COLORADO STATE UNIVERSITY

Cross-Lingual Transfer Robustness to Lower-Resource Languages on Adversarial Datasets

Shadi Manafi and Nikhil Krishnaswamy

LREC-COLING 2024, Turin, Italy

Introduction

https://medium.com/omnius/hallo-multilingual-bert-c%C3%B3mo-funcionas-2b3406cc4dc2

- **Multilingual Language Models (MLLMs)** exhibit robust **cross-lingual transfer** capabilities.
- Cross-lingual transfer: ability to leverage a information acquired in a fine-tuned on a task for a source language and apply it to a target language.
- Zero-shot learning may sometimes rely on “**vocabulary memorization**” rather than true language **understanding**.
- Realizing why this is the case for particular tasks is tough due to **language differences and specific domain variations**.
- Objectives:
  - How does zero-shot learning accuracy shift with **minor input variations**?
  - How do language features, like **shared vocabulary**, affect zero-shot learning?

Related Works

Input: pull[s] off the rare trick of recreating not only the look of a certain era , but also the feel .
Output: pull[s] off the seldom trick of recreating not only the look of a certain era , but also the feel .

https://aclanthology.org/2020.emnlp-demos.16.pdf

- **Adversarial data** creation for NLP
  - **Surface-level** text modifications
    - Inserting, deleting, or swapping words, characters, or sentences (Gao et al., 2018; Ribeiro et al., 2018; Jia and Liang, 2017)
  - **Semantic-level** alternative strategies
    - Paraphrasing (Iyyer et al., 2018)
    - Generating text with semantically analogous content using neural models (Zhao et al., 2018; Michel et al., 2019)
    - Human-in-the-loop interventions (Wallace et al., 2019)

Datasets

- **Named Entity Recognition (NER) task**: an **information extraction** (IE) from unstructured texts, encompassing the identification of individuals' names, organizations, geographical locations, etc.
- **WikiANN** dataset (Pan et al., 2017): a common multilingual NER dataset

Datasets

- **Section title prediction task**: a **proxy for document classification**, selection of the most appropriate title for a section text among the four presented choices.
- **We built** the section title prediction corpus (**WikiTitle**):

1. **Crawling** the Wikipedia pages corresponding to each specific language with at least 4 sections.
2. Using the **WikiExtractor** tool (Attardi, 2015) to systematically extract sections along with their associated second and third-level titles from the Wikipedia pages.
3. **Pairing** subsection text with four candidate titles, of which one is correct and the others are titles of other sections of the same article.
4. **Collecting** as many samples as possible for each language **up to a limit of 100,000**.

Datasets

- Focusing on 13 language pairs from a pool of 21 languages.
- Selecting language pairs usually consisting of
  - **High-Resource Language (HRL)**: One language with **greater** resources in the data
  - **Low-Resource Language (LRL)**: One with **fewer** resources
  - Substantial level of **overlap in the vocabulary**
    - **Areal** (French/Breton)
    - **Genetic** relationship (Czech/Slovak)
    - History of **borrowing** at large scale (Arabic/Farsi).
- Arabic/Hindi—serves as a kind of "**control**" group
  - Share a substantial amount of vocabulary due to borrowing
  - But use different native scripts, so low vocabulary overlap level
- Pairwise notation **L1/L2**: **L1** refers to the **HRL** and **L2** refers to **LRL**

HRL	Size	LRL	Size	Relationship
Arabic (ar)	100K	Farsi (fa)	100K	Borrowing
Arabic (ar)	100K	Hindi (hi)	42.6K	Borrowing
Czech (cs)	100K	Slovak (sk)	61.1K	Areal, Genetic
Dutch (nl)	100K	Afrikaans (af)	29.7K	Genetic
English (en)	100K	Scots (sco)	5.1K	Areal, Genetic
English (en)	100K	Welsh (cy)	15.2K	Areal, Borrowing
French (fr)	100K	Breton (br)	8.1K	Areal, Borrowing
French (fr)	100K	Occitan (oc)	13.7K	Areal, Genetic
Indonesian (id)	100K	Malay (ms)	60.3K	Areal, Genetic
Italian (it)	100K	Sicilian (scn)	1.4K	Areal, Genetic
Spanish (es)	100K	Aragonese (an)	5.1K	Areal, Genetic
Spanish (es)	100K	Asturian (ast)	85.5K	Areal, Genetic
Spanish (es)	100K	Catalan (ca)	100K	Areal, Genetic

Size of languages for section title prediction dataset, and relationship between languages in studied pair.

Methodology

- We evaluate two well-known **MLLMs**, which demonstrate strong **cross-lingual transfer** abilities for downstream tasks:
  - **MBERT**: bert-base-multilingual-cased (Devlin et al., 2019)
  - **XLM-R** : xlm-roberta-base (Conneau et al., 2020)
- For two tasks: **NER and section title prediction**
- We evaluate both models in different settings:
  - **Native** setting: they are fully fine-tuned in an LRL
  - **Transfer** setting: they are trained on an HRL and evaluated on the paired LRL
  - Under different **perturbations** of the data

Perturbation Methods

- **Four** main methods to generate adversarial sets:

