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 Krone1  He He1,3  Saab Mansour1

1 AWS AI Labs
2 University of Edinburgh
3 New York University

💻 EACL 2023
🇦🇷 Dubrovnik
Text Classification

Sentence-level classification task (eg. Intent Classification, XNLI, etc.):
- Croatia is such a lovely place → +ve

Token-level classification task (eg. NER, Slot-labeling):
- **Croatia** is such a lovely place → \{Croatia: Country\}
Text Classification

Sentence-level classification task (eg. Intent Classification, XNLI, etc.):
- Croatia is such a lovely place $\rightarrow$ +ve
- Croatia is suhc a lovely place $\rightarrow$ ?

Token-level classification task (eg. NER, Slot-labeling):
- **Croatia** is such a lovely place $\rightarrow$ {**Croatia**: Country}
- Croatia is suhc a lovely place $\rightarrow$ ?

🤔 What happens when faced with real-world noise?

In this work, we study this question for languages beyond English.
1. Related Work

2. Evaluation Mechanisms
   - Finding Noisy Data
   - Creating Evaluation Test-sets
   - Multilingual Noise Characteristics

3. Robust Contrastive Pretraining

4. Experiments
   - Experimental Setup
   - Robustness of Multilingual Models
   - A Study of Errors
Related Work

• **Works have investigated the impact of various noise types, mostly for English** – misspellings [BB17, KLEG19, MKS21], casing [vMvdLCFK20], paraphrases [EGMS19], morphological variance [TJKS20], synonyms [SKM21], dialectical variance [SLS+22]

• **Methods to improve robustness of SOTA models have considered** – Data augmentation during pre-training [TJKS20, SLS+22] or the task-training stage [PLZ+21], token-free models motivate robustness in multilingual settings [CGTW21, XBC+22, TTR+21], Adversarial Logit Pairing [EGMS19]

• **Our works is similar to works in computer vision that have used of Contrastive learning to boost model robustness** [FLC+21, GL21, JCCW20, KTH20]
Table of Contents

1 Related Work

2 Evaluation Mechanisms
   • Finding Noisy Data
   • Creating Evaluation Test-sets
   • Multilingual Noise Characteristics

3 Robust Contrastive Pretraining

4 Experiments
   • Experimental Setup
   • Robustness of Multilingual Models
   • A Study of Errors
Finding Noisy Data

There is a lack of benchmark to investigate the robustness of multilingual models. Why? 😐
Finding Noisy Data

There is a lack of benchmark to investigate the robustness of multilingual models. *Why? 😞*

Synthetic noise-generation methods (mostly developed for English) need linguistic expertise to create benchmarks for individual languages. *What to do? 😞*
Finding Noisy Data

There is a lack of benchmark to investigate the robustness of multilingual models. Why? 😕

Synthetic noise-generation methods (mostly developed for English) need linguistic expertise to create benchmarks for individual languages. What to do? 😞

Can we find a data source from where we can obtain such data?

💡 Wikipedia articles are continually updated/edited. Maybe we can mine these edits. (We also leverage other corpora such as Lang8.)
Similar to [TMKK20], we obtain sentence edit dictionaries and word-edit dictionaries.
Creating Evaluation Test-sets

💡 Use word-edit dictionaries for noisy test-set creation!

We note that this makes our test-data limited to work level edits. But, we can have multiple words manipulated in a single utterance.

\[ (w_t, w_{t+1}) \]

\{de: 
(del, 0.52), (se, 0.32), (do: 0.1), 
(dë, 0.04), (en, 0.02) \}

\( (t) \) vuelos de atlanta a seattle → 
\( (t') \) vuelos del atlanta a seattle
Creating Evaluation Test-sets – QA

We inject various degrees of noise and conduct evaluation to decide which test sets are more realistic.

We keep test-data on if they have < 5% unrealistic errors.

