Artificial Intelligence for Text Analytics

Multilingual Named Entity Recognition (NER)

1121AITA06
MBA, IM, NTPU (M5265) (Fall 2023)
Tue 2, 3, 4 (9:10-12:00) (B3F17)

https://meet.google.com/miy-fbif-max

Min-Yuh Day, Ph.D,
Associate Professor
Institute of Information Management, National Taipei University
https://web.ntpu.edu.tw/~myday

2023-10-25

Syllabus

Week	Date	Subject/Topics
1	2023/09/13	Introduction to Artificial Intelligence for Text Analytics
2	2023/09/20	Foundations of Text Analytics: Natural Language Processing (NLP)
3	2023/09/27	Python for Natural Language Processing
4	2023/10/04	Natural Language Processing with Transformers
5	2023/10/11	Case Study on Artificial Intelligence for Text Analytics I
6	2023/10/18	Text Classification and Sentiment Analysis

Syllabus

Week	Date	Subject/Topics
7	2023/10/25	Multilingual Named Entity Recognition (NER)
8	2023/11/01	Midterm Project Report
9	2023/11/08	Text Similarity and Clustering
10	2023/11/15	Text Summarization and Topic Models
11	2023/11/22	Text Generation with Large Language Models (LLMs)
12	2023/11/29	Case Study on Artificial Intelligence for Text Analytics II

Syllabus

Week	Date	Subject/Topics
13	2023/12/06	Question Answering and Dialogue Systems
14	2023/12/13	Deep Learning, Generative AI, Transfer Learning, Zero-Shot, and Few-Shot Learning for Text Analytics
15	2023/12/20	Final Project Report I
16	2023/12/27	Final Project Report II

Multilingual Named Entity Recognition (NER)

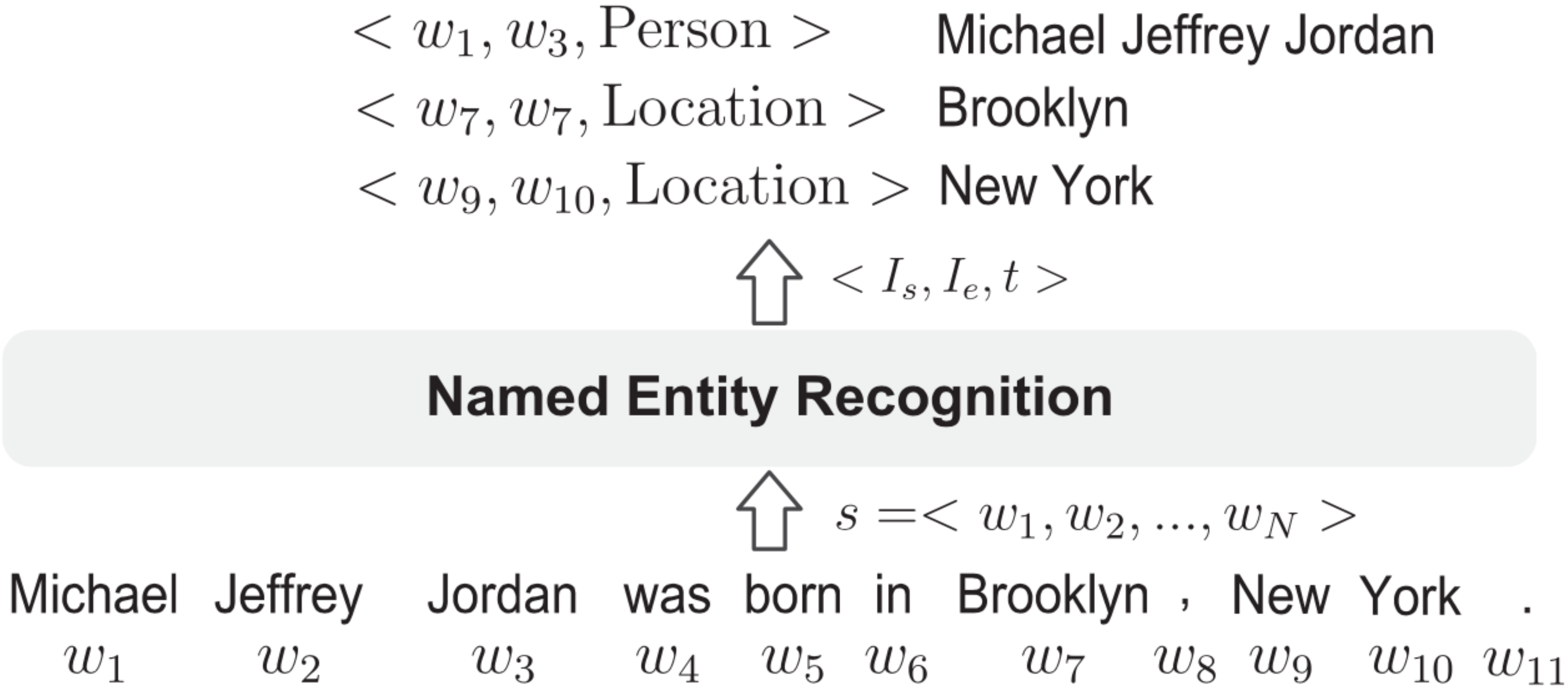
Outline

- Named Entities (NE)
  - represent real-world objects
  - people, places, organizations
  - proper names
- Named Entity Recognition (NER)
  - Entity chunking
  - Entity extraction
- Relation Extraction (RE)

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

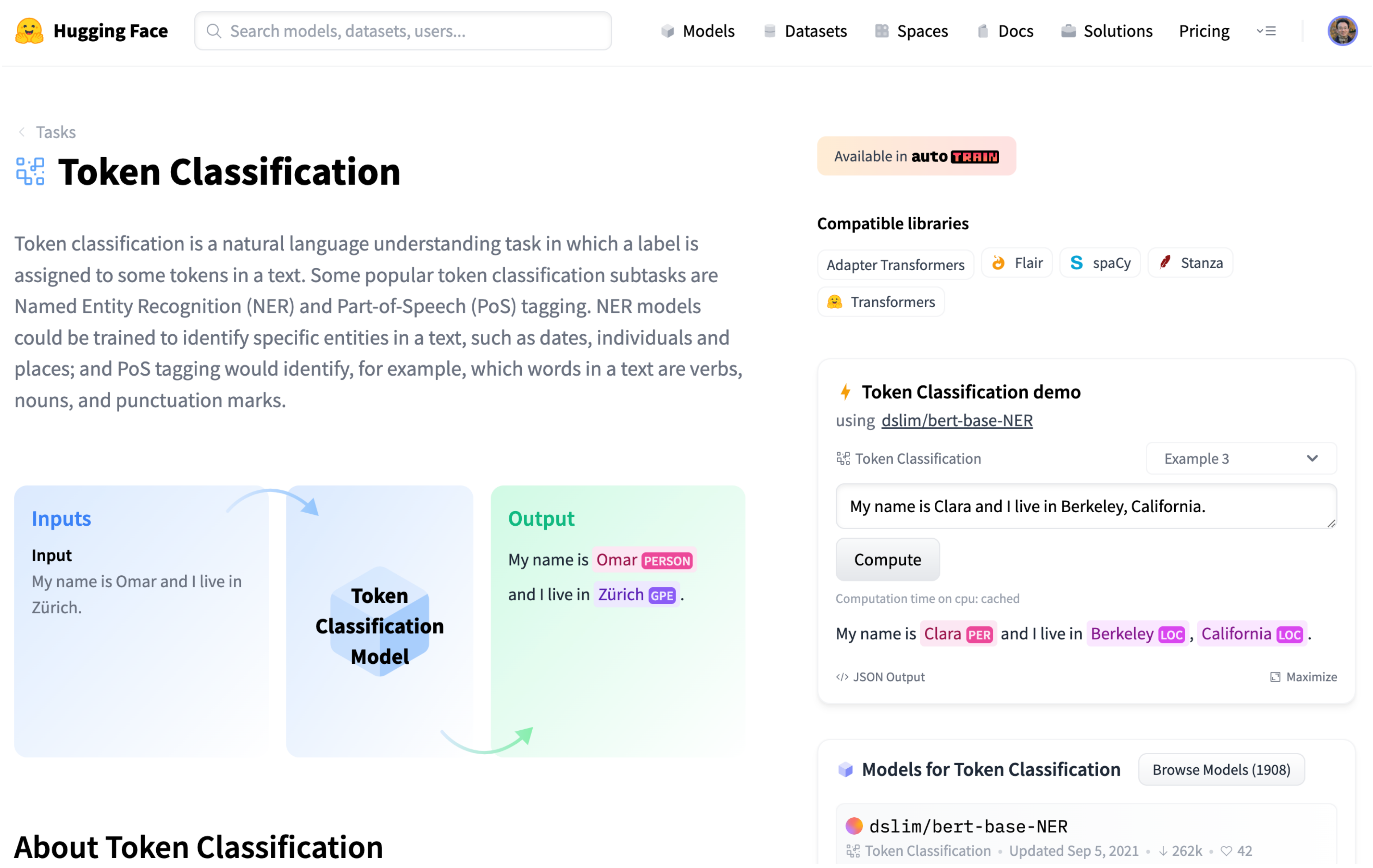
Named Entity Recognition (NER)

Michael Jeffrey Jordan was born in Brooklyn, New York.



Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

Token Classification (NER)



Hugging Face | Search models, datasets, users... | Models | Datasets | Spaces | Docs | Solutions | Pricing

Tasks

Token Classification

Token classification is a natural language understanding task in which a label is assigned to some tokens in a text. Some popular token classification subtasks are Named Entity Recognition (NER) and Part-of-Speech (PoS) tagging. NER models could be trained to identify specific entities in a text, such as dates, individuals and places; and PoS tagging would identify, for example, which words in a text are verbs, nouns, and punctuation marks.

About Token Classification

Available in autoTRAIN

Compatible libraries

Adapter Transformers | Flair | spaCy | Stanza | Transformers

Token Classification demo

using dslim/bert-base-NER

Token Classification | Example 3

My name is Clara and I live in Berkeley, California.

Compute

Computation time on cpu: cached

My name is Clara PER and I live in Berkeley LOC, California LOC.

JSON Output | Maximize

Models for Token Classification | Browse Models (1908)

dslim/bert-base-NER

Token Classification • Updated Sep 5, 2021 • 262k • 42

https://huggingface.co/tasks/token-classification

Named Entity Recognition (NER)

```python
from transformers import pipeline
import pandas as pd
classifier = pipeline("ner")
text = "My name is Michael and I live in Berkeley, California."
outputs = classifier(text)
pd.DataFrame(outputs)
```

	entity	score	index	word	start	end
0	I-PER	0.998874	4	Michael	11	18
1	I-LOC	0.997050	9	Berkeley	33	41
2	I-LOC	0.999170	11	California	43	53

https://tinyurl.com/aintpupython101

Multilingual Named Entity Recognition (NER)

```
#!pip install transformers
from transformers import pipeline
import pandas as pd
nlp = pipeline('ner', model="Babelscape/wikineural-multilingual-ner")
outputs = nlp("My name is Alan and I live in Taipei.")
pd.DataFrame(outputs)
```

	entity	score	index	word	start	end
0	B-PER	0.860065	4	Alan	11	15
1	B-LOC	0.999816	9	Taipei	30	36

https://tinyurl.com/aintpupython101

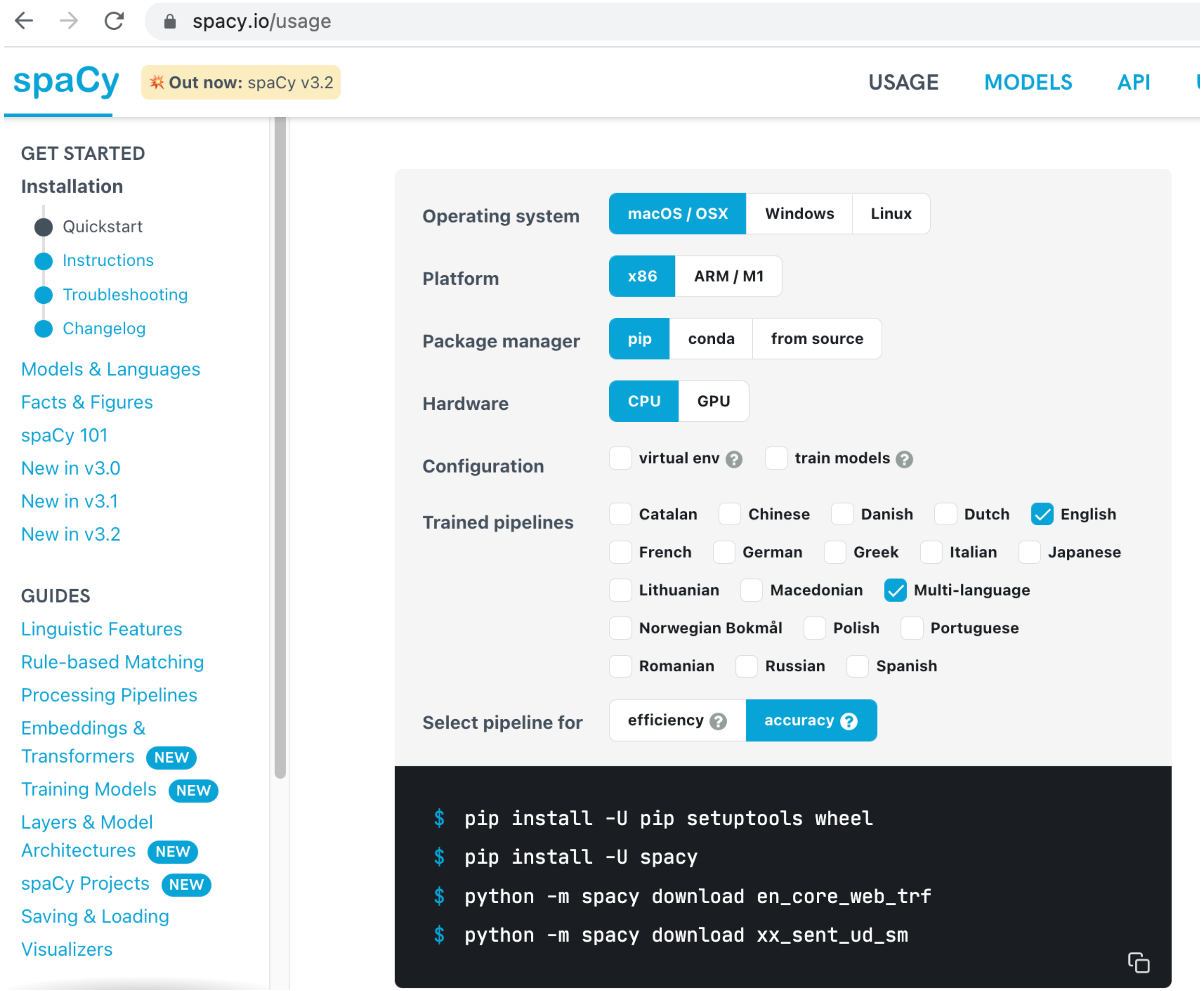
Multilingual Named Entity Recognition (NER)

```
#!pip install transformers
from transformers import pipeline
import pandas as pd
nlp = pipeline('ner', model="Babelscape/wikineural-multilingual-ner")
outputs = nlp("My name is Alan and I live in Taipei. 他是王小明，他住在台南")
pd.DataFrame(outputs)
```

	entity	score	index	word	start	end
0	B-PER	0.912095	4	Alan	11	15
1	B-LOC	0.999747	9	Taipei	30	36
2	B-PER	0.994766	13	王	40	41
3	I-PER	0.992879	14	小	41	42
4	I-PER	0.982183	15	明	42	43
5	B-LOC	0.999288	20	台	47	48
6	I-LOC	0.993408	21	南	48	49

https://tinyurl.com/aintpupython101

spaCy



https://spacy.io/usage

NER: OntoNotes 5 Named Entities (18)

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).

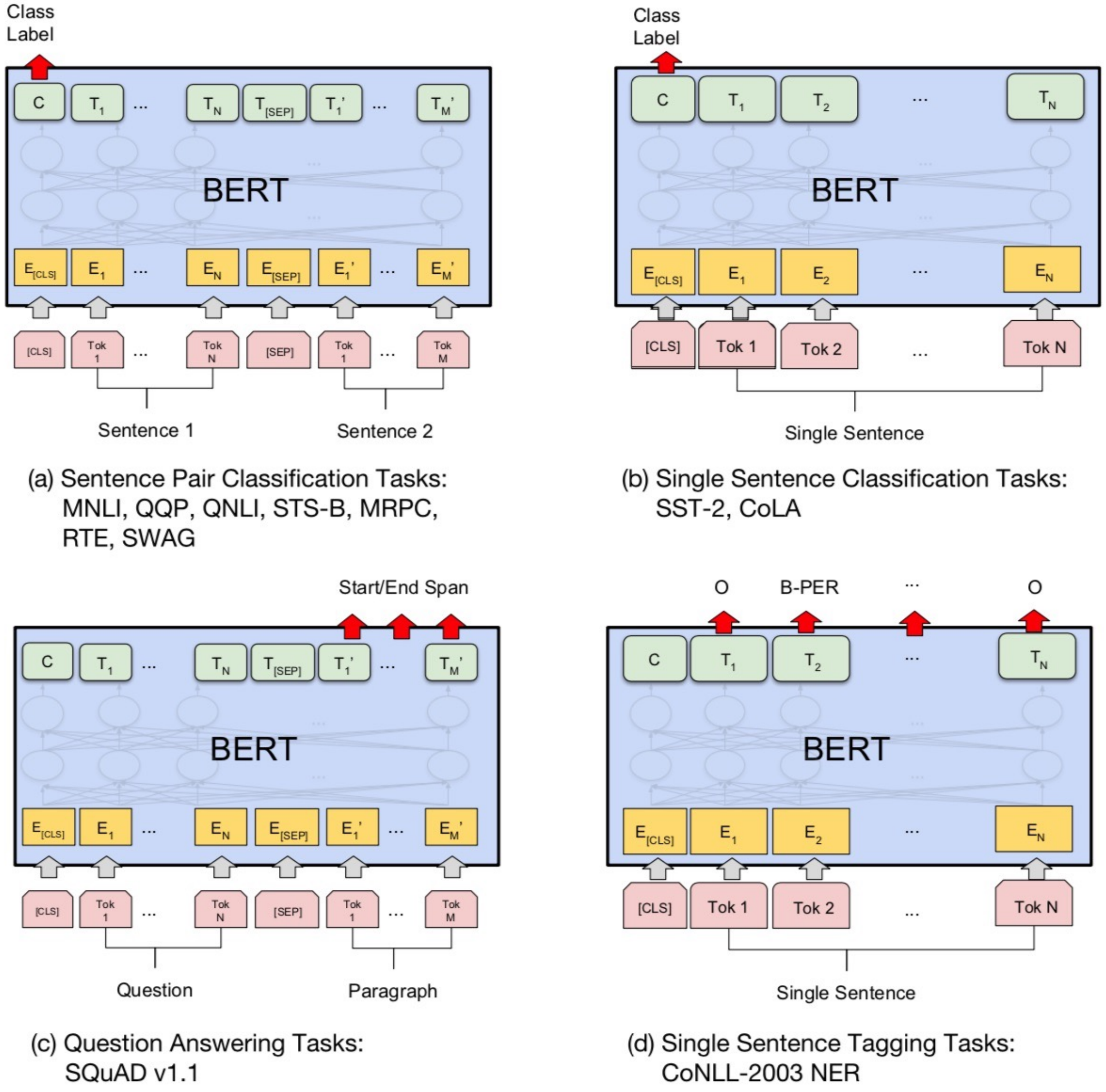
Source: https://spacy.io/api/annotation#named-entities

