Enhancing Natural Language Inference of Cross-lingual N-shot Transfer with Multilingual Data

Kuang Tseng, Chow-Sing Lin

Department of Computer Science and Information Engineering, National University of Tainan, Tainan, Taiwan
E-mail Address: dsif2012@gmail.com, mikelin@mail.nutn.edu.tw

2022 8th International Conference on Applied System Innovation (ICASI) | 978-1-6654-9650-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASI55125.2022.9774470

Abstract

Cross-lingual N-shot transfers are often used to solve low-resource language problems. However, the result of transfer in different languages can be inconsistent due to large fluctuations in accuracy. To reduce the inconsistency caused by the fluctuation of accuracy and improve the overall accuracy of Cross-lingual N-shot transfer, we propose the use of Multitasking Cross-lingual N-shot transfer. To the best of our knowledge, our propose Multitasking Cross-lingual N-shot transfer is the first method to combine Multitasking Learning and Cross-lingual N-shot transfer. By training with multiple languages simultaneously, the model can learn multilingual common semantic features from different languages, decreasing the fluctuations in accuracy, and also having a higher overall accuracy compared to traditional method with single high-resource language. The method presented in this paper can effectively reduce 93.2% of the accuracy fluctuations and improve the overall accuracy by 10.4% in NLI.

Key words: Natural Language Processing, low-resource languages, Cross-lingual N-shot transfer and Natural Language Inference

Introduction

There are more than 7,000 languages in the world, yet 94% of the population only speaks 4%, which is about 280, of these languages [1]. Among these, there are only about 20 high-resource languages with a large variety and number of datasets. In fact, most people are[1] still using low-resource languages. The number and types of datasets in low-resource languages are insufficient, resulting in poor performance in training Natural Language Processing (NLP) models. It is time-consuming and expensive to collect and label data from low-resource languages. Yet models trained in high-resource languages can not be easily applied to other languages, keeping low-resource language users from utilizing NLP techniques to their full potential [2].

To improve the poor performance of low-resource language NLP tasks, many studies have proposed various transfer learning techniques, such as Cross-lingual N-shot transfer [3]. Cross-lingual N-shot transfer is based on a multilingual pre-trained model which trains multilingual features through a large amount of data. Generally there are two types of Cross-lingual N-shot transfer, namely zero-shot transfer and few-shot transfer. Transferring without target domain data is called zero-shot transfer while retraining the model with a small input of target domain data is known as few-shot transfer. By fine-tuning the model with high-resource language and then converting it to low-resource languages model, Cross-lingual N-shot transfer can effectively improve the accuracy of various tasks in different languages. However there are shortcomings in this method whose accuracy is dramatically decreased when converting to certain languages. Such an instability is caused by structural differences in languages.

Most of the existing studies use only English as the language for Cross-lingual N-shot transfer [4]. We argue that for many low-resource languages, there are more suitable high-resource languages other than English. For example, Chinese and Japanese, both using Kanji character, may be more suitable for transferring to each other. Thus in this paper, we compare the Cross-lingual N-shot transfer accuracy of English with those of other languages.

To study the above-mentioned issues, we select Natural Language Inference (NLI) as the target. The standard NLI dataset provides premise and hypothesis, and the machine can successfully learn the relationship (Entailment, contradiction, or neutrality) between premise and hypothesis through a large amount of data. The main task of NLI is to understand the meaning of the sentence. The model is able to learn rich common semantic features by finding the connection between contexts. The previous research work reveals that NLI has good generality for most NLP tasks such as Question Answering, Summarization, Chatting, etc. These NLP tasks can be simplified into contextual relationships which can be understood by NLI. Therefore, We can expect the accuracy decrease of NLI is similar to that of most other NLP tasks on Cross-lingual N-shot transfer. Also, the performance of NLI can be easily measured by accuracy. Furthermore, a good NLI model can even be applied to more complex NLP tasks to improve their accuracy [5].

In this paper, we use mT5-large [6] as the pre-train model for cross-lingual N-shot transfer. The mT5-large then is fine-tuned with 15 different languages in Cross-lingual Natural Language Inference Corpus (XNLI) [7], and finally is transferred to 14 other languages. Through this experiment, we find that English is not the most suitable language for Cross-lingual N-shot transfer in our experimental settings. To address these issues, we propose the use of Multitasking Cross-lingual N-shot transfer to obtain better accuracy than Cross-lingual N-shot transfer with only using English. In our experimental results show that providing more data in different languages can significantly improve the overall accuracy of Cross-lingual N-shot transfer.

