GLUECoS : An Evaluation Benchmark for Code-Switched NLP

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, Monojit Choudhury

https://aka.ms/gluecos
What is Code-Mixing?

*Code-mixing or Code-Switching* is the mixing of two or more languages in a conversation or even an utterance.

Is generally associated with informal conversations

Is predominantly a spoken language phenomenon

VIJAYLAXMI
What happened?

RANI
*Meri* family call *kar rahi hai*  
(My family is calling)

VIJAYLAXMI
then talk..

RANI
*kaisey* ?  
(How?)
Why Code-Mixing?

- Monolingual as well as multilingual NLP systems break-down in the presence of code-mixing
- In public pages from Facebook
- Over 50M tweets analysed by (Rijhwani et al., 2017) in which 3.5% tweets are code-switched
- ALL sufficiently long threads were multilingual
- 17.2% of the comments/posts have code-mixing (Bali et al., 2014)

What is the world’s prettiest location?

World *ki sabse sundar* location *kya hai*?
Why do we need a benchmark?

Shared Tasks focus on one aspect alone

Commonly included tasks
- Language Identification
- POS Tagging
- Named Entity Recognition
- Sentiment Analysis

Single test-bed reflects if a model truly understands code-mixed languages
How do we choose our datasets?

- Complexity of tasks
- Typological variations: English-Spanish; English-Hindi
- Script variance
- Multiple datasets for each task
- Publicly available datasets
How do we choose our datasets?

- **Complexity of Tasks**
- **Typological Variations**
  - English-Spanish;
  - English-Hindi
- **Script Variance**
- **Multiple Datasets for Each Task**
- **Publicly Available Datasets**
How do we choose our datasets?

- Complexity of tasks
- Typological variations: English-Spanish; English-Hindi
- Script variance
- Multiple datasets for each task
- Publicly available datasets
How do we choose our datasets?

- Complexity of tasks
- Typological variations: English-Spanish; English-Hindi
- Script variance
- Multiple datasets for each task
- Publicly available datasets
How do we choose our datasets?

- Complexity of tasks
- Typological variations: English-Spanish; English-Hindi
- Script variance
- Multiple datasets for each task
- Publicly available datasets
What all tasks do we include?

**Word Level Tasks**
- Language Identification
- Part-of-speech Tagging
- Named Entity Recognition

**Sentence Level Tasks**
- Sentiment Analysis
- Question Answering

**Sentence Pair Tasks**
- Natural Language Inference

**Corpus Size**

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English-Hindi</strong></td>
<td></td>
</tr>
<tr>
<td>Lang Identification</td>
<td>3537 / 13.51</td>
</tr>
<tr>
<td>POS Tagging A</td>
<td>1814 / 14.17</td>
</tr>
<tr>
<td>POS Tagging B</td>
<td>2631 / 15.56</td>
</tr>
<tr>
<td>NER</td>
<td>3084 / 22.21</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>12601</td>
</tr>
<tr>
<td>QA</td>
<td>313</td>
</tr>
<tr>
<td>NLI</td>
<td>1300</td>
</tr>
<tr>
<td><strong>English-Spanish</strong></td>
<td></td>
</tr>
<tr>
<td>Lang Identification</td>
<td>14413 / 12.09</td>
</tr>
<tr>
<td>POS Tagging</td>
<td>2758 / 9.49</td>
</tr>
<tr>
<td>NER</td>
<td>34208 / 11.74</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>2103</td>
</tr>
</tbody>
</table>
Code-Mixing Metrics

<table>
<thead>
<tr>
<th>Corpus</th>
<th>CMI</th>
<th>SP Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-Hindi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lang Identification</td>
<td>78.26</td>
<td>4.47</td>
</tr>
<tr>
<td>POS Tagging A</td>
<td>136</td>
<td>4.98</td>
</tr>
<tr>
<td>POS Tagging B</td>
<td>68</td>
<td>5.5</td>
</tr>
<tr>
<td>NER</td>
<td>133</td>
<td>11.39</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>72.8</td>
<td>5.07</td>
</tr>
<tr>
<td>QA</td>
<td>142.28</td>
<td>3.96</td>
</tr>
<tr>
<td>NLI</td>
<td>149.95</td>
<td>66.74</td>
</tr>
<tr>
<td>English-Spanish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lang Identification</td>
<td>33.46</td>
<td>2.86</td>
</tr>
<tr>
<td>POS Tagging</td>
<td>123.06</td>
<td>1.67</td>
</tr>
<tr>
<td>NER</td>
<td>94.52</td>
<td>3.17</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>110.56</td>
<td>4.13</td>
</tr>
</tbody>
</table>
Welcome to our presentation on GLUECoS!

