GlotLID: Language Identification for Low-Resource Languages

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, Hinrich Schütze

Center for Information and Language Processing, LMU Munich, Germany
Munich Center for Machine Learning (MCML), Germany
Sorbonne Université, CNRS, ISIR, France

EMNLP 2023

Introduction

GlotLID

We introduce GlotLID, a **language identification** (LID) model that

(i) is open-source.
(ii) covers a wide range of languages, more than **1600 languages.**
(iii) is rigorously evaluated and reliable.
(iv) is efficient and easy to use.

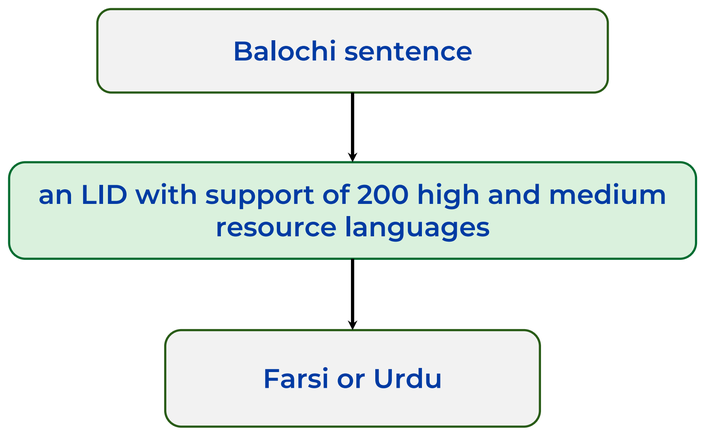
https://github.com/cisnlp/GlotLID

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

Background and Methodology

LID models in general don't have an ability to say they don't know a language.

LID should support a **broad coverage of languages** to minimize out-of model cousin errors.



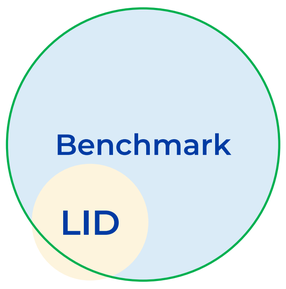
Model:
- We choose FastText model as the GlotLID architecture.
  - scalable, open-source, ease of use, efficient, provide confidence thresholds

Training Data:
- We only use sources we deem trustworthy for GlotLID training.
  - Wikipedia, religious texts, collaborative translations, academia, storybooks, and news sites.
  - This gives us a coverage of 1832 languages, more than any other public LID

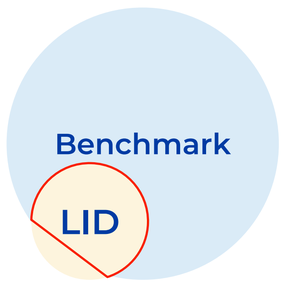
Evaluation Data:
- Flores-200
- UDHR (Universal Declaration of Human Rights)
- Our test set

Comparison Results of GlotLID with Off-the-shelf LIDs

			FLORES-200								UDHR							
			CLD3		FT176		OpenLID		NLLB		CLD3		FT176		OpenLID		NLLB	
			$\|L\|=96$		$\|L\|=108$		$\|L\|=195$		$\|L\|=188$		$\|L\|=100$		$\|L\|=124$		$\|L\|=159$		$\|L\|=172$	
	LID Model	θ	F1↑	FPR↓	F1↑	FPR↓	F1↑	FPR↓	F1↑	FPR↓	F1↑	FPR↓	F1↑	FPR↓	F1↑	FPR↓	F1↑	FPR↓
SET?	baselines	.0	.753	.0098	.775	.0090	.923	.0051	.947	.0053	.544	.0099	.566	.0079	.645	.0056	.641	.0051
SET?	baselines	θ_1	.779	.0081	.816	.0033	.923	.0050	.948	.0051	.576	.0081	.644	.0025	.676	.0046	.677	.0040
SET?	baselines	θ_2	.799	.0060	.796	**.0021**	**.923**	.0044	**.947**	.0047	.618	.0060	.647	**.0014**	.718	.0034	.717	.0030
SET?	GlotLID-M	.0	.978	.0051	.987	.0042	.916	.0043	**.947**	.0035	.868	.0033	.868	.0030	**.848**	.0020	**.847**	.0019
SET?	GlotLID-M	.3	.980	.0042	.987	.0037	.898	.0020	.927	.0019	.881	.0028	.879	.0026	.846	.0015	.844	.0015
SET?	GlotLID-M	.5	**.980**	**.0031**	**.987**	.0029	.886	**.0014**	.916	**.0013**	**.903**	**.0023**	**.890**	.0021	.847	**.0012**	.846	**.0011**
SET!	baselines	.0	.952	**.0104**	.881	**.0093**	**.923**	**.0051**	.950	**.0053**	.922	.0101	.739	**.0081**	.881	**.0063**	.854	**.0058**
SET!	GlotLID-M	.0	**.983**	**.0104**	**.991**	**.0093**	.922	**.0051**	**.954**	**.0053**	**.952**	**.0100**	**.927**	**.0081**	**.926**	.0064	**.925**	.0060



SET?: Benchmark is not known. Apply LID on the whole benchmark.



SET!: Benchmark is known. Apply LID on the intersection of LID supported languages and benchmark.

Θ is the confidence threshold. If the confidence score for a predicted label falls below the threshold, the model should label the input text as "undetermined".

Contact Us

https://arxiv.org/abs/2310.16248

amir@cis.lmu.de

@amir_nlp