Федеральное государственное автономное образовательное учреждение
высшего образования «Национальный исследовательский университет
«Высшая школа экономики»

Факультет гуманитарных наук
Основная образовательная программа
Фундаментальная и компьютерная лингвистика

КУРСОВАЯ РАБОТА

на тему

"Использование мультиязычных моделей для вопросно-ответных систем"

"Usage of multilingual models in Question-answering systems"

Выполнила студентка группы 181, 3 курса,
Матяш Дарья Сергеевна

Руководитель ВКР:
Доцент, кандидат технических наук,
Артемова Екатерина Леонидовна

Москва 2021

Содержание

1 Introduction

Task-oriented dialog systems are known as one of the most popular spheres in modern Natural language Processing (NLP). Such systems are geared towards helping users with such particular tasks as setting an alarm, getting to know about weather forecast, etc. Task-oriented dialog systems are made to understand what users mean speaking or texting in a chat-bot by identifying users' intents (what the user wants) and the corresponding slots (the key entities which are found relevant to the intent).

Example 1.*Remind me of my [paper] [on Thursday].*

In the Example 1 the user's intent is to set the reminder, so the name of the intent of the mentioned example is given as 'set reminder'. Slots or intent-arguments modify the intent, which in this example are 'paper' and 'on Thursday'. Intent-arguments are given a name, such as 'to do' and 'datetime' accordingly. In current work we firstly deal with the problem of building task-oriented dialogue system using unbalanced data. Our aim was to find whether it is possible to train a model in a joint fashion or two separate models to do slots filling and intent classification in both major and minor classes of slots and intents respectively and then transfer knowledge included in datasets made on English utterances and use models fine-tuned on English data in Spanish dataset.

For performing intent classification and slot filling we compared two approaches:

1 Using a joint model

   1.1. Fine-tuning pretrained multilingual uncased Joint BERT-base (Chen et al. 2019)

   1.2. Fine-tuning pretrained multilingual uncased XLM-Roberta-base (Liu et al. 2019)

2 Using two models: one for intent classification and one for slot classification,

for separate tasks (Chen et al. 2019)

The best performance was shown by the approach (2) but fine-tuned joint m-BERT and XLM-Roberta also had comparable results. Transfer learning from English data was quite more successful on Spanish data rather than Thai, two possible reasons of which was the difference in Spanish and Thai datasets' sizes and genealogical relationship.

In Section 2 there is a short observation of previous works' methods and results. Section 3. describes the data that was used in the experiments. In Section 4. we explain the chosen metrics in order to measure models' results and interpret them correctly. It also presents the chosen models' architecture and features and details the algorithms and methods used to solve slot filling and intent classification and transfer-learning tasks. The results gained after all conducted experiments are presented in Section 5. Section 6 describes future prospects and possibilities of improving current work's results. The overall conclusion is formulated in Section 8. Next section presents some additional illustrations to work. Previous work on the field of our research is located in Section 9.

2 Related work

For Intent Classification and slot Filling Task there were invented two main approaches: the use of two-headed models (like in (Chen et al. 2019)) or the use of two separate models, one of which classified intents and another filled the slots. In (Chen et al. 2019) joint BERT model slightly outperformed no joint model. The main tasks involved in the dialogue system are the classification of intentions and filling in slots, which can be formulated as the tasks of classification of sentences and marking of sequences respectively. These two tasks can be learned together or as two separate tasks. Most of the approaches proposed for successful strategies are based on deep learning algorithms (Chen et al., 2019; Dowlagar et al., 2021) and therefore require large training datasets. Transfer learning is a machine learning (ML) research problem that focuses on storing knowledge, solving one problem, and applying it to another related problem. Sometimes transfer learning is used for work with less-resource data like in Lacalle 2020. As Spanish and especially Thai datasets are significally smaller than English FMTOD we used transfer learning techniques on the mentioned ground. As the Spanish language belongs to the same Indo-European language family as English (though they belong to different brunches – Romance and Germanic respectively) and both English and Spanish languages use Latin alphabetical system, while the Thai language belongs to Kra–Dai languages and uses special Thai script (or Thai abugida) we consider Spanish data (which is also quite larger than Thai one) be better or not worse fine-tuned. There are several datasets related to the tasks of intent classification and slot filling. Most popular are both mono-domain datasets such as airline travel information system (ATIS) corpus (Tur et al.), CLINC (Larson et al.) and multi-domain data like SNIPS (Coucke et al.). Facebook Multilingual Task Oriented Dataset (FMTOD), which we chose for current work, is referred to as the former type.

