Bidirectional LSTM-CRF Models for Named Entity Recognition

안건이
고려대학교 DMQA 연구실
2019/01/31

Bidirectional LSTM-CRF Models for Sequence Tagging

Zhiheng Huang
Baidu research
huangzhiheng@baidu.com

Wei Xu
Baidu research
xuwei06@baidu.com

Kai Yu
Baidu research
yukai@baidu.com

Abstract

In this paper, we propose a variety of Long Short-Term Memory (LSTM) based models for sequence tagging. These models include LSTM networks, bidirectional LSTM (BI-LSTM) networks, LSTM with a Conditional Random Field (CRF) layer (LSTM-CRF) and bidirectional LSTM with a CRF layer (BI-LSTM-CRF). Our work is the first to apply a bidirectional LSTM CRF (denoted as BI-LSTM-CRF) model to NLP benchmark sequence tagging data sets. We show that the BI-LSTM-CRF model can efficiently use both past and future input features thanks to a bidirectional LSTM component. It can also use sentence level tag information thanks to a CRF layer. The BI-LSTM-CRF model can produce state of the art (or close to) accuracy on POS, chunking and NER data sets. In addition, it is robust and has less dependence on word embedding as compared to previous observations.

Keywords : BI-LSTM-CRF, Named Entity Recognition, Sequence Tagging

---

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

목차

- Introduction
- POS Tagging
- B-I-O Tagging
- Word Embedding
- Models
- Conditional Random Field (CRF)
- Result

Armstrong

Named Entity Recognition

Person

Armstrong?

ARMSTRONG
River Hawks
BOWLING

Sports Team

Armstrong
WORLD INDUSTRIES

Company

암스트롱 강
Armstrong River
미국
55731 미네소타
강

Location

Named Entity Recognition

Person

Armstrong
WORLD INDUSTRIES

Company

Armstrong

Armstrong landed the Eagle on the Moon

ARMSTRONG
River
Hawks
BOWLING

Sports Team

암스트롱 강
Armstrong River
미국
55731 미네소타
강

Location

Named Entity Recognition

Person

Armstrong
WORLD INDUSTRIES

Company

Armstrong

Armstrong landed the Eagle on the Moon

ARMSTRONG
River
Hawks
BOWLING

Sports Team

암스트롱 강
Armstrong River
미국
55731 미네소타
강

Location

Named Entity Recognition

Person

Armstrong WORLD INDUSTRIES

Company

Armstrong

Context

Armstrong landed the Eagle on the Moon

ARMSTRONG River Hawks BOWLING

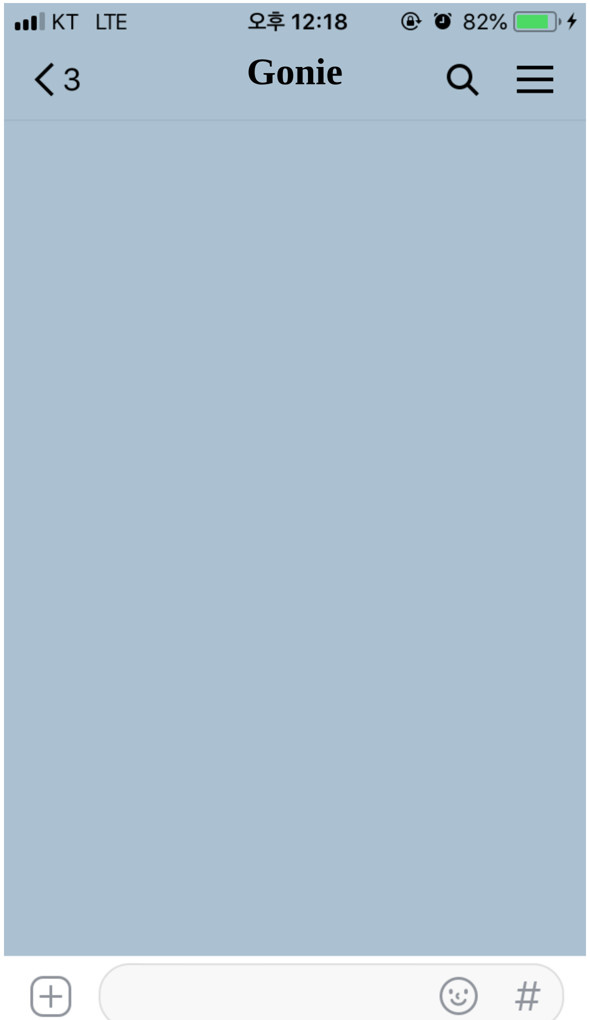
Sports Team

암스트롱 강
Armstrong River
미국
55731 미네소타
강

Location

Text Classification VS Named Entity Recognition

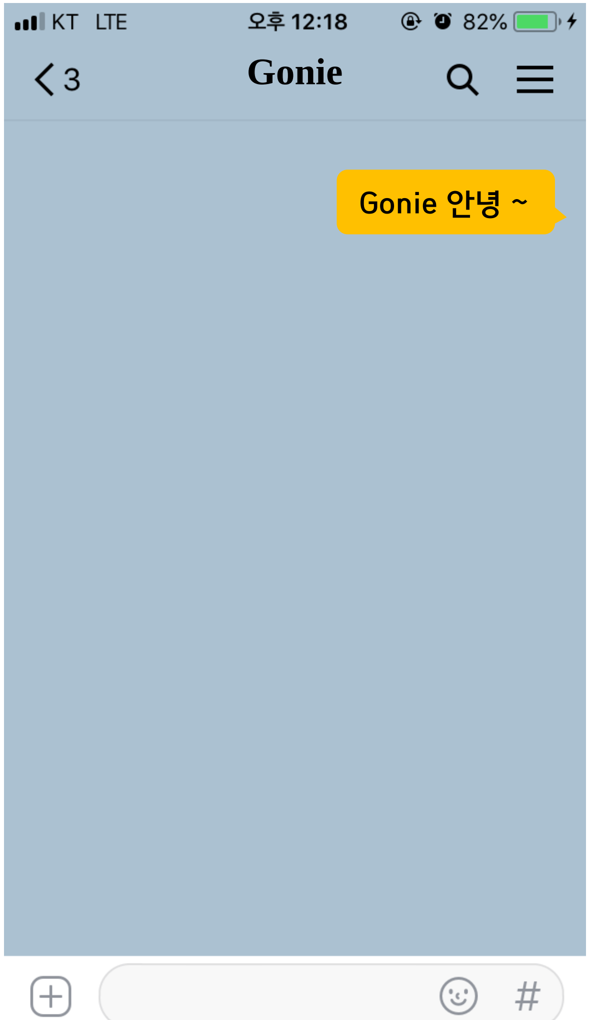
Gonie ChatBot



ChatBot Area

Text Classification VS Named Entity Recognition

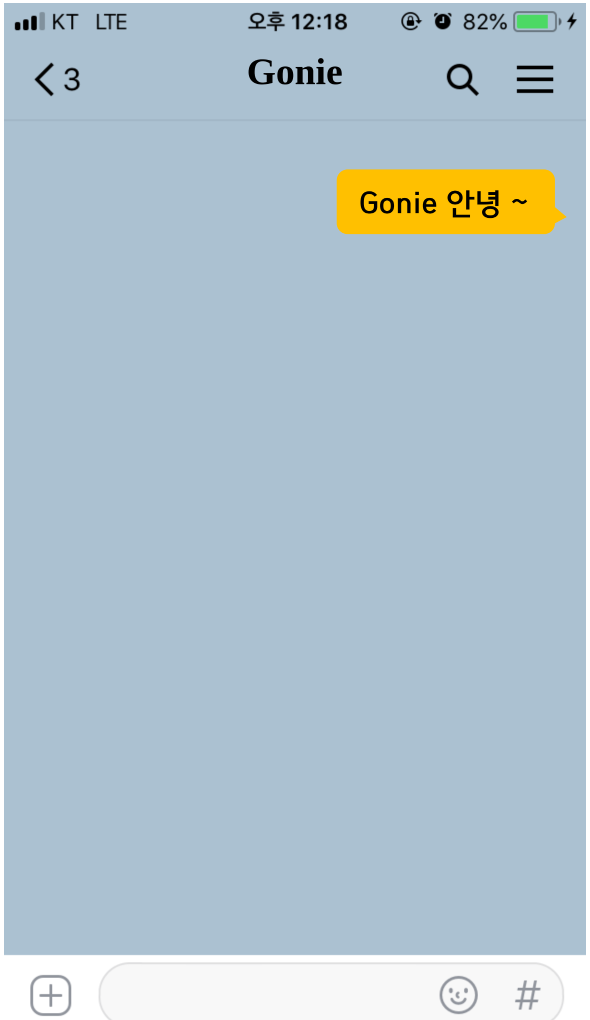
Gonie ChatBot



ChatBot Area

Text Classification VS Named Entity Recognition

Gonie ChatBot

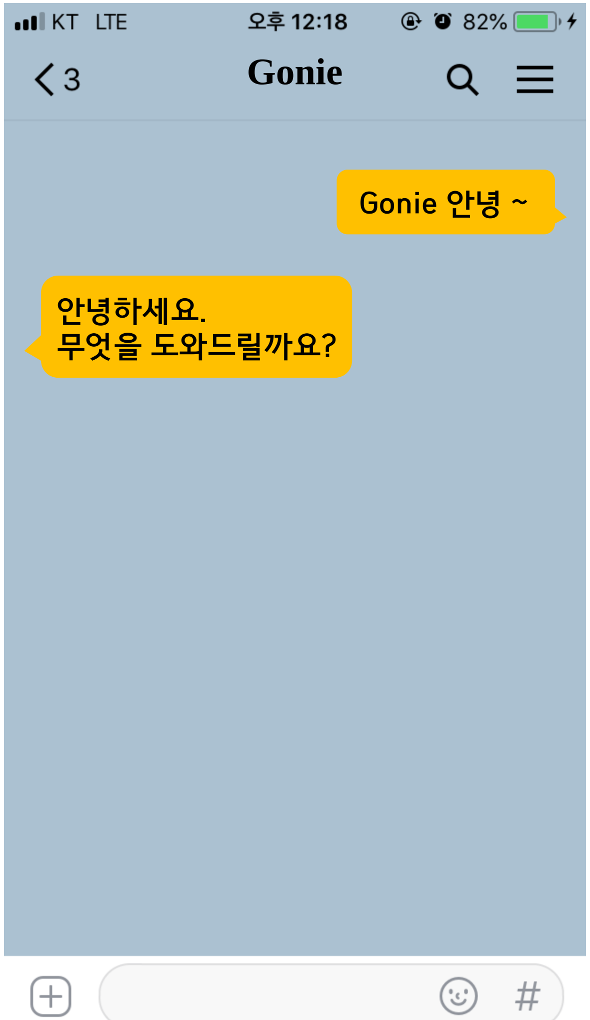


ChatBot Area

- 'Gonie 안녕 ~ ' → Text Classification → class_인사

Text Classification VS Named Entity Recognition

Gonie ChatBot



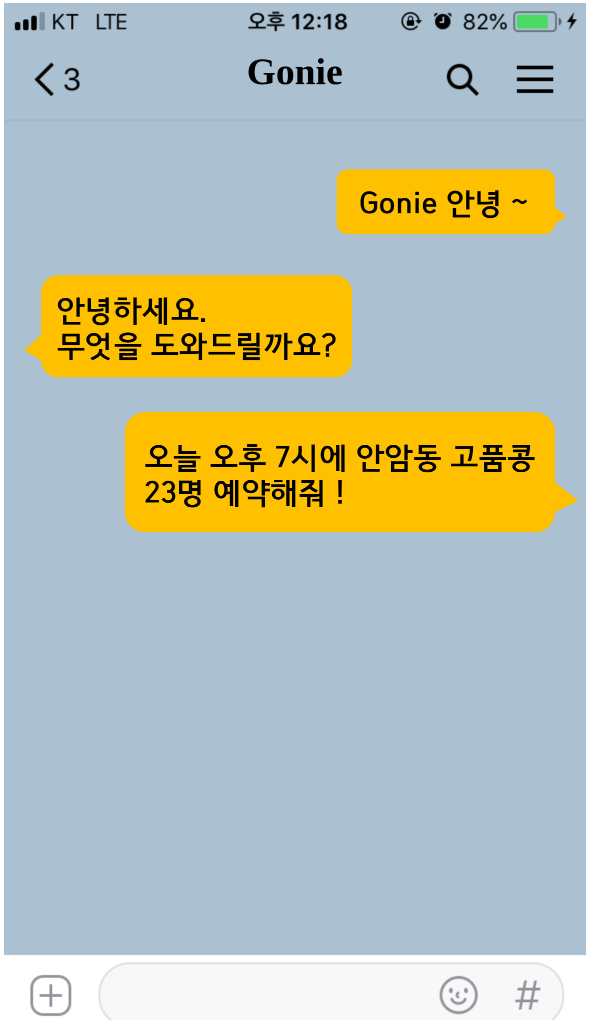
ChatBot Area

- 'Gonie 안녕 ~ ' → Text Classification → class_인사

  Action: 안녕하세요. 무엇을 도와드릴까요?

