NATURAL LANGUAGE INFERENCE IN TAMIL : DATASET AND EVALUATION

Submitted By

K. EZHILARASI (18MT8203)

Under The Guidance Of

L. JAYASREE

Assistant Professor (Ph.D)

Department Of Computer Science and Engineering

TABLE OF CONTENT

- Abstract
- Introduction
- Related works
- Background
- Textual Entailment
- Our work
- Conclusion
- Reference

NATURAL LANGUAGE INFERENCE IN TAMIL : DATASET AND EVALUATION

ABSTRACT

Natural Language Inference (NLI) has been believed to test a model's language understanding capability. Recent works like Multilingual BERT has raised significant interest in cross-lingual NLI in the Natural Language Processing (NLP) community. In this work, a new Cross-lingual Natural Language Inference (NLI) dataset for the Tamil Language is created by translating the Cross-Lingual Natural Language Inference (XNLI) test dataset. Further, the baselines on our dataset are provided. The newly created dataset would help improve the Natural Language Processing in Tamil, especially with the ongoing research in cross-lingual learning.

INTRODUCTION

- We choose the Tamil language because it is a good representative of the Dravidian family of languages
    - Existing cross-lingual NLI dataset doesn't contain any language from the Dravidian family
    - Further, we hypothesis that the performance of a cross-lingual technique or model in Tamil could be generalized to Telugu, Kannada, and Malayalam as well.
- We choose Textual Entailment, also known as Natural Language Inference, because it is believed that NLI tests a model's language understanding capabilities.
- We provide baselines using two existing models
    - Multilingual BERT (M-BERT)
    - Extended Multilingual BERT (E M-BERT)

RELATED WORKS

XNLI: Evaluating Cross-lingual Sentence Representations

- Provides cross-lingual NLI dataset on 14 languages.
- Covers only two Indian Language – Hindi and Urdu.
- Created by translating the test data of Multi-Genre NLI (Multi NLI).
- Popular dataset to evaluate cross-lingual performance of a model.

SYSTEM REQUIREMENT

SOFTWARE REQUIREMENT

- PyTorch 1.3.1+
- TensorFlow 2.0
- Transformers
- Python 3.6+

HARDWARE REQUIREMENTS

- GPUs or TPUs
- In principle we can run on a CPU but takes several days.

