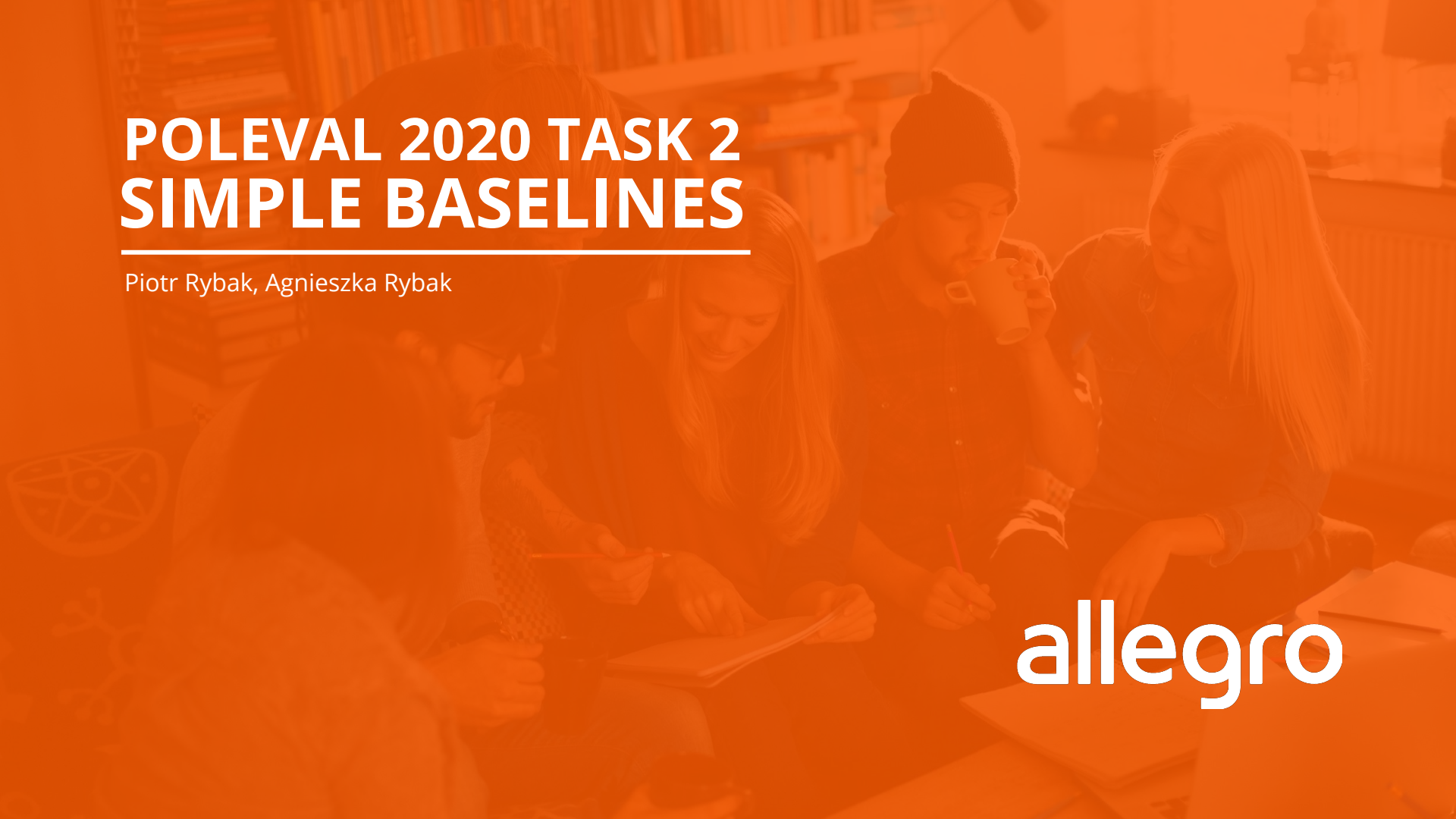
POLEVAL 2020 TASK 2
SIMPLE BASELINES

Piotr Rybak, Agnieszka Rybak



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'),
]
```

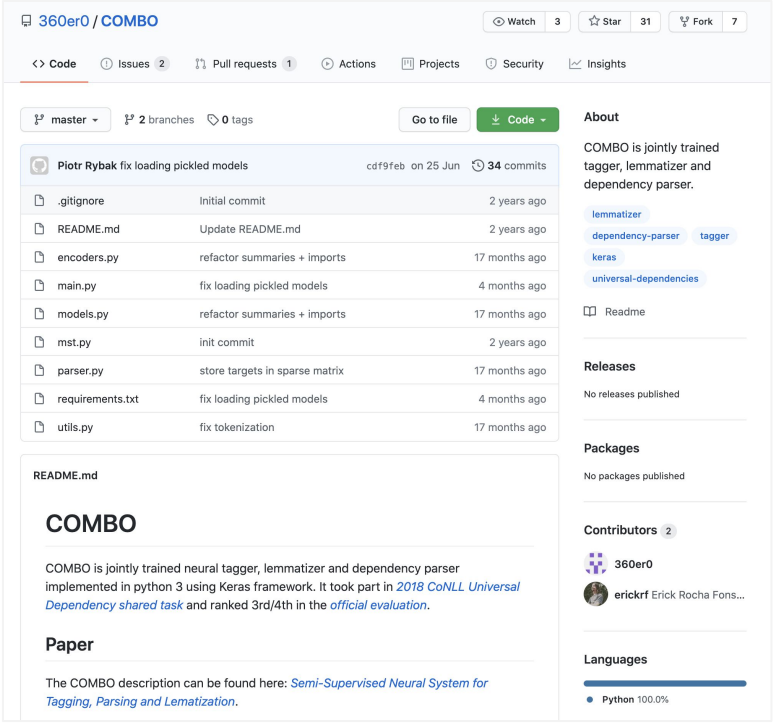
13 16 **harcem** harc **subst:sg:inst:m**

Predykcja

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

13	14	harc	harc	subst:sg:acc:m
14	16	em	być	aglt:sg:pri:imperf:wok

COMBO



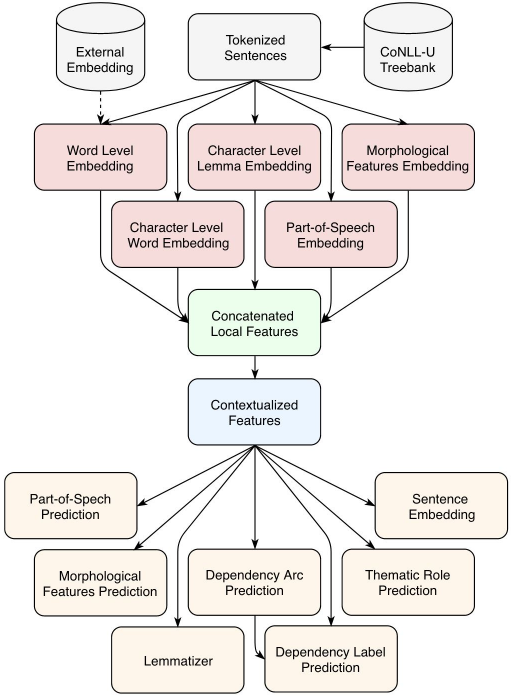
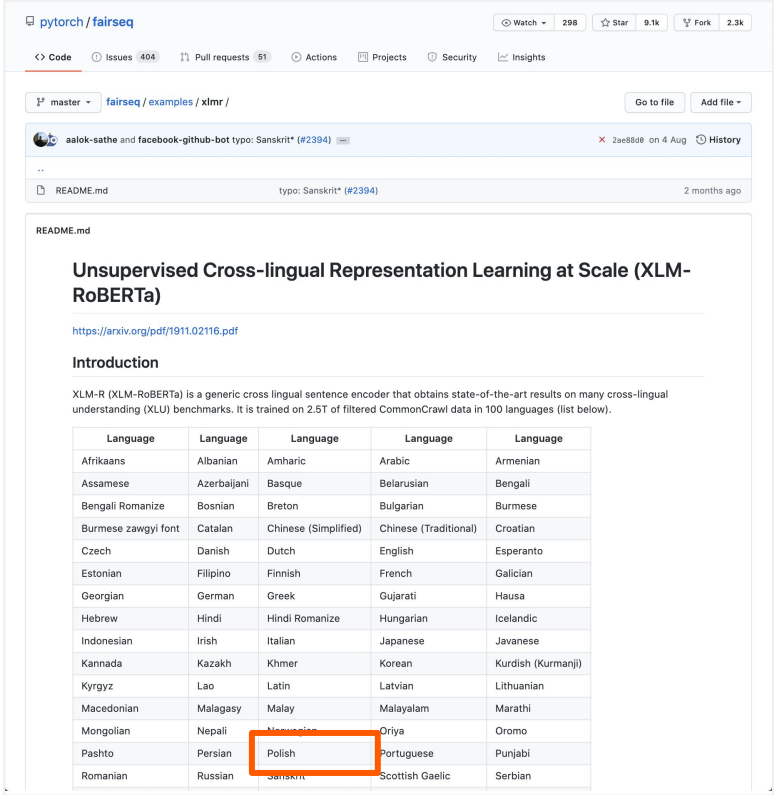


Figure 1: The schema of COMBO architecture.

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

XLM-RoBERTa



pytorch / fairseq

Watch 298 · Star 9.1k · Fork 2.3k

Code · Issues 404 · Pull requests 51 · Actions · Projects · Security · Insights

master · fairseq / examples / xlmr /

Go to file · Add file

aalok-sathe and facebook-github-bot typo: Sanskrit* (#2394) 2ae88d0 on 4 Aug · History

