

Vrednotenje jezikovnih tehnologij in SloBENCH

Slavko Žitnik

Marko Robnik-Šikonja, Simon Krek, Frenk Dragar

Jota, 21. september 2023

Vsebina

- Pregled načinov vrednotenja
- Predstavitev SloBENCH
- Primer izdelave novega vrednotnika v SloBENCH

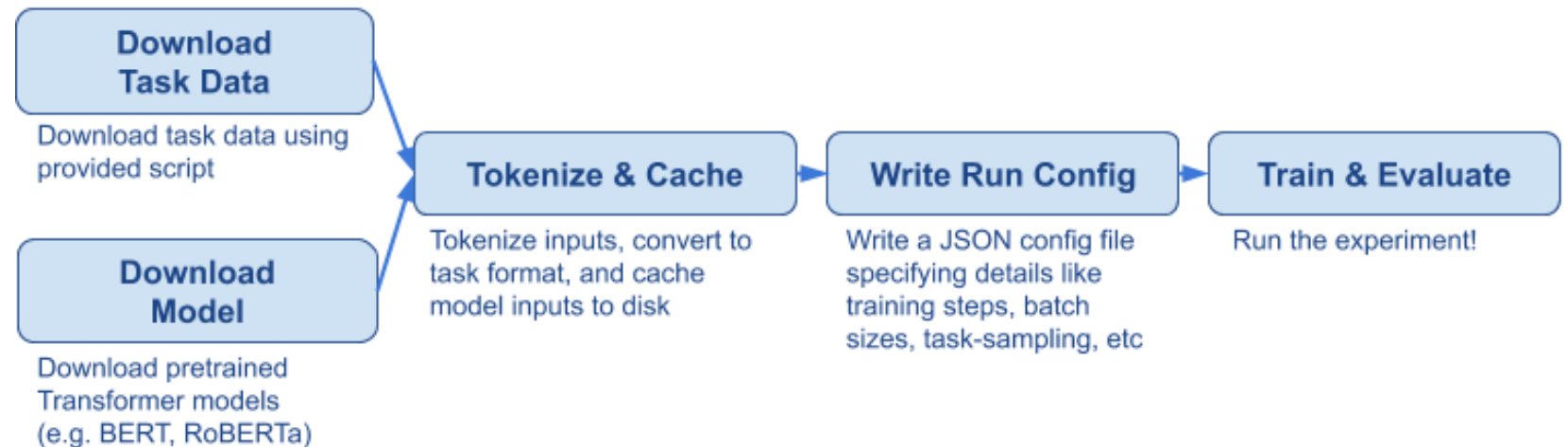
Pregled načinov vrednotenja

Razvoj vrednotnikov

- Testni korpus se lahko uporabi le 1x, sicer prilagajanje.
- Bogata zgodovina kampanj za neodvisno vrednotenje v ONJ/UI.
- Naslavljanje izzivov:
 - primerjalniki *slabijo* tekom časa
 - korpusi vsebujejo napake, pristranosti, ...
 - raziskovalci prekomerno prilagajajo modele
 - rezultati primerjav so lahko zavajajoči
 - vzpostavitev cikla izboljšav namesto zamenjav korpusov

Tipi vrednotnikov

- Primerjalniki
 - Podatki
 - Odprtost in transparentnost
 - Način oddaje (oznake, model, repozitorij, zahtevani podatki)
 - Omejitve (knjižnice, # oddaj / časovno enoto, format podatkov)
- Podporna ogrodja
 - *jiant*




Primeri primerjalnikov

- **GLUE (2018)**
- SuperGLUE (2019)
- Russian SuperGLUE (2020)
- XTREME (2020)
- XGlue (2020)
- **SQuAD (2018)**
- **dynabench (2020)**
- **explainaboard (2021)**
- decaNLP (2018)
- GEM benchmark (2021)
- GENIE benchmark (2021)
- **Allen AI Leaderboards (2021)**
- babuska-bench (2019)
- klej-benchmark (2020)
- nlpprogress.com (2018)
- **paperswithcode.com (Meta)**
- Therapeutics Data Commons (2021)
- CodaBench (2023, *CodaLab*)
- HuggingFace evaluation (2023?)

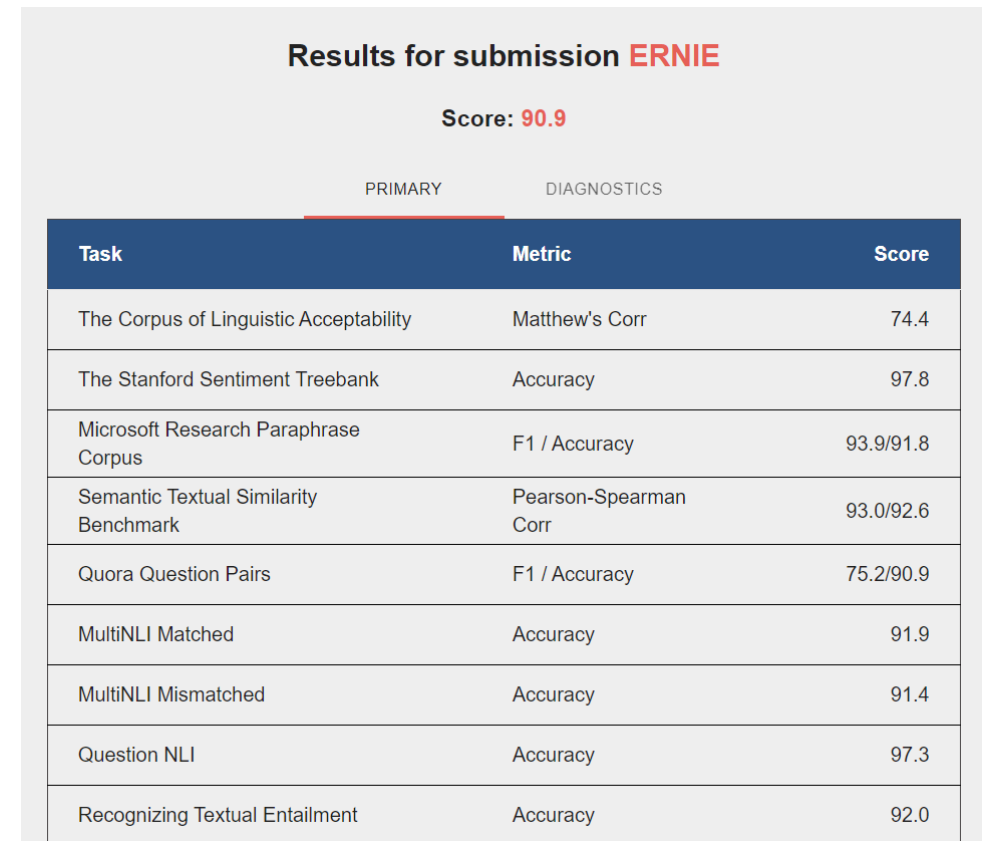
GLUE (2018)

