Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks

이봉석

•Author
- Daojian Zeng (Chinese Academy of Sciences)
- Kang Liu (Chinese Academy of Sciences)
- Yubo Chen (Chinese Academy of Sciences)
- Jun Zhao (Chinese Academy of Sciences)

•Title of Conference(Journal)
- EMNLP 2015

01. Introduction

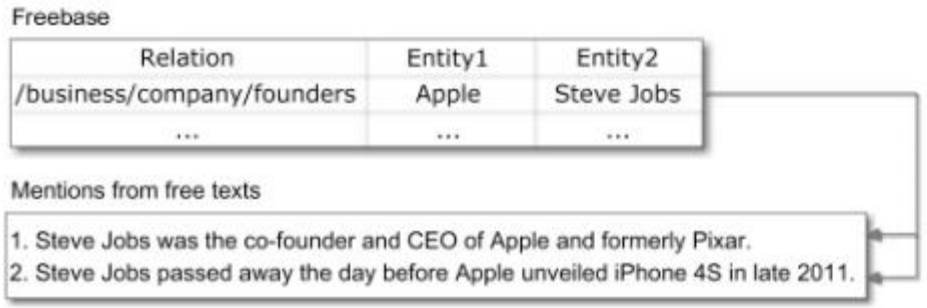


Figure 1: Training instances generated through distant supervision. Upper sentence: correct labeling; lower sentence: incorrect labeling.

•Relation extraction에서 하나의 도전 과제는 training examples을 만드는 것이다.
•Distant supervision이 이를 해결하는 하나의 방법이 될 수 있다.

•Distant supervision은 어떤 두 entity가 이미 알려진 knowledge base에서 relation을 가지고 있을 때, 이 두 entity가 등장한 모든 문장에 대해서 동일한 relation을 가진다고 가정하고 data를 만들어내는 것이다.

01. Introduction

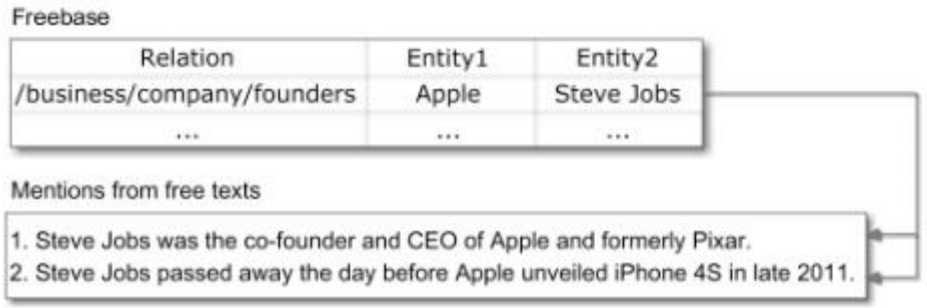


Figure 1: Training instances generated through distant supervision. Upper sentence: correct labeling; lower sentence: incorrect labeling.

- 문제점 1
  - distant supervision의 가정이 너무 강력하고 잘못된 label을 만들어낸다는 것이다. 즉, 두 entity가 언급된 문장이 반드시 knowledge base에서 나타내는 relation을 포함한다고 볼 수 없다.

Ex) Figure 1의 2 번째 문장에서 두 entitiy는 "/copany/founders"의 relation을 갖는다고 보기 어렵지만 distant supervision때문에 이런 noisy data가 생성된다.

01. Introduction

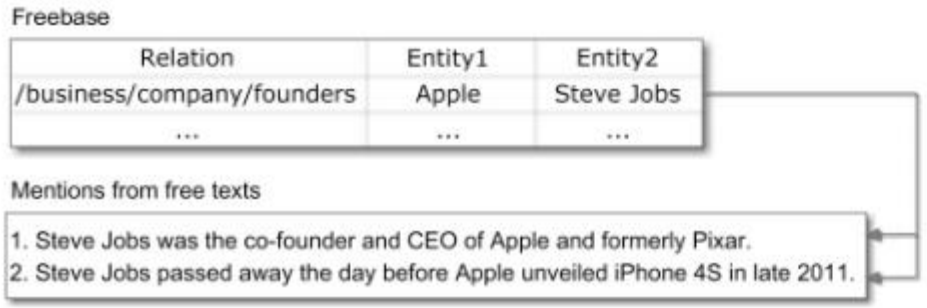


Figure 1: Training instances generated through distant supervision. Upper sentence: correct labeling; lower sentence: incorrect labeling.

