Journal of The Gujarat Research Society

ISSN: 0374-8588
Volume 21 Issue 16, December 2019

Natural Language Inference in Tamil: Dataset and Evaluation

[1]K. Ezhilarasi, [2] L. Jayasree

[1] PG Student, [2] Assistant professor

[1, 2] *Department of C.S.E, Sri Padmavathi Mahila Visvavidyalayam, Tirupati*

[1]ezhil.k@icloud.com,

[2]jayasreemohan15@gmail.com

***Abstract*—** *Natural Language Inference (NLI) has been believed to test a model's language understanding capability. Recent works like Multilingual BERT and XLM-Roberta has raised significant interest in zero-shot cross-lingual NLI in the Natural Language Processing (NLP) community. We observed that the current Cross-Lingual Natural Language Inference (XNLI) not having any language from the Dravidian family of languages. Therefore, in this work, we generate a new Cross-lingual Natural Language Inference (NLI) dataset for the Tamil Language through translation -- both human and machine translation -- the Cross-Lingual Natural Language Inference (XNLI) test dataset. Further, we provide baselines on our dataset. This dataset would help improve the Natural Language Processing in Tamil, especially with the ongoing research in cross-lingual learning.*

Keywords- *Natural Language Processing, Natural Language Inference, Textual Entailment, Cross Lingual Learning, NLP in Indian Language, Transformer Models*

I. INTRODUCTION

The recent development in transformer-based pretraining multilingual language models like Multilingual BERT (Devlin et al., 2018), elaborated Multilingual BERT (Wang et al., 2020), XLMRoberta (Conneau et al., 2019) has made a significant impact on the zero-shot cross-lingual learning (K et al., 2020) which is imperative especially for low and moderate resource languages as getting a huge training data in these languages is particularly harder. Although these models have made significant progress in the domain, the cross-lingual transfer depends a lot on the target language downstream task, amount of data in the pre-training, the typological similarity between the source and target language, etc..

There are not many datasets in Tamil for semantic tasks, particularly Natural language Inference (NLI) tasks which are believed to test the language understanding and reasoning ability of a model.

Among other semantic tasks, we choose Natural language Inference (NLI), (also known as Textual Entailment) because (1) NLI measures several aspects of natural language understanding such as Lexical Semantics, Predicate-Argument Structure, Logic, Knowledge, and Common sense (Wang et al., 2019) (2) also, most of the works benchmark their cross-lingual systems on a standard NLI dataset called Cross-Lingual Natural Language Inference (XNLI) (K et al., 2020; Devlin et al., 2018; Lample and Conneau, 2019).

Most of the Indian Languages belong to Indo-Aryan (mostly used in Northern India), Dravidian (mostly used in Southern India), and Sino-Tibetan family of languages. The existing XNLI dataset covers some of the Indo-Aryan languages (Hindi, Urdu) as well as the Sino-Tibetan language (Chinese), but it does not have any language from the Dravidian family of languages – Tamil, Telugu, Kannada, and Malayalam. Therefore, in this work, we created and evaluated Tamil NLI dataset which would be a good representative of the

Dravidian family of languages. To be consistent with the existing XLNI dataset, we translated the English XNLI data to the Tamil language.

We also provide a benchmark on our newly created Tamil NLI dataset using Multilingual BERT and XLM-Roberta pre-trained models, which are currently the most widely used and state-of-the-art models (for various asks and languages) for zero-shot cross-lingual learning.

The paper is organized as follows: We briefly discuss the related works and the background required to follow this paper. Then we discuss the dataset creation and evaluation. Finally, we conclude this paper with a brief discussion and potential future works.

The major contribution of this work is as follows: (1) We created the Tamil Natural Language Inference dataset (2) We provide a benchmark on our newly created dataset.

II. RELATED WORKS

Recent advances in natural language representation (Devlin et al., 2019) have a significant improvement on the zero-shot cross-lingual, in which we can train a NLP model using the supervised data of one language, mostly English, and then we can test (or use it) on other languages. It is particularly impressive for low-resource languages as obtaining annotated data in these languages is harder (both in terms of money and human resources). There are more than 4000 human languages in the world and getting a huge annotated corpus for each task, in each of the languages is very expensive as well as time-consuming. Using Zero-shot cross-lingual is very effective to solve this issue.

