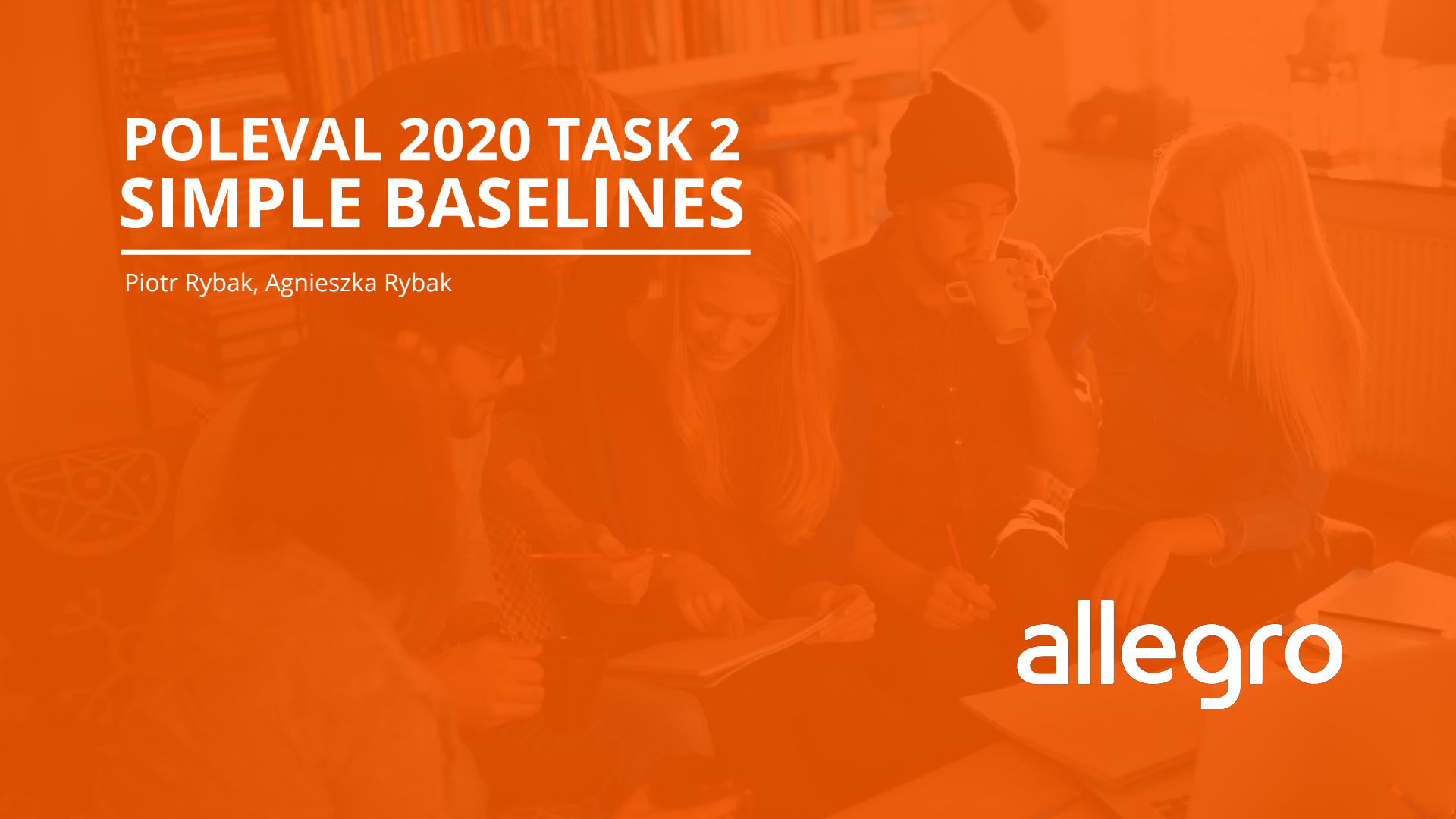


POLEVAL 2020 TASK 2

SIMPLE BASELINES

Piotr Rybak, Agnieszka Rybak

The background of the slide is a photograph of four students (three women and one man) sitting at a table in a library or study hall, working together on papers. The image is overlaid with a semi-transparent orange filter. The students are focused on their work, with one man drinking from a cup. Bookshelves filled with books are visible in the background.