- 문제점 2
  - distant supervision으로 데이터를 얻을 때 정교하게 디자인 된 feature를 가지고 model에 적용시킨다는 것이다.
  - 문제의 대표적 원인은 이미 존재하는 NLP tool을 사용하는데 있다.
  - 불가피하지만 NLP tool을 사용하면서 그에 내재된 error가 feature에 전달이 된다.

01. Introduction

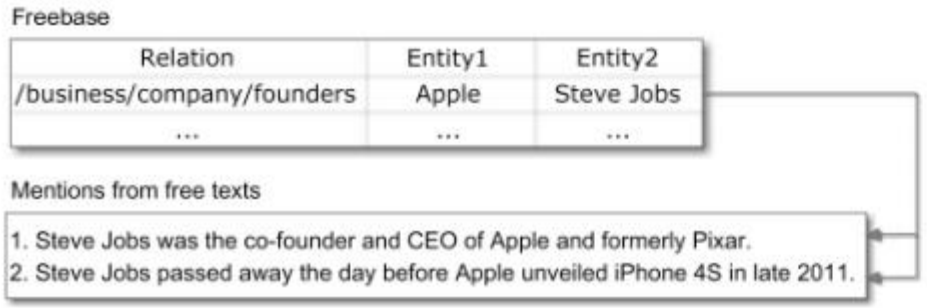


Figure 1: Training instances generated through distant supervision. Upper sentence: correct labeling; lower sentence: incorrect labeling.

- 문제점 1 -> multi-instance learning
- 문제점 2 -> convolutional architecture

해결!

