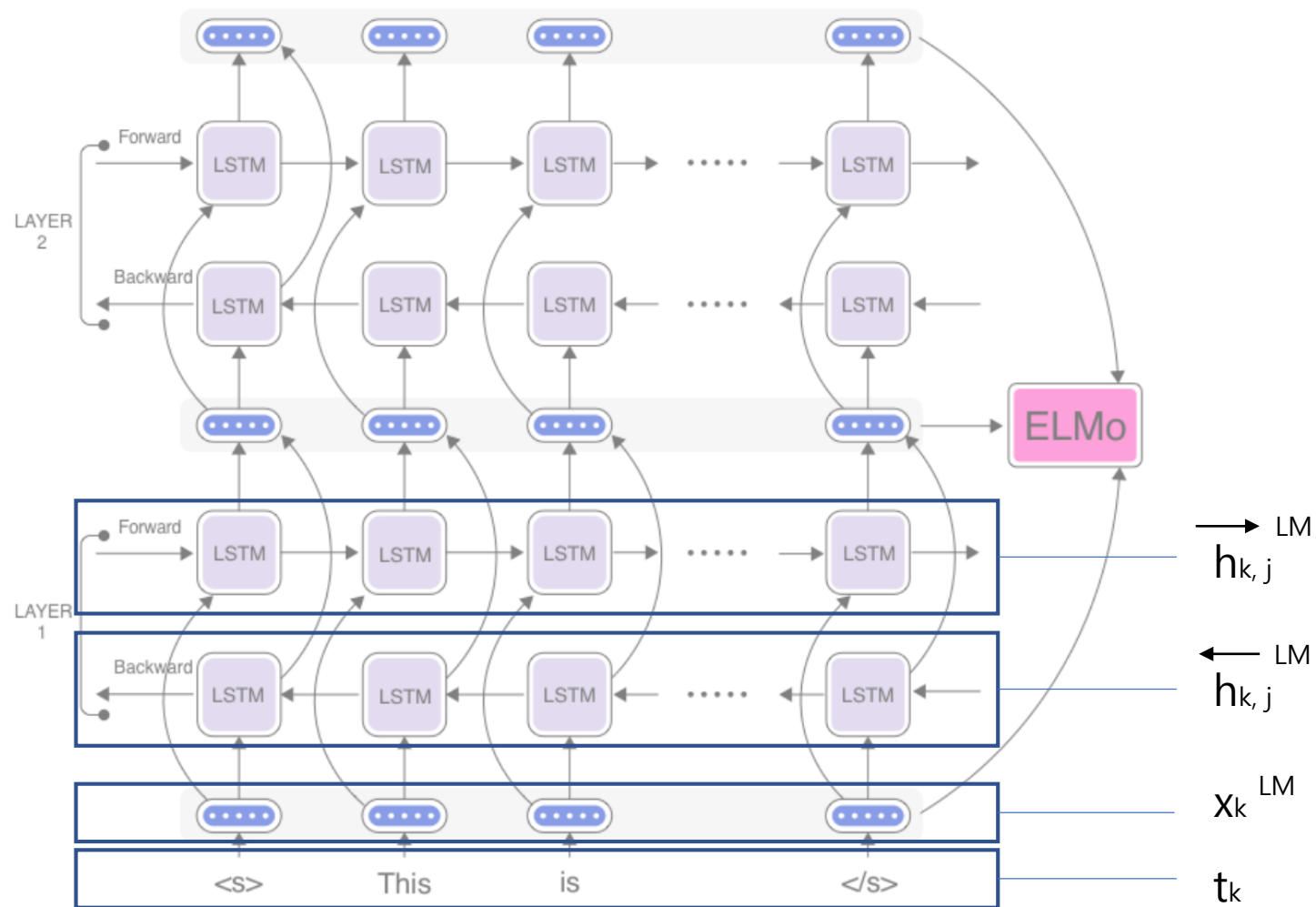
$$P(W) = P(w_1, w_2, w_3, w_4, w_5, \dots, w_n)$$

- 기본적으로는 조건부 확률을 통해 다음에 오는 단어를 구함

$$P(w_5 | w_1, w_2, w_3, w_4)$$

$$P(W) = P(w_1, w_2, w_3, w_4, w_5, \dots, w_n) = \prod_{i=1}^n P(w_i | w_1, \dots, w_{i-1})$$