The rest of this paper is organized as follows. The related work section provides background and issues on cross-lingual n-shot transfer itself. The architecture of the proposed method is described in the Multitasking cross-lingual n-shot transfer section. The experiments section introduces and analyzes experiments. Finally, in the conclusion part, we will summarize the conclusions of this article.

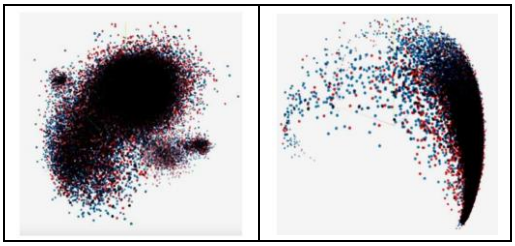
Authorized licensed use limited to: Institut Teknologi Bandung. Downloaded on February 10,2023 at 09:46:48 UTC from IEEE Xplore. Restrictions apply.

Related Work

Many scholars have studied Cross-lingual N-shot transfer since it was proposed [8]. In their work, they find that Cross-lingual N-shot transfer using solely English has the problem of accuracy degradation on transferring to other languages in all of the tasks. They use mBERT and XLM-R two different multilingual pre-traind model with five NLP tasks. On three NLP tasks, Dependency Parsing (DEP), Part-of-Speech Tagging (POS), and Named Entity Recognition (NER), the average accuracy of Cross-lingual zero-shot transfer to *ja* drops more than 50% by solely English training. In XNLI, while using mBERT or XLM-R as a pre-trained model, the accuracy on *sw* also drops by more than 20%. In XQuAD, the accuracy on *zh* also drops by more than 25%. From this experiment, it is shown that when English is use to transfer to low-resource languages, the transferring accuracy varies significantly in different languages.

Although English is considered as the most resourceful high-resource language and is most frequently used for cross-language N-shot transfer, this default choice has not been systematically vetted [9]. In their work, by using machine translation (MT) from *en* to other languages datasets to fine-tune the model, Cross-lingual N-shot transfer to other languages, and then testing with human-translated datasets (HT), it is shown that the transfer accuracy with most other languages such as *bg*, *zh*, *ru* is better than with English. Also, the accuracy fluctuation of cross-language N-shot transfer may have its own unique pattern, which is not likely to be affected by machine translation. They found that German and Russian were more effective at doing cross-language N-shot transfer, especially when the target language is diverse or unknown even if the training set is automatically translated from English.

To more intuitively confirm the impact of using different languages on fine-tuning of pre-trained models, it is shown in Fig.1 [10]. Fig.1(a) shows the word embedding of the model without fine-tuning projected to 2D space through Principal Component Analysis (PCA). Fig.1(b) shows the word embedding of the model after being fine-tuned in English and different colors dots represent different languages. According to the difference between Fig.1 (a) and Fig.1 (b), it is shown that even fine-tuning with a single language can inevitably affect all languages in pre-trained model. To further explore possible impacts of words or grammars, the authors substitute some of the English words when training the QA system. They find the model's EM and F1 evaluation metrics drop no matter what language replaces the original words or what the substitution rate is. Even if the positions of subject, verb, and object change, the effect on EM and F1 is always small. Although we have known that Cross-lingual N-shot transfer to the target language has its unique fluctuation pattern of accuracy, currently it is still impossible to utilize this regulation to further improve the accuracy of target languages in practice.



(a) (b)

Fig. 1 Model word-embedding before and after fine-tuning.

Multitasking Cross-lingual N-shot Transfer

From previous research, we find that fine-tuning the model even with only one language can affect all languages in the pre-trained model. Also English is probably not the most suitable language for Cross-lingual N-shot transfer in most of the time. Although Cross-lingual N-shot transfer has its unique fluctuation pattern of accuracy, the pattern cannot be clearly known due to insufficient existing experiments. In this paper, we propose multitasking cross-lingual n-shot transfer to solve the instability of accuracy fluctuation caused by the large difference of accuracy in order to further improve the overall accuracy.

Compared to traditional cross-lingual N-shot transfer, we fine-tune our model to become the multitasking model after simultaneously pre-processing the data with multiple languages. Multitasking Learning is to make the model learn all the different tasks simultaneously by providing a variety of different datasets to the model at the same time. However, this technique may lead to conflict in some tasks, and there is a problem of the accuracy drop compared to training with only a single task. Traditional multitasking learning use different tasks for training, but we only use the same task for different language training, we believe the accuracy drop should be tolerable. After training the multilingual model, we perform cross-lingual n-shot transfer to other low-resource languages by providing multilingual semantic features.