3 Dataset

Facebook Multilingual Task Oriented Dataset (FMTOD) is a multilingual multi-domain dataset, labelled for intent classification and slot filling tasks. FMTOD was manually generated and annotated for three languages: English, Spanish and Thai (Schuster et al. 2019). Originally there were around 43,000 utterances in English accumulated across three domains: ALARM, REMINDER, WEATHER. For the data in Spanish and Thai, native speakers of those languages respectively were asked to translate the phrases in English. Then some samples were discarded if the labels of two annotators that were to label slots in the given data did not correspond (for Thai data) or only one of two examples was left (for English and Spanish data). Table 1 depicts summary statistics of the dataset used in actual work. Note that the number of English examples substantially outweighs the amount of Spanish and Thai utterances. Each line in each dataset looks as it is shown in Example 1: every token in utterance is given a label (some of the slot types' labels or 'NoLabel') and every utterance is given an intent class. Some phrases may have an intent but there can be no label for tokens. There are examples of utterances with their intent marking and slot labelling in FMTOD below. Intents assigned to the phrases are on the right. All labelled slots are below the tokens.

Remind me of my	paper	on	Thursday	reminder/set_reminder
	B-reminder/todo	B-datetime	I-datetime	

Example 1: Example of utterance 'Remind me of my paper on Thursday' from FMTOD.

Give me the	latest	forecast	for	Half	Moon	Bay	weather/find
	B-datetime	B-weather/noun		B-location	I-location	I-location	

Example 2: Example of utterance 'Give me the latest forecast for Half Moon Bay' from FMTOD.

Number of utterances									
	English			Spanish			Thai		
Domain	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
Alarm	9,282	1,309	2,621	1,184	691	1,011	777	439	597
Reminder	6,900	943	1,960	1,207	647	1,005	578	336	442
Weather	14,339	1,929	4,040	1,226	645	1,027	801	460	653
Total	30,521	4,181	8,621	3,617	1,983	3,043	2,156	1,235	1,692

Table 1. Statistics of datasets used in the experiments.

As it was previously said, FMTOD is a multi-domain dataset and intent classes contain tree domains: ALARM, REMINDER, WEATHER. They are classified in accordance with 12 types of intentions and 11 types of arguments, which are illustrated in Table 2.

Domain	Intent type	Slot type
Alarm	6	2
Reminder	3	6
Weather	3	5
Total	12	11

Table 2. Domain, intent types and slot types which are included in FMTOD.

According to Table 3, it is obvious that intent types are unbalanced in train datasets in all languages.

		Number of utterances					
Domain	Intent type	English		Spanish		Thai	
		train	test	train	test	train	test
alarm	cancel_alarm	124	588	313	248	227	186
	modify_alarm	168	125	47	34	8	13
	set_alarm	14047	1387	588	543	376	272
	show_alarms	1151	294	143	121	101	85
	snooze_alarm	4743	118	52	35	34	21
	time_left_on_alarm	1006	109	41	30	31	20
reminder	cancel_reminder	2069	333	230	196	135	103
	set_reminder	439	1340	768	646	317	268
	show_reminders	4816	287	209	163	126	71
weather	checkSunrise	1142	34	3	2	801	653
	checkSunset	432	54	3	1025	227	186
	find	384	3952	1220	248	8	13

Table 3. Intent type hierarchy.

Figure 1 illustrates the distribution of utterances by domain in each language. According to Figure 1, the WEATHER domain is predominant, so the overbalance

in this domain prediction was expected, though it did not happen after all conducted experiments.