# • ELMo



$t_k$ : 토큰  
 $k$ :  $k$ 번째 토큰  
 $j$ :  $j$ 번째 레이어

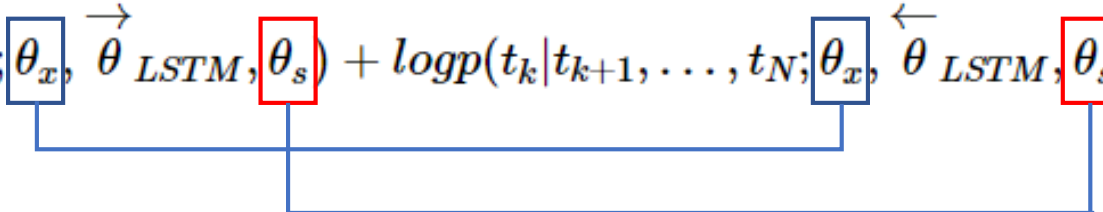
# • ELMo

- Bidirectional Language Models

$$p(t_1, t_2, \dots, t_N) = \sum_{k=1}^N p(t_k | t_1, t_2, \dots, t_{k-1})$$

$$p(t_1, t_2, \dots, t_N) = \sum_{k=1}^N p(t_k | t_{k+1}, t_{k+2}, \dots, t_N)$$

- 두 방향으로부터 나오는 로그 우도의 합을 최대화하도록 학습시킴

$$\sum_{k=1}^N (\log p(t_k | t_1, \dots, t_{k-1}; \theta_x, \vec{\theta}_{LSTM}, \theta_s) + \log p(t_k | t_{k+1}, \dots, t_N; \theta_x, \overleftarrow{\theta}_{LSTM}, \theta_s))$$


# • ELMo

여기서 말하는 태스크는 문서분류, QA, 개체명 인식 등의 임의의 태스크

$$ELMo_k^{task} = E(R_k; \theta^{task}) = \gamma^{task} \sum_{j=0}^L s_j^{task} h_{k,j}^{LM}$$

