GlotScript: A Resource and Tool for Low Resource Writing System Identification

***Amir Hossein Kargaran**, François Yvon, Hinrich Schütze*

CIS

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

SORBONNE UNIVERSITÉ

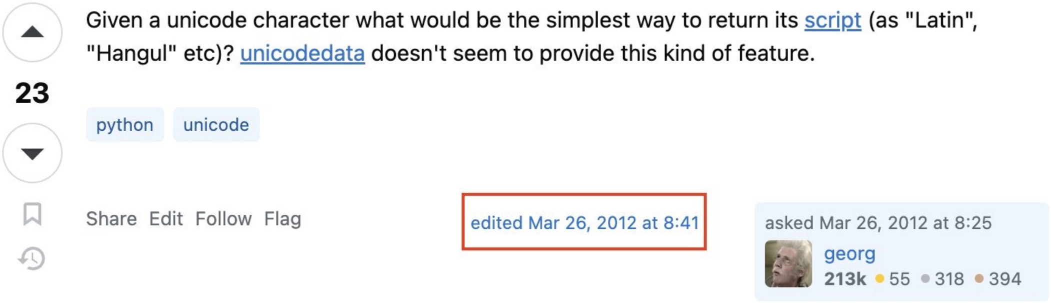
MCML Munich Center for Machine Learning

LREC-COLING 2024

Script (Writing System) Identification

Script Identification Task

- **Question**
  "Given a Unicode character, what would be the simplest way to return its script?"



Given a unicode character what would be the simplest way to return its script (as "Latin", "Hangul" etc)? unicodedata doesn't seem to provide this kind of feature.

edited Mar 26, 2012 at 8:41

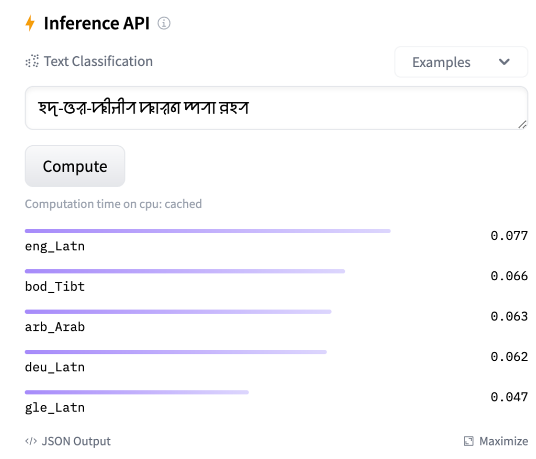
asked Mar 26, 2012 at 8:25

georg

213k ● 55 ● 318 ● 394

Script Identification in Language Identification Models

- **N-gram Language Identification**
  - Not all of the scripts are supported.
  - Hash collision



Inference API

Text Classification

Examples

হদ্-ওর-দ্বীনীস দ্বারগ দ্দসা রহস

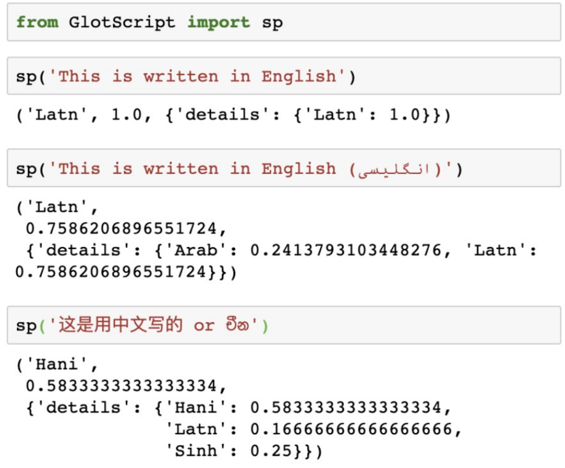
Compute

Computation time on cpu: cached

```
eng_Latn    0.077
bod_Tibt    0.066
arb_Arab    0.063
deu_Latn    0.062
gle_Latn    0.047
```

Script Identification Tool

- **Task**
  Identify script (writing system) of given text
- **Motivation**
  Language identification for low resource languages is prone to high error rates.
- **GlotScript-Tool**
  - Python library to identify script (writing system)
  - Supports 161 Unicode scripts, identified as ISO 15924 codes.



```
from GlotScript import sp
```

```
sp('This is written in English')
```

```
('Latn', 1.0, {'details': {'Latn': 1.0}})
```

```
sp('This is written in English (انگلیسی)')
```

```
('Latn',
 0.7586206896551724,
 {'details': {'Arab': 0.2413793103448276, 'Latn':
0.7586206896551724}})
```

```
sp('这是用中文写的 or ඊන')
```

```
('Hani',
 0.5833333333333334,
 {'details': {'Hani': 0.5833333333333334,
              'Latn': 0.16666666666666666,
              'Sinh': 0.25}})
```

Figure 1: How to use GlotScript-T: three examples. GlotScript-T returns a tuple consisting of the main script, the percentage of characters in the main script and detailed information on the distribution of scripts.