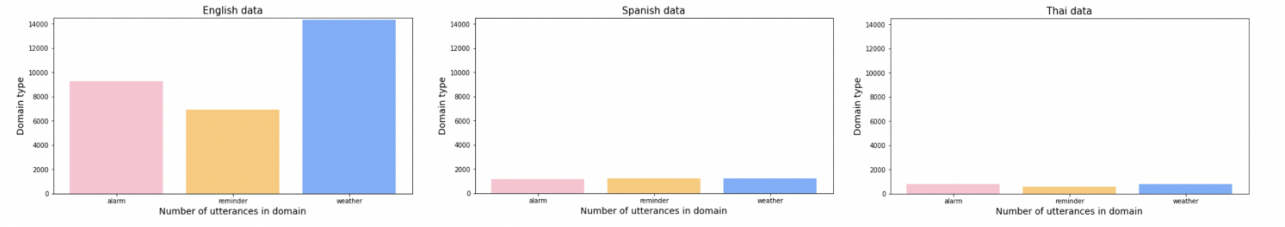

Figure 1. Distribution of utterances by domain in English, Spanish, Thai datasets.

Figure 2 more thoroughly depicts the distribution of phrases in each language by intent type. It is noticeable that in non-English datasets not all intent classes from the English data included but it was not a problem for transfer-learning tasks.

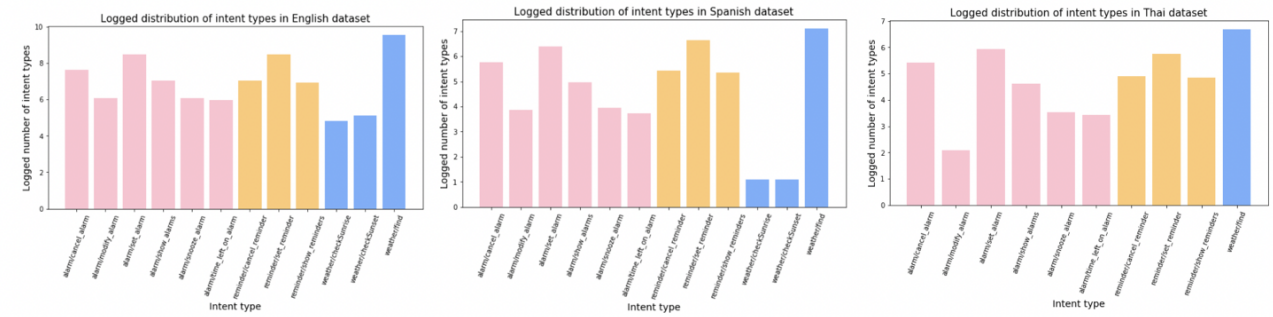

Figure 2. Distribution of utterances by intent type in English, Spanish, Thai datasets.

Figure 3 shows the distribution of utterances by slot type in English dataset. We can notice that slot types are quite unbalanced, too, so we did not expect high F1 or accuracy scores for minor classes, such as 'news/type', 'timer/attribute', 'timer/noun'. We also did not expect the same for the same slot classes in Spanish and Thai datasets for the same reason. At the same time, we consider major slot classes to be successfully detected due to the predominant quality of examples, quite enough for making comparatively accurate predictions. In order to tackle the problem connected with data containing unbalanced slot classes a few methods

were used. Firstly, all duplicated utterances in all datasets (training, evaluation and test) were left for over-sampling. Earliest tests showed reasonable differences in models' results' qualities, so all described in the current work models used data with duplicated utterances. Due to severe imbalance of intent and slot classes, we assume some difficulties during the model training. It is also a reason to use some special metrics in order to estimate models' performances and interpret their results correctly. We decided to use not only most used metrics for the same task as our such as F1-score and accuracy which are used, for example, in Schuster et al. 2019, Chen et al. 2019, but also Matthews correlation coefficient (MCC) for each intent and slot class and Cohen's kappa score. The two former metrics' results are included in Appendix (Section 8).

