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

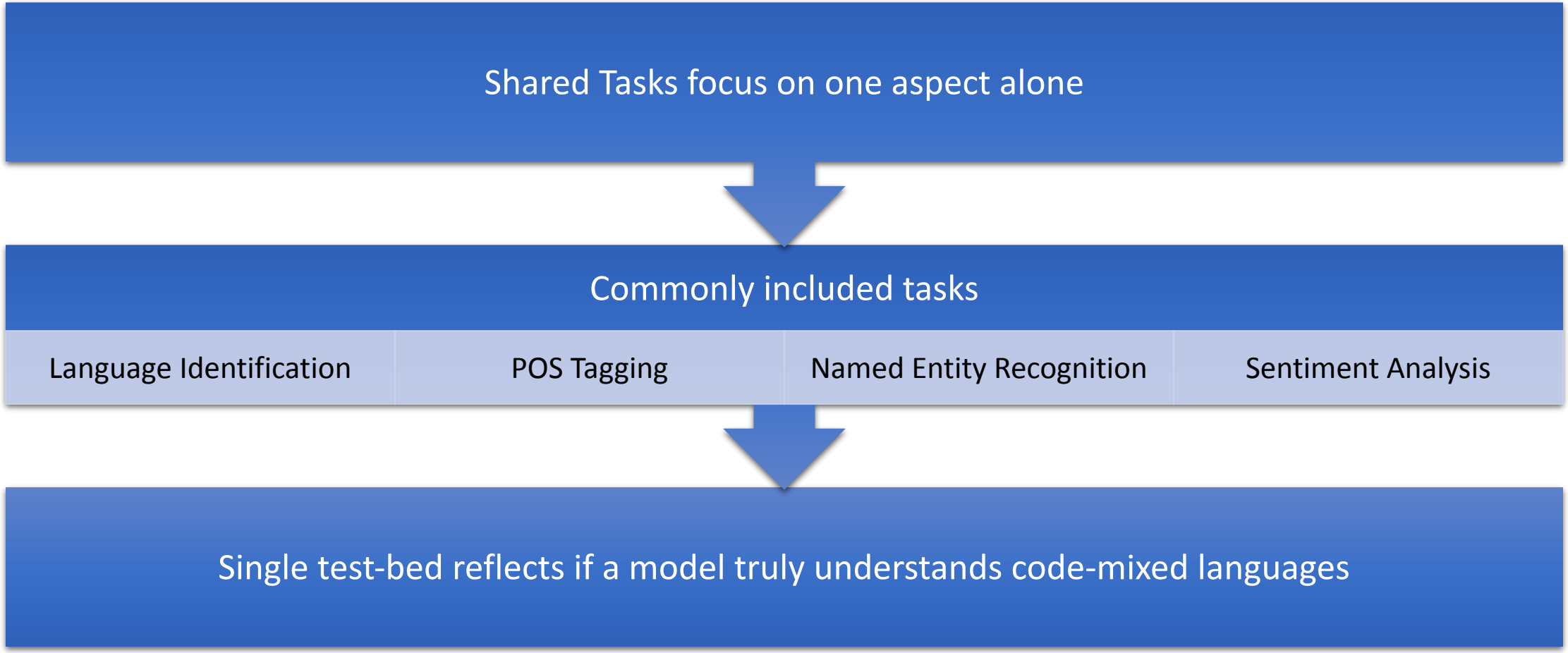
- ALL sufficiently long threads were multilingual
- 17.2% of the comments/posts have code-mixing (Bali et al., 2014)

Over 50M tweets analysed by (Rijhwani et al., 2017) in which 3.5% tweets are code-switched

What is the world's prettiest location?

World *ki sabse sundar* location *kya hai*?

Why do we need a benchmark?



How do we choose our datasets?

