Robustification of Multilingual Language Models to Real-world Noise in Crosslingual Zero-shot Settings with Robust Contrastive Pretraining

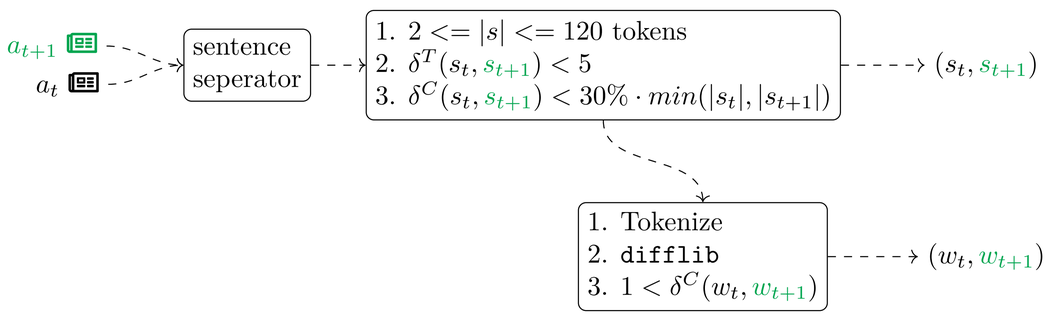
Asa Cooper Stickland[*1,2] **Sailik Sengupta**[*1] **Jason Krone**[1] **He He**[1,3] **Saab Mansour**[1]

[1]aws AI Labs [2]University of Edinburgh [3]New York University

Multilingual Language Models are **not inherently robust** to real-world noise.

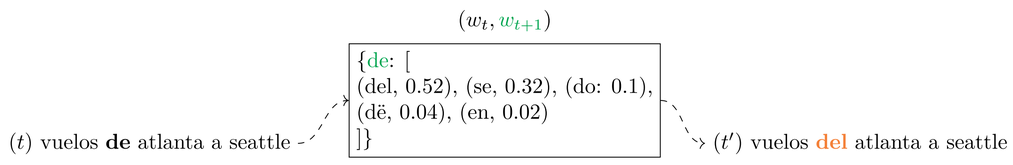
Edit corpus + Contrastive loss = 👆 robustness

Edit Corpus Mining



Evaluation Set Creation

Edit-data injection

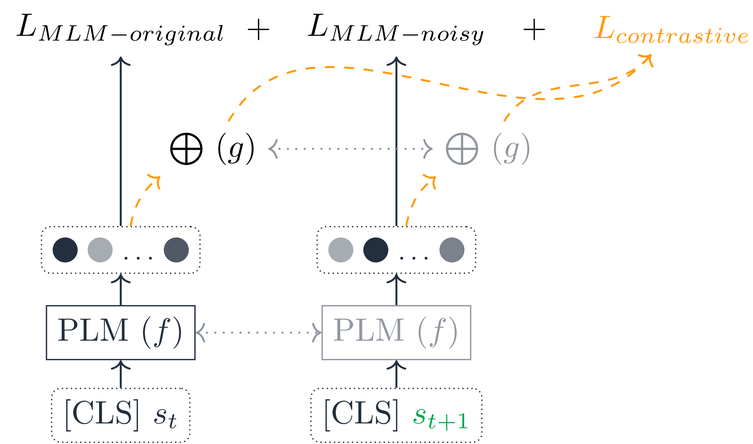


Quality Assurance with Language Experts

Language	Noise Injection Ratio	Realistic Utt. %	Realistic Examples (test-set)	Unrealistic Examples (test-set)
French (fr)	0.1	95.4%	Me montré les vols directs de Charlotte à Minneapolis mardi matin . Quelle compagnie aérienne fut YX	Me montré des vols entre Détroit er St. Louis sur Delta Northwest US Air est United Airlines . Lister des vols de Las Vegas à Son Diego
German (de)	0.2	94.5%	Zeige mir der Flüge zwischen Housten und Orlando Welche Flüge gibt es vom Tacoma nach San Jose	Zeige mit alle Flüge vor Charlotte nach Minneapolis zum Dienstag morgen Zeige mit Flüge an Milwaukee nach Washington DC v. 12 Uhr
Spanish (es)	0.1	96.9%	qué aerolíneas vuelan de baltimore a san francesc muéstrame vuelos entr toronto y san diego	necesito información de un vuelo y la tarifa de oakland a salt lake city para el jueves antes e sus 8 am de nuevo york a las vegas el domingo con la tarde
Hindi (hi)	0.05	95.4%	मुझे डेल्टा उड़ानों के बारे में बताइए जो कोच के यात्रियों को नाश्ता देता हों मुझे मेम्फिस से लास वेगास तक उड़ान की जरूरत है	सोमवार दोपहर ने लॉस एंजिल्स से पिट्सबर्ग रविवार दोपहर को मियामी में क्लीवलैंड
Japanese (jp)	0.1	**92.3%**	来F水曜日にカンザスシティ初シカゴ行きでシカゴ の 午後 7 時ごろ到着して、Fりのフライトが木曜日のフライト ワシントン を コロンバス間のすべてのフライトの運賃はいくら	シャFロット空港 の 土曜日 err 午后 1 時に出Fする US エアFのフライトをリストアップして 水曜日のフェニックスFミルウォFキF逝き
Chinese (zh)	0.1	**86.2%**	我需要 4 点 后 在 达拉斯起飞飞往旧金山的联程航班 请列出从纽瓦克飞往 洛杉机 的航班	然而 每天上午 10 点之前从密尔沃基飞往亚特兰大 拉瓜迪亚 了 豪华轿车服务要多少钱

Robust Contrastive Pretraining

Encourage the model to learn that s_t and s_{t+1} are similar (+ve pairs).



Results

Improves performance on noisy data and clean data!

Task	Metric	XLMR	XLMR +p(aug)	XLMR +t(En-aug)	XLMR +RCP (Ours)	XLMR +RCP+t (Ours)	Gain
	IC%	89.65	93.10	91.26	93.80	**94.57**	+4.92
	SL-F1	62.30	67.47	74.62	67.45	**80.68**	+18.38
	IC%	90.46	93.98	91.60	93.79	**94.53**	+4.07
	SL-F1	61.63	66.67	66.44	67.69	**70.20**	+8.57
Wiki-ann	NER-F1	69.48	72.32	-	**72.37**	-	+2.89
XNLI	NLI%	74.38	74.83	-	**75.06**	-	+0.68

Task	Metric	XLMR	Ours	Gain
	IC%	90.68	95.32	+4.64
	SL-F1	71.45	84.07	+12.62
	IC%	92.93	95.66	+2.73
	SL-F1	68.01	74.39	+6.38
Wiki-ann	NER-F1	74.14	76.34	+2.2
XNLI	NLI%	76.69	76.75	+0.06

We notice a significant improvement in slot-labeling performance.

Improvement in slot-label classification ($2\times$ `de`, $2.6\times$ `es, hi`, $4\times$ `fr`)

We see a sharp drop in *hallucination* errors across all languages.

N/O	Model	de	es	fr	hi
Noisy	XLMR	315	358	413	671
	XLMR+RCP+t	21	123	33	204
Original	XLMR	208	262	334	460
	XLMR+RCP+t	19	106	22	180

↑ Explicability of errors

👎 fromloc.airport_code → date
👍 fromloc.airport_code → toloc.airport_code

↓ Hallucination errors

Model identifies irrelevant tokens as slot values. Eg.

"Ichs brauche einen Flug von Memphis nach Tacoma, der uber Los Angeles fliegt."

👍 O (über) ⟶ 👎 airline_code (uber)

Scan me

amazon | science