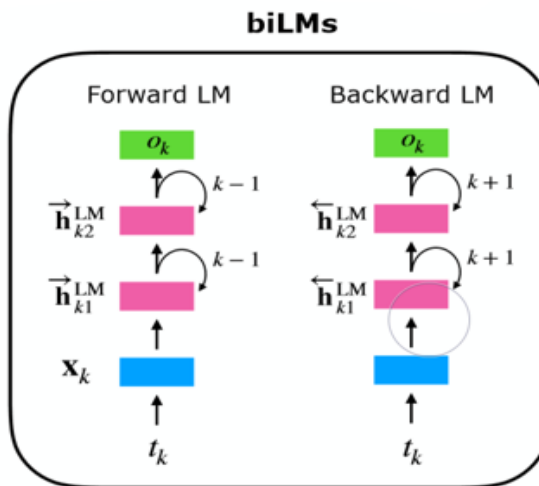


j 번째 레이어가 해당 태스크 수행에 얼마나 중요한지 가리키는 스칼라값

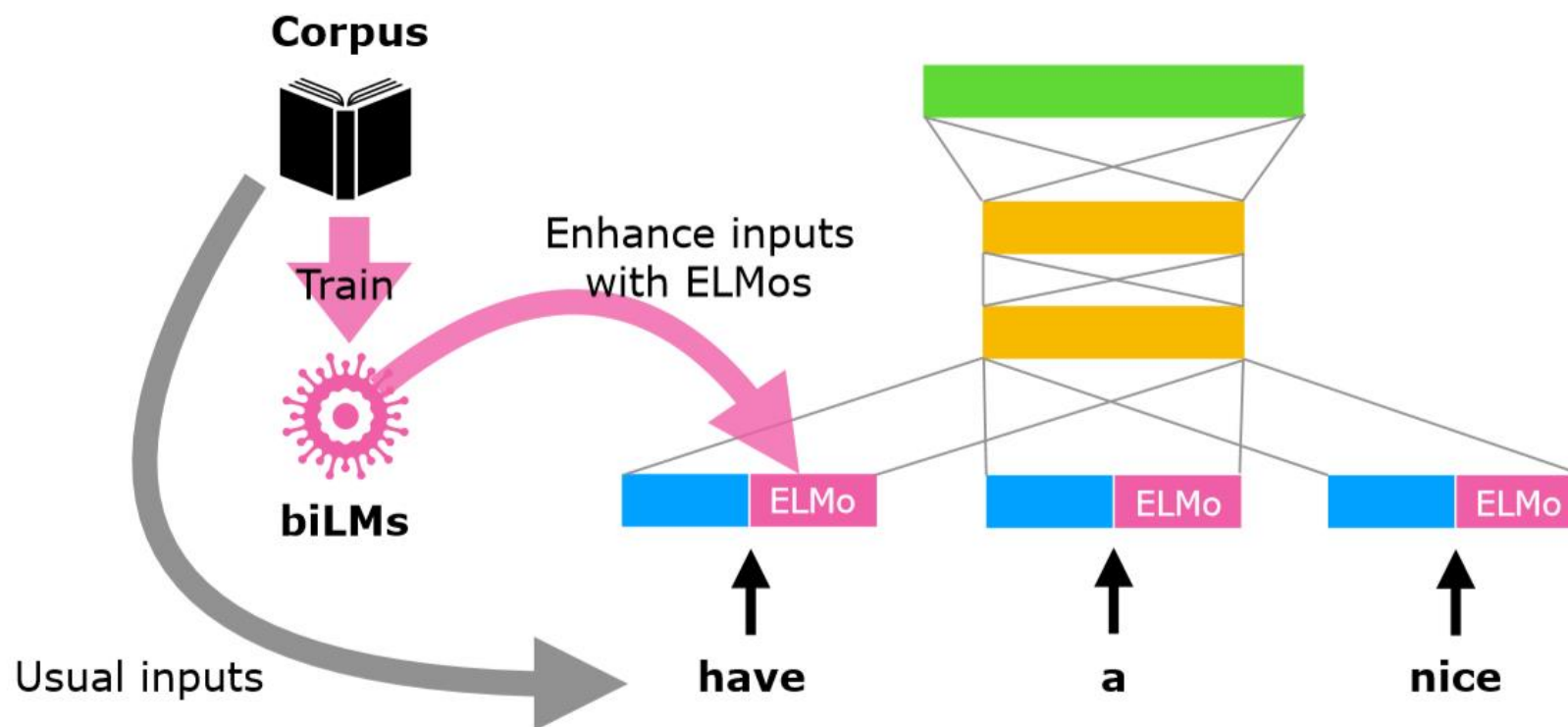
$$ELMo_k^{task} = \gamma^{task} \times \sum \left\{ \begin{array}{l} s_2^{task} \times h_{k2}^{LM} \\ s_1^{task} \times h_{k1}^{LM} \\ s_0^{task} \times h_{k0}^{LM} \end{array} \right. \quad \text{Concatenate hidden layers} \quad \left[ \vec{h}_{kj}^{LM}; \overleftarrow{h}_{kj}^{LM} \right]$$

( $(\mathbf{x}_k; \mathbf{x}_k)$ )

해당 태스크가 얼마나 중요한지를 뜻하는 가중치 - 전체적인 Vector의 크기를 결정하는 용도로 사용되며, 값이 커질수록 중요도가 높아지는 방식



- ELMo



## • 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	$88.7 \pm 0.17$	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)	$91.93 \pm 0.19$	90.15	$92.22 \pm 0.10$	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	$54.7 \pm 0.5$	3.3 / 6.8%