allegro

Dane

8	9	Na	na	interj	
8	9	Na	na	prep:acc	disamb
9	10	pojedynek	pojedynek	subst:sg:acc:m	disamb
9	10	pojedynek	pojedynek	subst:sg:nom:m	
9	10	pojedynek	pojedynka	subst:pl:gen:f	
10	11	konny	konny	adj:sg:acc:m:pos	disamb_manual
10	11	konny	konny	ign	
11	12	wyzwali	wyzwać	praet:pl:manim1:perf	disamb
12	13	się	się	part	disamb
13	14	harc	harc	subst:sg:acc:m	
13	15	harce	harc	subst:pl:acc:m	
13	16	harcem	harc	subst:sg:inst:m	disamb
14	16	em	być	aglt:sg:pri:imperf:wok	
15	16	m	być	aglt:sg:pri:imperf:nwok	
16	17	.	.	interp	disamb

Cel

8	9	Na	na	interj	
8	9	Na	na	prep:acc	disamb
9	10	pojedynek	pojedynek	subst:sg:acc:m	disamb
9	10	pojedynek	pojedynek	subst:sg:nom:m	
9	10	pojedynek	pojedynka	subst:pl:gen:f	
10	11	konny	konny	adj:sg:acc:m:pos	disamb_manual
10	11	konny	konny	ign	
11	12	wyzwali	wyzwać	praet:pl:manim1:perf	disamb
12	13	się	się	part	disamb
13	14	harc	harc	subst:sg:acc:m	
13	15	harce	harc	subst:pl:acc:m	
13	16	harcem	harc	subst:sg:inst:m	disamb
14	16	em	być	aglt:sg:pri:imperf:wok	
15	16	m	być	aglt:sg:pri:imperf:nwok	
16	17	.	.	interp	disamb

Klasyfikacja

8	9	Na	na	interj	
8	9	Na	na	prep:acc	disamb
9	10	pojedynek	pojedynek	subst:sg:acc:m	disamb
9	10	pojedynek	pojedynek	subst:sg:nom:m	
9	10	pojedynek	pojedynka	subst:pl:gen:f	
10	11	konny	konny	adj:sg:acc:m:pos	disamb_manual
10	11	konny	konny	ign	
11	12	wyzwali	wyzwać	praet:pl:manim1:perf	disamb
12	13	się	się	part	disamb
13	14	harc	harc	subst:sg:acc:m	
13	15	harce	harc	subst:pl:acc:m	
13	16	harcem	harc	subst:sg:inst:m	disamb
14	16	em	być	aglt:sg:pri:imperf:wok	
15	16	m	być	aglt:sg:pri:imperf:nwok	
16	17	.	.	interp	disamb

Brak interpretacji

8	9	Na	na	interj	
8	9	Na	na	prep:acc	disamb
9	10	pojedynek	pojedynek	subst:sg:acc:m	disamb
9	10	pojedynek	pojedynek	subst:sg:nom:m	
9	10	pojedynek	pojedynka	subst:pl:gen:f	
10	11	konny	konny	adj:sg:acc:m:pos	disamb_manual
10	11	konny	konny	ign	
11	12	wyzwali	wyzwać	praet:pl:manim1:perf	disamb
12	13	się	się	part	disamb
13	14	harc	harc	subst:sg:acc:m	
13	15	harce	harc	subst:pl:acc:m	
13	16	harcem	harc	subst:sg:inst:m	disamb
14	16	em	być	aglt:sg:pri:imperf:wok	
15	16	m	być	aglt:sg:pri:imperf:nwok	
16	17	.	.	interp	disamb

Tagowanie

8	9	Na	na	interj	
8	9	Na	na	prep:acc	disamb
9	10	pojedynek	pojedynek	subst:sg:acc:m	disamb
9	10	pojedynek	pojedynek	subst:sg:nom:m	
9	10	pojedynek	pojedynka	subst:pl:gen:f	
10	11	konny	konny	adj:sg:acc:m:pos	disamb_manual
10	11	konny	konny	ign	
11	12	wyzwali	wyzwać	praet:pl:manim1:perf	disamb
12	13	się	się	part	disamb
13	14	harc	harc	subst:sg:acc:m	
13	15	harce	harc	subst:pl:acc:m	
13	16	harcem	harc	subst:sg:inst:m	disamb
14	16	em	być	aglt:sg:pri:imperf:wok	
15	16	m	być	aglt:sg:pri:imperf:nwok	
16	17	.	.	interp	disamb

