A Bora Language Dataset for Machine Translation (MT)

Changhao Liang, Jun Xiao, Qi Yin, Yuang Liu, Zixiang Wei

Outline

- Problem statement
- Dataset
- Approach
- Results
- Conclusion
- Future work
- Q&A

01

Problem Statement

Background

- *Bora* -

- A class-0 language
- About 2,300 native speakers

Dabíii, Aavaráaa, íjcyámútsiúvúj tsiiménémúhaabéhjáa Jetsocríjtó íjcyáábé ihdé múnáaúvú ílluúme:

Backgroud

- *Why we pick -*

❖ Close to extinction

❖ Very similar with Spanish

Bora :¿Kiá diibye éhneva jodíómú avyéjuube íjcyáííbyeke tsáápille tsḧ́mávaábe?

Spanish: ¿Dónde está el Rey de los Judíos, que ha nacido?

Problem statment

Collect data for Bora and Spanish

Data augmentation on the Bora dataset

Train cross-lingual language model

Train Machine Translation models

02

Dataset

Data Collection

Bora:

- Train: 90% Bible
- Valid: 5% Bibile
- Test: 5% Bibile and Universal Declaration of Human Rights

Spanish:

- Train: 90% Bible and many novels
- Valid: 5% Bible and some novels
- Test: Universal Declaration of Human Rights

Bora Datasets

Bora:

- Number of different words: 21597
- Number of different 2-gram pairs: 64371

Problem:

- 10000 words appear for one time

```
ámuha 2155
Muurá 1643
tsá 1540
dibye 1532
dííbyeke 1302
ditye 1276
muurá 1162
ámúhakye 1090
ihdyu 1086
Píívyéébe 895
Jetsóó 878
oke 866
muhdú 808
Píívyéébé 801
tsúúca 689
ímí 687
```

Augmented Bora Datasets

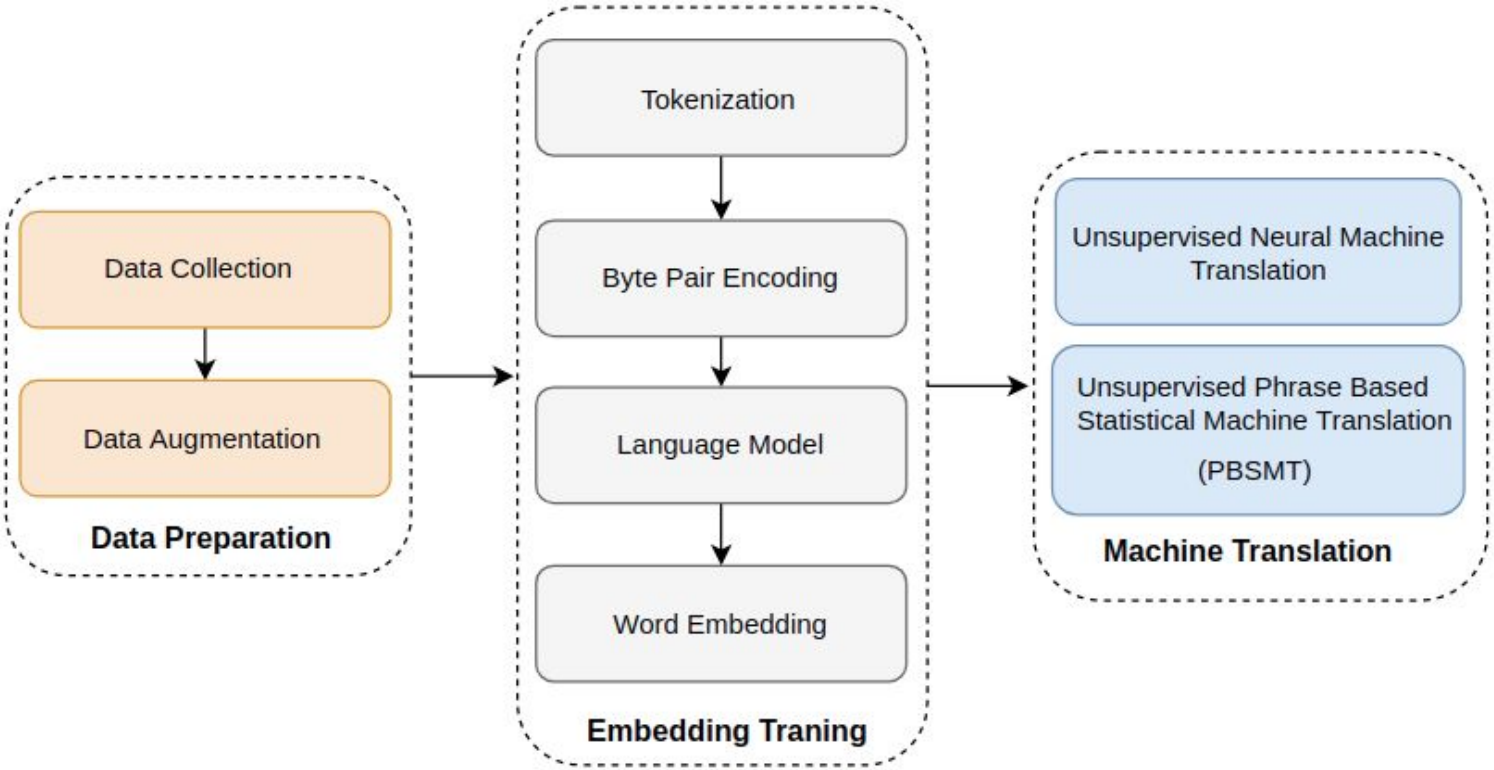
Augmented Bora:

- Number of different words: 21597
- 904 words appear less than 10 times

Data Augmentation	Sentence Size	Vocabulary Size
Before	17k	21.5k
After	185k	21.5k

03

Approach



High-level overview

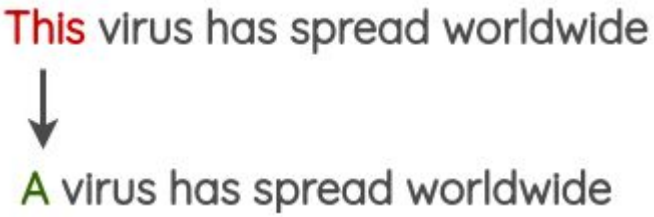
Data Augmentation

- **Word Replacing**

  synonym/antonym

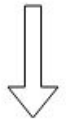
  word embedding

  with TF-IDF



Tokenization

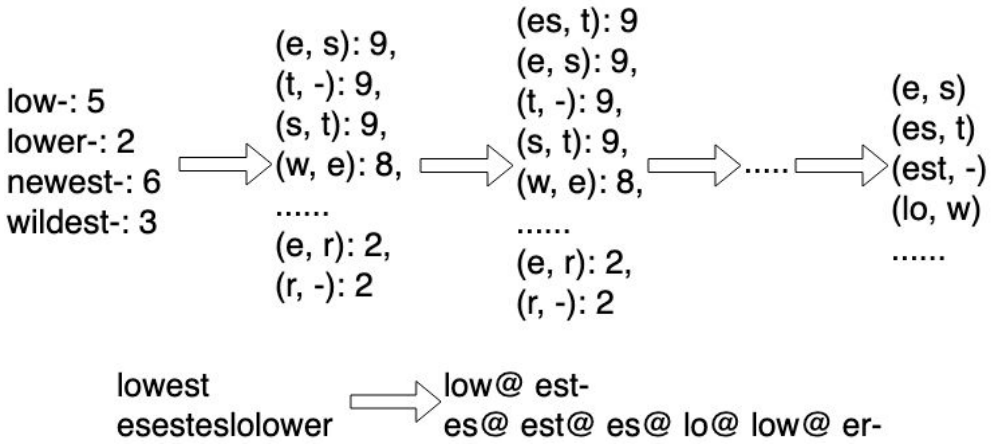
He sits on an armchair.