- General Language Understanding Evaluation
- Agnostično glede na model (*jiant*)
- Vrednotenje, primerjava, analiza



GLUE SuperGLUE Paper </> Code Tasks Leaderboard FAQ Diagnostics Submit Login

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	M	
1	ERNIE Team - Baidu	ERNIE	🔗	90.9	74.4	97.8	93.9/91.8	93.0/92.6	75.2/90.9		
2	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	🔗	90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8		
3	HFL IFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6		
+	4	Alibaba DAMO NLP	StructBERT + TAPT	🔗	90.6	75.3	97.3	93.9/91.9	93.2/92.7	74.8/91.0	
+	5	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	
6	T5 Team - Google	T5	🔗	90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6		
7	Microsoft D365 AI & MSR AI & GATECHMT-DNN-SMART		🔗	89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2		
+	8	Huawei Noah's Ark Lab	NEZHA-Large		89.8	71.7	97.3	93.3/91.0	92.4/91.9	75.2/90.7	
+	9	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)	🔗	89.7	70.5	97.5	93.4/91.2	92.6/92.3	75.4/90.7	



Results for submission **ERNIE**

Score: **90.9**

PRIMARY DIAGNOSTICS

Task	Metric	Score
The Corpus of Linguistic Acceptability	Matthew's Corr	74.4
The Stanford Sentiment Treebank	Accuracy	97.8
Microsoft Research Paraphrase Corpus	F1 / Accuracy	93.9/91.8
Semantic Textual Similarity Benchmark	Pearson-Spearman Corr	93.0/92.6
Quora Question Pairs	F1 / Accuracy	75.2/90.9
MultiNLI Matched	Accuracy	91.9
MultiNLI Mismatched	Accuracy	91.4
Question NLI	Accuracy	97.3
Recognizing Textual Entailment	Accuracy	92.0

SQuAD (2018)

- Stanford Question Answering Dataset
- Vrednotenje enega korpusa
- Nalaganje modela z izvorno kodo na CodaLab
- Testni podatki (vključno z besedili) nedostopni
- Primeri skript za samostojen preskus

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Feb 21, 2021	FPNet (ensemble) Ant Service Intelligence Team	90.871	93.183
2 May 16, 2021	IE-NetV2 (ensemble) RICOH_SRCB_DML	90.860	93.100
3 Feb 24, 2021	IE-Net (ensemble) RICOH_SRCB_DML	90.758	93.044
4 Apr 06, 2020	SA-Net on Albert (ensemble) QIANXIN	90.724	93.011
5 May 05, 2020	SA-Net-V2 (ensemble) QIANXIN	90.679	92.948

dynabench (2020)

- dinamično dodajanje podatkov
- podpora več univerz (Harvard, Stanford, ...) in podjetij (Google, Meta, ...)
- več podpornih orodij, CLI za oddajo celotnega modela

DynaBench About Tasks

SENTIMENT ANALYSIS

Find examples that fool the model

Your goal: enter a **negative** statement that fools the model into predicting positive.

Please pretend you are reviewing a place, product, book or movie.

This year's NAACL was very different because of Covid

Model prediction: **positive**

Well done! You fooled the model.

Optionally, provide an explanation for your example: [Draft. Click out of input box to save.](#)

Covid is clearly not a good thing

The model probably doesn't know what Covid is

Model Inspector

#s This year 's NA AC L was very different because of Cov id #/s

The model inspector shows the [layer integrated gradients](#) for the input token layer of the model.

93.79% 6.21%

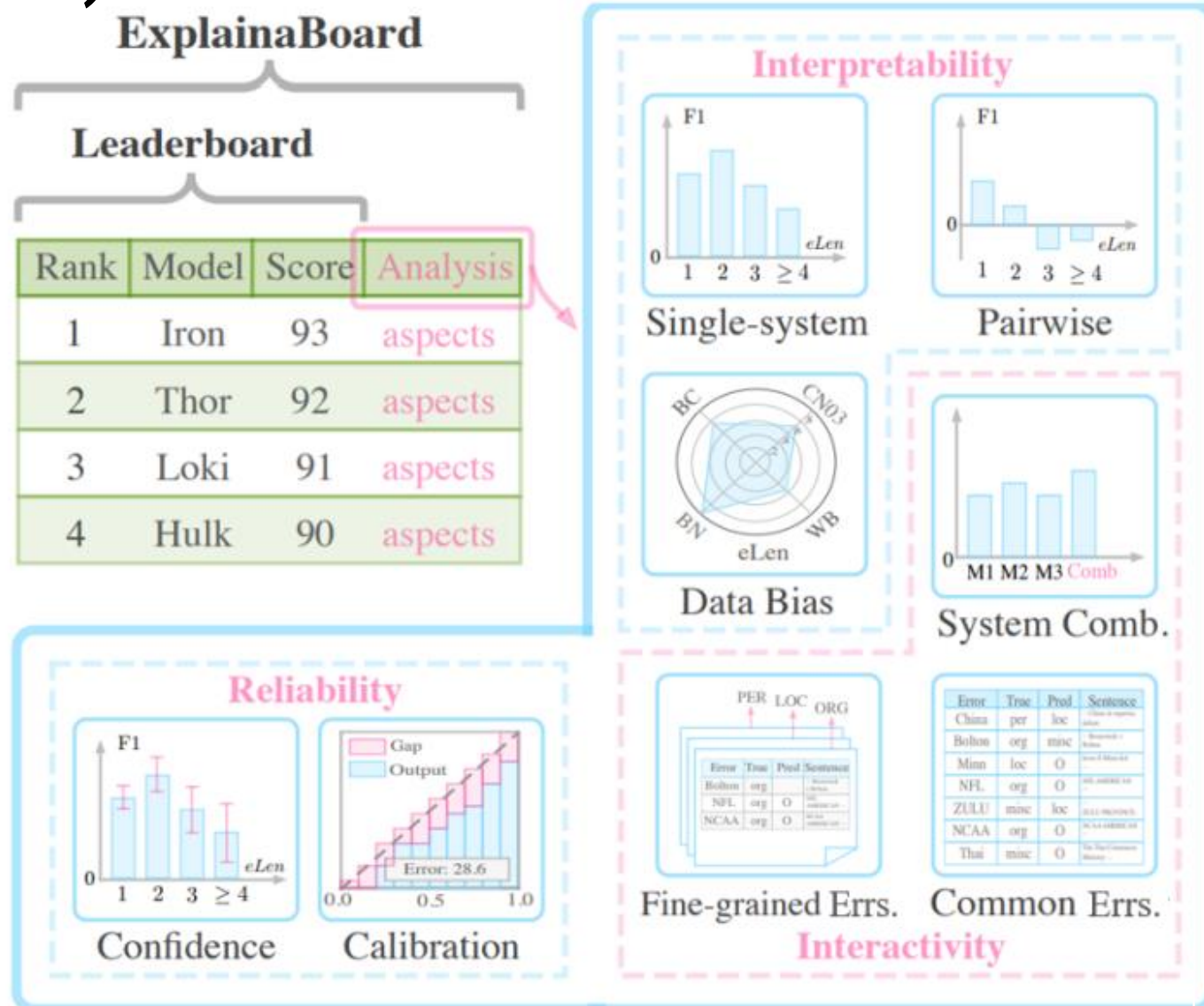
Retract Flag Inspect

This year's NAACL was very different because of Covid

Submit

explainaboard (2021)