02. Method

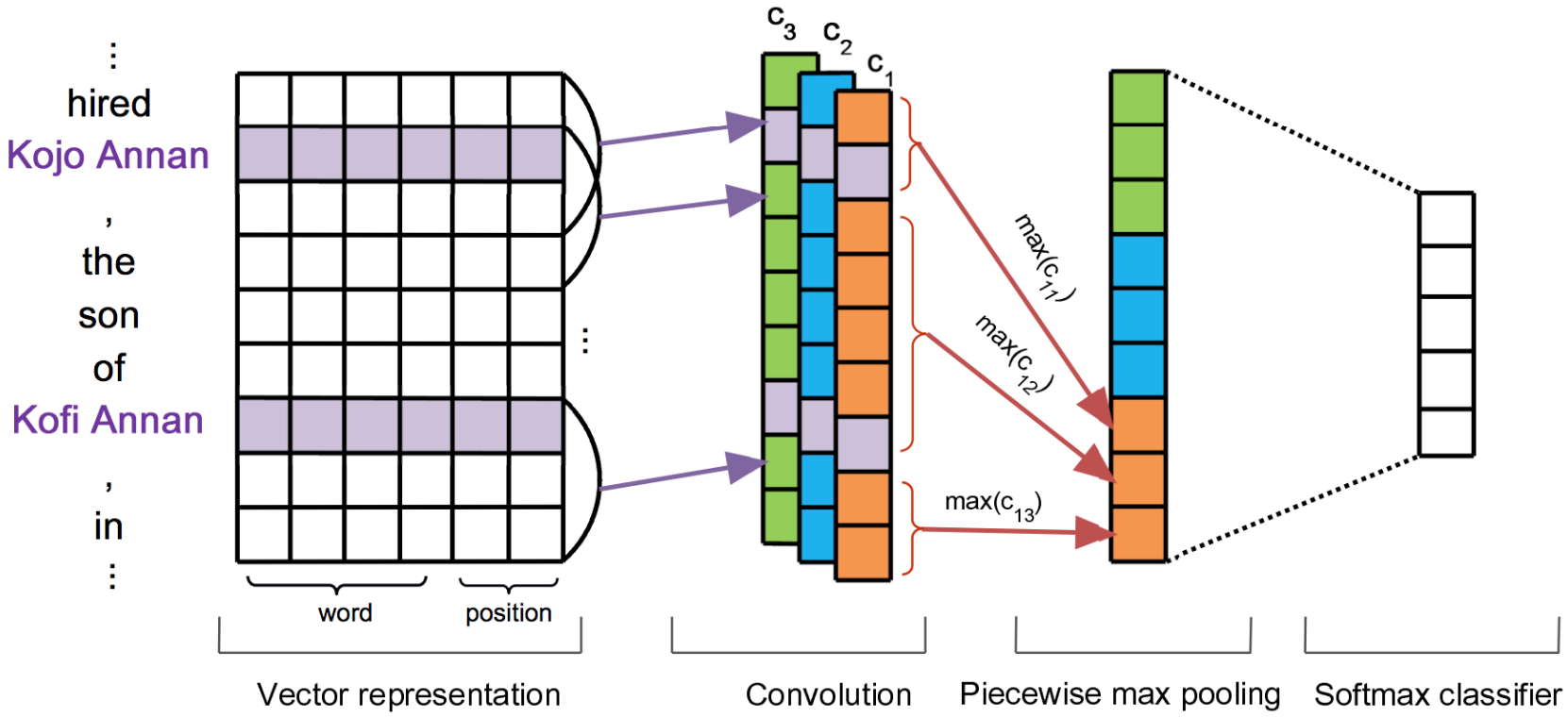
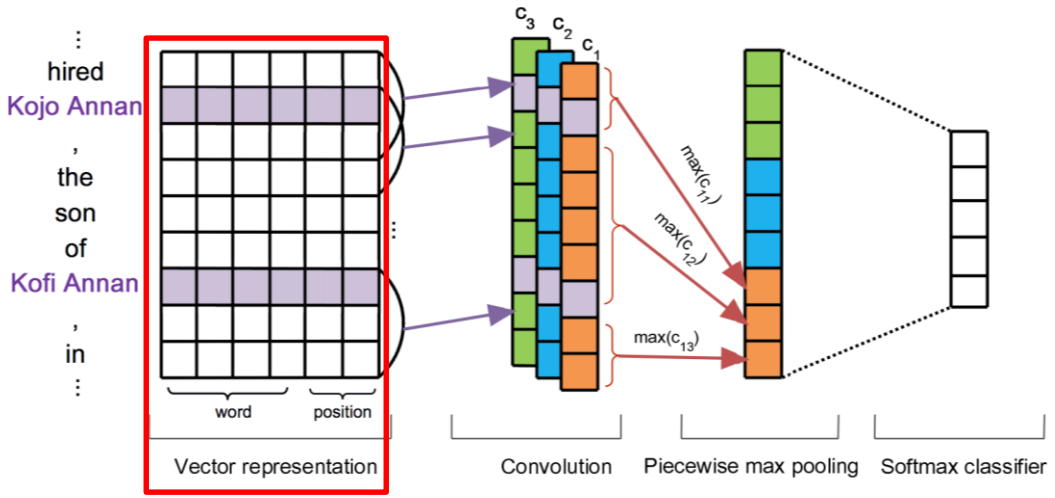


Figure 3: The architecture of PCNNs (better viewed in color) used for distant supervised relation extraction, illustrating the procedure for handling one instance of a bag and predicting the relation between *Kojo Annan* and *Kofi Annan*.

02. Method

- Vector representation
  - word -> word embedding
  - position -> position embedding



- Position embedding

문장 내 두 개체와 개체가 아닌 단어들 간의 상대적 거리를 n 차원의 벡터로 임베딩 한 것이다. 예를 들어, 그림 3에서 보는바와 같이 '설립'이라는 단어가 '마크_저커버그' 개체로 부터 3단어 만큼 떨어져 있고, '페이스북' 개체로부터 1단어만큼 떨어져있다. 이 상대적 거리를 n 차원의 벡터로 임베딩하고 , 그 값을 모델 학습의 입력 벡터 중 일부로 사용한다.