Text Classification VS Named Entity Recognition

Gonie ChatBot



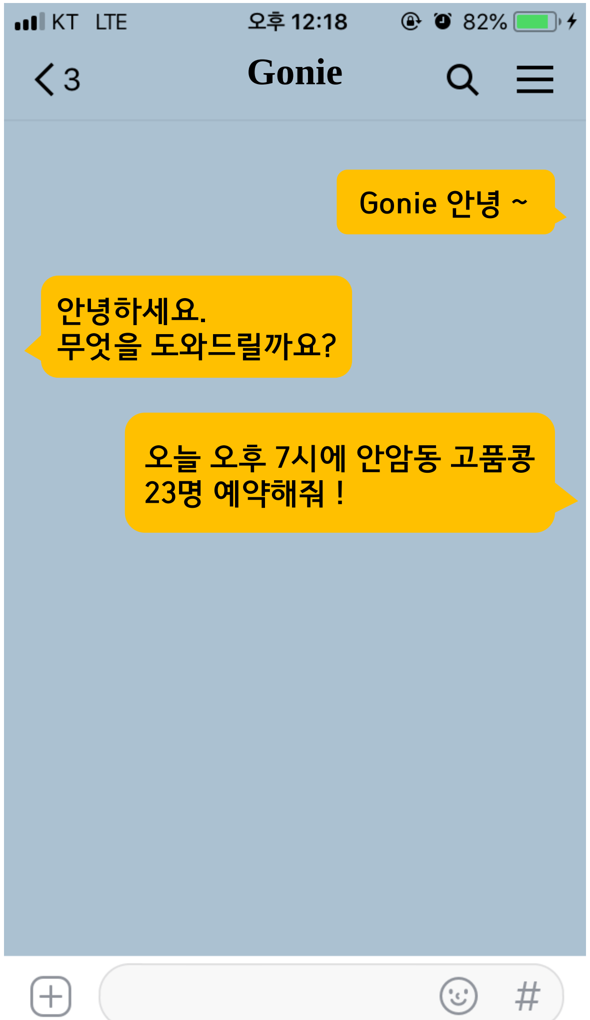
ChatBot Area

- 'Gonie 안녕 ~ ' → Text Classification → class_인사

  Action: 안녕하세요. 무엇을 도와드릴까요?

Text Classification VS Named Entity Recognition

Gonie ChatBot



ChatBot Area

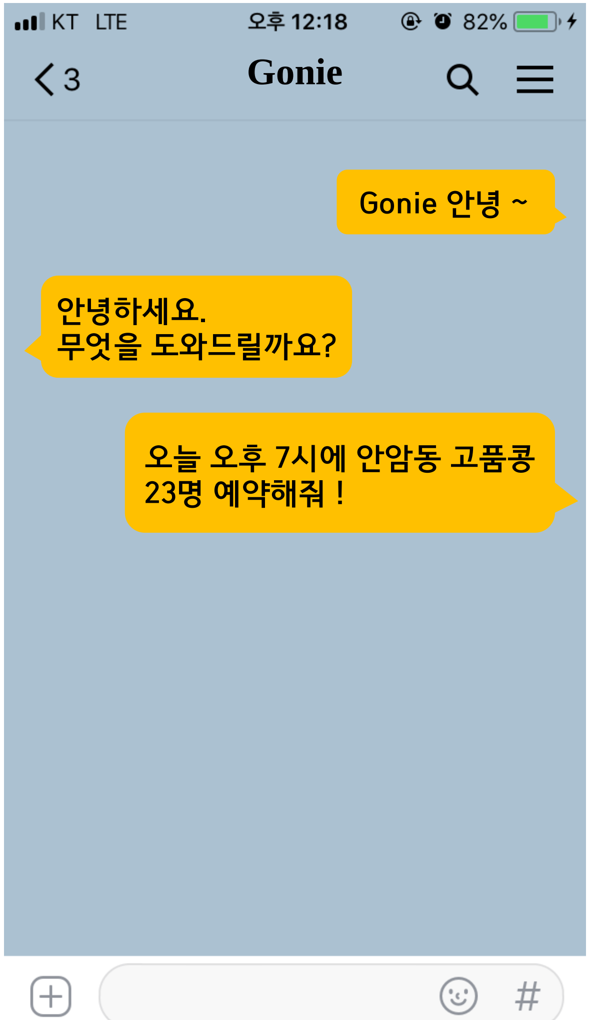
- ‘Gonie 안녕 ~ ’ → Text Classification → class_인사

  Action: 안녕하세요. 무엇을 도와드릴까요?

- ‘오늘 오후 7시에 안암동 고품콩 23명 예약해줘 !’ → Text Classification → class_예약

Text Classification VS Named Entity Recognition

Gonie ChatBot



ChatBot Area

- 'Gonie 안녕 ~ ' → Text Classification → class_인사

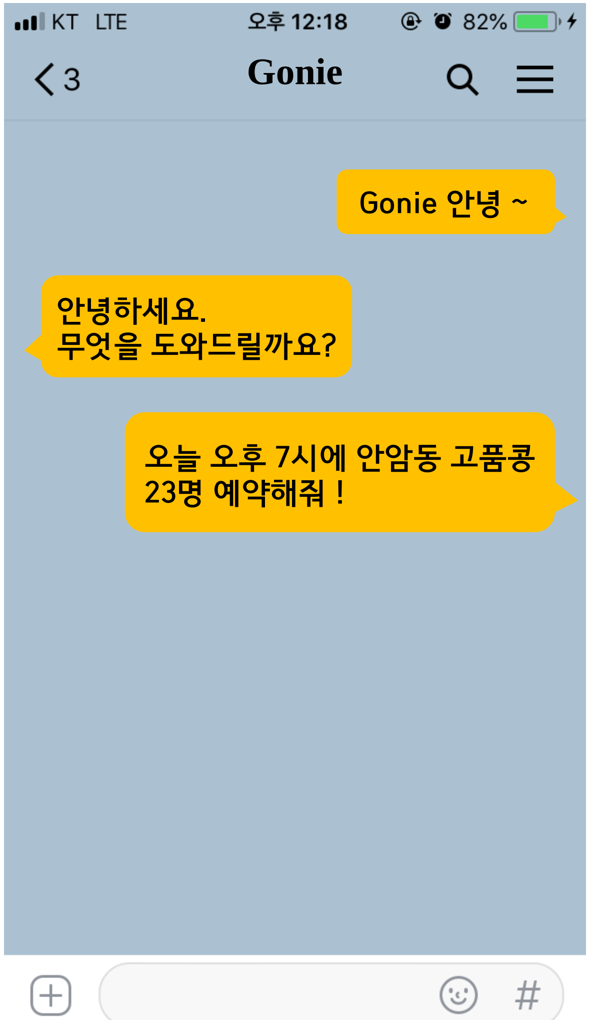
  Action: 안녕하세요. 무엇을 도와드릴까요?

- '오늘 오후 7시에 안암동 고품콩 23명 예약해줘 !' → Text Classification → class_예약

  * 시간:
  * 장소:
  * 음식점:
  * 사람 수:

Text Classification VS Named Entity Recognition

Gonie ChatBot



ChatBot Area

- 'Gonie 안녕 ~ ' → Text Classification → class_인사

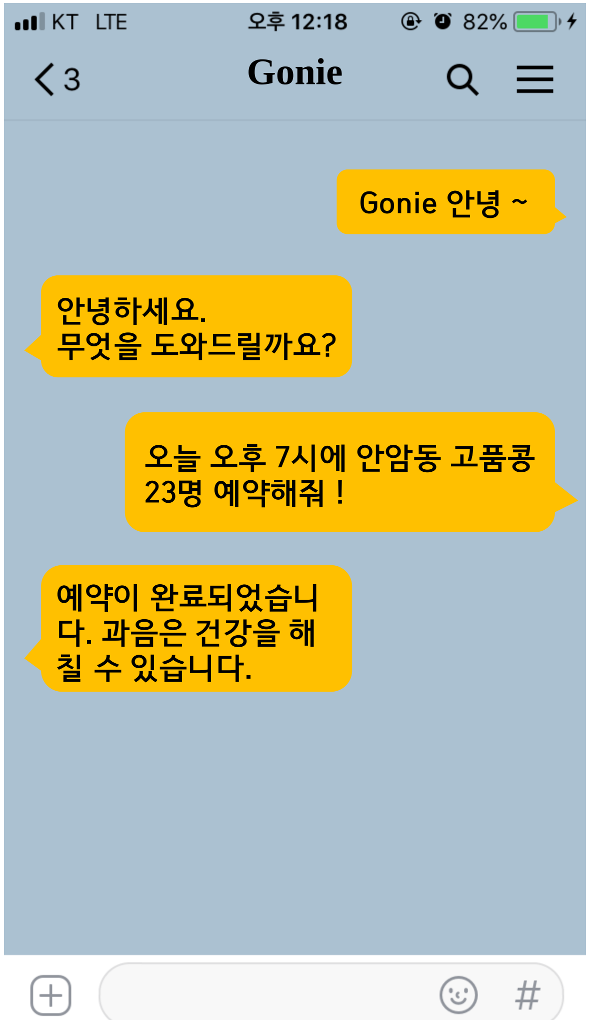
  Action: 안녕하세요. 무엇을 도와드릴까요?

- '오늘 오후 7시에 안암동 고품콩 23명 예약해줘 !' → Text Classification → class_예약 → NER 추출

  * 시간: '오늘 오후 7시'
  * 장소: '안암동'
  * 음식점: '고품콩'
  * 사람 수: '23명'

Text Classification VS Named Entity Recognition

Gonie ChatBot



ChatBot Area

- 'Gonie 안녕 ~ ' → Text Classification → class_인사

  Action: 안녕하세요. 무엇을 도와드릴까요?

- '오늘 오후 7시에 안암동 고품콩 23명 예약해줘 !' → Text Classification → class_예약 → NER 추출

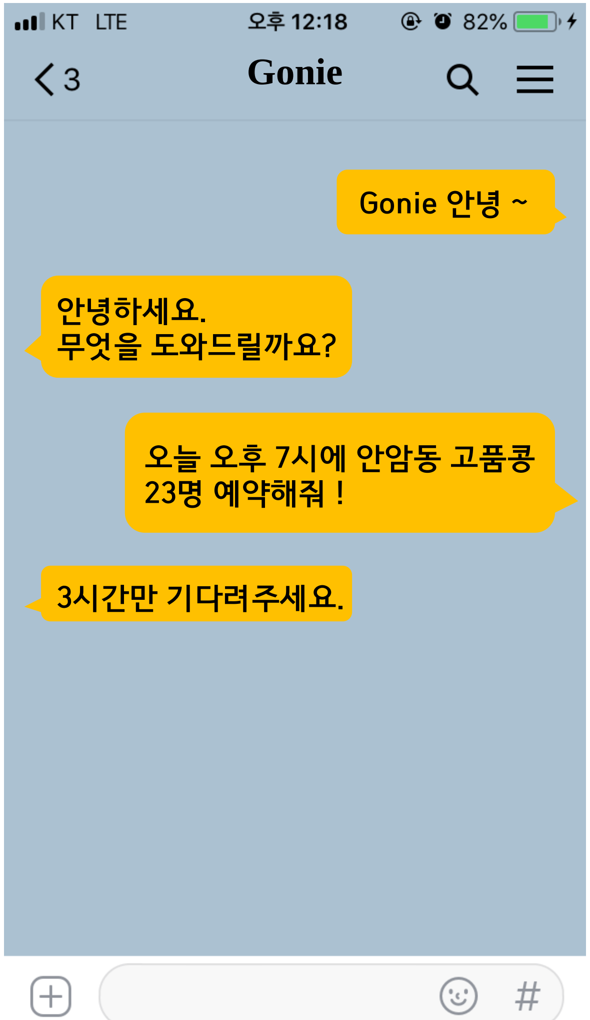
  * 시간: '오늘 오후 7시'
  * 장소: '안암동'
  * 음식점: '고품콩'
  * 사람 수: '23명'

  API..

