Information Extraction

Tutor: Rahmad Mahendra

rahmad.mahendra@cs.ui.ac.id

Natural Language Processing & Text Mining
Short Course

Pusat Ilmu Komputer UI
22 – 26 Agustus 2016

pusilkom ui
quality results

HANDS ON POS TAGGER & NER

NER Task

- Given a sentence, get the list of token with the POS tag and Entity tag.
- Example:
  - Input:
    - “George Washington was unanimously elected the first President of the United States”
  - Output:

```
(S
  (PERSON George/NNP Washington/NNP)
  was/VBD
  unanimously/RB
  elected/VBN
  the/DT
  first/JJ
  President/NNP
  of/IN
  the/DT
  (GPE United/NNP States/NNPS)
)
```

POS & NER on Python: Tools Preparation

1. Install Anaconda Python
2. Install nltk module: pip install nltk
3. Download nltk module:
   - Enter python interactive mode
   - Type:
     - import nltk;
     - nltk.download();
   - Download:
     - Model: maxent_ne_chunker
     - Model: maxent_treebank_pos
     - Model: tagset
     - Model: punkt
     - Corpora: words

pusilkom ui
quality results

POS & NER on Python

- Step 1: Tokenization
  - Split each token (word) on a sentence

```
tok_sentence = nltk.word_tokenize(sentence);
print tok_sentence;
```

POS & NER on Python

- Step 2: POS Tagging
  - Identify the Part of Speech (POS) for each token

```
tagged_sentence = nltk.pos_tag(tok_sentence);
print tagged_sentence;
```

  - To see the list of tag set

```
nltk.help.upenn_tagset()
```

POS & NER on Python

- Step 3: NE Recognition
  - Given words and its POS, identify the entity of each token

```
tree = nltk.ne_chunk(tagged_sentence);
print tree;
```

POS & NER on Python

Make file D:\NER\ner.py

```
import nltk;

sentence = "George Washington was unanimously elected the first
President of the United States";

tok_sentence = nltk.word_tokenize(sentence);
print tok_sentence;

tagged_sentence = nltk.pos_tag(tok_sentence);
print tagged_sentence;

tree = nltk.ne_chunk(tagged_sentence);
print tree;

tree.draw();
```

POS & NER on Python

- Error on reconition

```
(S
  (PERSON George/NNP)
  (GPE Washington/NNP)
  was/VBD
  unanimously/RB
  elected/VBN
  the/DT
  first/JJ
  President/NNP
  of/IN
  the/DT
  (GPE United/NNP States/NNPS)
)
```

- Why this can be happened?

- How to fix (improve) the NER?

How to Train POS Model

1. Prepare the corpus
   - IO / IOB Format
   - Make the *.iob files to your corpus directory
     - C:/POS/corpus

How to Train POS Model

2. Build POS model using your own corpus
   - Download the script for train NER model:
   - Extract the zip file to
     - C:/POS/
   - Run the following commands:

```
>>> C:
>>> cd POS\trainer
>>> python train_tagger.py "C:\POS\corpus" --reader
nltk.corpus.reader.tagged.TaggedCorpusReader --sequential b --
classifier Maxent --filename C:\POS\model.pickles --no-eval
```

How to Train POS Model

3. Use the model
   - Copy:
     - C:\POS\trainer\nltk_trainer
       to
     - C:\POS\
   - Make: C:\POS\pos.py

```
import nltk;
import pickle;
tagger = pickle.load(open("C:\POS\model.pickles"))
sentence = "John Kuhns eats pizza";
tok_sentence = nltk.word_tokenize(sentence);
tagged_sentence = tagger.tag(tok_sentence);
print tagged_sentence;
```

How to Train NER Model

1. Prepare the corpus
   - IO / IOB Format
   - Make the *.iob files to your corpus directory
     - C:/NER/corpus

How to Train NER Model

2. Build NER model using your own corpus
   - Download the script for train NER model:
   - Extract the zip file to
     - C:/NER/
   - Run the following commands:

```
>>> C:
>>> cd NER\trainer
>>> python train_chunker.py "C:\NER\corpus" --reader
nltk.corpus.reader.ChunkedCorpusReader --fraction 0.85 --sequential
'' --classifier Maxent --filename C:\NER\model.pickles --no-eval
```

pusilkom ui
quality results

How to Train NER Model

3. Use the model
   - Copy:
     - C:\NER\trainer\nltk_trainer
       to
     - C:\NER\
   - Make: C:\NER\ner2.py

```
import nltk;
import pickle;
chunker = pickle.load(open("C:\NER\model.pickles"))
sentence = "John Kuhns eats pizza";
tok_sentence = nltk.word_tokenize(sentence);
tagged_sentence = nltk.pos_tag(tok_sentence);
tree = chunker.parse(tagged_sentence);
tree.draw();
```

How to Train NER Model

- Try for another sentence

```
sentence = "George Washington eats pizza";
```

- What happened?
- Take a moment to fix your model.

```
George NNP B-PERS
Washington NNP I-PERS
was VBD O
the DT O
first JJ O
president NN O
of IN O
America NNP B-LOC
```

Using Real Data

- Build your own model using large data.
- You can make your own IOB file or generate it from available corpus.
  - http://schwa.org/projects/resources/wiki/Wikiner
  - https://github.com/famrashel/idn-tagged-corpus/blob/master/Indonesian_Manually_Tagged_Corpus.tsv

Thank you