- avtomatska primerjava modelov preko več kategorij – razlaga, zanesljivost, interakcije
- oddaja zip datotek



Allen AI Leaderboards (2021)

- nadgradnja GENIE
- opis modela / povezava do članka obvezna
- splošno ogrodje za vrednotenje v okolju Docker - na voljo

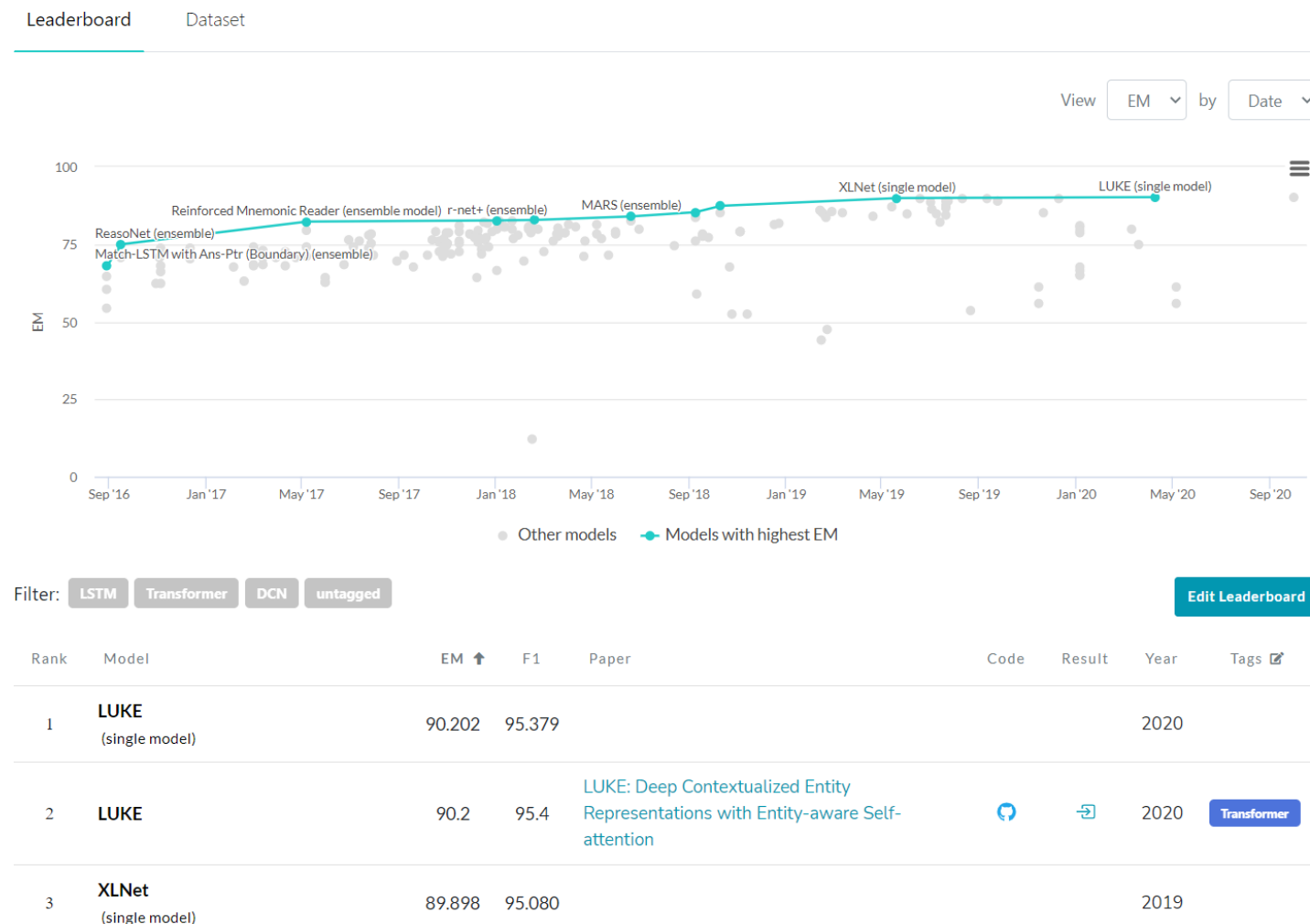
The screenshot shows the Allen AI Leaderboards interface for the A-OKVQA benchmark. At the top, there is a dark blue header with the Allen Institute for AI logo and the word "Leaderboards". A "Log In" button is visible in the top right. Below the header, the benchmark name "A-OKVQA" is prominently displayed next to its logo. A description states: "A-OKVQA is a new knowledge-based visual question answering benchmark comprised of 25K questions that require a bro... more". A "+ Create Submission" button is located to the right. Below this, there are navigation links for "Public Submissions", "Getting Started", and "About". The main content is a table of public submissions with columns for Rank, Submission, Created, DA VQA Score, and MC Accuracy. A "Download" button is in the top right of the table area.

Rank	Submission	Created	DA VQA Score	MC Accuracy
1	QAPrompt-Clip-8-13B Fudan University	08/13/2023	0.6879	0.5043
2	QAprompt-Clip-8 Fudan University	07/14/2023	0.6810	0.4948
3	test2 fdu	07/13/2023	0.6669	0.4576
4	test fdu	07/14/2023	0.6548	0.4505

paperswithcode.com (Meta)

- področja -> naloge -> primerjalniki
- brez avtomatskega vrednotenja, le *ročni* vnosi

Question Answering on SQuAD1.1



Predstavitev SloBENCH

Glavni poudarki

- Ciljna skupina: študenti, podjetja, posamezniki, raziskovalci (angl.)
- Splošna podpora za poljubno nalogo ONJ
 - transparentnost ocenjevalnega mehanizma
- Modularnost in razširljivost
- Proces oddaje rezultatov
 - omejitev oddaj; oddaja le oznak; skriti označeni testni korpusi
- Dodatne značke
 - objave; *baseline*; izvorna koda; preverjeno delovanje; ...