  Action: (예약 완료한 후)
  예약이 완료되었습니다.
  과음은 건강을 해칠 수 있습니다.

Text Classification VS Named Entity Recognition

Gonie ChatBot



ChatBot Area

- **'Gonie 안녕 ~ ' → Text Classification → class_인사**

  Action: 안녕하세요. 무엇을 도와드릴까요?

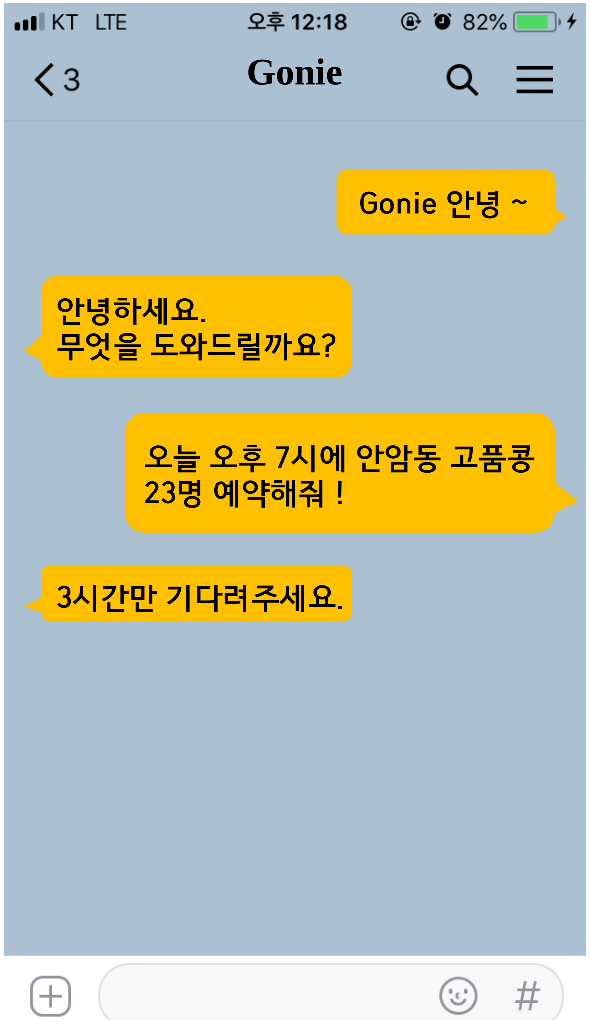
- **'오늘 오후 7시에 안암동 고품콩 23명 예약해줘 !' → Text Classification → class_예약**

시간		장소		음식점		사람 수
00:00 ~ 24:00	**X**	가락동 송파동 ~ 종암동 안암동	**X**	고품콩 쿠마 ~ 한추 희야	**X**	1 명 ~ 100 명
1,440		**2100**		**60만**		**100**

∞

Text Classification VS Named Entity Recognition

Gonie ChatBot



ChatBot Area

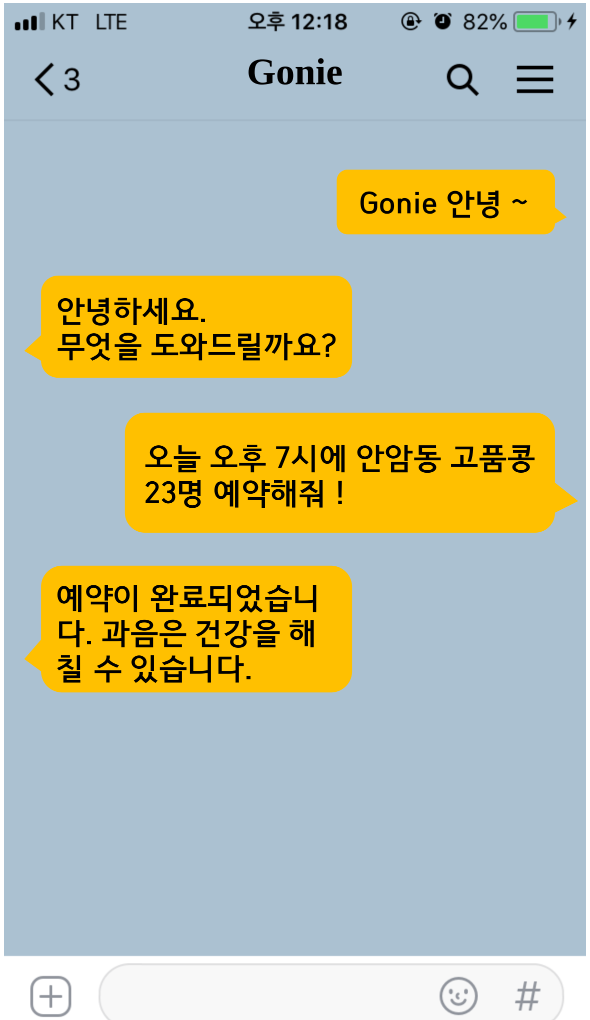
- 'Gonie 안녕 ~ ' → Text Classification → class_인사

  Action: 안녕하세요. 무엇을 도와드릴까요?

- '오늘 오후 7시에 안암동 고품콩 23명 예약해줘 !' → Text Classification → class_예약

Text Classification VS Named Entity Recognition

Gonie ChatBot



ChatBot Area

- 'Gonie 안녕 ~ ' → Text Classification → class_인사

  Action: 안녕하세요. 무엇을 도와드릴까요?

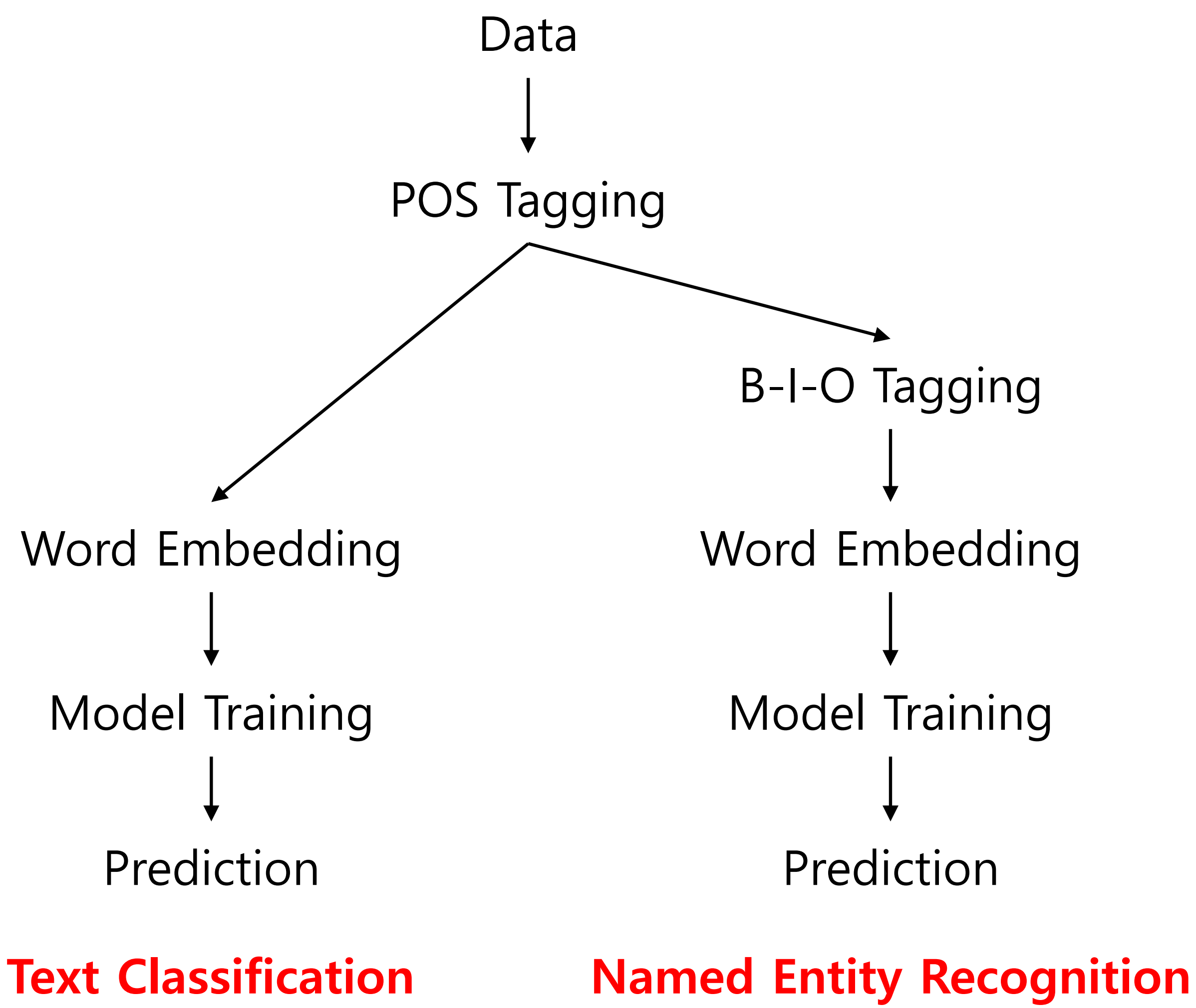
- '오늘 오후 7시에 안암동 고품콩 23명 예약해줘 !' → Text Classification → class_예약 → NER 추출

  * 시간: '오늘 오후 7시'
  * 장소: '안암동'
  * 음식점: '고품콩'
  * 사람 수: '23명'

  API..

  Action: (예약 완료한 후)
  예약이 완료되었습니다.
  과음은 건강을 해칠 수 있습니다.

Process


Data
POS Tagging
B-I-O Tagging
Word Embedding
Word Embedding
Model Training
Model Training
Prediction
Prediction
Text Classification
Named Entity Recognition

POS Tagging

Korean POS Tagger

Sejong project

Sim Gwangsub project

Twitter Korean Text

Komoran

Mecab-ko

Kkma

Hannanum

오늘 오후 7시에 안암동 고품콩 23명 예약해줘!

POS Tagging

('오늘', 'Noun')
('오후', 'Noun')
('7시', 'Number')
('에', 'Josa')
('안암동', 'Noun')
('고품콩', 'Noun')
('23명', 'Number')
('예약', 'Noun')
('해줘', 'Verb')
('!', 'Punctuation')

POS Tagging

Korean POS Tagger

Sejong project

Sim Gwangsub project

Twitter Korean Text

Komoran

Mecab-ko

Kkma

Hannanum

Text Classification

오늘 오후 7시에 안암동 고품콩 23명 예약해줘!

POS Tagging

제거

('오늘', 'Noun')
('오후', 'Noun')
('7시', 'Number')
('에', 'Josa')
('안암동', 'Noun')
('고품콩', 'Noun')
('23명', 'Number')
('예약', 'Noun')
('해줘', 'Verb')
('!', 'Punctuation')

POS Tagging

Korean POS Tagger

Sejong project

Sim Gwangsub project

Twitter Korean Text

Komoran

Mecab-ko

Kkma

Hannanum

Named Entity Recognition

오늘 오후 7시에 안암동 고품콩 23명 예약해줘!

POS Tagging

제거 X

('오늘', 'Noun')
('오후', 'Noun')
('7시', 'Number')
('에', 'Josa')
('안암동', 'Noun')
('고품콩', 'Noun')
('23명', 'Number')
('예약', 'Noun')
('해줘', 'Verb')
('!', 'Punctuation')

B-I-O Tagging

❖ 개체명은 대체로 하나 이상의 단어가 결합되어 구성됨

오늘 오후 7시에 안암동 고품콩 23명 예약해줘!

POS Tagging

('오늘', 'Noun')
('오후', 'Noun')
('7시', 'Number')
('에', 'Josa')
('안암동', 'Noun')
('고품콩', 'Noun')
('23명', 'Number')
('예약', 'Noun')
('해줘', 'Verb')
('!', 'Punctuation')

] * 시간
] * 장소
] * 음식점
] * 사람 수

B-I-O Tagging

❖ 개체명은 대체로 하나 이상의 단어가 결합되어 구성됨.

