Test Time Augmentation for Cross-lingual Text Classification

Artsem Zhyvalkouski
Tokyo City University
artsem20@ipl.cs.tcu.ac.jp

Shiho Hoshi Nobesawa
Tokyo City University
shiho@tcu.ac.jp

Abstract

Recently, in order to face the problem of the lack of labeled data for languages other than English, language models based on BERT were extended to tackle the task of cross-lingual transfer. Such multilingual models have been known to perform well on the target language data without actually being trained on it, but the need to improve their accuracy still exists. In this paper, we present an approach to improve the accuracy of such models by utilizing machine translation and test time augmentation, widely used in computer vision, but not thoroughly researched in natural language processing. We show that our method demonstrates improvement without training additional models or collecting more labeled data.

1 Introduction

1.1 Multilingual Language Models

Recently, a highly accurate masked language model BERT[1] is widely used. It is pre-trained with a large amount of unlabeled data, such as Wikipedia[2]. Training is performed by randomly removing a word from a sentence and making the model predict what the removed word is. It can be said that this is a method of extracting features of sentences, such as word2vec[3] or FastText[4]. As an application of this model, firstly it is trained on a large amount of unlabeled data in advance. Next, it is trained further with data for tasks such as text classification or question answering.

Another recent trend is to train multilingual models that extend BERT to multiple languages. Typical models include mBERT[1], XLM[5] and XLM-R[6]. The background of multilingual models is that there is currently a lack of non-English labeled data. Unlike BERT, mBERT is pre-trained with a large amount of data in 104 languages. The application process is illustrated on Figure 1. As a concrete application, firstly it is trained with a large amount of unlabeled multilingual data, then it is trained with labeled English data for a particular task. And finally, the inference is performed in the target language, for example, French. By doing so, it is known that high accuracy can be achieved even in the target language without actually having any labeled data for the target language. As another example, we can take a multilingual pre-trained model, fine-tune it on English data for Question Answering and then perform inference in Japanese, so the model can answer questions in Japanese, without actually being trained on them. One of the purposes of this research is to improve the accuracy of such multilingual models.

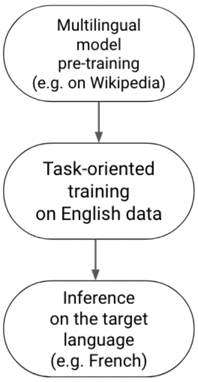
Multilingual model pre-training (e.g. on Wikipedia)
Task-oriented training on English data
Inference on the target language (e.g. French)

Figure 1 Application of multilingual models

1.2 Test Time Augmentation

Another concept that is used in this research is Test Time Augmentation. Let us start with Train Time Augmentation. This is a method to increase the training data in an artificial manner. For example, in the case of natural language processing, there is back-translation[7]. An example of the back-translation method is shown on Figure 2. With this method machine translation from the original language A to another language B is performed. Then, once again machine translation from language B to the original language A is performed. For instance, you can translate En-

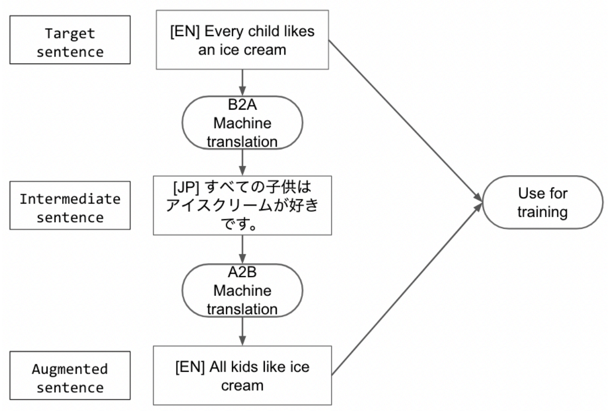

Figure 2 Back-translation algorithm[7]

glish into Japanese and then Japanese into English again to create different sentences without changing their meaning. Xie et al. insisted that the accuracy can be significantly improved by using this method[7].

Unlike Train Time Augmentation, Test Time Augmentation is performed for models that have already been trained. As Shorten et al. and Buslaev et al. described this is a commonly used technique in image processing to improve accuracy [8, 9]. An example of the Test Time Augmentation technique for images is shown on Figure 3. For example, firstly an image to be predicted is mirrored or rotated. Then the inference is performed for the resulting images separately, and finally, the average is taken to summarize the predicted values.