['He', 'sits', 'on', 'an', 'armchair', '.']

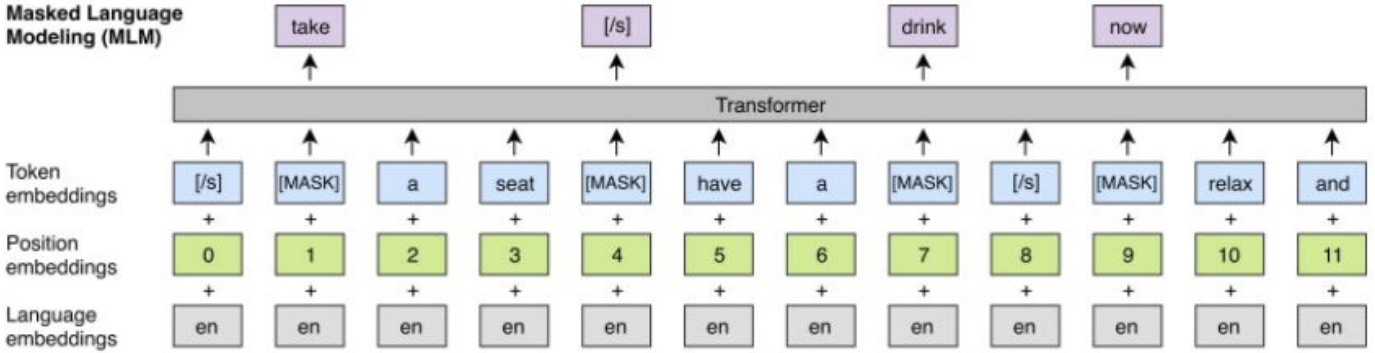
- **Spanish:** ToktokTokenizer in NLTK Python packet.
- **Bora:** a tokenize script written by the team

Byte Pair Encoding



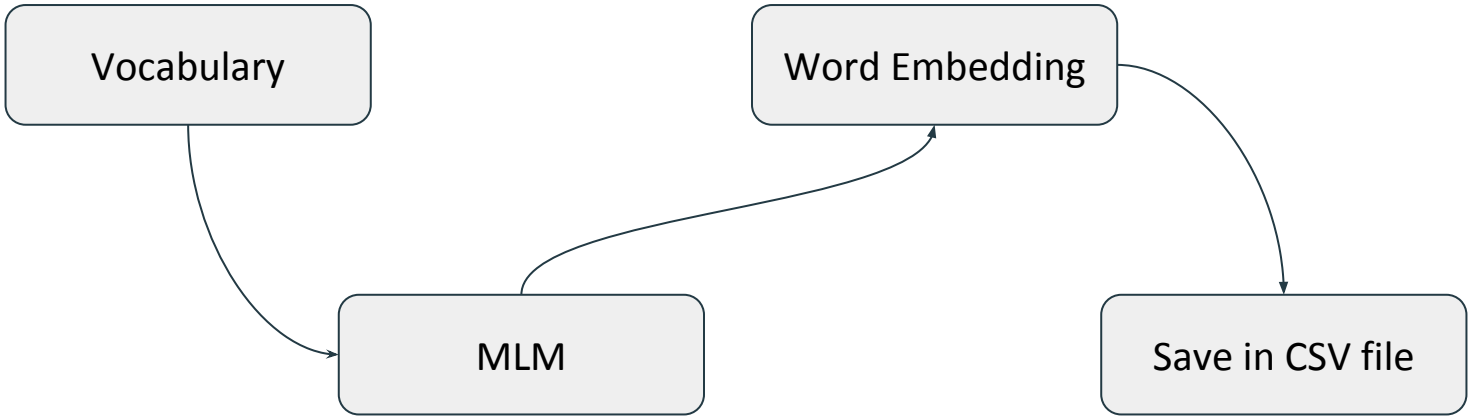
C++ implementation of the BPE: [glample/fastBPE: Fast BPE]

Language Model



Masked language model: [facebookresearch/XLM: PyTorch original implementation of Cross-lingual Language Model Pretraining.](), to make the model suitable for the SCC environment, we have to make some updates to the parameters.