COMPLEXITY OF TASKS

TYPOLOGICAL VARIATIONS ENGLISH-SPANISH; ENGLISH-HINDI

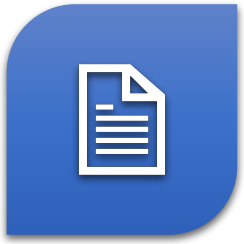
SCRIPT VARIANCE

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

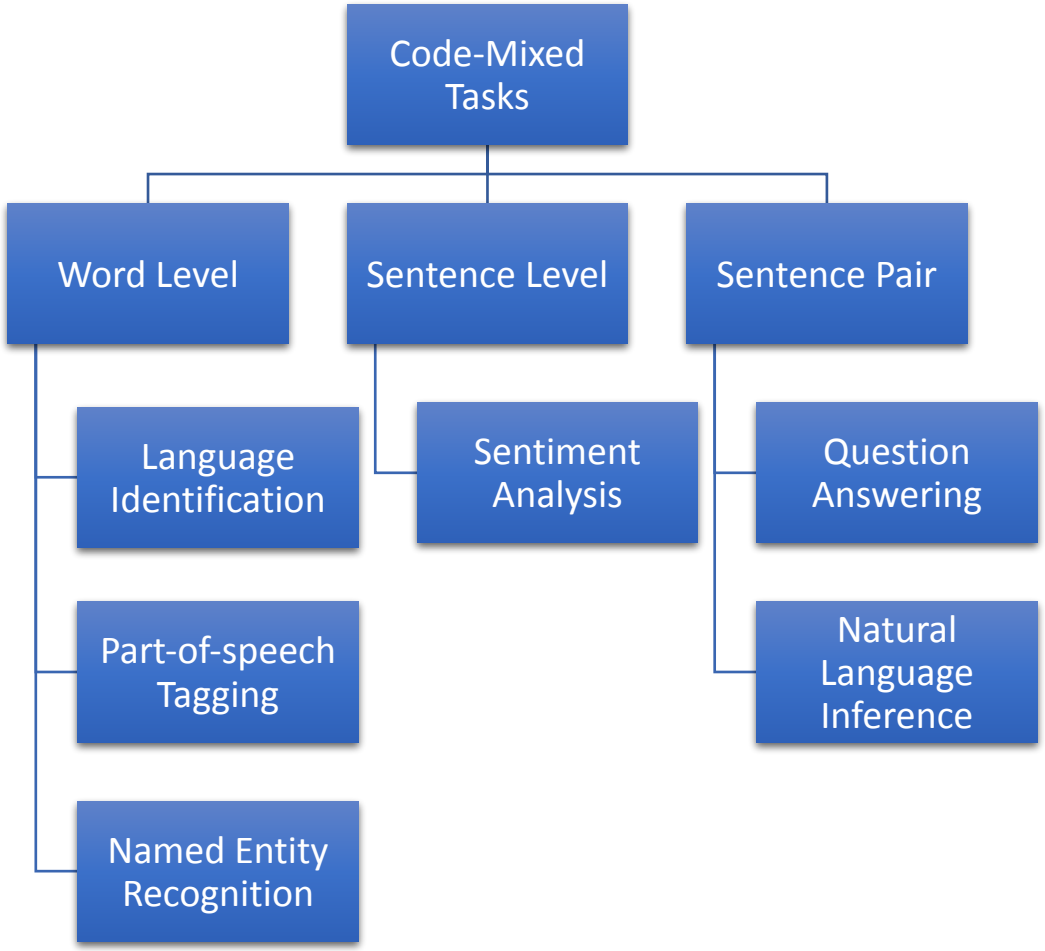
TYPOLOGICAL VARIATIONS
ENGLISH-SPANISH;
ENGLISH-HINDI

SCRIPT VARIANCE

MULTIPLE DATASETS FOR
EACH TASK

PUBLICLY AVAILABLE
DATASETS

What all tasks do we include?



Corpus	Size
English-Hindi	
Lang Identification	3537 / 13.51
POS Tagging A	1814 / 14.17
POS Tagging B	2631 / 15.56
NER	3084 / 22.21
Sentiment Analysis	12601
QA	313
NLI	1300
English-Spanish	
Lang Identification	14413 / 12.09
POS Tagging	2758 /9.49
NER	34208 / 11.74
Sentiment Analysis	2103