Obstoječi vrednotniki in SloBENCH

	GLUE/SuperGLUE	KLEJ	RussianSuperGLUE	XTREME	SQuAD	GENIE	GEM	Dynabench	Explainaboard	SloBench
Online leaderboard	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Automatic (online) evaluation						✓		✓		✓
Model-agnostic	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Hidden test dataset labels	✓	✓		✓	✓	✓		✓	✓	✓
Public evaluation scripts	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Number of tasks	10	9	9	4	1	4	13	7	13	10
Extendable						✓	✓	✓	✓	✓
Advanced analysis									✓	
Human-in-the-loop						✓		✓		

0 SloBENCH

- CLARIN projekt 2021
- Diplomaska naloga FRI
- Frenk Dragar

- Rezultati
 - Portal CJVT
 - Ogradje za vrednotenje CLARIN.si
 - Testni korpusi

SloBENCH: Zasnova in izdelava ogrodja za merjenje uspešnosti orodij za jezikovne tehnologije
Projektni predlog za "Razpis CLARIN.SI za projekte 2021"

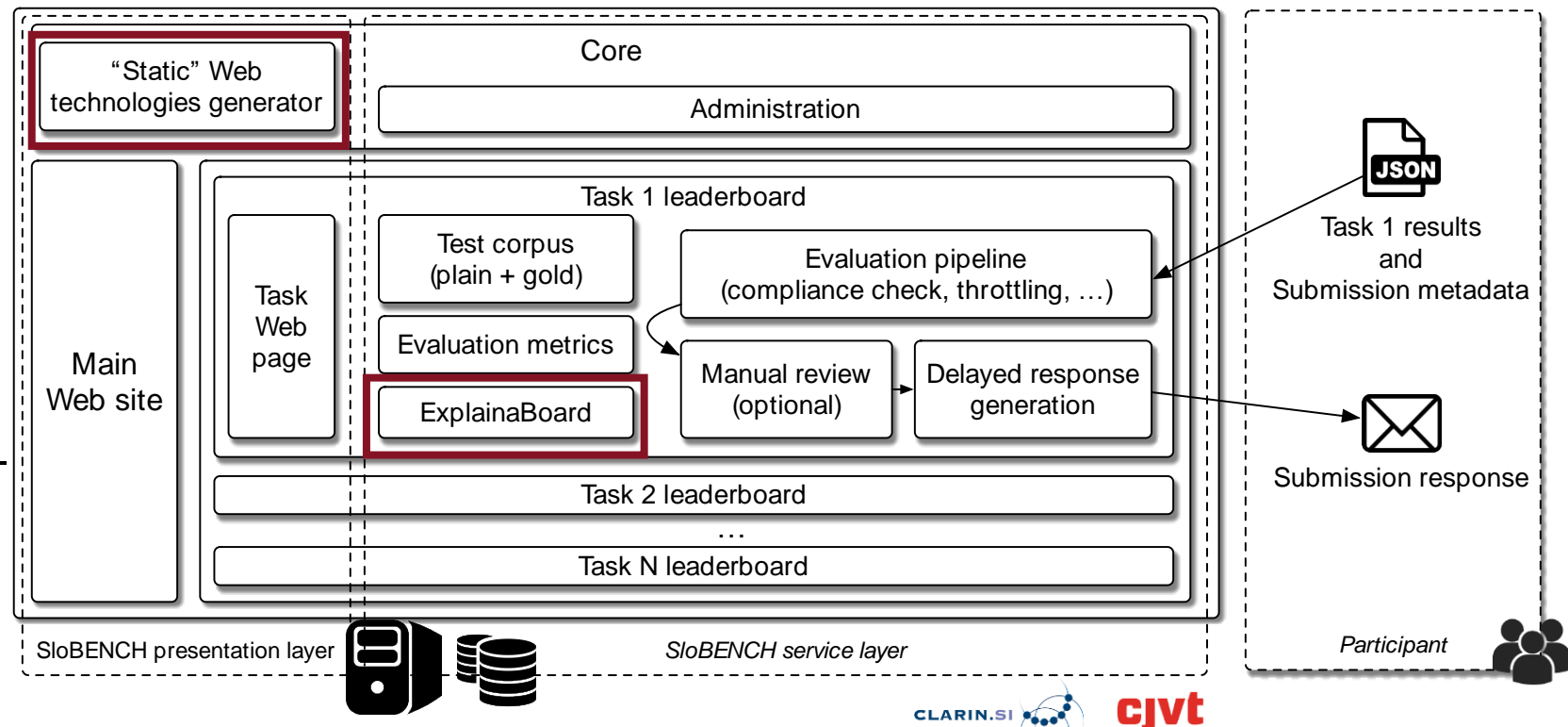
Ime predlaganega projekta: SloBENCH: Zasnova in izdelava ogrodja za merjenje uspešnosti orodij za jezikovne tehnologije

Prijavitelji:
Slavko Žitnik, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Simon Krek, Univerza v Ljubljani, Center za jezikovne vire in tehnologije
Marko Robnik-Sikonja, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