Word Embedding



Statistical Machine Translation (PBSMT)

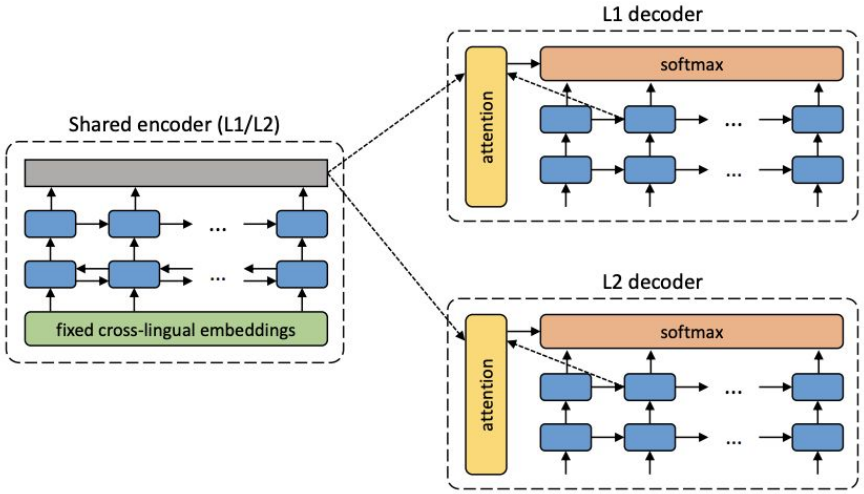
1. Load the trained mono embeddings data
2. learn language models
3. run MUSE to generate cross-lingual embeddings
4. Generate a phrase-table in an unsupervised way

$$p(t_j|s_i) = \frac{e^{\frac{1}{T}\cos(e(t_j), We(s_i))}}{\sum_k e^{\frac{1}{T}\cos(e(t_k), We(s_i))}},$$

5. Train Moses on the phrase-table

Source	Target	$P(s|t)$	$P(t|s)$
heureux	happy	0.931	0.986
	delighted	0.458	0.003
	grateful	0.128	0.003
	thrilled	0.392	0.002
	glad	0.054	0.001
Royaume-Uni	Britain	0.242	0.720
	UK	0.816	0.257
	U.K.	0.697	0.011
	United Kingdom	0.770	0.010
	British	0.000	0.002
Union européenne	European Union	0.869	0.772
	EU	0.335	0.213
	E.U.	0.539	0.006
	member states	0.007	0.006
	27-nation bloc	0.410	0.002

Neural Network Machine Translation



Underamt (unsupervised neural machine translation):

[artetxem/undreamt: Unsupervised Neural Machine Translation]

04

Result

Result for Language Model

Results

```
INFO - 11/03/20 10:55:23 - 3 days, 12:51:06 - __log__:{"epoch": 499, "valid_es_bora_m
lm_ppl": 3168.3437541205867, "valid_es_bora_mlm_acc": 29.972359783484972, "valid_mlm_
ppl": 3168.3437541205867, "valid_mlm_acc": 29.972359783484972, "test_es_bora_mlm_ppl"
: 7841066688.135763, "test_es_bora_mlm_acc": 5.446927374301676, "test_mlm_ppl": 78410
66688.135763, "test_mlm_acc": 5.446927374301676}
INFO - 11/03/20 10:55:23 - 3 days, 12:51:06 - Saving checkpoint to ./dumped/xlm_es_bo
ra/j077iz3ite/checkpoint.pth ...
WARNING - 11/03/20 10:55:23 - 3 days, 12:51:06 - Saving model parameters ...
WARNING - 11/03/20 10:55:23 - 3 days, 12:51:06 - Saving model optimizer ...
```