We choose Natural language Inference as it is one of the most popular semantic and natural language understanding tasks. Some of the most popular NLI datasets are Stanford Natural Language Inference (SNLI) (Bowman et al., 2015), Multi-Genre Natural Language Inference (MultiNLI), Cross-Lingual Natural language Inference (XNLI) (Conneau et al., 2018), Recognizing Textual Entailment (Wang et al., 2019; Dagan et al., 2013; Giampiccolo et al., 2007). Among them only XNLI is cross-lingual (test and validation of data exists in distinct languages but the training data exists particularly in English), others are available only for the English language. There has been a lot of interest in the NLI task as it involves several kinds of language exploration (linguistic inference, logical inference, commonsense, world knowledge, etc). Cross-Lingual Natural language Inference (XNLI) (Conneau et al., 2018) is closely related to out work where they provide human translated NLI test datasets on 15 languages. However, the XNLI doesn't contain any language from the Dravidian family of languages.

In this work, we rely on the zero-shot cross-lingual transferability of the transformer (Vaswani et al., 2017) based models (K et al., 2020), especially Multilingual BERT and XLM-Roberta. Both are transformer-based bidirectional language model which are pre-trained on huge unsupervised data.

The major advantage of using transformer-based models is that they do not require any parallel corpus or dictionary (Even though there are unsupervised word-alignment based methods where we first learn token representation for each of the languages independently and then align them, like MUSE or VecMap, transformer-based methods generally outperform them). Another advantage with models like M-BERT or XLM-Roberta is that they are massively multilingual, i.e. one model can be used for a lot of languages. All these Transformers models are based on the idea of a multi-head self-attention mechanism

III. BACKGROUND

3.1 Zero-Shot Cross Lingual Learning

Instead of training a fully supervised model with annotated data in low resource languages, we are able to train through the data from high resource language(s), typically English, and then test on a low resource language like Tamil. This is called zero-shot cross-lingual learning because we don't use any supervised or annotated data in the target language. Broadly, there are two paradigms for cross-lingual learning (1) Cross-Lingual alignment and (2) Joint training (Wang* et al., 2020).

In cross-lingual alignment, we first learn word/token/sentence embeddings in each of the source and target language independently and then learn the alignment using a bilingual dictionary, parallel corpora or comparable corpora, etc.. There are semi-supervised and unsupervised techniques to align as well. Some of the popular cross-lingual alignment methods are the Bilingual Skip-Gram Model (BiSkip) (Upadhyay et al., 2016; Luong et al., 2015), Bilingual Compositional Model (BiCVM) (Hermann and Blunsom, 2014), Bilingual Correlation Based Embeddings (BiCCA) (Faruqui and Dyer, 2014), Bilingual Vectors from Comparable Data (BiVCD) (Vulic´ and Moens, 2015), MUSE (Lample et al., 2017), VecMap (Artetxe et al., 2018).

In joint training, the representation of both the languages is learned jointly using some language modeling objective like Masked Language Model (requires no parallel data) or Translation Language Model (requires parallel corpus). Recent transformer-based Multilingual models like Multilingual BERT (Devlin et al., 2018), XLM (Lample and Conneau, 2019), XLM-Roberta (Conneau et al., 2019), Extended Multilingual BERT (Wang et al.,2020) falls under this category. One of the main advantages of such models is often they are massively multilingual (i.e. one model can work for many languages).

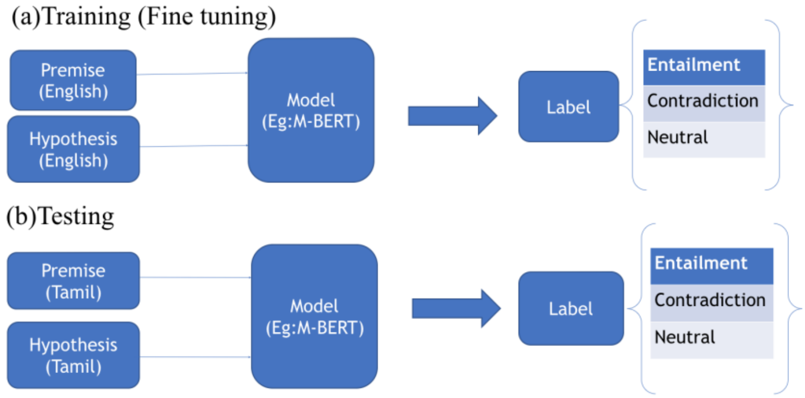


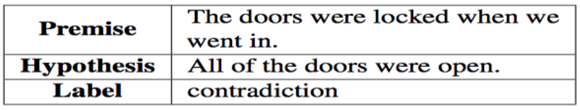
Figure 1 illustrates the cross-lingual textual entailment in which the model is first trained using English examples and then tested on Tamil examples