Experiments

We used XNLI for our experiments which is based on the The Stanford Natural Language Inference(SNLI) dataset [11], containing 5000 validation sets and 2500 testing sets. These sets are manually translated into 14 languages, including French, Spanish, German, etc. However, we do not use the method recommended in the study [7] which used the SNLI dataset to train the model, and transfer to target language by cross-lingual N-shot transfer. To effectively control external variables, we only use XNLI. Because all the data in XNLI are manually translated from English, the length of data are aligned and the contents are the same. In addition, transfer learning such as Cross-lingual N-shot transfer is often mainly used for training low-resource languages. For many low-resource languages, 7500 sets of data may not be easily found on the public network. To utilize the XNLI dataset specifically for training, we divided all 7500 sets of data in XNLI into 5000 training sets, 500 validation sets, and 2000 testing sets.

ISBN: 978-1-6654-9650-6

Authorized licensed use limited to: Institut Teknologi Bandung. Downloaded on February 10,2023 at 09:46:48 UTC from IEEE Xplore. Restrictions apply.

We selected mT5-large for pre-train model in our experiments. mT5 has a total of 101 different languages, ranging from high-resource languages to low-resource languages. mT5 can become a multilingual pre-trained model by being pre-trained with mC4 dataset. The mT5 also has a variety of specifications for selecting different capability, such as mT5-XXL. Although the mT5-XXL has the highest accuracy, we decided to use mT5-large in our experiments since it is the best balance between performance and computing cost [6]. In this research work, mT5-large is also used for the training baseline of the XNLI dataset. Therefore, we follow the recommended practice and use mT5-large as the pre-trained model.

Table 1 shows Cross-lingual N-shot transfer to all language average accuracy in XNLI. As shown in Table 1, the average accuracy of English is only at the 8th, suggesting English is not the most suitable language in this case. The comparison between our experiments using HT and the previous study using MT, English average accuracy drops from the 8th to the 13th [9]. We believe that because the translation already has the features of English combined with the machine translated target language, the translated language can have a better performance than English alone in most cases. According to this result, providing more language features during Cross-lingual N-shot transfer may effectively improve the overall accuracy.

TABLE I
CROSS-LINGUAL N-SHOT TRANSFER TO ALL LANGUAGE AVERAGE ACCURACY IN XNLI

Lan	avg	No	*En* mt	No
ar	47.0%	11	0.5%	8
bg	46.6%	12	3.0%	1
de	50.3%	7	2.2%	4
el	48.7%	10	1.7%	5
en	50.2%	8	0.0%	11(13)
es	45.1%	14	1.4%	6
fr	50.7%	6	-0.3%	12(14)
hi	49.8%	9	0.2%	9(11)
ru	51.3%	5	2.5%	3
sw	46.4%	13	0.5%	8
th	52.0%	2	0.5%	8
tr	52.7%	1	1.2%	7
ur	51.8%	3	-0.7%	13(15)
vi	42.9%	15	0.1%	10(12)
zh	51.6%	4	2.8%	2

Table 2 shows Multitasking Cross-lingual N-shot transfer using 15 languages in XNLI. We fine-tuned the pre-trained model using 15 languages in XNLI before training it into a multitasking model. Compared with models fine-tuned with only a single language, training the model with multiple languages does not result in any significant drop of accuracy. Because of the increase in the amount of data and language semantic features, the overall average accuracy can be even increased by 10% .

TABLE II
MULTITASKING MODEL CROSS-LINGUAL N-SHOT TRANSFER TO 15 LANGUAGES IN XNLI

Lan	ours	original	accuracy	increase(%)
ar	58.9%	50.2%	8.7%	17.3%
bg	57.7%	48.0%	9.6%	20.1%
de	56.3%	50.9%	5.4%	10.7%
el	57.2%	51.3%	5.9%	11.4%
en	60.6%	55.2%	5.3%	9.6%
es	57.6%	49.6%	8.0%	16.0%
fr	56.2%	51.4%	4.8%	9.3%
hi	54.8%	51.5%	3.4%	6.5%
ru	57.4%	51.9%	5.6%	10.7%
sw	53.6%	50.3%	3.3%	6.5%
th	58.4%	54.4%	4.0%	7.4%
tr	56.3%	53.1%	3.2%	6.0%
ur	54.0%	51.1%	3.0%	5.8%
vi	55.3%	48.9%	6.4%	13.1%
zh	64.4%	61.0%	3.4%	5.5%
avg	57.2%	51.9%	5.3%	10.4%

Table 3 shows Multitasking Cross-lingual N-shot transfer accuracy. As shown in Table 3, by using 14 languages simultaneously for Multitasking Cross-lingual N-shot transfer, the average accuracy in low-resource languages is increased by 10.4\% on average. It shows that give model more language features can improve overall accuracy. Other than that, table 4 shows Comparison of average accuracy drop rate. As shown in Table 4, our average decreased of accuracy is much better than *en*, *de*, and *ru* by 93.2%, 66.7%, and 55.2%, respectively.