BACKGROUND

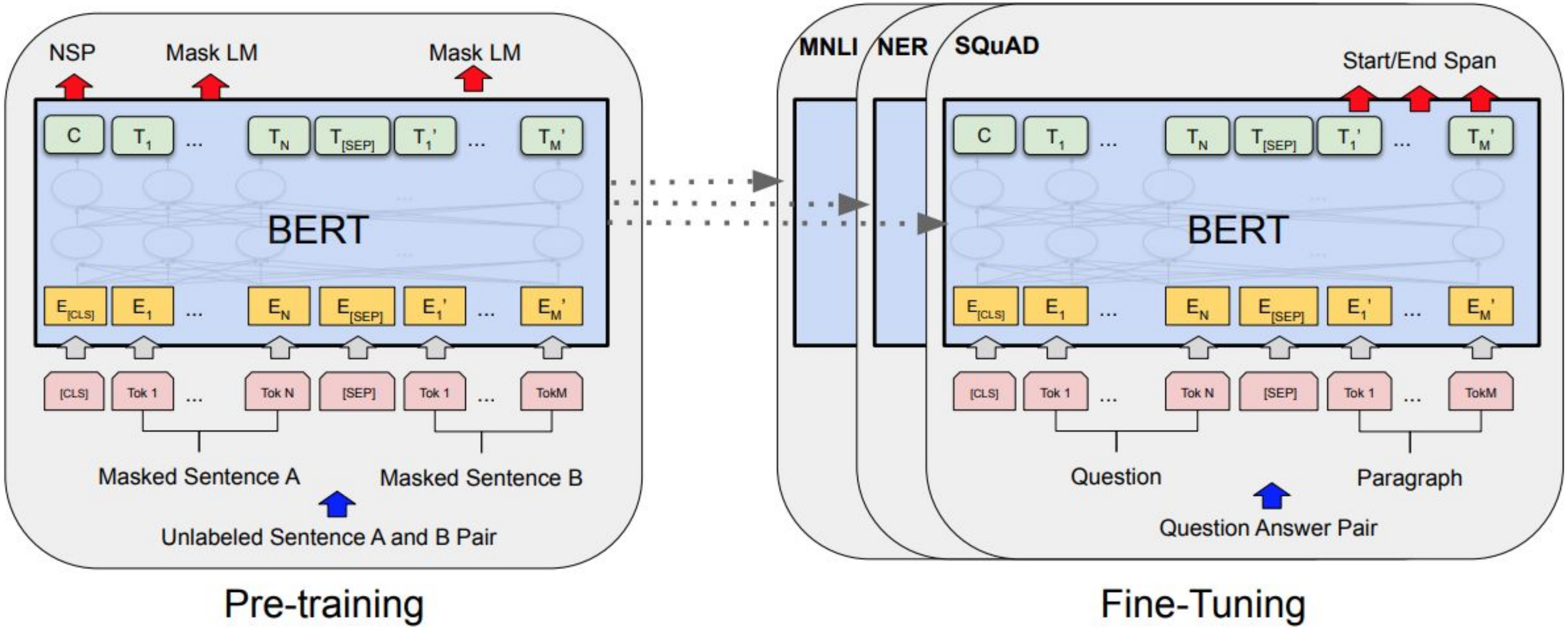
1. BERT

2. Multilingual BERT

3. Extended Multilingual BERT

1.Bidirectional Encoder Representations from Transformers **(BERT)**

- BERT is a Transformer-based pre-trained language representation model trained on English Wikipedia data.
- BERT is pre-trained using Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) Objective.
- Input to BERT is a pair of sentences A and B, such that half of the time B comes after A in the original text and the rest of the time B is a randomly sampled sentence.
- Some tokens from the input are randomly masked, and the MLM objective is to predict them.
- NSP objective is to predict whether the sentence B is actually next sentence or not.
- Typically, BERT is finetuned on the down-stream task.



Pre-training

Fine-Tuning

2.Multilingual BERT

- Multilingual BERT is pre-trained in the same way as monolingual BERT except using Wikipedia text from the top 104 languages.

- To account for the differences in the size of Wikipedia, some languages are sub-sampled, and some are super-sampled using exponential smoothing.

- Multilingual BERT works cross-lingually.

- It’s worth mentioning that there are no cross-lingual objectives specifically designed nor any cross-lingual data, e.g. parallel corpus, used.

3.Extended M-BERT

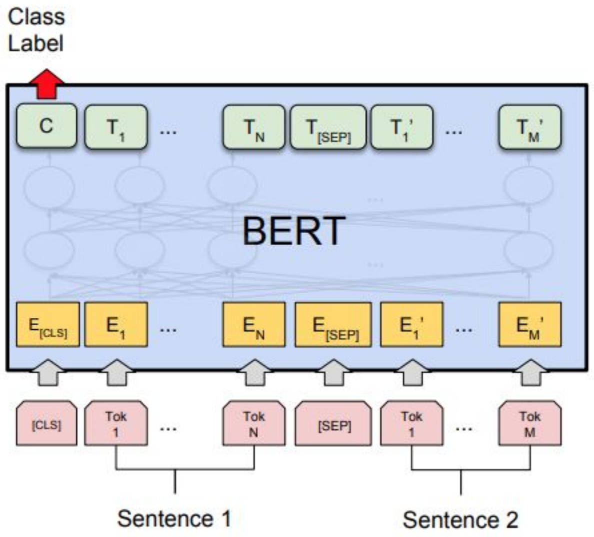
- Major disadvantage of M-BERT is it may not work for low resources languages.

- Extended M-BERT works by enlarging the vocabulary of M-BERT to accommodate the new language and then continue pre-training on this language.

- The perfomance improves significantly over M-BERT for both languages that are in M-BERT and new languages.