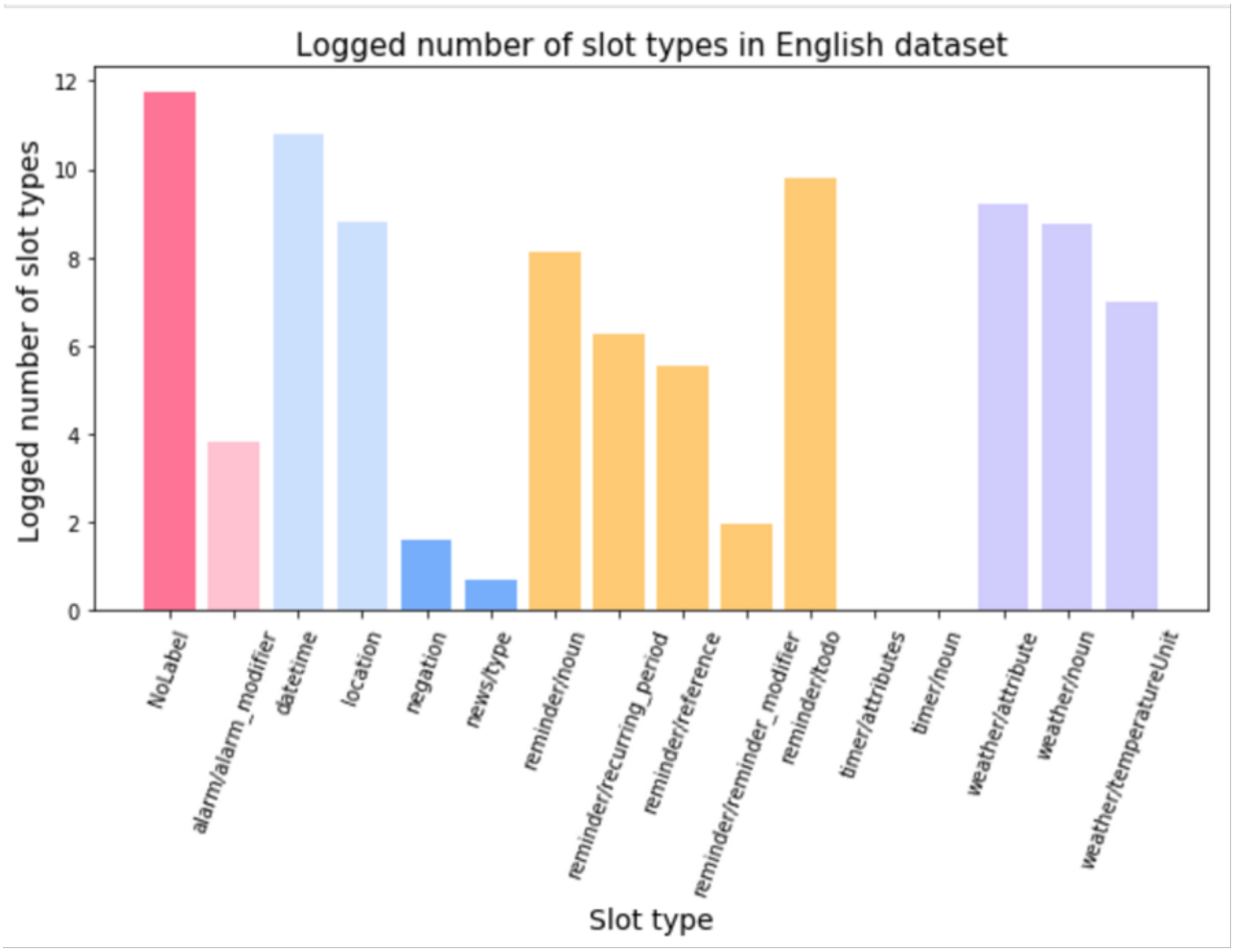

Figure 3. Distribution of utterances by intent type in English, Spanish, Thai datasets.