3.2 *Multilingual BERT*

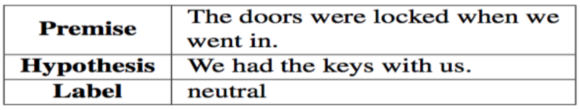
BERT (Devlin et al., 2019) is a contextual representation model based on deep transformers, trained using Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) objective. Unlike ELMo (Peters et al., 2018) or GPT (Radford, 2018), BERT is a bidirectional representation model. BERT considers two sentences together with few tokens as input so that 50 percent of time while the other sentence is the remaining 50 percent of time is a random sentence. The aim of MLM is to forecast masked token precisely, on the other hand the motto of NSP objective is forecasting whether the 2nd sentence is following the 1st sentence or not. BERT is trained using English Wikipedia data. BERT has two training phases called pre-training and fine-tuning. MLM and NSP are used during the pretraining phase to learn good language representation, then BERT is fine-tuned for the specific task. Often a new task-specific layer is added (which is randomly initialized) during the fine-tuning stage.Training objective and model architecture of Multilingual-BERT (Devlin et al., 2018) is the same as BERT, however it is trained on data of top 104 Wikipedia languages. They also sub-sample the high-resource languages and super-sample the low resource languages so that the difference in data sizes are taken into consideration.

3.3 Natural Language Inference

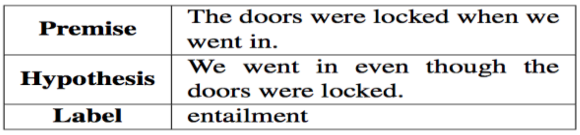
Natural Language Inference, also known as Textual Entailment is a natural language understanding task that was given two sentences called premise and hypothesis the objective is to predict whether the premise entails or contradicts or is not related to the hypothesis. In other words, we need to predict whether we can infer (or contradict) the hypothesis from the premise or not. It is widely believed that the performance on Natural language inference is a good measure of the language analyzing capacity of the model as NLI comprises different types of language understanding ranges from lexical semantics to logic and commonsense reasoning.



Premise	The doors were locked when we went in.
Hypothesis	All of the doors were open.
Label	contradiction

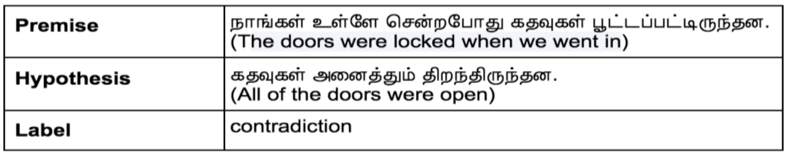


Premise	The doors were locked when we went in.
Hypothesis	We had the keys with us.
Label	neutral

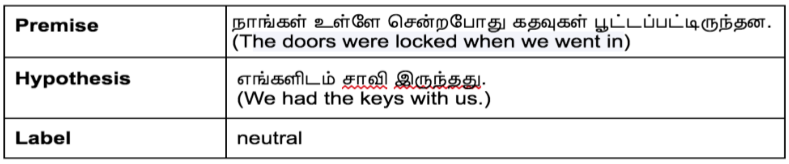


Premise	The doors were locked when we went in.
Hypothesis	We went in even though the doors were locked.
Label	entailment

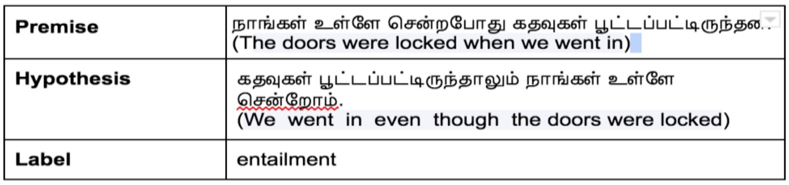
3.4 *Cross-lingual natural language inference (XNLI)*



Premise	நாங்கள் உள்ளே சென்றபோது கதவுகள் பூட்டப்பட்டிருந்தன. (The doors were locked when we went in)
Hypothesis	கதவுகள் அனைத்தும் திறந்திருந்தன. (All of the doors were open)
Label	contradiction