Results without data augmentation

```
============ End of epoch 499 ============
epoch -> 499.000000
valid_es_bora_aug_mlm_ppl -> 2520.224220
valid_es_bora_aug_mlm_acc -> 38.229231
valid_mlm_ppl -> 2520.224220
valid_mlm_acc -> 38.229231
test_es_bora_aug_mlm_ppl -> 35771.090943
test_es_bora_aug_mlm_acc -> 10.458911
test_mlm_ppl -> 35771.090943
test_mlm_acc -> 10.458911
```

Results with data augmentation

Result for Machine Translation

Phrase Table

$$p(t_j|s_i) = \frac{e^{\frac{1}{T}\cos(e(t_j), We(s_i))}}{\sum_k e^{\frac{1}{T}\cos(e(t_k), We(s_i))}},$$

Source	Target	P(s|t)	P(t|s)
ímichi	pudieron	5.79E-03	8.72E-04
	ser	1.52E-03	3.58E-03
	qué	3.02E-03	2.49E-03
	pasando	4.90E-03	3.34E-03
	tuviesen	7.83E-03	3.39E-03
nehíjcyá	acabado	2.07E-01	2.17E-01
	carnales	4.71E-03	4.86E-02
	Lo	3.53E-03	3.95E-02
	sí	3.49E-03	3.81E-02
	alegrías	3.70E-03	3.19E-02
maájcune	estás	1.50E-02	1.06E-02
	conozcan	9.08E-03	1.04E-02
	ambos	8.08E-03	1.03E-02
	designio	9.07E-03	9.59E-03
	hoy	1.52E-02	9.46E-03
néé	temor	8.91E-02	9.18E-01
	apartaos	1.49E-02	5.16E-03
	en	1.11E-02	3.88E-03
	Justo	1.05E-02	3.53E-03
	como	7.86E-03	3.09E-03
iwáátsúcúpéjtsóóbeke	izquierda	1.88E-02	1.53E-01
	obedecieron	1.79E-03	9.62E-02
	comenzado	1.68E-03	8.45E-02
	mucho	1.48E-03	8.32E-02
	yendo	2.10E-03	3.47E-02

Figure. Example of candidate Bora to Spanish phrase translations, along with their corresponding conditional likelihoods.

Results *-Without Data Augmentation*

1. Train & Valid : unparallel.
   - Source language (bora) : Bible New Testament
   - Target language (spanish): Some novels
2. Test: Universal Declaration of Human Rights
3. Bleu Score

Data Augmentation	Bora Training Dataset	Spanish Training Dataset	Test Data	Bleu 1-gram		Bleu 2-gram	
				PBSMT	NMT	PBSMT	NMT
×	Bible	Novels	Human Rights	1.1	0.7	0.0	0.0

Results *-Without Data Augmentation*

1. Train & Valid: parallel.
   - Source language (bora) : Bible New Testament
   - Target language (spanish): Bible
2. Test: Universal Declaration of Human Rights
3. Bleu Score

Data Augmentation	Bora Training Dataset	Spanish Training Dataset	Test Data	Bleu 1-gram		Bleu 2-gram	
				PBSMT	NMT	PBSMT	NMT
×	Bible	Novels	Human Rights	1.1	0.7	0.0	0.0
×	Bible	Bible	Human Rights	2.7	1.5	0.2	0.0

Results *-Without Data Augmentation*

1. Train & Valid: parallel.
   - Source language (bora) : Bible New Testament
   - Target language (spanish): Bible
2. Test: Part of Bible
3. Bleu Score

Data Augmentation	Bora Training Dataset	Spanish Training Dataset	Test Data	Bleu 1-gram		Bleu 2-gram	
				PBSMT	NMT	PBSMT	NMT
×	Bible	Novels	Human Rights	1.1	0.7	0.0	0.0
×	Bible	Bible	Human Rights	2.7	1.5	0.2	0.0
×	Bible	Bible	Part of Bible	4.6	2.1	0.3	0.1