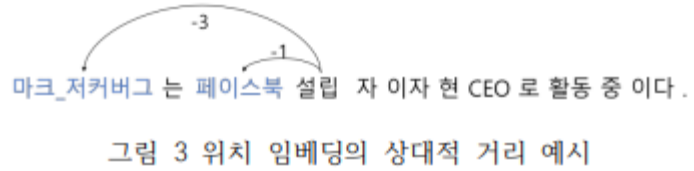
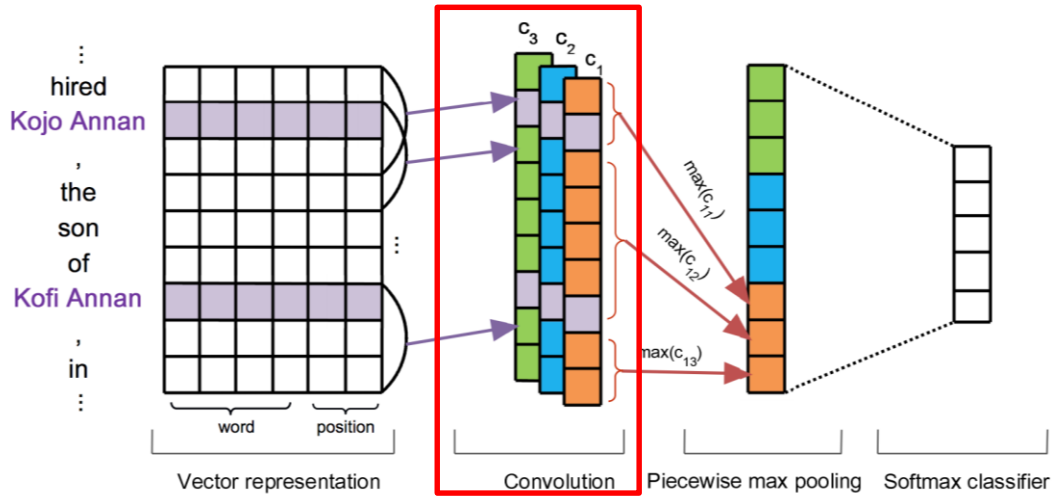


그림 3 위치 임베딩의 상대적 거리 예시

02. Method

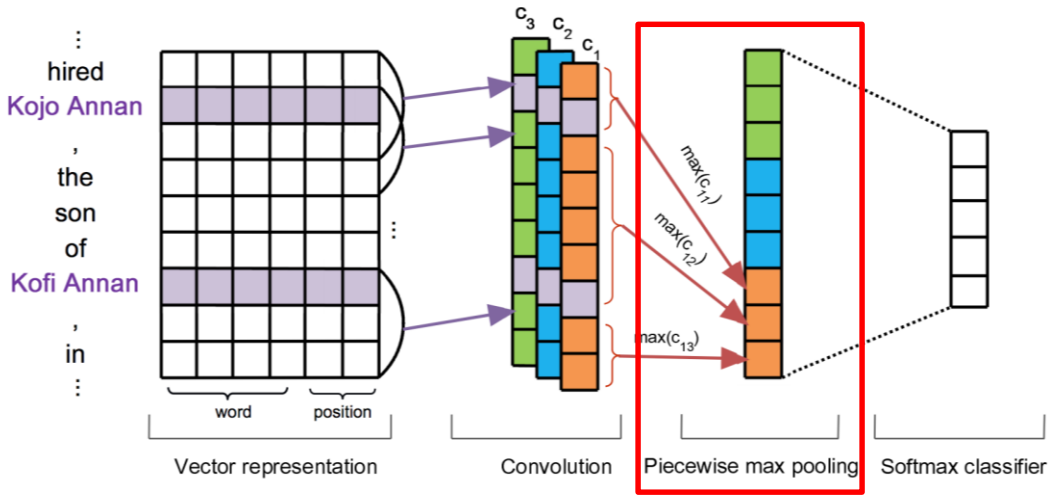
- Convolution



- NLP tool를 사용하지 않으므로 그에 내재된 error가 feature에 전달되는 것을 방지한다.

02. Method

- Piecewise Max Pooling



- Single max pooling는 relation extraction을 하기에 불충분하다.
- Hidden layer의 size 가 급격하고 거칠게 줄어들어 고운 feature를 얻기 힘들다.