❖ B-I-O Tagging으로 하나 이상의 단어를 하나로 인식하게 함

- B (Begin): 개체명이 시작되는 단어에 "B_~"를 Tagging 함
- I (Inside): 개체명의 시작이 아니지만 "B_~"에 속하는 일부 단어일 때 " I " Tagging을 함
- O (Outside): 개체명 외 " O " Tagging 함

오늘 오후 7시에 안암동 고품콩 23명 예약해줘!

*** 시간: B_TIME**
*** 장소: B_LOC**
*** 음식점: B_RES**
*** 사람 수: B_NUM**

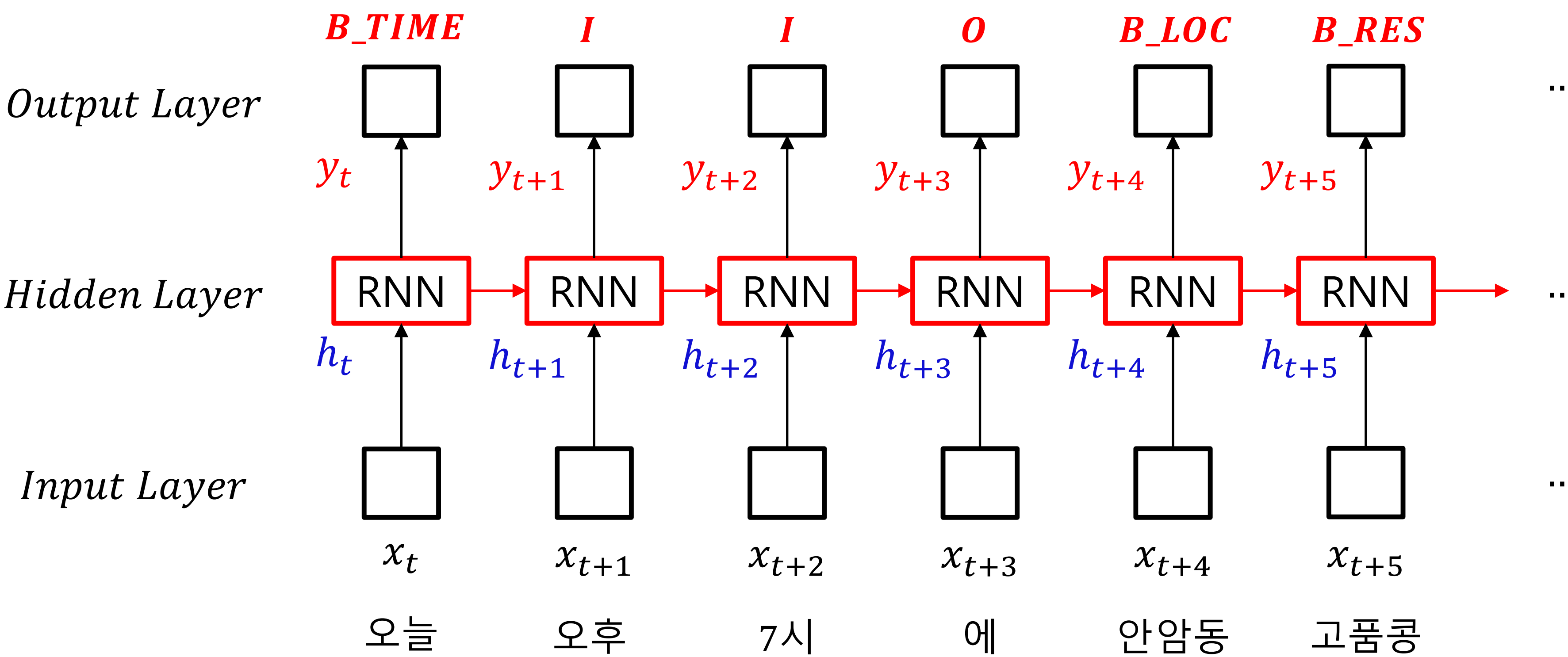
POS Tagging

('오늘', 'Noun')
('오후', 'Noun')
('7시', 'Number')
('에', 'Josa')
('안암동', 'Noun')
('고품콩', 'Noun')
('23명', 'Number')
('예약', 'Noun')
('해줘', 'Verb')
('!', 'Punctuation')

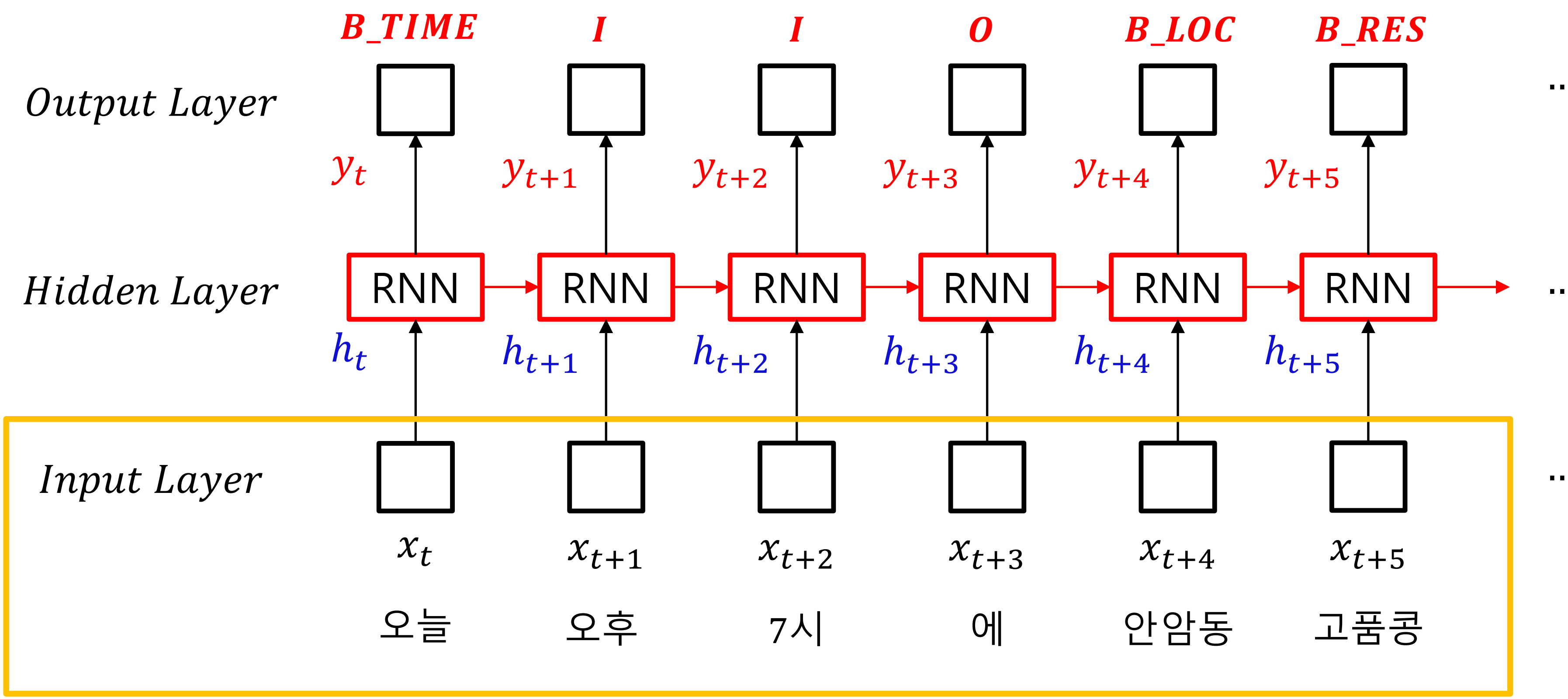
POS Tagging + B-I-O Tagging

('오늘', 'Noun', 'B_TIME')
('오후', 'Noun', 'I')
('7시', 'Number', 'I')
('에', 'Josa', 'O')
('안암동', 'Noun', 'B_LOC')
('고품콩', 'Noun', 'B_RES')
('23명', 'Number', 'B_NUM')
('예약', 'Noun', 'O')
('해줘', 'Verb', 'O')
('!', 'Punctuation', 'O')