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

Fine-tuning BERT on NLP Tasks



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

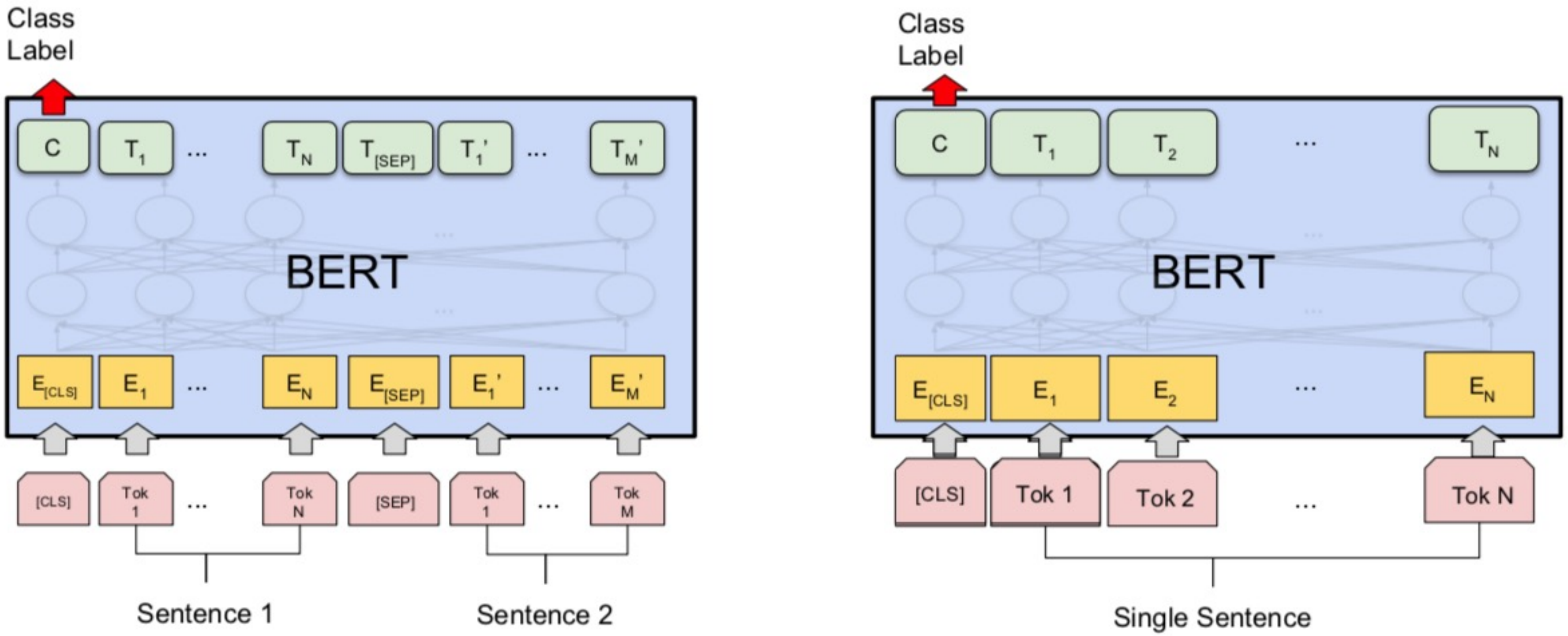
(b) Single Sentence Classification Tasks: SST-2, CoLA

(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

BERT Sequence-level tasks

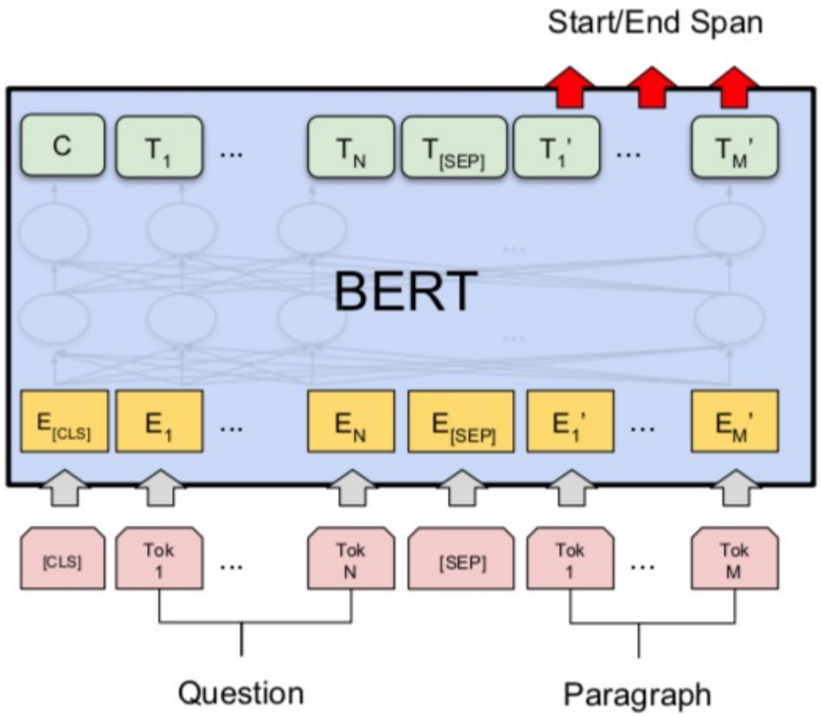


(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

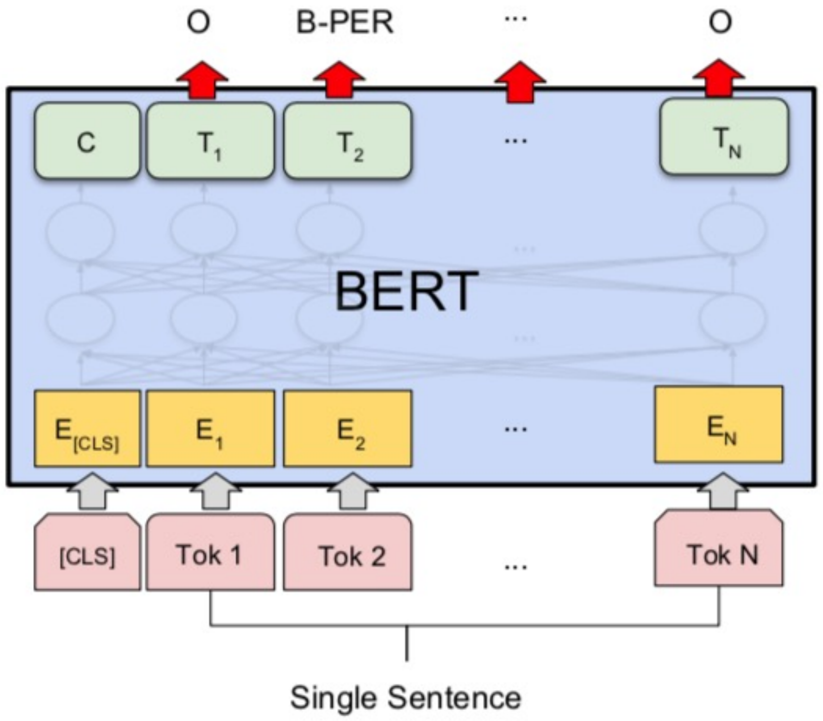
(b) Single Sentence Classification Tasks: SST-2, CoLA

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



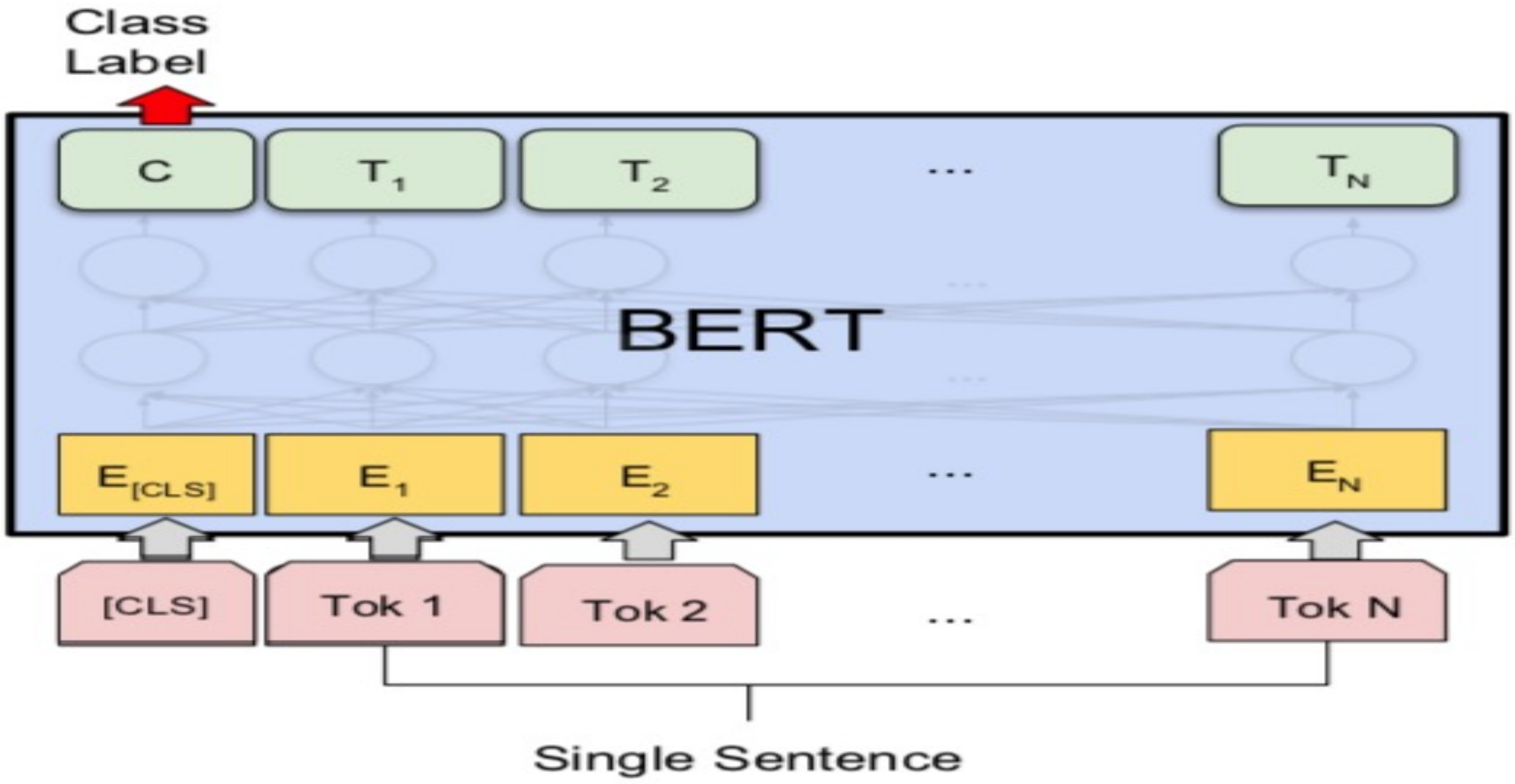
(c) Question Answering Tasks:
SQuAD v1.1



(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

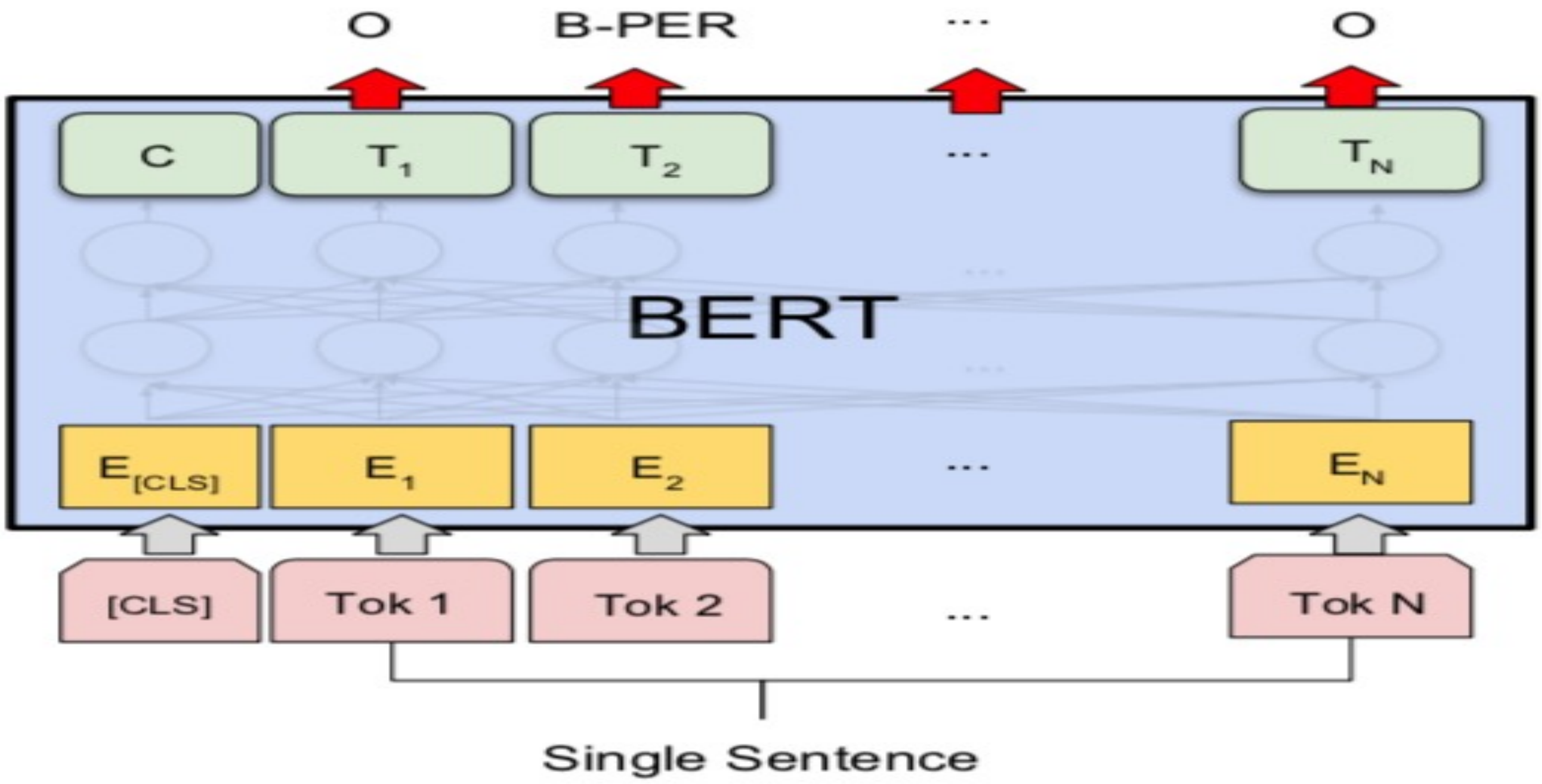
Sentiment Analysis: Single Sentence Classification



(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

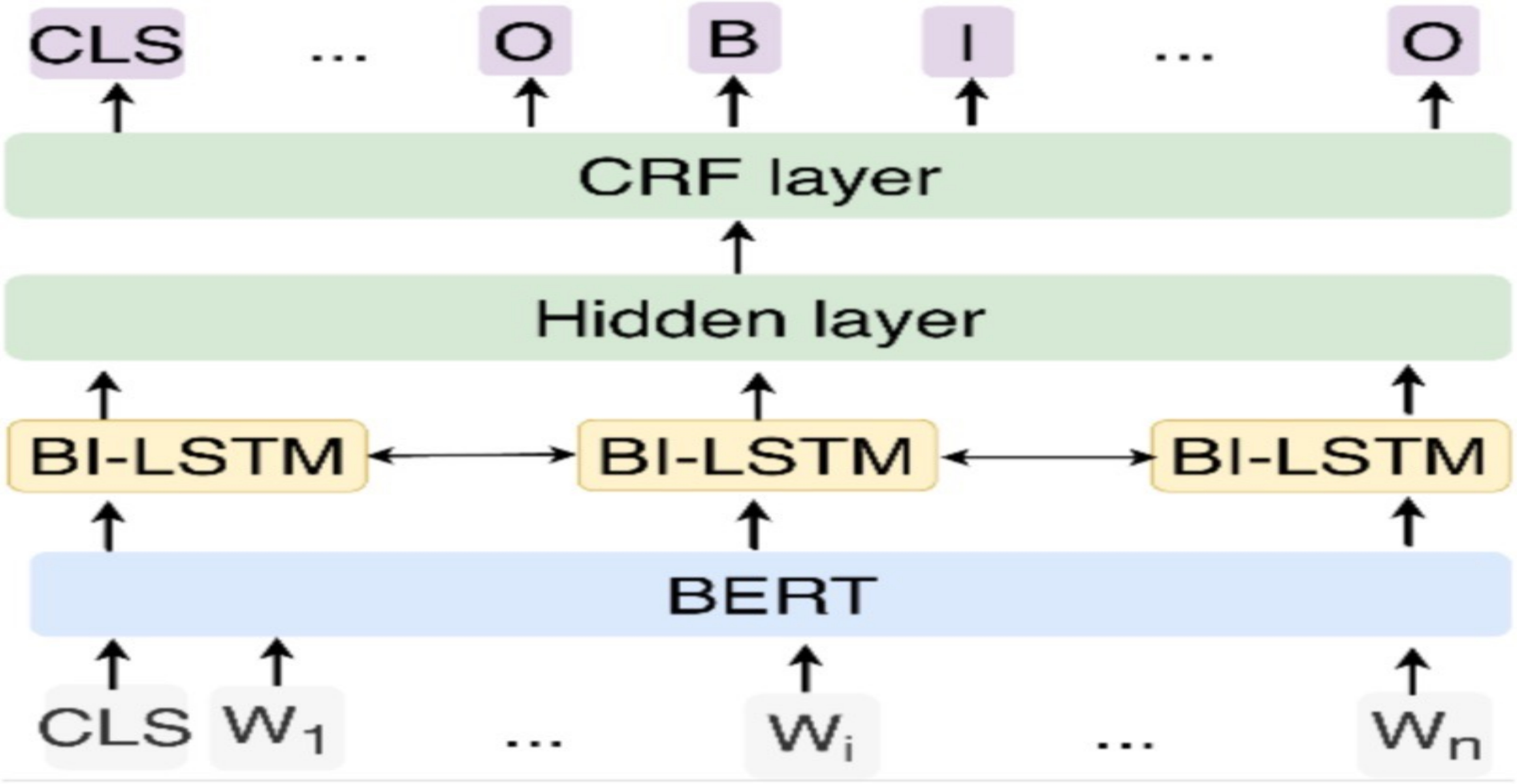
NER: Single Sentence Tagging



(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.

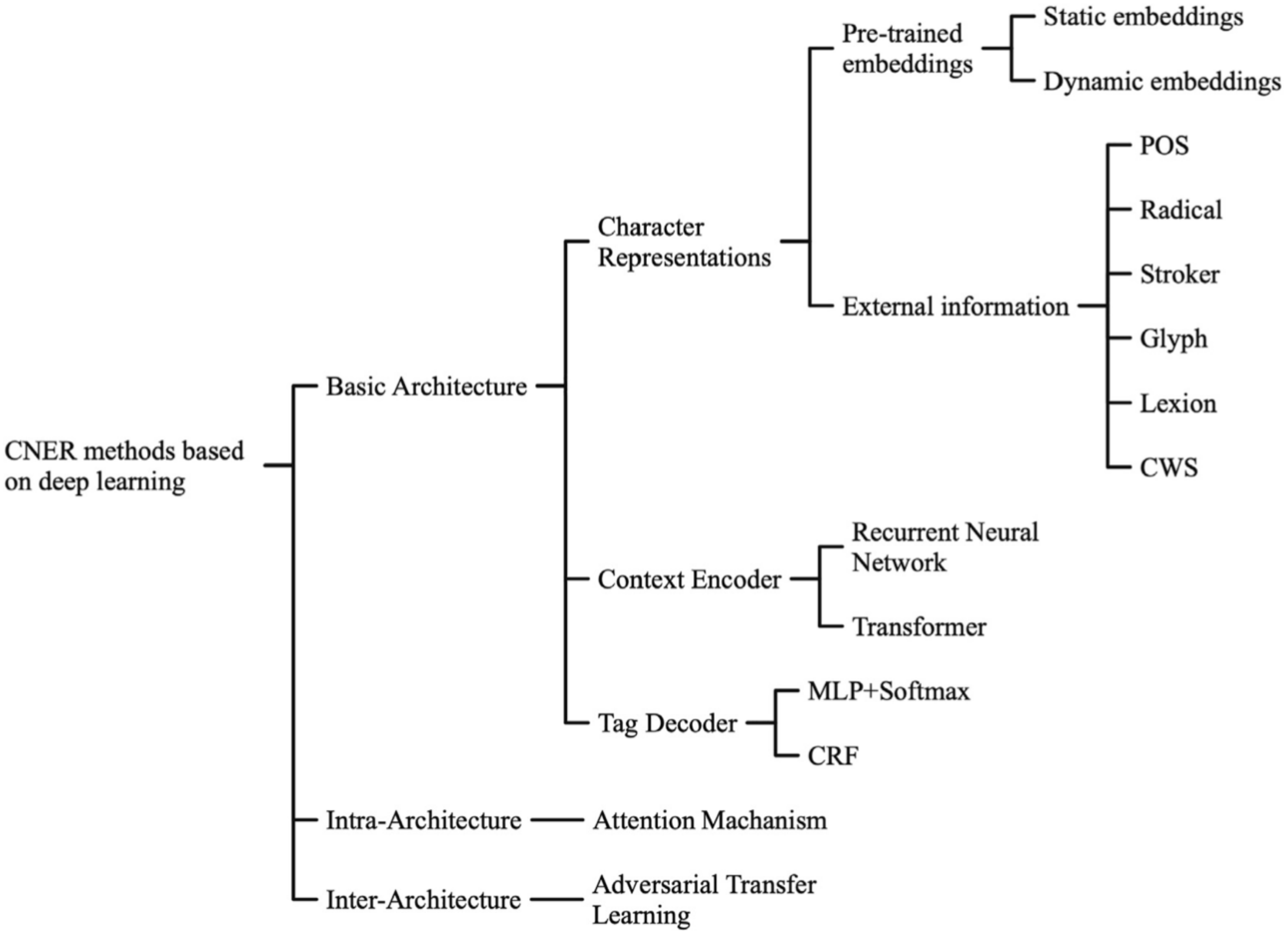
Named Entity Recognition (NER)

Statistical-based methods and Deep learning-based methods

	Statistical-based methods	Deep learning-based methods
Character Representations	Handcrafted features (orthographic, prefixes, suffixes, etc.)	Distributed representations (Word2vec, RNN, ELMo, BERT, etc.)
Machine learning models	Statistical-based models (HMM, ME, CRF, SVM, etc.)	Encoder (LSTM, GRU, Transformer, etc.) Decoder (CRF, Transformer, etc.)

Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

The taxonomy of CNER methods based on deep learning



Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

List of Annotated Datasets for English NER

Corpus	Year	Text Source	#Tags	URL
MUC-6	1995	Wall Street Journal	7	https://catalog.ldc.upenn.edu/LDC2003T13
MUC-6 Plus	1995	Additional news to MUC-6	7	https://catalog.ldc.upenn.edu/LDC96T10
MUC-7	1997	New York Times news	7	https://catalog.ldc.upenn.edu/LDC2001T02
CoNLL03	2003	Reuters news	4	https://www.clips.uantwerpen.be/conll2003/ner/
ACE	2000 - 2008	Transcripts, news	7	https://www.ldc.upenn.edu/collaborations/past-projects/ace
OntoNotes	2007 - 2012	Magazine, news, web, etc.	18	https://catalog.ldc.upenn.edu/LDC2013T19
W-NUT	2015 - 2018	User-generated text	6/10	http://noisy-text.github.io
BBN	2005	Wall Street Journal	64	https://catalog.ldc.upenn.edu/LDC2005T33
WikiGold	2009	Wikipedia	4	https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500
WiNER	2012	Wikipedia	4	http://rali.iro.umontreal.ca/rali/en/winer-wikipedia-for-ner
WikiFiger	2012	Wikipedia	112	https://github.com/xiaoling/figer
HYENA	2012	Wikipedia	505	https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/hyena/
N^3	2014	News	3	http://aksw.org/Projects/N3NERNEDNIF.html
Gillick	2016	Magazine, news, web, etc.	89	https://arxiv.org/e-print/1412.1820v2
FG-NER	2018	Various	200	https://fgner.alt.ai/
NNE	2019	Newswire	114	https://github.com/nickyringland/nested_named_entities
GENIA	2004	Biology and clinical text	36	http://www.geniaproject.org/home
GENETAG	2005	MEDLINE	2	https://sourceforge.net/projects/bioc/files/
FSU-PRGE	2010	PubMed and MEDLINE	5	https://julielab.de/Resources/FSU_PRGE.html
NCBI-Disease	2014	PubMed	1	https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/
BC5CDR	2015	PubMed	3	http://bioc.sourceforge.net/
DFKI	2018	Business news and social media	7	https://dfki-lt-re-group.bitbucket.io/product-corpus/

"#Tags" refers to the number of entity types.

Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

Named Entity Recognition (NER)

NER System	URL
StanfordCoreNLP	https://stanfordnlp.github.io/CoreNLP/
OSU Twitter NLP	https://github.com/aritter/twitter_nlp
Illinois NLP	http://cogcomp.org/page/software/
NeuroNER	http://neuroner.com/
NERsuite	http://nersuite.nlplab.org/
Polyglot	https://polyglot.readthedocs.io
Gimli	http://bioinformatics.ua.pt/gimli

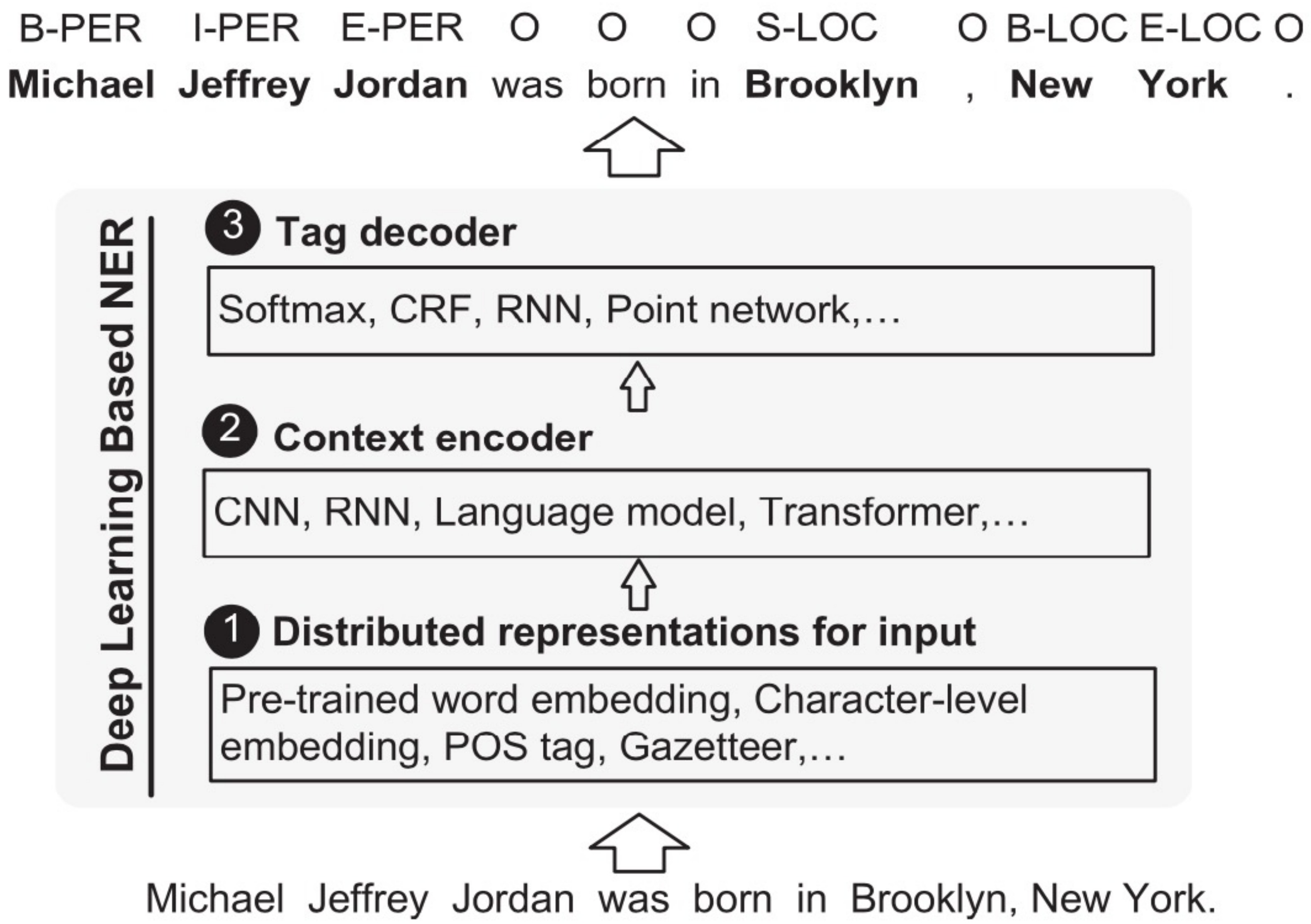
Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

Named Entity Recognition (NER)

NER System	URL
spaCy	https://spacy.io/api/entityrecognizer
NLTK	https://www.nltk.org
OpenNLP	https://opennlp.apache.org/
LingPipe	http://alias-i.com/lingpipe-3.9.3/
AllenNLP	https://demo.allennlp.org/
IBM Watson	https://natural-language-understanding-demo.ng.bluemix.net/
FG-NER	https://fgner.alt.ai/extractor/
Intellexer	http://demo.intellexer.com/
Repustate	https://repustate.com/named-entity-recognition-api-demo/
AYLIEN	https://developer.aylien.com/text-api-demo
Dandelion API	https://dandelion.eu/semantic-text/entity-extraction-demo/
displaCy	https://explosion.ai/demos/displacy-ent
ParallelDots	https://www.paralleldots.com/named-entity-recognition
TextRazor	https://www.textrazor.com/named_entity_recognition

Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

Named Entity Recognition (NER)



Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

Deep Learning for Named Entity Recognition (NER)

- Distributed Representations for Input
  - Hybrid Representation
- Context Encoder Architectures
  - Deep Transformer
- Tag Decoder Architectures
  - Conditional Random Fields (CRF)

Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

Deep Learning for Named Entity Recognition (NER)

- Distributed Representations for Input
  - Word-Level Representation
  - Character-Level Representation
  - Hybrid Representation

Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

Deep Learning for Named Entity Recognition (NER)

- Context Encoder Architectures
  - Convolutional Neural Networks
  - Recurrent Neural Networks
  - Recursive Neural Networks
  - Neural Language Models
  - Deep Transformer

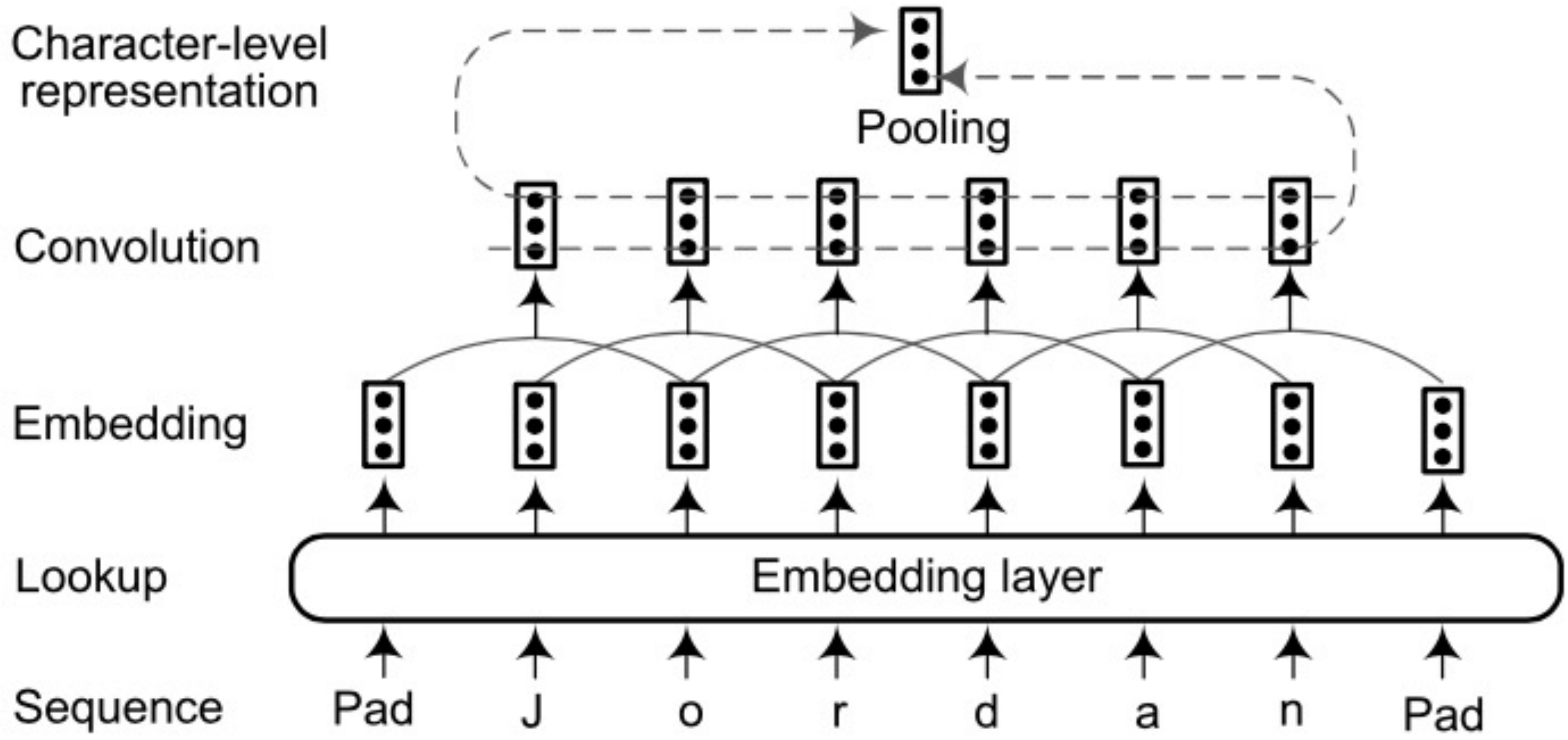
Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

Deep Learning for Named Entity Recognition (NER)

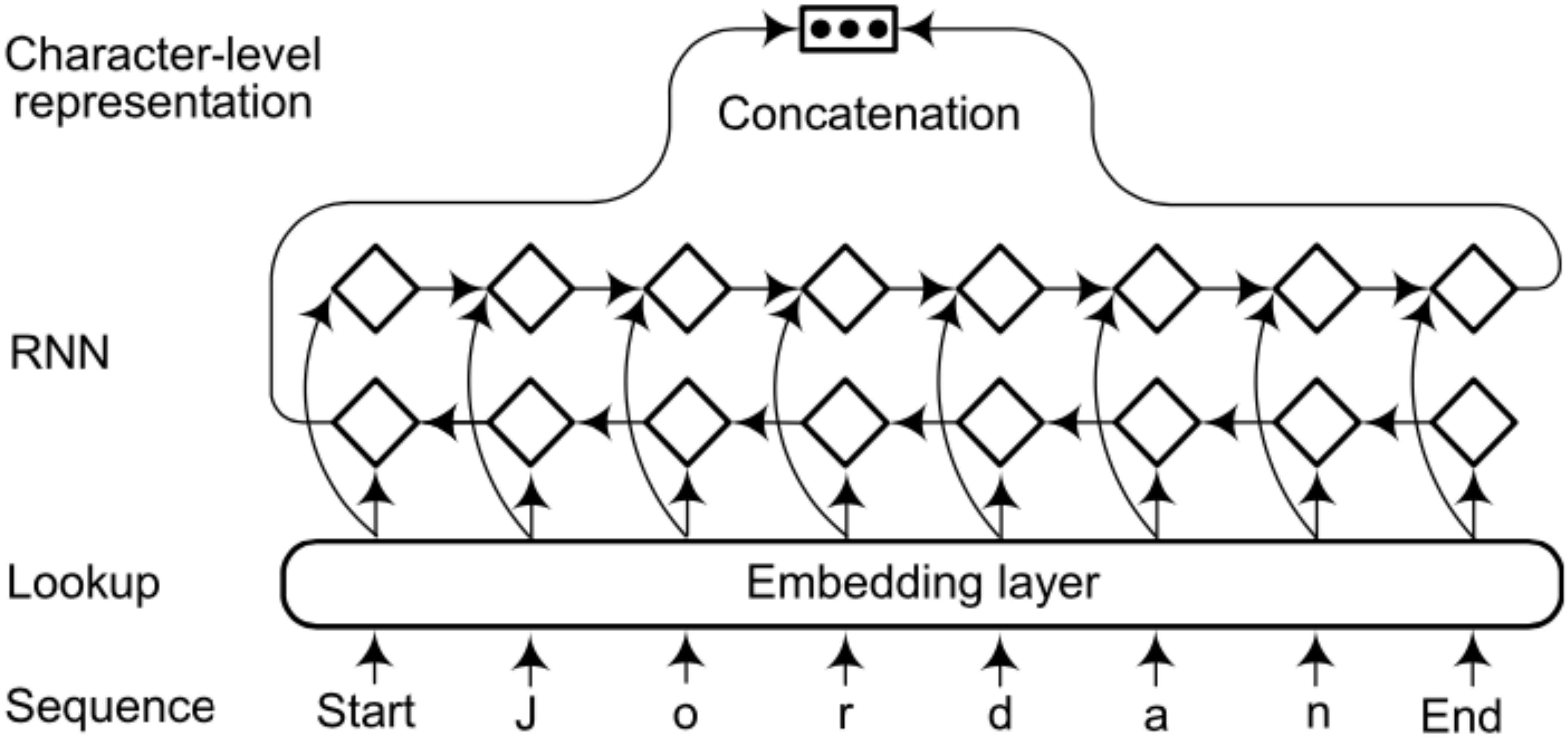
- Tag Decoder Architectures
  - Multi-Layer Perceptron + Softmax
  - Conditional Random Fields (CRF)
  - Recurrent Neural Networks
  - Pointer Networks

Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

CNN-based character-level representation

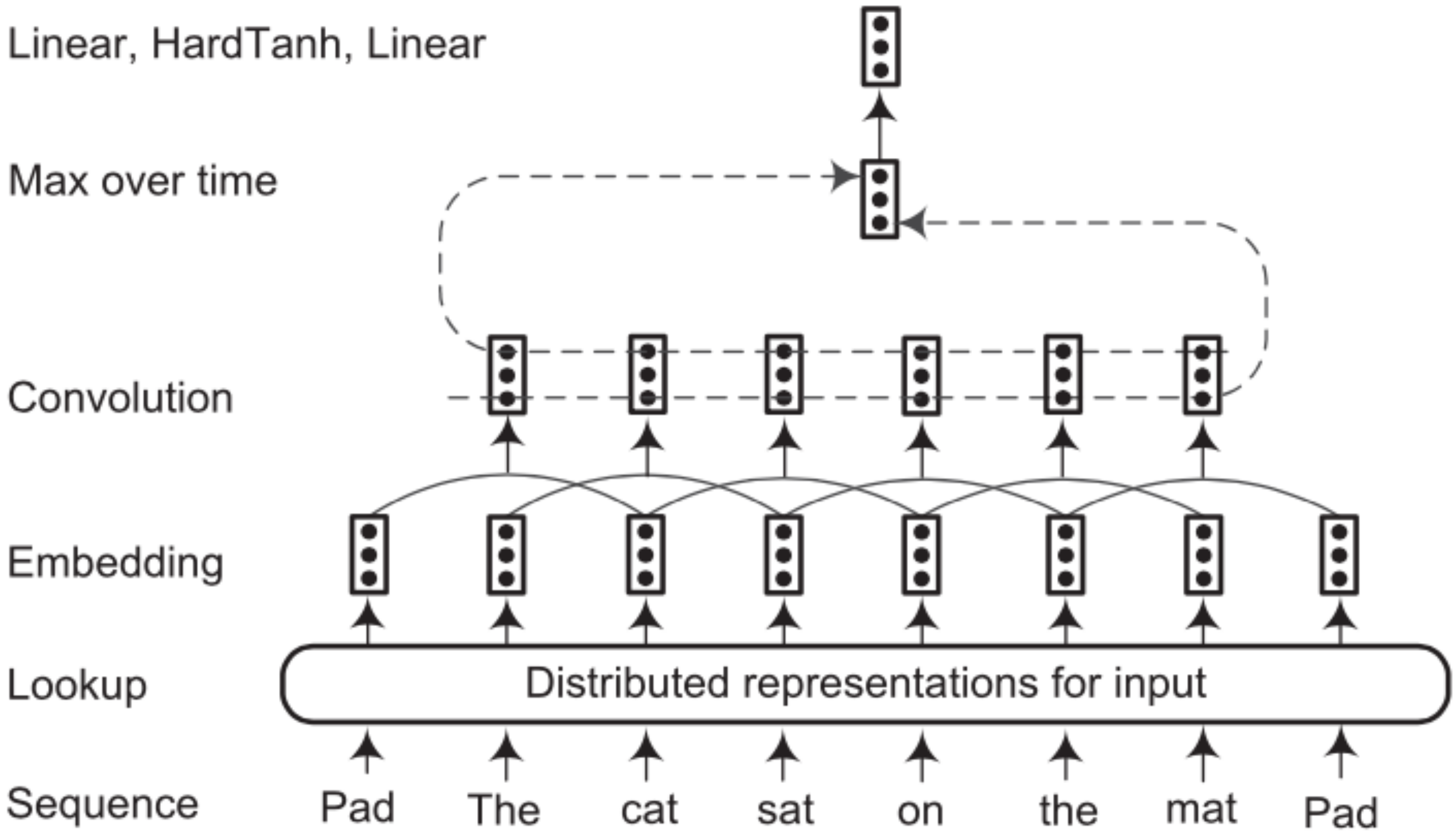


RNN-based character-level representation



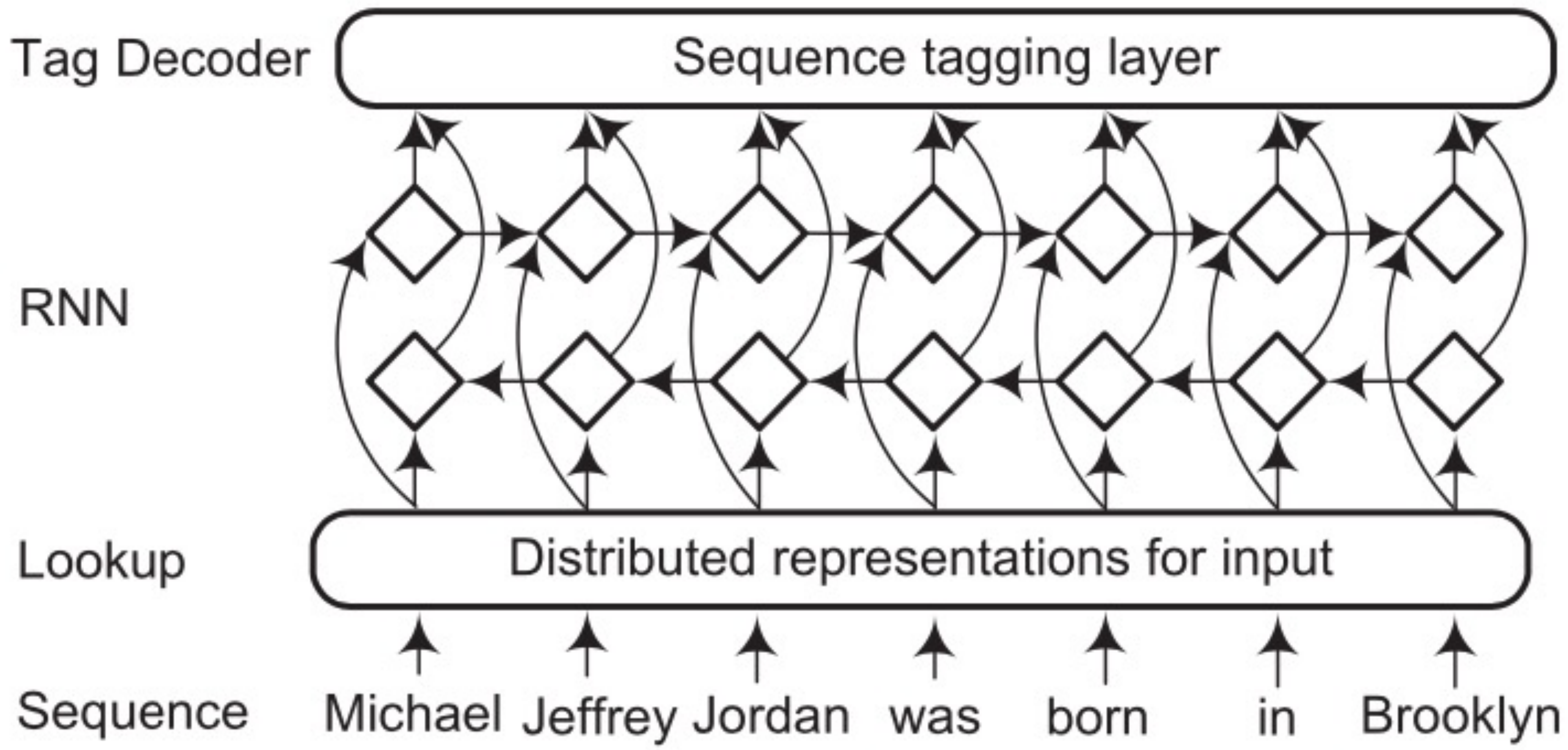
Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