<table>
<thead>
<tr>
<th>Language</th>
<th>Noise Injection Ratio</th>
<th>Realistic Utt. %</th>
<th>Realistic Examples (test-set)</th>
<th>Unrealistic Examples (test-set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>French (fr)</td>
<td>0.1</td>
<td>95.4%</td>
<td>Me montré les vols directs de Charlotte à Minneapolis mardi matin.</td>
<td>Me montré des vols entre Détroit et St. Louis sur Delta Northwest US Air est United Airlines.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Quelle compagnie aérienne fut YX</td>
<td>Lister des vols de Las Vegas à Son Diego</td>
</tr>
<tr>
<td>German (de)</td>
<td>0.2</td>
<td>94.5%</td>
<td>Zeige mir der Flüge zwischen Housten und Orlando</td>
<td>Zeige mit alle Flüge vor Charlotte nach Minneapolis zum Dienstag morgen</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Welche Flüge gibt es vom Tacoma nach San Jose</td>
<td>Zeige mit Flüge an Milwaukee nach Washington DC v. 12 Uhr</td>
</tr>
<tr>
<td>Spanish (es)</td>
<td>0.1</td>
<td>96.9%</td>
<td>qué aerolíneas vuelan de baltimore a san francisco</td>
<td>necesito información de un vuelo y la tarifa de oakland a salt lake city para el jueves antes e sus 8 am</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>muestrame vuelos entr toronto y san diego</td>
<td>de nuevo york a las vegas el domingo con la tarde</td>
</tr>
<tr>
<td>Hindi (hi)</td>
<td>0.05</td>
<td>95.4%</td>
<td>मूसे डेल्टा उड़ानों के बारे में बताइए जो कोष के \ यात्रियों को गाजरा देता है</td>
<td>सोमवार दोपहर ने लॉस एंजिल्स से पिट्सबर्ग</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>मूसे मेनजर से नाम बेगाम तक उड़ान की जारीत है</td>
<td>ग्रिल्वार दोपहर को मियामी में कर्नीकोड</td>
</tr>
<tr>
<td>Japanese (jp)</td>
<td>0.1</td>
<td>92.3%</td>
<td>来国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、再び水曜日のフライト</td>
<td>シャロット空港の土曜日 err 午後 1 時に出発する US エアのフライトをリストアップして</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>のフライトが木曜日のフライト</td>
<td>水曜日のフェニックス国ミルウォekteキ去</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>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</td>
<td>然而 每天上午10点之前从密尔沃基飞往亚特兰大</td>
</tr>
<tr>
<td>Chinese (zh)</td>
<td>0.1</td>
<td>86.2%</td>
<td>我需要4点后在达拉斯起飞往旧金山的联程航班</td>
<td>拉瓜迪亚了豪华轿车服务要多少钱</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Noise Injection Ratio</th>
<th>Realistic Utt. %</th>
<th>Realistic Examples (test-set)</th>
<th>Unrealistic Examples (test-set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>French (fr)</td>
<td>0.1</td>
<td>95.4%</td>
<td>Me montré les vols directs de Charlotte à Minneapolis mardi matin.</td>
<td>Me montré des vols entre Détroit et St. Louis sur Delta Northwest US Air est United Airlines.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Quelle compagnie aérienne fut YX</td>
<td>Lister des vols de Las Vegas à Son Diego</td>
</tr>
<tr>
<td>German (de)</td>
<td>0.2</td>
<td>94.5%</td>
<td>Zeige mir der Flüge zwischen Housten und Orlando</td>
<td>Zeige mit alle Flüge vor Charlotte nach Minneapolis zum Dienstag morgen</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Welche Flüge gibt es vom Tacoma nach San Jose</td>
<td>Zeige mit Flüge an Milwaukee nach Washington DC v. 12 Uhr</td>
</tr>
<tr>
<td>Spanish (es)</td>
<td>0.1</td>
<td>96.9%</td>
<td>qué aerolíneas vuelan de baltimore a san francisco</td>
<td>necesito información de un vuelo y la tarifa de oakland a salt lake city para el jueves antes e sus 8 am</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>muestrame vuelos entr toronto y san diego</td>
<td>de nuevo york a las vegas el domingo con la tarde</td>
</tr>
<tr>
<td>Hindi (hi)</td>
<td>0.05</td>
<td>95.4%</td>
<td>मूसे डेल्टा उड़ानों के बारे में बताइए जो कोष के \ यात्रियों को गाजरा देता है</td>
<td>सोमवार दोपहर ने लॉस एंजिल्स से पिट्सबर्ग</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>मूसे मेनजर से नाम बेगाम तक उड़ान की जारीत है</td>
<td>ग्रिल्वार दोपहर को मियामी में कर्नीकोड</td>
</tr>
<tr>
<td>Japanese (jp)</td>
<td>0.1</td>
<td>92.3%</td>
<td>来国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、再び水曜日のフライト</td>
<td>シャロット空港の土曜日 err 午後 1 時に出発する US エアのフライトをリストアップして</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>のフライトが木曜日のフライト</td>
<td>水曜日のフェニックス国ミルウォekteキ去</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>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</td>
<td>然而 每天上午10点之前从密尔沃基飞往亚特兰大</td>
</tr>
<tr>
<td>Chinese (zh)</td>
<td>0.1</td>
<td>86.2%</td>
<td>我需要4点后在达拉斯起飞往旧金山的联程航班</td>
<td>拉瓜迪亚了豪华轿车服务要多少钱</td>
</tr>
</tbody>
</table>
Human evaluation of injected noise surfaces many interesting insights.

- Certain noise types are language specific (e.g. `jp` has conversion, `tr` has anglicization errors).
- Certain noise types are common across languages (although `zh` has less typos due to pinyin style keyboards).

See our paper for more [Sec 3.3].
# Table of Contents

1. Related Work

2. Evaluation Mechanisms
   - Finding Noisy Data
   - Creating Evaluation Test-sets
   - Multilingual Noise Characteristics

3. Robust Constrastive Pretraining

4. Experiments
   - Experimental Setup
   - Robustness of Multilingual Models
   - A Study of Errors
Use sentence-edit dictionaries to pre-train multilingual models!