Named Entity Recognition



Named Entity Recognition



Word Embedding

❖ **One-Hot Encoding based word representation**

: TF-IDF, Random, Senna …

❖ **Model based word representation**

: NNLM, Word2Vec, GloVe, FastText, …

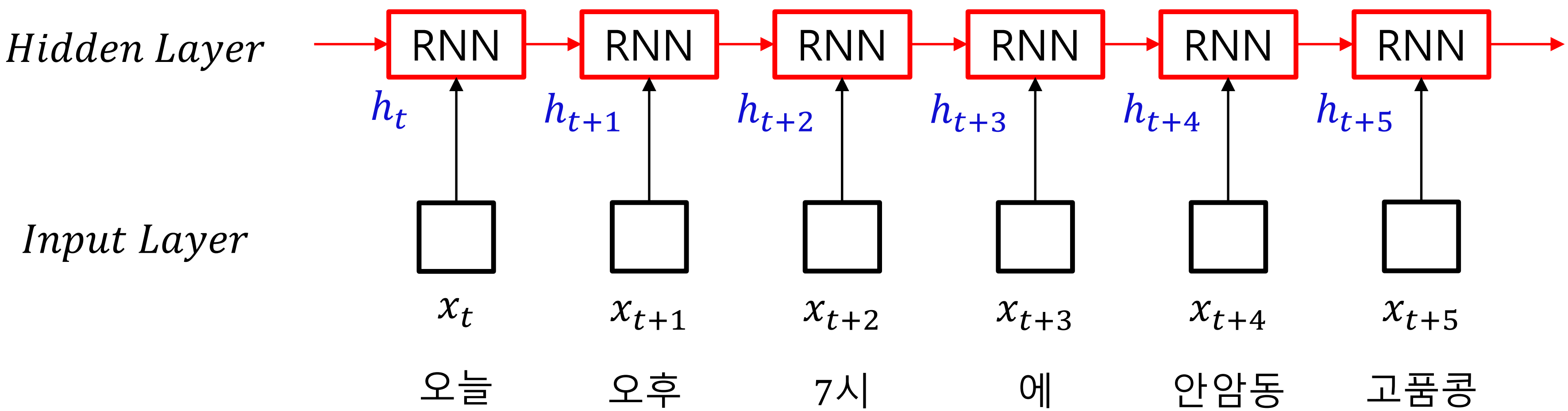
Word Embedding

- **One-Hot Encoding based word representation**

  : TF-IDF, Random, Senna …

- **Model based word representation**

  : NNLM, Word2Vec, GloVe, FastText, …



X1	X2	…	X29	X30
1.225	2.11	…	0.812	-1.21

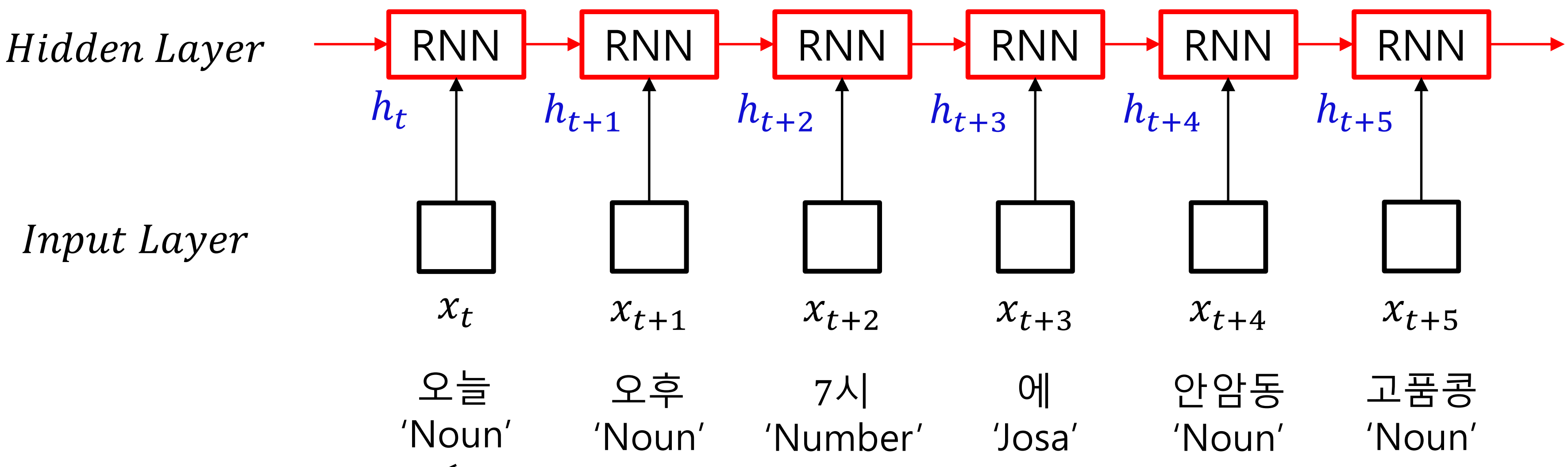
Word Embedding

❖ **One-Hot Encoding based word representation**

: TF-IDF, Random, Senna …

❖ **Model based word representation**

: NNLM, Word2Vec, GloVe, FastText, …



X1	X2	…	X29	X30
1.225	2.11	…	0.812	-1.21

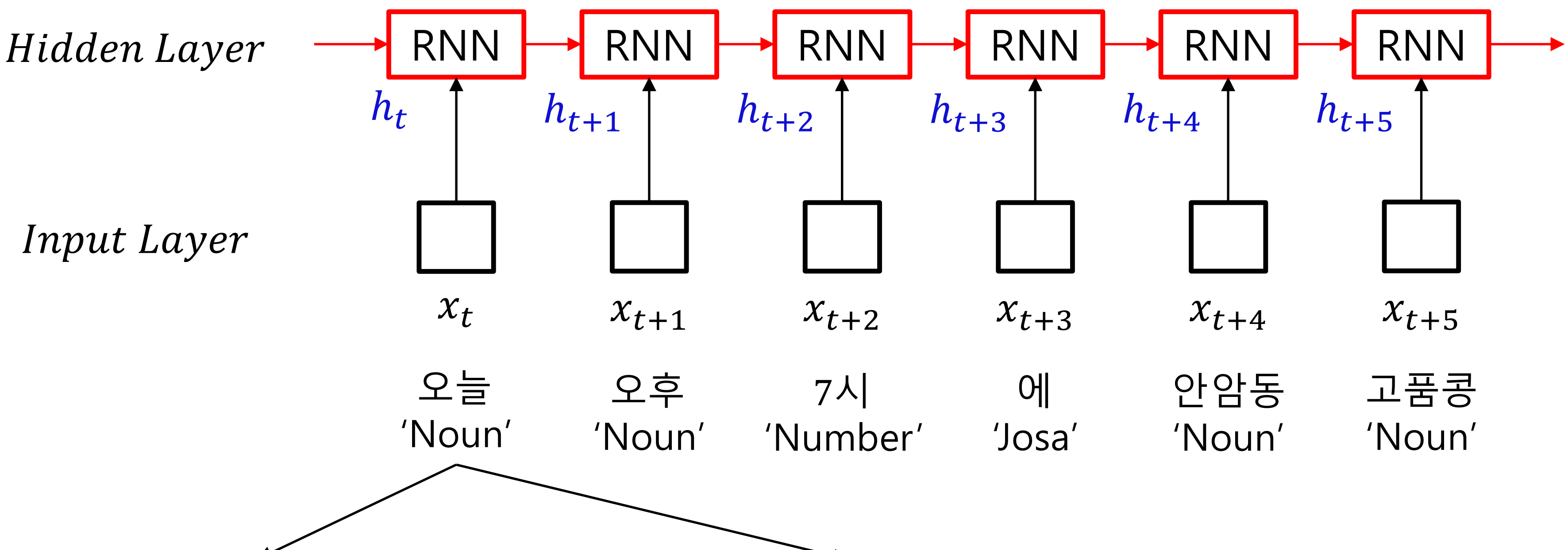
Word Embedding

❖ **One-Hot Encoding based word representation**

: TF-IDF, Random, Senna …

❖ **Model based word representation**

: NNLM, Word2Vec, GloVe, FastText, …



X1	X2	…	X29	X30
1.225	2.11	…	0.812	-1.21

+

Noun	Alpha	…	Adjective	Verb
1	0	…	0	0

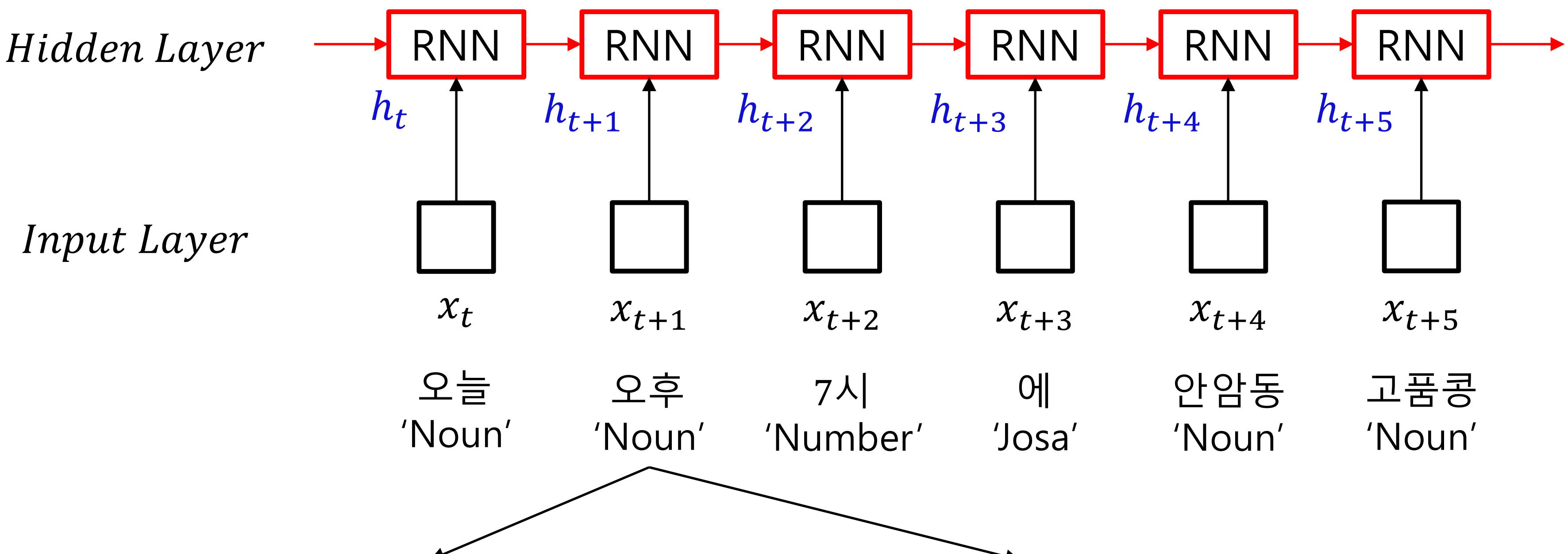
Word Embedding

- **One-Hot Encoding based word representation**

  : TF-IDF, Random, Senna …

- **Model based word representation**

  : NNLM, Word2Vec, GloVe, FastText, …

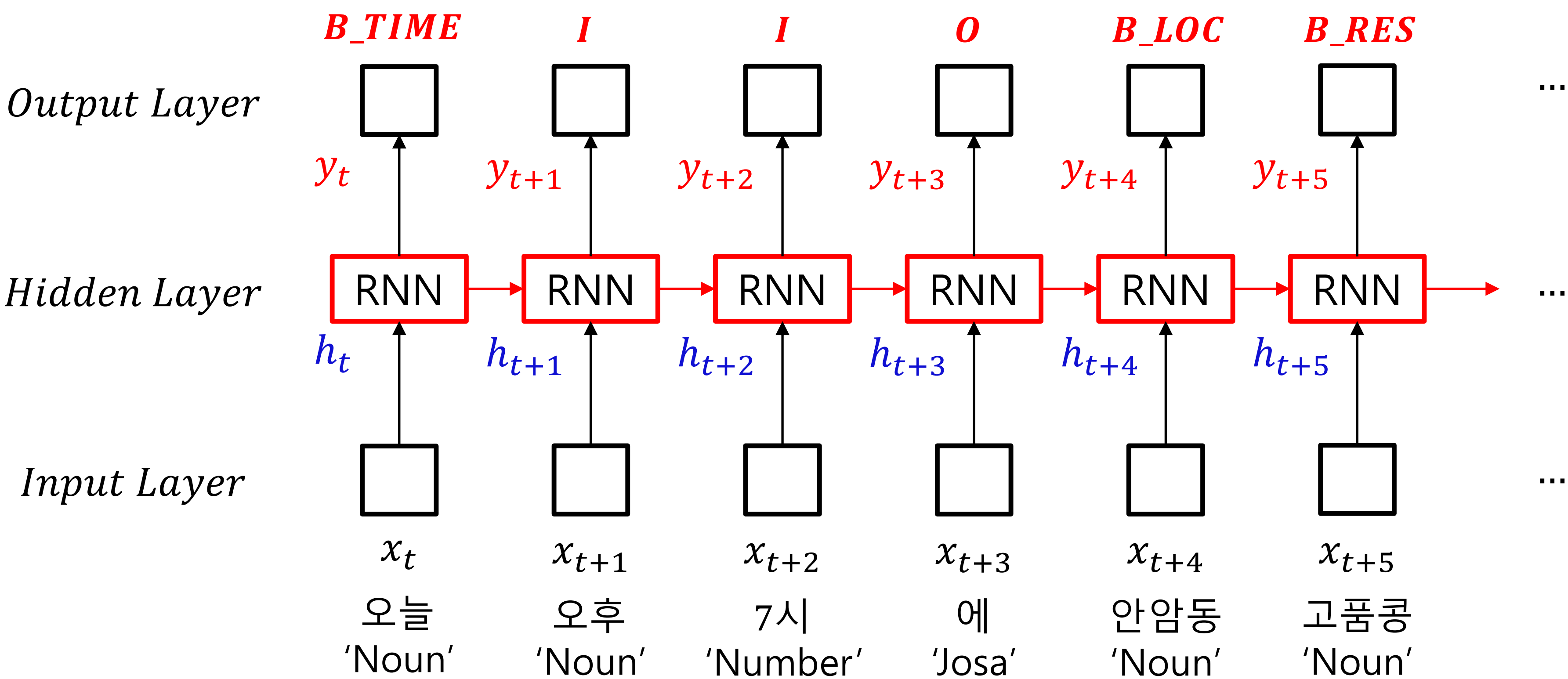


X1	X2	…	X29	X30
0.845	3.88	…	-1.141	0.99

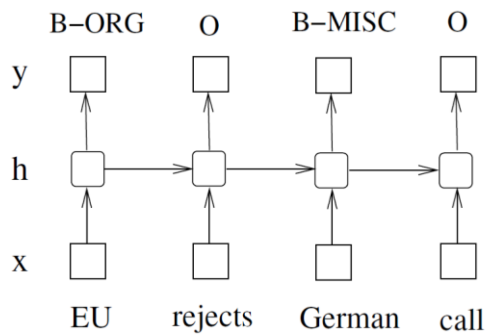
\+

Noun	Alpha	…	Adjective	Verb
1	0	…	0	0

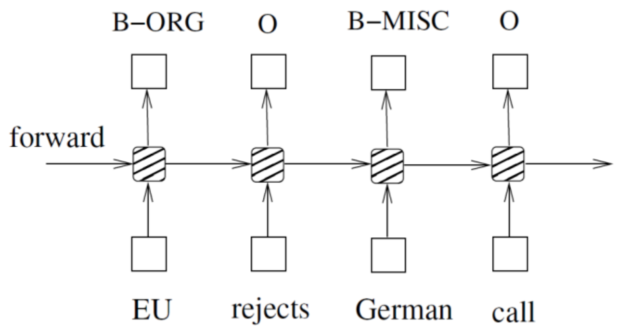
Named Entity Recognition



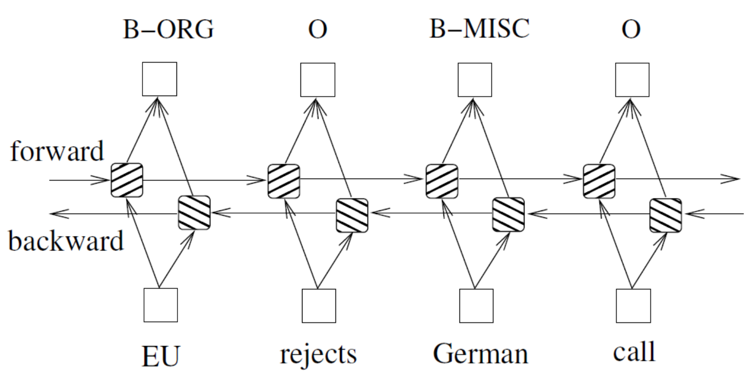
Models



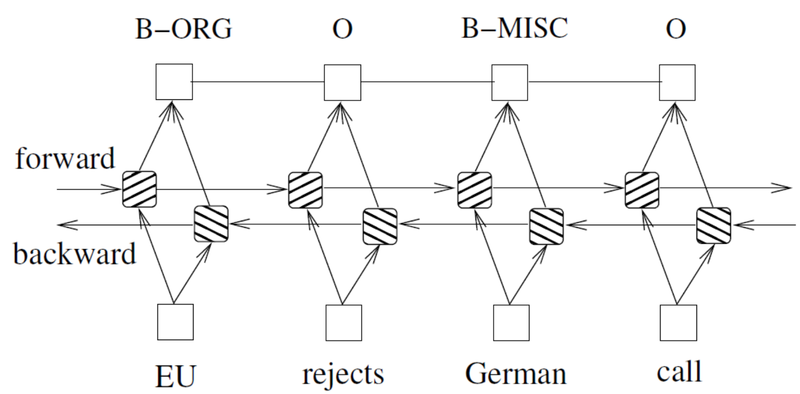
1. RNN Model



2. LSTM Model



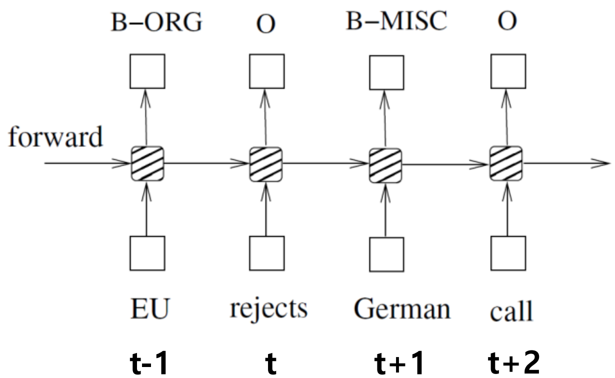
3. Bi-LSTM Model



4. Bi-LSTM-CRF Model

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

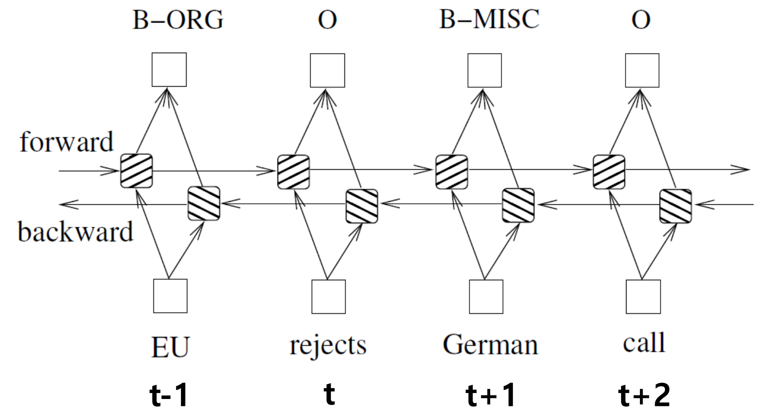
Bidirectional LSTM



2. LSTM Model

$$\overrightarrow{h_t} = \sigma\left(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right) \quad (1)$$

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + b_y \quad (2)$$



3. Bi-LSTM Model

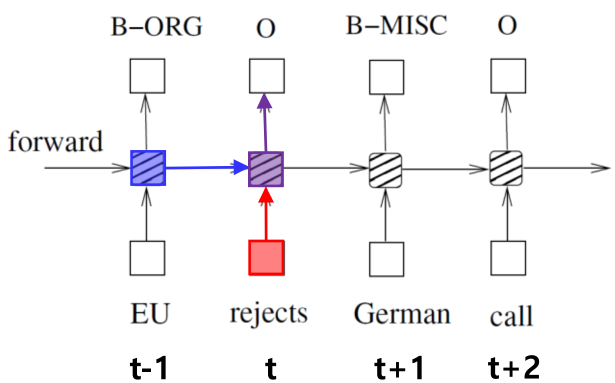
$$\overrightarrow{h_t} = \sigma\left(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right) \quad (1)$$

$$\overleftarrow{h_t} = \sigma\left(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}\right) \quad (2)$$

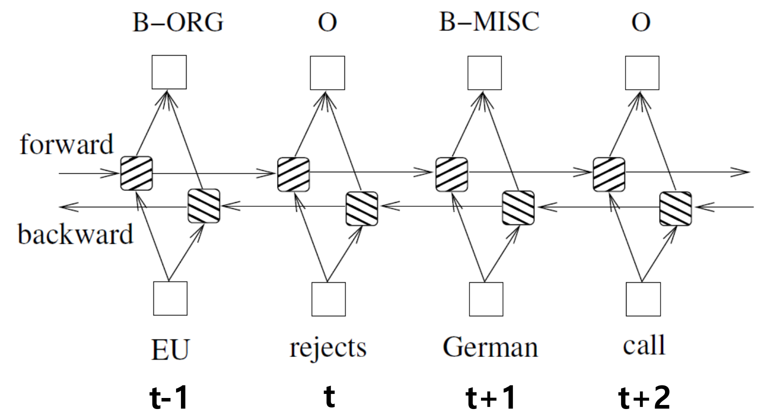