TEXTUAL ENTAILMENT

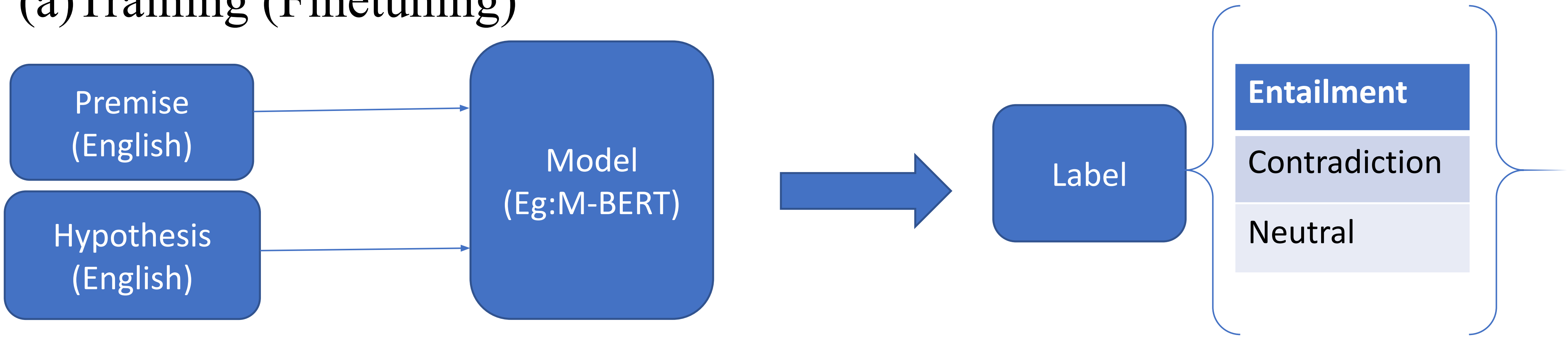
- Textual Entailment is also known as Natural Language Inference (NLI) is a sentence pair classification task.
- Given a premise and a Hypothesis we need to classify whether Premise entails Hypothesis.
- In other words, we need to classify whether we can infer Hypothesis from Premise.
- We use M-BERT and Extended M-BERT to train our models.



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

Cross-Lingual Textual Entailment (**CL M-BERT**)

(a)Training (Finetuning)



Premise (English)

Hypothesis (English)

Model (Eg:M-BERT)

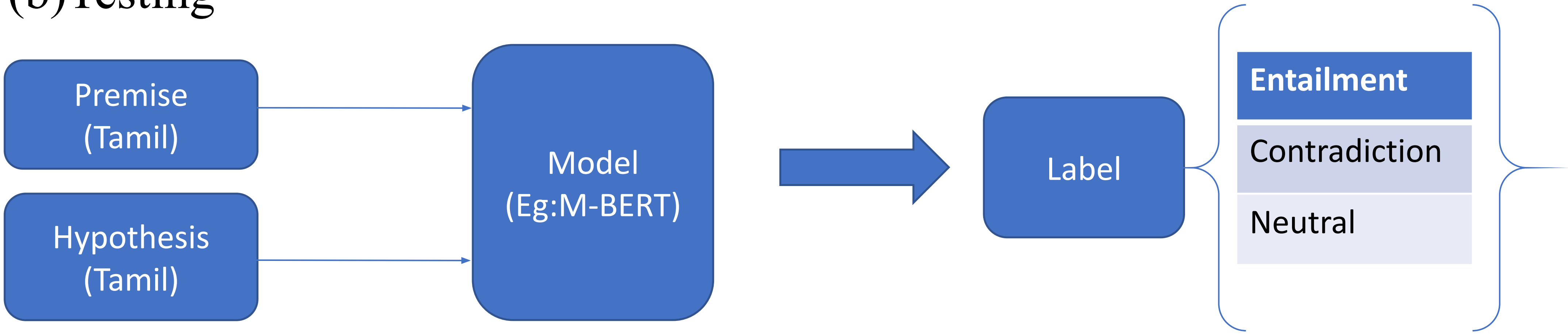
Label

Entailment

Contradiction

Neutral

(b)Testing



Premise (Tamil)

Hypothesis (Tamil)

Model (Eg:M-BERT)

Label

Entailment

Contradiction

Neutral

MOTIVATION

- Recent works show that NLI can be used for reasoning and zero-shot classification.
- **Problem:** Given a document we need to classify whether we can infer that people are suffering from water crisis.
- Instead of thinking it as a classification problem, think it as an Entailment or inference problem
  - Premise: *Given document*
  - Hypothesis: *People are suffering from water crisis*
  - Label: Can we infer hypothesis from Premis
- We will see a demo by AllenNLP to understand it better https://demo.allennlp.org/textual-entailment

OUR WORK - DATASET CREATION

- We used XNLI Test data in English and Translated them to Tamil
  - Training data is XNLI English data which is same as MultiNLI training data (**MultiNLI**)

- Out of 5000 sentence pairs, we human translated 1000 sentence pairs and Google translated all of them.
  - Human Translated – 1000
  - Google Translated – 5000

- Further, our dataset could be used to compare human translation and Google translation.