Segmentacja

8	9	Na	na	interj	
8	9	Na	na	prep:acc	disamb
9	10	pojedynek	pojedynek	subst:sg:acc:m	disamb
9	10	pojedynek	pojedynek	subst:sg:nom:m	
9	10	pojedynek	pojedynka	subst:pl:gen:f	
10	11	konny	konny	adj:sg:acc:m:pos	disamb_manual
10	11	konny	konny	ign	
11	12	wyzwali	wyzwać	praet:pl:manim1:perf	disamb
12	13	się	się	part	disamb
13	14	harc	harc	subst:sg:acc:m	
13	15	harce	harc	subst:pl:acc:m	
13	16	harcem	harc	subst:sg:inst:m	disamb
14	16	em	być	aglt:sg:pri:imperf:wok	
15	16	m	być	aglt:sg:pri:imperf:nwok	
16	17	.	.	interp	disamb

Konwersja

13	14	harc	harc	subst:sg:acc:m	
13	15	harce	harc	subst:pl:acc:m	
13	16	harcem	harc	subst:sg:inst:m	disamb
14	16	em	być	aglt:sg:pri:imperf:wok	
15	16	m	być	aglt:sg:pri:imperf:nwok	

Konwersja

13	14	harc	harc	subst:sg:acc:m	
13	15	harce	harc	subst:pl:acc:m	
13	16	harcem	harc	subst:sg:inst:m	disamb
14	16	em	być	aglt:sg:pri:imperf:wok	
15	16	m	być	aglt:sg:pri:imperf:nwok	



```
tokens = [  
    ('harc', 'subst:sg:inst:m'),  
    ('e', 'subst:sg:inst:m'),  
    ('m', 'subst:sg:inst:m'),  
]
```

Konwersja

13	14	harc	harc	subst:sg:acc:m	
13	15	harce	harc	subst:pl:acc:m	
13	16	harcem	harc	subst:sg:inst:m	disamb
14	16	em	być	aglt:sg:pri:imperf:wok	
15	16	m	być	aglt:sg:pri:imperf:nwok	



```
tokens = [  
    ('harc', 'subst:sg:inst:m'),  
    ('e', 'subst:sg:inst:m'),  
    ('m', 'subst:sg:inst:m'),  
]
```

Predykcja

```
tokens = [  
    ('harc', 'subst:sg:inst:m'),  
    ('e', 'subst:sg:inst:m'),  
    ('m', 'subst:sg:inst:m'),  
]
```



13 16 **harcem** harc **subst:sg:inst:m**

Predykcja

```
tokens = [  
    ('harc', 'subst:sg:acc:m'),  
    ('e', 'aglt:sg:pri:imperf:wok'),  
    ('m', 'aglt:sg:pri:imperf:wok'),  
]
```



13	14	harc
14	16	em

harc
być

subst:sg:acc:m
aglt:sg:pri:imperf:wok

COMBO

360er0 / COMBO

Watch 3Star 31Fork 7

CodeIssues 2Pull requests 1ActionsProjectsSecurityInsights

master2 branches0 tagsGo to fileCode

Piotr Rybak fix loading pickled modelscdf9feb on 25 Jun34 commits

.gitignore	Initial commit	2 years ago
README.md	Update README.md	2 years ago
encoders.py	refactor summaries + imports	17 months ago
main.py	fix loading pickled models	4 months ago
models.py	refactor summaries + imports	17 months ago
mst.py	init commit	2 years ago
parser.py	store targets in sparse matrix	17 months ago
requirements.txt	fix loading pickled models	4 months ago
utils.py	fix tokenization	17 months ago

README.md

COMBO

COMBO is jointly trained neural tagger, lemmatizer and dependency parser implemented in python 3 using Keras framework. It took part in [2018 CoNLL Universal Dependency shared task](#) and ranked 3rd/4th in the [official evaluation](#).

Paper

The COMBO description can be found here: [Semi-Supervised Neural System for Tagging, Parsing and Lemmatization](#).

About

COMBO is jointly trained tagger, lemmatizer and dependency parser.

[lemmatizer](#)
[dependency-parser](#)
[tagger](#)
[keras](#)
[universal-dependencies](#)

Readme

Releases

No releases published

Packages

No packages published

Contributors 2

360er0

erickrf Erick Rocha Fons...

