ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing
Mark Neumann, Daniel King, Iz Beltagy, Waleed Ammar
9 Oct 2019

Vladislav Panferov

Novosibirsk State University

v.panferov@g.nsu.ru

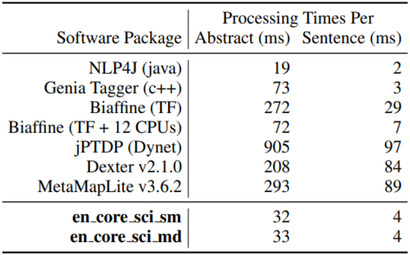
March 16, 2021

Overview

1. Introduction
2. Definitions
3. POS
4. Named Entity Recognition
5. Smart thought

Introduction

ScispaCy is a Python package containing spaCy models for processing biomedical, scientific or clinical text.
SpaCy is a free open-source library for Natural Language Processing in Python. It features NER, POS tagging, dependency parsing, word vectors and more.



	Processing Times Per	
Software Package	Abstract (ms)	Sentence (ms)
NLP4J (java)	19	2
Genia Tagger (c++)	73	3
Biaffine (TF)	272	29
Biaffine (TF + 12 CPUs)	72	7
jPTDP (Dynet)	905	97
Dexter v2.1.0	208	84
MetaMapLite v3.6.2	293	89
en_core_sci_sm	32	4
en_core_sci_md	33	4

Figure: Wall clock comparison of different publicly available biomedical NLP pipelines. All experiments run on a single machine with 12 Intel(R) Core(TM) i7-6850K CPU @ 3.60GHz and 62GB RAM.

Named Entity Recognition

Named Entity Recognition (NER) - is the task of identifying and categorizing key information (entities) in text.

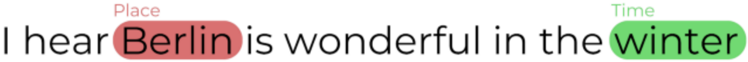


Figure: Named Entity Recognition.

Part-of-speech tagging

Part-of-speech (POS) tagging - is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context.

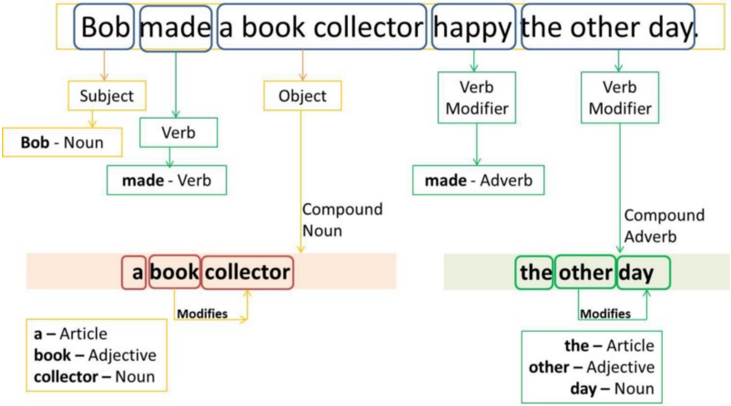


Figure: Part-of-speech tagging.

Dependency parsing

Dependency parsing - is the process of analyzing the grammatical structure of a sentence based on the dependencies between the words in a sentence.

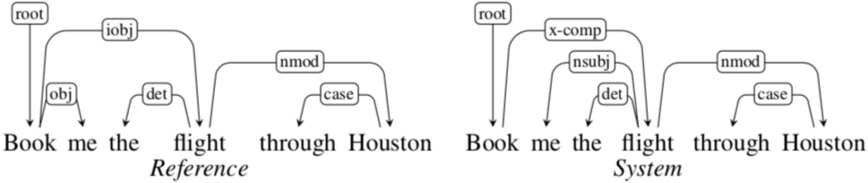

root
iobj
obj
det
nmod
case
Book me the flight through Houston
Reference
root
x-comp
nsubj
det
nmod
case
Book me the flight through Houston
System

Figure: Dependency parsing.

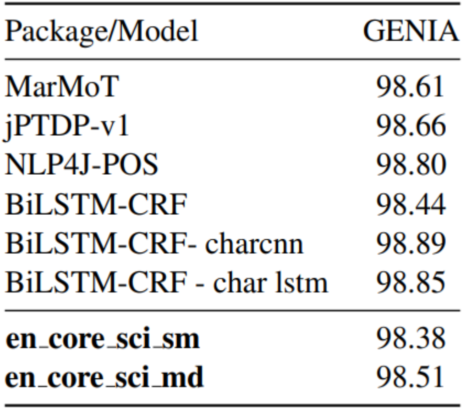
How it Works - POS

The joint POS tagging and dependency parsing model in **spaCy** is an arc-eager transition-based parser trained with a dynamic oracle, similar to Goldberg and Nivre (2012). Features are CNN representations of token features and shared across all pipeline models.

To increase the robustness of the dependency parser and POS tagger to generic text, they make use of the **OntoNotes 5.0** corpus when training the dependency parser and part of speech tagger.

The **OntoNotes** corpus consists of multiple genres of text, annotated with syntactic and semantic information, but they only use POS and dependency parsing annotations in this work.

