國立臺北護理健康大學
National Taipei University of Nursing and Health Sciences

Tamkang University
淡江大學

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: 國立臺北護理健康大學 (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

Topics

1. 自然語言處理核心技術與文字探勘
   (Core Technologies of Natural Language Processing and Text Mining)
2. 人工智慧文本分析基礎與應用
   (Artificial Intelligence for Text Analytics: Foundations and Applications)
3. 文本表達特徵工程
   (Feature Engineering for Text Representation)
4. **語意分析和命名實體識別**
   (Semantic Analysis and Named Entity Recognition; NER)
5. 深度學習和通用句子嵌入模型
   (Deep Learning and Universal Sentence-Embedding Models)
6. 問答系統與對話系統
   (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.

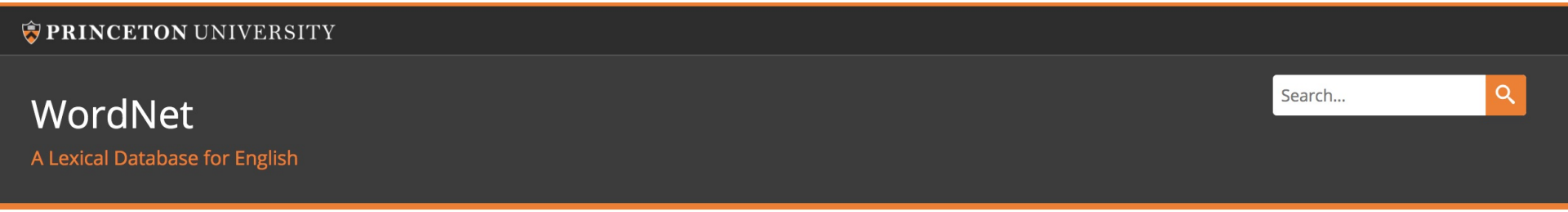
Semantic Analysis and NER

- WordNet and synsets
  - Analyzing lexical semantic relations
  - Word sense disambiguation
- Named entity recognition
- Analyzing semantic representations

Source: Dipanjan Sarkar (2019), Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second Edition. APress.

WordNet

A Lexical Database for English



- What is WordNet
- People
- News
- Use Wordnet Online
- Download
- Citing WordNet
- License and Commercial Use
- Related Projects
- Documentation
- Publications
- Frequently Asked

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

Note

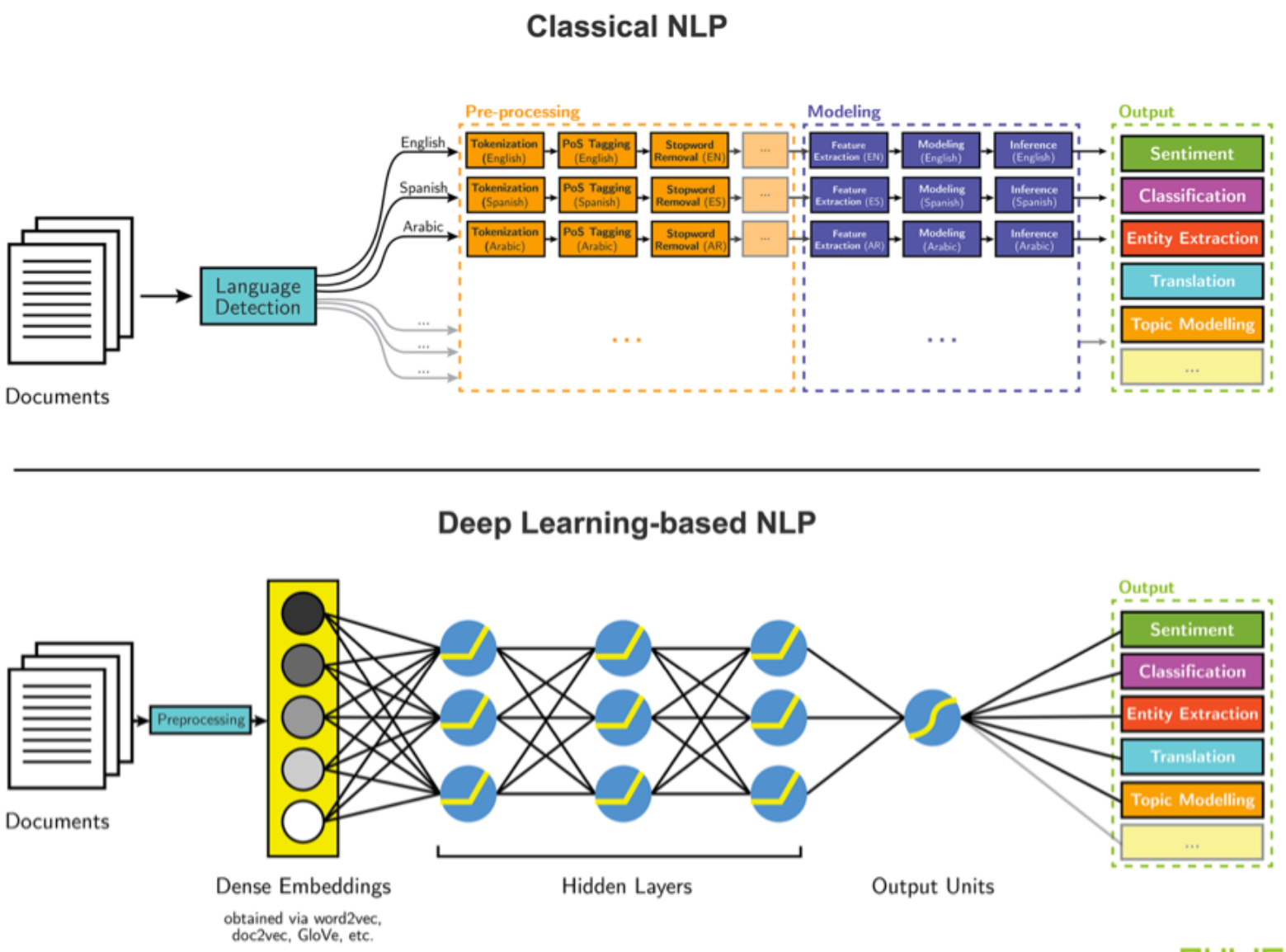
Due to funding and staffing issues, we are no longer able to accept comment and suggestions.