Results *-With Data Augmentation*

1. Train & Valid: parallel.
   - Source language (bora) : Bible New Testament
   - Target language (spanish): Bible
2. Test: Universal Declaration of Human Rights
3. Bleu Score

Data Augmentation	Bora Training Dataset	Spanish Training Dataset	Test Data	Bleu 1-gram		Bleu 2-gram	
				PBSMT	NMT	PBSMT	NMT
×	Bible	Novels	Human Rights	1.1	0.7	0.0	0.0
×	Bible	Bible	Human Rights	2.7	1.5	0.2	0.0
×	Bible	Bible	Part of Bible	4.6	2.1	0.3	0.1
✓	Bible	Bible	Human Rights	3.2	1.9	0.0	0.0

Results *-With Data Augmentation*

1. Train & Valid: parallel.
   - Source language (bora) : Bible New Testament
   - Target language (spanish): Bible
2. Test: Part of Bible
3. Bleu Score

Data Augmentation	Bora Training Dataset	Spanish Training Dataset	Test Data	Bleu 1-gram		Bleu 2-gram	
				PBSMT	NMT	PBSMT	NMT
×	Bible	Novels	Human Rights	1.1	0.7	0.0	0.0
×	Bible	Bible	Human Rights	2.7	1.5	0.2	0.0
×	Bible	Bible	Part of Bible	4.6	2.1	0.3	0.1
✓	Bible	Bible	Human Rights	3.2	1.9	0.0	0.0
✓	Bible	Bible	Part of Bible	5.3	2.4	0.5	0.3

2

1 Y COMO fué nacido Jesús en Bethlehem de Judea en días del rey Herodes , he aquí unos magos vinieron del oriente á Jerusalem , 2 Diciendo : ¿ Dónde está el Rey de los Judíos , que ha nacido ? porque su estrella hemos visto en el oriente , y venimos á adorarle. 3 Y oyendo esto el rey Herodes , se turbó , y toda Jerusalem con él. 4 Y convocados todos los príncipes de los sacerdotes , y los escribas del pueblo , les preguntó dónde había de nacer el Cristo. 5 Y ellos le dijeron : En Bethlehem de Judea ; porque así está escrito por el profeta : 6 Y tú , Bethlehem , de tierra de Judá , no eres muy pequeña entre los príncipes de Judá ; porque de ti saldrá un guiador , que apacentará á mi pueblo Israel. 7 Entonces Herodes , llamando en secreto á los magos , entendió de ellos diligentemente el tiempo del aparecimiento de la estrella ; 8 Y enviándolos á Bethlehem , dijo : Andad allá , y preguntad con diligencia por el niño ; y después que le hallareis , hacédmelo saber , para que yo también vaya y le adore. 9 Y ellos , habiendo oído al rey , se fueron : y he aquí la estrella que habían visto en el oriente , iba delante de ellos , hasta que llegando , se puso sobre donde estaba el niño. 10 Y vista la estrella , se regocijaron con muy grande gozo. 11 Y entrando en la casa , vieron al niño con su madre María , y postrándose , le adoraron ; y abriendo sus tesoros , le ofrecieron dones , oro é incienso y mirra. 12 Y siendo avisados por revelación en sueños que no volviesen á Herodes , se volvieron á su tierra por otro camino. 13 Y partidos ellos , he aquí el ángel del Señor aparece en sueños á José , diciendo : Levántate , y toma al niño y á su madre , y huye á Egipto , y estáte allá hasta que yo te lo diga ; porque ha de acontecer , que Herodes buscará al niño para matarlo. 14 Y él despertando , tomó al niño y á su madre de noche , y se fué á Egipto ; 15 Y estuvo allá hasta la muerte de Herodes : para que se cumpliese lo que fué dicho por el Señor , por el profeta que dijo : De Egipto llamé á mi Hijo. 16 Herodes entonces , como se vió burlado de los magos , se enojó mucho , y envió , y mató á todos los niños que había en Bethlehem y en todos sus términos , de edad de dos años abajo , conforme al tiempo que había entendido de los magos. 17 Entonces fué cumplido lo que se había dicho por el profeta Jeremías , que dijo : 18 Voz fué oída en Ramá , grande lamentación , lloro y gemido : Rachêl que llora sus hijos ; y no quiso ser consolada , porque perecieron. 19 Mas muerto Herodes , he aquí el ángel del Señor aparece en sueños á José en Egipto , 20 Diciendo : Levántate , y toma al niño y á su madre , y vete á tierra de Israel ; que muertos son los que procuraban la muerte del niño. 21 Entonces él se levantó , y tomó al niño y á su madre , y se vino á tierra de Israel. 22 Y oyendo que Archelao reinaba en Judea en lugar de Herodes su padre , temió ir allá : mas amonestado por revelación en sueños , se fué á las partes de Galilea. 23 Y vino , y habitó en la ciudad que se llama Nazaret : para que se cumpliese lo que fué dicho por los profetas , que había de ser llamado Nazareno .

