Multilingual Symptom Prediction by Simultaneous Learning using BERT

Chencheng Zhu
Department of Artificial Intelligence
Kyushu Institute of Technology
Fukuoka, Japan
zhu.chencheng822@mail.kyutech.jp

Niraj Pahari
Department of Artificial Intelligence
Kyushu Institute of Technology
Fukuoka, Japan
nirajpahari@gmail.com

Kazutaka Shimada
Department of Artificial Intelligence
Kyushu Institute of Technology
Fukuoka, Japan
shimada@ai.kyutech.ac.jp

***Abstract*—Natural language processing (NLP) techniques have become increasingly significant in the medical domain. However, the amount of relevant medical text data remains limited. In this work, we propose a BERT-based multilingual simultaneous learning (MSL) model for reducing the problem of scarce data. We evaluate the benefit of MSL on the NTCIR-13 MedWeb multi-label symptoms classification task. The results indicate that the MSL model performs slightly better than Single-Task Learning (STL) models. Additionally, it shows that the similarity between languages has an impact on the performance of the MSL model.**

***Index Terms*—multi-task learning, cross-lingual classification, social media**

I. Introduction

The advancement of digitalization in healthcare has facilitated the promotion of EHRs, which leads to a significant increase in textual data. There is a growing interest in natural language processing (NLP) techniques within the medical domain [1]. However, the labeled data available for research is still limited due to the domain specificity. Since medical experts are required to annotate the corpus, time and labor costs are substantial. Moreover, numerous datasets are not publicly available due to the sensitive nature of patient privacy concerns [2]. This circumstance is more prevalent in non-English data, and progress in corresponding research has been sluggish. The imbalance in language resources can be remedied by utilizing techniques suitable for multilingual environments.

Multi-task learning (MTL) has demonstrated its effectiveness in various machine learning applications [3]. For a variety of NLP problems, it has been applied to handle the limited labeled datasets and to learn the common representations between the relative tasks [4], [5].

In this work, we are inspired by MTL, utilizing its framework to learn the multilingual corpus consisting of Japanese, Chinese, and English simultaneously. In contrast to conventional MTL, the tasks handled in this work are different only at the sentence level. In particular, the input sentences are in three different languages, while the output labels for each sentence in each language are the same. We name this learning process Multilingual Simultaneous Learning (MSL).

Our contributions are as follows:

- We propose a BERT-based model for simultaneous learning on the multilingual corpus.
- We demonstrate the effectiveness of MSL on a multilingual symptom classification task over single-task learning (STL).
- Additional experiments applying the bilingual simultaneous learning (BSL) model are conducted to examine the effect of language similarity.
- A comprehensive error analysis is provided to understand the limitations of the MSL model.

II. Related work

As mentioned above, our method, MSL, is inspired by MTL. Therefore, we provide the explanation of MTL first. MTL is known as a form of transfer learning. It utilizes more valuable information from multiple related tasks. Thus it empowers these tasks to attain superior performance than STL. Besides, the generalization performance of the model can be improved as well [6].

In the context of healthcare, MTL has been applied to a variety of tasks in NLP. Joshi et al. [7] have proposed an MTL model based on BiLSTM to perform three health informatics prediction tasks on tweets. The experiment demonstrated the benefit of MTL in comparison with STL. Hartmann et al. [8] have proposed a multilingual MTL model with hard parameter sharing. The model combined three tasks: negation scope resolution in clinical text, negation scope resolution in product reviews, and detection of negated events. Furthermore, they examined the cross-lingual transfer ability of the model. The results showed that zero-shot scope resolution in the clinical text is possible.

III. Dataset

We use the NTCIR-13 Medical Natural Language Processing for Web Document (MedWeb) dataset [9] in this work. MedWeb provides manually created pseudo-Twitter messages, covering three languages: Japanese, English, and Chinese. Both English and Chinese corpora are translated from the original Japanese messages. Each sentence is annotated by eight labels: Influenza, Diarrhea, Hay fever, Cough, Headache, Fever, Runny nose, and Cold. Table I shows examples of pseudo-tweets for each language. A positive (p) or negative (n) status is given to each symptom/disease label. Since a single

979-8-3503-3078-6/23/$31.00

TABLE I
EXAMPLES OF PSEUDO-TWEETS FOR EACH LANGUAGE.