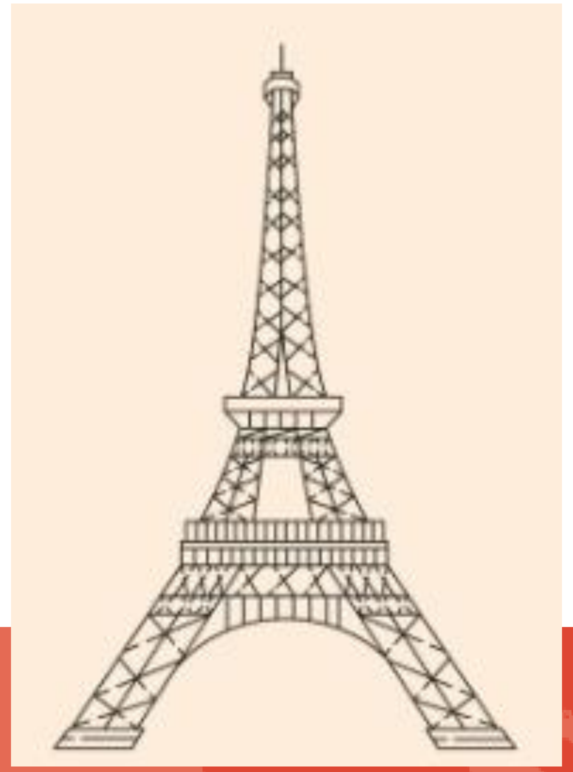
1. **Perturbation #1 (P1)**: Change **given names** (first element) of all PER entities to randomly-chosen elements of the given names dataset in the same language.
   - A dataset of given names for each target language scraped from the **[Language]_given_names** category of Wiktionary.
2. **Perturbation #2 (P2)**: Change **location names** of all LOC entities to randomly-chosen elements of the placenames dataset in the same language.
   - A dataset of places for each target language scraped from its **Places** category in Wiktionary.

Perturbation Methods

3. **Perturbation #3 (P3)**: Replace **named entities** shared between L2 test file and L1 training file with **named entities with the same tag** unique to L2**.**
   - ***Eiffel***: same in French and Breton. Replaced with ***Bolz-enor Pariz*** (Arc de Triomphe), which is the same NER type, but non-overlapping.
4. **Perturbation #4 (P4)**: take ***surrounding* words** shared between L2 test file and L1 training file with the **highest cosine similarity** with the original word unique to L2.
   - "An tour Eiffel", the word "**tour**": same in French and Breton. Replaced with a semantically-similar Breton word, not existing in French like "**kastell**".

Content word	MBERT option	XLM-R option
channels	shots	broadcasts
bred	lived	assistant
population	parted	people
serve	carried	arrangement
place	event	there
journalist	lawyer	activist
female	woman	woman
hijackers	triumphs	males
defeated	won	defeating

Sample of highest cosine-similarity alternatives existing in the test split of the English dataset.



Eiffel Tower: English
Tour Eiffel: French, Breton

https://aclanthology.org/2020.emnlp-demos.16.pdf

Computing Vocabulary Overlap - NER

- Extracting all labeled NER chunks.
- $\%\ \text{overlap L1/L2} = \dfrac{\text{number of } \textbf{shared} \text{ entities with similar tags between L1 and L2}}{\textbf{total} \text{ number of entities in L2}}$

L1	L2	% overlap
ar	hi	4.88
ar	fa	19.94
cs	sk	39.55
nl	af	31.57
en	sco	25.19
en	cy	22.07
fr	br	23.33
fr	oc	23.61
it	scn	43.17
id	ms	41.87
es	an	46.26
es	ast	47.66
es	ca	36.77

Computing Vocabulary Overlap – Section Title

- % overlap L1/L2 =

  $$\frac{\text{number of } \textbf{shared} \text{ words between between L1 and L2}}{\textbf{total} \text{ number of words in L2}}$$

- Considering only the first 128 tokens from each section.
- Due to variances in tokenization between MBERT and XLM-R, the overlap percentage would be different.

L1	L2	Model	% overlap
ar	hi	MBERT	2.12
ar	hi	XLM-R	1.98
ar	fa	MBERT	14.65
ar	fa	XLM-R	15.01
cs	sk	MBERT	24.26
cs	sk	XLM-R	24.18
nl	af	MBERT	22.63
nl	af	XLM-R	22.57
en	sco	MBERT	29.22
en	sco	XLM-R	29.19
en	cy	MBERT	17.31
en	cy	XLM-R	17.08
fr	br	MBERT	9.50
fr	br	XLM-R	9.44
fr	oc	MBERT	23.09
fr	oc	XLM-R	23.04
id	ms	MBERT	36.34
id	ms	XLM-R	36.34
it	scn	MBERT	25.99
it	scn	XLM-R	25.86
es	an	MBERT	24.80
es	an	XLM-R	24.77
es	ast	MBERT	29.59
es	ast	XLM-R	29.65
es	ca	MBERT	17.12
es	ca	XLM-R	17.20

Results – Native and Transfer

- Most Initial HRL→LRL **transfer** performance do **not reach** the **native** LRL fine-tuning, falling below by **~1-30%** F1/accuracy.
- Cross-lingual transfer goes **closer** to native for closer language pairs **geographically** and **genetically**

