Min-Yuh Day
Associate Professor
Dept. of Information Management, Tamkang University

http://mail.tku.edu.tw/myday/
2020-05-13
課程大綱 (Syllabus)

週次 (Week) 日期 (Date)  內容 (Subject/Topics)
1 2020/03/04 人工智慧文本分析課程介紹
   (Course Orientation on Artificial Intelligence for Text Analytics)
2 2020/03/11 文本分析的基礎：自然語言處理
   (Foundations of Text Analytics: Natural Language Processing; NLP)
3 2020/03/18 Python自然語言處理
   (Python for Natural Language Processing)
4 2020/03/25 處理和理解文本
   (Processing and Understanding Text)
5 2020/04/01 文本表達特徵工程
   (Feature Engineering for Text Representation)
6 2020/04/08 人工智慧文本分析個案研究 I
   (Case Study on Artificial Intelligence for Text Analytics I)
課程大綱 (Syllabus)

週次 (Week)  日期 (Date)  內容 (Subject/Topics)
7  2020/04/15  文本分類 (Text Classification)
8  2020/04/22  文本摘要和主題模型
               (Text Summarization and Topic Models)
9  2020/04/29  期中報告 (Midterm Project Report)
10  2020/05/06  文本相似度和分群 (Text Similarity and Clustering)
11  2020/05/13  語意分析和命名實體識別
                (Semantic Analysis and Named Entity Recognition; NER)
12  2020/05/20  情感分析 (Sentiment Analysis)
<table>
<thead>
<tr>
<th>週次 (Week)</th>
<th>日期 (Date)</th>
<th>內容 (Subject/Topics)</th>
</tr>
</thead>
</table>
| 13          | 2020/05/27  | 人工智慧文本分析個案研究Ⅱ  
(Case Study on Artificial Intelligence for Text Analytics II) |
| 14          | 2020/06/03  | 深度學習和通用句子嵌入模型  
(Deep Learning and Universal Sentence-Embedding Models) |
| 15          | 2020/06/10  | 問答系統與對話系統  
(Question Answering and Dialogue Systems) |
| 16          | 2020/06/17  | 期末報告 I (Final Project Presentation I) |
| 17          | 2020/06/24  | 期末報告 II (Final Project Presentation II) |
| 18          | 2020/07/01  | 教師彈性補充教學 |
Semantic Analysis and Named Entity Recognition (NER)
Outline

• Semantic Analysis
  • WordNet
  • Word sense disambiguation
• Named Entity Recognition (NER)
Semantic Analysis

• Semantics
  – the study of meaning

• Linguistic semantics
  – the study of meaning in natural language.

Semantic Analysis and NER

• WordNet and synsets
  – Analyzing lexical semantic relations
  – Word sense disambiguation

• Named entity recognition

• Analyzing semantic representations

What is WordNet?

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the creators of WordNet and do not necessarily reflect the views of any funding agency or Princeton University.

When writing a paper or producing a software application, tool, or interface based on WordNet, it is necessary to properly cite the source. Citation figures are critical to WordNet funding.

About WordNet

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet is also freely and publicly available for download. WordNet’s structure makes it a useful tool for computational linguistics and natural language processing.

WordNet superficially resembles a thesaurus, in that it groups words together based on their

https://wordnet.princeton.edu/
NLP

Classical NLP

Deep Learning-based NLP

Source: http://blog.aylien.com/leveraging-deep-learning-for-multilingual/
Modern NLP Pipeline

Modern NLP Pipeline
Deep Learning NLP

Documents → Preprocessing → Dense Word Embeddings → Deep Neural Network

Task / Output
- Classification
- Sentiment Analysis
- Entity Extraction
- Topic Modeling
- Document Similarity

Pre-generated Lookup OR Generated in 1st level of NeuralNet

Source: http://mattfortier.me/2017/01/31/nlp-intro-pt-1-overview/
Natural Language Processing (NLP) and Text Mining

1. Raw text
2. Sentence Segmentation
3. Tokenization
4. Part-of-Speech (POS)
5. Stop word removal
6. Stemming / Lemmatization
7. Dependency Parser
8. String Metrics & Matching

- word’s stem:
  - am → am
  - having → hav

- word’s lemma:
  - am → be
  - having → have

Source: Nitin Hardeniya (2015), NLTK Essentials, Packt Publishing; Florian Leitner (2015), Text mining - from Bayes rule to dependency parsing
Analyzing Lexical Semantic Relationships

• Entailments
• Homonyms and Homographs
• Synonyms and Antonyms
• Hyponyms and Hypernyms
• Holonyms and Meronyms
• Semantic Relationships and Similarity

Word Sense Disambiguation

• Lesk algorithm (Lesk, 1986)

  – leverage dictionary or **vocabulary definitions** for a word we want to disambiguate in a body of text and compare the words in these **definitions** with a section of text surrounding our word of interest.