Script Identification Resource

- **Task**
  What is the attested writing systems for each language?
- **GlotScript-Resource**
  - Provide metadata of each language CORE writing system and AUXILLARY ones
  - Supports more than **7,000** languages
- **Source**
  Agreement over multiple metadata such as van Esch et al. (2022), Wikipedia, SIL

Language	CORE	AUXILIARY
Turkish	Latn	Arab, Cyrl, Grek
Thai	Thai	Latn
Vietnamese	Latn	Hani

GlotScript Use Case:

Corpus Cleaning/Quality Assessment

Corpus Cleaning/Quality

Assessment

- **Evaluation Corpora**

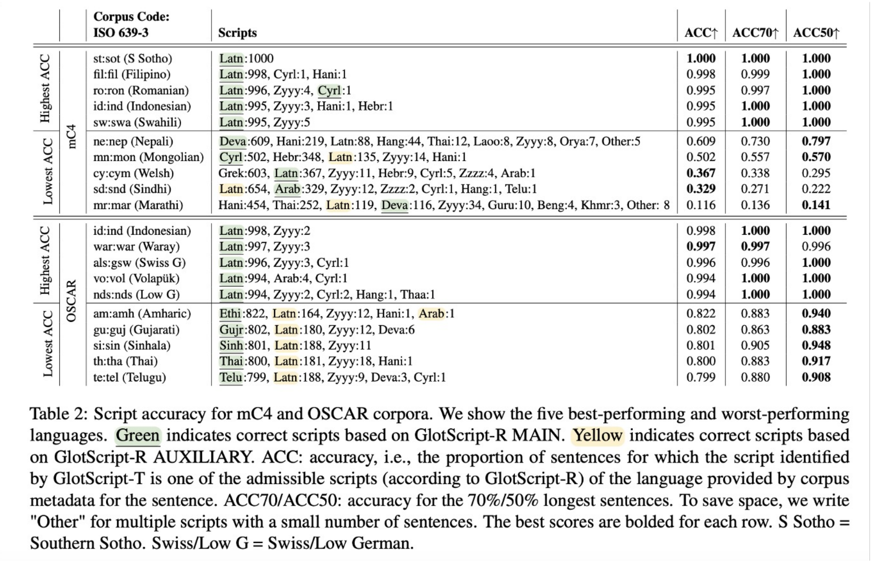
  mC4 and OSCAR2201

- **Evaluation Method**

  We take 1000 sentences from each language of each corpus

- **Table**

  We here show 5 best/worst performing languages.