README.md typo: Sanskrit* (#2394) 2 months ago

README.md

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	Norwegian	Oriya	Oromo
Pashto	Persian	Polish	Portuguese	Punjabi
Romanian	Russian	Sanskrit	Scottish Gaelic	Serbian

Model

RoBERTa Base & Large

Korpus

Common Crawl

Tokenizacja

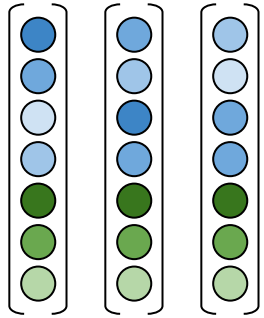
```
['Sło', 'wa', 'cki', 'wielki',
'm', 'poet', 'ą', 'był']
```

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

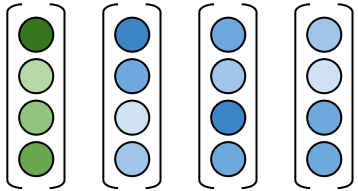
Model	Accuracy			
	overall	known	unknown	manual not ign
COMBO	92.82	93.59	56.87	57.66
COMBO + Word-level Augmentation	93.08	93.88	55.67	57.52

```
['Ala', 'ma', 'kota']
['1993', '1993', '1993']
```



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

```
['1993', 'Ala', 'ma', 'kota']
```



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