  – The main objective is to return the **synset** with the maximum number of overlapping words or terms between the context sentence and the **different definitions** from each **synset** for the word we target for **disambiguation**.
Named Entity Recognition (NER)

• Named entities
  – represent real-world objects
  – people, places, organizations
  – proper names

• Named entity recognition
  – Entity chunking
  – Entity extraction

# NER: OntoNotes 5 Named Entities

<table>
<thead>
<tr>
<th>SID</th>
<th>TYPE</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PERSON</td>
<td>People, including fictional.</td>
</tr>
<tr>
<td>2</td>
<td>NORP</td>
<td>Nationalities or religious or political groups.</td>
</tr>
<tr>
<td>3</td>
<td>FAC</td>
<td>Buildings, airports, highways, bridges, etc.</td>
</tr>
<tr>
<td>4</td>
<td>ORG</td>
<td>Companies, agencies, institutions, etc.</td>
</tr>
<tr>
<td>5</td>
<td>GPE</td>
<td>Countries, cities, states.</td>
</tr>
<tr>
<td>6</td>
<td>LOC</td>
<td>Non-GPE locations, mountain ranges, bodies of water.</td>
</tr>
<tr>
<td>7</td>
<td>PRODUCT</td>
<td>Objects, vehicles, foods, etc. (Not services.)</td>
</tr>
<tr>
<td>8</td>
<td>EVENT</td>
<td>Named hurricanes, battles, wars, sports events, etc.</td>
</tr>
<tr>
<td>9</td>
<td>WORK_OF_ART</td>
<td>Titles of books, songs, etc.</td>
</tr>
<tr>
<td>10</td>
<td>LAW</td>
<td>Named documents made into laws.</td>
</tr>
<tr>
<td>11</td>
<td>LANGUAGE</td>
<td>Any named language.</td>
</tr>
<tr>
<td>12</td>
<td>DATE</td>
<td>Absolute or relative dates or periods.</td>
</tr>
<tr>
<td>13</td>
<td>TIME</td>
<td>Times smaller than a day.</td>
</tr>
<tr>
<td>14</td>
<td>PERCENT</td>
<td>Percentage, including ”%“</td>
</tr>
<tr>
<td>15</td>
<td>MONEY</td>
<td>Monetary values, including unit.</td>
</tr>
<tr>
<td>16</td>
<td>QUANTITY</td>
<td>Measurements, as of weight or distance.</td>
</tr>
<tr>
<td>17</td>
<td>ORDINAL</td>
<td>“first”, “second”, etc.</td>
</tr>
<tr>
<td>18</td>
<td>CARDINAL</td>
<td>Numerals that do not fall under another type.</td>
</tr>
</tbody>
</table>

Source: [https://spacy.io/api/annotation#named-entities](https://spacy.io/api/annotation#named-entities)
# NER: Wikipedia Named Entities

<table>
<thead>
<tr>
<th>SID</th>
<th>TYPE</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PER</td>
<td>Named person or family.</td>
</tr>
<tr>
<td>2</td>
<td>LOC</td>
<td>Name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains).</td>
</tr>
<tr>
<td>3</td>
<td>ORG</td>
<td>Named corporate, governmental, or other organizational entity.</td>
</tr>
<tr>
<td>4</td>
<td>MISC</td>
<td>Miscellaneous entities, e.g. events, nationalities, products or works of art.</td>
</tr>
</tbody>
</table>
# NER IOB Scheme

<table>
<thead>
<tr>
<th>TAG</th>
<th>ID</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;I&quot;</td>
<td>1</td>
<td>Token is <em>inside</em> an entity.</td>
</tr>
<tr>
<td>&quot;O&quot;</td>
<td>2</td>
<td>Token is <em>outside</em> an entity.</td>
</tr>
<tr>
<td>&quot;B&quot;</td>
<td>3</td>
<td>Token <em>begins</em> an entity.</td>
</tr>
<tr>
<td>&quot;&quot;</td>
<td>0</td>
<td>No entity tag is set (missing value).</td>
</tr>
</tbody>
</table>