Lang	Pseudo-tweets	Flu	Diarrhea	Hay fever	Cough	Headache	Fever	Runny nose	Cold
en	I have a fever but I don't think it's the kind of cold that will make it to my stomach.	n	n	n	n	n	p	n	p
ja	熱は出てるけどお腹に来る風邪じゃなさそう.								
zh	虽然发烧，但是好像不是肚子着凉的感冒。								

message could contain multiple symptoms, the positive status may be given to multiple labels.

IV. METHOD

In this study, BERT is utilized as the base model. We construct the STL models and the multilingual simultaneous learning (MSL) model to perform the multi-label classification task.

The STL Models: STL is employed as the baseline. We apply different BERT variants to perform the classification tasks respectively for each language. The utilized variations of BERT will be explicated in Section V.

The MSL Model: We construct the MSL model with a soft parameter sharing method. This study utilizes BERT as the base model. We divide the 12 Transformer layers into the bottom, middle, and top layers. Generally, the bottom layers of Transformers have the most information about linear word order [10]. The middle layers are most predictive of dependencies [11] and are the most transferable across tasks [12]. The top layers learn the task-specific features [13].

Figure 1 shows the overview of our MSL model. Freeze layers are the layers that will not update the weights during fine-tuning. Share layers share the parameters between the three models. The weights are updated depending on the three tasks. Individual layers learn the task-specific features of each language dataset. The weights are updated according to the corresponding language's own classification task. The F-S-I (Freeze - Share - Individual) combination has been shown to be the most effective in a study [14]. Therefore, we apply this setting is applied in this study as well: the bottom layers of BERT as the freeze layers, the middle layers as the share layers, and the top layer as the individual layers. A linear layer is added above the 12 Transformer layers to perform the multi-label classification for each language.

V. EXPERIMENTAL SETTINGS

The following describes the models and parameters utilized in this experiment.

a) BERT variants: For each language, we utilize both the multilingual pre-trained models and the language-specific monolingual BERT. For the multilingual pre-trained models, we employ mBERT [15] and LaBSE [16]. For the language-specific monolingual BERT, we employ *bert-base-uncased* [1] for English, *cl-tohoku/bert-base-japanesewhole-word-masking* [2] for Japanese, and *bert-base-chinese* [3] for Chinese.

[1]https://huggingface.co/bert-base-uncased
[2]https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking
[3]https://huggingface.co/bert-base-chinese

b) The STL Models: We conduct the experiments for each language with three different STL models: $Single_{mBERT}$, $Single_{LaBSE}$, $Single_{mono}$. The subscript "mBERT", "LaBSE", and "mono" represent the BERT variants used for each dataset.

c) The MSL Models: As shown in Figure 1, three MSL models are constructed: MSL_{mBERT}, MSL_{LaBSE}, MSL_{mono}. For each MSL model, we perform the tests with various numbers of layers of F-S-I to figure out the best combination. The initial combination is 4-4-4 (Freeze layers: L1-4, Share layers: L5-8, Individual layers: L9-12). Using the knowledge from a paper [14], the combination 1-7-4 and 1-4-7 are the most effective ones. Therefore, we examine three combinations in total in the experiments.

The AdamW optimizer and binary cross-entropy were used as the loss function. The learning rate was set as 5e-5 for transformer layers and 5e-3 for linear layers. We set the number of training epochs to 10. To prevent overfitting, we applied the EarlyStopping mechanism. For each language, 1920 sentences were used for training and 640 for testing. The performance was evaluated based on exact match accuracy.

