ELMo

Matthew E. Peters et al., Deep contextualized word representations, 2018.

허무지 AI Robotics KR NLP Study

1. Introduction

ELMo:

Embeddings from **L**anguage **Mo**dels

Neural Network

Embedding

Deep Contextualized Word Representations

y=f(word, context)

슬라이드 출처https://youtu.be/9JfGxKkmBc0

1. Introduction

ELMo의 목표 : 입력을 위한 임베딩을 할 때에도 단어의 맥락을 파악

(1) 문법론, 의미론에서 단어의 사용이 복잡하다는 특징
(2) 문맥에서 단어의 다의적인 의미를 파악

f(단어)

손

f(단어, 맥락)

손을 씻고 밥을 먹어야지.

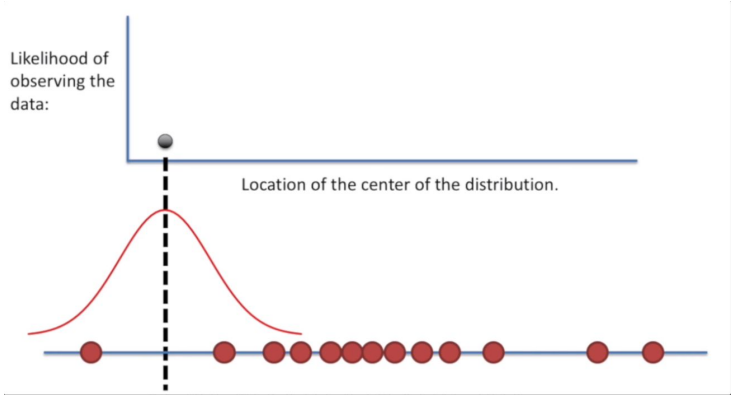
손이 크니 계란도 한 판을 삶는구나.

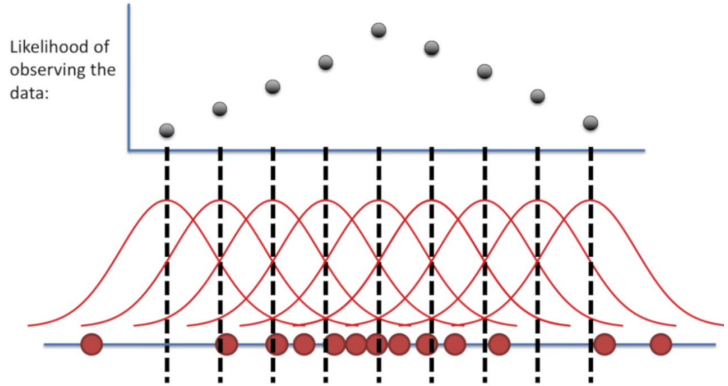
2. Background

Maximum Likelihood Estimation

우도(가능도):
관측값이 해당 확률분포에서 나왔을 확률

최대우도법:
관측값에 대한 총 가능도가 최대가 되게 하는 분포





※정규분포라고 가정한 그림.

출처 https://jjangjjong.tistory.com/41

2. Background

Maximum Likelihood Estimation

1) 모델을 설정한다.
2) 모델에서 데이터의 발생확률 식을 정한다. 언어 생성 모델 확률
3) 확률을 최대로 높이는 모델 변수를 구한다. 파라미터(모수) Θ

출처 https://youtu.be/sOtkPm_1GYw

3.1 Bidirectional language models

2) 모델에서 데이터의 발생확률 식을 정한다.

토큰이 주어졌을 때 Language model에서 문장 발생 확률 식

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

Forward language model

t_k

모가지가 길어서 슬픈 짐승이여 언제나 점잖은 편 말이 없구나

Backward language model

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

3.1 Bidirectional language models.

3) 확률을 최대로 높이는 모델 변수를 구한다.

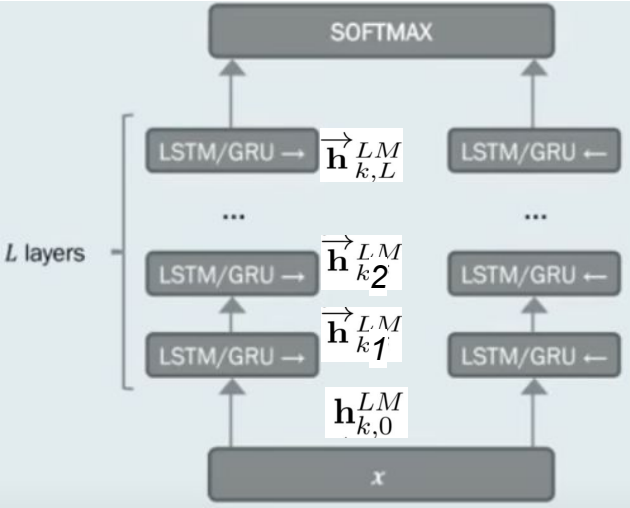
$$\sum_{k=1}^{N} \Big(\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \Big)$$

; 앞의 발생 확률 식에 따른 파라미터를 ; 뒤에 나타낸다.

최대우도법에서는 log를 씌워서 계산을 편하게 한다.
(logx를 미분하면 1/x)

토큰, 소프트맥스 파라미터는 방향에 관계없이 묶었다.

3.2 ELMo



ELMo는
biLM에서 등장하는 중간 매체 layer의 표현들을 특별하게 합친 것을 의미한다.