We get numerous questions regarding topics that are addressed on our FAQ page. If you have a problem or question regarding something you downloaded from the "Related projects" page, you must contact the developer directly.

Please note that any changes

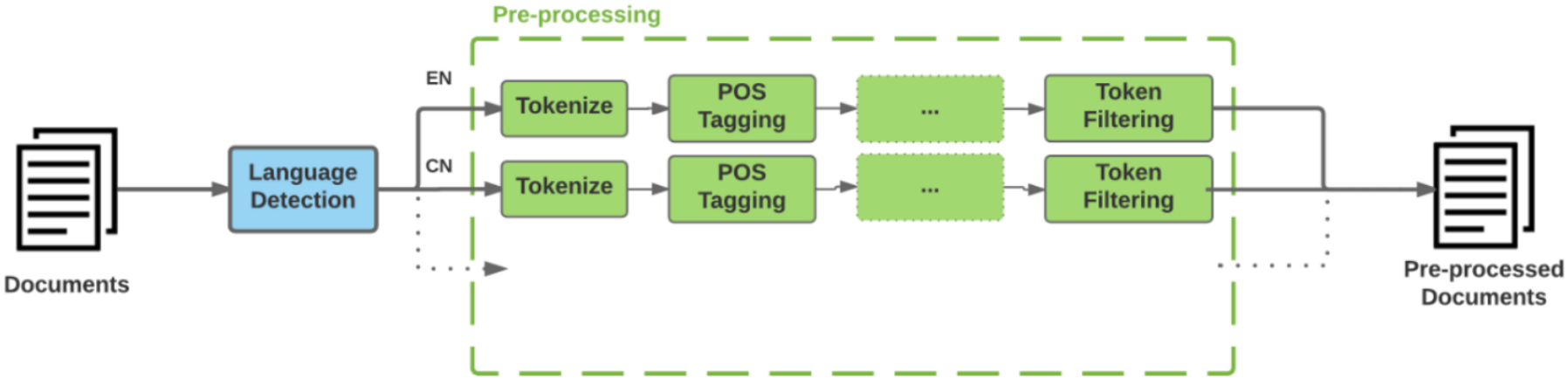
https://wordnet.princeton.edu/