VI. RESULTS

1) The suitable BERT variant and layer combination: Table II shows the exact match accuracy of the three MSL models with different F-S-I layer combinations. The first column denotes the size of the freeze, share, and individual layers we set. The boldface values are the highest accuracy for each language. The values with "*" denote the highest accuracy for each language in each model. The MSL model using LaBSE provided high results on the average of the three languages. In terms of layer combination, the 1-7-4 setting showed the best overall performance. It is apparent from Table 2 that 1-7-4 attained the highest results for most languages with any BERT variants.

TABLE II
THE ACCURACY OF THE THREE MTL MODELS WITH DIFFERENT F-S-I LAYER COMBINATIONS.

F-S-I	en	ja	zh
MSL_{mBERT}			
4-4-4	0.827	0.863*	0.845
1-7-4	0.834*	0.850	0.856*
1-4-7	0.830	0.861	0.834
MSL_{LaBSE}			
4-4-4	0.827	0.855	0.873*
1-7-4	**0.848***	**0.869***	0.866
1-4-7	0.836	0.866	0.864
MSL_{mono}			
4-4-4	0.822	0.847	0.859
1-7-4	0.817	0.850*	**0.877***
1-4-7	0.836*	0.830	0.855

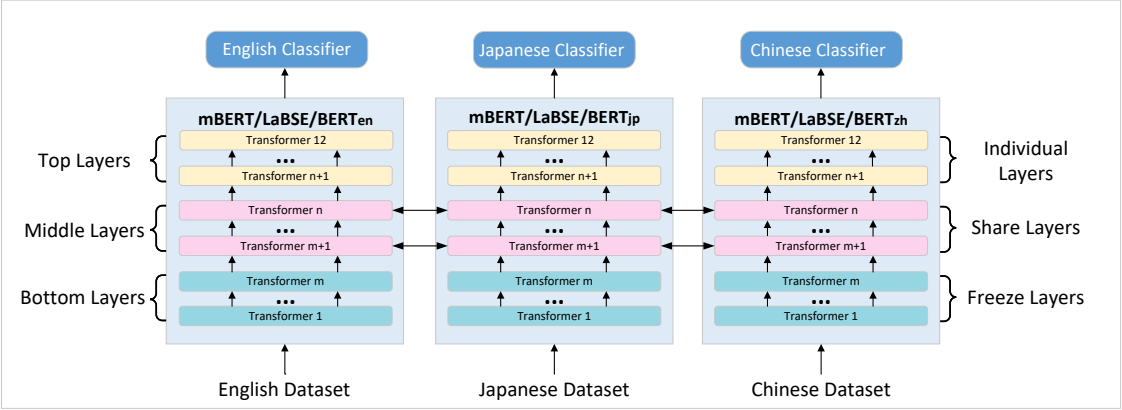

Fig. 1. The architecture of the MSL model.

2) Comparison with STL: Since the 1-7-4 layer combination was the best, we used the MSL models with this setting to compare with the STL models. The results are shown in Table III. The boldface values denote the highest accuracy for each language. First, we compared the MSL models with the STL models that employed the corresponding BERT variants (MSL_{mBERT} vs. $Single_{mBERT}$, MSL_{LaBSE} vs. $Single_{LaBSE}$, MSL_{mono} vs. $Single_{mono}$). We calculated the difference between the performance of the MSL models and the STL models (Δ). The reported Δ are averaged across all the languages. Overall, MSL tended to outperform STL, but the effect was not significant. Next, all the MSL models were compared to $Single_{mono}$ because it was the best-performing model among the STL models. We reported Δ_{mono} for the comparison. The results show that only the MSL model employed LaBSE was better than the $Single_{mono}$. Despite that, our results surpass the previous work [17] that utilized the BERT model pre-trained with Japanese clinical test.

TABLE III
THE COMPARISON OF THE ACCURACY BETWEEN MSL AND STL.