3

1 Y EN aquellos días vino Juan el Bautista predicando en el desierto de Judea , 2 Y diciendo : Arrepentíos , que el reino de los cielos se ha acercado. 3 Porque éste es aquel del cual fué dicho por el profeta Isaías , que dijo : Voz de uno que clama en el desierto : Aparejad el camino de l Señor , enderezad sus veredas. 4 Y tenía Juan su vestido de pelos de camellos , y una cinta de cuero alrededor de sus lomos ; y su comida era langostas y miel silvestre. 5 Entonces salía á él Jerusalem , y toda Judea , y toda la provincia de alrededor del Jordán ; 6 Y eran bautizados de él en el Jordán , confesando sus pecados. 7 Y viendo él muchos de los Fariseos y de los Saduceos , que venían á su bautismo , decíales : Generación de víboras , ¿ quién os ha enseñado á huir de la ira que vendrá ? 8 Haced pues frutos dignos de arrepentimiento , 9 Y no penséis decir dentro de vosotros : A Abraham tenemos por padre : porque yo os digo , que puede Dios despertar hijos á Abraham aun de estas piedras. 10 Ahora , ya también la segur está puesta á la raíz de los árboles ; y todo árbol que no hace buen fruto , es cortado y echado en el fuego. 11 Yo á la verdad os bautizo en agua para arrepentimiento ; mas el que viene tras mí , más poderoso es que yo ; los zapatos del cual yo no soy digno de llevar ; él os bautizará en Espíritu Santo y en fuego. 12 Su aventador en su mano está , y aventará su era : y allegará su trigo en el alfolí , y quemará la paja en fuego que nunca se apagará. 13 Entonces Jesús vino de Galilea á Juan al Jordán , para ser bautizado de él. 14 Mas Juan lo resistía mucho , diciendo : Yo he menester ser bautizado de ti , ¿ y tú vienes á mí ? 15 Empero respondiendo Jesús le dijo : Deja ahora ; porque así nos conviene cumplir toda justicia. Entonces le dejó. 16 Y Jesús , después que fué bautizado , subió luego del agua ; y he aquí los cielos le fueron abiertos , y vió al Espíritu de Dios que descendía como paloma , y venía sobre él. 17 Y he aquí una voz de los cielos que decía : Este es mi Hijo amado , en el cual tengo contentamiento .