Sentences/Avg. tokens per sentence

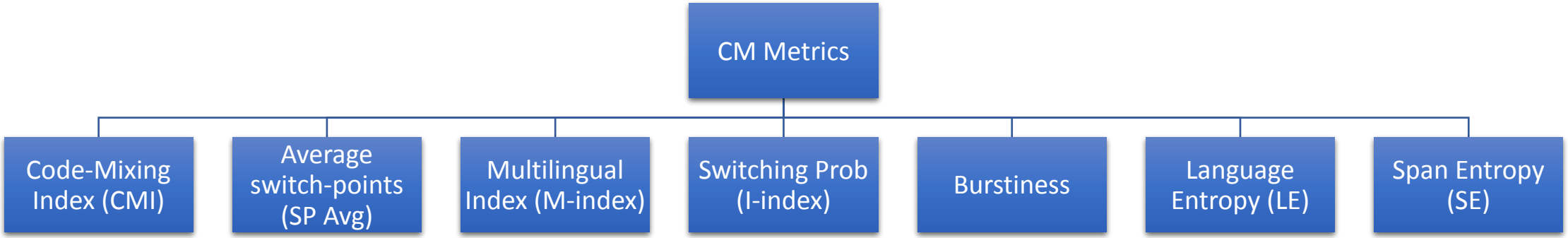
Sentences

Sentence pairs

Sentences/Avg. tokens per sentence

Sentences

Code-Mixing Metrics



Corpus	CMI	SP Avg
English-Hindi		
Lang Identification	78.26	4.47
POS Tagging A	136	4.98
POS Tagging B	68	5.5
NER	133	11.39
Sentiment Analysis	72.8	5.07
QA	142.28	3.96
NLI	149.95	66.74

Corpus	CMI	SP Avg
English-Spanish		
Lang Identification	33.46	2.86
POS Tagging	123.06	1.67
NER	94.52	3.17
Sentiment Analysis	110.56	4.13

Language Identification

Welcome to our presentation on GLUECoS!

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

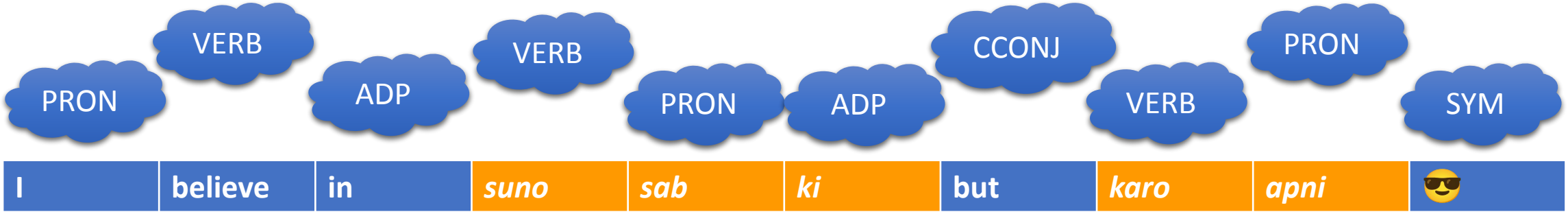
EN-HI

- FIRE 2013 dataset for the transliterated search subtask (Roy et al., 2013)

EN-ES

- Shared task @ EMNLP 2014 (Solorio et al., 2014)

Part-of-speech Tagging



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.

EN-HI

- 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)

EN-ES

- 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

EN-HI

- Twitter NER corpus (Singh et al., 2018)

EN-ES

- CALCS shared task @ ACL 2018 (Aguilar et al., 2019)

Sentiment Analysis

neutral

***Hay siete continentes* in the world**

There are seven continents in the world

negative

Coronavirus *ki vajah se log bohot pareshan hai* :'(

People are very troubled because of the coronavirus

Sentence classification task wherein each sentence is labeled to be expressing a positive, negative or neutral sentiment

EN-HI

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

EN-ES

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

Question Answering

internet explorer *ko*
kis browser *se* replace *kiya* m
icrosoft *ne*?

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

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

EN-HI

- Conversational NLI (Khanuja et al., 2020)

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 rahi thi*.
(Devi worked with Yadav in her previous job)

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		
Task	**Synth. CM w2v**	**mBERT**
Lang Identification	93.64	**95.87**
POS Tagging A	77.84	**87.16**
POS Tagging B	61.03	**63.42**
NER	72.37	**74.96**
Sentiment Analysis	50.01	**58.24**
QA	62.78	**71.96**
NLI	-	**61.09**

English-Spanish		
Task	**Synth. CM w2v**	**mBERT**
Lang Identification	92.42	**95.97**
POS Tagging	89.37	**93.33**
NER	53.57	**59.69**
Sentiment Analysis	62.89	**66.03**