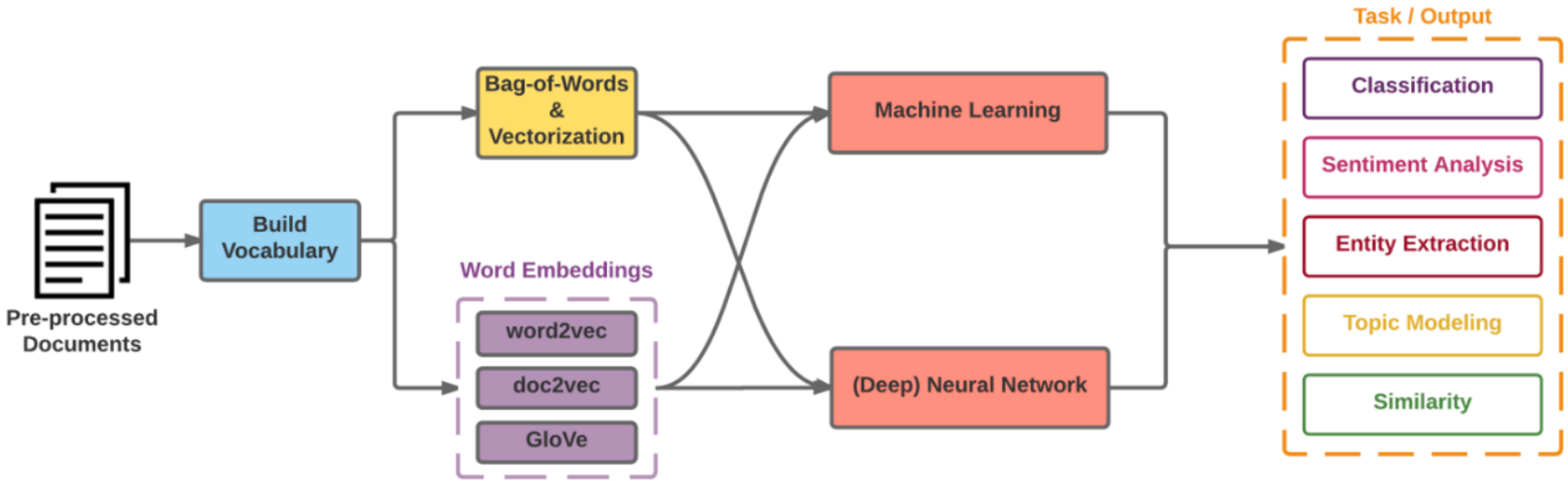
NLP



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

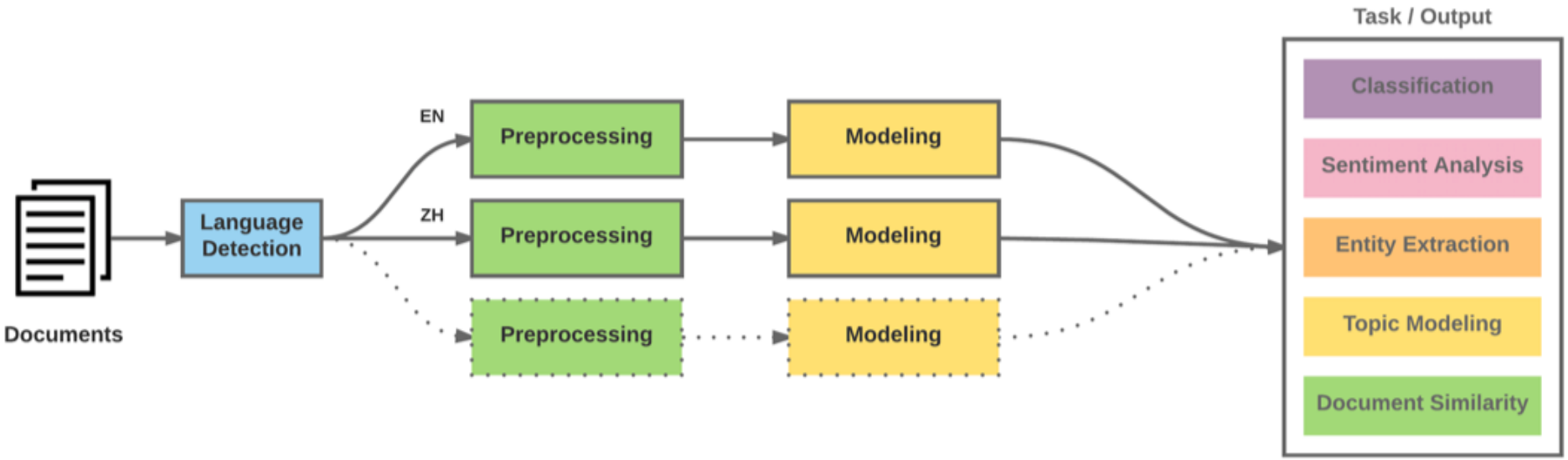
Modern NLP Pipeline





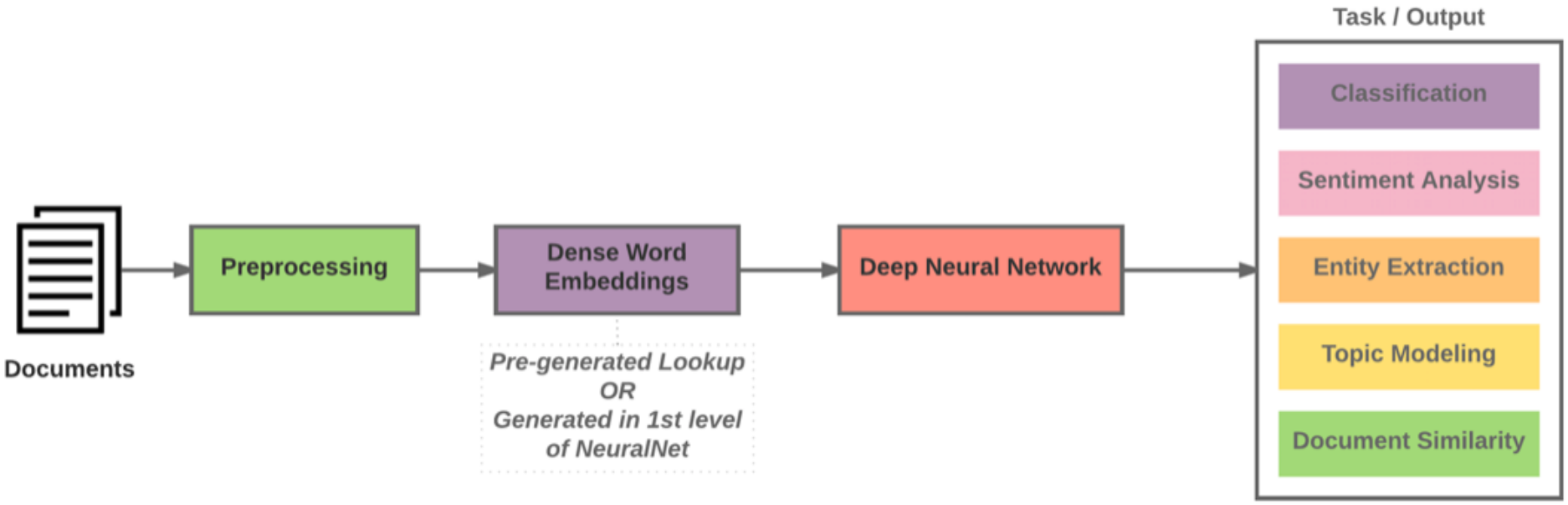
Source: https://github.com/fortiema/talks/blob/master/opendata2016sh/pragmatic-nlp-opendata2016sh.pdf