		Corpus Code: ISO 639-3	Scripts	ACC↑	ACC70↑	ACC50↑
Highest ACC	mC4	st:sot (S Sotho)	Latn:1000	**1.000**	**1.000**	**1.000**
		fil:fil (Filipino)	Latn:998, Cyrl:1, Hani:1	0.998	0.999	**1.000**
		ro:ron (Romanian)	Latn:996, Zyyy:4, Cyrl:1	0.995	0.997	**1.000**
		id:ind (Indonesian)	Latn:995, Zyyy:3, Hani:1, Hebr:1	0.995	**1.000**	**1.000**
		sw:swa (Swahili)	Latn:995, Zyyy:5	0.995	**1.000**	**1.000**
Lowest ACC		ne:nep (Nepali)	Deva:609, Hani:219, Latn:88, Hang:44, Thai:12, Laoo:8, Zyyy:8, Orya:7, Other:5	0.609	0.730	**0.797**
		mn:mon (Mongolian)	Cyrl:502, Hebr:348, Latn:135, Zyyy:14, Hani:1	0.502	0.557	**0.570**
		cy:cym (Welsh)	Grek:603, Latn:367, Zyyy:11, Hebr:9, Cyrl:5, Zzzz:4, Arab:1	**0.367**	0.338	0.295
		sd:snd (Sindhi)	Latn:654, Arab:329, Zyyy:12, Zzzz:2, Cyrl:1, Hang:1, Telu:1	**0.329**	0.271	0.222
		mr:mar (Marathi)	Hani:454, Thai:252, Latn:119, Deva:116, Zyyy:34, Guru:10, Beng:4, Khmr:3, Other: 8	0.116	0.136	**0.141**
Highest ACC	OSCAR	id:ind (Indonesian)	Latn:998, Zyyy:2	0.998	**1.000**	**1.000**
		war:war (Waray)	Latn:997, Zyyy:3	**0.997**	**0.997**	0.996
		als:gsw (Swiss G)	Latn:996, Zyyy:3, Cyrl:1	0.996	0.996	**1.000**
		vo:vol (Volapük)	Latn:994, Arab:4, Cyrl:1	0.994	**1.000**	**1.000**
		nds:nds (Low G)	Latn:994, Zyyy:2, Cyrl:2, Hang:1, Thaa:1	0.994	**1.000**	**1.000**
Lowest ACC		am:amh (Amharic)	Ethi:822, Latn:164, Zyyy:12, Hani:1, Arab:1	0.822	0.883	**0.940**
		gu:guj (Gujarati)	Gujr:802, Latn:180, Zyyy:12, Deva:6	0.802	0.863	**0.883**
		si:sin (Sinhala)	Sinh:801, Latn:188, Zyyy:11	0.801	0.905	**0.948**
		th:tha (Thai)	Thai:800, Latn:181, Zyyy:18, Hani:1	0.800	0.883	**0.917**
		te:tel (Telugu)	Telu:799, Latn:188, Zyyy:9, Deva:3, Cyrl:1	0.799	0.880	**0.908**

Table 2: Script accuracy for mC4 and OSCAR corpora. We show the five best-performing and worst-performing languages. Green indicates correct scripts based on GlotScript-R MAIN. Yellow indicates correct scripts based on GlotScript-R AUXILIARY. ACC: accuracy, i.e., the proportion of sentences for which the script identified by GlotScript-T is one of the admissible scripts (according to GlotScript-R) of the language provided by corpus metadata for the sentence. ACC70/ACC50: accuracy for the 70%/50% longest sentences. To save space, we write "Other" for multiple scripts with a small number of sentences. The best scores are bolded for each row. S Sotho = Southern Sotho. Swiss/Low G = Swiss/Low German.

GlotScript Use Case:

Towards a better language identification

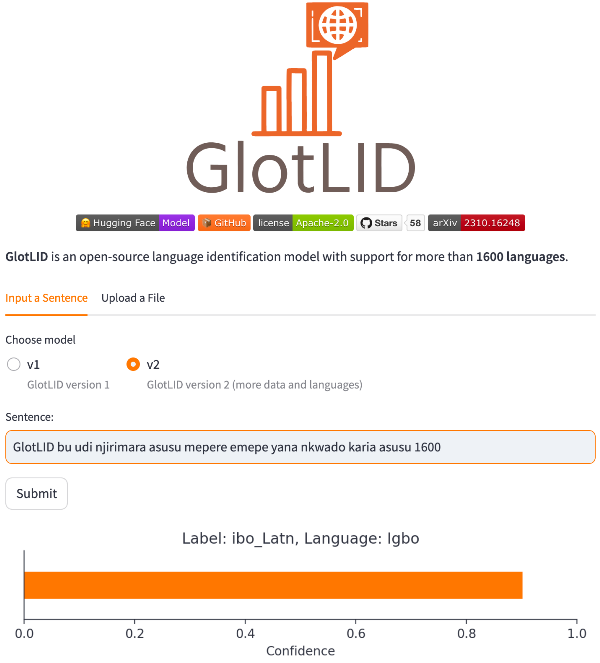
Towards a Better Language Identification

GlotLID is an open-source language identification model with support for more than 2000 languages.

https://arxiv.org/abs/2310.16248

https://github.com/cisnlp/GlotLID

https://huggingface.co/spaces/cis-lmu/glotlid-space



GlotLID is an open-source language identification model with support for more than **1600 languages**.