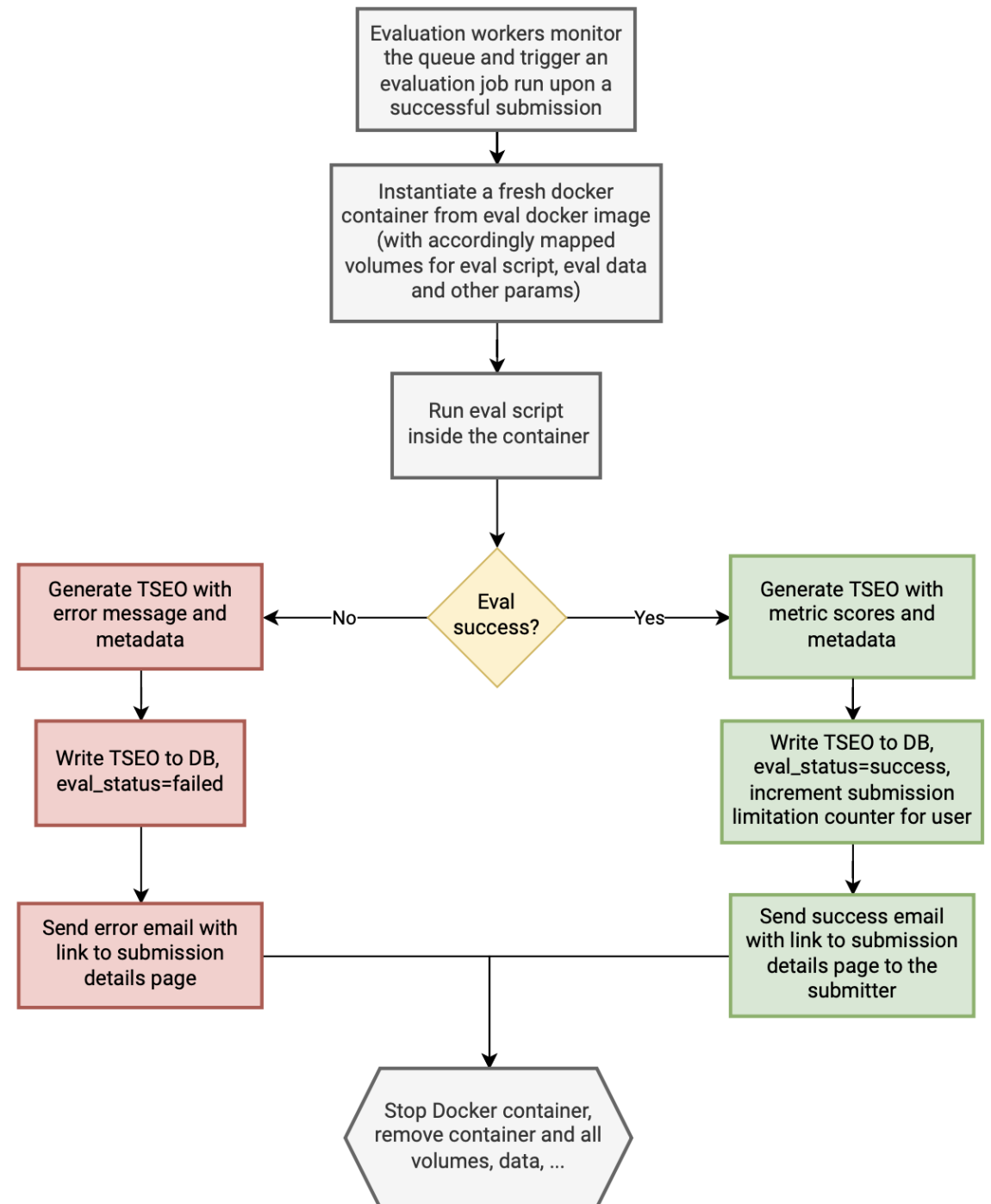
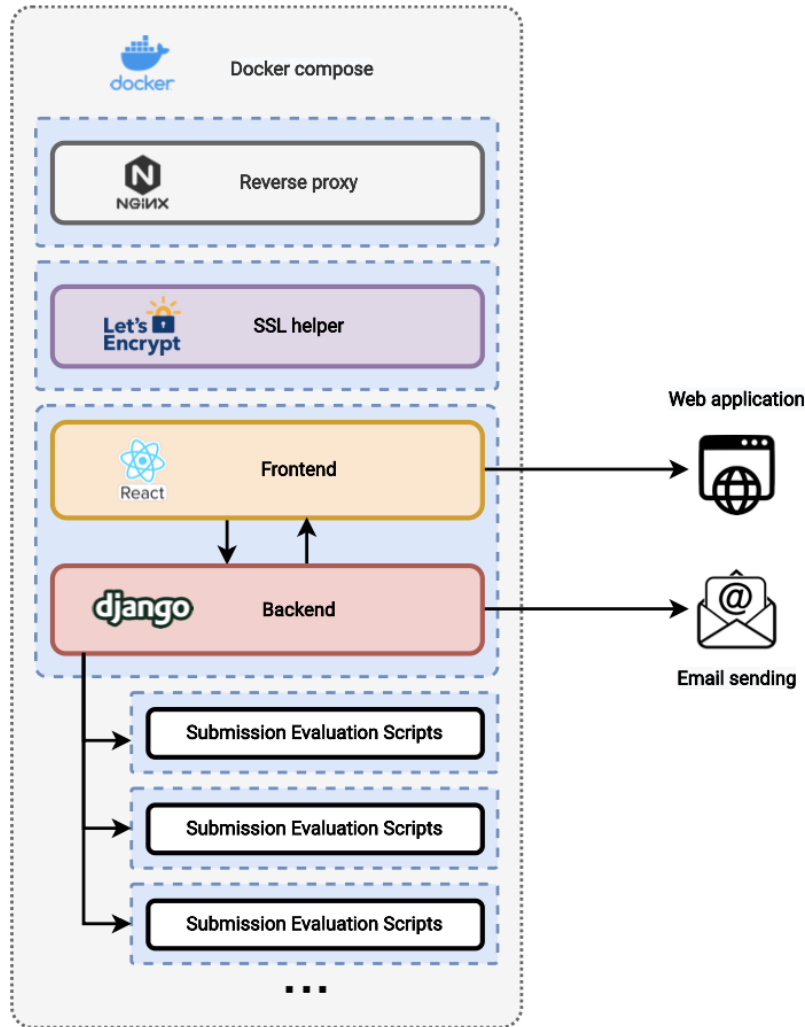
Opis projekta:
Na področju obdelave naravnega jezika obstaja kar nekaj nalog, ki predstavljajo pomembne dele opremljenosti posameznega jezika z jezikovnimi tehnologijami in viri. Primeri takšnih nalog so avtomatsko povzemanje, prevajanje, oblikoslovno označevanje ali tehnike ekstrakcije informacij. Jezikovni viri in tehnologije so dostopne preko različnih resursov (npr. repozitorij Clarin), vendar se njihova objektivna primerjava ne izvaja celovito ali enotno.

Z namenom objektivne primerjave se za določene naloge organizirajo »tekmovanja« (angl. shared tasks), kjer so oznake testnih podatkov nedostopne. Neoznačena testna množica je tekmovalcem na voljo šele kratak čas pred oddajo, ko morajo oddati napovedi svojih algoritmov in na podlagi katerih organizatorji rangirajo udeležence. Nekatere označene testne množice niso razkrite niti po koncu tekmovanja, vendar pa so zagotovljene storitve, preko katere lahko raziskovalci preverijo svoj sistem.

SloBENCH - High level architecture



Delovanje SloBENCH



Leaderboards

Question answering (SuperGLUE) >

This question answering leaderboard is a Slovene (translated) version of an existing English version of SuperGLUE benchmark.



15 February 2022



1

1.0

Machine Translation (ENG -> SLO) >

This machine translation leaderboard is measuring a success of automatic machine translation from English to Slovene language.



21 February 2022



3

1.0

Machine Translation (SLO -> ENG) >

This machine translation leaderboard is measuring a success of automatic machine translation from Slovene to English language.



21 February 2022



3

1.0

Frequently Asked Questions

How can I make a submission? ∨

I have questions or want to report bugs ∨

How can my submission get special tags - "Verified", ...? ∨

About SloBench - Slovenian NLP Benchmark

SloBENCH is an evaluation platform for benchmarking the development of Slovene natural language processing technologies. The main goal of SloBENCH evaluation is that evaluation data labels are not publicly published. Additionally, submission limitation are defined to avoid overfitting of models. Still we ask all the submitters to respect ethics and to not manually annotate the test data.


The evaluation scripts are publicly available. To support the community we propose to open-source as much of the submitted systems as possible. Initially, up to ten different leaderboards will be set up and gradually updated later.

The system was developed as one of Clarin.si's 2021 projects and was proposed by Slavko Žitnik, Marko Robnik-Šikonja and Simon Krek. The implementation was done by Frenk Dragar.

Edit Leaderboard 

New Submission +

Train Dataset 

Test Dataset (Unlabeled) 

Sample Submission 

Leaderboard

Natural language inference (SI-NLI) 1.0

Description

This leaderboard is measuring the performance of natural language inference systems in Slovene. Given a premise and a hypothesis, the task is to detect whether the hypothesis entails, contradicts, or is neutral in relation to the premise. The test data contains 998 pairs. The training, validation and (unlabeled) test set are available on the [CLARIN.si repository](#).