-> 전체 문장을 한순간에 하나의 값으로 뭉뚱그리기 때문에 feature가 뭉개진다.

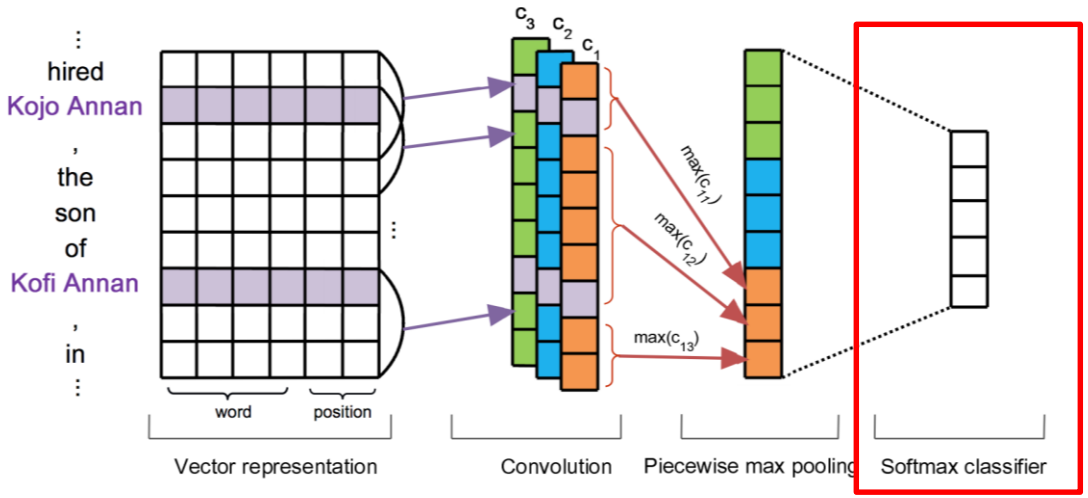
- 또한 두 entity에 대한 structural information을 잡아내기 어렵다.
- piecewise max pooling은 위의 단점을 보완하는 방법으로서, 두 entity를 기준으로 문장을 3개의 segment로 나눈 뒤 각 segment 별로 max pooling을 해서 3차원의 벡터를 얻을 수 있음
- 이렇게 추출된 3차원 벡터들을 쭉 이어 붙이고(concatenate) non-linear activation function(tanh)를 거쳐 다음 layer에 전달된다.

02. Method

- Multi-instance Learning

Algorithm 1 Multi-instance learning

1: Initialize θ. Partition the bags into mini-batches of size b_s.
2: Randomly choose a mini-batch, and feed the bags into the network one by one.
3: Find the j-th instance m_i^j $(1 \leq i \leq b_s)$ in each bag according to Eq. (9).
4: Update θ based on the gradients of m_i^j $(1 \leq i \leq b_s)$ via Adadelta.
5: Repeat steps 2-4 until either convergence or the maximum number of epochs is reached.



- BAGS=> T bags $\{M_1, M_2, \cdots, M_T\}$ $M_i = \{m_i^1, m_i^2, \cdots, m_i^{q_i}\}$. T : 릴레이션의 갯수
- batch를 구성하는 모든 instance들은 network를 거쳐서 output(probability by softmax)을 구한다.
- 각 i번째 bag에서 i번째 label의 output(확률)이 가장 큰 instance에 대해 cross-entropy를 구해서 네트워크 parameter를 업데이트 함

=> 확실한 instance와 relation에 대해서만 학습을 진행하게 되므로 wrong label problem을 완화시킬 수 있다.

03. Experiments

Top N	Mintz	MultiR	MIML	PCNNs+MIL
Top 100	0.77	0.83	0.85	**0.86**
Top 200	0.71	0.74	0.75	**0.80**
Top 500	0.55	0.59	0.61	**0.69**
Average	0.676	0.720	0.737	**0.783**

- Mintz represents a traditional distant supervision-based model that was proposed by (Mintz et al., 2009).
- MultiR is a multi-instance learning method that was proposed by (Hoffmann et al., 2011).
- MIML is a multi-instance multilabel model that was proposed by (Surdeanu et al., 2012).