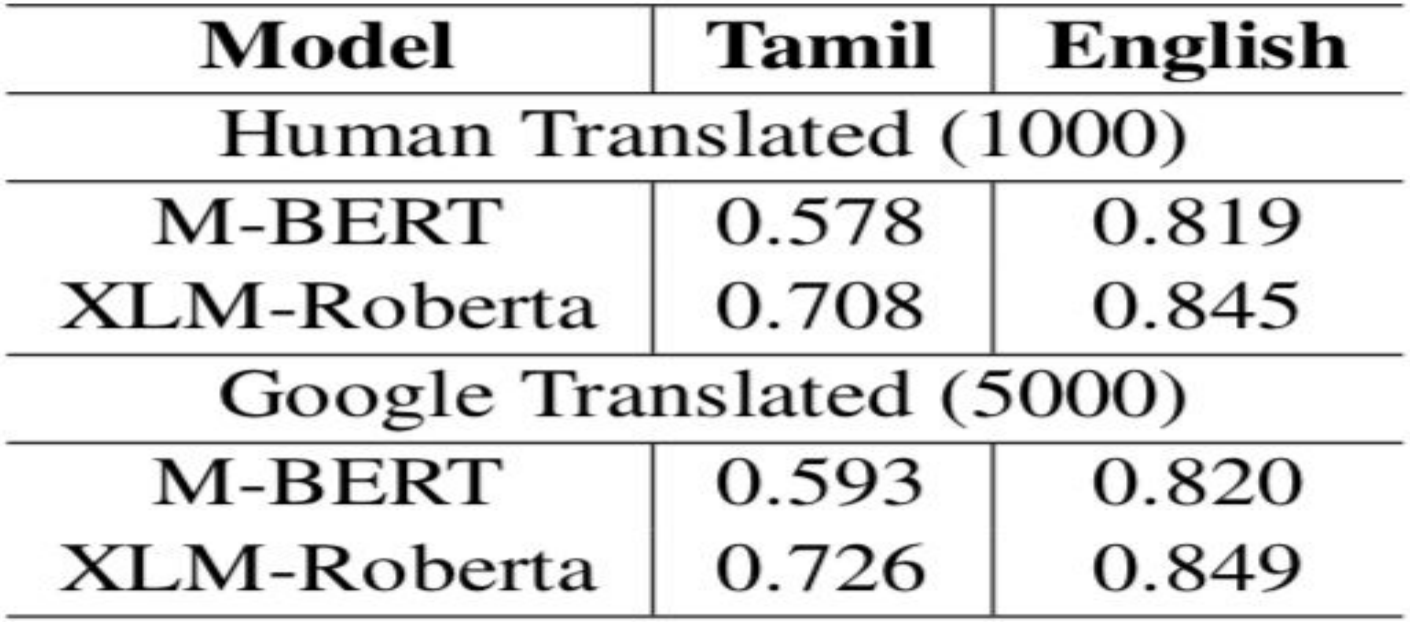
OUR WORK - EVALUATION

- We finetune M-BERT and Extended M-BERT on English MultiNLI dataset.

- We evaluate on the cross-lingual performance
  - 1000 human translated pairs
  - 1000 Google Translated pairs (same sentence pairs as human translated)
  - 5000 Google translated pairs

- Further, we Google Translate the human translated data from Tamil to English and then evaluate its performance using M-BERT

RESULT



Model	Tamil	English
Human Translated (1000)		
M-BERT	0.578	0.819
XLM-Roberta	0.708	0.845
Google Translated (5000)		
M-BERT	0.593	0.820
XLM-Roberta	0.726	0.849

Performance Evaluation: We report the accuracy on Tamil and English test set for both human and Google translated data (and its corresponding English data). We use pre-trained Multilingual BERT and XLM-Roberta as out initial models.

CONCLUSION

Until now, there exists no Cross-Lingual Natural Language Inference dataset for any of the Dravidian family of languages. I created a new NLI dataset for the Tamil language by translating the standard XNLI data from English to Tamil. I evaluated its performance using a state-of-the-art method called Extended Multilingual BERT as well as standard Multilingual BERT. Further, I compare performance on both humans and Google translated input to understand the quality of existing commercial translation models.

REFERENCE

- **M-BERT** - Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Multilingual bert - r, 2018. URL https://github.com/google-research/bert/blob/master/ multilingual.md
- **CL M-BERT** - K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. URL https://openreview.net/pdf?id=HJeT3yrtDr
- **BERT** - Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 4171– 4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.
- **E M-BERT** - Zihan Wang and Karthikeyan K and Stephen Mayhew and Dan Roth. Extending Multilingual BERT to Low-Resource Languages. URL https://arxiv.org/abs/2004.13640
- **MultiNLI** - Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn A. Walker, Heng Ji, and Amanda Stent (eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pp. 1112–1122. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1101. URL https://doi.org/10.18653/v1/n18-1101
- **XNLI** - Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: evaluating cross-lingual sentence representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pp. 2475–2485. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1269. URL https://doi.org/10.18653/v1/d18-1269