Model	en	ja	zh	Δ	Δ_{mono}
Baseline (STL)					
$Single_{mBERT}$	0.794	0.855	0.852		
$Single_{LaBSE}$	0.805	0.861	0.844		
$Single_{mono}$	0.838	0.856	0.873		
MSL(1-7-4)					
MSL_{mBERT}	0.834	0.850	0.856	+0.014	-0.009
MSL_{LaBSE}	**0.848**	**0.869**	0.866	+0.024	+0.005
MSL_{mono}	0.817	0.850	**0.877**	-0.008	-0.008

$^{\Delta}$ denotes the difference between the average performance of the MSL models and the STL models employed the corresponding BERT variants.
$^{\Delta_{mono}}$ denotes the difference between the average performance of the MSL models and $Single_{mono}$.

Table IV shows the label-wise F1 score for each language of the MSL_{LaBSE}(1-7-4). It can be seen that the F1 values of runny nose and cold for English were lower than those for Japanese and Chinese. This point will be discussed in the error analysis in Section VIII-B.

To sum up, the proposed model (MSL) does not always outperform the models that are specifically fine-tuned with each language (e.g., the accuracy of *zh* in MSL_{LaBSE} and $Single_{mono}$). However, the experimental result implies the potential merits of the models based on MSL.

VII. ADDITIONAL EXPERIMENTS ON BILINGUAL COMBINATIONS

Conneau et al. [18] showed that language similarity affects the cross-lingual transfer capability of multilingual pre-training models such as mBERT. Therefore, in this study, additional experiments were conducted to investigate the effect of language similarity on the MSL models. We created the Bilingual Simultaneous Learning (BSL) models by combining the three languages two by two. Figure 2 shows the architecture of the BSL model (Japanese-Chinese pair).

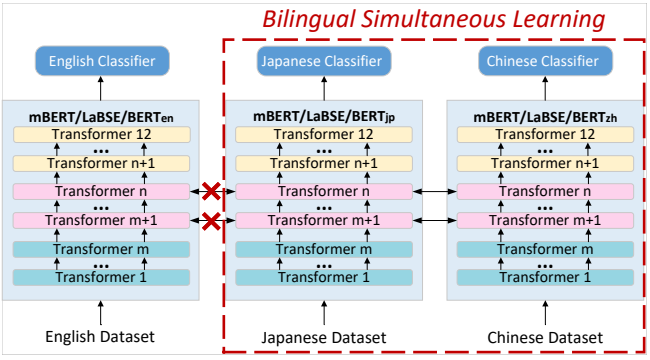

Fig. 2. The architecture of the BSL model.

Linguistically, English and Chinese have similar grammatical structures. Japanese and Chinese have similar surface forms. English and Japanese have the lowest language similarity among the combinations. We performed three sets of tests: English-Chinese, Japanese-Chinese, and English-Japanese. The results are shown in Table V. We only report the performance of the MSL(1-7-4) *without en*, as no significant improvement in performance was observed in *without ja* and *without zh*. The Japanese-Chinese pair demonstrated

TABLE IV
THE LABEL-WISED F1 SCORE FOR EACH LANGUAGE OF THE MSL_{LaBSE}(1-7-4).

Lang	Flu	Diarrhea	Hay fever	Cough	Headache	Fever	Runny nose	Cold
en	0.692	0.968	0.864	0.963	0.968	0.825	0.868	0.889
ja	0.769	0.954	0.876	0.951	0.954	0.872	0.926	0.905
zh	0.644	0.968	0.863	0.958	0.974	0.864	0.909	0.895

TABLE V
THE ACCURACY OF THE BSL MODELS.

Model	ja	zh	Δ	Δ_{mono}
Baseline				
$Single_{mono}$	0.856	0.873		
MSL(1-7-4)				
MSL_{LaBSE}	0.869	0.866		+0.002
MSL_{mono}	0.850	**0.877**		-0.002
BSL(MSL without en)				
BSL_{LaBSE}	**0.873**	0.866	+0.005	**+0.005**
BSL_{mono}	0.869	0.870	**+0.006**	**+0.005**

The values of $Single_{mono}$, MSL_{LaBSE}, MSL_{mono} is the same as Table III.
Δ denotes the difference between the average performance of the BSL models and the original MSL models employed the corresponding BERT variants.
Δ_{mono} denotes the difference between the average performance of the BSL models and $Single_{mono}$.

the highest exact match accuracy among all the groups that were tested. The average accuracy of the BSL was higher compared to that of the MSL combined three languages Δ. Compared with the best baseline, $Single_{mono}$, it also obtained higher performance (Δ_{mono}). In addition, the BSL model with LaBSE showed the highest accuracy in Japanese among all experiments. These results suggest that the degree of correlation between languages, especially the similarity of the surface forms, has an impact on the performance of the MSL model.