The intuition is that this will teach these multilingual models to represent incorrect and edited sentence closer to one another.

\[
L_{MLM-original} + L_{MLM-noisy} + L_{contrastive} \quad [Soh16]
\]

\[
\oplus (g) \quad [GNWB21, RG19]
\]
Table of Contents

1 Related Work

2 Evaluation Mechanisms
   • Finding Noisy Data
   • Creating Evaluation Test-sets
   • Multilingual Noise Characteristics

3 Robust Contrastive Pretraining

4 Experiments
   • Experimental Setup
   • Robustness of Multilingual Models
   • A Study of Errors
The training data is the original English training set of each task. Test data had two splits for each language—the original test set (Original) and the noise-added test set (Noisy).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>Training size (only en)</th>
<th>Languages (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiATIS++ [XHM20]</td>
<td>IC/SL</td>
<td>5k</td>
<td>de,en,es,fr,hi</td>
</tr>
<tr>
<td>+ training data aug. (en)</td>
<td></td>
<td>18k</td>
<td>de,en,es,fr,hi</td>
</tr>
<tr>
<td>MultiSNIPS</td>
<td>IC/SL</td>
<td>13k</td>
<td>en,es,fr,hi</td>
</tr>
<tr>
<td>+ training data aug. (en)</td>
<td></td>
<td>72k</td>
<td>en,es,fr,hi</td>
</tr>
<tr>
<td>WikiANN [PZM+17]</td>
<td>NER</td>
<td>20k</td>
<td>de,en,es,fr,hi,te</td>
</tr>
<tr>
<td>XNLI [CRL+18]</td>
<td>NLI</td>
<td>392k</td>
<td>de,es,fr,hi,te</td>
</tr>
</tbody>
</table>

**Multilingual Model Robustness (as-is)**

\[
XLM-R_{base} [CKG+20] > m-BERT [DCLT19] > Canine-c [CGTW21]
\]
### Robustness of Multilingual Models

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

RCP ↑ model robustness across all tasks metrics – Accuracy of IC & XNLI, F1-score for SL & NER (avg across languages).

Gains ↑↑ when agg. English noise data [SKM21] is used during task-time augmentation.
Robustness of Multilingual Models

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

RCP ↑ model robustness across all tasks metrics – Accuracy of IC & XNLI, F1-score for SL & NER (avg across languages).

Gains ↑↑ when agg. English noise data [SKM21] is used during task-time augmentation.

RCP ↑ model performance on clean data too!
A Study of Errors (on MultiATIS++)

Improvement in slot-label classification \((2 \times \text{de}, 2.6 \times \text{es}, \text{hi}, 4 \times \text{fr})\)

Our model is better | Baseline (XLMR) is better | Equal

<table>
<thead>
<tr>
<th>Language</th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>de</td>
<td>24</td>
<td>12</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>es</td>
<td>31</td>
<td>12</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fr</td>
<td>31</td>
<td>8</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hi</td>
<td>32</td>
<td>12</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

↑ Explicability of errors [OSK20]

- fromloc.airport_code → date
- fromloc.airport_code → toloc.airport_code
A Study of Errors (on MultiATIS++)

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

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

<table>
<thead>
<tr>
<th>N/O</th>
<th>Model</th>
<th>de</th>
<th>es</th>
<th>fr</th>
<th>hi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy</td>
<td>XLMR</td>
<td>315</td>
<td>358</td>
<td>413</td>
<td>671</td>
</tr>
<tr>
<td></td>
<td>XLMR+RCP+t</td>
<td>21</td>
<td>123</td>
<td>33</td>
<td>201</td>
</tr>
<tr>
<td>Original</td>
<td>XLMR</td>
<td>208</td>
<td>262</td>
<td>334</td>
<td>460</td>
</tr>
<tr>
<td></td>
<td>XLMR+RCP+t</td>
<td>19</td>
<td>106</td>
<td>22</td>
<td>180</td>
</tr>
</tbody>
</table>

↑ Explicability of errors [OSK20]

- 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 über Los Angeles fliegt."

👍 O (über) → 👎 airline_code (uber)


- Multilingual test data to evaluate the robustness of multilingual models to noise.
- Performance of existing multilingual language models deteriorates on four tasks when tested on the noisy test data.
- Robust Contrastive Pretraining (RCP) can boost the robustness of existing multilingual language models.

Data & Code

https://github.com/amazon-science/multilingual-robust-contrastive-pretraining
Conclusion

- Multilingual test data to evaluate the robustness of multilingual models to noise.
- Performance of existing multilingual language models deteriorates on four tasks when tested on the noisy test data.
- Robust Contrastive Pretraining (RCP) can boost the robustness of existing multilingual language models.

Data & Code

https://github.com/amazon-science/multilingual-robust-contrastive-pretraining

Questions?