Sentence approach network based on CNN



Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

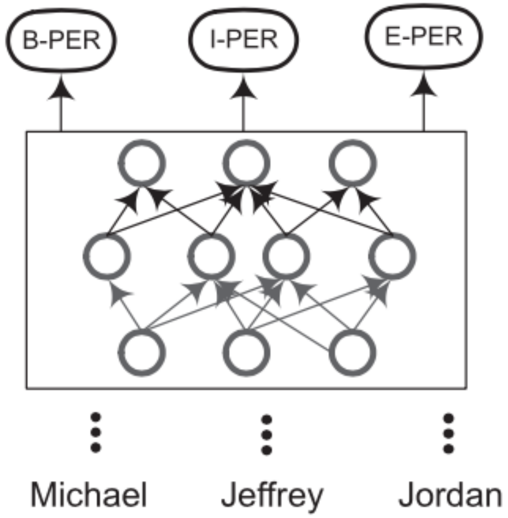
The architecture of RNN-based context encoder



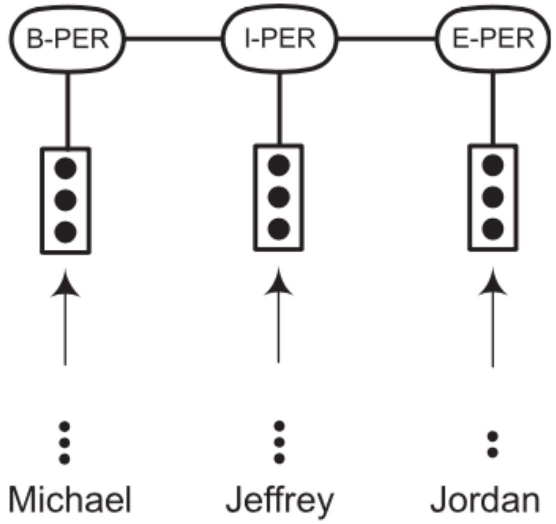
Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

Named Entity Recognition (NER)

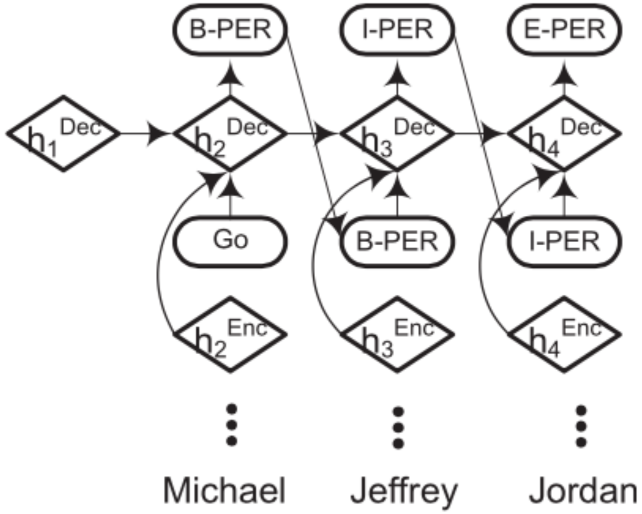
Four tag decoders: MLP+Softmax, CRF, RNN, and Pointer network



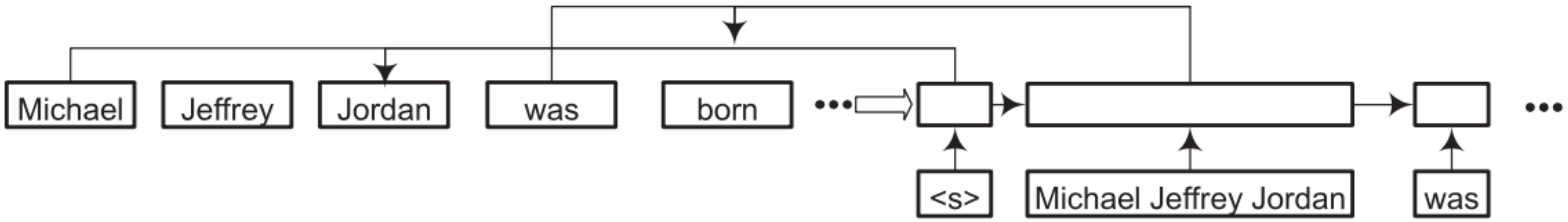
(a) MLP+Softmax



(b) CRF



(c) RNN



(d) Pointer Network

Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

Named Entity Recognition (NER)

Work	Input representation			Context encoder	Tag decoder	Performance (F-score)
	Character	Word	Hybrid			
[93]	-	Trained on PubMed	POS	CNN	CRF	GENIA: 71.01%
[88]	-	Trained on Gigaword	-	GRU	GRU	ACE 2005: 80.00%
[94]	-	Random	-	LSTM	Pointer Network	ATIS: 96.86%
[89]	-	Trained on NYT	-	LSTM	LSTM	NYT: 49.50%
[90]	-	SENNA	Word shape	ID-CNN	CRF	CoNLL03: 90.65%; OntoNotes5.0: 86.84%
[95]	-	Google word2vec	-	LSTM	LSTM	CoNLL04: 75.0%
[99]	LSTM	-	-	LSTM	CRF	CoNLL03: 84.52%
[96]	CNN	GloVe	-	LSTM	CRF	CoNLL03: 91.21%
[104]	LSTM	Google word2vec	-	LSTM	CRF	CoNLL03: 84.09%
[19]	LSTM	SENNA	-	LSTM	CRF	CoNLL03: 90.94%
[105]	GRU	SENNA	-	GRU	CRF	CoNLL03: 90.94%
[97]	CNN	GloVe	POS	BRNN	Softmax	OntoNotes5.0: 87.21%
[106]	LSTM-LM	-	-	LSTM	CRF	CoNLL03: 93.09%; OntoNotes5.0: 89.71%
[102]	CNN-LSTM-LM	-	-	LSTM	CRF	CoNLL03: 92.22%
[17]	-	Random	POS	CNN	CRF	CoNLL03: 89.86%
[18]	-	SENNA	Spelling, n-gram, gazetteer	LSTM	CRF	CoNLL03: 90.10%
[20]	CNN	SENNA	capitalization, lexicons	LSTM	CRF	CoNLL03: 91.62%; OntoNotes5.0: 86.34%
[115]	-	-	FOFE	MLP	CRF	CoNLL03: 91.17%
[100]	LSTM	GloVe	-	LSTM	CRF	CoNLL03: 91.07%
[112]	LSTM	GloVe	Syntactic	LSTM	CRF	W-NUT17: 40.42%
[101]	CNN	SENNA	-	LSTM	Reranker	CoNLL03: 91.62%
[113]	CNN	Twitter Word2vec	POS	LSTM	CRF	W-NUT17: 41.86%
[114]	LSTM	GloVe	POS, topics	LSTM	CRF	W-NUT17: 41.81%
[117]	LSTM	GloVe	Images	LSTM	CRF	SnapCaptions: 52.4%
[108]	LSTM	SSKIP	Lexical	LSTM	CRF	CoNLL03: 91.73%; OntoNotes5.0: 87.95%

Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

Named Entity Recognition (NER)

Work	Input representation			Context encoder	Tag decoder	Performance (F-score)
	Character	Word	Hybrid			
[118]	-	WordPiece	Segment, position	Transformer	Softmax	CoNLL03: 92.8%
[120]	LSTM	SENNA	-	LSTM	Softmax	CoNLL03: 91.48%
[123]	LSTM	Google Word2vec	-	LSTM	CRF	CoNLL03: 86.26%
[21]	GRU	SENNA	LM	GRU	CRF	CoNLL03: 91.93%
[125]	LSTM	GloVe	-	LSTM	CRF	CoNLL03: 91.71%
[141]	-	SENNA	POS, gazetteers	CNN	Semi-CRF	CoNLL03: 90.87%
[142]	LSTM	GloVe	-	LSTM	Semi-CRF	CoNLL03: 91.38%
[87]	CNN	Trained on Gigaword	-	LSTM	LSTM	CoNLL03: 90.69%; OntoNotes5.0: 86.15%
[109]	-	GloVe	ELMo, dependency	LSTM	CRF	CoNLL03: 92.4%; OntoNotes5.0: 89.88%
[107]	CNN	GloVe	ELMo, gazetteers	LSTM	Semi-CRF	CoNLL03: 92.75%; OntoNotes5.0: 89.94%
[132]	LSTM	GloVe	ELMo, POS	LSTM	Softmax	CoNLL03: 92.28%
[136]	-	-	BERT	-	Softmax	CoNLL03: 93.04%; OntoNotes5.0: 91.11%
[137]	-	-	BERT	-	Softmax +Dice Loss	CoNLL03: 93.33%; **OntoNotes5.0: 92.07%**
[133]	LSTM	GloVe	BERT, document-level embeddings	LSTM	CRF	CoNLL03: 93.37%; OntoNotes5.0: 90.3%
[134]	CNN	GloVe	BERT, global embeddings	GRU	GRU	CoNLL03: 93.47%
[131]	CNN	-	Cloze-style LM embeddings	LSTM	CRF	**CoNLL03: 93.5%**
[135]	-	GloVe	Plooled contextual embeddings	RNN	CRF	CoNLL03: 93.47%

Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

Applied Deep Learning for Named Entity Recognition (NER)

- Deep Multi-Task Learning for NER
- Deep Transfer Learning for NER
- Deep Active Learning for NER
- Deep Reinforcement Learning for NER
- Deep Adversarial Learning for NER
- Neural Attention for NER

Source: Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.

Named Entity Recognition (NER)

Message Understanding Conference (MUC) Corpus

Year	Conf.	Language	Source Type	Data Sources	Task
1987	MUC1	English	Military reports	Fleet Operations	Open ended (no pre-defined template)
1989	MUC2	English	Military reports	Fleet Operations	IE in form of pre-provided template
1991	MUC3	English	Reports from News	Acts of terrorism in Latin America	IE in form of pre-provided template
1992	MUC4	English	Reports from News	Acts of terrorism in Latin America	IE in form of pre-provided template
1993	MUC5	English, Japanese	Reports from News	Corporate Joint Ventures, Microelectronic production	IE in form of pre-provided template
1995	MUC6	English	Reports from News	Negotiation of Labor Disputes and Corporate Management Succession	NER, Coreference Resolution, Description of NEs and scenarios
1997	MUC7	English	Reports from News	Reports on various aerial crashes, launch report of various missiles and rockets	NER, Coreference Resolution, Description of NEs and scenarios

Source: Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik (2022). "Named entity recognition and relation extraction: State-of-the-art." ACM Computing Surveys (CSUR) 54, no. 1 (2022): 1-39.

Named Entity Recognition (NER)
Automatic Content Extraction (ACE) corpus

Corpus	Tasks	Language	Data Source
ACE 2002	EDT, RDC	English	Newswire
ACE 2003	EDT, RDC	English	Newswire, Broadcast
	EDT	Arabic	
ACE 2004	EDT, RDC, LNK	English, Arabic, Chinese	Newswire, Broadcast
ACE 2005	EDT, EDC, RDC, LNK, Time-Stamping	English, Chinese	Newswire, Newsgroups, Weblogs Broadcast
	EDT, EDC, RDC, LNK	Arabic	
ACE 2007	EDT, EDC, RDC, LNK	Arabic, Spanish	Newswire, Weblogs

Source: Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik (2022). "Named entity recognition and relation extraction: State-of-the-art." ACM Computing Surveys (CSUR) 54, no. 1 (2022): 1-39.

Named Entity Recognition (NER)

Conference on Computational Natural Language Learning (CoNLL) Corpus

Dataset Name	Year	Language	Source Type	Data Source
CoNLL'02	2002	Dutch	Newswire Articles	Belgian newspaper "De Morgen"
		Spanish	Newswire Articles	Spanish EFE News Agency
CoNLL'03	2003	English	Newswire Articles	Reuters Corpus
		German	Newswire Articles	Frankfurter Rundschau

Source: Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik (2022). "Named entity recognition and relation extraction: State-of-the-art." ACM Computing Surveys (CSUR) 54, no. 1 (2022): 1-39.

Named Entity Recognition (NER) OntoNotes

Dataset Name	Year	Source Type	Language	Data Source
OntoNotes 1.0	2007	Newswire Articles	English	Wall Street Journal
	2007	Newswire Articles	Mandarin Chinese	Xinhua News Agency and Sinorama Magazine
OntoNotes 2.0	2008	Broadcast News	English	VoA, Public Radio International, NBC, CNN and ABC
	2008	Broadcast News	Mandarin Chinese	VoA, China Television System, China Broadcasting System, China Central TV, and China National Radio
OntoNotes 3.0	2009	Broadcast Conversation	English	Phoenix TV and China Central TV
	2009	Broadcast Conversation	Mandarin Chinese or Chinese	Phoenix TV and China Central TV
	2009	Newswire Articles	Standard Arabic or Arabic	An-Nahar
OntoNotes 4.0	2011	Weblogs, Newsgroups	English	English P2.5
	2011	Weblogs, Newsgroups	Mandarin Chinese or Chinese	Dev09, P2.5
	2011	Newswire Articles	Standard Arabic or Arabic	An-Nahar
OntoNotes 5.0	2013	Telephone, Pivot	English	English CallHome, New Testament, Old Testament
	2013	Telephone	Mandarin Chinese or Chinese	Chinese CallHome
	2013	Newswire Articles	Arabic	An-Nahar

Source: Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik (2022). "Named entity recognition and relation extraction: State-of-the-art." ACM Computing Surveys (CSUR) 54, no. 1 (2022): 1-39.

Named Entity Recognition (NER)
Other Datasets

Dataset Name	Language	Data Source
MET [9]	Spanish, Japanese	MUC-6 dataset
IJCNLP [10]	Telugu, Bengali, Urdu, Hindi, Oriya	History of India including places and festivals
KPU-NE [11]	Urdu	Fifteen various sources including Education, Health, Science, Novels
Weibo [12]	Chinese	1,890 messages from social service provider “Weibo” with four entities GPE, person, location, and organization
Evalita	Italian	Tweets
		525 News stories taken from “L’Adige”
IREX	Japanese	Mainichi Newspaper
Mongolian [13]	Mongolian	33,209 sentences from news website

Source: Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik (2022). "Named entity recognition and relation extraction: State-of-the-art." ACM Computing Surveys (CSUR) 54, no. 1 (2022): 1-39.

Named Entity Recognition (NER) and Relation Extraction (RE)

Study Type	Pre 2000	2001–2005		2006–2010		2011–2015		Post 2015	
	NER	NER	RE	NER	RE	NER	RE	NER	RE
Rule-based	0	0	0	1	2	0	0	1	0
Supervised	2	3	4	4	2	2	1	4	0
Semi Supervised	1	1	3	0	5	2	4	0	0
Distant Supervised	0	0	0	0	2	0	3	0	0
Unsupervised	0	1	2	1	2	1	1	1	0
Deep Learning	0	0	0	1	0	4	2	18	10
Joint Modeling	0	0	0	1	3	0	2	0	2
Transfer Learning	0	0	0	0	0	1	0	10	2
Survey	0	0	0	1	1	4	1	4	4
Total	3	5	9	9	17	14	14	38	18

Source: Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik (2022). "Named entity recognition and relation extraction: State-of-the-art." ACM Computing Surveys (CSUR) 54, no. 1 (2022): 1-39.

Named Entity Recognition (NER)

	Technique[1]	Features/ Properties				Typ[2]	Results			Lang.	Dataset
		E	W	C	O		P	R	F		
[21]	HMM, MEMM	-	Y	Y		HR				English	CONLL
										German	CONLL
[22]	Semi-CRF, JM	Y	Y		Brown Clusters, Wiki	HR	91.5	91.4	91.2	English	CONLL
[26]	US	Y			Heuristics		Low	High			MUC-7
[27]	MLP		Y		Sliding Window	HR	87.41	86.15	86.76	English	Commercial offers
							85.57	86.22	85.95		Seminar Announ.
[28]	MLP	Y			Skip-gram	HR			90.9	English	CONLL
									82.3		OntoNotes
[29]	RNN		Y	Y	Language Model	HR			91.93	English	CONLL
[30]	Bi-LSTM		Y	Y	Language Model	HR			92.22	English	CONLL
[32]	Neuro-CRF		Y			HR			89.62	English	CONLL
[33]	Neuro-CRF		Y	Y	Bi-LSTM	HR				English	CONLL
[54]	Neuro-CRF		Y	Y	Bi-LSTM	HR	Multiple languages are used.				
[34]	Neuro-CRF		Y		CNN, Iterated Dilation	HR			90.65	English	CONLL
									84.53		OntoNotes5
[35]	Neuro-CRF		Y		Memory Network				89.5	English	CONLL
[42]	HMM	Y	Y	Y	Lexicalized HMM	OTH				Chinese	Multiple Chinese Datasets
[11]	MLP	-	Y	-	Context Window	OTH	81.05	87.54	84.17	Urdu	KPU-NE
[47]	SS				TBL	OTH	76.45	99.20	86.36	Filipino	Asian Hist. Ref.
[48]	SS		Y	Y	Bootstrapping, linguistic rules	OTH	73.03	71.62	72.31	Dutch	CONLL
							78.19	76.14	77.15	Spanish	CONLL

Source: Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik (2022). "Named entity recognition and relation extraction: State-of-the-art." ACM Computing Surveys (CSUR) 54, no. 1 (2022): 1-39.

Named Entity Recognition (NER)

	Technique[1]	Features/ Properties				Typ[2]	Results			Lang.	Dataset
		E	W	C	O		P	R	F		
[49]	SS	Y	Y		Iterative	OTH				Indon-esian	75 Wikipedia Articles
[74]	RNN	Y	Y		Early Stopping, Weight Decay		85.69	80.10	82.81	Italian	Evalita (Tweets and News)
[50]	DNN		Y	Y	Bi-GRU, AdaGrad	OTH			89.92	Czech	News
[52]	DNN		Y	Y	Co-training	OTH			94.56	Vietnam	VLSP
[53]	Neuro-CRF		Y	Y	LSTM, GRU, SCRN	OTH			90.89	Korean	ETRI
[72]	Heuristic	D	-	-		DOM	99.57	93.75	96.52	English	Dietary Recom.
[73]	CRF	D	Y			DOM	67.81	52.52	58.46	English	Micropost Twitter
[87]	US				Phrase Chunking	DOM			15.2	English	GENIA
									26.5		Pittsburgh
[74]	RNN	Y	Y		Early Stopping, Weight Decay	DOM	85.69	80.10	82.81	Italian	Evalita
[88]	LSTM		Y	Y		DOM	82.70	86.70	84.60	English	Pubmed Abstracts
[75]	LSTM	Y	Y	Y	Cross domain learning	DOM			59.78	Chinese	Social Media
[79]	CNN		Y		One vs rest approach	DOM			88.64	Chinese	Discharge Summ.
									91.13		Progress Note
[80]	Neuro-CRF				Document level features		87.38	87.38	87.38	Chinese	Marriage Judge.
							94.49	88.60	91.45		Contract Judge.