Modern NLP Pipeline



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

Deep Learning NLP



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

Natural Language Processing (NLP) and Text Mining

- Raw text
- Sentence Segmentation
- Tokenization
- Part-of-Speech (POS)
- Stop word removal
- Stemming / Lemmatization
- Dependency Parser
- String Metrics & Matching

word's stem	word's lemma
am → am	am → be
having → hav	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.

Source: Dipanjan Sarkar (2019), Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second Edition. APress.

Named Entity Recognition (NER)

- **Named entities**
  - represent real-world objects
  - people, places, organizations
  - proper names
- **Named entity recognition**
  - Entity chunking
  - Entity extraction

Source: Dipanjan Sarkar (2019), Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second Edition. APress.

NER: OntoNotes 5 Named Entities

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

Source: https://spacy.io/api/annotation#named-entities

NER: Wikipedia Named Entities

SID	TYPE	DESCRIPTION
1	PER	Named person or family.
2	LOC	Name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains).
3	ORG	Named corporate, governmental, or other organizational entity.
4	MISC	Miscellaneous entities, e.g. events, nationalities, products or works of art.

Source: https://spacy.io/api/annotation#named-entities

NER IOB Scheme

TAG	ID	DESCRIPTION
"I"	1	Token is inside an entity.
"O"	2	Token is outside an entity.
"B"	3	Token begins an entity.
""	0	No entity tag is set (missing value).

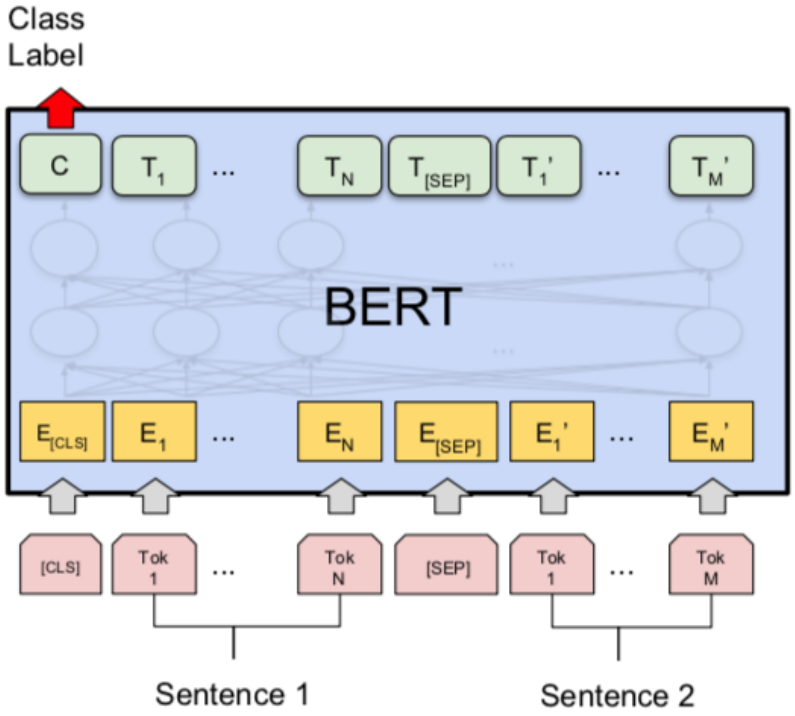
Source: https://spacy.io/api/annotation#named-entities

NER BILUO Scheme

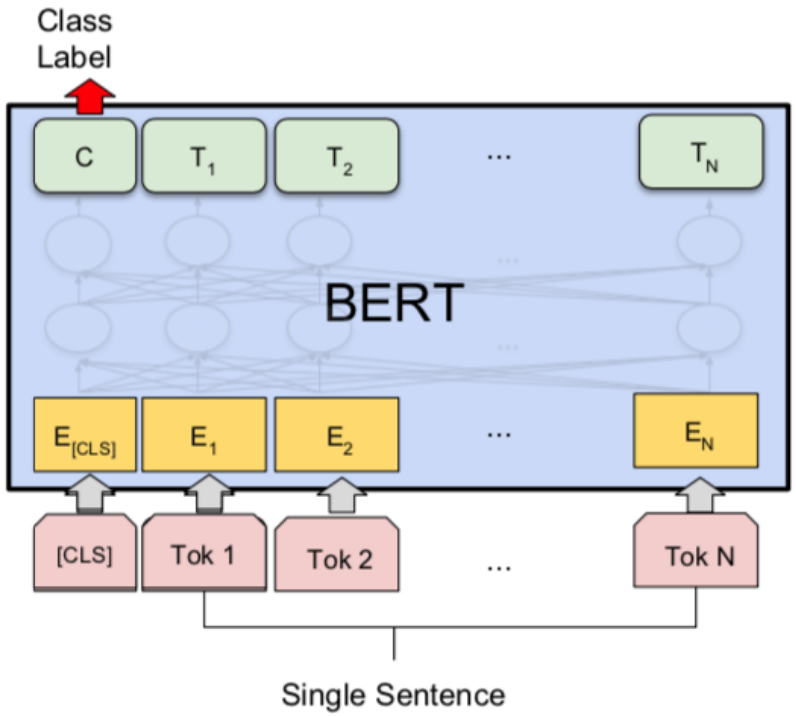
TAG	DESCRIPTION
BEGIN	The first token of a multi-token entity.
IN	An inner token of a multi-token entity.
LAST	The final token of a multi-token entity.
UNIT	A single-token entity.
OUT	A non-entity token.