Premise	நாங்கள் உள்ளே சென்றபோது கதவுகள் பூட்டப்பட்டிருந்தன. (The doors were locked when we went in)
Hypothesis	எங்களிடம் சாவி இருந்தது. (We had the keys with us.)
Label	neutral



Premise	நாங்கள் உள்ளே சென்றபோது கதவுகள் பூட்டப்பட்டிருந்தன. (The doors were locked when we went in)
Hypothesis	கதவுகள் பூட்டப்பட்டிருந்தாலும் நாங்கள் உள்ளே சென்றோம். (We went in even though the doors were locked)
Label	entailment

This (Conneau et al., 2018) is one of the most popular Natural language inference datasets. XNLI uses MultiNLI (Williams et al., 2018) training data, which is a huge NLI corpus in English. MultiNLI spans over different domains like Face-to-face, Government, Fiction, Letters, Telephone Speech, 9/11 Report. XNLI has test and validation in 15 languages all these data are generated by translating the English test and validation dataset. Languages covered by XNLI are: English (en), French(fr), Spanish(es), German(de), Greek(el), Bulgarian(bg), Russian(ru), Turkish(tr), Arabic(ar), Vietnamese(vi), Thai(th), Chinese(zh)

Hindi(hi), Swahili(sw), Urdu(ur), Among the 15 languages Swahili and Urdu are relatively low resource languages. In each language, XNLI has about 5000 tests and 2500 dev premise-hypothesis pairs.

IV. DATASET

4.1 Tamil Cross-Lingual NLI Data Creation

In this section, we describe the step-by-step procedure on how we created our Tamil Natural Language Inference (NLI) dataset. Please refer table 2 for the Tamil NLI sample.

1. First we took 5000 English XNLI test data
2. Then we translated all of them from English to Tamil using Google Translate API.
3. Among the 5000 Google translated examples, we choose 1000 examples
4. We used Native speakers to verify and correct each of the 1000 translated examples. We found that using the native speakers to translate directly from English to Tamil is less effective (more time consuming) as well as diverse when compared with the the case where we provided them with google translated examples (along with the original English examples).

4.2 Data Statistics

In this section, we present the complete data statistics of the newly created Tamil NLI dataset. In table 1 we present the mean and standard deviation of the number of words in premise and hypothesis for both English and Tamil. We can see that on average Tamil sentences are shorter in word length when compared with English sentences. Further, in table 2 we also present the mean and standard deviation of many characters. While contrast to length of the word, We are able to identify that the characters inTamil sentences are lenghthy. In table 3 we also present the number of examples for each the 3 labels – entailment, contradiction and neutral. We can see that the dataset is balanced.

Language	Premise Mean(Std)	Hypothesis Mean(Std)
Human Translated (1000)		

Tamil
English
	12.04(6.13)
	17.30(8.78)
		6.41(2.54)
		8.75(3.56)

Google Translated (5000)		

Tamil
English
	13.46(6.13)
	18.22(8.11)
		6.79(2.65)
		9.30(3.59)

Table 1: Number of Words Comparison: We reported the mean and standard deviation (inside parenthesis) of number of words in English and Tamil data for both machine translated (5000 examples) and human translated (1000 examples) Tamil Dataset and their corresponding English data

V. EVALUATION & OBSERVATION

We start with pre-trained multilingual BERT and XLM-Roberta, both of which has been pretrained on Tamil and English (along with many other languages) and then fine-tune using English MultiNLI dataset. We used MultiNLI development set to choose the best model. To select the best model we tried different learning rates – 5e −6 , 1e −5 , 2e −5 , 3e −5 , 5e −5 and trained each of them for up to 3 epochs. We also repeated the experiment with different random initialization of the newly created added weights (weights from [CLS] layer to output layer are newly added to Multilingual BERT)

Language	Premise Mean(Std)	Hypothesis Mean(Std)
Human Translated (1000)		

Tamil
English
	99.24(52.47)
	86.34(45.72)
		54.10(22.12)
		44.03(17.91)

Google Translated (5000)		

Tamil
English
	119.18(56.71)
	106.53(49.96)
		60.64(24.34)
		52.15(20.76)

Table 2: Number of Characters Comparison: We reported the mean and standard deviation (inside parenthesis) of number of characters in English and Tamil data for both machine translated (5000 examples) and human translated (1000 examples) Tamil Dataset and their corresponding English data.