Source: Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik (2022). "Named entity recognition and relation extraction: State-of-the-art." ACM Computing Surveys (CSUR) 54, no. 1 (2022): 1-39.

Named Entity Recognition (NER)

	Technique[1]	Features/ Properties				Typ[2]	Results			Lang.	Dataset
		E	W	C	O		P	R	F		
[38]	HMM	-	Y	-	-	MUL	96.00	93.00	94.47	English	MUC-6
									90.00	Spanish	MET-1
[40]	MEMM	Y	Y		Reference Resolution	MUL			90.25	English	MUC-7
									83.80	Japanese	MET-2
									77.37	Japanese	IREX
[41].	CRF	Y	Y						84.04	English	CONLL
									68.11	German	CONLL
study [43]	CLM	Y	Y	Y	Language Models, CogCompNLP	MUL	Performance at par with recent DL frameworks			Tagalog, Somali, Hindi, Farsi, Bengali, Arabic, Amharic and English	
[61]	Neuro-CRF	Y	Y		Bi-LSTM and CRF for NER, CNN for word features	MUL			70.90	Marathi	
									55.57	Bengali	
									64.27	Malayalam	
									60.25	Tamil	
[60]	TL	Y	Y		Wikipedia, Translation of lexical resources, Cross-lingual NER	MUL	Training of each model using English and one relevant language			Dutch, German, Spanish, Turkish, Bengali, Tamil, Yoruba, Uyghur	

[1]SS, US, TL denote semi-supervised, unsupervised, respectively, and transfer learning.
[2]HR, OTH, MUL denotes high-resource, others, and multiple languages, respectively.

Source: Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik (2022). "Named entity recognition and relation extraction: State-of-the-art." ACM Computing Surveys (CSUR) 54, no. 1 (2022): 1-39.

Relation Extraction (RE)

Study	Technique	Evaluation Metrics			Features/ Model Properties	Dataset/ Genre
		P	R	F		
[105]	MEMM			52.8	Lexical, Semantic and Syntactic	ACE'02
				55.2		ACE'03
[88]	Bi-LSTM	67.5	75.8	71.4	Stacked LSTM Model	PubMed abstracts
[99]	Heuristic	68	83	75	Conjunction, Negation	LLL'05 workshop
[101]	Heuristic	75.5	62.1	68.1	Syntactic Parser, DBPedia	Quaero News
[107]	SVM	77.2	60.7	68.0	Lexical, Semantic, Syntactic, External Lexicon	ACE'03
[109]	SVM	82.7	91.3	86.0	Kernels and voted perceptron	200 newswire and publications
[110]	SVM	70.3	26.3	38.0	Tree Kernel	ACE
[111]	SVM	76.1	68.4	72.1	Tree Kernel	ACE'03
[113]	Bootstrapping with SVM	63.2	61.5	60.3	Radial Bias Kernel	Self-annotated
[115]	CRF	73.4	56.1	63.6	Relational pattern features. Word, external	Wikipedia articles
[94]	SS				BootStrapping, Ontology	Sports and Companies web pages
[117]	SVM				Semantic Classes, Partial Pattern	TSUBAKI
[118]	SS	57.0			KBs, Tensor Decomposition	New York Times dataset [119]
[119]	Collaborative Filtering	69.0			KB, Universal Schemas	New York Times dataset

Source: Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik (2022). "Named entity recognition and relation extraction: State-of-the-art." ACM Computing Surveys (CSUR) 54, no. 1 (2022): 1-39.

Relation Extraction (RE)

Study	Technique	Evaluation Metrics			Features/ Model Properties	Dataset/ Genre
		P	R	F		
[116]	Multi-class Logistic Regression	68.0			KBs, Lexical and Syntactical Features	Self-annotated using Mechanical Turk
[121]	DS	87.0			SampleRank, CRF, FreeBase	New York Times dataset
[123]	Logistic Regression	78.2	68.2	66.7	Freebase, EM	Wikipedia articles
[144]	LSTM				Attention Mechanism	NYT
[145]	CNN	Accuracy: 86.2 77.3			Semantic Jaccard	Wikipedia Articles New York Times
[146]	Piece-wise CNN	46.9	44.5	45.7	Word-level attention model	NYT [121]
[147]	Multi-path CNN	77.0			Word and sentence level attention model	NYT [121]
[154]	Clustering	77.5	78.5	77.5	Hierarchy NER, Complete Linkage	NYT
[125]	US				Unsupervised Feature Subset Selection, K-means	
[126]	US	Accuracy: 79.5			Chunking Information, Hierarchical Clustering	News
[127]	HMM	85.1				Web-pages
[128]	US	89.7	68.4	77.6	Hierarchical Clustering	Cluewebset'09

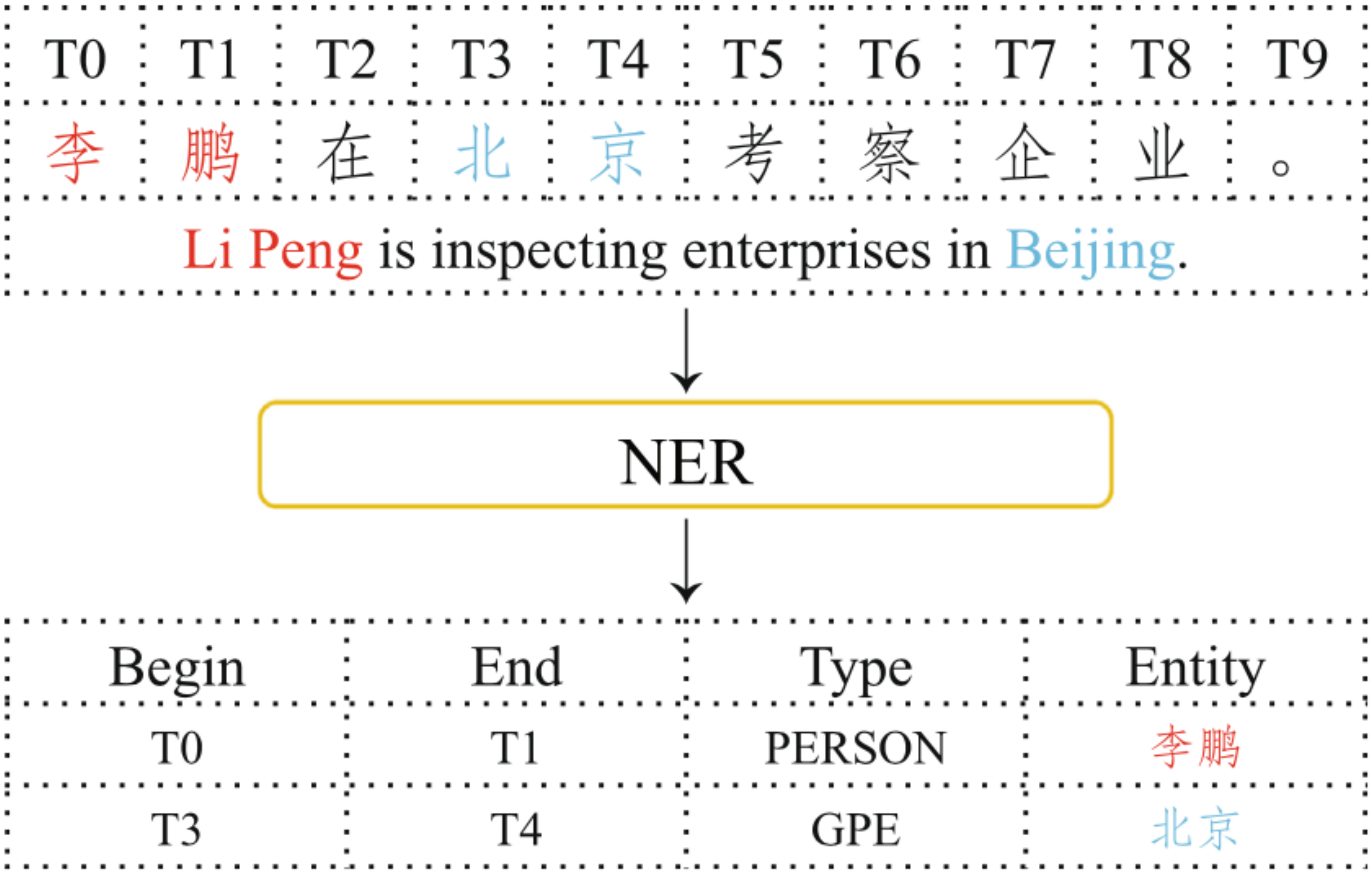
Source: Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik (2022). "Named entity recognition and relation extraction: State-of-the-art." ACM Computing Surveys (CSUR) 54, no. 1 (2022): 1-39.

Relation Extraction (RE)

Study	Technique	Evaluation Metrics			Features/ Model Properties	Dataset/ Genre
		P	R	F		
[138]	RNN	82.4		82.4	POS Tags, NER Tags, Wordnet Hypernyms	SemEval 2010
[139]	CNN	82.7		82.7	Wordnet	SemEval 2010
[140]	CNN	88.0		88.0	Multi-level Attention Model	SemEval 2010
[141]	RNN	79.0		79.0	Skip-gam-based Word Vectors	SemEval 2010
[142]	LSTMs	72.9	70.8	67.9	Dynamic models	CONLL'04
[129]	Viterbi	54.0	68.4	58.14	Inferencing	TREC documents
[130]	Joint Model	90.1 73.0	91.8 62.7	91.3 66.0	POS Tags, Context Words, Hybrid Model including SVM, CYK-Parsing	TREC documents [129]
[155]	Joint Model	94.0 76.0			Graph	New York Times data
[131]	Joint Model	93.4 72.6	93.4 64.3	93.4 68.2	BootStrapping with Markov Models and CRF, Joint Model	Wikipedia
[148]	Joint Model	83.5 64.7	76.2 38.5	79.7 48.3	Casing, Gazetteer, Relation Features, Perceptron	ACE'04
		85.2 68.9	76.9 41.9	80.8 52.1		ACE'05
[132]	Joint Model	92.4 83.7	92.4 59.9	92.4 69.8	History Info., Structured Learning	TREC documents [129]
[149]	Joint Model	80.8 48.7	82.9 48.1	81.8 48.4	Bi-directional LSTM	ACE'04
		82.9 57.2	83.9 54.0	83.4 55.6		ACE'05
[143]	LSTM, Capsule Networks	30.8	63.7	41.6	Attention re-routing, position embedding	NYT
				84.5		SemEval-2010
[150]	Transfer Learning				Knowledge bases	Wiki-KBP NYT

Source: Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik (2022). "Named entity recognition and relation extraction: State-of-the-art." ACM Computing Surveys (CSUR) 54, no. 1 (2022): 1-39.

Chinese Named Entity Recognition (CNER)



An illustration of NER task. The sample sentence is from People's daily dataset, and GPE means Geo-Political Entity.

Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

Named Entity Recognition (NER)

Language	Ref.	Year	Topic
Universal	[1]	2007	A survey of named entity recognition and classification
Universal	[2]	2008	Named entity recognition approaches
Universal	[3]	2013	Techniques for named entity recognition: a survey
Universal	[4]	2018	An overview of named entity recognition
Universal	[5]	2018	Recent named entity recognition and classification techniques: a systematic review
Universal	[6]	2019	A survey on named entity recognition
Universal	[7]	2020	A survey on deep learning for named entity recognition
Universal	[8]	2020	A survey of named-entity recognition methods for food information extraction

Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

Named Entity Recognition (NER)

Language	Ref.	Year	Topic
Arabic	[13]	2017	A comparative review of machine learning for Arabic named entity recognition
Arabic	[14]	2019	Arabic named entity recognition using deep learning approach
Arabic	[15]	2019	Arabic named entity recognition: What works and what's next
Indian	[16]	2010	A survey of named entity recognition in English and other Indian languages
Indian	[17]	2011	A survey on named entity recognition in Indian languages with particular reference to Telugu
Indian(Assamese)	[18]	2014	A survey of named entity recognition in Assamese and other Indian languages
Indian(Hindi)	[19]	2016	Survey of named entity recognition systems with respect to Indian and foreign languages
Indian	[20]	2017	Survey of named entity recognition techniques for various Indian regional languages
Indian(Hindi)	[21]	2019	Named entity recognition for Hindi language: A survey
Indian	[22]	2019	Named entity recognition: A survey for Indian languages
Indian(Hindi)	[23]	2020	A survey on various methods used in named entity recognition for hindi language
English	[24]	2013	Named entity recognition in english using hidden markov model
Marathi	[25]	2016	Issues and Challenges in Marathi Named Entity Recognition
Turkish	[26]	2017	Named entity recognition in Turkish: Approaches and issues
Spanish	[27]	2020	Named entity recognition in Spanish biomedical literature: Short review and bert model

Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

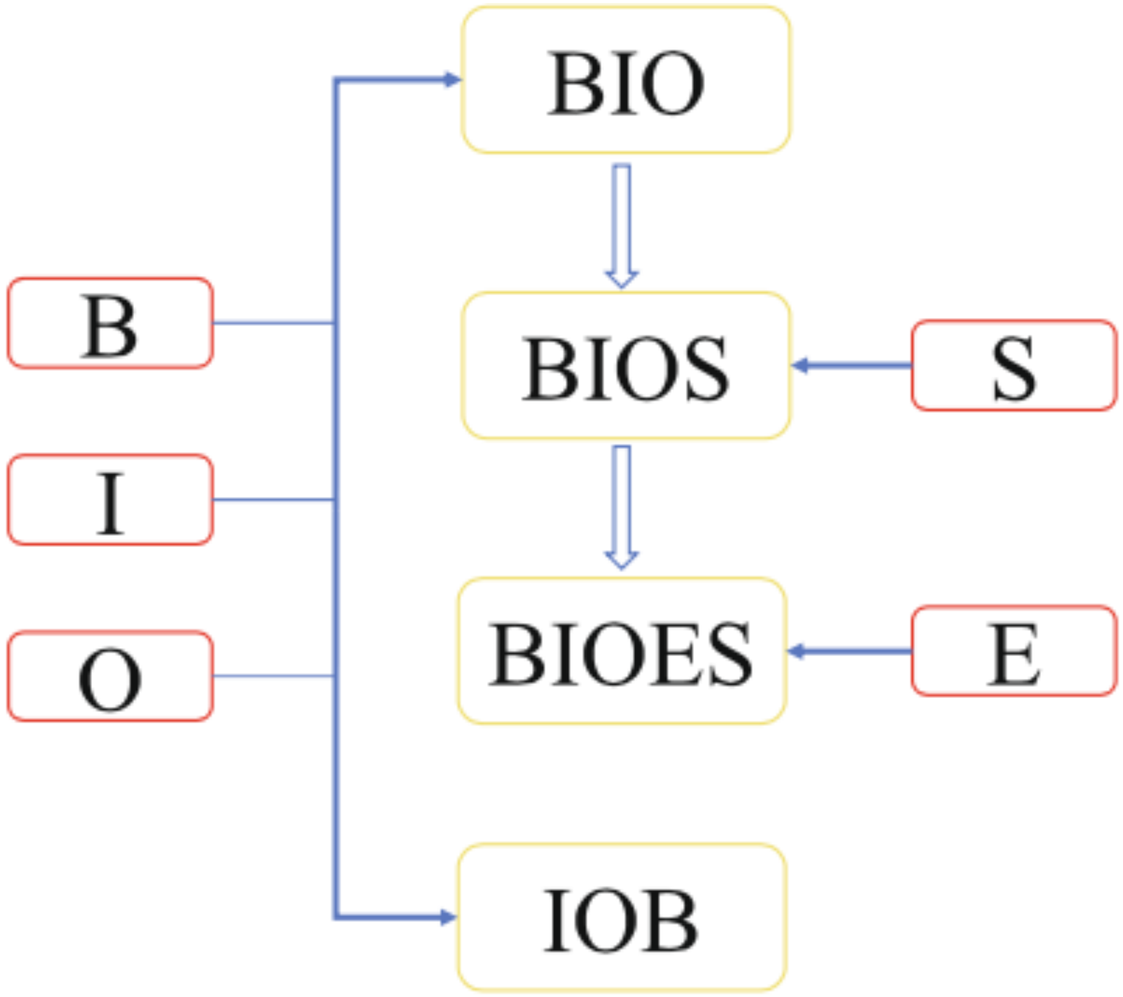
Public datasets of Chinese NER

Corpus	#Tags	Entity types	URL
WEIBO	4	Person, Location, Organization and Geo-political	https://github.com/hltcoe/golden–horse
MSRA	3	Person, Location, Organization	https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER/MSRA
People's Daily	4	Person, Organization, Geo-political, Date	https://github.com/GuocaiL/nlp_corpus/tree/main/open_ner_data/people_daily
bosonNLP	6	Person, Location, Organization, Company, Product, Time	https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER/boson
RESUME	8	Person, Location, Organization, Country, Education, Profession, Race, Title	https://github.com/GuocaiL/nlp_corpus/tree/main/open_ner_data/ResumeNER
OntoNotes Release 5.0	18	Preson, NORP, Facility, Organization, GPE, Location, Product, Event, Work of art, Law, Language, Date, Time, Percent, Money, Quantity, Ordinal, Cardinal	https://doi.org/10.35111/xmhb-2b84
CLUENER 2020	10	Address, Book, Company, Game, Government, Movie, Name, Organization, Position, Scene	https://github.com/CLUEbenchmark/CLUENER2020

#Tags: the number of entity types

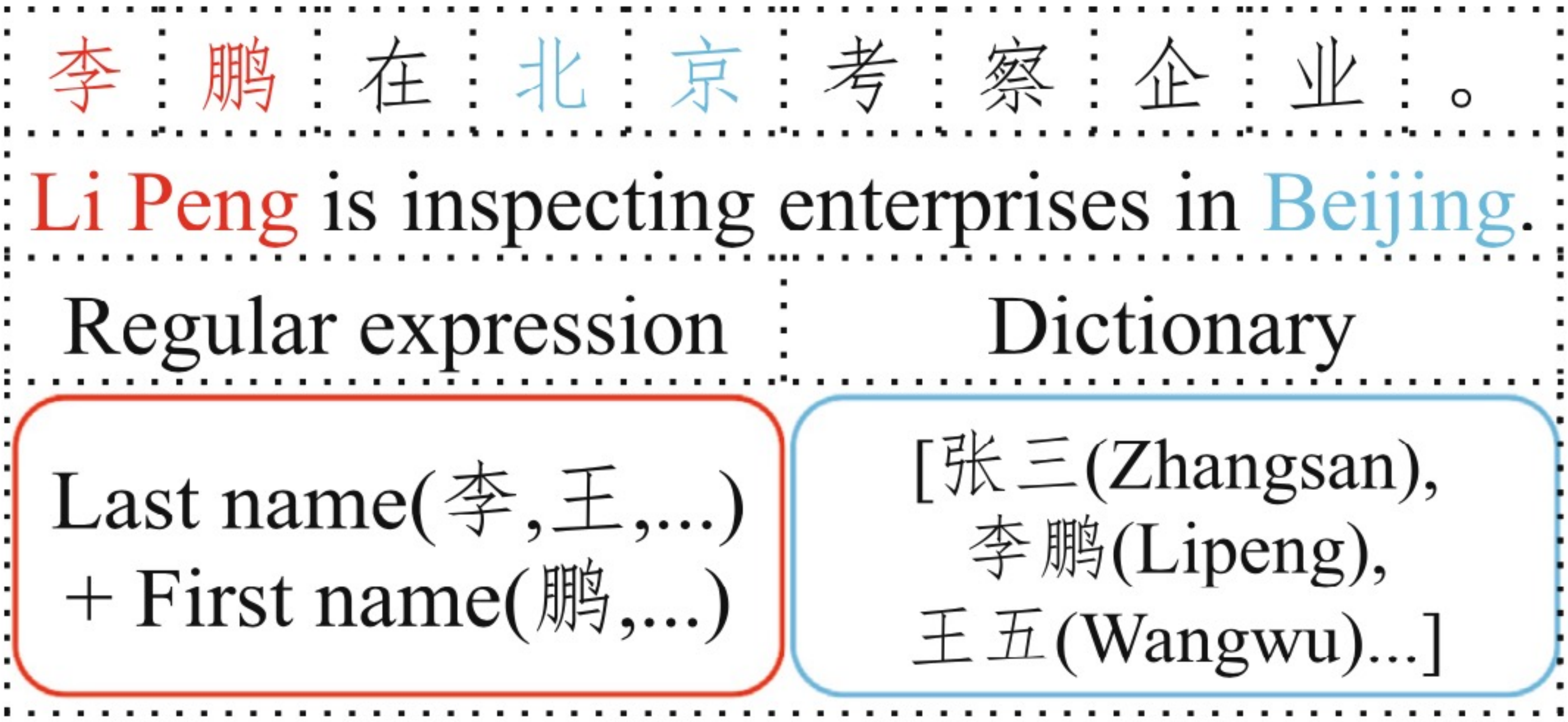
Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

Evolution of four commonly used tag schemes



Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

Named Entity Recognition (NER)



An illustration of rule-based methods.