$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h_t} + b_y \quad (3)$$

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Bidirectional LSTM



2. LSTM Model



3. Bi-LSTM Model

$$\overrightarrow{h}_t = \sigma\left(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right) \quad (1)$$

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + b_y \quad (2)$$

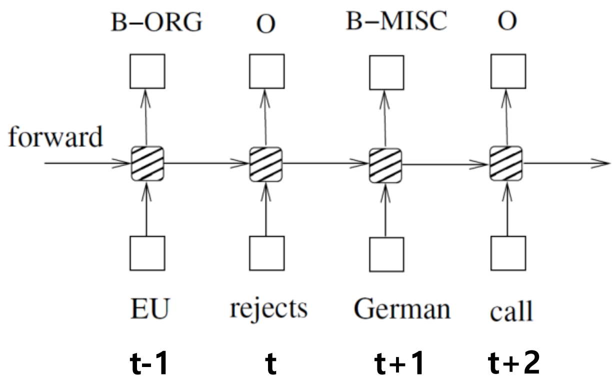
$$\overrightarrow{h}_t = \sigma\left(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right) \quad (1)$$

$$\overleftarrow{h}_t = \sigma\left(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}\right) \quad (2)$$

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \quad (3)$$

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

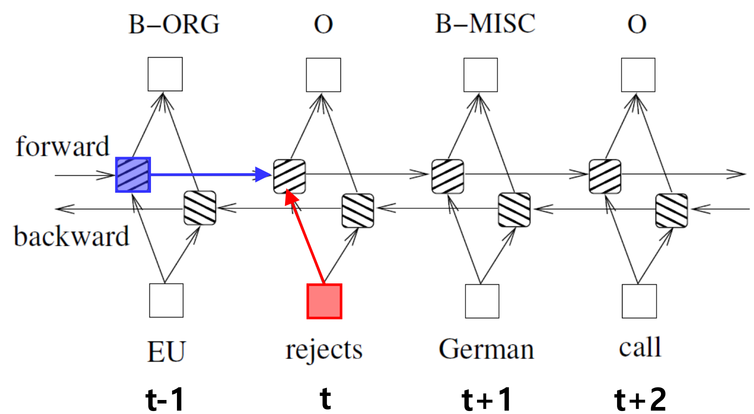
Bidirectional LSTM



2. LSTM Model

$$\overrightarrow{h_t} = \sigma\left(W_{x\vec{h}} x_t + W_{\vec{h}\vec{h}} \overrightarrow{h}_{t-1} + b_{\vec{h}}\right) \quad (1)$$

$$y_t = W_{\vec{h}y} \overrightarrow{h}_t + b_y \quad (2)$$



3. Bi-LSTM Model

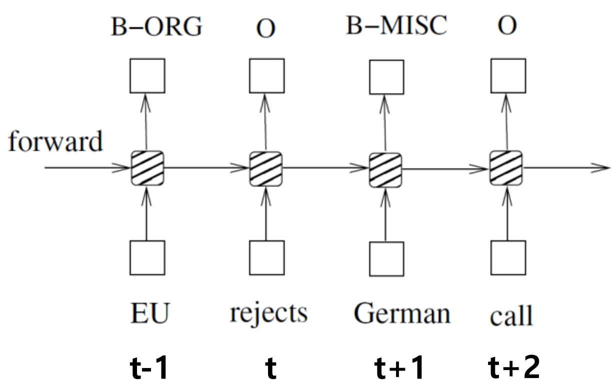
$$\overrightarrow{h_t} = \sigma\left(W_{x\vec{h}} x_t + W_{\vec{h}\vec{h}} \overrightarrow{h}_{t-1} + b_{\vec{h}}\right) \quad (1)$$

$$\overleftarrow{h_t} = \sigma\left(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}\right) \quad (2)$$

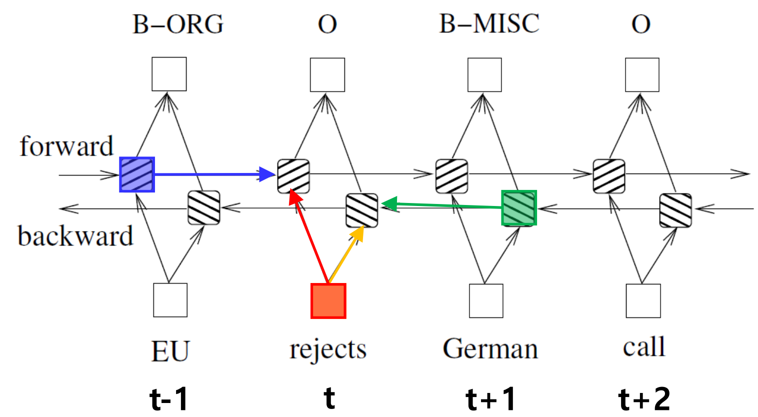
$$y_t = W_{\vec{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h_t} + b_y \quad (3)$$

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Bidirectional LSTM



2. LSTM Model



3. Bi-LSTM Model

$$\overrightarrow{h_t} = \sigma\left(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right) \quad (1)$$

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + b_y \quad (2)$$

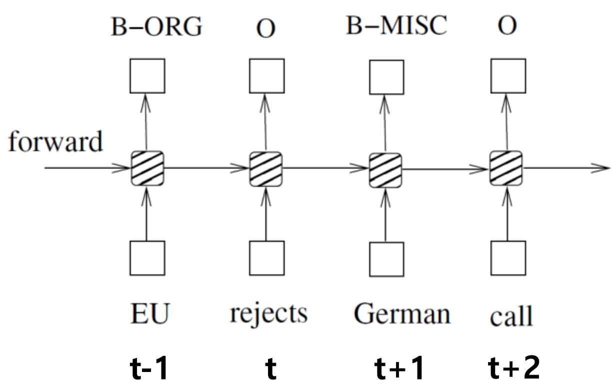
$$\overrightarrow{h_t} = \sigma\left(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right) \quad (1)$$

$$\overleftarrow{h_t} = \sigma\left(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}\right) \quad (2)$$

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h_t} + b_y \quad (3)$$

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

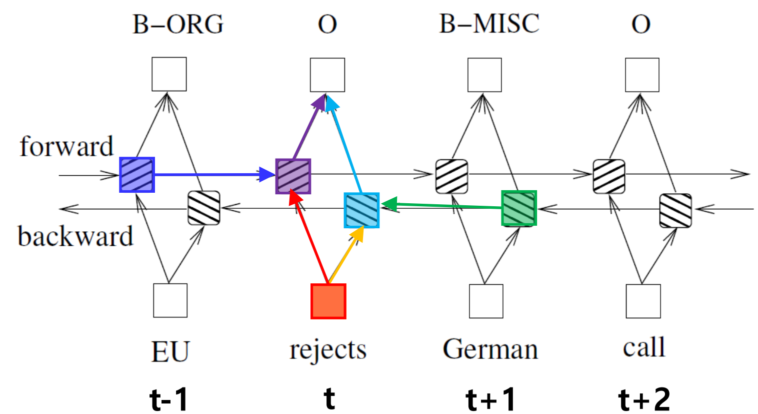
Bidirectional LSTM



2. LSTM Model

$$\overrightarrow{h}_t = \sigma\left(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right) \qquad (1)$$