Language Identification

LID is the task of obtaining word-level language labels for code-switched sentences.

- FIRE 2013 dataset for the transliterated search subtask (Roy et al., 2013)
- Shared task @ EMNLP 2014 (Solorio et al., 2014)
I believe in listening to everyone but doing your own 😎

POS tagging includes labelling at the word level, grammatical part of speech tags such as noun, verb, adjective, pronoun etc.

- Universal Dependency parsing dataset (Bhat et al., 2018)
- ICON 2016 Tool Contest on POS Tagging for Code-Mixed Indian Social Media Text (Jamatia et al., 2016)
- Bangor Miami corpus (AlGhamdi et al., 2016)
Named Entity Recognition

NER involves recognizing named entities such as person, location, organization etc. in a segment of text

Kohli should be given two World Cups

Twitter NER corpus (Singh et al., 2018)

CALCS shared task @ ACL 2018 (Aguilar et al., 2019)
Sentence classification task wherein each sentence is labeled to be expressing a positive, negative or neutral sentiment.

Hay siete continentes in the world

There are seven continents in the world

Coronavirus ki vajah se log bohot pareshan hai :'( People are very troubled because of the coronavirus

Sentiment Analysis for Indian Languages (SAIL – Patra et al., 2018)

Sentiment annotated Twitter dataset (Vilares et al., 2016)
Question Answering

internet explorer ko kis browser se replace kiya main?

Microsoft Edge

QA is the task of answering a question based on the given context or world knowledge

EN-HI

- Code-mixed QA challenge, CALCS @ ACL 2018
Natural Language Inference

**PREMISE**

BABLU: You teach well!  
*Pichhli job ka kya huya?*  
(What happened to your previous job?)

DEVI: Work *achha nahin tha*  
(The work wasn’t good)

BABLU: I know *aap Yadav ke udhar thi* ...  
(I know you were at Yadav’s)

**HYPOTHESIS**

DEVI *pichhli job mein Yadav ke sa-ath kaam kar rahì thi.*  
(Devi worked with Yadav in her previous job)

---

NLI is the task of inferring a positive (entailed) or negative (contradicted) relationship between a premise and hypothesis.

- Conversational NLI (Khanuja et al., 2020)
What models have been used for CM Tasks?

- **Most common**
  - Adapting cross-lingual models for code-mixing

- **Less common**
  - Using synthetic code-mixed data to train embeddings

- **Our method**
  - Combines the best of both these methods
Cross-Lingual models for Code-Mixing

- Treat code-mixing as a special case of cross-lingual NLP
  - Is this the way?

- Cross-Lingual Word Embeddings (Ruder et al., 2017)
  - Trained to predict an L2 word given a context of L1 words
  - BiCVM, BiSkip, MUSE

- Not to forget – Multilingual BERT
  - Originally evaluated on cross-lingual tasks like XNLI
  - Can this be adapted for Code-Mixing?
Embeddings for Code-Mixing

Train word2vec embeddings on synthetic code-mixed data

Pratapa et al., 2018b showed that these outperform cross-lingual embeddings on 2 code-mixed tasks
### Initial Results

#### English-Hindi

<table>
<thead>
<tr>
<th>Task</th>
<th>Synth. CM w2v</th>
<th>mBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lang Identification</td>
<td>93.64</td>
<td>95.87</td>
</tr>
<tr>
<td>POS Tagging A</td>
<td>77.84</td>
<td>87.16</td>
</tr>
<tr>
<td>POS Tagging B</td>
<td>61.03</td>
<td>63.42</td>
</tr>
<tr>
<td>NER</td>
<td>72.37</td>
<td>74.96</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>50.01</td>
<td>58.24</td>
</tr>
<tr>
<td>QA</td>
<td>62.78</td>
<td>71.96</td>
</tr>
<tr>
<td>NLI</td>
<td>-</td>
<td>61.09</td>
</tr>
</tbody>
</table>