A person’s name is matched by a regular expression and a dictionary.

Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

Named Entity Recognition (NER)

Lexicon	朝阳 (morning sun)	明朝 (Ming Dynasty)	朝鲜半岛 (Korean Peninsula)	朝夕 (morning and evening)
Glyph	Oracle Bone Script	Bronze Script	Clerical Script	朝 Regular Script
Radical	十 (ten)	日 (sun)	十 (ten)	月 (moon)
Stoker	一 丨	丨 𠃌 一 一	一 丨	丿 𠃌 一 一

The illustration of some external information of the character “朝”

Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

Named Entity Recognition (NER)

Improvement brought by BERT in different works

Work	Dataset	Model	F1(%)	Improvement(%)	Year
[63]	MSRA	Word2Vec + radical + BGRU-CRF	90.45	4.97	2019
		BERT + radical + BGRU-CRF	95.42		
[74]	MSRA	PLTE	93.26	1.27	2020
		PLTE[BERT]	94.53		
	Ontonotes	PLTE	74.60	6.00	
		PLTE[BERT]	80.60		
	Weibo	PLTE	55.15	14.08	
		PLTE[BERT]	69.23		
[75]	MSRA	SoftLexicon(LSTM)	93.66	1.76	2020
		SoftLexicon(LSTM)+BERT	95.42		
	Ontonotes	SoftLexicon(LSTM)	75.64	7.17	
		SoftLexicon(LSTM)+BERT	82.81		
	Weibo	SoftLexicon(LSTM)	61.42	9.08	
		SoftLexicon(LSTM)+BERT	70.50		
[76]	CCKS2018	Word2Vec + CRF	69.01	21.53	2020
		BERT + CRF	90.54		
		Word2Vec + BILSTM-CRF	75.60	15.83	
		BERT + BILSTM-CRF	91.43		

Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

Named Entity Recognition (NER)
The effect of POS and radical information

Work	Dataset	Model	F1(%)	Improvement(%)	Year
[55]	MSRA	random + dropout	88.91	0.53	2016
		random + radical + dropout	89.44		
[58]	CCKS2018	LSTM-CRF	67.32	11.62	2019
		POS + LSTM-CRF	78.94		
		SM-LSTM-CRF	69.91	10.16	
		POS + SM-LSTM-CRF	80.07		
[60]	CCKS2017	BILSTM-CRF	88.78	Baseline	2019
		BILSTM-CRF + radical	89.64	0.86	
		BILSTM-CRF + POS	89.06	0.28	
		BILSTM-CRF + radical + POS	90.12	1.34	
		Att-BILSTM-CRF	90.11	Baseline	
		Att-BILSTM-CRF + radical	90.96	0.85	
		Att-BILSTM-CRF + POS	90.81	0.70	
		Att-BILSTM-CRF + radical + POS	91.35	1.24	
[61]	CCKS2017	CRF	85.14	1.87	2019
		POS + CRF	87.01		
		BILSTM-CRF	89.66	-0.11	
		POS + BILSTM-CRF	89.55		
	CCKS2018	CRF	82.49	0.93	
		POS + CRF	83.42		
		BILSTM-CRF	84.13	-0.17	
		POS + BILSTM-CRF	83.96		

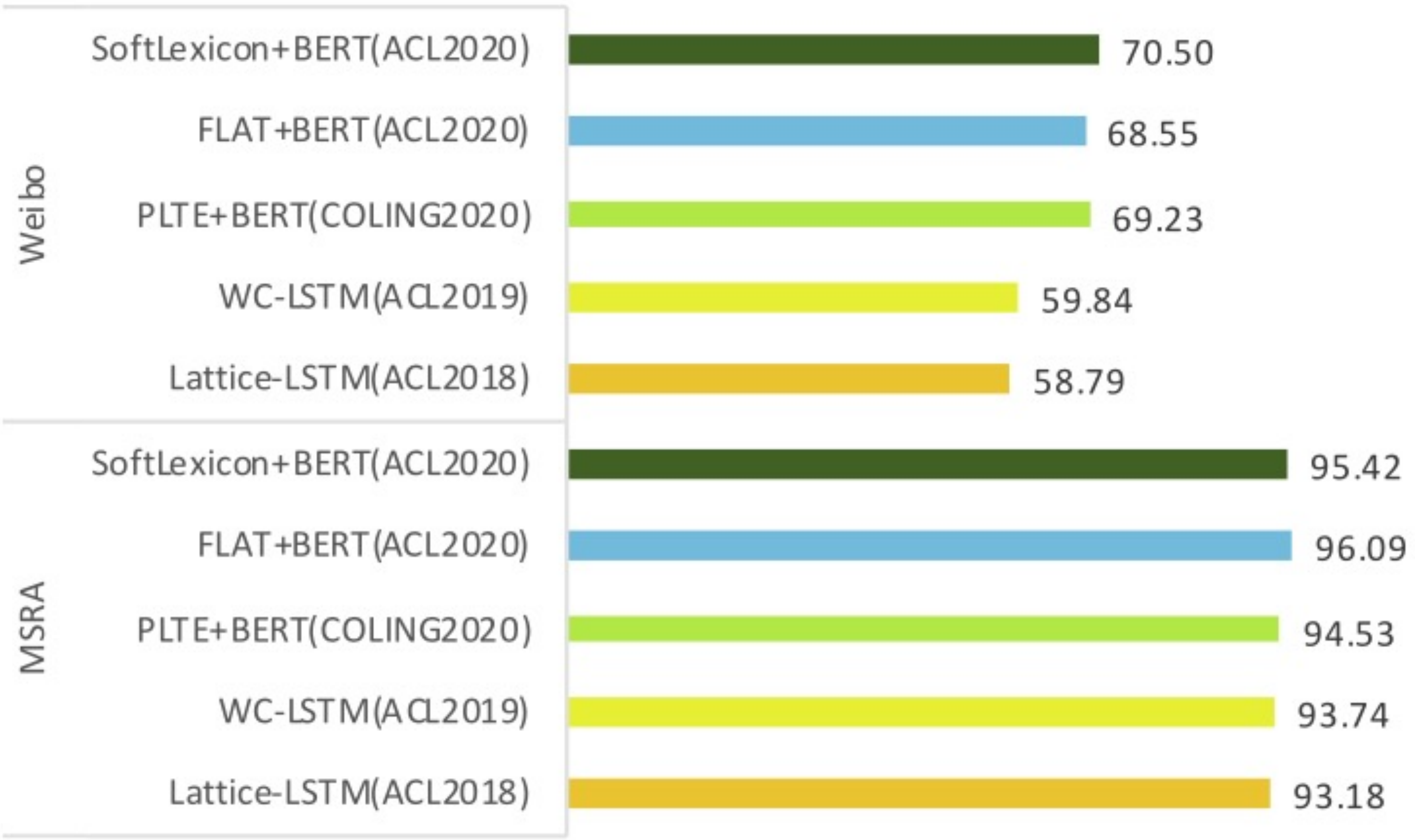
Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

Named Entity Recognition (NER) Improvement brought by Glyph information

Work	Dataset	Model	F1(%)	Improvement(%)	Year
[81]	MSRA	BERT	94.80	0.74	2019
		BERT + Glyce	95.54		
		Lattice-LSTM	93.18	0.71	
		Lattice-LSTM + Glyce	93.89		
[57]	MSRA	BILSTM-CRF	89.94	1.14	2019
		BILSTM-CRF + glyph embeddings	91.08		
[83]	MSRA	BERT + BILSTM-CRF	95.30	1.19	2019
		BERT + BILSTM-CRF + GLYNN	96.49		

Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

Named Entity Recognition (NER)



Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

Named Entity Recognition (NER)

The illustration of representations of the character “朝”

	Pre-trained character embeddings		External Information					
	Lookup tables	Pre-trained language models	POS	Radical Information	Stroke Information	Glyph Information	Lexicon Information	Chinese Word Segmentation
Illustrations	Word2vec, Glove, FastText, etc.	BERT, ELMo, NEZHA, etc.	Noun	十 日 十 月	一丨丨㇆一 一一丨丿㇆ 一一	朝	朝阳, 明朝, 朝夕, 朝鲜	明朝/的/皇帝
methods	Static embeddings	Contextual embeddings	embeddings	embeddings and RNN	RNN	CNN	Lattice	CWS tools

Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

Chinese Named Entity Recognition (CNER)

	Character representation		Attention		Attention			
Work	Character embeddings	External Information	Input -> Encoder	Context Encoder	Encoder -> Decoder	Tag Decoder	Performance (F1-score)	Year
[55]		Radical		LSTM		CRF	MSRA:89.78%	2016
	Word2vec						MSRA:90.95%	
[58]	√	POS		LSTM		CRF	CCKS2018:78.94%	2019
			√				CCKS2018:80.07%	
[59]	√			LSTM		CRF	CCKS2018:86.68%	2019
					√		CCKS2018:87.26%	
[60]	√	POS, Radical		LSTM		CRF	CCKS2017:90.12%	2019
					√		CCKS2017:91.35%	
[61]	√	POS, Dictionary		LSTM	√	CRF	CCKS2017:90.48%	2019
							CCKS2018:86.11%	
[80]	Word2vec	CWS, Radical, Lexicon, Stroker		LSTM		CRF	CCKS2017:91.75%	2020
							CCKS2018:90.05%	
[76]	Word2vec			O		CRF	CCKS2018:69.01%	2020
	BERT						CCKS2018:90.54%	
	ERNIE						CCKS2018:93.37%	
	ALBERT						CCKS2018:87.68%	
	NEZHA						CCKS2018:93.58%	
	Word2vec			LSTM			CCKS2018:75.60%	
	BERT						CCKS2018:91.43%	
	ERNIE						CCKS2018:93.11%	
	ALBERT						CCKS2018:90.12%	
	NEZHA						CCKS2018:95.08%	
[56]	Sogou news	Radical		LSTM		CRF	Peoples'Daily:92.06%	2019
	Word2vec				√		Peoples'Daily:94.37%	
[62]	√	Position, segmentation		GRU		CRF	WEIBO:53.80% MSRA:90.32%	2019
			Concolution - attention				WEIBO:55.91% MSRA:92.34%	
					√		WEIBO:59.31% MSRA:92.97%	

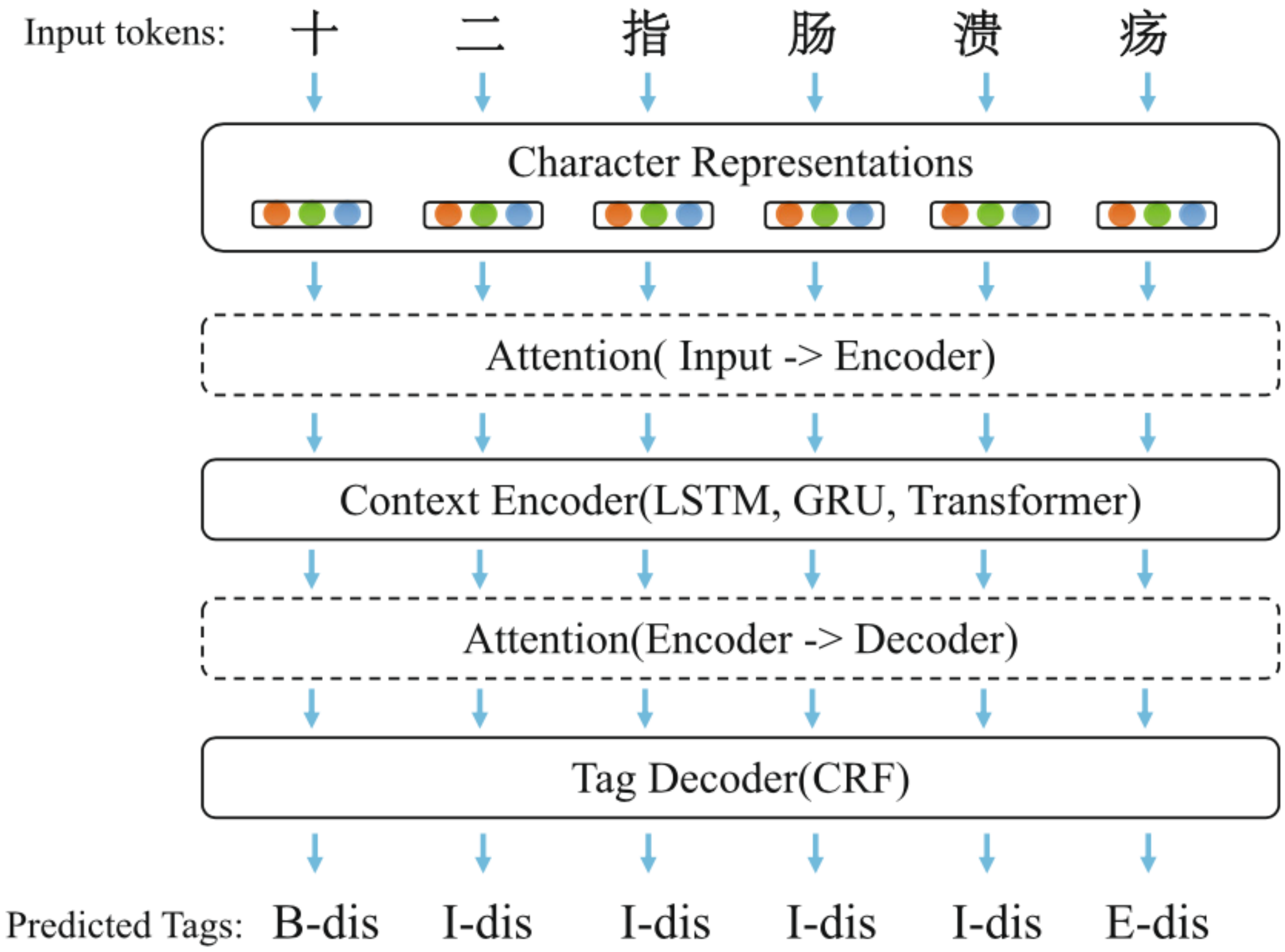
Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

Chinese Named Entity Recognition (CNER)

	Character representation		Attention		Attention			
Work	Character embeddings	External Information	Input -> Encoder	Context Encoder	Encoder -> Decoder	Tag Decoder	Performance (F1-score)	Year
[63]	Word2vec	Radical		GRU		CRF	MSRA:90.45%	2019
	BERT						MSRA:95.42%	
[77]	Conv-GRU Embedding	Word, Radical		GRU		CRF	WEIBO:68.93% MSRA:91.45%	2019
[88]	√	Dictionary		LSTM		CRF	CCKS2017:91.24%	2019
[64]	√	Lexicon, Word		LSTM		CRF	WEIBO:63.09% MSRA:93.47%	2019
[65]	√	Word, Position		LSTM	√	CRF	WEIBO:59.5% MSRA:92.99%	2020
[81]	BERT	Glyph		Transformer		CRF	WEIBO:67.60% MSRA:95.54%	2019
[57]	Wikipedia GloVe	Glyph		LSTM		CRF	MSRA:91.11%	2019
[82]	BERT	Radical, Glyph		LSTM		CRF	WEIBO:70.01% MSRA:95.51%	2020
[83]	BERT	Glyph		LSTM		CRF	WEIBO:71.81% MSRA:96.49%	2019
[66]	√	Radical, Word	√	GRU		CRF	WEIBO:71.86% MSRA:92.71%	2020
[67]	√	Adapted GGNN Gazetteers		LSTM		CRF	WEIBO:59.5% MSRA:94.4%	2020
[85]	√	Lexicon		Lattice-LSTM		CRF	WEIBO:58.79% MSRA:93.18%	2018
[86]	√	Lexicon		WC-LSTM		CRF	WEIBO:59.84% MSRA:93.36%	2019
[74]	√	Lexicon		PLTE		CRF	WEIBO:55.15% MSRA:93.26%	2019
	BERT						WEIBO:69.23% MSRA:94.53%	
[87]	BERT			MLP		CRF	WEIBO:68.20% MSRA:94.95%	2020
	√	Lexicon		FLAT			WEIBO:63.42% MSRA:94.35%	
	BERT						WEIBO:68.55% MSRA:96.09%	
[75]	√	SoftLexicon		LSTM		CRF	WEIBO:61.42% MSRA:93.66%	2020
	BERT						WEIBO:70.50% MSRA:95.42%	
[99]	BERT	Lexicon, radical		Transformer	√	CRF	WEIBO:70.43% MSRA:96.24%	2021

Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

Named Entity Recognition (NER)



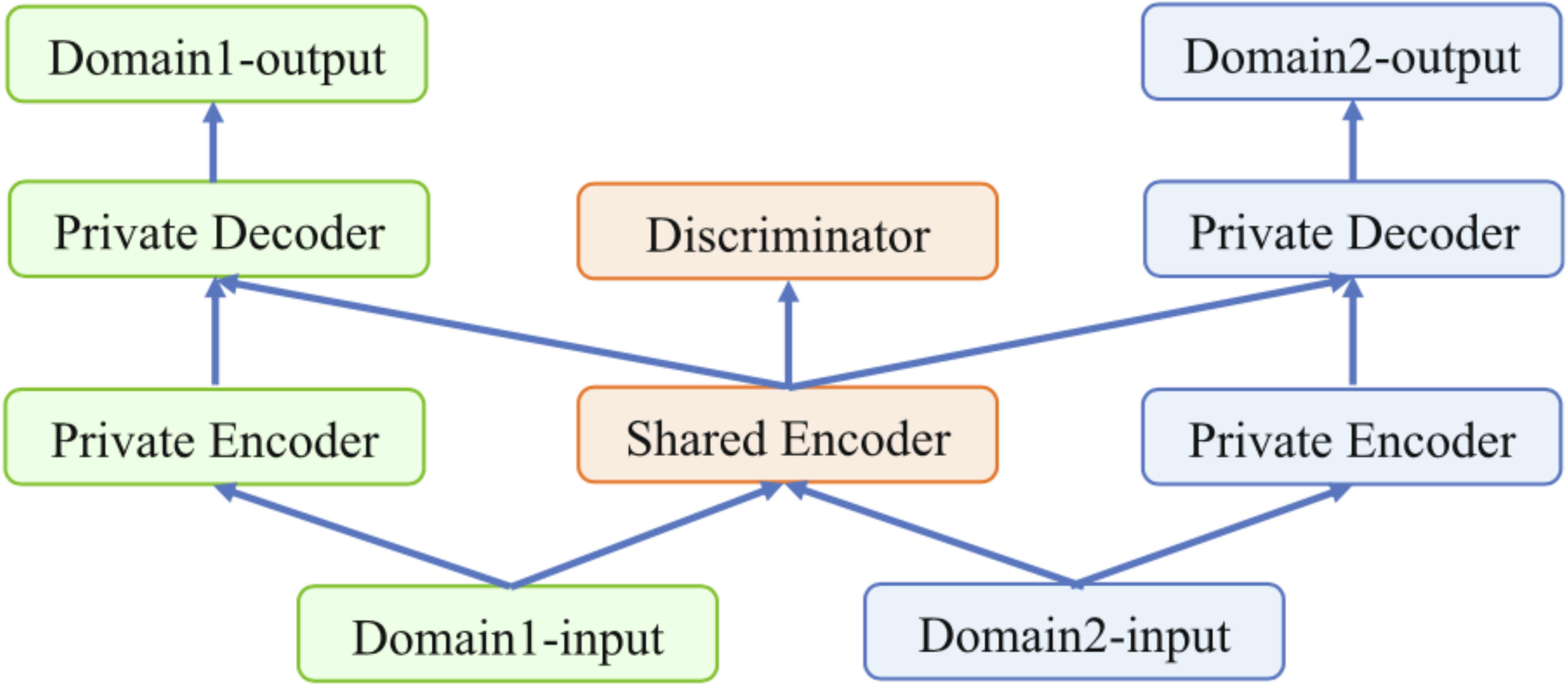
Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