Choose model

○ v1 GlotLID version 1

● v2 GlotLID version 2 (more data and languages)

Sentence: GlotLID bu udi njirimara asusu mepere emepe yana nkwado karia asusu 1600

Label: ibo_Latn, Language: Igbo

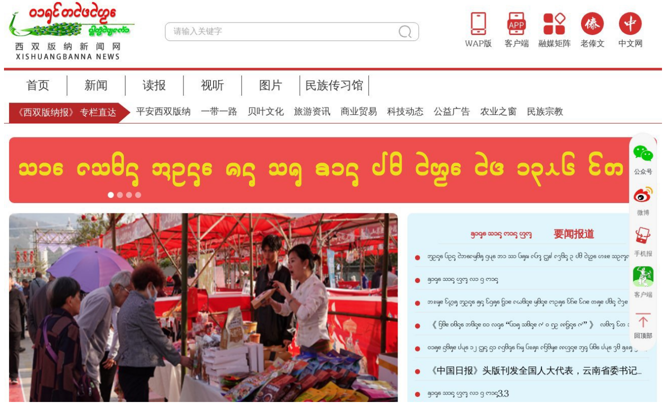
GlotScript Use Case:

Find Languages/Build Corpora

Find Languages/Build Corpora

- **Languages with Specific Scripts**
  - khb_Talu; Tai Lue
    - dw12.com
  - mon_Mong; Mongolian (Inner Mongolia)
    - mongolian.news.cn
  - syl_Sylo; Sylheti
    - surmafarorkhobor.com

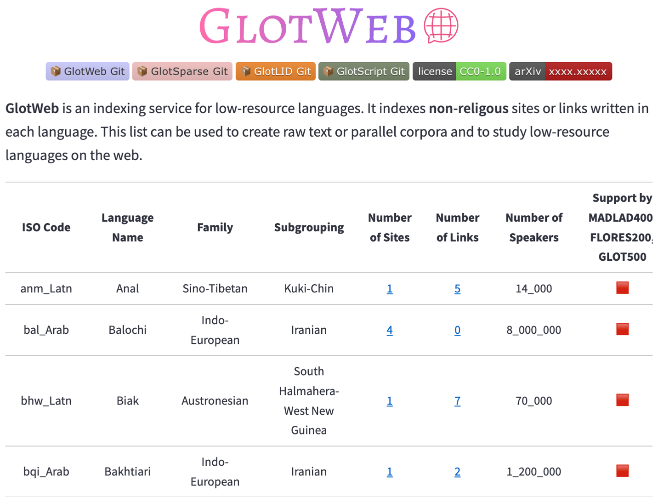
  …



《中国日报》头版刊发全国人大代表，云南省委书记...

Find Languages/Build Corpora

- GlotWeb: Indexing service for low-resource languages
- https://huggingface.co/spaces/cis-lmu/GlotWeb



GlotWeb is an indexing service for low-resource languages. It indexes **non-religous** sites or links written in each language. This list can be used to create raw text or parallel corpora and to study low-resource languages on the web.

ISO Code	Language Name	Family	Subgrouping	Number of Sites	Number of Links	Number of Speakers	Support by MADLAD400 FLORES200, GLOT500
anm_Latn	Anal	Sino-Tibetan	Kuki-Chin	1	5	14_000	🟥
bal_Arab	Balochi	Indo-European	Iranian	4	0	8_000_000	🟥
bhw_Latn	Biak	Austronesian	South Halmahera-West New Guinea	1	7	70_000	🟥
bqi_Arab	Bakhtiari	Indo-European	Iranian	1	2	1_200_000	🟥

GlotScript Use Case:

Analysis of Pre-trained Models

Tokenizer vocabulary

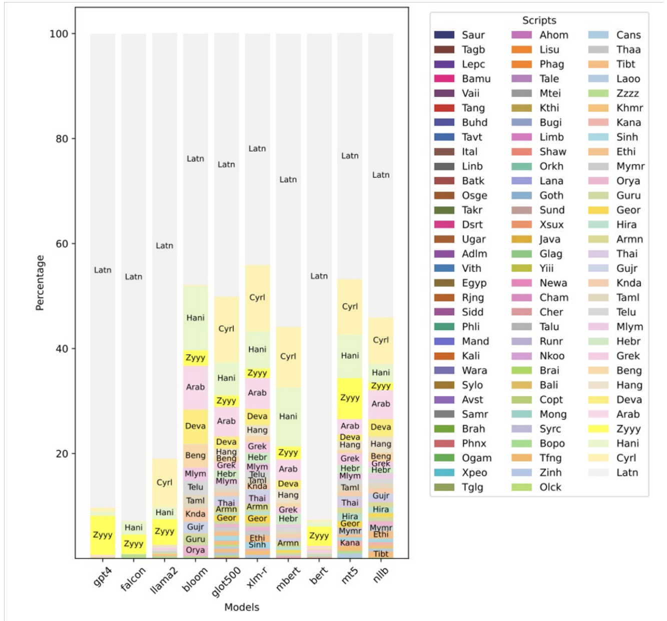


Figure 2: The percentage of each script in the vocabulary of model tokenizers. Scripts with a presence of more than 1% in each tokenizer are text-labeled in the figure.

