Efficacy of ByT5 in Multilingual Translation of Biblical Texts for Underrepresented Languages

C. Aars,* L. Adams,* X. Tian,* Z. Wang,* C. Wismer,* J. Wu,* P. Rivas, K. Sooksatra, and M. Fendt

*Equal Contribution. Department of Computer Science, School of Engineering and Computer Science, Baylor University.

Motivating Problem

- Traditional human translation of sacred texts into **underrepresented languages** is very **time-consuming** and resource-intensive.
- The **scarcity of skilled translators** for these languages delays the availability of important cultural & religious texts.
- Advanced models like **ByT5 can enhance the accuracy of these translations**, improving accessibility to sacred literature.

Contributions

- Developed a **ByT5-based multilingual** translation model tailored for translating the Bible into **underrepresented languages**.
- Trained the model on the **Johns Hopkins University Bible Corpus** to capture intricate linguistic nuances of character-based and morphologically rich languages.
- Achieved significant improvements in **translation quality and accessibility** of sacred texts for underrepresented languages using the BLEU score and sample translations.

Methodology Overview

1. **Model Selection and Training:**
   - The ByT5 model was chosen for its proficiency in byte-level tokenization, which is crucial for handling subtle linguistic differences in underrepresented languages.
   - The dataset used was the Johns Hopkins University Bible Corpus, encompassing a wide array of linguistic variations, providing a comprehensive foundation for the model.
2. **Parameter Tuning:**
   - The training hyperparameters were optimized through an iterative experimental protocol, focusing on adjusting the learning rate (LR), early stopping, patience, and batch size.
   - Final configuration: an LR of 0.0002, a scheduler factor of 0.5, patience of 10, a batch size of 48, and 50 max epochs, with early stopping typically occurring around epoch 43.
3. **Data Handling:**
   - The dataset comprised 3 million pairs of source-target translations, cleaned and aligned to ensure verse-parallel structure, which is ideal for machine learning applications.
   - A byte-level tokenization enabled the model to handle character-based and morphologically rich languages, leading to better generalization across languages.
4. **Evaluation Metrics:**
   - The translation quality was assessed using the BLEU score, comparing the model's outputs with human translations.
   - Additional qualitative analysis was conducted through a series of sample translations highlighting its capability.

- The translations below illustrate ByT5's ability to produce accurate translations across multiple languages, demonstrating its effectiveness in maintaining semantic integrity and handling diverse linguistic structures, particularly for underrepresented languages.

Language	Target text	Translated text
English	Jesus heard that John was put in prison , so he went back to Galilee.	When Jesus heard that John had been arrested, he went back to Galilee.
	And He warned them to tell no one about Him.	And he charged them that they should tell no man of him.
	Today your Savior was born in David's town. He is Christ, the Lord.	for today in the city of David a Saviour was born to you, which is Christ the Lord.
Spanish	porque decía : « Si toco tan sólo su manto , seré salva . »	Porque decía: Si tocare tan solamente su vestido, seré salva.
	y diciendo : « Tú , el que derribas el Templo y en tres días lo reedificas , sálvate a ti mismo . Si eres Hijo de Dios , desciende de la cruz . »	y diciendo: Tú, el que derribas el templo, y en tres días lo reedificas, sálvate a ti mismo. Si eres Hijo de Dios, desciende de la cruz.
	Jesús salió llevando su cruz , para ir al llamado « Lugar de la Calavera » (que en hebreo se llama Gólgota) .	Y llevando su cruz, salió al lugar que se dice de la Calavera, y en hebreo Gólgota;
Marathi	म्हणून ज्याच्याविषयी तो बोलतो तो कोण आहे हे आम्हास सांग, असे शिमोन पेत्राने त्यास खुणावून म्हटले.	शिमोन पेत्राने त्या शिष्याला खुणावले आणि म्हणाला, " ज्याच्याविषयी तो बोलतो तो कोण आहे हे विचार.
	आणि तेथे काही मेंढपाळ रात्रीच्या वेळी शेतामध्ये राहून आपले कळप राखीत होते.	आणि तेथे काही मेंढपाळ रात्रीच्या वेळी शेतामध्ये राहून आपले कळप राखीत होते.
	तर आता काय करावे? तुम्ही आला आहात हे ते खातरीपूर्वक ऐकतील.	" मग आता काय केले पाहिजे? तू येथे आला आहेस हे त्यांना नक्की कळेल.
Chinese	说完，他又告诉他们:"我们的朋友拉撒路睡着了，我去叫醒他。"	说完，他又告诉他们:"我们的朋友拉撒路睡着了，我去叫醒他。"
	当他发现，这么做很讨好犹太人时，便索性又把彼得也抓了起来。（当时正值逾越节期间）	当他发现，这么做很讨好犹太人时，便索性又把彼得也抓了起来。（当时正值逾越节期间）
	因为我有五个兄弟，让拉撒路告诫他们，以便将来他们不会到这个受尽折磨的地方来。'	因为我有五个兄弟，让拉撒路告诫他们，以便将来他们不会到这个受尽折磨的地方来。'

- Below are additional sample translations generated by ByT5, accurately translating complex religious narratives across a variety of languages, including those that are character-centric and underrepresented.