		MBERT								XLM-R							
		NER						WikiTitle		NER						WikiTitle	
Train	Test	Base	P1	P2	P3	P4	P5	Base	P4	Base	P1	P2	P3	P4	P5	Base	P4
ar	hi	67.2	64.2	68.9	67.2	67.2	67.2	63.6	63.0	67.3	67.4	70.7	67.3	67.3	67.3	75.8	75.0
hi	hi	86.7	86.5	87.2	71.3	79.0	66.7	73.8	72.5	87.5	87.2	88.1	76.6	80.7	68.3	77.8	77.1
ar	fa	45.0	43.0	44.7	45.0	45.0	44.9	79.3	77.1	43.6	42.8	40.1	43.6	43.5	43.4	78.0	73.9
fa	fa	90.3	88.0	89.1	86.5	60.8	56.7	81.6	79.1	89.4	88.2	87.4	85.5	78.2	74.1	81.0	76.5
cs	sk	82.9	82.4	87.0	78.4	82.5	77.9	80.3	75.6	78.0	77.2	86.1	73.4	78.1	73.5	80.3	73.3
sk	sk	92.6	91.7	91.0	86.4	92.1	85.0	83.5	78.5	91.5	91.1	89.8	81.5	88.6	77.5	82.3	75.1
nl	af	81.2	81.0	83.8	78.4	81.2	78.6	78.5	71.6	79.9	80.0	81.5	77.8	79.3	76.9	75.4	71.6
af	af	92.2	91.6	92.1	81.1	89.5	78.5	81.3	74.3	89.8	90.0	90.8	77.9	86.2	76.0	76.8	66.9
en	sco	78.3	77.9	72.0	71.0	78.2	71.7	85.7	76.2	62.4	62.0	60.6	60.6	63.2	61.3	75.5	62.5
sco	sco	93.4	93.0	83.2	81.0	91.4	79.2	88.6	80.8	90.2	89.6	82.5	79.6	87.5	75.0	71.5	60.2
en	cy	62.5	61.8	65.3	61.3	62.4	61.6	67.5	63.6	61.5	61.2	64.9	60.4	61.4	60.4	61.7	58.8
cy	cy	92.6	91.9	87.1	77.0	89.5	75.0	76.6	73.5	90.9	90.4	85.1	76.1	83.1	67.8	72.1	67.3
fr	br	74.3	71.8	73.5	73.3	74.2	72.8	66.6	63.1	66.3	64.2	66.6	64.7	66.3	64.5	59.3	54.0
br	br	92.8	88.4	88.2	84.5	88.8	79.9	71.1	66.1	89.1	85.8	87.1	81.3	82.8	74.1	59.3	55.2
fr	oc	83.9	83.7	89.1	83.5	83.7	83.4	76.6	71.9	72.5	72.3	78.8	71.8	72.3	71.9	66.5	59.1
oc	oc	95.3	94.9	95.8	92.3	87.8	83.9	79.1	75.2	93.8	93.0	94.6	91.5	92.6	89.8	67.0	61.3
id	ms	68.7	67.7	76.7	64.8	68.5	64.8	79.9	68.4	69.7	69.5	79.9	66.2	69.5	65.8	78.3	58.4
ms	ms	92.4	92.6	83.5	81.7	81.8	70.5	82.7	71.8	92.4	91.9	89.1	71.7	79.7	59.5	80.3	62.4
it	scn	63.7	63.3	80.2	58.4	49.5	45.4	71.0	66.2	60.8	60.7	74.0	55.3	50.4	45.5	60.7	46.8
scn	scn	92.9	91.1	88.1	79.8	74.4	64.9	64.3	57.1	90.5	88.2	82.8	79.7	72.4	62.5	40.0	39.0
es	an	**88.0**	87.9	84.8	85.4	80.7	77.5	86.1	76.3	**86.1**	86.2	86.4	83.3	75.3	72.9	77.0	55.0
an	an	95.8	95.8	88.4	85.6	90.9	**79.1**	83.4	76.8	94.2	93.6	92.5	79.8	80.4	**66.1**	72.6	59.4
es	ast	**90.4**	90.2	86.0	85.1	89.6	84.6	84.1	77.5	**84.3**	84.2	86.0	77.0	84.1	76.3	76.7	59.6
ast	ast	93.6	92.8	90.1	82.7	93.3	**79.7**	85.2	78.4	89.6	89.2	90.1	77.7	90.0	**76.4**	80.3	68.0
es	ca	**85.1**	84.3	87.2	84.0	85.1	84.0	79.3	75.9	**82.6**	82.8	83.9	80.8	82.3	79.8	72.8	66.2
ca	ca	92.3	91.5	91.6	87.3	91.6	**86.5**	85.9	83.0	89.4	89.6	88.0	83.3	88.6	**82.1**	83.9	78.0