Source: https://spacy.io/api/annotation#named-entities

BERT Sequence-level tasks



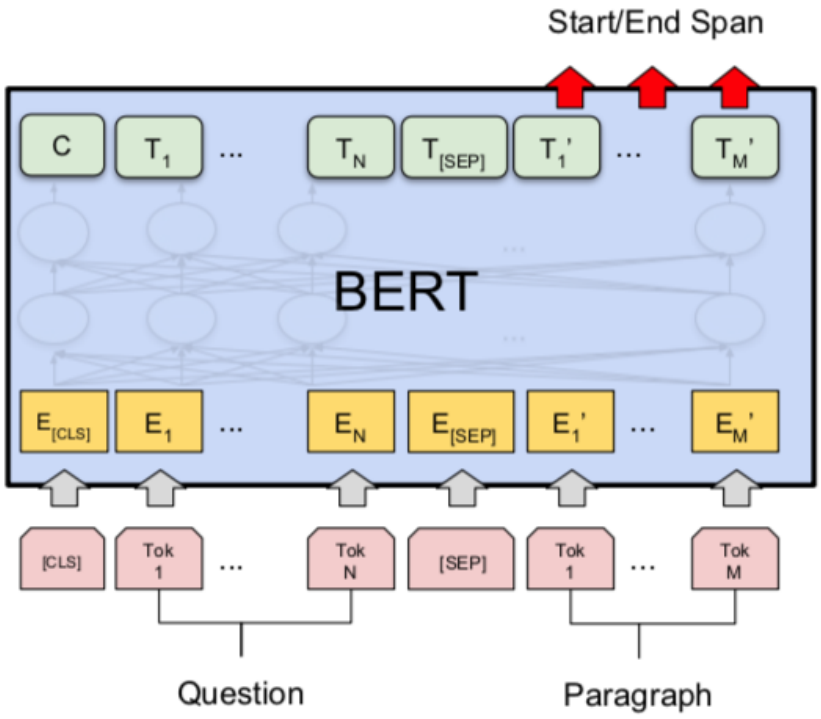
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



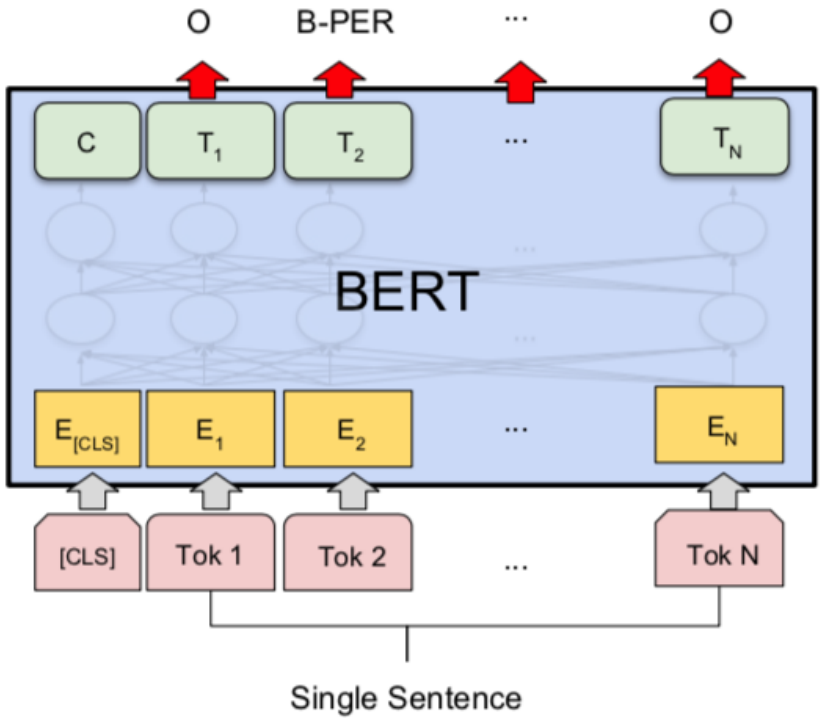
(b) Single Sentence Classification Tasks: SST-2, CoLA

Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805

BERT Token-level tasks



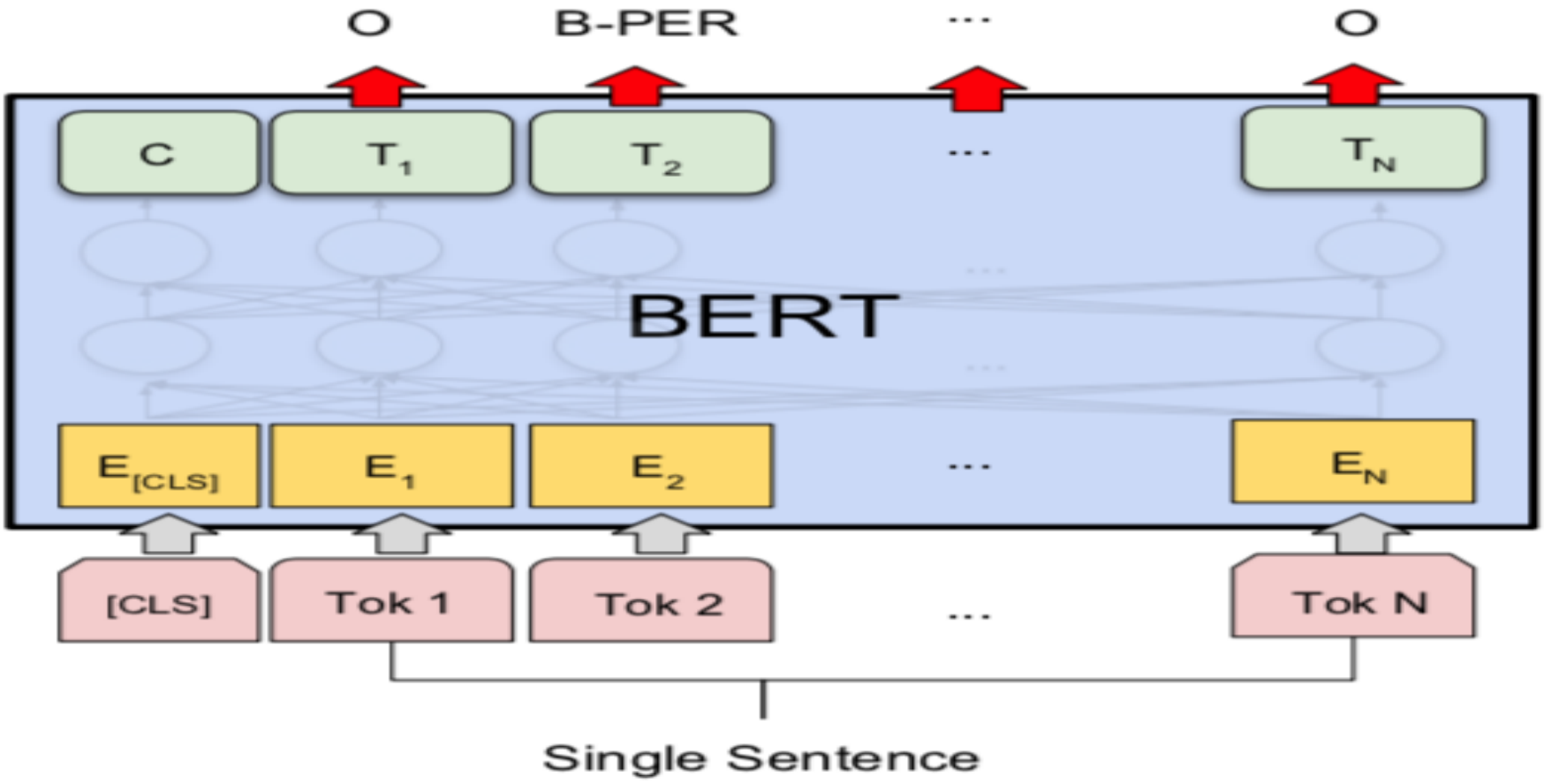
(c) Question Answering Tasks: SQuAD v1.1