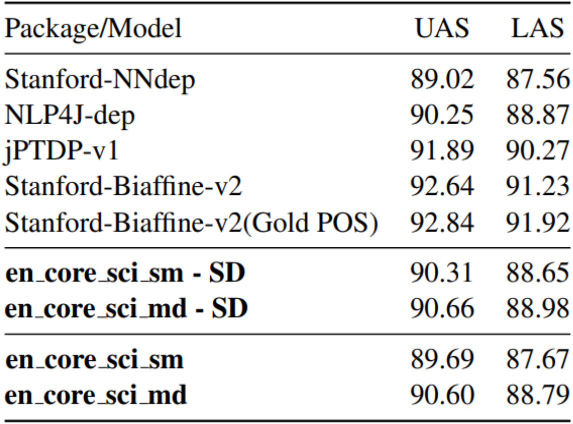
GENIA Experiment



Package/Model	GENIA
MarMoT	98.61
jPTDP-v1	98.66
NLP4J-POS	98.80
BiLSTM-CRF	98.44
BiLSTM-CRF- charcnn	98.89
BiLSTM-CRF - char lstm	98.85
en_core_sci_sm	98.38
en_core_sci_md	98.51

Figure: Part of Speech tagging results on the GENIA Test set.

GENIA Experiment



Package/Model	UAS	LAS
Stanford-NNdep	89.02	87.56
NLP4J-dep	90.25	88.87
jPTDP-v1	91.89	90.27
Stanford-Biaffine-v2	92.64	91.23
Stanford-Biaffine-v2(Gold POS)	92.84	91.92
en_core_sci_sm - SD	90.31	88.65
en_core_sci_md - SD	90.66	88.98
en_core_sci_sm	89.69	87.67
en_core_sci_md	90.60	88.79

Figure: Dependency Parsing results on the GENIA 1.0 corpus converted to dependencies using the Stanford Universal Dependency Converter. They additionally provide evaluations using Stanford Dependencies(SD) in order for comparison relative to the results reported in (Nguyen and Verspoor, 2018).

Robustness Experiment

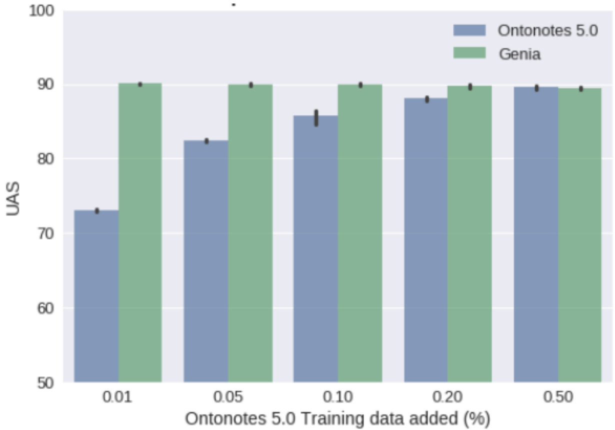


Figure: Unlabeled attachment score (UAS) performance for a model trained with increasing amounts of web data incorporated. Table shows mean of 3 random seeds.

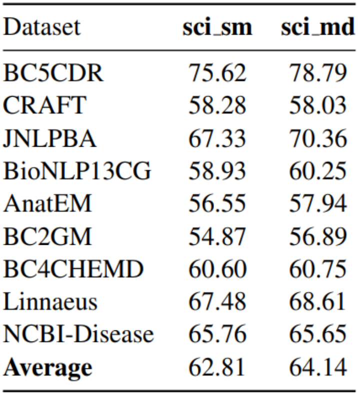
How it Works - Named Entity Recognition

The NER model in spaCy is a transition-based system based on the chunking model from Lample et al. (2016). Tokens are represented as hashed, embedded representations of the prefix, suffix, shape and lemmatized features of individual words.

The main NER model in both released packages in **scispaCy** is trained on the mention spans in the **MedMentions** dataset.
In order to provide for users with more specific requirements around entity types, they release four additional packages with finer-grained NER models trained on:

- **BC5CDR** - for chemicals and diseases;
- **CRAFT** - for cell types, chemicals, proteins, genes;
- **JNLPBA** - for cell lines, cell types, DNAs, RNAs, proteins;
- **BioNLP13CG** - for cancergenetics;

Experiment



Dataset	**sci_sm**	**sci_md**
BC5CDR	75.62	78.79
CRAFT	58.28	58.03
JNLPBA	67.33	70.36
BioNLP13CG	58.93	60.25
AnatEM	56.55	57.94
BC2GM	54.87	56.89
BC4CHEMD	60.60	60.75
Linnaeus	67.48	68.61
NCBI-Disease	65.76	65.65
Average	62.81	64.14

Figure: Recall on the test sets of 9 specialist NER datasets, when the base mention detector is trained on **MedMentions**.

Smart thought

Greeted by clothes.

? Vstrechayut po odezhke.

References

📄 Mark Neumann, Daniel King, Iz Beltagy, Waleed Ammar (9 Oct 2019)
ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing

The End