

튜토리얼 – Sequence Tagging

Named Entity Recognition



KOREA
UNIVERSITY



Natural Language
Processing
& Artificial Intelligence



Background

Brief Introduction to
Named Entity Recognition and
Neural Networks



Named Entity Recognition

- NER Goal
 - It is to identify all Named Entities (NEs)
 1. Finding NEs
 2. Identify the type of NE found
 - When performing text mining in a specific field, it is better to learn Tagger and Recognizer by using a corpus suitable for it

Jim bought 300 shares of Acme Corp. in 2006.

-> Jim bought 300 shares of Acme Corp. in 2006.

Person

Organization

Time



BIO Tagging Scheme

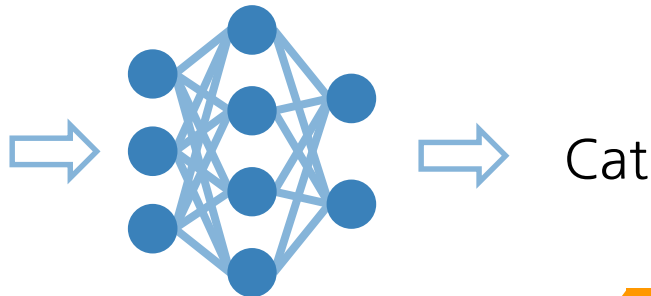
- Beginning-Inside-Outside (BIO)
- Used for correctly tag multi-word entities

| | | | | | | | |
|-----|----------|--------|-------|---------|-----|---------|-----|
| ... | Minister | Loyola | de | Palacio | had | earlier | ... |
| ... | O | B-PER | I-PER | I-PER | O | O | ... |



Neural Networks

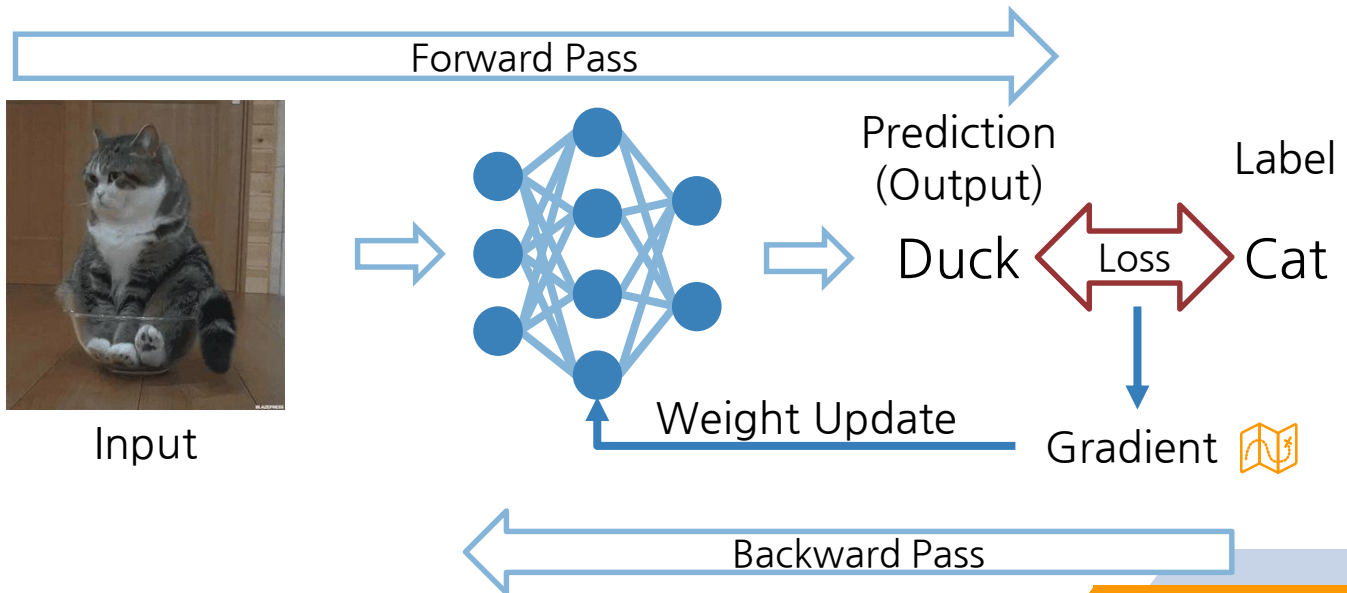
- Transforms input to output
- NN is a “set of weights(parameters)”
- Need to change the weights(train) to make it do what we want