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

Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805

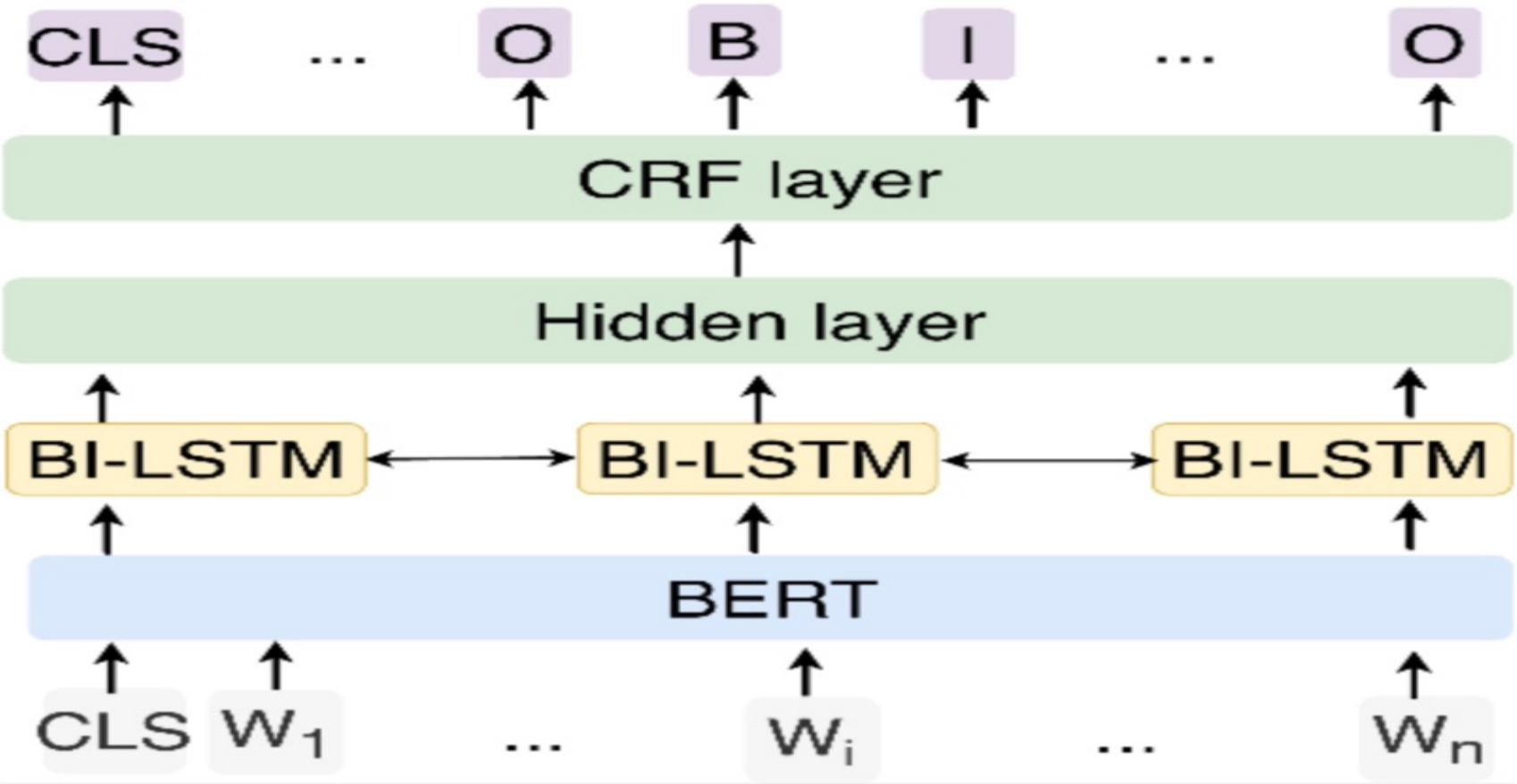
NER: Single Sentence Tagging



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

Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805

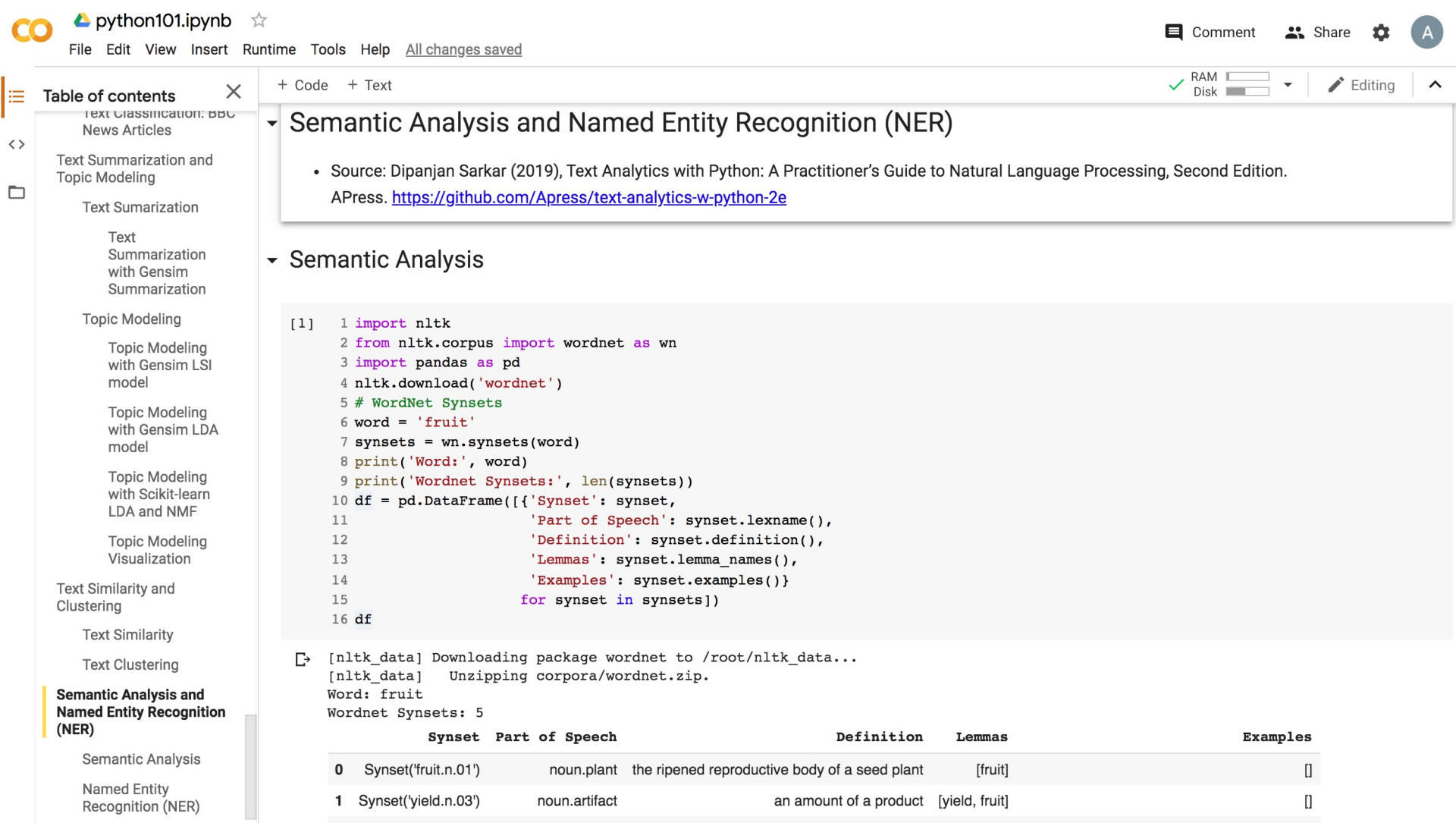
NER: Fine-tuning BERT with Bi-LSTM CRF



Source: Zhang, Xiaohui, Yaoyun Zhang, Qin Zhang, Yuankai Ren, Tinglin Qiu, Jianhui Ma, and Qiang Sun. "Extracting comprehensive clinical information for breast cancer using deep learning methods." International Journal of Medical Informatics 132 (2019): 103985.

Python in Google Colab (Python101)

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



	Synset	Part of Speech	Definition	Lemmas	Examples
0	Synset('fruit.n.01')	noun.plant	the ripened reproductive body of a seed plant	[fruit]	[]
1	Synset('yield.n.03')	noun.artifact	an amount of a product	[yield, fruit]	[]

https://tinyurl.com/imtkupython101

Python in Google Colab (Python101)

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

https://tinyurl.com/imtkupython101

NLP Benchmark Datasets

Task	Dataset	Link
Machine Translation	WMT 2014 EN-DE WMT 2014 EN-FR	http://www-lium.univ-lemans.fr/~schwenk/cslm_joint_paper/
Text Summarization	CNN/DM Newsroom DUC Gigaword	https://cs.nyu.edu/~kcho/DMQA/ https://summari.es/ https://www-nlpir.nist.gov/projects/duc/data.html https://catalog.ldc.upenn.edu/LDC2012T21
Reading Comprehension Question Answering Question Generation	ARC CliCR CNN/DM NewsQA RACE SQuAD Story Cloze Test NarativeQA Quasar SearchQA	http://data.allenai.org/arc/ http://aclweb.org/anthology/N18-1140 https://cs.nyu.edu/~kcho/DMQA/ https://datasets.maluuba.com/NewsQA http://www.qizhexie.com/data/RACE_leaderboard https://rajpurkar.github.io/SQuAD-explorer/ http://aclweb.org/anthology/W17-0906.pdf https://github.com/deepmind/narrativeqa https://github.com/bdhingra/quasar https://github.com/nyu-dl/SearchQA