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + b_y \qquad (2)$$



3. Bi-LSTM Model

$$\overrightarrow{h}_t = \sigma\left(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right) \qquad (1)$$

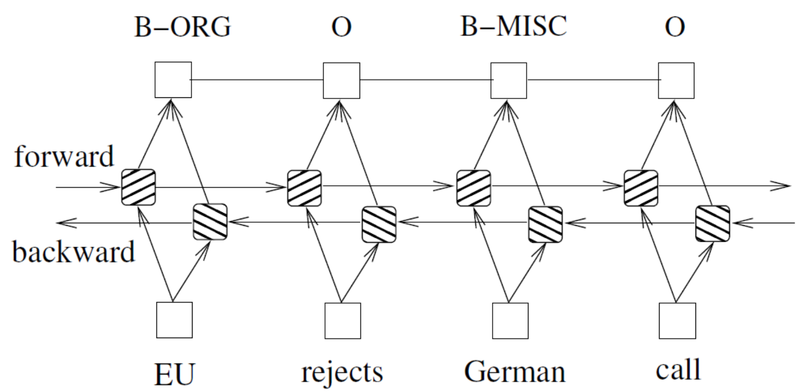
$$\overleftarrow{h}_t = \sigma\left(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}\right) \qquad (2)$$

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \qquad (3)$$

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Bidirectional LSTM-CRF

❖ **Bidirectional Long and Short Term Memory Conditional Random Field (Bi-LSTM-CRF)**



4. Bi-LSTM-CRF Model

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Sequence Prediction via Conditional Random Field (CRF)

- x = (새, 가, 운다)
- Y = (B_An, O, O)

$P(B_{An}, O, O \mid$새, 가, 운다$)$

- x = (삼성,전자, 는, 좋다)
- Y = (B_Com, I, O, O)

$P(B_{Com}, I, O, O \mid$ 삼성,전자, 는, 좋다$)$

- x = (고양이, 는, 귀엽다)
- Y = (B_An, O, O)

$P(B_{An}, O, O \mid$고양이, 는, 귀엽다$)$

- x = (오후, 7시, 에, 시작한다)
- Y = (B_Time, I, O, O)

$P(B_{Time}, I, O, O \mid$오후, 7시, 에, 시작한다$)$

⋮

- x = (개, 좀, 안, 짖게, 해라)
- Y = (B_An, O, O, O, O)

$P(B_{An}, O, O, O, O \mid$개, 좀, 안, 짖게, 해라$)$

Sequence Prediction via Conditional Random Field (CRF)

- x = (새, 가, 운다)
- Y = (B

$P(B$ … O, O |새, 가, 운다)

X and Y Dependent – CRF

Y and Y Dependent – CRF

- x = (신
- Y = (B

…는다)

- x = (고
- Y = (B_An, O, O)

- x = (오후, 7시, 에, 시작한다)
- Y = (B_Time, I, O, O)

$P(B_{Time}, I, O, O$ |오후, 7시, 에, 시작한다)

⋮

- x = (개, 좀, 안, 짖게, 해라)
- Y = (B_An, O, O, O, O)

$P(B_{An}, O, O, O, O$ |개, 좀, 안, 짖게, 해라)

Sequence Prediction via Conditional Random Field (CRF)

- x = (새, 가, 운다)
- Y = (B

- x = (
- Y = (B

다)

- x = (
- Y = (B_An, O, O)

- x = (오후, 2시, 에, 시작한다)
- Y = (B

작한다)

X and Y Dependent – CRF

Y and Y Dependent – CRF

X and X Dependent – Bi-LSTM

X and Y Dependent – Bi-LSTM

- x = (
- Y = (B_An, O, O, O, O)

, 해라)

Sequence Prediction via Conditional Random Field (CRF)

- x = (새, 가, 운다)
- Y = (B_An, O, O)

$P(B_{An}, O, O\,|$새, 가, 운다$)$

- x = (삼성,전자, 는, 좋다)
- Y = (B_Com, I, O, O)

$P(B_{Com}, I, O, O\,|$ 삼성,전자, 는, 좋다$)$

- x = (고
- Y = (B_A

- x = (오
- Y = (B_

한다)

X and X Dependent – Bi-LSTM

X and Y Dependent – Bi-LSTM, CRF

Y and Y Dependent – CRF

⋮

- x = (개, 좀, 안, 짖게, 해라)
- Y = (B_An, O, O, O, O)

$P(B_{An}, O, O, O, O\,|$개, 좀, 안, 짖게, 해라$)$

Sequence Prediction via Conditional Random Field (CRF)

- x = (새, 가, 운다)
- Y = (B_An, O, O)

$P(B_{An}, O, O \mid$새, 가, 운다$)$

- x = (삼성,전자, 는, 좋다)
- Y = (B_Com, I, O, O)

$P(B_{Com}, I, O, O \mid$ 삼성,전자, 는, 좋다$)$

- x = (고양이, 는, 귀엽다)
- Y = (B_An, O, O)

$P(B_{An}, O, O \mid$고양이, 는, 귀엽다$)$

- x = (오후, 7시, 에, 시작한다)
- Y = (B_Time, I, O, O)

$P(B_{Time}, I, O, O \mid$오후, 7시, 에, 시작한다$)$

⋮

- x = (개, 좀, 안, 짖게, 해라)
- Y = (B_An, O, O, O, O)

$P(B_{An}, O, O, O, O \mid$개, 좀, 안, 짖게, 해라$)$

Training

Sequence Prediction via Conditional Random Field (CRF)

$P(?\,,?\,,?\,,?$ | 오후, 10시, 에, 축구한다)

↓

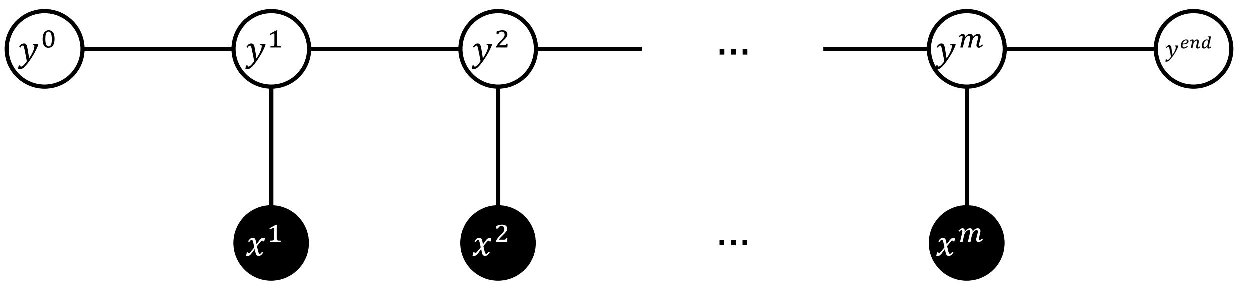
Trained CRF

↓

$P(B_{Time}, I, O, O$ | 오후, 10시, 에, 축구한다)

Sequence Prediction via Conditional Random Field (CRF)

- Input: x = $(x^1, \ldots, x^m)$
- Predict: y = $(y^1, \ldots, y^m)$
  - Each y^i one of L labels



- P(x^i | y^i)
- P(y^i | y^{i-1})
- P(y^1 | y^0)
- P(End | y^m) – Not always used

Sequence Prediction via Conditional Random Field (CRF)

$$P(y^1, \dots, y^m | x^1, \dots, x^m) = \frac{1}{Z(x)} \exp\left\{\sum_{j=1}^{m} \left(P\left(x^j \middle| y^j\right) + P\left(y^j \middle| y^{j-1}\right)\right)\right\}$$

Sequence Prediction via Conditional Random Field (CRF)

$$P(y^1, \dots, y^m | x^1, \dots, x^m) = \frac{1}{Z(x)} \exp\left\{\sum_{j=1}^{m} \left(P\left(x^j \middle| y^j\right) + P\left(y^j \middle| y^{j-1}\right) \right)\right\}$$

$$P(y^1, y^2, y^3 | x^1, x^2, x^3)$$

Sequence Prediction via Conditional Random Field (CRF)

$$P(y^1, \dots, y^m | x^1, \dots, x^m) = \frac{1}{Z(x)} \exp\left\{\sum_{j=1}^{m} \left(P\left(x^j \middle| y^j\right) + P\left(y^j \middle| y^{j-1}\right)\right)\right\}$$

$$P(y^1, y^2, y^3 | x^1, x^2, x^3) = \frac{1}{Z(x)} \exp\left\{\begin{matrix} P(x^1|y^1) + P(y^1|y^0) + \\ P(x^2|y^2) + P(y^2|y^1) + \\ P(x^3|y^3) + P(y^3|y^2) \end{matrix}\right\}$$

Sequence Prediction via Conditional Random Field (CRF)

$$P(y^1, \dots, y^m | x^1, \dots, x^m) = \frac{1}{Z(x)} \exp\left\{\sum_{j=1}^{m} \left(P\left(x^j \middle| y^j\right) + P\left(y^j \middle| y^{j-1}\right)\right)\right\}$$

$$P(y|x) = \frac{1}{Z(x)} \exp\left\{\sum_{j=1}^{m} \left(A_{y^j, y^{j-1}} + O_{y^j, x^j}\right)\right\}$$

$F(y, x) = \sum_{j=1}^{m} (A_{y^j, y^{j-1}} + O_{y^j, x^j})$ → Scoring function

Scoring transitions **Scoring input features**

$Z(x) = \sum_{y' \in Y} \exp\{F(y', \mathrm{x})\}$ → aka "Partition Function"

All possible label sequence

Sequence Prediction via Conditional Random Field (CRF)

- x = "Gonie Sleep"
- y = (B, O)

$$P(y|x) = \frac{1}{Z(x)} \exp\left\{\sum_{j=1}^{m} \left(A_{y^j, y^{j-1}} + O_{y^j, x^j}\right)\right\}$$

Scoring transitions

$A_{O,B}$

	$A_{B,*}$	$A_{O,*}$
$A_{*,B}$	-2	2
$A_{*,O}$	1	-2
$A_{*,start}$	1	-1

Scoring input features

$O_{B,Gonie}$

	$O_{B,*}$	$O_{O,*}$
$O_{*,Gonie}$	2	1
$O_{*,Sleep}$	1	0

$$P(B,O \mid \text{"Gonie Sleep"}) = \frac{1}{Z(x)} \exp\{\underbrace{A_{B,start}}_{1} + \underbrace{A_{O,B}}_{2} + \underbrace{O_{B,Gonie}}_{2} + \underbrace{O_{O,Sleep}}_{0}\} = \frac{1}{Z(x)} \exp\{5\} = 0.93$$

$$\begin{aligned} Z(x) = {} & P(B,B \mid \text{"Gonie Sleep"}) + \\ & P(B,O \mid \text{"Gonie Sleep"}) + \\ & P(O,B \mid \text{"Gonie Sleep"}) + \\ & P(O,O \mid \text{"Gonie Sleep"}) \end{aligned} \Rightarrow \begin{aligned} Z(x) = {} & \exp(1-2+2+0) + \\ & \exp(1+2+2+0) + \\ & \exp(-1+1+1+1) + \\ & \exp(-1-2+1+0) \end{aligned} \Rightarrow Z(x) = 158.65$$

Conditional Random Field Parameter Estimation

$$P(y|x) = \frac{1}{Z(x)} \exp\left\{\sum_{j=1}^{m}(A_{y^j,y^{j-1}} + O_{y^j,x^j})\right\}$$

$F(y,x) = \sum_{j=1}^{m}(A_{y^j,y^{j-1}} + O_{y^j,x^j})$ → Scoring function

Scoring transitions **Scoring input features**

$Z(x) = \sum_{y' \in Y} \exp\{F(y',\mathrm{x})\}$ → aka “Partition Function”

All possible label sequence

$$P(y|x) = \frac{\exp(F(y,x))}{\sum_{y' \in Y} \exp\{F(y',\mathrm{x})\}}$$

Conditional Random Field Parameter Estimation

$$P(y|x) = \frac{\exp(F(y,x))}{\sum_{y' \in Y} \exp\{F(y',\mathrm{x})\}}$$

$$P(y|x\,;w) = \frac{\exp\{\sum_j \sum_i w_i f_i\left(y_{j-1}, y_j, x, i\right)\}}{\sum_{y' \in Y} \exp\{\sum_j \sum_i w_i f_i(y'_{j-1}, y'_j, x, i)\}}$$

Conditional Random Field Parameter Estimation

$$P(y|x) = \frac{\exp(F(y,x))}{\sum_{y' \in Y} \exp\{F(y',\mathrm{x})\}}$$

Sum over all scoring function

Weight for given scoring function

scoring function without weight

Length of Sequence x

$$P(y|x\,;w) = \frac{\exp\{\sum_j \sum_i w_i f_i\big(y_{j-1}, y_j, x, i\big)\}}{\sum_{y' \in Y} \exp\{\sum_j \sum_i w_i f_i(y'_{j-1}, y'_j, x, i)\}}$$

Sum over all possible label sequence

Conditional Random Field Parameter Estimation

$$P(y|x\,;w) = \frac{\exp\{\sum_j \sum_i w_i f_i\big(y_{j-1}, y_j, x, i\big)\}}{\sum_{y' \in Y} \exp\{\sum_j \sum_i w_i f_i(y'_{j-1}, y'_j, x, i)\}}$$

Conditional Random Field Parameter Estimation

$$P(y|x\,;w) = \frac{\exp\{\sum_j \sum_i w_i f_i\big(y_{j-1}, y_j, x, i\big)\}}{\sum_{y' \in Y} \exp\{\sum_j \sum_i w_i f_i(y'_{j-1}, y'_j, x, i)\}}$$

$$\log P(y|x\,;w) = \sum_j \sum_i w_i f_i\big(y_{j-1}, y_j, x, i\big) - \log[\sum_{y' \in Y} exp\{\sum_j \sum_i w_i f_i(y'_{j-1}, y'_j, x, i)\}]$$

Gradient ascent !!