VIII. ERROR ANALYSIS

We investigated the prediction errors made by the MSL model. The error analysis was conducted on MSL_{LaBSE} (1-7-4) since it demonstrated superior performance in the previous experiment. We extracted all the incorrectly predicted sentences and corresponding labels and states from the test dataset for each language. It can be found that the model incorrectly classified some identical sentences in all three languages. These common errors represent the shared error patterns across the languages. There are also some errors that are specific to each language. These errors highlight the influence of linguistic features and translation nuances on the classification performance. We manually categorized all these error sentences, resulting in 12 distinct cases. Table VI shows the cause and corresponding example of each case, while Table VII reports the statistics of the 12 cases.

The sentences in the MedWeb dataset were annotated based on three principles: Factuality (whether the Twitter user has a particular symptom or not), Tense (whether the symptoms exist within 24 hours or not), Location (whether the symptoms described are experienced by the Twitter user or someone nearby). We manually categorized eight FP (False-Positive) error cases and four FN (False-Negative) error cases based on these principles. The following part will analyze each case in detail.

A. Common Error

Case 1: Overlapping. This type of error occurred due to the inability to understand words with multiple meanings. In example 1, the word headache is used metaphorically to express a complaint to the boss rather than express a clinical symptom. The MSL model misunderstood the sentence and incorrectly predicted that "Headache" is positive.

Case 2: Symptoms mentioned in general topics. As example 2, some tweets just express health-related discussions, roasts, or questions, rather than stating the tweeter's own symptoms. The MSL model could not distinguish between symptoms described as general topics and those actually occurring in people.

Case 3: Denied symptoms. This case pertains to the instances in which sentences contain negative expressions about symptoms. The MSL model failed to judge the negative expression or the scope of negation.

Case 4: Suspected symptoms. This type of error usually occurs in "Flu". In example 4, the tweeter suspected of having the flu, but it did not necessarily imply experiencing symptoms. The MSL model was unable to identify the symptoms that were only suspected.

Case 5: Fully recovered symptoms. According to the MedWeb annotation criteria, symptoms were labeled positive if they were in the recovery process and as negative if they had been fully recovered. The MSL model could not recognize the representation such as "went away" in example 5, and still predicted "Diarrhea" as positive.

Case 6: Past symptoms. This was categorized as Tense. In example 6, the tweeter stated an experience where he/she had a cold, and the symptoms had already passed. This error shows that the MSL model was insensitive to the past tense.

Case 9: Symptoms that are directly expressed. Despite that the sentences directly express the symptoms, the MSL model failed to predict them correctly.

Case 10: Implied symptoms. In contrast to Case 9, sentences in Case 10 employ allusions and metaphors to convey the symptoms indirectly. In example 10a, although the tweeter did not directly state that he/ she experienced hay fever, it can be inferred through the context. It was difficult for the MSL model to detect the symptoms that needed to be inferred.

Case 11: Symptoms that are in the recovery process. Based on the MedWeb annotation criteria, symptoms in the

TABLE VI
THE DETAILS OF 12 CASES OF ERRORS.