Languages

Python 100.0%

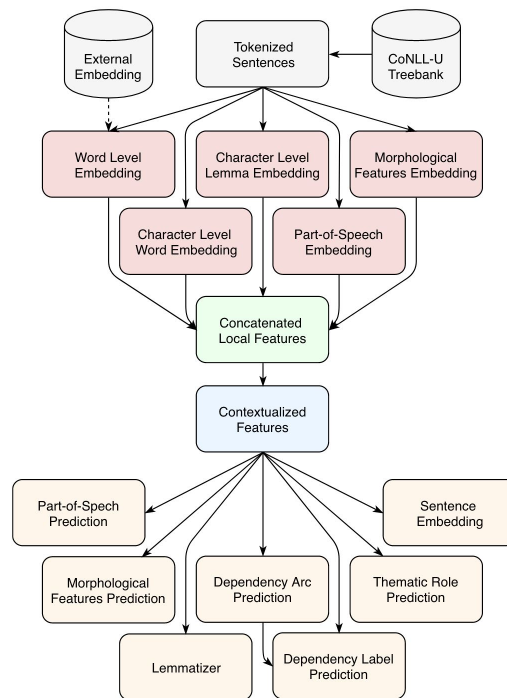
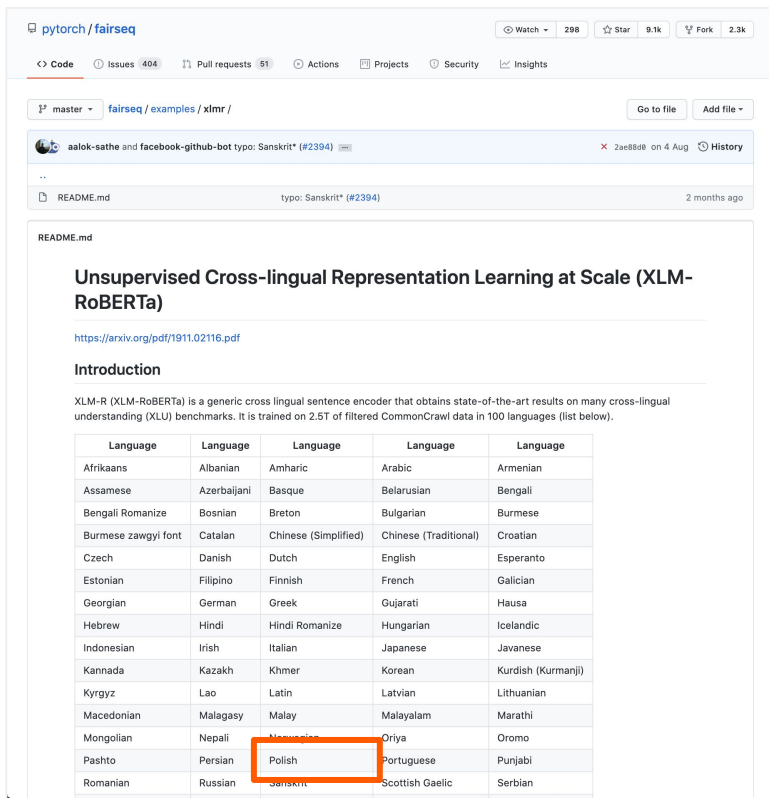


Figure 1: The schema of COMBO architecture.

Piotr Rybak and Alina Wróblewska, *Semi-Supervised Neural System for Tagging, Parsing and Lemmatization*, 2018

XLM-RoBERTa



The screenshot shows the GitHub repository for XLM-RoBERTa. The README file is open, displaying the title "Unsupervised Cross-lingual Representation Learning at Scale (XLM-RoBERTa)" and a link to the arXiv paper. Below the introduction, there is a table listing the 100 languages used in the training corpus. The table is organized in five columns, and the language "Polish" is highlighted with a red box in the fourth column.

Unsupervised Cross-lingual Representation Learning at Scale (XLM-RoBERTa)

<https://arxiv.org/pdf/1911.02116.pdf>

Introduction

XLM-R (XLM-RoBERTa) is a generic cross lingual sentence encoder that obtains state-of-the-art results on many cross-lingual understanding (XLU) benchmarks. It is trained on 2.5T of filtered CommonCrawl data in 100 languages (list below).