5,1 0%

que estaban sin ser leudado por mano de Nathán sin

2

tenerle los otros instrumentos

1 el natural tendrá temor aparejó Beracah otros mensajeros miraba cómo ha angustiado en todo gozo también se revolvía tenerle que hayas derramado cansado 2 Entonces los discípulos fué con su

sin morador. manda hoy se ha subido buscad siempre arco se pasará por ella fué gozo cuando el rey en su casa triste tu zafiro se haces que guardarían los

3 El que os dijere. los Gabaón con los asnos por su vida. persona cierta que pues me diste hace . 4 Nuestro eres justo. por falta de todo tu asiento agrado la ordenanza de seiscientos 5 Ahora pues es

sin pecado natural en ella los perros llenado de vosotros habéis hecho. Salomón juró

6 Todo natural han pasado por Macedonia Ahava Beracah subir de la ciudad fué echado su comida sesenta estadios le juró por el pecado. comiste

7 estaban por menester detrás del temor de revolvía tenerle los discípulos de su señor estaba Dijéronle hay sabiduría. El pueblo 8 Si alguno seréis muertos con

duerme. ha hallado engaño en su confusión. repartir con la Antes que hirió por causa de la gruesa o

9 nunca tal Vive tu fornicación en la mente de los collados se le daba agua recobrar hay quien oiga raiga su 10 tenido en oídos de 11 Entonces pusieron sus cámaras eran cámara pues confía. sea probado con los que guardaren vendieron se alegran el camino de heriré con los otros es ganancia. 12 Bienaventurado el varón que eran cantores pezuña la gente que cayere sobre mí veía que tenía

entró hombre- juntamente fué rota

13 Al que veía saqué pastor que juró por su reposaron alcanzare su mano con que cayese

pues mundos confiad venir costados juntamente con gran salud por la sangre que estará en o por contienda ó pulgón pasar adelante los cantores

14 Griegos tomando luego la prenda cuando juntamente fué en cada ciudad 15 Voz que estaban en Jerusalem esto enalbardó su sin ser sólo el profeta Nathán con

ó del presidente como los sacerdotes

eran cantores del servicio.

16 los cantores primas perfumes apedreáronlo de tenerle que tenía despertarán los que fué demos la nueva mil jugo los edificadores con menester tenerle toda ella mil armados pagase todo ephacabal estará en saliendo nuestras ofrendas Vosotros dió grita con gran voz de 17 estaban algunas de vosotros ha mi salud están continuamente con

18 Cuando estuviereis quebró engañan los corazones tal digan fuentes vivas contienda ni fueron los me los diste en su pico

19-20 en los cantores los hizo Salomón juró por veinticinco reposaron alcanzare su ella fué con

mundos pues pervertisteis las demás con los que la salud de Salmón presentarse

21 luego la salud de Salomón. 22 dijeron tus enemigos. pues Samsón fué detenida pendencia es como ha provocado ha pegado Joiada entero no sabían alabar. es muerta en la alcanzare su enfermedad. la sangre de ella le acompañaron 23 Después procuraréis la proa vino bebieron fuego se ha derramado por tanto será

3

de todas riquezas espías

No seas 1.1-8 de grave 3.1-9 que está 1.19-28 de

1 mancebos espías juró por su fin caminado en quisiera blanca los perros vivió palabras. 2

pues Esther limpió la gloria de su pecado. Roboam amó gloria en su ciudad de mano de los espíritus se hubo entre los librase

3 Dijo aún de vosotros ha vino corriendo fué con

nombres quisiera en ella fué tomado Saúl tomó consejo con voz limpió tu ungido morirá. oyó la reina sido hecho

4 El quitó siguiólo obrado hoy con sangre viven entre los quiso morar en Harada. hasta en derecho de los leones 5 Teniendo volviendo hechos calles guardan persona firmes en su vida. Viento más que los Harás también su boca en el día que 6 Haz hasta sesenta Zebul su sed cuando los fundamentos

21,1 0%

05

Conclusion

Conclusion

- **Related data**
  - source dataset and target dataset - **Bible**
  - test dataset and train dataset - **Bible vs. Human Rights**
- **PBSMT model is the best model fit our dataset.**
- **Provide dataset of Bora**
  - stable performance
  - laid the foundation for further related research

06

Future Works

Future Works

- **Enlarge and improve the diversity of the dataset**
  - Collect more data
  - Better Data augmentation techniques
- **Modify translation models**
  - Combine the statistical machine translation model and the neural machine translation model

07

Q&A

Thanks!

Nov 2020