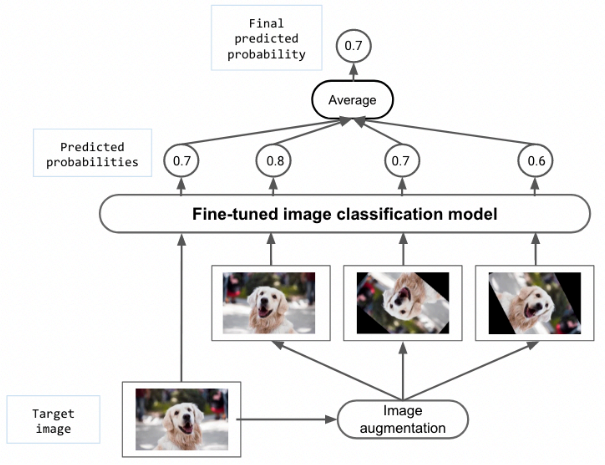

Figure 3 Test Time Augmentation in image processing

In the field of image processing, this method is also known to improve accuracy[9]. However, the situation is not so much considered in the field of natural language processing. We can see its investigation in the research by Howard et al.[10] where they utilize back-translation in test time for classification with the ULMFiT model[11], which is based on LSTM[12]. Liu mentioned an empirical result where name swapping augmentation was used in test time to improve accuracy on the Gender Pronoun Resolution task[13]. However, there seems to be very little research done on using Test Time Augmentation with multilingual data or on cross-lingual tasks. As another purpose of this research, we also aim to draw more attention to Test Time Augmentation in natural language processing.

2 Test Time Augmentation for Cross-lingual Text Classification

2.1 Overview

The flow of the proposed method is presented on Figure 4. Since multilingual models can handle various languages, we believe that data augmentation is possible by machine translation. First, we translate the text in the target language into texts in various other languages. Then we make separate predictions for all the texts and finally put together all the predictions. As a way of summarizing, we are going to use averaging and weighted averaging. For weighted averaging, we utilize the Nelder-Mead optimization method[14] to find the best weights for the validation dataset and then use the obtained weights on the test dataset. We suppose that simple averaging may not be enough since we cannot be sure about the most appropriate augmentation language.

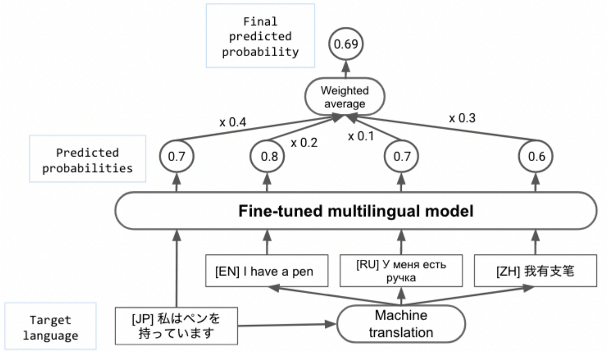

Figure 4 Proposed method flow

2.2 Algorithm

Here we explain the proposed algorithm step by step.

1. Translate the text T_L in the target language L with the machine translation model $M^{trans}_{A_i}$ into the text T_{A_i} in the additional language A_i. N is a number of

additional languages.

$$T_{A_i} = M_{A_i}^{trans}(T_L),\ 0 \le i < N$$

2. Obtain the predicted probability vectors p_L and p_{A_i} for the texts T_L and T_{A_i} respectively by using the fine-tuned multilingual model M^{multi} for text classification.

$$p_L = M^{multi}(T_L),\ p_{A_i} = M^{multi}(T_{A_i}),\ 0 \le i < N$$

3. Obtain the final predicted probability vector p_w by performing weighted averaging using the weights vector w optimized for the validation dataset using the Nelder-Mead method[14].

$$p_w = w_N p_L + \Sigma_{j=0}^{N-1} w_j p_{A_j}$$

4. Predict the final class C by using the highest probability in the vector p_w.

$$C = argmax\ p_w$$

As a result, we get the final predicted class C, which is in our case is one of the contradiction, neutral or entailment classes.

3 Dataset

We use the XNLI dataset[15]. A few samples from the dataset are presented in Table 1.