Semantic Parsing	AMR parsing ATIS (SQL Parsing) WikiSQL (SQL Parsing)	https://amr.isi.edu/index.html https://github.com/jkkummerfeld/text2sql-data/tree/master/data https://github.com/salesforce/WikiSQL
Sentiment Analysis	IMDB Reviews SST Yelp Reviews Subjectivity Dataset	http://ai.stanford.edu/~amaas/data/sentiment/ https://nlp.stanford.edu/sentiment/index.html https://www.yelp.com/dataset/challenge http://www.cs.cornell.edu/people/pabo/movie-review-data/
Text Classification	AG News DBpedia TREC 20 NewsGroup	http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html https://wiki.dbpedia.org/Datasets https://trec.nist.gov/data.html http://qwone.com/~jason/20Newsgroups/
Natural Language Inference	SNLI Corpus MultiNLI SciTail	https://nlp.stanford.edu/projects/snli/ https://www.nyu.edu/projects/bowman/multinli/ http://data.allenai.org/scitail/
Semantic Role Labeling	Proposition Bank OneNotes	http://propbank.github.io/ https://catalog.ldc.upenn.edu/LDC2013T19

Source: Amirsina Torfi, Rouzbeh A. Shirvani, Yaser Keneshloo, Nader Tavvaf, and Edward A. Fox (2020). "Natural Language Processing Advancements By Deep Learning: A Survey." arXiv preprint arXiv:2003.01200.

Summary

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

References

- Dipanjan Sarkar (2019),
  Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second Edition. APress. https://github.com/Apress/text-analytics-w-python-2e
- Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda (2018),
  Applied Text Analysis with Python, O'Reilly Media.
  https://www.oreilly.com/library/view/applied-text-analysis/9781491963036/
- HuggingFace (2020), Transformers Notebook,
  https://huggingface.co/transformers/notebooks.html
- The Super Duper NLP Repo, https://notebooks.quantumstat.com/
- Min-Yuh Day (2020), Python 101, https://tinyurl.com/imtkupython101