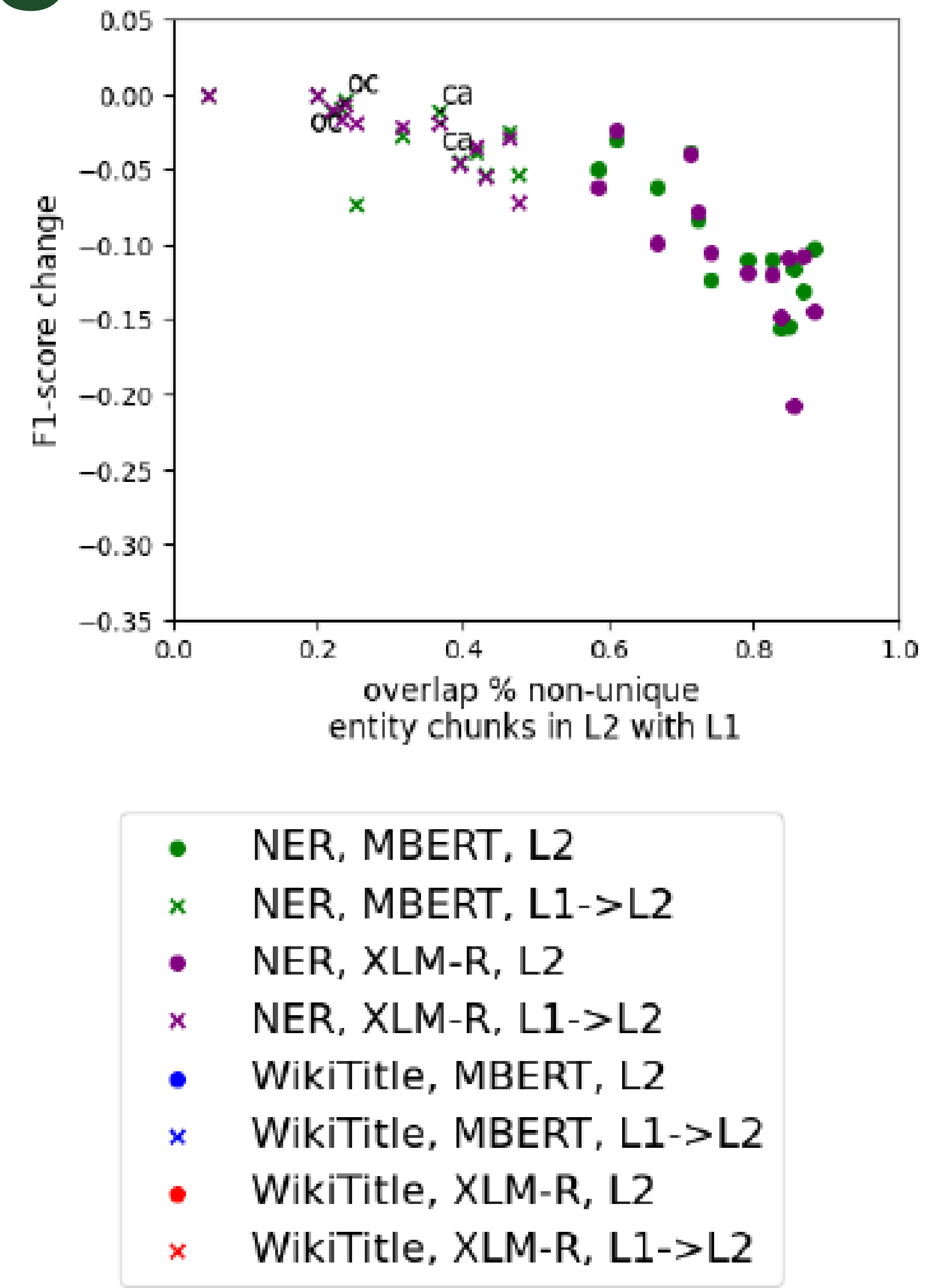
Results – Native and Transfer

- P5: **combination of P3 and P4**
- For three pairs involving **Spanish** (Spanish/Aragonese, Spanish/Asturian, and Spanish/Catalan), P5 brings the **native** model **down** to the performance level of the **unperturbed cross-lingual transfer** model.
- This also suggests that on these LRLs, MLLMs may be **leveraging** their **capabilities in Spanish** to achieve their **initial performances**.

		MBERT								XLM-R							
		NER						WikiTitle		NER						WikiTitle	
Train	**Test**	**Base**	**P1**	**P2**	**P3**	**P4**	**P5**	**Base**	**P4**	**Base**	**P1**	**P2**	**P3**	**P4**	**P5**	**Base**	**P4**
ar	hi	67.2	64.2	68.9	67.2	67.2	67.2	63.6	63.0	67.3	67.4	70.7	67.3	67.3	67.3	75.8	75.0
hi	hi	86.7	86.5	87.2	71.3	79.0	66.7	73.8	72.5	87.5	87.2	88.1	76.6	80.7	68.3	77.8	77.1
ar	fa	45.0	43.0	44.7	45.0	45.0	44.9	79.3	77.1	43.6	42.8	40.1	43.6	43.5	43.4	78.0	73.9
fa	fa	90.3	88.0	89.1	86.5	60.8	56.7	81.6	79.1	89.4	88.2	87.4	85.5	78.2	74.1	81.0	76.5
cs	sk	82.9	82.4	87.0	78.4	82.5	77.9	80.3	75.6	78.0	77.2	86.1	73.4	78.1	73.5	80.3	73.3
sk	sk	92.6	91.7	91.0	86.4	92.1	85.0	83.5	78.5	91.5	91.1	89.8	81.5	88.6	77.5	82.3	75.1
nl	af	81.2	81.0	83.8	78.4	81.2	78.6	78.5	71.6	79.9	80.0	81.5	77.8	79.3	76.9	75.4	71.6
af	af	92.2	91.6	92.1	81.1	89.5	78.5	81.3	74.3	89.8	90.0	90.8	77.9	86.2	76.0	76.8	66.9
en	sco	78.3	77.9	72.0	71.0	78.2	71.7	85.7	76.2	62.4	62.0	60.6	60.6	63.2	61.3	75.5	62.5
sco	sco	93.4	93.0	83.2	81.0	91.4	79.2	88.6	80.8	90.2	89.6	82.5	79.6	87.5	75.0	71.5	60.2
en	cy	62.5	61.8	65.3	61.3	62.4	61.6	67.5	63.6	61.5	61.2	64.9	60.4	61.4	60.4	61.7	58.8
cy	cy	92.6	91.9	87.1	77.0	89.5	75.0	76.6	73.5	90.9	90.4	85.1	76.1	83.1	67.8	72.1	67.3
fr	br	74.3	71.8	73.5	73.3	74.2	72.8	66.6	63.1	66.3	64.2	66.6	64.7	66.3	64.5	59.3	54.0
br	br	92.8	88.4	88.2	84.5	88.8	79.9	71.1	66.1	89.1	85.8	87.1	81.3	82.8	74.1	59.3	55.2
fr	oc	83.9	83.7	89.1	83.5	83.7	83.4	76.6	71.9	72.5	72.3	78.8	71.8	72.3	71.9	66.5	59.1
oc	oc	95.3	94.9	95.8	92.3	87.8	83.9	79.1	75.2	93.8	93.0	94.6	91.5	92.6	89.8	67.0	61.3
id	ms	68.7	67.7	76.7	64.8	68.5	64.8	79.9	68.4	69.7	69.5	79.9	66.2	69.5	65.8	78.3	58.4
ms	ms	92.4	92.6	83.5	81.7	81.8	70.5	82.7	71.8	92.4	91.9	89.1	71.7	79.7	59.5	80.3	62.4
it	scn	63.7	63.3	80.2	58.4	49.5	45.4	71.0	66.2	60.8	60.7	74.0	55.3	50.4	45.5	60.7	46.8
scn	scn	92.9	91.1	88.1	79.8	74.4	64.9	64.3	57.1	90.5	88.2	82.8	79.7	72.4	62.5	40.0	39.0
es	an	**88.0**	87.9	84.8	85.4	80.7	77.5	86.1	76.3	**86.1**	86.2	86.4	83.3	75.3	72.9	77.0	55.0
an	an	95.8	95.8	88.4	85.6	90.9	**79.1**	83.4	76.8	94.2	93.6	92.5	79.8	80.4	**66.1**	72.6	59.4
es	ast	**90.4**	90.2	86.0	85.1	89.6	84.6	84.1	77.5	**84.3**	84.2	86.0	77.0	84.1	76.3	76.7	59.6
ast	ast	93.6	92.8	90.1	82.7	93.3	**79.7**	85.2	78.4	89.6	89.2	90.1	77.7	90.0	**76.4**	80.3	68.0
es	ca	**85.1**	84.3	87.2	84.0	85.1	84.0	79.3	75.9	**82.6**	82.8	83.9	80.8	82.3	79.8	72.8	66.2
ca	ca	92.3	91.5	91.6	87.3	91.6	**86.5**	85.9	83.0	89.4	89.6	88.0	83.3	88.6	**82.1**	83.9	78.0