Leaderboard Editors

Slavko Žitnik, slavko@zitnik.si

Matej Klemen, Matej.Klemen@fri.uni-lj.si

Related Leaderboards

Natural language inference (SI-NLI) (v1) 

Leaderboard Submissions

Rank	Title	Authors and Affiliations	Accuracy	P(entailment)	R(entailment)	F1(entailment)	P(neutral)	R(neutral)	F1(neutral)	P(contradiction)	R(contradiction)
1	cross-encoder-s...	Jan Štihec	0.7715	0.7937	0.7427	0.7674	0.7097	0.7378	0.7235	0.8131	0.8354
2	SloBERTa	Matej Klemen, A...	0.7355	0.7939	0.6871	0.7367	0.6955	0.7591	0.7259	0.7267	0.7622

Natural language inference (SI-NLI)

Edit Make Verified Make Baseline Make Private Delete

Submission

cross-encoder-sloberta-si-nli-snli-mnli

Public

Description

SentenceTransformers CrossEncoder class model based on SloBERTa. It was trained on the SI-NLI and the slovene-snli-mnli (https://huggingface.co/datasets/jacintes/slovene_mnli_snli) datasets. All training details including the training script are available at: <https://github.com/jacintes/slovene-nli-benchmark>. The model is available at: <https://huggingface.co/jacintes/cross-encoder-sloberta-si-nli-snli-mnli>

Authors and Affiliations

Jan Štihec

Code Link

`</>` <https://github.com/jacintes/slovene-nli-benchmark>

Licence

[Creative Commons - Attribution-ShareAlike 4.0 International \(CC BY-SA 4.0\)](#)

Date Submitted

20 April 2023, 09:10

Submission Context Table

Rank	Title	Accuracy	P(entailment)	R(entailment)	F1(entailment)
1	cross-encoder-s...	0.7715	0.7937	0.7427	0.7674
2	SloBERTa	0.7355	0.7939	0.6871	0.7367

Accuracy

P(entailment)

1 0.77154200

1 0.79375000

Natural language inference (SI-NLI)

[Train Dataset](#) ⬇

[Test Dataset](#) ⬇

[Sample Submission](#) ⬇

New Submission

Title

e.g. 'My NER Model'

Description

Describe your submission here

Authors and Affiliations

e.g. 'John Doe, FRI & Jana Novak, IJS'

Paper Link

e.g. 'https://arxiv.org/...'

Optional (but desired)

Code Link

e.g. 'https://github.com/...'

Optional (but desired)

Submission File

Add File

Submission Visibility

Submissions on this leaderboard are private by default.

Evaluation Progress Notification

Because the evaluation can take some time, you will receive an email after evaluation.

Submission Limitations

0/5 submissions this week

Submission Instructions

Submit a zip with a text file submission.txt containing one of "entailment", "neutral" or "contradiction" in each line. Make sure that each line corresponds to the instances in the test dataset.



no-reply@slobench.cjvt.si

Evaluation for submission 'CLASSLA 1.1.0' has been completed

To: Slavko Žitnik

Bin - zitnik.si 11:45

Evaluation for submission 'CLASSLA 1.1.0' has been completed

The submission 'CLASSLA 1.1.0' has been completed. You can check out the results on the site.

[See Submission](#)

Primer izdelave novega vrednotnika v SloBENCH

Postopek izdelave novega vrednotnika

- Korak 1: Izdelajte testno množico (označeno in javno) in javen primer označene množice.
- Korak 2: Definirajte način ocenjevanja in ga implementirajte.
- Korak 3: Vgradite ocenjevanje v ločeno vejo ogrodja za vrednotenje SloBENCH.
- Korak 4: Pridobite dostop do SloBENCH (urednik) in objavite sliko Docker za ocenjevanje.
- Korak 5: Vpišite nov vrednotnik v SloBENCH
- Korak 6: Naložite *baseline* oddajo in preverite delovanje.

Korak 1: Izdelajte testno množico (označeno in javno) in javen primer označene množice

- Testna množica (.zip)
 - označena za namene avtomatskega vrednotenja
 - javna brez oznak za označevanje sodelujočih
- Primer označene množice (.zip)
 - manjši javen primer množice za testiranje delovanja
- Učna množica (opcijsko)
 - povezava na primer vira z učnimi podatki

```
17 Deutsche B-ORG
18 Telekom I-ORG
19 ima 0
20 v 0
21 večinski 0
22 lasti 0
23 hrvaški 0
24 T B-ORG
```

```
17 Deutsche
18 Telekom
19 ima
20 v
21 večinski
22 lasti
23 hrvaški
```

Korak 2: Definirajte način ocenjevanja in ga implementirajte.

■ Vhoda

- mapa z razširjeno oddano datoteko zip
- mapa z razširjeno referenčno datoteko zip

■ Postopek

- poljuben programski jezik, algoritmi, knjižnice, ...
- kasneje zapakirano v Docker