Some observations:

(1) The Cyrillic representation in the BLOOM tokenizer is relatively scarce compared to other models.

(2) The BERT tokenizer supports not only Latin scripts but also recognizes Hani, Arabic, Cyrillic and some tokens in an additional 12 scripts.

(3) Glot500 encompasses the highest number of scripts, totaling 88. Following that, mT5 supports 66 scripts. However, a significant portion of these scripts in both models has limited presence.

(4) Llama2's second most prominent script is Cyrillic.

(5) Falcon's second most prominent script is Hani.

(6) The GPT-4 tokenizer vocabulary includes representations for 18 scripts, albeit not very comprehensively compared to its coverage of Latin.

(7) In all tokenizer models combined, a total of 92 scripts has some presence.

UDHR tokenization

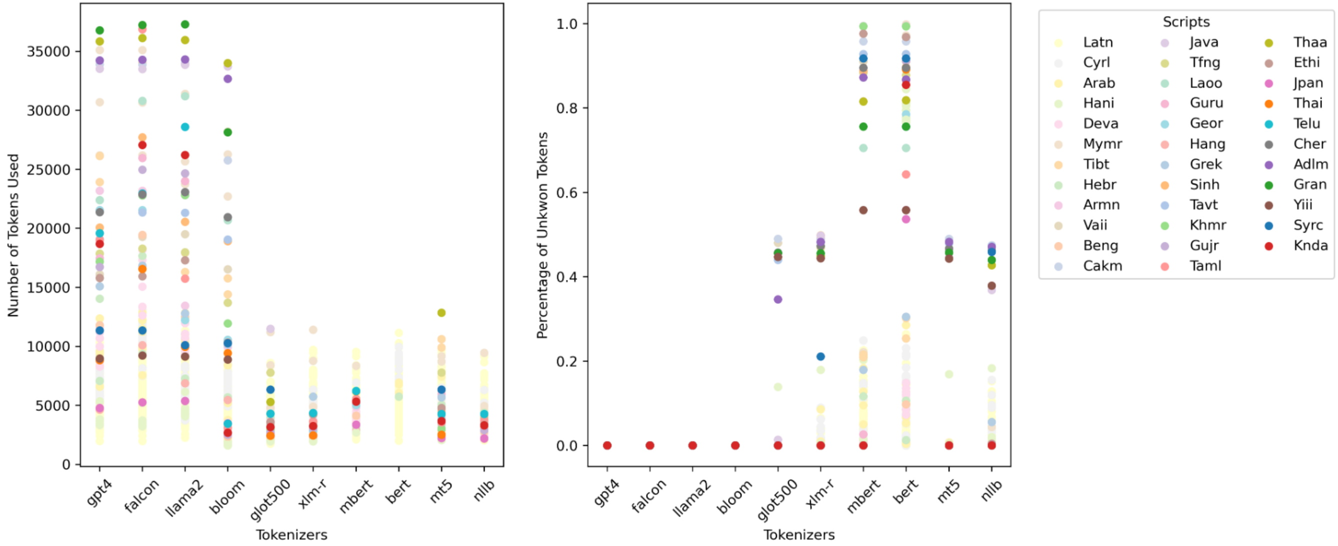


Figure 3: Analysis of the multilinguality of the tokenization of ten language models. This analysis was performed on 396 UDHR translations. Left: the number of tokens into which the UDHR translation is tokenized. We omit a pair of tokenizer and translation with more than 5% unknown tokens. Right: the percentage of unknown tokens generated for a pair of tokenizer and translation.

Conclusion

- We published **GlotScript-R**, an extensive resource covering writing systems for over 7,000 languages.
- We open source **GlotScript-T**, a script identification tool that supports all 161 scripts in Unicode 15.0.

https://arxiv.org/abs/2309.13320

https://github.com/cisnlp/GlotScript