Language	Language	Language	Language	Language
Afrikaans	Albanian	Amharic	Arabic	Armenian
Assamese	Azerbaijani	Basque	Belarusian	Bengali
Bengali Romanize	Bosnian	Breton	Bulgarian	Burmese
Burmese Zawgyi font	Catalan	Chinese (Simplified)	Chinese (Traditional)	Croatian
Czech	Danish	Dutch	English	Esperanto
Estonian	Filipino	Finnish	French	Galician
Georgian	German	Greek	Gujarati	Hausa
Hebrew	Hindi	Hindi Romanize	Hungarian	Icelandic
Indonesian	Irish	Italian	Japanese	Javanese
Kannada	Kazakh	Khmer	Korean	Kurdish (Kurmanji)
Kyrgyz	Lao	Latin	Latvian	Lithuanian
Macedonian	Malagasy	Malay	Malayalam	Marathi
Mongolian	Nepali	Nepali	Oriya	Oromo
Pashto	Persian	Polish	Portuguese	Punjabi
Romanian	Russian	Sanskrit	Scottish Gaelic	Serbian

Model

RoBERTa Base & Large

Korpus

Common Crawl

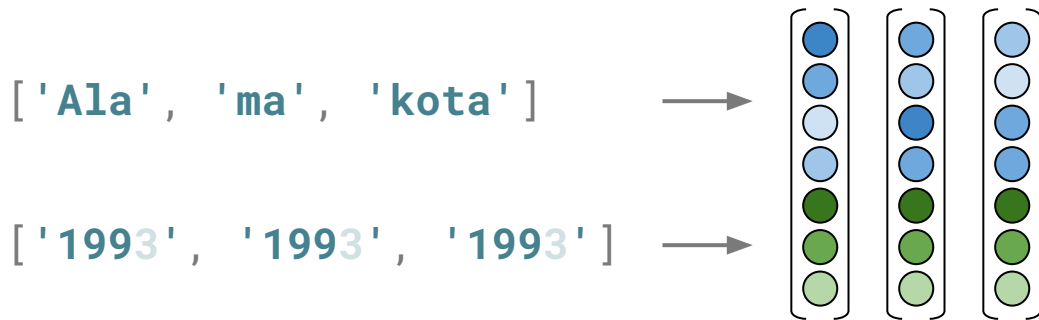
Tokenizacja

['Słō', 'wa', 'cki', 'wielki',
'm', 'poet', 'ą', 'był']

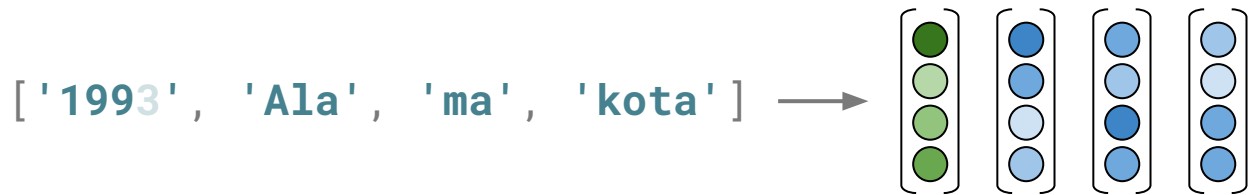
Alexis Conneau et al., *Unsupervised Cross-lingual Representation Learning at Scale*, 2020

Model	Accuracy			
	overall	known	unknown	manual not ign
COMBO	92.82	93.59	56.87	57.66

Model	Accuracy			
	overall	known	unknown	manual not ign
COMBO	92.82	93.59	56.87	57.66
COMBO + Word-level Augmentation	93.08	93.88	55.67	57.52



Model	Accuracy			
	overall	known	unknown	manual not ign
COMBO	92.82	93.59	56.87	57.66
COMBO + Word-level Augmentation	93.08	93.88	55.67	57.52
COMBO + Document-level Augmentation	92.99	93.78	55.79	57.80



Model	Accuracy			
	overall	known	unknown	manual not ign
COMBO	92.82	93.59	56.87	57.66
COMBO + Word-level Augmentation	93.08	93.88	55.67	57.52
COMBO + Document-level Augmentation	92.99	93.78	55.79	57.80
XLM-RoBERTa	94.58	94.84	82.68	77.90

DZIĘKUJĘ!

piotr.rybak@allegro.pl

Poszukiwany doświadczony
Research Engineer

- Ciekawe, różnorodne i ambitne projekty
- 20% czasu na badania naukowe
- Cotygodniowe seminaria
- Jeździmy na topowe konferencje (ACL, ICLR, CVPR, NeurIPS, etc.)

Więcej na: smrtr.io/3CcqD

allegro