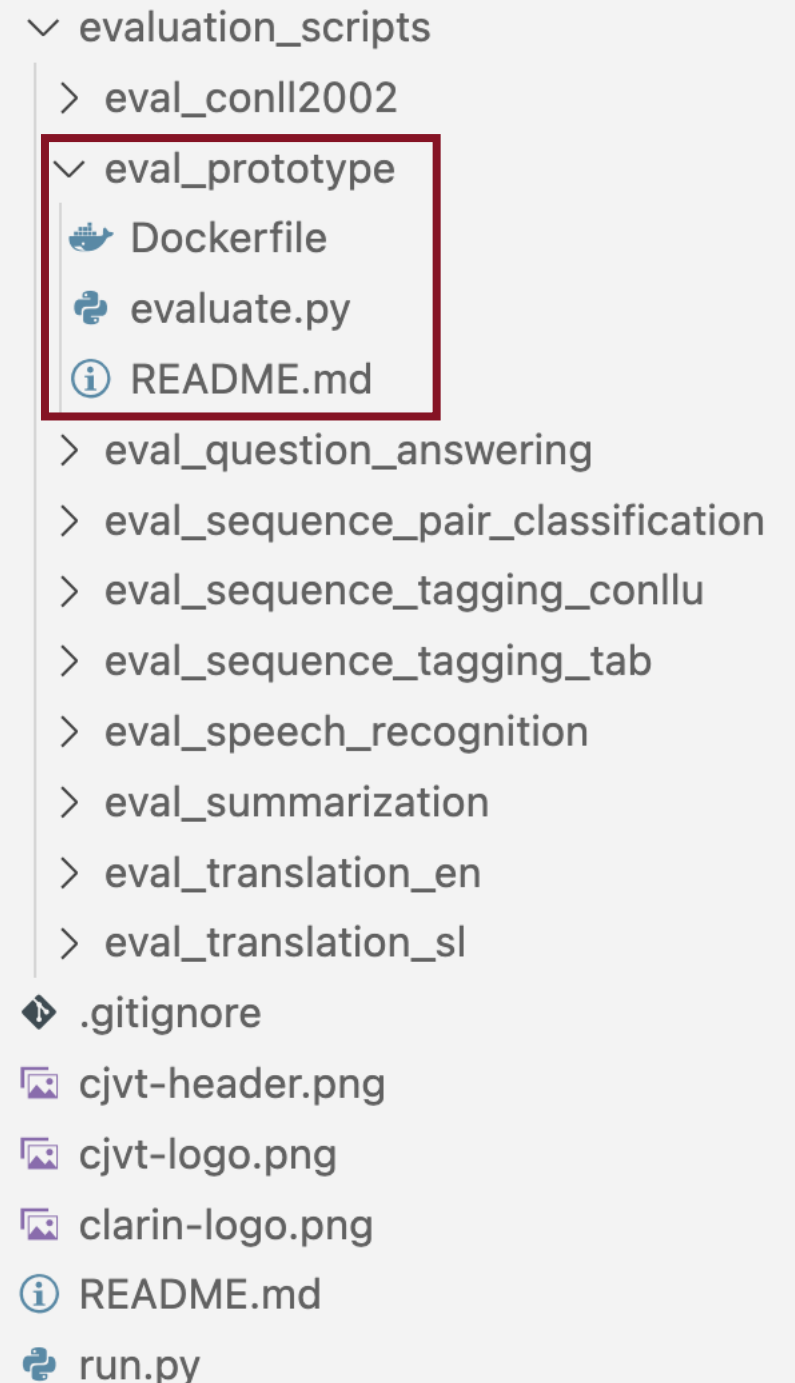
■ Izhod

- Python slovar (*string* -> *int*)

```
{  
  "status": "S",  
  "metrics": {  
    "Precision": 91.51159,  
    "Recall": 89.1484184,  
    "F1 score": 90.329677,  
  },  
  "evaluation_time": 0.044729,  
  "error_report": ""  
}
```

Korak 3: Vgradite ocenjevanje v ločeno vejo ogrodja za vrednotenje SloBENCH.

- “Fork” <https://github.com/clarinsi/slobench-eval-docker>
- Implementacija v novi mapi ...



evaluation_scripts

> eval_conll2002

✓ eval_prototype

Dockerfile

evaluate.py

README.md

> eval_question_answering

> eval_sequence_pair_classification

> eval_sequence_tagging_conllu

> eval_sequence_tagging_tab

> eval_speech_recognition

> eval_summarization

> eval_translation_en

> eval_translation_sl

.gitignore

cjvt-header.png

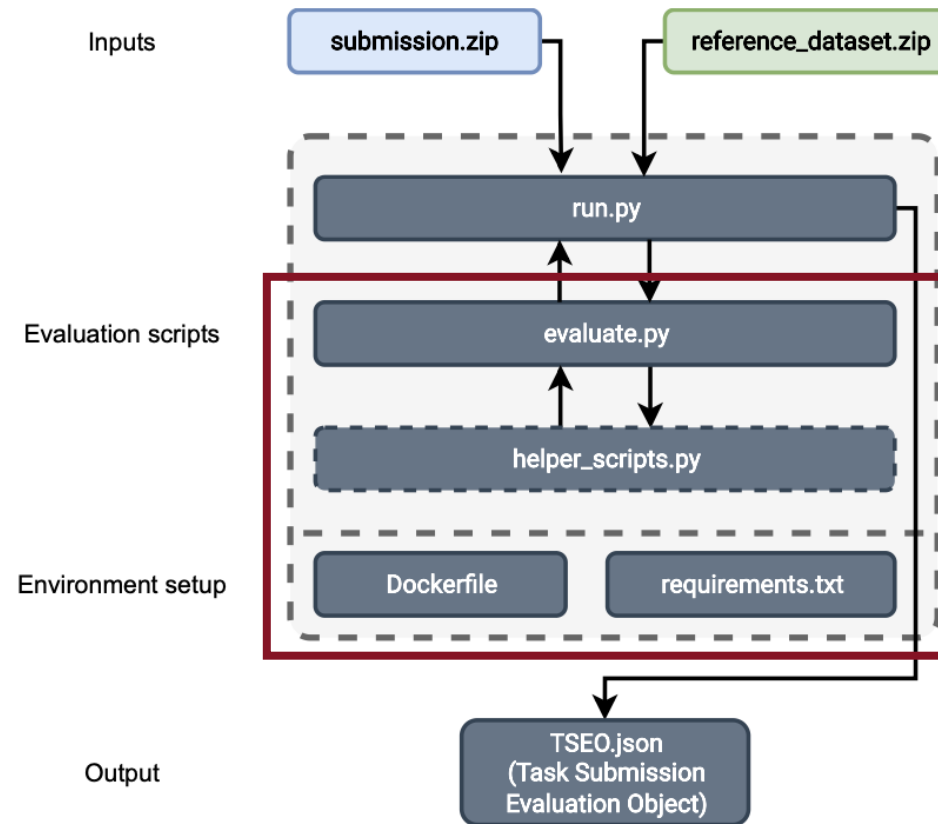
cjvt-logo.png

clarin-logo.png

README.md

run.py

Korak 3: .. implementacija ...



```
{
  "status": "S",
  "metrics": {
    "Precision": 91.51159,
    "Recall": 89.1484184,
    "F1 score": 90.329677,
  },
  "evaluation_time": 0.044729,
  "error_report": ""
}
```

Korak 3: ..., primer NER.

- Pregled izdelanega primera
 - https://github.com/clarinsi/slobench-eval-docker/blob/main/evaluation_scripts/eval_conll2002

HANDS-ON

Korak 4: Pridobite dostop do SloBENCH (urednik) in objavite sliko Docker za ocenjevanje

- Registrirajte se na <https://slobench.cjvt.si>
 - dostop uporabnika za oddajanje nalog
- Kontaktirajte skrbnika (*Slavko Žitnik*)
 - Skrbnik dodeli status urednika (obstaja še status administratorja).
 - Skrbnik preveri vejo (*fork*) repozitorija, zahteva spremembe/jih uredi sam in sprejme *pull request*.
 - Skrbnik lokalno preveri delovanje in objavi novo sliko na DockerHub
 - <https://hub.docker.com/repository/docker/slobench/eval/general>

New Leaderboard

Title

Named entity recognition

Version

1.0

Character limit: 24/80

Short Description

A named entity recognition benchmark for Slovene language.

Character limit: 58/200

Description

This leaderboard is measuring the performance of named entity recognition in Slovene. Given tokenized text, the task is to annotate each token with an appropriate label. The test data contains 81 document. Currently largest named entity recognition corpus - Training corpus SUK 1.0, is available at the [CLARIN.si repository](http://hdl.handle.net/11356/1747).

Check out the [markdown guide](#) for a styling guide

Submission Instructions

A submission file should be one zip file having **annotated** files only! Prior submission, you can check whether output of your system is correct (e.g., by testing on training data outputs) by following instrucions in the repository [https://github.com/clarinsi/slobench-eval-docker] (https://github.com/clarinsi/slobench-eval-docker).