Results – Vocabulary Overlap

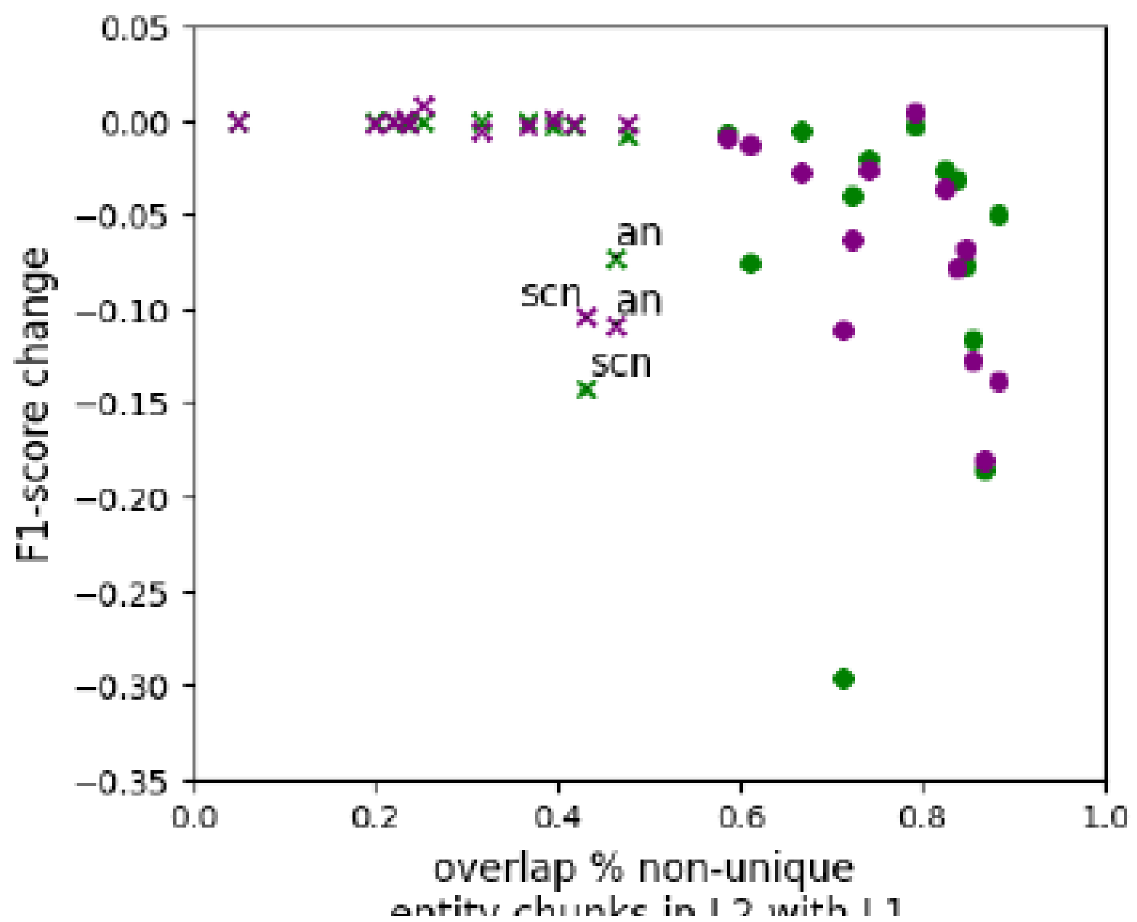
- P3: clear **correlation** between the **vocabulary** overlap percentage and the **performance** degradation for replacing **named entities**.
- This suggests that multilingual models' NER performance for LRLs depends to some extent on **word memorization.**
- Model may not be recognizing a named entity in L2, but its ability in L1 is riding for L2 due to vocabulary overlap (or memorization).



