語意分析和命名實體識別
(Semantic Analysis and Named Entity Recognition; NER)

Time: 2020/06/05 (Fri) (9:10 -12:00)
Place: 國立台北護理健康大學 (台北市明德路365號) G210
Host: 祝國忠 院長 (健康科技學院院長)

Min-Yuh Day
戴敏育
Associate Professor
副教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail.tku.edu.tw/myday/
2020-06-05
1. Core Technologies of Natural Language Processing and Text Mining
2. Artificial Intelligence for Text Analytics: Foundations and Applications
3. Feature Engineering for Text Representation
4. Semantic Analysis and Named Entity Recognition; NER
5. Deep Learning and Universal Sentence-Embedding Models
6. Question Answering and Dialogue Systems
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

Documents

Language Detection

English

Spanish

Arabic

Pre-processing

Tokenization (English)

POS Tagging (English)

Stopword Removal (EN)

...

...

...

Modeling

Feature Extraction (EN)

Feature Extraction (ES)

Feature Extraction (AR)

...

...

...

Output

Sentiment

Classification

Entity Extraction

Translation

Topic Modelling

...

...

...

Deep Learning-based NLP

Documents

Preprocessing

Dense Embeddings

obtained via word2vec, doc2vec, GloVe, etc.

Hidden Layers

Output Units

Output

Sentiment

Classification

Entity Extraction

Translation

Topic Modelling

...

...

...

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

Pre-processing

Documents

Language Detection

Tokenize

POS Tagging

... 

Token Filtering

Pre-processed Documents

Documents

Build Vocabulary

Pre-processed Documents

Bag-of-Words & Vectorization

Word Embeddings

word2vec

doc2vec

GloVe

Machine Learning

(Deep) Neural Network

Task / Output

Classification

Sentiment Analysis

Entity Extraction

Topic Modeling

Similarity

Modern NLP Pipeline

Documents → Language Detection → Preprocessing → Modeling

EN: Preprocessing → Modeling
ZH: Preprocessing → Preprocessing → Modeling

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

Source: http://mattfortier.me/2017/01/31/nlp-intro-pt-1-overview/
Deep Learning NLP

Documents → Preprocessing → Dense Word Embeddings → Deep Neural Network

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

Task / Output

- Classification
- Sentiment Analysis
- Entity Extraction
- Topic Modeling
- Document Similarity

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

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Segmentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tokenization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part-of-Speech (POS)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stop word removal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stemming / Lemmatization</td>
<td>Transform word to its base form</td>
<td>am → am, having → hav</td>
</tr>
<tr>
<td>Dependency Parser</td>
<td></td>
<td></td>
</tr>
<tr>
<td>String Metrics &amp; Matching</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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>

Source: [https://spacy.io/api/annotation#named-entities](https://spacy.io/api/annotation#named-entities)
## 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 inside an entity.</td>
</tr>
<tr>
<td>&quot;O&quot;</td>
<td>2</td>
<td>Token is outside an entity.</td>
</tr>
<tr>
<td>&quot;B&quot;</td>
<td>3</td>
<td>Token begins 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>BEGIN</td>
<td>The first token of a multi-token entity.</td>
</tr>
<tr>
<td>IN</td>
<td>An inner token of a multi-token entity.</td>
</tr>
<tr>
<td>LAST</td>
<td>The final token of a multi-token entity.</td>
</tr>
<tr>
<td>UNIT</td>
<td>A single-token entity.</td>
</tr>
<tr>
<td>OUT</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