Check out the [markdown guide](#) for a styling guide

Train Dataset URL

http://hdl.handle.net/11356/1747

This field is optional

Korak 5: Vpišite nov vrednotnik v SloBENCH

Evaluation Docker Image

slobench/eval:conll2002_3.0

Visible Metrics (as valid JSON array)

```
[  
  {  
    "F1 (MISC)": "F1 (MISC)"  
  }  
]
```

Test Dataset (Unlabeled)

Add File

Reference Dataset

Add File

Sample Submission

Add File

Submission Limitation Number

5

Time Unit

Week



Submission Default Visibility



Private by default

Korak 5: Vpišite nov vrednotnik v SloBENCH

Korak 5: Vpišite nov vrednotnik v SloBENCH

Send Users an Email upon Submission Evaluation Task Finish

Send Email

Related Leaderboards

1 × Related Leaderboards ▾

Editors


1 × Editors ▾

Leaderboard Visibility

Visible


Submit

Korak 5: Vpišite nov vrednotnik v SloBENCH

Edit Leaderboard 

New Submission +

Train Dataset 

Test Dataset (Unlabeled) 

Sample Submission 

Leaderboard

Named entity recognition 1.0

Description

This leaderboard is measuring the performance of named entity recognition in Slovene. Given tokenized text, the task is to annotate each token with an appropriate label. The test data contains 81 document. Currently largest named entity recognition corpus - Training corpus SUK 1.0, is available at the [CLARIN.si repository](#).

Leaderboard Editors

Slavko Žitnik, slavko@zitnik.si

Leaderboard Submissions

This leaderboard contains no public and successful submissions.

Related Leaderboards

Named entity recognition (v1) 

Korak 6: Naložite baseline oddajo in preverite delovanje

Leaderboard

Named entity recognition

Train Dataset ↓

Test Dataset ↓

Sample Submission ↓

New Submission

Title

CLASSLA 1.1.0

Description

Using pretokenized CLASSLA seting against the test data.

Authors and Affiliations

Slavko Žitnik, UL FRI

Submission Visibility

Submissions on this leaderboard are private by default.

Evaluation Progress Notification

Because the evaluation can take some time, you will receive an email after evaluation.

Submission Limitations

0/5 submissions this week

Submission Instructions

Submit a zip with annotated files following *conll2002 format*. Make sure the submission file exactly corresponds the reference dataset file (e.g., number of files, tokenization preserved).

A submission file should be one zip file having *annotated* files only! Prior submission, you can check whether output of your system is correct (e.g., by testing on training data outputs) by following instrucions in the repository <https://github.com/clarinsi/slobench-eval-docker>.

Korak 6: Naložite baseline oddajo in preverite delovanje

Paper Link

e.g. 'https://arxiv.org/...'

Optional (but desired)

Code Link

e.g. 'https://github.com/...'

Optional (but desired)

Submission File

Add File

classla_reference_submission.zip ×

Code Licence

Apache License 2.0



Check out the [detailed licence descriptions here](#).

Licence agreement

I agree that by uploading data through this form, I give the SloBench team a worldwide license to use, host, store, reproduce and modify such data for the purpose of evaluating and verifying the results on this benchmark task.

I agree to the licence agreement

Submit

Named entity recognition

Edit

Make Verified

Make Baseline

Make Public

Delete

Submission

CLASSLA 1.1.0

Private

Description

Using pretokenized CLASSLA setting against the test data.

Authors and Affiliations

Slavko Žitnik, UL FRI

Licence

Apache License 2.0

Date Submitted

21 September 2023, 10:09

F1 score



Precision



Recall



F1 (PER)



F1 (DERIV-PER)



F1 (LOC)



F1 (ORG)



F1 (MISC)



Evaluation Results



Evaluation in progress ...

Korak 6: Naložite *baseline* oddajo in preverite delovanje

Named entity recognition

Edit | Make Verified | Make Baseline | Make Private | Delete

Submission
CLASSLA 1.1.0 Public

Description
Using pretokenized CLASSLA seting against the test data.

Authors and Affiliations
Slavko Žitnik, UL FRI

Licence
Apache License 2.0

Date Submitted
21 September 2023, 10:09

F1 score	Precision
87.67027593	89.10188143
Recall	F1 (PER)
86.28394637	92.95454545
F1 (DERIV-PER)	F1 (LOC)
78.43137255	91.87609585
F1 (ORG)	F1 (MISC)
82.39266763	30.00000000

Submission Context Table

Rank	Title	F1 score	Precision	Recall	F1 (PER)	F1 (DERIV-PER)	F1 (LOC)
1	CLASSLA 1.1.0	87.6703	89.1019	86.2839	92.9545	78.4314	91.8761

Korak 6: Naložite *baseline* oddajo in preverite delovanje

Korak 6: Naložite *baseline* oddajo in preverite delovanje

Edit Leaderboard ↗

New Submission +

Train Dataset ⬇

Test Dataset (Unlabeled) ⬇

Sample Submission ⬇

Leaderboard

Named entity recognition 1.0

Description

This leaderboard is measuring the performance of named entity recognition in Slovene. Given tokenized text, the task is to annotate each token with an appropriate label. The test data contains 81 document. Currently largest named entity recognition corpus - Training corpus SUK 1.0, is available at the [CLARIN.si repository](#).

Leaderboard Editors

Slavko Žitnik, slavko@zitnik.si


Related Leaderboards

Named entity recognition (v1) ▾

Leaderboard Submissions

Rank	Title	Authors and Affiliations	F1 score	Precision	Recall	F1 (PER)	F1 (DERIV-PER)	F1 (LOC)	F1 (ORG)	F1 (MISC)	Tag
1	CLASSLA 1.1.0	Slavko Žitnik, UL FRI	87.6703	89.1019	86.2839	92.9545	78.4314	91.8761	82.3927	30.0000	Baseline


Korak 6: Naložite *baseline* oddajo in preverite delovanje

 **no-reply@slobench.cjvt.si** Inbox - zitnik.si 12:09

Evaluation for submission 'CLASSLA 1.1.0' has been completed
To: Slavko Žitnik

Evaluation for submission 'CLASSLA 1.1.0' has been completed

The submission 'CLASSLA 1.1.0' has been completed. You can check out the results on the site.

 [See Submission](#)

SloBENCH v prihodnje ...

- Možnosti za nadaljnji razvoj
 - odprtokodnost spletišča; nadgradnje (avtomatske analize, izboljšave, ...)
- Promocija uporabe in nalaganja modelov/vrednotnikov

Created by



Published by

Univerza v Ljubljani



Managed by



Availability

The evaluation scripts for this project are available under Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)



Contact

Centre for Language
Resources and
Technologies

Večna pot 113
SI-1000 Ljubljana

Phone
+386 1 479 82 99

Email
info@cjvt.si