Tradutor

Building a Variety Specific Translation Model

Hugo Sousa[1, 2]*, Satya Almasian[3]*, Ricardo Campos[2, 4], Alípio Jorge[1, 2]

[1] University of Porto, Portugal [2] INESC TEC, Portugal [3] Hidelberg University, Germany [4] University of Beira Interior, Portugal

Association for the Advancement of Artificial Intelligence

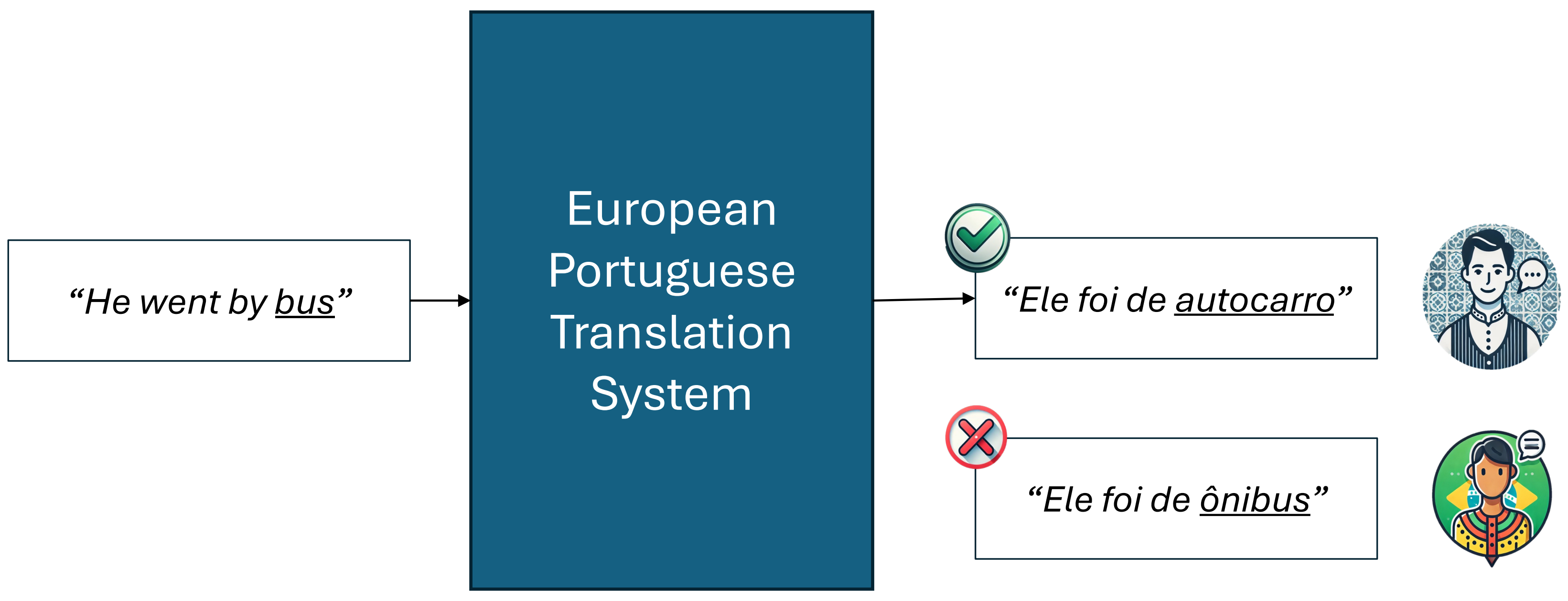
The 39th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, PENNSYLVANIA, USA