Table 1 XNLI dataset examples

Language	**Premise - Hypothesis**	**Label**
English	Fun for adults and children. Fun for only children.	Contradiction
English	He turned and smiled at Vrenna. He smiled at Vrenna who was walking slowly behind him with her mother.	Neutral
English	Postal Service were to reduce delivery frequency. The postal service could deliver less frequently.	Entailment
Swahili	Bosi wangu alikua mcheshi na thabiti. Bwana, ni kama ako nafsi mbili tofauti.	Contradiction
German	Wir hatten ein tolles Gespräch. Nun, daran dachte ich nicht einmal, aber ich war so frustriert, dass ich am Ende doch mit ihm redete.	Neutral
French	J’avais l’impression que j’étais le seul à avoir ce numéro dans le domaine de carrière de l’AFFC Air Force. Et je pensais que c’était un privilège, et ça l’est toujours, toujours, j’étais le seul 922 Ex-O, c’est là que j’avais effectué ma carrière dans l’AFFC Air Force.	Entailment

It has classification data for 15 languages: English, French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu. There are about 5,000 test and 2,500 validation pairs for each language, so it results in 112,500 annotated pairs. The pairs are obtained by human translation from English. The task is called textual entailment, which is a 3 class classification of two sentences called premise and hypothesis. Classes include contradiction, neutral and entailment. We use this dataset since it is a standard for evaluating Cross-lingual Text Classification used in the state-of-the-art research [5, 6].

4 Results

We perform two different experiments using a pre-trained mBERT model, which also trained on the English XNLI data, thus evaluating the model in a cross-lingual transfer manner.

4.1 Experiment 1: Evaluation on a Single Additional Language

In this setting, we use the original English translations released in the XNLI dataset and simply average predictions in the original language with predictions using the English translations. The results for each language are shown on Figure 5.

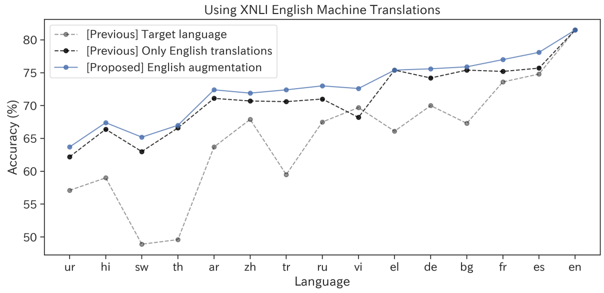

Figure 5 Experiment 1: Comparison with previous methods and our method using evaluation on a single additional language

We compare our method with previous methods, such as using the target language and using English translations[6].

At first, we can notice that the performance on English translations is higher for most of the languages, except for Vietnamese. This can be explained by the following facts: the model was fine-tuned on the English samples; the largest amount of the pre-training data is in English. As for Vietnamese, it is hard to draw any conclusions, but we suppose it may relate to the quality of the given English translations. We can see that by using the proposed method and aggregating both predictions the accuracy for all of the languages either improves or is equal to the best previous method. This indicates that both the original texts and English translations provide a meaningful signal to the model. From the results, we may suppose that in a setting when the target language training data is not available, it is recommended to utilize English translations with our method to improve the accuracy without collecting more data or fine-tuning additional models.

4.2 Experiment 2: Evaluation on Multipule Additional Languages

In this setting we use the Marian MT models[16] to translate the original language into not only into English, but also into *close* and *far* languages. We use our own definition of *close* and *far* languages. The results for each language are shown on Figure 6. As a *close* language we use the language from the same group which has the highest number of articles in Wikipedia. For a *close* language we use the language from a different family which has the highest number of articles in Wikipedia. The reasoning behind choosing *close* and *far* languages is to account for the trade-off between translation quality and signal diversity. A *close* language can be easier to translate, so the data quality is higher, but it will provide presumably less diversity since the syntax structure will be similar. Vice versa, a *far* language might be more difficult to translate, although the syntax structure could be different and the translation might have more diverse word usage with similar semantics. For both languages, we use the one with the highest number of articles on Wikipedia because the model was pre-trained on it so we can obtain more meaningful representations. It is needed to mention that this reasoning is a pure assumption and in the perfect scenario the most useful augmentation language must be found using tuning on the validation set. As in Experiment 1, we compare our method with previous methods: using the target language and using English translations. For the proposed methods we use two settings: English augmentation and augmentation with English, *close* and *far* languages.