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

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.lema().
12                  'Definition': synset.definition(),
13                  'Lemmas': synset.lemma_names(),
14                  'Examples': synset.examples()}
15              for synset in synsets])
16 df
```

- [nltk_data] Downloading package wordnet to /root/nltk_data...
- [nltk_data] Unzipping corpora/wordnet.zip.

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>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>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 within Facebook.”

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

https://tinyurl.com/imtkupython101
## NLP Benchmark Datasets

<table>
<thead>
<tr>
<th>Task</th>
<th>Dataset</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WMT 2014 EN-FR</td>
<td></td>
</tr>
<tr>
<td><strong>Text Summarization</strong></td>
<td>CNN/DM</td>
<td><a href="https://cs.nyu.edu/~kcho/DMQA/">https://cs.nyu.edu/~kcho/DMQA/</a></td>
</tr>
<tr>
<td></td>
<td>Newsroom</td>
<td><a href="https://summari.es/">https://summari.es/</a></td>
</tr>
<tr>
<td></td>
<td>Gigaword</td>
<td><a href="https://catalog.ldc.upenn.edu/LDC2012T21">https://catalog.ldc.upenn.edu/LDC2012T21</a></td>
</tr>
<tr>
<td><strong>Reading Comprehension</strong></td>
<td>ARC</td>
<td><a href="http://data.allenai.org/arc/">http://data.allenai.org/arc/</a></td>
</tr>
<tr>
<td><strong>Question Answering</strong></td>
<td>CliCR</td>
<td><a href="http://aclweb.org/anthology/N18-1140">http://aclweb.org/anthology/N18-1140</a></td>
</tr>
<tr>
<td><strong>Question Generation</strong></td>
<td>CNN/DM</td>
<td><a href="https://cs.nyu.edu/~kcho/DMQA/">https://cs.nyu.edu/~kcho/DMQA/</a></td>
</tr>
<tr>
<td></td>
<td>NewsQA</td>
<td><a href="https://datasets.maluuba.com/NewsQA">https://datasets.maluuba.com/NewsQA</a></td>
</tr>
<tr>
<td></td>
<td>RACE</td>
<td><a href="http://www.qizhexie.com/data/RACE_leaderboard">http://www.qizhexie.com/data/RACE_leaderboard</a></td>
</tr>
<tr>
<td></td>
<td>SQuAD</td>
<td><a href="https://rajpurkar.github.io/SQuAD-explorer/">https://rajpurkar.github.io/SQuAD-explorer/</a></td>
</tr>
<tr>
<td></td>
<td>NarrativeQA</td>
<td><a href="https://github.com/deepmind/narrativeqa">https://github.com/deepmind/narrativeqa</a></td>
</tr>
<tr>
<td></td>
<td>Quasar</td>
<td><a href="https://github.com/bd">https://github.com/bd</a> Bowling/quasar</td>
</tr>
<tr>
<td></td>
<td>SearchQA</td>
<td><a href="https://github.com/nyu-dl/SearchQA">https://github.com/nyu-dl/SearchQA</a></td>
</tr>
<tr>
<td><strong>Semantic Parsing</strong></td>
<td>AMR parsing</td>
<td><a href="https://amr.isi.edu/index.html">https://amr.isi.edu/index.html</a></td>
</tr>
<tr>
<td></td>
<td>ATIS (SQL Parsing)</td>
<td><a href="https://github.com/jkkummerfeld/text2sql-data/tree/master/data">https://github.com/jkkummerfeld/text2sql-data/tree/master/data</a></td>
</tr>
<tr>
<td></td>
<td>WikiSQL (SQL Parsing)</td>
<td><a href="https://github.com/salesforce/WikiSQL">https://github.com/salesforce/WikiSQL</a></td>
</tr>
<tr>
<td><strong>Sentiment Analysis</strong></td>
<td>IMDB Reviews</td>
<td><a href="http://ai.stanford.edu/~amaas/data/sentiment/">http://ai.stanford.edu/~amaas/data/sentiment/</a></td>
</tr>
<tr>
<td></td>
<td>SST</td>
<td><a href="https://nlp.stanford.edu/sentiment/index.html">https://nlp.stanford.edu/sentiment/index.html</a></td>
</tr>
<tr>
<td></td>
<td>Yelp Reviews</td>
<td><a href="https://www.yelp.com/dataset/challenge">https://www.yelp.com/dataset/challenge</a></td>
</tr>
<tr>
<td></td>
<td>Subjectivity Dataset</td>
<td><a href="http://www.cs.cornell.edu/people/pabo/movie-review-data/">http://www.cs.cornell.edu/people/pabo/movie-review-data/</a></td>
</tr>
<tr>
<td><strong>Text Classification</strong></td>
<td>AG News</td>
<td><a href="http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html">http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html</a></td>
</tr>
<tr>
<td></td>
<td>DBpedia</td>
<td><a href="https://wiki.dbpedia.org/Datasets">https://wiki.dbpedia.org/Datasets</a></td>
</tr>
<tr>
<td></td>
<td>TREC</td>
<td><a href="https://trec.nist.gov/data.html">https://trec.nist.gov/data.html</a></td>
</tr>
<tr>
<td></td>
<td>20 NewsGroup</td>
<td><a href="http://qwone.com/~jason/20Newsgroups/">http://qwone.com/~jason/20Newsgroups/</a></td>
</tr>
<tr>
<td><strong>Natural Language Inference</strong></td>
<td>SNLI Corpus</td>
<td><a href="https://nlp.stanford.edu/projects/snli/">https://nlp.stanford.edu/projects/snli/</a></td>
</tr>
<tr>
<td></td>
<td>MultiNLI</td>
<td><a href="https://www.nyu.edu/projects/bowman/multinli/">https://www.nyu.edu/projects/bowman/multinli/</a></td>
</tr>
<tr>
<td></td>
<td>SciTail</td>
<td><a href="http://data.allenai.org/scitail/">http://data.allenai.org/scitail/</a></td>
</tr>
<tr>
<td><strong>Semantic Role Labeling</strong></td>
<td>Proposition Bank</td>
<td><a href="http://propbank.github.io/">http://propbank.github.io/</a></td>
</tr>
<tr>
<td></td>
<td>OneNotes</td>
<td><a href="https://catalog.ldc.upenn.edu/LDC2013T19">https://catalog.ldc.upenn.edu/LDC2013T19</a></td>
</tr>
</tbody>
</table>

Summary

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


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