Conditional Random Field Parameter Estimation

$$P(y|x\,;w) = \frac{\exp\{\sum_j \sum_i w_i f_i(y_{j-1}, y_j, x, i)\}}{\sum_{y' \in Y} \exp\{\sum_j \sum_i w_i f_i(y'_{j-1}, y'_j, x, i)\}}$$

$$\log P(y|x\,;w) = \sum_j \sum_i w_i f_i(y_{j-1}, y_j, x, i) - \log[\sum_{y' \in Y} exp\{\sum_j \sum_i w_i f_i(y'_{j-1}, y'_j, x, i)\}]$$

Gradient ascent !!

$$\frac{\partial}{\partial w_i} \log P(y|x\,;w) = \sum_j f_i(y_{j-1}, y_j, x, i) - \sum_{y' \in Y} P(y'|x) \sum_j f_i(y'_{j-1}, y'_j, x, i)$$

$$\lambda_i = \lambda_i + \alpha \cdot [\sum_j f_i(y_{j-1}, y_j, x, i) - \sum_{y' \in Y} P(y'|x) \sum_j f_i(y'_{j-1}, y'_j, x, i)]$$

Conditional Random Field Parameter Estimation

$$P(y|x\,;w) = \frac{\exp\{\sum_j \sum_i w_i f_i(y_{j-1}, y_j, x, i)\}}{\sum_{y' \in Y} \exp\{\sum_j \sum_i w_i f_i(y'_{j-1}, y'_j, x, i)\}}$$

$$\log P(y|x\,;w) = \sum_j \sum_i w_i f_i(y_{j-1}, y_j, x, i) - \log[\sum_{y' \in Y} exp\{\sum_j \sum_i w_i f_i(y'_{j-1}, y'_j, x, i)\}]$$

Gradient ascent !!

$$\frac{\partial}{\partial w_i} \log P(y|x\,;w) = \sum_j f_i(y_{j-1}, y_j, x, i) - \sum_{y' \in Y} P(y'|x) \sum_j f_i(y'_{j-1}, y'_j, x, i)$$

Learning rate

$$\lambda_i = \lambda_i + \alpha \cdot [\sum_j f_i(y_{j-1}, y_j, x, i) - \sum_{y' \in Y} P(y'|x) \sum_j f_i(y'_{j-1}, y'_j, x, i)]$$

Direction of the gradient

Parameter Estimation !!!

Conditional Random Field Prediction

$$\underset{y*}{arg\max}\, P(y|x)$$

$$Z(x) = 158.65$$

$$P(B,B|\,"Gonie\ Sleep") = \frac{1}{Z(x)} \exp\{A_{B,start} + A_{B,B} + O_{B,Gonie} + O_{B,Sleep}\} = \frac{1}{Z(x)} \exp\{1\} = 0.02$$

$$P(B,O\,|\,"Gonie\ Sleep") = \frac{1}{Z(x)} \exp\{A_{B,start} + A_{O,B} + O_{B,Gonie} + O_{O,Sleep}\} = \frac{1}{Z(x)} \exp\{5\} = 0.93$$

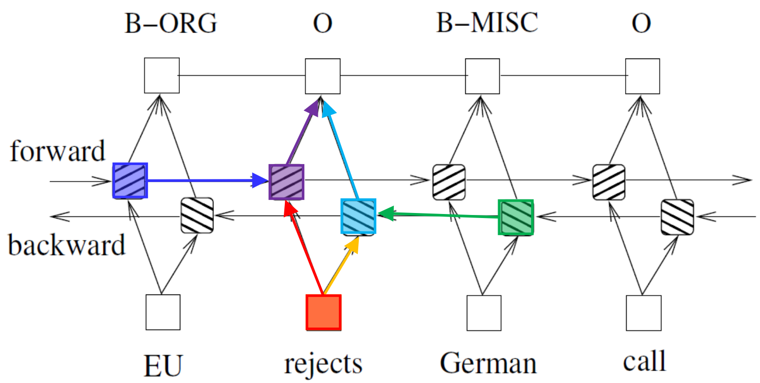
$$P(O,B\,|\,"Gonie\ Sleep") = \frac{1}{Z(x)} \exp\{A_{O,start} + A_{B,O} + O_{O,Gonie} + O_{B,Sleep}\} = \frac{1}{Z(x)} \exp\{2\} = 0.05$$

$$P(O,O\,|\,"Gonie\ Sleep") = \frac{1}{Z(x)} \exp\{A_{O,start} + A_{O,O} + O_{O,Gonie} + O_{O,Sleep}\} = \frac{1}{Z(x)} \exp\{-2\} = 0.00$$

$$\underset{y*}{arg\max}\, P(y|x) = (B,O)$$

Bidirectional LSTM-CRF

❖ **Bidirectional Long and Short Term Memory Conditional Random Field (Bi-LSTM-CRF)**



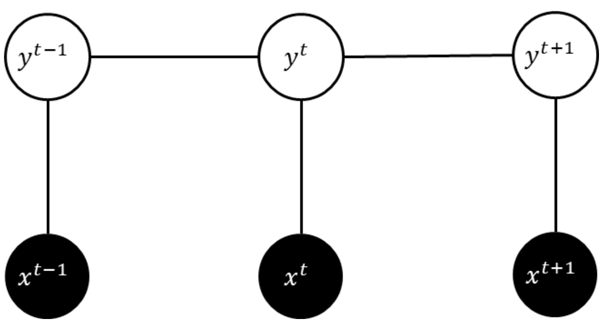
4. Bi-LSTM-CRF Model

$$\overrightarrow{h_t} = \sigma(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}) \quad (1)$$

$$\overleftarrow{h_t} = \sigma(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}) \quad (2)$$

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h_t} + b_y \quad (3)$$

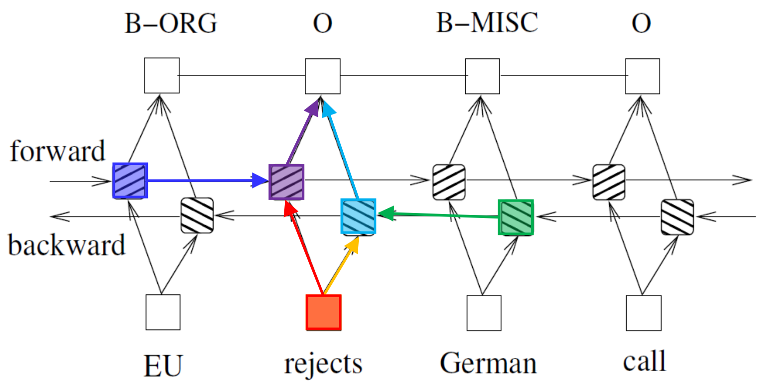
Conditional Random Field



Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Bidirectional LSTM-CRF

❖ Bidirectional Long and Short Term Memory **Conditional Random Field** (Bi-LSTM-CRF)



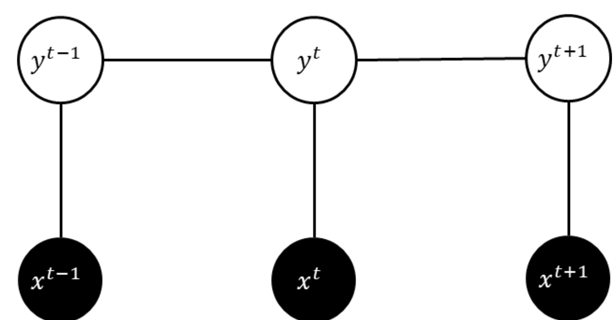
4. Bi-LSTM-CRF Model

Conditional Random Field

$$\overrightarrow{h}_t = \sigma\left(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right) \quad (1)$$

$$\overleftarrow{h}_t = \sigma\left(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}\right) \quad (2)$$

$$y_t = \sigma(W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y) \quad (3)$$



Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Experiment Result

❖ Data

Table 1: Size of sentences, tokens, and labels for training, validation and test sets.

		POS	CoNLL2000	CoNLL2003
training	sentence #	39831	8936	14987
	token #	950011	211727	204567
validation	sentence #	1699	N/A	3466
	token #	40068	N/A	51578
test	sentences #	2415	2012	3684
	token #	56671	47377	46666
	label #	45	22	9

❖ Result

Table 2: Comparison of tagging performance on POS, chunking and NER tasks for various models.

		POS	CoNLL2000	CoNLL2003
Random	Conv-CRF (Collobert et al., 2011)	96.37	90.33	81.47
	LSTM	97.10	92.88	79.82
	BI-LSTM	97.30	93.64	81.11
	CRF	97.30	93.69	83.02
	LSTM-CRF	**97.45**	93.80	84.10
	BI-LSTM-CRF	97.43	**94.13**	**84.26**
Senna	Conv-CRF (Collobert et al., 2011)	97.29	94.32	88.67 (89.59)
	LSTM	97.29	92.99	83.74
	BI-LSTM	97.40	93.92	85.17
	CRF	97.45	93.83	86.13
	LSTM-CRF	97.54	94.27	88.36
	BI-LSTM-CRF	**97.55**	**94.46**	**88.83 (90.10)**

Gazetteer(지명 사전) feature: set of lists containing names of entities such as cities, organizations, days of the week, etc.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Experiment Result

❖ 다른 모델과 비교 – POS Tagging

Table 4: Comparison of tagging accuracy of different models for POS.

System	accuracy	extra data
Maximum entropy cyclic dependency network (Toutanova et al., 2003)	97.24	No
SVM-based tagger (Gimenez and Marquez, 2004)	97.16	No
Bidirectional perceptron learning (Shen et al., 2007)	97.33	No
Semi-supervised condensed nearest neighbor (Soegaard, 2011)	97.50	Yes
CRFs with structure regularization (Sun, 2014)	97.36	No
Conv network tagger (Collobert et al., 2011)	96.37	No
Conv network tagger (senna) (Collobert et al., 2011)	97.29	Yes
BI-LSTM-CRF (ours)	**97.43**	No
BI-LSTM-CRF (Senna) (ours)	**97.55**	Yes

Extra data: Spelling features and Context features (Stanford NER tool (Finkel et al., 2005; Wang and Manning, 2013))

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Experiment Result

❖ 다른 모델과 비교 – NER Tagging

Table 6: Comparison of F1 scores of different models for NER.

System	accuracy
Combination of HMM, Maxent etc. (Florian et al., 2003)	88.76
MaxEnt classifier (Chieu., 2003)	88.31
Semi-supervised model combination (Ando and Zhang., 2005)	89.31
Conv-CRF (Collobert et al., 2011)	81.47
Conv-CRF (Senna + Gazetteer) (Collobert et al., 2011)	89.59
CRF with Lexicon Infused Embeddings (Passos et al., 2014)	**90.90**
BI-LSTM-CRF (ours)	84.26
BI-LSTM-CRF (Senna + Gazetteer) (ours)	90.10

Extra data: Spelling features and Context features (Stanford NER tool (Finkel et al., 2005; Wang and Manning, 2013))

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

감사합니다