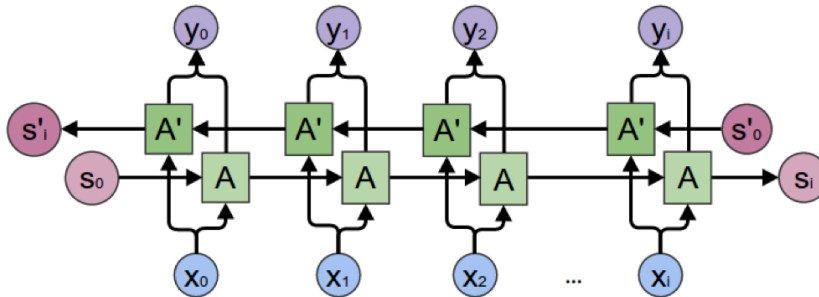
How to Train a NN

- With loss and gradient





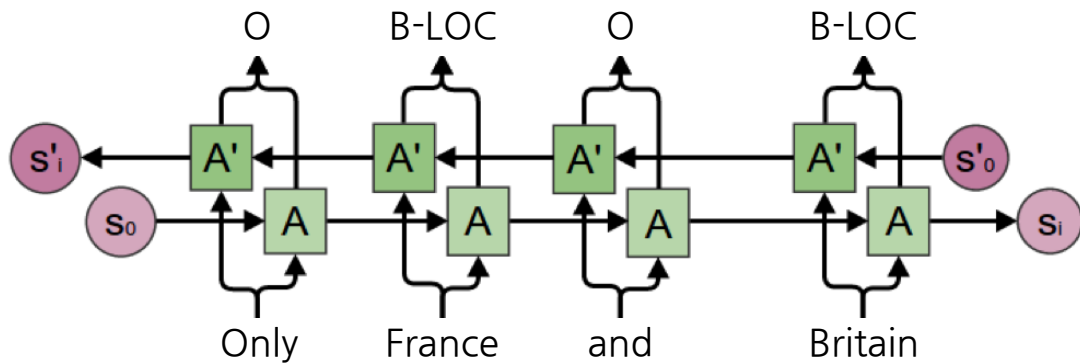
Model - Bidirectional RNN



- RNNs that combine two RNNs in different directions to enable bidirectional dependence
- Output y_t has input $[x_0, x_1, \dots, x_{t-1}]$ and $[x_{t+1}, x_{t+2}, \dots, x_N]$ is reflected

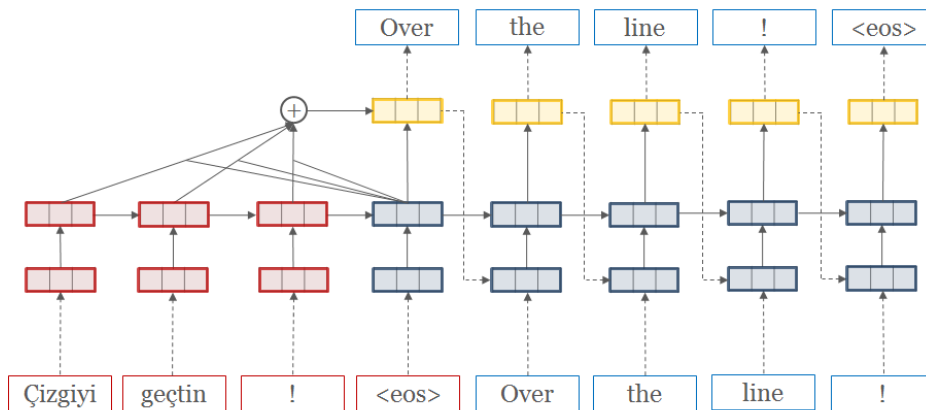


Model - Bidirectional RNN





Neural Network and Language



Human
-> Characters



Neural Network
-> Numbers



One-hot Encoding

$V = \{\text{cat, fat, mat, sat, the, on}\}$

cat = [1, 0, 0, 0, 0, 0]

fat = [0, 1, 0, 0, 0, 0]

mat = [0, 0, 1, 0, 0, 0]

sat = [0, 0, 0, 1, 0, 0]

the = [0, 0, 0, 0, 1, 0]

on = [0, 0, 0, 0, 0, 1]



Converting Characters to Numbers

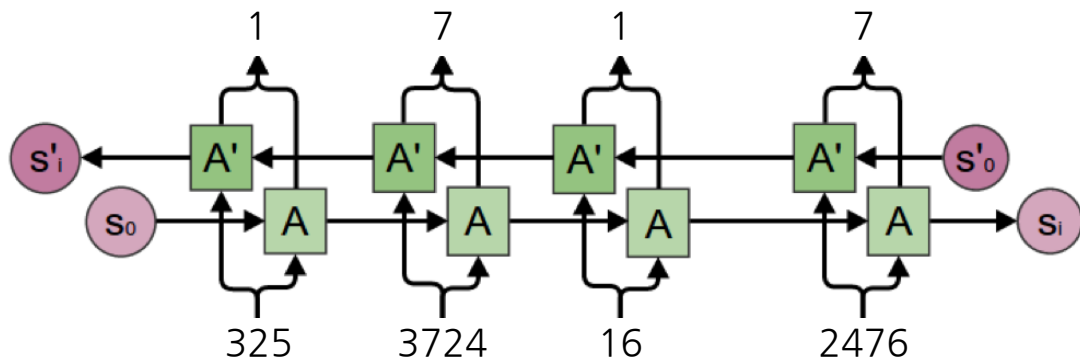
“... .. The fat cat sat
on the mat.”



“... .. 32 832 561 634
6132 32 565”



Model - Final





실습 - Sequence Tagger

Named Entity Recognition



Dataset - CoNLL-2003

- The SIGNLL Conference on Computational Natural Language Learning (CoNLL)
- Most frequently used open dataset for NER
- Four kind of entities - LOC(location), ORG(organization), PER(person), MISC(miscellaneous)
- With BIO scheme, nine possible tags - B-LOC, B-MISC, B-ORG, B-PER, I-LOC, I-MISC, I-ORG, I-PER, O



Files

- “CoNLL-2003” - Dataset
- “outdir” - Logger (debug, info)
- “save” - Trained model
- “sequence_tagger_blank.ipynb”
- 뼈대 코드
- “pieces.ipynb” - 코드 조각들



Links

- “sequence_tagger_blank.ipynb”
https://bit.ly/ku_sequence_tagger_blank
- “pieces.ipynb”
https://bit.ly/ku_sequence_tagger_pieces