Can we develop a ***language variety translation*** *system?*

* Only having a corpus in the language variety

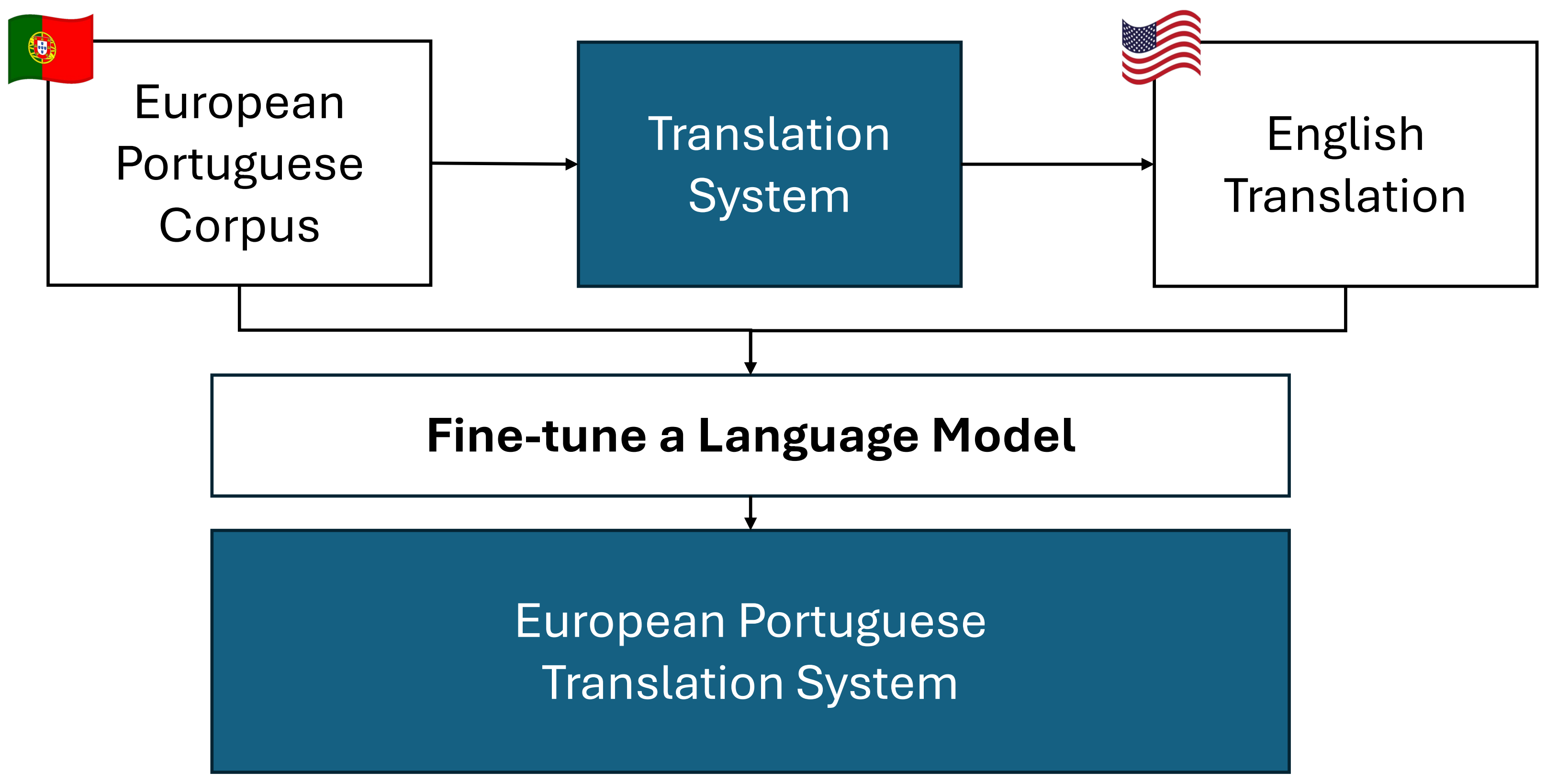
Portuguese as a Case Study

Although Portuguese has a lot of linguistic resources, most are dominated by Brazilian Portuguese, leaving other varieties underrepresented. This bias affects machine translation systems, which typically default to Brazilian Portuguese, resulting in suboptimal translations for European Portuguese. To address this, we introduce Tradutor, the first open-source translation model specifically tailored for European Portuguese.



Approach

We start with a European Portuguese corpus, translate it into English using a system that does not distinguish between varieties, and use the resulting parallel data to fine-tune a language model. This retro-translation approach overcomes data scarcity, enabling accurate translations tailored to European Portuguese.