Source: [https://spacy.io/api/annotation#named-entities](https://spacy.io/api/annotation#named-entities)
# NER BILUO Scheme

<table>
<thead>
<tr>
<th>TAG</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BEGIN</strong></td>
<td>The first token of a multi-token entity.</td>
</tr>
<tr>
<td><strong>IN</strong></td>
<td>An inner token of a multi-token entity.</td>
</tr>
<tr>
<td><strong>LAST</strong></td>
<td>The final token of a multi-token entity.</td>
</tr>
<tr>
<td><strong>UNIT</strong></td>
<td>A single-token entity.</td>
</tr>
<tr>
<td><strong>OUT</strong></td>
<td>A non-entity token.</td>
</tr>
</tbody>
</table>

Source: [https://spacy.io/api/annotation#named-entities](https://spacy.io/api/annotation#named-entities)
BERT Sequence-level tasks

(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

BERT Token-level tasks

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

NER: Single Sentence Tagging

NER: Fine-tuning BERT with Bi-LSTM CRF

Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT

Semantic Analysis and Named Entity Recognition (NER)


Semantic Analysis

```python
1 import nltk
2 from nltk.corpus import wordnet as wn
3 import pandas as pd
4 nltk.download('wordnet')
5 # WordNet Synsets
6 word = 'fruit'
7 synsets = wn.synsets(word)
8 print('Word:', word)
9 print('Wordnet Synsets:', len(synsets))
10 df = pd.DataFrame([['Synset': synset,
11         'Part of Speech': synset.lexname(),
12         'Definition': synset.definition(),
13         'Lemmas': synset.lemma_names(),
14         'Examples': synset.examples()]
15         for synset in synsets])
16 df

[1] Downloading package wordnet to /root/nltk_data...
Word: fruit
Wordnet Synsets: 5

<table>
<thead>
<tr>
<th>Synset</th>
<th>Part of Speech</th>
<th>Definition</th>
<th>Lemmas</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synset('fruit.n.01')</td>
<td>noun.plant</td>
<td>the ripened reproductive body of a seed plant</td>
<td>[fruit]</td>
<td>[]</td>
</tr>
<tr>
<td>Synset('yield.n.03')</td>
<td>noun.artifact</td>
<td>an amount of a product</td>
<td>[yield, fruit]</td>
<td>[]</td>
</tr>
</tbody>
</table>
```

https://tinyurl.com/imtkupython101
November with the intention of hearing from Zuckerberg. Since the Cambridge Analytica scandal broke, the Facebook chief has only appeared in front of two legislatures: the American Senate and House of Representatives, and the European parliament. Facebook has consistently rebuffed attempts from others, including the UK and Canadian parliaments, to hear from Zuckerberg. He added that an article in the New York Times on Thursday, in which the paper alleged a pattern of behaviour from Facebook to “delay, deny and deflect” negative news stories, “raises further questions about how recent data breaches were allegedly dealt with within Facebook.”

Three more countries have joined an “international grand committee” of parliaments, adding to calls for Facebook’s boss, Zuckerberg.

text_nlp: Three more countries have joined an “international grand committee” of parliaments, adding to calls for Facebook’s boss, Zuckerberg.

```python
# print named entities in article
ger_tagged = [(word.text, word.ENT_TYPE_) for word in text_nlp]
print(ner_tagged)
```

```python
# visualize named entities
from spacy import displacy
ner_tagged = [(word.text, word.ENT_TYPE_) for word in text_nlp]
displacy.render(ner_tagged, style='ent', jupyter=True)
```

Three more countries have joined an “international grand committee” of parliaments, adding to calls for Facebook’s boss, Zuckerberg, to give evidence on misinformation to the coalition. Brazil, Latvia, and Singapore bring the total to eight different parliaments across the world, with plans to send representatives to London on 27 November with the intention of hearing from Zuckerberg. Since the Cambridge Analytica scandal broke, the Facebook chief has only appeared in front of two legislatures: the American Senate and House of Representatives, and the European parliament. Facebook has consistently rebuffed attempts from others, including the UK and Canadian parliaments, to hear from Zuckerberg. He added that an article in the New York Times on Thursday, in which the paper alleged a pattern of behaviour from Facebook to “delay, deny and deflect” negative news stories, “raises further questions about how recent data breaches were allegedly dealt with within Facebook.”
Summary

• Semantic Analysis
  • WordNet
  • Word sense disambiguation
• Named Entity Recognition (NER)
References


• The Super Duper NLP Repo, https://notebooks.quantumstat.com/