#### English-Spanish

<table>
<thead>
<tr>
<th>Task</th>
<th>Synth. CM w2v</th>
<th>mBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lang Identification</td>
<td>92.42</td>
<td>95.97</td>
</tr>
<tr>
<td>POS Tagging</td>
<td>89.37</td>
<td>93.33</td>
</tr>
<tr>
<td>NER</td>
<td>53.57</td>
<td>59.69</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>62.89</td>
<td>66.03</td>
</tr>
</tbody>
</table>

mBERT outperforms all word embedding based methods

- Data that mBERT was pretrained on is much larger and spans 104 languages
- mBERT was exposed to no code-mixing during training
Our New Model

Modified mBERT

- Take mBERT and perform MLM finetuning on code-mixed data
- One model per language pair

2 stage curriculum

- First on large corpus of synthetic code-mixed data (method from Pratapa et al., 2018a)
- Next on a smaller corpus of non-synthetic code-mixed data
Results and Analysis

<table>
<thead>
<tr>
<th></th>
<th>English-Hindi</th>
<th>English-Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Task</strong></td>
<td><strong>Stock mBERT</strong></td>
<td><strong>Modified mBERT</strong></td>
</tr>
<tr>
<td>Lang Identification</td>
<td>95.87</td>
<td>96.60</td>
</tr>
<tr>
<td>POS Tagging A</td>
<td>87.16</td>
<td>88.06</td>
</tr>
<tr>
<td>POS Tagging B</td>
<td><strong>63.42</strong></td>
<td>63.31</td>
</tr>
<tr>
<td>NER</td>
<td>74.96</td>
<td><strong>78.21</strong></td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>58.24</td>
<td>59.35</td>
</tr>
<tr>
<td>QA</td>
<td><strong>71.96</strong></td>
<td>68.01</td>
</tr>
<tr>
<td>NLI</td>
<td>61.09</td>
<td><strong>63.10</strong></td>
</tr>
</tbody>
</table>

- Modified mBERT outperforms the stock version of mBERT on most tasks
- Varying performance across language pairs
  - Gains in En-Es are larger

- Within a language pair
  - Some tasks are much easier than the other
  - Different datasets for same task show varying performance numbers
Takeaways

- Code-mixing cannot be solved by just applying cross-lingual techniques
- Requires a method designed with code-mixing in mind
- Our technique produces such a model that makes a step in this direction
- We are making the GLUECoS testbed available to everyone - A platform to evaluate your model on multiple code-mixing tasks
To find out more

- Check out the benchmark at https://aka.ms/gluecos
- Submit your model(s) for evaluation on the leaderboard
- Contribute any code-mixed datasets you have
We would like to thank:

- Rishiraj Saha Roy, Prasenjit Majumder and Komal Agarwal – Language Identification (En-Hi)
- Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg and Alison Chang – Language Identification (En-Es)
- Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Manish Shrivastava and Dipti Misra Sharma – POS Tagging (En-Hi)
- Anupam Jamatia, Björn Gambäck and Amitava Das – POS Tagging (En-Hi)
- Fahad AlGhamdi, Giovanni Molina, Mona Diab, Thamar Solorio, Abdelati Hawwari, Victor Soto and Julia Hirschberg – POS Tagging (En-Es)
- Vinay Singh, Deepanshu Vijay, Syed Sarfaraz Akhtar and Manish Shrivastava – NER (En-Hi)
- Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg and Thamar Solorio – NER (En-Es)
- Braja Gopal Patra, Dipankar Das and Amitava Das – Sentiment Analysis (En-Hi)
- David Vilares, Miguel A Alonso and Carlos Gómez-Rodríguez - Sentiment Analysis (En-Hi)
- Khyathi Chandu, Ekaterina Loginova, Vishal Gupta, Josef van Genabith, Gunter Neumann, Manoj Chinnakotla, Eric Nyberg and Alan W Black – Question Answering (En-Hi)