Integrating Polish Language Tools and Resources in **spaCy**

Ryszard Tuora, Łukasz Kobyliński

spaCy

is a popular, open-source Python framework for NLP. It aims at ease of access and breadth of possible applications.

Text processing

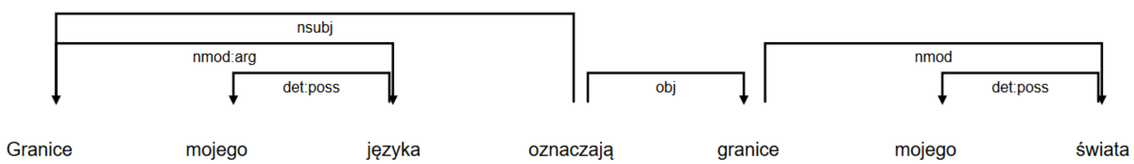
- Tokenization
- POS Tagging
- Lemmatization
- Dep. Parsing
- NER

Usage:

```
>>> import pandas # for visualization
>>> import spacy
>>> nlp = spacy.load("pl_spacy_model")

# Tokenization, Tagging, Lemmatization and Dependency Parsing
>>> sent1 = "Granice mojego języka oznaczają granice mojego świata" # ~Wittgenstein
>>> parse1 = nlp(sent1)
>>> attribs = ['orth_', 'lemma_', 'tag_', 'pos_', 'dep_', 'head']
>>> table = [{att:tok.__getattribute__(att) for att in attribs} for tok in parse1]
>>> df = pandas.DataFrame(table)
>>> print(df[attribs])
```

	orth_	lemma_	tag_	pos_	dep_	head
0	Granice	granica	SUBST	NOUN	nsubj	oznaczają
1	mojego	mój	ADJ	ADJ	det:poss	języka
2	języka	język	SUBST	NOUN	nmod:arg	Granice
3	oznaczają	oznaczać	FIN	VERB	ROOT	oznaczają
4	granice	granica	SUBST	NOUN	obj	oznaczają
5	mojego	mój	ADJ	ADJ	det:poss	świata
6	świata	świat	SUBST	NOUN	nmod	granice



NER: **87.52** F-score on a test subset of NKJP

```
# NER
>>> sent2 = "Selekcjoner reprezentacji Polski Jerzy Brzęczek \
        powołał piłkarzy na mecze eliminacji mistrzostw Europy 2020,\
        6 września Polacy zagrają w Lublanie ze Słowenią." # ~Rzeczpospolita
>>> parse2 = nlp(sent2)
```

Selekcjoner reprezentacji Polski PLACENAME Jerzy PERSNAME Brzęczek PERSNAME powołał piłkarzy na mecze eliminacji mistrzostw Europy GEOGNAME 2020 DATE , 6 września DATE Polacy PLACENAME zagrają w Lublanie PLACENAME ze Słowenią PLACENAME .

Two versions

- A default one, without any external dependencies, here labelled as **IPI PAN**
- A version utilizing **Morfeusz 2** for tokenization, tagging and lemmatization

The future

- Further optimization
- Additional components (e.g. a chunker, sentiment analysis component)
- Models of different sizes (e.g. a 20 MB model for quick and easy tasks)
- Better integration with Morfeusz
- Tell us!

Evaluation

We've evaluated our two models, and a previous model by Sigmoidal, against the three Polish treebanks available in UD. These results list scores on tokenization, NKJP (XPOS) and UD tagset POS tagging, lemmatization, and UAS and LAS for dependency parsing.

	name	Tokens f1	XPOS f1	UPOS f1	Lemma f1	UAS f1	LAS f1
IPI PAN	lfg_test	96.7%	96.0%	83.2%	90.7%	84.3%	79.3%
	pdb_test	98.8%	90.9%	83.9%	91.0%	86.1%	83.2%
	pud_test	98.2%	90.9%	78.5%	87.7%	85.9%	82.6%
+MORFEUSZ	lfg_test	99.5%	98.9%	86.1%	94.6%	90.6%	85.4%
	pdb_test	98.3%	92.7%	85.9%	93.7%	90.7%	87.8%
	pud_test	99.2%	92.6%	80.3%	93.4%	89.8%	85.8%
SIGMOIDAL	lfg_test	96.7%	95.4%	82.8%	73.9%	85.9%	83.2%
	pdb_test	98.8%	90.4%	83.6%	72.7%	75.1%	68.0%
	pud_test	98.2%	90.7%	78.3%	69.5%	74.4%	66.4%

Institute of Computer Science, Polish Academy of Sciences
Linguistic Engineering Group

zil.ipipan.waw.pl