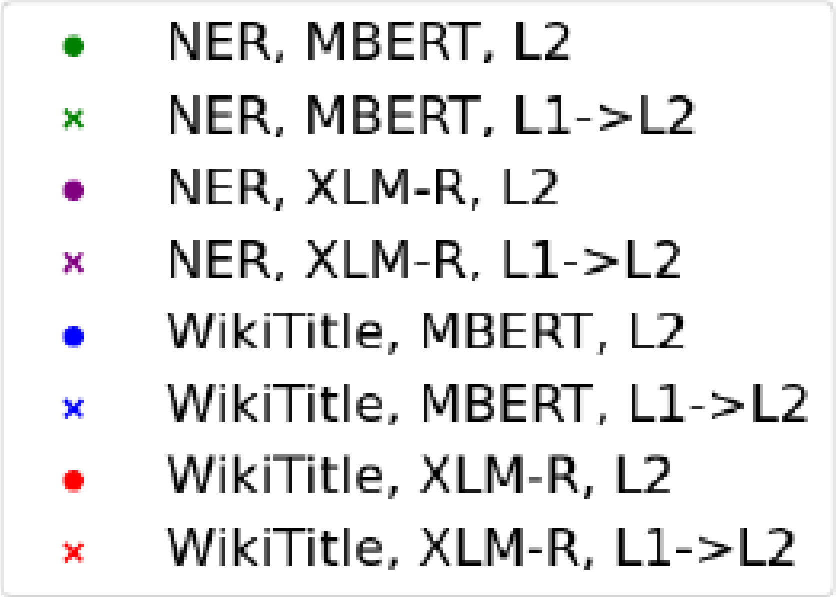
NER F1 changes in P3 perturbation

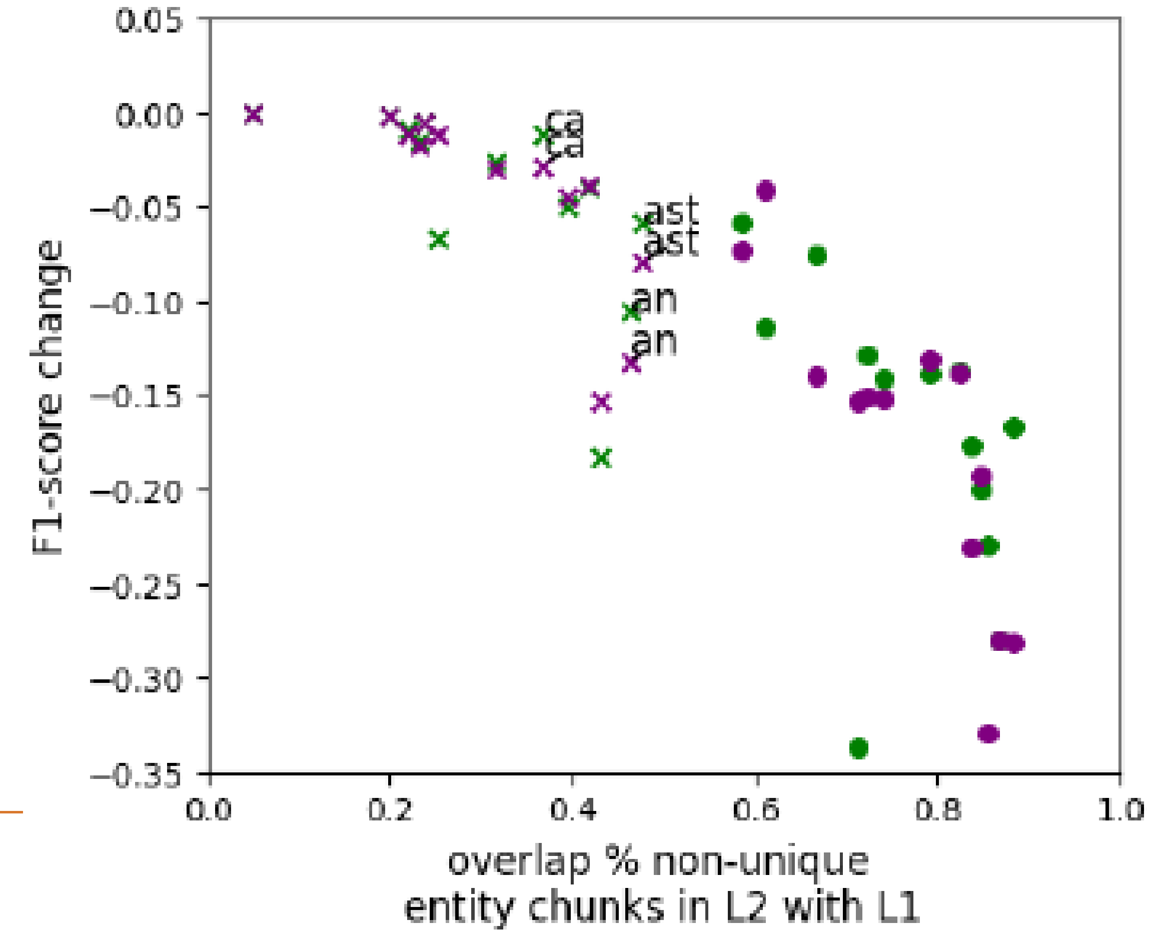
Results – Vocabulary Overlap

- P4: interestingly, the cross-lingual **transfer** models appear to be more robust to **certain perturbations**, such as perturbing **context words.**
- P5: NER performance suffers a significant drop.