TABLE III
MULTITASKING CROSS-LINGUAL N-SHOT TRANSFER TO LOW-RESOURCE LANGUAGES IN XNLI

zero-shot	original	ours	accuracy	increase(%)
ar	54.0%	60.0%	6.1%	11.2%
bg	52.3%	60.0%	7.7%	14.6%
el	51.8%	57.2%	5.4%	10.3%
hi	50.6%	56.0%	5.4%	10.7%
sw	46.2%	51.0%	4.7%	10.3%
th	52.8%	58.6%	5.8%	10.9%
tr	52.1%	56.2%	4.1%	7.9%
ur	51.8%	55.5%	3.7%	7.1%
avg	51.4%	56.8%	5.3%	10.4%

TABLE IV
COMPARISON OF AVERAGE ACCURACY DROP RATE

	en	de	ru	ours
ar	-5.6%	2.5%	6.9%	5.0%
bg	-8.5%	1.5%	-1.6%	4.9%
el	-9.4%	0.8%	-1.3%	0.0%
hi	-11.5%	-1.9%	-3.4%	-2.0%
sw	-19.2%	-11.1%	-3.3%	-10.8%
th	-3.8%	-0.9%	-0.9%	2.4%
tr	-8.8%	-3.2%	-1.2%	-1.7%
ur	-9.4%	-3.3%	-6.7%	-3.0%
avg	-9.5%	-1.9%	-1.4%	-0.6%

ISBN: 978-1-6654-9650-6

Authorized licensed use limited to: Institut Teknologi Bandung. Downloaded on February 10,2023 at 09:46:48 UTC from IEEE Xplore. Restrictions apply.

Conclusion

In this paper, we observe that by giving the model more multilingual features at the same time, it can effectively improve the model's accuracy. To address the instability caused by the large difference in the accuracy to further improve the overall accuracy, we propose using Multitasking Cross-lingual N-shot transfer in low-resource languages. The experimental results demonstrate that our average accuracy on Multitasking Cross-lingual N-shot transfer can be significantly increased, compared to the traditional use of only *en*.

References

[1] A. Magueresse, V. Carles, and E. Heetderks, "Low-resource Languages: A Review of Past Work and Future Challenges," *ArXiv200607264 Cs*, Jun. 2020.

[2] F. Nooralahzadeh, G. Bekoulis, J. Bjerva, and I. Augenstein, "Zero-Shot Cross-Lingual Transfer with Meta Learning," *ArXiv200302739 Cs*, Oct. 2020.

[3] K.-H. Huang, W. U. Ahmad, N. Peng, and K.-W. Chang, "Improving Zero-Shot Cross-Lingual Transfer Learning via Robust Training," *ArXiv210408645 Cs*, Sep. 2021.

[4] J. Phang *et al.*, "English Intermediate-Task Training Improves Zero-Shot Cross-Lingual Transfer Too," *ArXiv200513013 Cs*, Sep. 2020.

[5] S. Bowman and X. Zhu, "Deep Learning for Natural Language Inference," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, Minneapolis, Minnesota, Jun. 2019, pp. 6–8. doi: 10.18653/v1/N19-5002.

[6] L. Xue *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer," *ArXiv201011934 Cs*, Mar. 2021.

[7] A. Conneau *et al.*, "XNLI: Evaluating Cross-lingual Sentence Representations," *ArXiv180905053 Cs*, Sep. 2018.

[8] A. Lauscher, V. Ravishankar, I. Vulić, and G. Glavaš, "From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers," *ArXiv200500633 Cs*, May 2020.

[9] I. Turc, K. Lee, J. Eisenstein, M.-W. Chang, and K. Toutanova, "Revisiting the Primacy of English in Zero-shot Cross-lingual Transfer," *ArXiv210616171 Cs*, Jun. 2021.

[10] T. Hsu, C. Liu, and H. Lee, "Zero-shot Reading Comprehension by Cross-lingual Transfer Learning with Multi-lingual Language Representation Model," *ArXiv190909587 Cs Stat*, Sep. 2019.

[11] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *ArXiv150805326 Cs*, Aug. 2015.

ISBN: 978-1-6654-9650-6

Authorized licensed use limited to: Institut Teknologi Bandung. Downloaded on February 10,2023 at 09:46:48 UTC from IEEE Xplore. Restrictions apply.