Applied Text Analytics & Natural Language Processing

with Dr. Mahdi Roozbahani
& Wafa Louhichi

Sequence Labelling
Part-of-Speech (POS) Tagging

Learning Objectives

In this lesson, you will learn a sequence labeling method named Part-of-Speech (POS) tagging

- Average Perceptron Tagger
- BERT

Part-of-Speech (POS) Tagging

- Assigning a part-of-speech to a word in a sentence
  - Words can have more than one POS. *This is the sitting (adjective) area. I was sitting (verb) at my desk.*

  - Essential in NLP:
    - In MT (Machine Translation), helps when we need to reorder words in a phrase. E.g., *a spacious car (English) => une voiture spacieuse (French)*
    - In text-to-speech, helps with the pronunciation of the words, because it recognizes the verb tense. E.g., *I read the book everyday (/reed/), I read the book yesterday (/red/)*
    - Utilized by Automatic Speech Recognition (ASR) systems to calculate the performance of the transcription algorithm

https://web.stanford.edu/~jurafsky/slp3/slides/8_POSNER_intro_May_6_2021.pdf
https://en.wikipedia.org/wiki/Word_order

Part-of-Speech (POS) Tagging

- Assigning a part-of-speech to a word in a sentence
  - Words can have more than one POS. *This is the sitting (adjective) area. I was sitting (verb) at my desk.*

  - Essential in NLP:
    - In MT (Machine Translation), helps when we need to reorder words in a phrase. E.g., *a spacious car (English) => une voiture spacieuse (French)*
    - In text-to-speech, helps with the pronunciation of the words, because it recognizes the verb tense. E.g., *I read the book everyday (/reed/), I read the book yesterday (/red/)*
    - Utilized by Automatic Speech Recognition (ASR) systems to calculate the performance of the transcription algorithm

https://web.stanford.edu/~jurafsky/slp3/slides/8_POSNER_intro_May_6_2021.pdf
https://en.wikipedia.org/wiki/Word_order

Part-of-Speech (POS) Tagging

```
import nltk //a Python library for NLP tasks
from nltk.tokenize import word_tokenize
from nltk.tag import pos_tag

text = 'Professor Roozbahani traveled from Atlanta to Paris via United Airlines to visit the
Louvre Museum' // text to be processed

raw_words = word_tokenize(text) //divides strings into lists of substrings
pos_tags = pos_tag(raw_words) // attaches a POS tag to each word
print(pos_tag)
```

Output:

```
>> [('Professor', 'NNP'), ('Roozbahani', 'NNP'), ('traveled', 'VBD'), ('from', 'IN'), ('Atlanta',
'NNP'), ('to', 'TO'), ('Paris', 'NNP'), ('via', 'IN'), ('United', 'NNP'), ('Airlines', 'NNPS'),
('to', 'TO'), ('visit', 'VB'), ('the', 'DT'), ('Louvre', 'NNP'), ('Museum', 'NNP')]
```

Part-of-Speech (POS) Tagging

```
import nltk //a Python library for NLP tasks
from nltk.tokenize import word_tokenize
from nltk.tag import pos_tag

text = 'Professor Roozbahani traveled from Atlanta to Paris via United Airlines to visit the
Louvre Museum' // text to be processed

raw_words = word_tokenize(text) //divides strings into lists of substrings
pos_tags = pos_tag(raw_words) // attaches a POS tag to each word
print(pos_tag)
```

Output:

```
>> [('Professor', 'NNP'), ('Roozbahani', 'NNP'), ('traveled', 'VBD'), ('from', 'IN'), ('Atlanta',
'NNP'), ('to', 'TO'), ('Paris', 'NNP'), ('via', 'IN'), ('United', 'NNP'), ('Airlines', 'NNPS'),
('to', 'TO'), ('visit', 'VB'), ('the', 'DT'), ('Louvre', 'NNP'), ('Museum', 'NNP')]
```

Part-of-Speech (POS) Tagging – BERT Model

Using pre-trained BERT language model from Hugging Face

```
!pip install transformers //Library from Hugging Face with all the models
from transformers import AutoTokenizer, AutoModelForTokenClassification //importing tokenizer
and token classification model from hugging face transformer library
from transformers import pipeline


text = 'Professor Roozbahani traveled from Atlanta to Paris via United Airlines to visit the
Louvre Museum' // text to be processed

model_name = “QCRI/bert-base-multilingual-cased-pos-English” // a model fine-tuned for POS task
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForTokenClassification.from_pretrained(model_name)
classifier = pipeline(‘token-classification’, model = model, tokenizer = tokenizer)

classifier(text) //running the model on the example text
```

https://huggingface.co/QCRI/bert-base-multilingual-cased-pos-english

Part-of-Speech (POS) Tagging – BERT Model

Using pre-trained BERT model from Hugging Face

Output from POS task on the same text example:

```
>> [{'entity': 'NNP', 'score': 0.999403, 'index': 1, 'word': 'Professor', 'start': 0, 'end': 9}, {'entity': 'NNP',
'score': 0.99948466, 'index': 2, 'word': 'R', 'start': 10, 'end': 11}, {'entity': 'NNP', 'score': 0.99902, 'index': 3,
'word': '##oo', 'start': 11, 'end': 13}, {'entity': 'NNP', 'score': 0.99918383, 'index': 4, 'word': '##z', 'start': 13,
'end': 14}, {'entity': 'NNP', 'score': 0.99906355, 'index': 5, 'word': '##bahan', 'start': 14, 'end': 19}, {'entity':
'NNP', 'score': 0.99928385, 'index': 6, 'word': '##i', 'start': 19, 'end': 20}, {'entity': 'VBD', 'score': 0.99896204,
'index': 7, 'word': 'traveled', 'start': 21, 'end': 29}, {'entity': 'IN', 'score': 0.999567, 'index': 8, 'word': 'from',
'start': 30, 'end': 34}, {'entity': 'NNP', 'score': 0.9975382, 'index': 9, 'word': 'Atlanta', 'start': 35, 'end': 42},
{'entity': 'TO', 'score': 0.9998747, 'index': 10, 'word': 'to', 'start': 43, 'end': 45}, {'entity': 'NNP', 'score':
0.99850154, 'index': 11, 'word': 'Paris', 'start': 46, 'end': 51}, {'entity': 'IN', 'score': 0.99976784, 'index': 12,
'word': 'via', 'start': 52, 'end': 55}, {'entity': 'NNP', 'score': 0.9988519, 'index': 13, 'word': 'United', 'start': 56,
'end': 62}, {'entity': 'NNPS', 'score': 0.77026224, 'index': 14, 'word': 'Airlines', 'start': 63, 'end': 71}, {'entity':
'TO', 'score': 0.999897, 'index': 15, 'word': 'to', 'start': 72, 'end': 74}, {'entity': 'VB', 'score': 0.9995158, 'index':
16, 'word': 'visit', 'start': 75, 'end': 80}, {'entity': 'DT', 'score': 0.9998246, 'index': 17, 'word': 'the', 'start':
81, 'end': 84}, {'entity': 'NNP', 'score': 0.99938285, 'index': 18, 'word': 'Louvre', 'start': 85, 'end': 91}, {'entity':
'NNP', 'score': 0.9989183, 'index': 19, 'word': 'Museum', 'start': 92, 'end': 98}]
```

Note: Since we're using a pre-trained model, it is splitting the proper noun 'Roozbahani' because the language model might not recognize its origin. This is why often the language models are fine-tuned for domain-specific tasks.

Summary

- We learned about POS tagging using
- NLTK (Average perceptron tagger)
- BERT (Encoder-based transformer model)