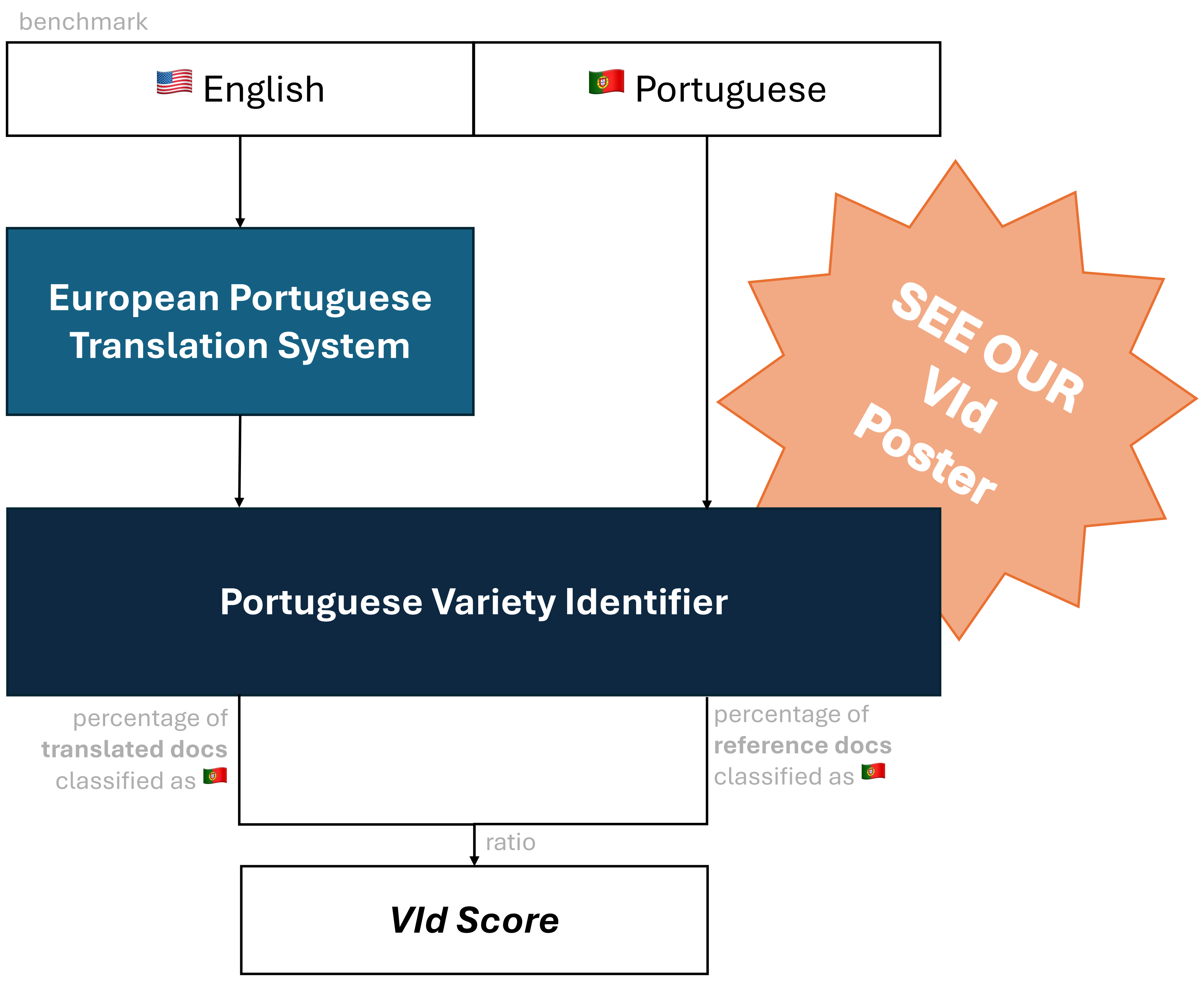

European Portuguese Corpus
Translation System
English Translation
Fine-tune a Language Model
European Portuguese Translation System

Dataset

Domain	# Docs	# Tokens PT	# Tokens EN
		Total	Total
Journalistic	1,250,982	253,767,361	188,072,054
Literature	12,082	1,461,651	1,085,296
Web	9,006	2,024,062	1,504,751
Politics	477	116,836	81,801
Legal	282,870	24,635,676	18,346,240
Social Media	163,585	11,622,673	9,025,327
DSL-TL (news)	1,734	110,334	81,821
All	1,719,002	293,628,259	218,115,469

Variety Identification Score

To evaluate whether our system produces translations in European Portuguese, we use a language variety classification model to assess the percentage of translated texts labeled as European Portuguese. We compare this with the percentage in the reference translations, computing the VId score as their ratio. This metric quantifies how well our model preserves the intended language variety.


benchmark
English
Portuguese
European Portuguese Translation System
SEE OUR VId Poster
Portuguese Variety Identifier
percentage of translated docs classified as
percentage of reference docs classified as
ratio
VId Score

Results

Our results show that fine-tuning significantly improves translation quality while ensuring linguistic alignment with European Portuguese. Our best model outperforms open-source baselines on all metrics and approaches industry-level systems.

	Model	FRMT			
		BLEU	ROUGE-L	COMET	VID
Close Baselines	Google_{br}	43.20	68.43	87.44±0.25	0.445
	Google_{pt}	47.81	71.66	87.87±0.25	0.956
	DeepL	49.77	72.44	88.48±0.23	0.999
Open Baselines	Argos	38.39	65.07	83.99±0.35	0.511
	Opus-MT	40.41	66.25	85.67±0.31	0.413
Zero-shot	Gemma-2	25.37	49.56	75.66±0.51	0.807
	Phi-3	17.59	43.99	57.90±0.56	0.942
	LLaMA-3	31.47	60.61	82.95±0.40	0.811
LoRA	Gemma-2	19.83	56.87	79.62±0.64	0.530
	Phi-3	24.70	53.34	72.19±0.58	**1.178**
	LLaMA-3	25.42	51.51	74.06±0.56	1.092
FFT	Gemma-2	33.76	66.41	85.25±0.35	1.066
	Phi-3	38.16	66.31	85.35±0.34	1.055
	LLaMA-3	**41.12**	**66.92**	**86.12**±0.28	0.968

PAPER

MODEL

CODE