Language	Target text	Translated text
English	Since that harbor was not a good place for the ship to stay for the winter , most of the men decided that the ship should leave . They hoped we could go to Phoenix and stay there for the winter . Phoenix , a city on the island of Crete , had a harbor which faced southwest and northwest .	And because the harbor was not commodious to winter in, the more part decided to put to sea from there, if somehow they could reach Phoenix, and winter there; which is an harbor of Crete, facing southwest and northwest, and spend the winter there.
	And when they had entered, they went up to the upper room, where they were staying, Peter and John and James and Andrew, Philip and Thomas, Bartholomew and Matthew, James the son of Alphaeus and Simon the Zealot and Judas the son of James.	And when they were come in, they went up into the upper chamber, where they were abiding; both Peter and John and James and Andrew, Philip and Thomas, Bartholomew and Matthew, James the son of Alphaeus, and Simon the Zealot, and Judas the son of James.
German	Und diese Rede von ihm verbreitete sich in ganz Judäa und in die ganze Umgegend .	Und diese Rede über ihn erscholl in ganz Judäa und in alle umliegenden Länder.
	Es kamen auch Zöllner , um sich taufen zu lassen , und sagten zu ihm : " Meister , was sollen wir tun ? "	Es kamen aber auch Zöllner, daß sie sich taufen ließen, und sprachen zu ihm : Meister, was sollen denn wir tun?
France	Et ce qui est tombé entre des épines , ce sont ceux qui ayant ouï la parole , et s'en étant allés , sont étouffés par les soucis , par les richesses , et par les voluptés de cette vie , et ils ne rapportent point de fruit à maturité .	Et ce qui est tombé parmi les épines, ce sont ceux qui, ayant entendu la parole, et s'en étant allés, sont étouffés par les soucis, par les richesses et par les voluptés de cette vie, et ils ne portent point de fruit à maturité.
	Si donc , méchants comme vous l'êtes , vous savez donner de bonnes choses à vos enfants , à combien plus forte raison le Père céleste donnera-t-il le Saint-Esprit à ceux qui le lui demandent .	Si donc vous, qui êtes méchants, savez donner à vos enfants des choses bonnes, combien plus votre Père céleste donnera-t-il le Saint-Esprit à ceux qui le lui demandent?
Russian	но писал вам , братия , с некоторою смелостью , отчасти как бы в напоминание вам , по данной мне отБога благодати	Но я довольно смело писал вам о некоторых делах, которые мне хотелось бы, чтобы вы запомнили. Я сделал это, потому что даровано мне было по
	И в другом месте Писания говорится: «Они будут смотреть на Того, Которого пронзили»†.	И ещё в другом [месте] Писания говорится : воззрят на Того, Которого пронзили.
Japanese	だから、兄弟たちよ、この事を承知しておくがよい。すなわち、このイエスによる罪のゆるしの福音が、今やあなたがたに宣べ伝えられている。そして、モーセの律法では義とされることができなかったすべての事についても、	だから、兄弟たちよ、この事を承知しておくがよい。このことによる罪のゆるしの福音が、今あなたがたに宣べ伝えられている。そして、モーセの律法では義とされることがないので
	これらのことを話したのは、あなたがたがわたしによって平和を得るためである。あなたがたには世で苦難がある。しかし、勇気を出しなさい。わたしは既に世に勝っている。」	これらのことをあなたがたに話したのは、わたしにあって平安を得るためである。あなたがたは、この世ではなやみがある。しかし、勇気を出しなさい。わたしは世に勝っている」。

Conclusions

- The ByT5 model achieved a BLEU score of 0.27, correctly handling translations of underrepresented languages.
- Advanced NLP models can significantly improve the accessibility and cultural preservation of sacred texts for communities with limited linguistic resources.

Work partiall funded by the NSF under grants CNS-2136961 and CNS-2210091.

SCAN ME

NAACL 2024

LXAI

NSF

BU ENGINEERING & COMPUTER SCIENCE