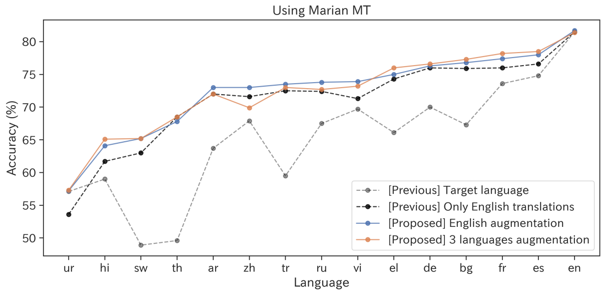

Figure 6 Experiment 2: Comparison with previous methods and our method using evaluation on multiple additional languages

From the results, we can see that for each language one of the proposed settings improves the accuracy or performs on par. The 3-language setting improves the accuracy or performs on par for all of the languages except for Mandarin, for which the English augmentation setting performs best. This may be explained by the quality of translations from Mandarin to other languages. The fact that the highest accuracy improvement is observed for low-resource languages such as Urdu, Hindi and Swahili is presumably due to the quality of the model's representations for their translations. We also conclude that choosing the most suitable augmentation language for each target language may be an appropriate step during the model tuning.

5 Discussion

We can notice that for each language the proposed method either outperforms or performs on par with the previous methods. We also can see that by using the proposed technique the accuracy on the XNLI dataset can be improved by around 2% for each language on average. Especially for low-resource Hindi, Urdu and Swahili, the performance is increased significantly (4–7%).

It is needed to mention that in this research we proposed only using one-step machine translation as augmentation, whereas back-translation or any other text augmentation technique can be utilized. We also observe that for each language the best additional language and its performance seem to be different so that may be a topic for future research. Moreover, in this research we examined the natural language task of cross-lingual text classification, a similar Test Time Augmentation method might be applied to various other tasks such as question answering or named-entity recognition. The model we used is mBERT which is smaller than XLM-R or recent XLM-R XL[17], which outperforms even priorly best monolingual RoBERTa[18] on English. This implies that larger and more accurate models can also be investigated with our method.

6 Conclusion

Due to the lack of labeled data for other languages than English, multilingual models are being widely used recently. Although they are already achieving relatively high accuracy, we propose a method that can further improve the accuracy of such models by machine translation and test time augmentation. We show that our method improves the accuracy by 2-7% on the XNLI dataset which is a standard for evaluating multilingual models. As for the practical usage of our method, the accuracy of cross-lingual classification models can be improved without additionally collecting more data or training new models.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

[2] Wikipedia contributors. Plagiarism — Wikipedia, the free encyclopedia, 2004. [Online; accessed 22-July-2004].

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. Enriching word vectors with subword information. **CoRR**, Vol. abs/1607.04606, , 2016.

[5] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. **CoRR**, Vol. abs/1901.07291, , 2019.

[6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[7] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation, 2019. cite arxiv:1904.12848.

[8] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. **Information**, Vol. 11, No. 2, 2020.

[9] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. **Journal of Big Data**, Vol. 6, pp. 1–48, 2019.

[10] Sam Shleifer. Low resource text classification with ulmfit and backtranslation. **CoRR**, Vol. abs/1903.09244, , 2019.

[11] Jeremy Howard and Sebastian Ruder. Fine-tuned language models for text classification. **CoRR**, Vol. abs/1801.06146, , 2018.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. **Neural computation**, Vol. 9, pp. 1735–80, 12 1997.

[13] Bo Liu. Anonymized BERT: an augmentation approach to the gendered pronoun resolution challenge. **CoRR**, Vol. abs/1905.01780, , 2019.

[14] John Ashworth Nelder and Roger Mead. A Simplex Method for Function Minimization. **The Computer Journal**, Vol. 7, No. 4, pp. 308–313, 01 1965.

[15] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations, 2018.

[16] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In **Proceedings of ACL 2018, System Demonstrations**, pp. 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[17] Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. Larger-scale transformers for multilingual masked language modeling. **CoRR**, Vol. abs/2105.00572, , 2021.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. **CoRR**, Vol. abs/1907.11692, , 2019.