Error	No.	Cause of the error	Example sentence	Incorrect prediction
FP	1	Overlapping	Difficult bosses are one kind of headache.	Headache pos.
	2	Symptoms mentioned in general topics	Not a lot of people stay home from school due to allergies.	Hay fever pos. Runny nose pos.
	3	Denied symptoms	I thought I had the flu so I went to the doctor, but I got tested and I was wrong.	Influenza pos. Fever pos.
	4	Suspected symptoms	Seems I've had the flu since last night.	Influenza pos. Fever pos.
	5	Fully recovered symptoms	My diarrhea went away when I played soccer.	Diarrhea pos.
	6	Past symptoms	My cold this round was awful. It started with a high temperature, and a cough that wouldn't go away..	Cough pos. Fever pos. Cold pos.
	7	Co-occurring symptoms	(en) With my out-of-it head, I just figured it out, I have a cold.	Runny nose pos.
			(ja) 風邪ひいて鼻水が止まるようくすり飲んだよ.	-
			(zh) 感冒了为了止住鼻涕吃了药。	-
	8	Symptoms not yet occur	(en) I have allergies, so I'm already super scared of next spring.	-
			(ja) 俺は花粉症なので今から来春が超怖い.	Runny nose neg.
			(zh) 我有花粉症所以从现在开始到明年春天超恐怖。	Runny nose neg.
FN	9	Symptoms that are directly expressed	Allergy season is so exhausting.	Hay fever neg. Runny nose neg.
	10 a	Implied symptoms	The people who are saying there's not a lot of pollen today don't have allergies, so...	Hay fever neg. Runny nose neg.
	10 b		(en) I caught a cold and took a decongestant.	Runny nose neg.
			(ja) 風邪ひいて鼻水が止まるようくすり飲んだよ.	-
			(zh) 感冒了为了止住鼻涕吃了药。	-
	11	Symptoms that are in the recovery process	I took medicine and my congestion stopped like that.	Hay fever neg. Runny nose neg.
	12	Misunderstanding about denied expression	(en) Please no diarrhea today, I have an important interview.	Diarrhea neg.
			(ja) 今日は大事な面接日なので下痢みたいなものは勘弁してくれ.	-
			(zh) 今天是重要的面试的日子所以像拉肚子这样的事就饶了我吧。	-

TABLE VII
THE STATISTIC OF 12 CASES OF ERRORS.

No.	Category	Number of errors						
		Common	Language specific			Total		
			en	ja	zh	en	ja	zh
1	Factuality	1	1	2	-	3	4	3
2	Factuality	**11**	1	**3**	**6**	14	**19**	**24**
3	Factuality	3	-	-	1	3	5	6
4	Factuality	8	-	1	1	8	9	9
5	Factuality	4	2	-	1	7	5	5
6	Tense	2	-	2	-	3	3	3
7	Factuality	-	4	-	-	4	-	-
8	Tense	-	-	1	-	-	2	2
9	Factuality	4	8	**3**	4	13	10	10
10	Factuality	5	**16**	**3**	3	**23**	14	14
11	Factuality	3	1	-	-	4	3	3
12	Factuality	-	2	1	-	2	1	-

recovery process would be annotated as positive. The MSL model could not predict the duration of having symptoms.

B. Language-specific errors

1) Language-specific errors for English:

Case 7: Co-occurring symptoms. Example 7 expresses that the cold is positive, but the MSL model detected both cold and runny nose as positive. Interestingly, we find that this type only occurred in the English corpus. According to the MedWeb annotation, there are 2 types of sentences when annotating cold. In type 1, only cold was labeled as positive. While in type 2, cold and other symptoms (mostly runny nose) were positive simultaneously. For these 2 types of annotation, both Japanese and Chinese have corresponding keywords to represent cold. In Japanese, there are "鼻風邪 (nose cold)" and "風邪 (cold)" to distinguish different types of cold. Specifically, when there is "鼻風邪", both cold and runny nose are positive. When there is "風邪", only cold is positive. Similarly, in Chinese, there are "伤风" (wind damage) and "感冒" (cold). However, English has only one expression of "cold" for both types. This is one possible reason why English has lower performance for cold and runny nose than Japanese and Chinese.

Case 12: Misunderstanding about denied expression. In example 12, it is evident from the context that the tweeter was suffering from diarrhea. However, the MSL model misunderstood the denied word "no" and predicted it as negative. This type of error is more likely to occur in English because Chinese and Japanese have a broader lexicon for expressing negation, rather than directly using the negation words such as "no" and "not".