Label	Human Translated	Google Translated
Entailment Contradiction Neutral	333 333 334	1670 1670 1670l

Table 3: Label Summary: We provide the number of examples for each of the three labels for both human translated (1000) and Google translated (5000) data.

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

Table 4: 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

From table 4, we can observe that XLM-Roberta performs significantly better than M-BERT; this is expected as XLM-Roberta is trained on a huge multilingual pre-training corpus then Multilingual BERT. It is interesting to note that the difference between the English performance of XLM-Roberta and M-BERT is quite small when compared to the difference between their Tamil performance. Also note that Both M-BERT and XLM-Roberta performs better on Google translated data than Human Translated data, which indicates that it is slightly difficult to comprehend the human written Tamil. When we compare the XLM-R results with other languages reported on their paper Hindi(72.4), Swahili(66.5) and Urdu(68.3) we can see that the performance on Tamil is as good as Hindi (on Human translated its is slightly lower) and better than the performance on Swahili and Urdu, which are comparatively low resource languages

VI. CONCLUSION

One of the reasons we choose Natural language inference is that it can be utilized to solve a wide range of classification problems. Recent works (Yin et al., 2019) have shown that any text classification problem can be converted to NLI problem. Further, it is also known that Question-Answering can also be solved using NLI or Textual Entailment (Harabagiu and Hickl, 2006; Negri and Kouylekov, 2009). We further believe that NLI can be used as a probing mechanism to understand Natural Language Processing Models. Therefore, we hope that having an NLI dataset in Tamil language could improve the NLP research in Tamil for a wide range of tasks.

In this work, we choose both premise and hypothesis to be in Tamil (or English), but it has been shown that for some languages (Spanish,Hindi and Russian) transformer based models might not work when premise and hypothesis are in different languages (K et al., 2020). Our dataset can be utilized in further study how the system performs when premise is in English and Hypothesis is in Tamil or the vice versa. In this work, we experimented with only 2 models M-BERT and XLM-Roberta, it would be nice to explore a wide range of models like Extended Multilingual BERT, RNN (Long short-term memory (LSTM) or Gated Recurrent Unit (GRU) based) with different cross-lingual alignment based embeddings (Upadhyay et al., 2016).

Further, it would also be interesting to study different aspects of structural similarity (K et al., 2020) of Tamil with other languages and see if some other languages transfers well (i.e. is it beneficial if we have training data in some other languages like Hindi). It would also be interesting to study how the system performs with few-shot training (Lauscher et al., 2020) (i.e. Can we improve the model significantly by using a few Tamil examples?).

REFERENCES

1. “Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. “A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings”. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 789–798.
2. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettle- ´ moyer, and Veselin Stoyanov. 2019. “Unsupervised cross-lingual representation learning at scale”. arXiv preprint arXiv:1911.02116.
3. Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating crosslingual sentence representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.
4. Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. Synthesis Lectures on Human Language Technologies, 6(4):1–220.
5. J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT.
6. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Multilingual bert - r.
7. Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

8. Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing, pages 1–9.
9. Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. CoRR, abs/2003.11080.
10. Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In International Conference on Learning Representations.
11. Guillaume Lample and Alexis Conneau. 2019. Crosslingual language model pretraining. Advances in Neural Information Processing Systems (NeurIPS).
12. Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. arXiv preprint arXiv:1711.00043.
13. Anne Lauscher, V. Ravishankar, Ivan Vulic, and Goran Glavas. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. ArXiv, abs/2005.00633.
14. Matteo Negri and Milen Kouylekov. 2009. Question answering over structured data: an entailment-based approach to question analysis. In Proceedings of the International Conference RANLP-2009, pages 305– 311, Borovets, Bulgaria. Association for Computational Linguistics.
15. Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proc. of NAACL.
16. Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996– 5001, Florence, Italy. Association for Computational Linguistics.
17. A. Radford. 2018. Improving language understanding by generative pre-training.
18. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need
19. Ivan Vulic and Marie-Francine Moens. 2015. ´ Bilingual word embeddings from non-parallel documentaligned data applied to bilingual lexicon induction
20. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 719–725, Beijing, China. Association for Computational Linguistics.
21. Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.
22. Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual bert to lowresource languages.
23. Zirui Wang*, Jiateng Xie*, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In International Conference on Learning Representations.
24. Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122. Association for Computational Linguistics.
25. Wenpeng Yin, Jamaal Hay, and D. Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. ArXiv, abs/1909.001"