맥락을 고려하기 위해 독립적인 토큰이 아니라, 문장을 반영한 각 은닉층을 계산한다.

높은 LSTM 레이어는 문맥을,
낮은 LSTM은 문법적 의미를 가진다.

출처 https://youtu.be/9JfGxKkmBc0

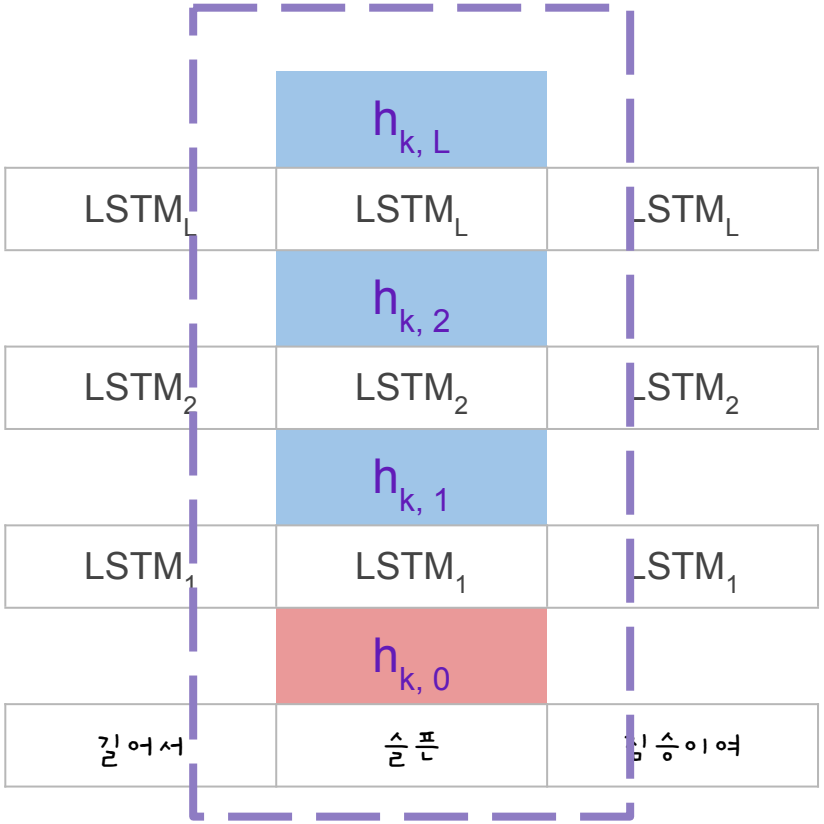
3.2 ELMo

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

$$\mathbf{x}_k^{LM} \longrightarrow \mathbf{h}_{k,0}^{LM}$$

$$\mathbf{h}_{k,j}^{LM} = [\overrightarrow{\mathbf{h}}_{k,j}^{LM}; \overleftarrow{\mathbf{h}}_{k,j}^{LM}]$$

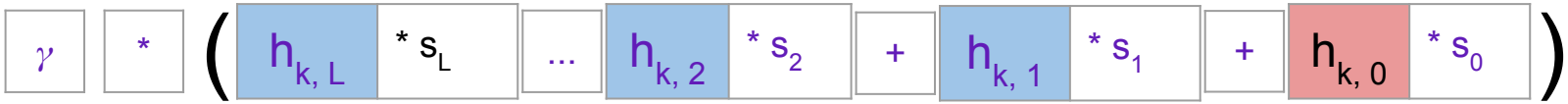
$$= \{\mathbf{h}_{k,j}^{LM} \mid j = 0, \ldots, L\}$$



3.2 ELMo

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

- : γ^{task} 전체 ELMo vector의 규모를 맞추는 파라미터
- s_j^{task} Softmax 정규화 가중치



3.3 Using biLMs for supervised NLP tasks

$$[\mathbf{x}_k; \mathbf{ELMo}_k^{task}]$$

대부분의 NLP model은 가장 낮은 층의 구조가 비슷한다. (토큰 x_k)

biLSTM의 가중치를 고정한 뒤,
ELMo vector와 토큰을 연결하여 입력값으로 사용한다.

Dropout을 사용하면 결과가 더 좋아짐.

출처 http://giyeon.blogspot.com/2019/01/elmo-deep-contextualized-word.html

5.2 Where to include ELMo?

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

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

- SQuAD (질의응답), SNLI(가설/전제 추론) Input과 Output 모두에 ELMo를 붙이는 경우 정확도가 더 높게 나타났다.
  Attention layers를 사용하여 ELMo를 연결하면 직접적으로 biLM의 표현을 사용한다.

- SRL(의미역 부착)에서는 Input에만 ELMo를 붙일 때 정확도가 더 높아졌다.
  구체적인 작업의 맥락이 biLM의 사용보다 중요하다.

5.3 the biLM's representations information

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*

Table 5: All-words fine grained WSD F_1. For CoVe and the biLM, we report scores for both the first and second layer biLSTMs.

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

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.

- 표 5 WSD: Word sense disambiguation

  단어 의미 중의성 해소
  동음 이의어를 구별할 수 있음.
  두 번째 레이어의 값이 더 높으므로
  문맥의 의미를 더 잘 반영한다.

- 표 6 POS tagging

  품사 추출
  첫 번째 레이어의 값이 더 높음.
  문법적인 의미는 낮은 레이어에서 더 잘 파악됨.

The end of document.