Chinese Named Entity Recognition (CNER) using attention modules

Work	Dataset	Model	F1(%)	Improvement(%)
[58]	CCKS2018	LSTM-CRF	67.32	2.59
		SM-LSTM-CRF	69.91	
		POS + LSTM-CRF	78.94	1.13
		POS + SM-LSTM-CRF	80.07	
[60]	CCKS2017	BILSTM-CRF	88.78	1.33
		Att-BILSTM-CRF	90.11	
		BILSTM-CRF + radical	89.64	1.32
		Att-BILSTM-CRF + radical	90.96	
		BILSTM-CRF + POS	89.06	1.75
		Att-BILSTM-CRF + POS	90.81	
		BILSTM-CRF + radical + POS	90.12	1.23
		Att-BILSTM-CRF + radical + POS	91.35	
[59]	CCKS2018	BILSTM-CRF	86.68	0.58
		Attention-BILSTM-CRF	87.26	
		BILSTM-CRF + dictionary	87.71	0.58
		Attention-BILSTM-CRF + dictionary	88.29	
[56]	CCKS2018	char	86.09	3.17
		char + attention	89.26	
		char + word	90.74	3.74
		char + word + attention	94.48	

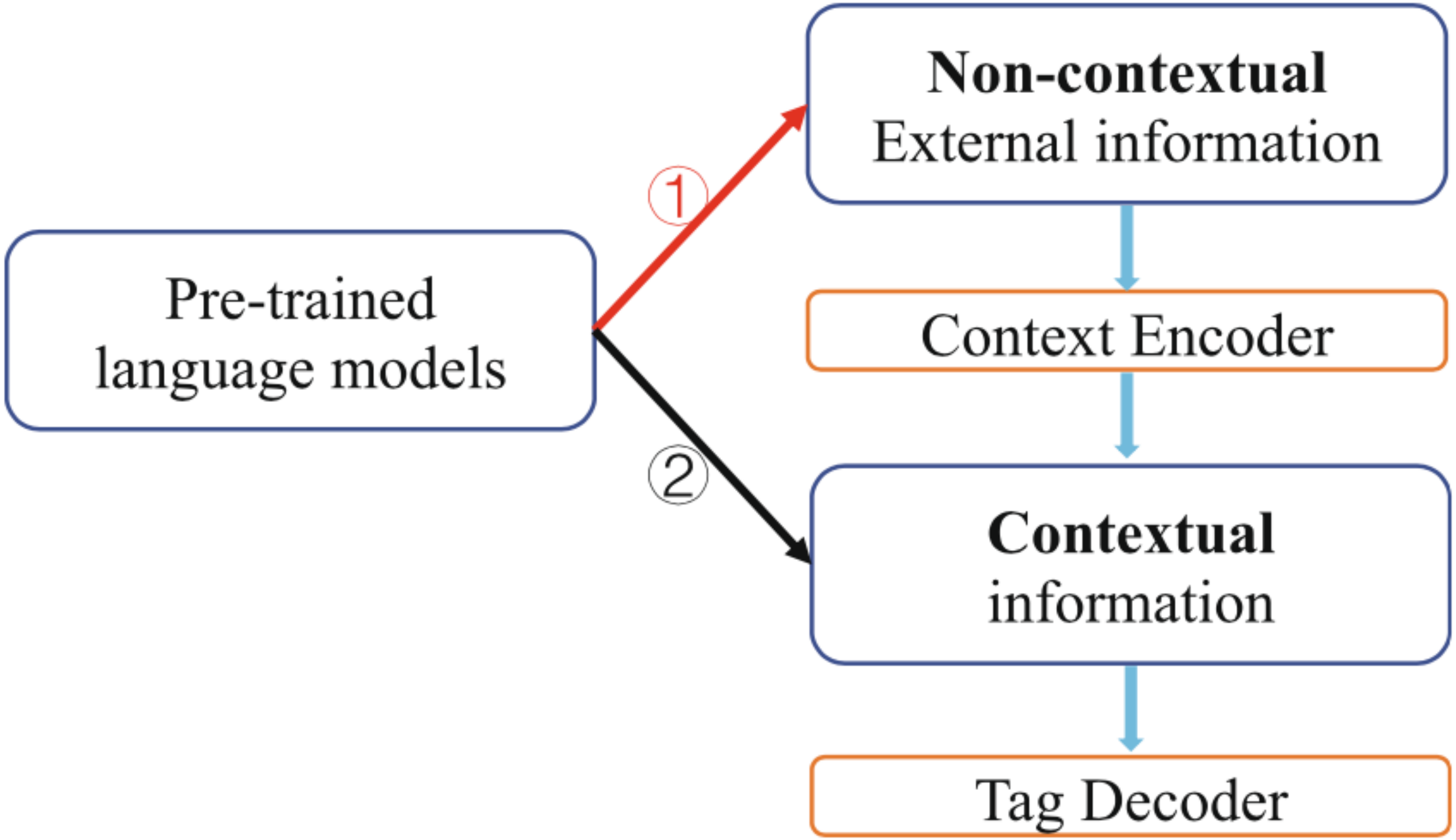
Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

Schematic diagram of cross-domain adversarial transfer learning



Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

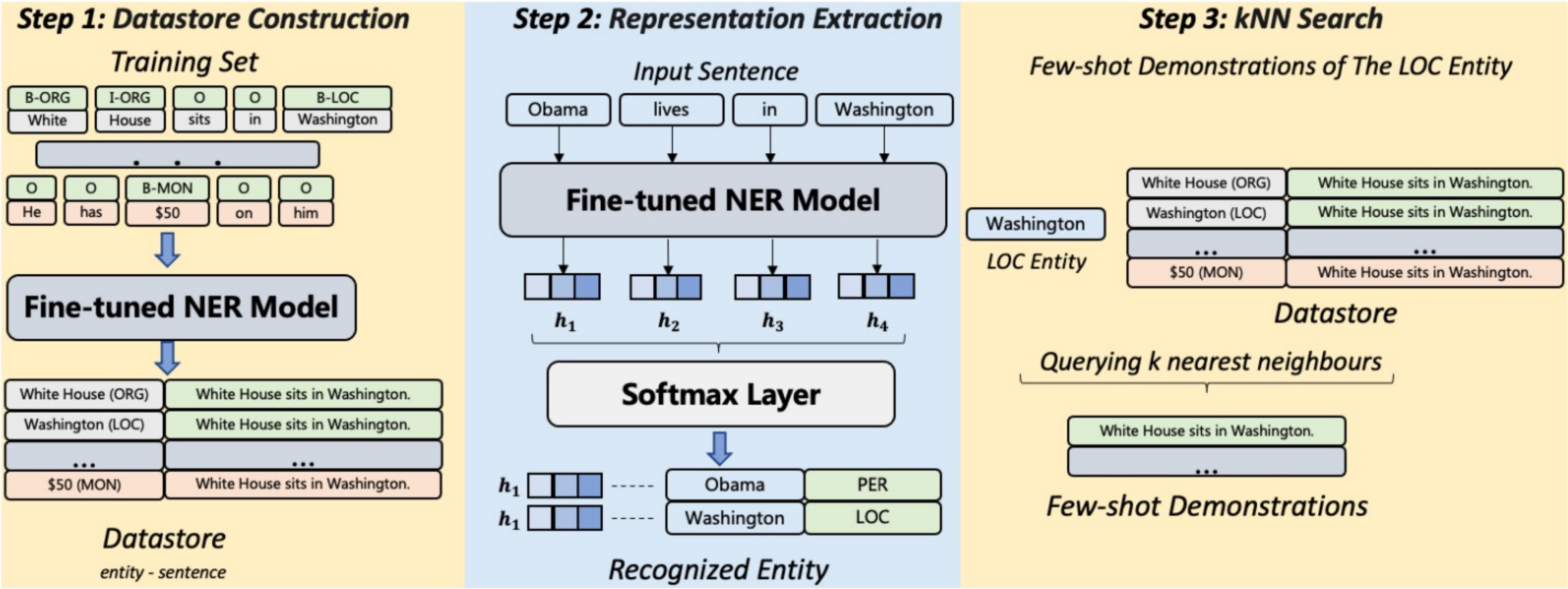
Two ways of concatenating the representations of pre-trained language models and external information



Source: Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.

GPT-NER: Named Entity Recognition (NER) via LLM

Entity-level embedding to retrieve few-shot demonstrations



Source: Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang (2023). "Gpt-ner: Named entity recognition via large language models." arXiv preprint arXiv:2304.10428 (2023).

GPT-NER: Named Entity Recognition (NER) via LLM

English CoNLL2003 (Sampled 100)			
Model	**Precision**	**Recall**	**F1**
Baselines (Supervised Model)			
ACE+document-context (Wang et al., 2020)	97.8	98.28	**98.04** (SOTA)
GPT-NER			
GPT-3 + *random retrieval*	88.18	78.54	83.08
GPT-3 + *sentence-level embedding*	90.47	95	92.68
GPT-3 + *entity-level embedding*	94.06	96.54	95.3
Self-verification (zero-shot)			
+ GPT-3 + *random retrieval*	88.95	79.73	84.34
+ GPT-3 + *sentence-level embedding*	91.77	96.36	94.01
+ GPT-3 + *entity-level embedding*	94.15	96.77	95.46
Self-verification (few-shot)			
+ GPT-3 + *random retrieval*	90.04	80.14	85.09
+ GPT-3 + *sentence-level embedding*	92.92	95.45	94.17
+ GPT-3 + *entity-level embedding*	94.73	96.97	95.85

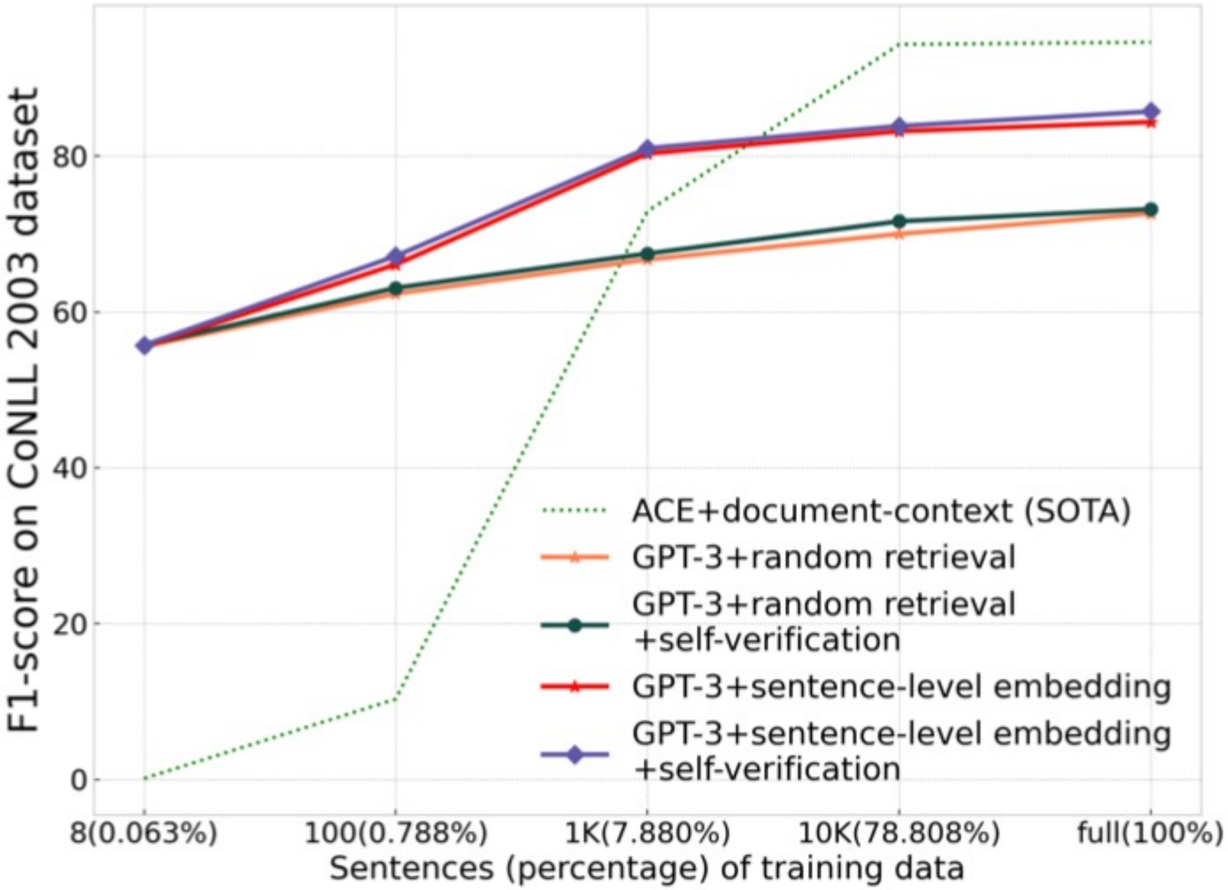
Source: Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang (2023). "Gpt-ner: Named entity recognition via large language models." arXiv preprint arXiv:2304.10428 (2023).

GPT-NER: Named Entity Recognition (NER) via LLM

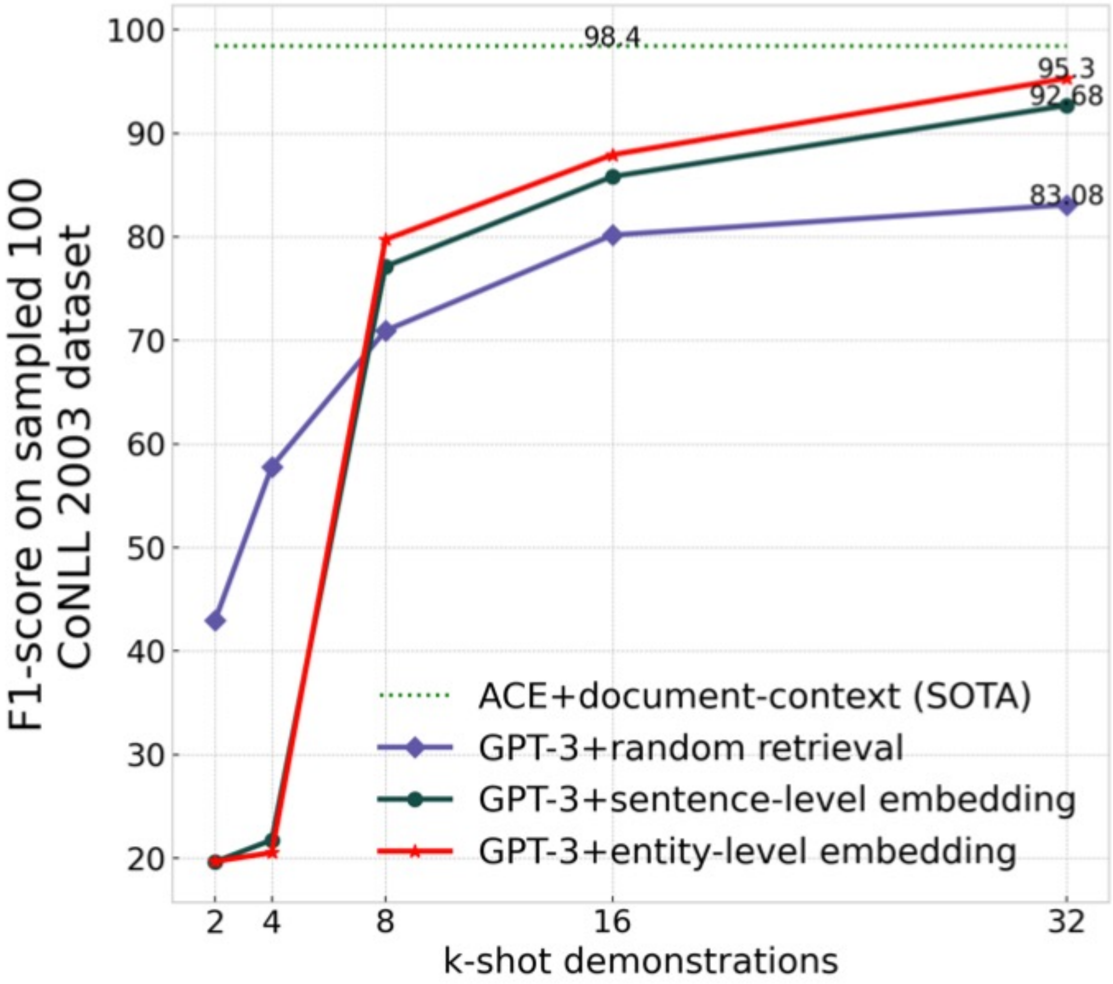
English CoNLL2003 (FULL)			
Model	**Precision**	**Recall**	**F1**
Baselines (Supervised Model)			
BERT-Tagger (Devlin et al., 2018)	-	-	92.8
BERT-MRC (Li et al., 2019a)	92.33	94.61	93.04
GNN-SL (Wang et al., 2022)	93.02	93.40	93.2
ACE+document-context (Wang et al., 2020)	-	-	**94.6 (SOTA)**
GPT-NER			
GPT-3 + *random retrieval*	77.04	68.69	72.62
GPT-3 + *sentence-level embedding*	81.04	88.00	84.36
GPT-3 + *entity-level embedding*	88.54	91.4	89.97
Self-verification (zero-shot)			
+ GPT-3 + *random retrieval*	77.13	69.23	73.18
+ GPT-3 + *sentence-level embedding*	83.31	88.11	85.71
+ GPT-3 + *entity-level embedding*	89.47	91.77	90.62
Self-verification (few-shot)			
+ GPT-3 + *random retrieval*	77.50	69.38	73.44
+ GPT-3 + *sentence-level embedding*	83.73	88.07	85.9
+ GPT-3 + *entity-level embedding*	89.76	92.06	90.91

Source: Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang (2023). "Gpt-ner: Named entity recognition via large language models." arXiv preprint arXiv:2304.10428 (2023).

GPT-NER: Named Entity Recognition (NER) via LLM



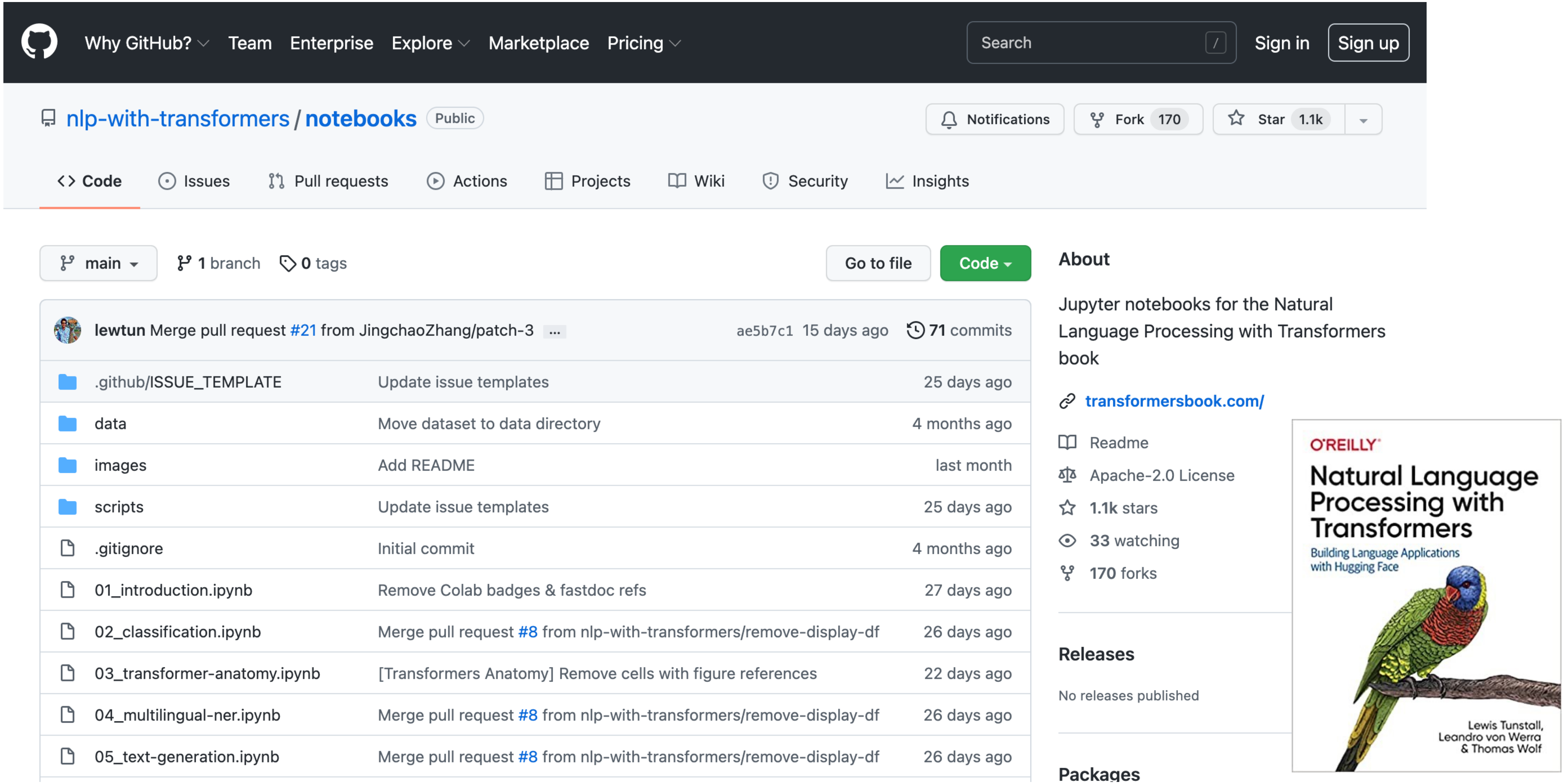
Low-resource comparisons on CoNLL2003 dataset.