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

English-Hindi		
Task	**Stock mBERT**	**Modified mBERT**
Lang Identification	95.87	**96.60**
POS Tagging A	87.16	**88.06**
POS Tagging B	**63.42**	63.31
NER	74.96	**78.21**
Sentiment Analysis	58.24	**59.35**
QA	**71.96**	68.01
NLI	61.09	**63.10**

English-Spanish		
Task	**Stock mBERT**	**Modified mBERT**
Lang Identification	95.97	**96.24**
POS Tagging	93.33	**93.62**
NER	59.69	**61.77**
Sentiment Analysis	66.03	**69.31**

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

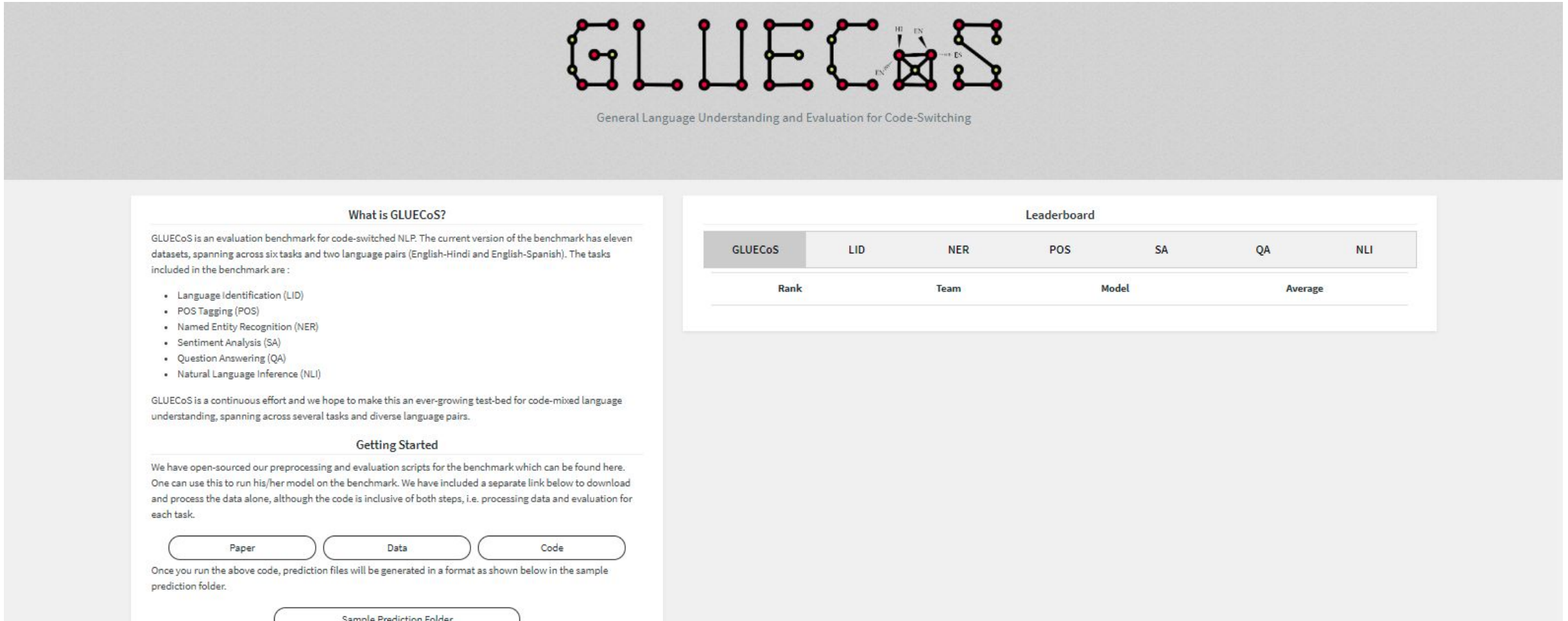
- Within a language pair
  - Some tasks are much easier than the other
  - Different datasets for same task show varying performance numbers

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, Bj¨orn Gamb¨ack 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 G´omez-Rodr´ıguez - Sentiment Analysis (En-Hi)
- Khyathi Chandu, Ekaterina Loginova, Vishal Gupta,Josef van Genabith, G¨unter Neumann, Manoj Chinnakotla, Eric Nyberg and Alan W Black – Question Answering (En-Hi)