NER F1 changes in P4 perturbation

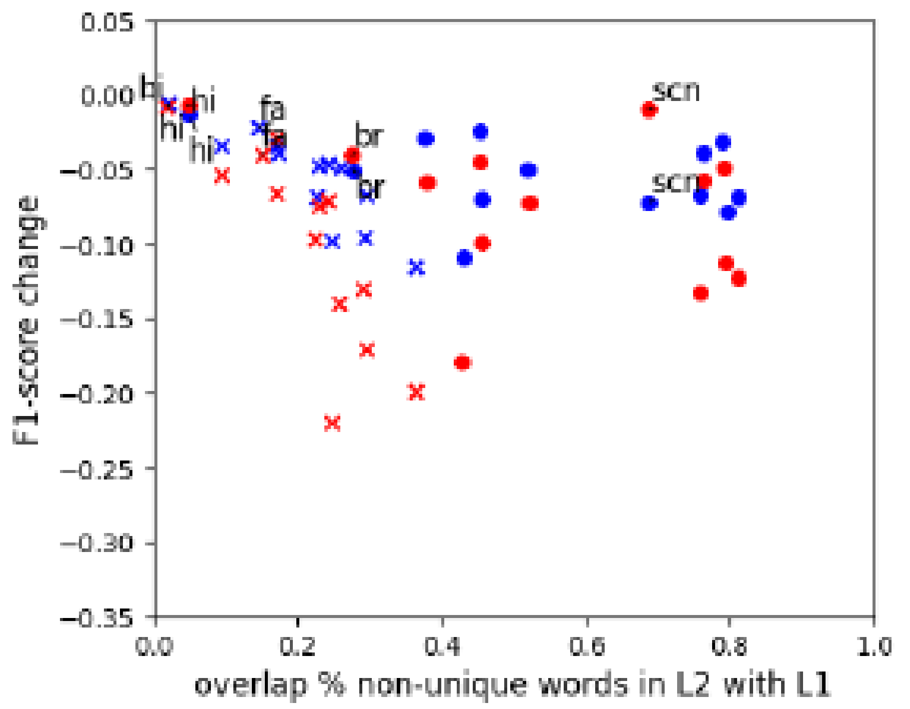




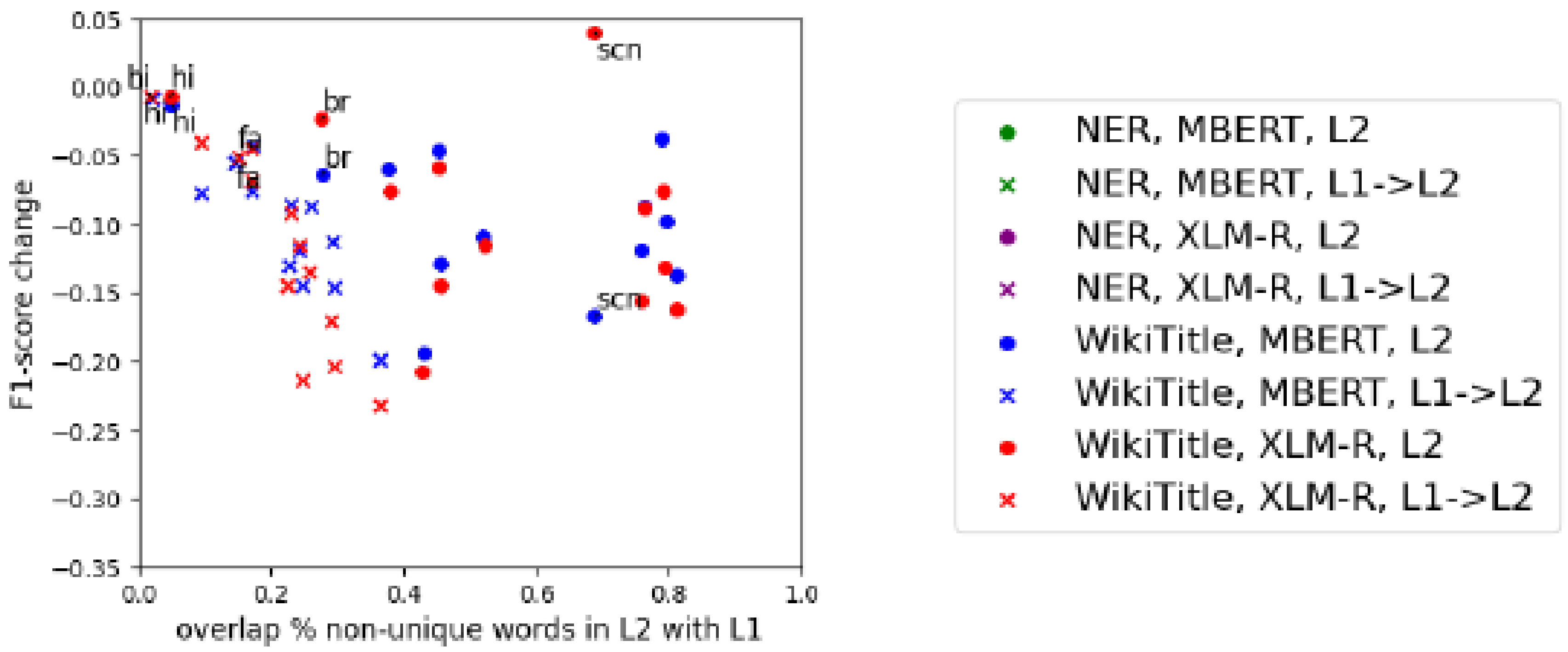
NER F1 changes in combination of P3 and P4 perturbations

Results – Vocabulary Overlap

- For **low overlap** like French/Breton, we would expect **performance** under perturbation to remain **relatively unchanged** (compare Hindi), but Breton still suffers a performance loss of ~4–5 points.
- This suggests that **title section** task **relies heavily on word memorization** of the training data, as a similar drop in performance is observed when words are substituted randomly.
  - The semantic **similarities of the substitute words under P4 seem to not matter**.



Section Title accuracy changes in cosine P4 perturbation



Section Title accuracy changes in random P4 perturbation

Results – Vocabulary Overlap

- **None** of the **perturbations** for **Arabic/Hindi** have much **effect** in the cross-lingual setting.
  - This is expected because Arabic/Hindi languages use **different native scripts**, so there is a **low default token overlap** and consequently very **minor changes**.
- In the case of **Arabic/*Persian***, which do share the same script, **the same is true**.
  - Because words **appearance** are so different in them.
- But **Arabic/Persian** cross-lingual **transfer** on NER is substantially **lower** than on **Arabic/ Hindi**.
  - *mark* (“brand”) vs. *mârd* (“evil”), or *sardard* (“headache”) vs. *sard* (“story”), while they **are not semantically similar**.

		MBERT								XLM-R							
		NER						WikiTitle		NER						WikiTitle	
Train	**Test**	**Base**	**P1**	**P2**	**P3**	**P4**	**P5**	**Base**	**P4**	**Base**	**P1**	**P2**	**P3**	**P4**	**P5**	**Base**	**P4**
ar	hi	67.2	64.2	68.9	67.2	67.2	67.2	63.6	63.0	67.3	67.4	70.7	67.3	67.3	67.3	75.8	75.0
hi	hi	86.7	86.5	87.2	71.3	79.0	66.7	73.8	72.5	87.5	87.2	88.1	76.6	80.7	68.3	77.8	77.1
ar	fa	45.0	43.0	44.7	45.0	45.0	44.9	79.3	77.1	43.6	42.8	40.1	43.6	43.5	43.4	78.0	73.9
fa	fa	90.3	88.0	89.1	86.5	60.8	56.7	81.6	79.1	89.4	88.2	87.4	85.5	78.2	74.1	81.0	76.5