We gave an example of slot type distribution in English dataset, it should be mentioned that all tokens are labelled with the inside – outside – beginning (IOB)

format (Ramshaw et al. 1995), the "industry standard" encoding (Carpenter 2009) which is used in all the used data. We have thorough results' metrics in appendix for each label in dataset, but we count mean F1-score for each whole slot type (including B-label and I-label) as in, for example, in Schuster et al. 2019, Chen et al. 2019.

4 Approach

4.1 Models

4.1.1 Joint multilingual BERT (m-BERT)

The architecture of the BERT model is a multilayer bi-directional Transformer encoder which is based on the original Transformer model (Vaswani et al., 2017). It is trained on 102 languages from the Wikipedia and Books corpora and has more 110 million parameters taking into account two classifying heads for slot filling and intent classification.

The input representation is a concatenation of positional embeddings, the segment embedding, and WordPiece embeddings (Wu et al., 2016). Especially for single sentence classification and tagging tasks, there is no discrimination in sentence embedding. As the first token a special classification token ([CLS]) is put before the beginning of the sentence and a special token ([SEP]) is inserted as the final token. Given an input token sequence $x = (x_1, ..., x_T)$, the output of BERT is $H = (h_1, ..., h_T)$. Figure 4 (Chen 2019) shows the example input query "play the song little robin redbreast".

BERT model is pre-trained model using two strategies for large-scale unlabeled text: the masked language model and next sentence prediction. The pretrained BERT model provides a powerful embeddings' presentation dependent on context of sentence representation and can be used for a variety of target tasks one of which can be classifying intents and filling slots, using a fine-tuning procedure, similar to that used for other NLP tasks.

BERT model can be extended to a joint intent classification and slot filling model. Based on the hidden state of the first special token ([CLS]), denoted h_1, the intent is predicted as:

$$y_i = softmax(W^i h_1 + b^i).$$

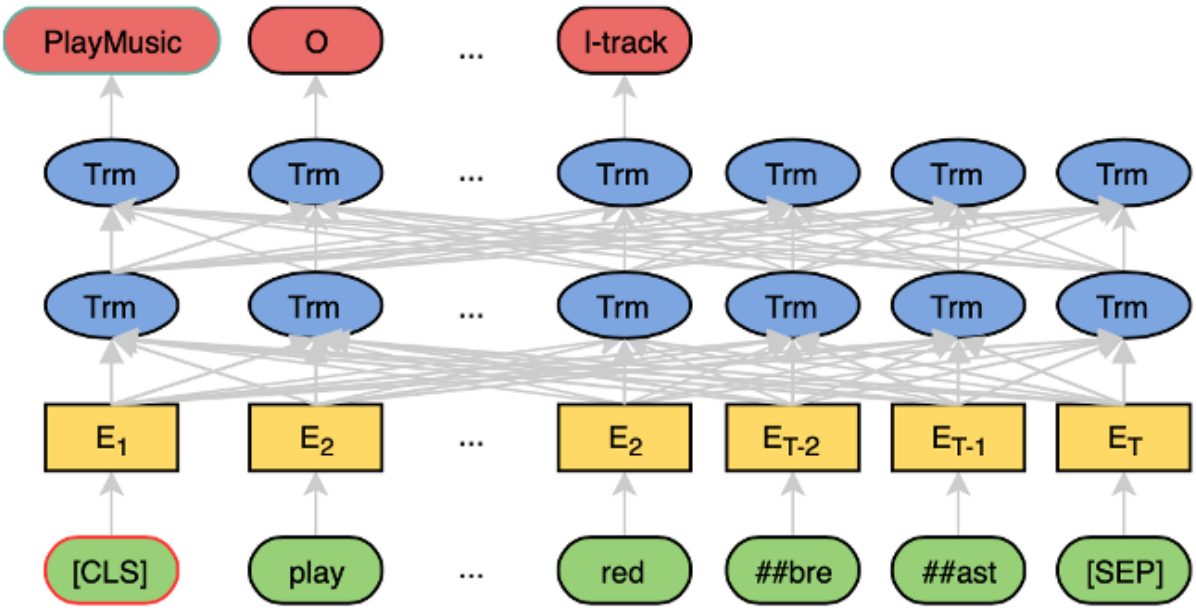

Figure 4. A high-level view of the proposed model. The input query is "play the song little robin redbreast".