As Table VII shows, the English-specific error in FN for implied symptoms (Case 10) corresponds to a considerable number. As shown in example 10b, the sentences that were incorrectly predicted contain more expressions of taking the medicine. In the training data, Japanese and Chinese tend to use the symptom name directly. Moreover, the expressions for medicine are often in a broad sense without specifying the medication type. In contrast, English tends to use a specific type of medicine to imply the symptoms. This type of error shows the lack of medicine-related knowledge in the MSL model.

2) Language-specific errors for Japanese and Chinese:

Case 8: Symptoms not yet occur. This type of error was categorized as Tense. In example 8, it can be inferred that the tweeter is not yet showing the symptoms of hay fever, but the MSL model could not recognize the tense.

Overall discussion for errors: The analysis above reveals that errors related to the factuality of symptoms constitute a significant proportion. The reason why errors related to tense and location are less common may be because MedWeb lacks the relevant data. Overall, FP for symptoms mentioned in general topics is the most frequent type of error. The MSL model encountered challenges in making a judgment about whether tweeters were stating their own symptoms or talking about general topics. Additionally, the total counts in Table VII show that Japanese and Chinese are more likely to have this type of error than English. In terms of linguistic features, Japanese and Chinese tend to omit the subject whenever describing their own condition or discussing general topics. English, on the other hand, tends to use "I" to state personal symptoms.

While for the English corpus, the most frequent error is FN for implied symptoms. Among the total of 23 error sentences, a majority contained a representation of taking medicine. Compared to the other 2 languages, English tends to imply symptoms with the specific medicine. Since the model lacked medicine-related knowledge, it failed to detect those implied symptoms.

IX. Conclusion

In this paper, we proposed the Multilingual Simultaneous Learning (MSL) model using BERT with soft parameter sharing. We evaluated the MSL model on a multi-labeled symptom classification task. The results suggested that the MSL model performed better than the STL models. However, in some cases, it could not surpass the language-specific monolingual BERT. In addition, we conducted additional experiments and created Bilingual Simultaneous Learning (BSL) models. The Japanese-Chinese pair exhibited the best performance compared to all other tests. This result indicated that the similarity of surface forms between languages affects the performance of the MSL model.

A detailed error analysis was conducted to reveal the limitations of the MSL model. We manually categorized 12 cases of errors and analyzed them from both common and language-specific perspectives. The analysis results implicated the importance of improving sentence subject recognition and medical understanding. Further work needs to be done to verify the effectiveness of the model on data of different sizes in other languages.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 23K11368.

References

[1] Eiji Aramaki, Shoko Wakamiya, Shuntaro Yada, and Yuta Nakamura. Natural language processing: from bedside to everywhere. *Yearbook of Medical Informatics*, 2022.

[2] Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. Real-mednlp: Overview of real document-based medical natural language processing task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, pages 285–296, 2022.

[3] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[4] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[5] Anders Søgaard and Yoav Goldberg. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, 2016.

[6] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[7] Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cécile Paris, and C Raina MacIntyre. Does multi-task learning always help?: An evaluation on health informatics. In *Proceedings of the the 17th annual workshop of the Australasian language technology association*, pages 151–158, 2019.

[8] Mareike Hartmann and Anders Søgaard. Multilingual negation scope resolution for clinical text. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 7–18, 2021.

[9] Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. Overview of the ntcir-13: Medweb task. In *NTCIR*, 2017.

[10] Yongjie Lin, Yi Chern Tan, and Robert Frank. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy, August 2019. Association for Computational Linguistics.

[11] Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy, August 2019. Association for Computational Linguistics.

[12] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

[13] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Visualizing and understanding the effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4143–4152, Hong Kong, China, November 2019. Association for Computational Linguistics.

[14] Niraj Pahari and Kazutaka Shimada. Multi-task learning using bert with soft parameter sharing between layers. In *2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS)*, pages 1–6. IEEE, 2022.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[16] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[17] Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. A clinical specific bert developed using a huge japanese clinical text corpus. *Plos one*, 16(11):e0259763, 2021.

[18] Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online, July 2020. Association for Computational Linguistics.