Comparisons by varying k-shot demonstrations.

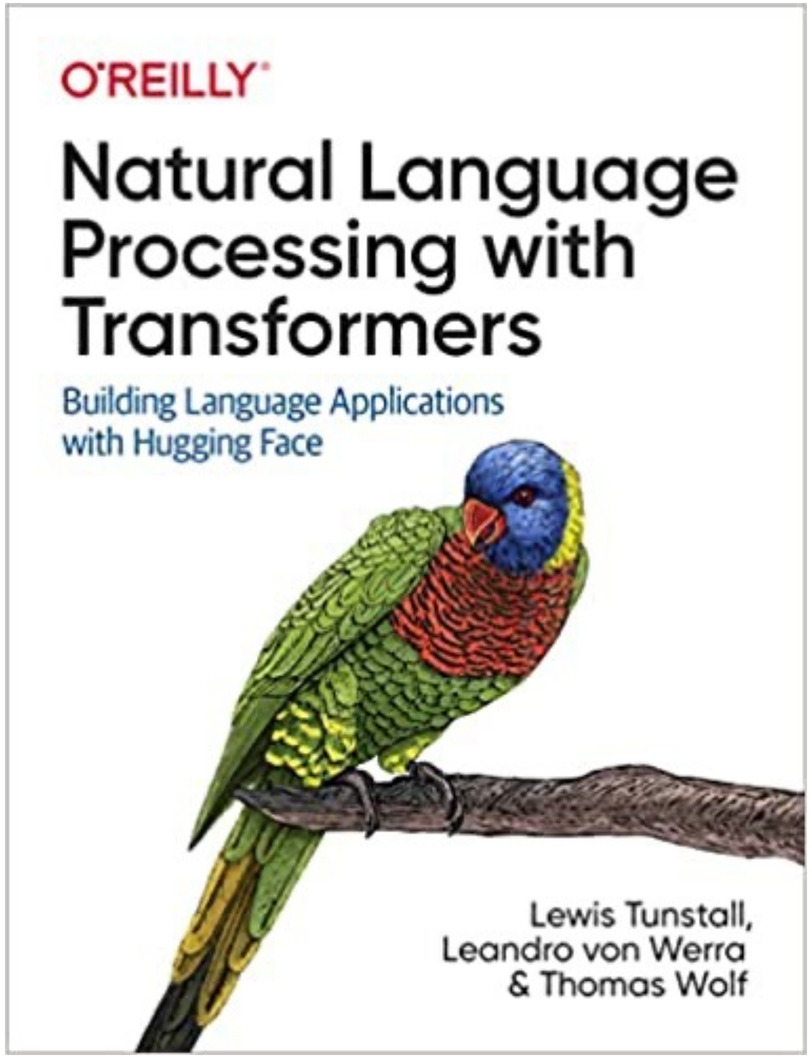
Source: Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang (2023). "Gpt-ner: Named entity recognition via large language models." arXiv preprint arXiv:2304.10428 (2023).

NLP with Transformers Github



https://github.com/nlp-with-transformers/notebooks

NLP with Transformers Github Notebooks



Running on a cloud platform

To run these notebooks on a cloud platform, just click on one of the badges in the table below:

Chapter	Colab	Kaggle	Gradient	Studio Lab
Introduction	Open in Colab	Open in Kaggle	Run on Gradient	Open Studio Lab
Text Classification	Open in Colab	Open in Kaggle	Run on Gradient	Open Studio Lab
Transformer Anatomy	Open in Colab	Open in Kaggle	Run on Gradient	Open Studio Lab
Multilingual Named Entity Recognition	Open in Colab	Open in Kaggle	Run on Gradient	Open Studio Lab
Text Generation	Open in Colab	Open in Kaggle	Run on Gradient	Open Studio Lab
Summarization	Open in Colab	Open in Kaggle	Run on Gradient	Open Studio Lab
Question Answering	Open in Colab	Open in Kaggle	Run on Gradient	Open Studio Lab
Making Transformers Efficient in Production	Open in Colab	Open in Kaggle	Run on Gradient	Open Studio Lab
Dealing with Few to No Labels	Open in Colab	Open in Kaggle	Run on Gradient	Open Studio Lab
Training Transformers from Scratch	Open in Colab	Open in Kaggle	Run on Gradient	Open Studio Lab
Future Directions	Open in Colab	Open in Kaggle	Run on Gradient	Open Studio Lab

Nowadays, the GPUs on Colab tend to be K80s (which have limited memory), so we recommend using Kaggle, Gradient, or SageMaker Studio Lab. These platforms tend to provide more performant GPUs like P100s, all for free!

https://github.com/nlp-with-transformers/notebooks

Python in Google Colab (Python101)

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

python101.ipynb

File Edit View Insert Runtime Tools Help Last edited on May 13

Comment Share

Table of contents

+ Code + Text Connect Editing

```
text = "Steve Jobs and Steve Wozniak incorporated Apple Computer on January 3, 1977, in Cupertino, California."
doc = nlp(text)
displacy.render(doc, style="ent", jupyter=True)
```

Steve Jobs PERSON and Steve Wozniak PERSON incorporated Apple Computer ORG on January 3, 1977 DATE , in Cupertino GPE , California GPE .

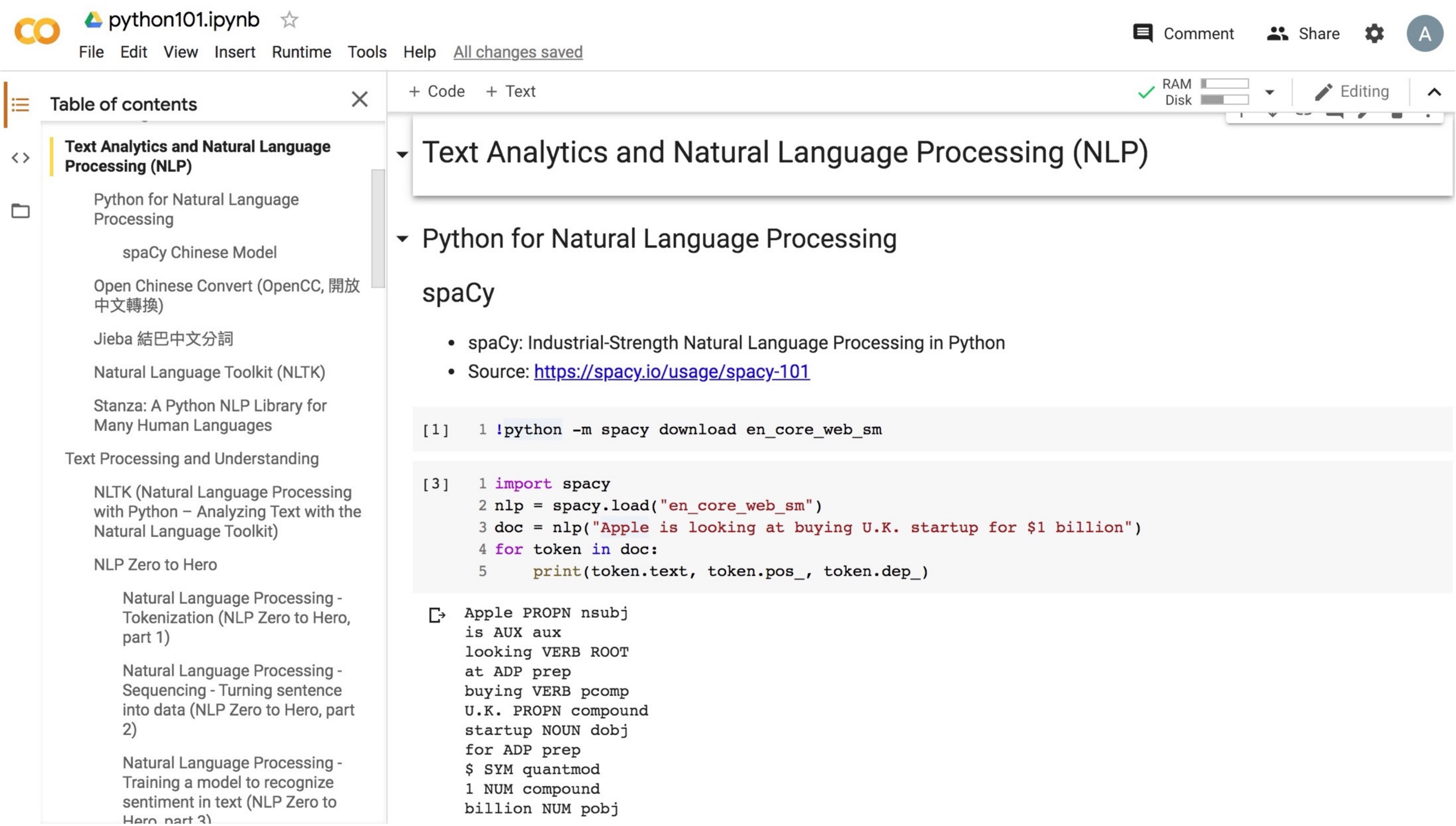
```
import spacy
nlp = spacy.load("en_core_web_sm")
doc = nlp("Stanford University is located in California. It is a great university.")
import pandas as pd
cols = ("text", "lemma", "pos", "tag", "pos_explain", "stopword")
rows = []
for t in doc:
    row = [t.text, t.lemma_, t.pos_, t.tag_,spacy.explain(t.pos_), t.is_stop]
    rows.append(row)
df = pd.DataFrame(rows, columns=cols)
df
```

	text	lemma	pos	tag	pos_explain	stopword
0	Stanford	Stanford	PROPN	NNP	proper noun	False
1	University	University	PROPN	NNP	proper noun	False
2	is	be	VERB	VBZ	verb	True
3	located	locate	VERB	VBN	verb	False
4	in	in	ADP	IN	adposition	True
5	California	California	PROPN	NNP	proper noun	False
6	.	.	PUNCT	.	punctuation	False
7	It	-PRON-	PRON	PRP	pronoun	True

https://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

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



https://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

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

python101.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

```
import spacy
nlp = spacy.load("en_core_web_sm")
doc = nlp("Apple is looking at buying U.K. startup for $1 billion")
import pandas as pd
cols = ("text", "lemma", "POS", "explain", "stopword")
rows = []
for t in doc:
    row = [t.text, t.lemma_, t.pos_, spacy.explain(t.pos_), t.is_stop]
    rows.append(row)
df = pd.DataFrame(rows, columns=cols)
df
```

	text	lemma	POS	explain	stopword
0	Apple	Apple	PROPN	proper noun	False
1	is	be	VERB	verb	True
2	looking	look	VERB	verb	False
3	at	at	ADP	adposition	True
4	buying	buy	VERB	verb	False
5	U.K.	U.K.	PROPN	proper noun	False
6	startup	startup	NOUN	noun	False
7	for	for	ADP	adposition	True
8	$	$	SYM	symbol	False
9	1	1	NUM	numeral	False
10	billion	billion	NUM	numeral	False

https://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

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

python101.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

```python
import spacy
nlp = spacy.load("en_core_web_sm")
doc = nlp("Stanford University is located in California. It is a great university.")
import pandas as pd
cols = ("text", "lemma", "POS", "explain", "stopword")
rows = []
for t in doc:
    row = [t.text, t.lemma_, t.pos_, spacy.explain(t.pos_), t.is_stop]
    rows.append(row)
df = pd.DataFrame(rows, columns=cols)
df
```

	text	lemma	POS	explain	stopword
0	Stanford	Stanford	PROPN	proper noun	False
1	University	University	PROPN	proper noun	False
2	is	be	VERB	verb	True
3	located	locate	VERB	verb	False
4	in	in	ADP	adposition	True
5	California	California	PROPN	proper noun	False
6	.	.	PUNCT	punctuation	False
7	It	-PRON-	PRON	pronoun	True
8	is	be	VERB	verb	True
9	a	a	DET	determiner	True
10	great	great	ADJ	adjective	False
11	university	university	NOUN	noun	False
12	.	.	PUNCT	punctuation	False

https://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

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

python101.ipynb

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

```
[ ]
import spacy
nlp = spacy.load("en_core_web_sm")
text = "Stanford University is located in California. It is a great university."
doc = nlp(text)
for ent in doc.ents:
    print(ent.text, ent.label_)
```

```
Stanford University ORG
California GPE
```

```
[ ]
from spacy import displacy
text = "Stanford University is located in California. It is a great university."
doc = nlp(text)
displacy.render(doc, style="ent", jupyter=True)
```

Stanford University **ORG** is located in California **GPE** . It is a great university.

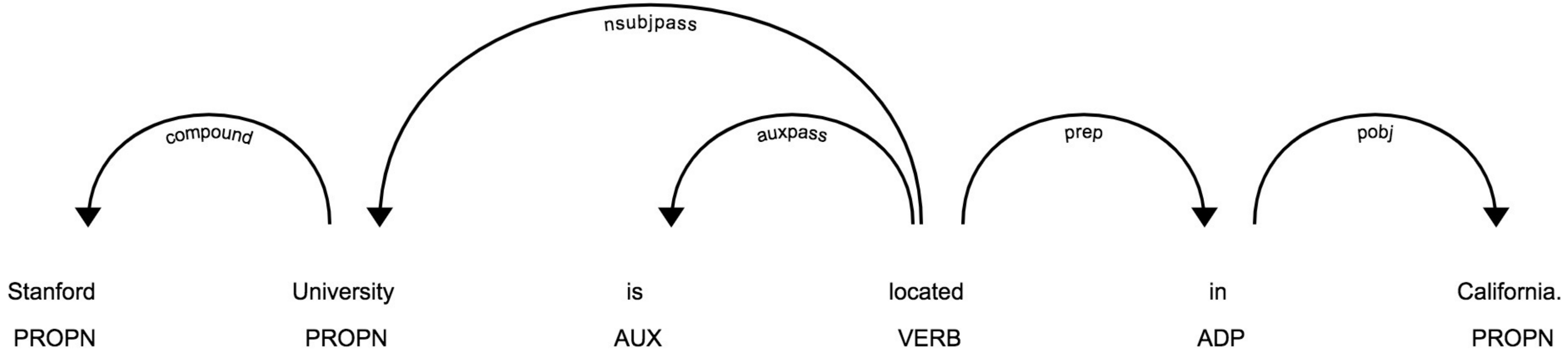
https://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

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

```python
from spacy import displacy
text = "Stanford University is located in California. It is a great university."
doc = nlp(text)
displacy.render(doc, style="ent", jupyter=True)
displacy.render(doc, style="dep", jupyter=True)
```

Stanford University **ORG** is located in California **GPE** . It is a great university.



https://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

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

python101.ipynb ☆

File Edit View Insert Runtime Tools Help Last edited on May 13

Comment Share A

Table of contents

+ Code + Text Connect Editing

```
text = "Steve Jobs and Steve Wozniak incorporated Apple Computer on January 3, 1977, in Cupertino, California."
doc = nlp(text)
displacy.render(doc, style="ent", jupyter=True)
```

Steve Jobs PERSON and Steve Wozniak PERSON incorporated Apple Computer ORG on January 3, 1977 DATE , in Cupertino GPE , California GPE .

```
import spacy
nlp = spacy.load("en_core_web_sm")
doc = nlp("Stanford University is located in California. It is a great university.")
import pandas as pd
cols = ("text", "lemma", "pos", "tag", "pos_explain", "stopword")
rows = []
for t in doc:
    row = [t.text, t.lemma_, t.pos_, t.tag_,spacy.explain(t.pos_), t.is_stop]
    rows.append(row)
df = pd.DataFrame(rows, columns=cols)
df
```

	text	lemma	pos	tag	pos_explain	stopword
0	Stanford	Stanford	PROPN	NNP	proper noun	False
1	University	University	PROPN	NNP	proper noun	False
2	is	be	VERB	VBZ	verb	True
3	located	locate	VERB	VBN	verb	False
4	in	in	ADP	IN	adposition	True
5	California	California	PROPN	NNP	proper noun	False
6	.	.	PUNCT	.	punctuation	False
7	It	-PRON-	PRON	PRP	pronoun	True

https://tinyurl.com/aintpupython101

Python in Google Colab (Python101)

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

python101.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

Table of contents

Semantic Analysis and Named Entity Recognition (NER)

- Source: 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

Semantic Analysis

[1]

```
import nltk
from nltk.corpus import wordnet as wn
import pandas as pd
nltk.download('wordnet')
# WordNet Synsets
word = 'fruit'
synsets = wn.synsets(word)
print('Word:', word)
print('Wordnet Synsets:', len(synsets))
df = pd.DataFrame([{'Synset': synset,
                    'Part of Speech': synset.lexname(),
                    'Definition': synset.definition(),
                    'Lemmas': synset.lemma_names(),
                    'Examples': synset.examples()}
                   for synset in synsets])
df
```

```
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Unzipping corpora/wordnet.zip.
Word: fruit
Wordnet Synsets: 5
```

	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/aintpupython101

Python in Google Colab (Python101)

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

python101.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

Multilingual Named Entity Recognition (NER)

- Source: Lewis Tunstall, Leandro von Werra, and Thomas Wolf (2022), Natural Language Processing with Transformers: Building Language Applications with Hugging Face, O'Reilly Media.
- Github: https://github.com/nlp-with-transformers/notebooks

```
#NER: https://huggingface.co/tasks/token-classification
!pip install transformers
from transformers import pipeline
classifier = pipeline("ner")
classifier("Hello I'm Omar and I live in Zürich.")
```

```
from transformers import pipeline
classifier = pipeline("ner")
classifier("Hello I'm Omar and I live in Zürich.")
```

```
No model was supplied, defaulted to dbmdz/bert-large-cased-finetuned-conll03-english (https://huggingface.co/dbmdz/bert-large-cased-finetuned-conll03-eng
[{'end': 14,
  'entity': 'I-PER',
  'index': 5,
  'score': 0.99770516,
  'start': 10,
  'word': 'Omar'},
 {'end': 35,
  'entity': 'I-LOC',
  'index': 10,
  'score': 0.9968976,
  'start': 29,
  'word': 'Zürich'}]
```

https://tinyurl.com/aintpupython101

Named Entity Recognition (NER)

```
from transformers import pipeline
import pandas as pd
classifier = pipeline("ner")
text = "My name is Michael and I live in Berkeley, California."
outputs = classifier(text)
pd.DataFrame(outputs)
```

	entity	score	index	word	start	end
0	I-PER	0.998874	4	Michael	11	18
1	I-LOC	0.997050	9	Berkeley	33	41
2	I-LOC	0.999170	11	California	43	53

https://tinyurl.com/aintpupython101

Summary

- Named Entities (NE)
  - represent real-world objects
  - people, places, organizations
  - proper names
- Named Entity Recognition (NER)
  - Entity chunking
  - Entity extraction
- Relation Extraction (RE)

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

References

- Lewis Tunstall, Leandro von Werra, and Thomas Wolf (2022), Natural Language Processing with Transformers: Building Language Applications with Hugging Face, O'Reilly Media.
- Denis Rothman (2021), Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more, Packt Publishing.
- Savaş Yıldırım and Meysam Asgari-Chenaghlu (2021), Mastering Transformers: Build state-of-the-art models from scratch with advanced natural language processing techniques, Packt Publishing.
- Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta (2020), Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems, O'Reilly Media.
- Jing Li, Aixin Sun, Jianglei Han, and Chenliang LI (2022). "A survey on deep learning for named entity recognition." IEEE Transactions on Knowledge and Data Engineering 34, no. 1 (2022): 50-70.
- Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik (2022). "Named entity recognition and relation extraction: State-of-the-art." ACM Computing Surveys (CSUR) 54, no. 1 (2022): 1-39.
- Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li (2022). "Chinese named entity recognition: The state of the art." Neurocomputing 473 (2022): 37-53.
- Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang (2023). "Gpt-ner: Named entity recognition via large language models." arXiv preprint arXiv:2304.10428 (2023).
- Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson.
- Dipanjan Sarkar (2019), Text Analytics with Python: A Practitioner’s Guide to Natural Language Processing, Second Edition. APress.
- Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda (2018), Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning, O’Reilly.
- Charu C. Aggarwal (2018), Machine Learning for Text, Springer.
- Gabe Ignatow and Rada F. Mihalcea (2017), An Introduction to Text Mining: Research Design, Data Collection, and Analysis, SAGE Publications.
- Rajesh Arumugam (2018), Hands-On Natural Language Processing with Python: A practical guide to applying deep learning architectures to your NLP applications, Packt.
- Jake VanderPlas (2016), Python Data Science Handbook: Essential Tools for Working with Data, O'Reilly Media.
- 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.
- The Super Duper NLP Repo, https://notebooks.quantumstat.com/
- Jay Alammar (2018), The Illustrated Transformer, http://jalammar.github.io/illustrated-transformer/
- Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/
- NLP with Transformer, https://github.com/nlp-with-transformers/notebooks
- Min-Yuh Day (2023), Python 101, https://tinyurl.com/aintpupython101