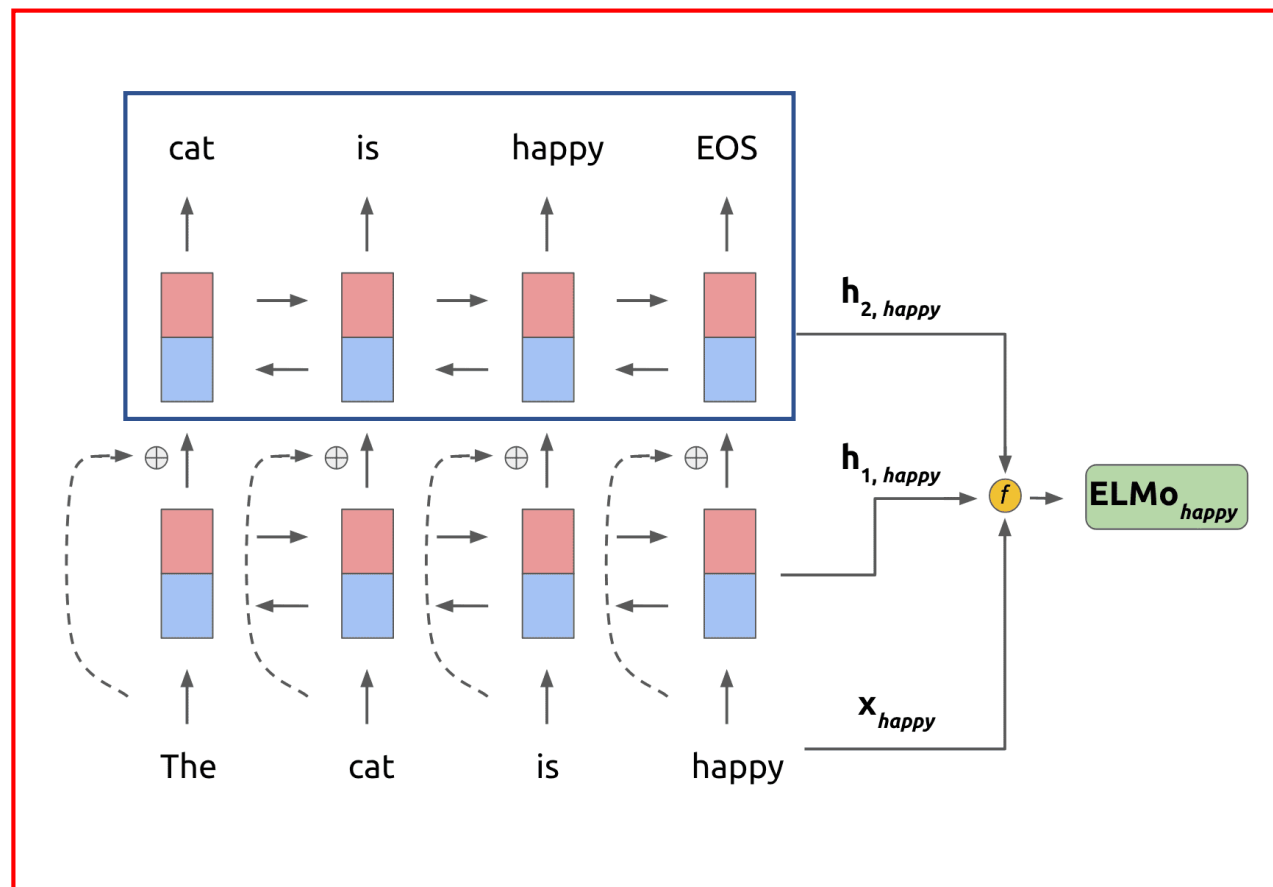
Table 1: Test set comparison of ELMo enhanced neural models with state-of-the-art single model baselines across six benchmark NLP tasks. The performance metric varies across tasks – accuracy for SNLI and SST-5;  $F_1$  for SQuAD, SRL and NER; average  $F_1$  for Coref. Due to the small test sizes for NER and SST-5, we report the mean and standard deviation across five runs with different random seeds. The “increase” column lists both the absolute and relative improvements over our baseline.



# • Analysis

Task	Baseline	Last Only	All layers	
			$\lambda=1$	$\lambda=0.001$
SQuAD	80.8	84.7	85.0	<b>85.2</b>
SNLI	88.1	89.1	89.3	<b>89.5</b>
SRL	81.6	84.1	84.6	<b>84.8</b>

Table 2: Development set performance for SQuAD, SNLI and SRL comparing using all layers of the biLM (with different choices of regularization strength  $\lambda$ ) to just the top layer.



---

- Analysis

- ELMo의 위치

Task	Input Only	Input & Output	Output Only
SQuAD	85.1	<b>85.6</b>	84.8
SNLI	88.9	<b>89.5</b>	88.7
SRL	<b>84.7</b>	84.3	80.9

Table 3: Development set performance for SQuAD, SNLI and SRL when including ELMo at different locations in the supervised model.

## • Analysis

### • 기존 워드 임베딩과의 비교

Source		Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular <u>play</u> 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 <u>play</u> .
	Olivia De Havilland signed to do a Broadway <u>play</u> for Garson {...}	{...} they were actors who had been handed fat roles in a successful <u>play</u> , 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.

# • Analysis

## • ELMo에서 representation이 내포하는 것

Second Layer가  
semantic한 정보를 더  
잘 알아냄

Model	F <sub>1</sub>
WordNet 1st Sense Baseline	65.9
Raganato et al. (2017a)	69.9
Iacobacci et al. (2016)	<b>70.1</b>
CoVe, First Layer	59.4
CoVe, Second Layer	64.7
biLM, First layer	67.4
biLM, Second layer	69.0

Table 5: All-words fine grained WSD F<sub>1</sub>. For CoVe and the biLM, we report scores for both the first and second layer biLSTMs.

First Layer는 syntactic한 정보를 더  
내포함

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

Table 6: Test set POS tagging accuracies for PTB. For CoVe and the biLM, we report scores for both the first and second layer biLSTMs.

---

## • Conclusion

- biLMs로부터 높은 성능의 문맥 의존적인 representations를 제안
- 기존의 워드 임베딩에 ELMo 표현을 더해주는 것만으로 높은 성능을 보임
- 모든 BiLM 레이어들을 이용해 정보 손실을 최소화하여 syntactic한 정보와 semantic한 정보를 얻음