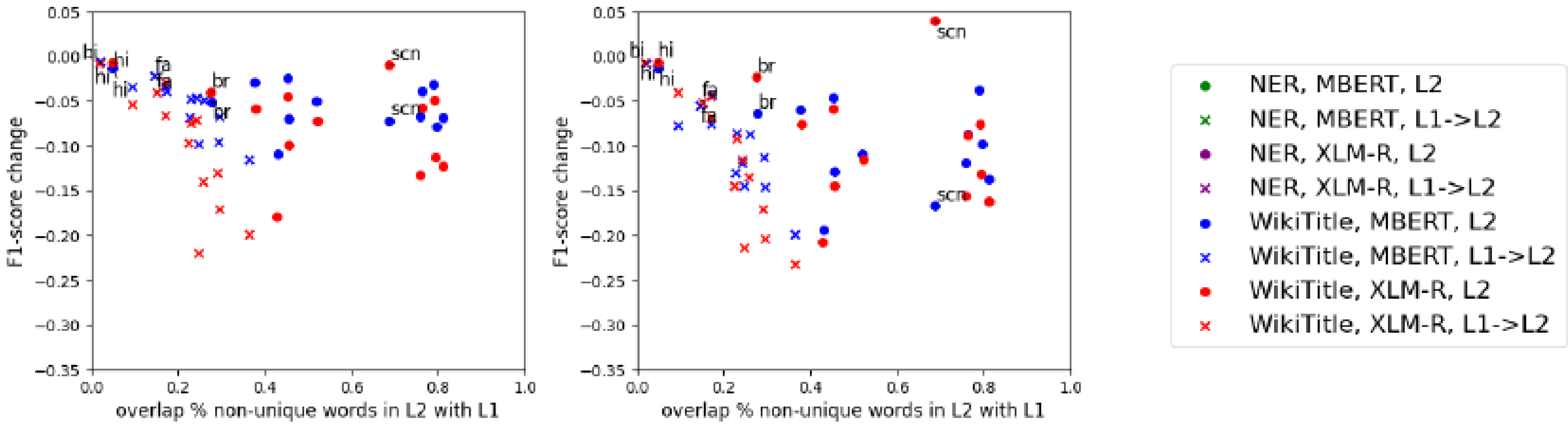
Results – MBERT and XLMR

- **XLM-R** is more robust to random replacement of **B-PER tags (P1)**.
- On **average**, **MBERT** appears more robust to the perturbations we applied
- Note that even the simple perturbation of **changing context words** in the **title selection** task **degraded** performance **universally**.

	MBERT				XLM-R			
	NER: L2	avg. Δ F_1	NER: L1→L2	avg. Δ F_1	NER: L2	avg. Δ F_1	NER: L1→L2	avg. Δ F_1
P1	$p = 0.0118$	-1.00	$p = 0.0046$	-0.92	$p = 0.0116$	-0.80	$p = 0.0655$	-0.34
P2	$p = 0.0033$	-3.65	$p = 0.2096$	2.15	$p = 0.0165$	-2.33	$p = 0.0246$	3.42
P3	$p < 0.0001$	-9.66	$p = 0.0013$	-2.72	$p < 0.0001$	-10.46	$p = 0.0013$	-2.52
P4	$p = 0.0105$	-7.07	$p = 0.1500$	-1.80	$p = 0.0004$	-6.73	$p = 0.1499$	-1.69
P5	$p < 0.0001$	-16.71	$p = 0.0106$	-4.36	$p < 0.0001$	-17.62	$p = 0.0090$	-4.26
	Titles: L2	avg. Δ acc.	Titles: L1→L2	avg. Δ acc.	Titles: L2	avg. Δ acc.	Titles: L1→L2	avg. Δ acc.
P4	$p < 0.0001$	-5.38	$p < 0.0001$	-5.54	$p = 0.0002$	-7.57	$p = 0.0003$	-9.52

Results – MBERT and XLMR

- For **Section title task, native** Sicilian performance in **MBERT** substantially exceeds **XLM-R**, but also suffers more **under perturbation**.
  - **Sicilian training data is included** in the **pretraining** data for **MBERT** but not for **XLM-R**.
- The much lower performance of the native Sicilian XLM-R model on title selection compared to NER suggests that NER fine-tuning can leverage other representations (e.g., common named entities between Italian and Sicilian).



Conclusion

- The first time such an experimental set has been performed with an explicit focus on **LRLs and cross-lingual transfer from HRLs**.
- We conducted evaluations on **21 languages**, encompassing both high and low-resource languages, employing two widely recognized multilingual models, **MBERT and XLM-R**.
- Results exhibit **variations across different languages**, influenced by their **linguistic structures and similarities**.
- Our core findings can be summarized as follows:
  - There is a pronounced effect of **vocabulary overlap on NER performance**.
  - Although models utilizing cross-lingual **transfer** typically exhibit **lower** numerical performance than models trained in a **native** LRL setting, they are **often somewhat more robust** to certain types of perturbations of the input.
  - **Title selection** in LRLs appears **heavily rely on word memorization**.

Discussion

- This research has been conducted on encoder models.
  - Encoder models are **older** and **smaller**, typically demand **fewer computational resources**, allowing us to perform more experiments.
  - Unlike SOTA **decoder** models like **GPT-4**, most encoder model **weights and processing pipelines are freely available** on platforms like HuggingFace, meaning that we can directly access the embedding spaces to inform our perturbation techniques.
  - Most open-weight generative models (e.g., LLaMA 2) are **not multilingual**.
  - However, since our techniques are general, they could be applied to open-source multilingual generative models like XGLM. We do note that multilingual generative models still **do not necessarily contain all the required languages**.

Future Work

- These proposed test sets have the potential for further exploration, particularly in challenging **tokenizers** directly.
- For example, the Persian examples suggest that, although **BPE tokenization** methods should help LRL performance by not biasing toward HRL, **similarity between sub-word tokens overvalued** when optimizing the embedding space.
- This motivates an **equitable consideration of lower-resource languages** in building NLP models.

COLORADO STATE UNIVERSITY

Thank You!

ShadiM@ColoState.edu
Nikhil.Krishnaswamy@ColoState.edu