To solve the slot filling task, we feed the final hidden states of other tokens $h_2, ..., h_T$ into a softmax layer to classify over the slot filling labels. We feed each tokenized input word into a WordPiece tokenizer and use the hidden state corresponding to the first sub-token as input to the softmax classifier in order to make this procedure compatible with the WordPiece tokenization.

$$y_n^s = softmax(W^s h_n + b^s).$$

where h_n is the hidden state corresponding to the first sub-token of word x_n. For jointly modeling intent classification and slot filling, the objective is formulated as:

$$p(y^i, y^s|x) = p(y^i|x) \prod_{n=1}^{N} p(y_n^s|x).$$

The learning objective is to maximize the conditional probability $p(y_i, y_s|x)$. The model is fine-tuned end-to-end via minimizing the cross-entropy loss.

4.1.2 Joint XLM-Roberta

XLM-Roberta is transformer-based multilingual masked language model pre-trained on text in 100 languages, which obtains state-of-the-art performance on cross-lingual classification, sequence labeling and question answering (Conneau 2020). It is trained on CommonCrawl, which contains several orders of magnitude more data for many languages. It has, as m-BERT-base, 768 hidden states, 12 attention heads, 12 layers but more than 270 million parameters. That is why it was twice more time-consuming for fine-tuning. As we solve the problem of simultaneous classification of intents and slot filling in a sentence using a single model, the model has two outputs, as it was explained above for joint BERT architecture, the first predicts the intents, the second predicts the labels of words.

4.1.3 Token Classification Model + Text Classification Model

To solve intent classification and slot filling tasks as two separated tasks, we used pretrained m-BERT for Sequence Classification and Token Classification from Hugging Face AI separately. As we used multilingual BERT-base in all cases, both models had 768 hidden states, 12 attention head, 12 layers and 110 million parameters.

Each of the models was trained with the same hyperparameters - 15 epochs (except the former one – it was pre-trained only with 5 epochs) on a training sample with a training step length of 10-5 and a batch size of 16 objects. Cross-entropy was used as the error function:

$$L = -\frac{1}{n}\sum_{i}^{n} y \log \hat{y}$$

4.2 Optimizer

Adam optimizer was chosen as an optimizer of all the mentioned models. It combines both the idea of accumulating movement and the idea of a weaker

update of the weights for typical features. We freezed all the layers in pretrained m-BERT and XLM-Roberta except last classifying layers on resource-economy purposes. Both joint-fashioned models performed quite successfully but it took us less time and sources to reach such result.

4.3 Number of epochs

The number of epochs was chosen according to resource-and-success balance seen in (Chen 2019) and (Schuster 2019). All models converged successfully, all losses looked normal, in Appendix some examples of training and validation losses could be found.

4.4 Evaluation metrics

For models' token labelling evaluation, when the IOB schema is adopted, where each of the words are tagged with their position in the slot: beginning (B), in (I) or other (O), recall and precision values are computed for each of the slots. A slot is considered to be correct if its range and type are correct. The F-Measure is defined as the harmonic mean of recall and precision. We also used accuracy and F1 scores for estimate models' ability to classify intents.

$$\text{F-measure} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$

$$\text{Recall} = \frac{\#\text{ correct slots found}}{\#\text{ true slots}}$$

$$\text{Precision} = \frac{\#\text{ correct slots found}}{\#\text{ found slots}}$$

We also used Matthews correlation coefficient (MCC), when the current slot or intent type is labelled as positive class and others as negative for lighter and more correct results' interpretation. Matthews correlation coefficient can show more accurately how an algorithm classifies small or large classes than classical accuracy or F1-score as accuracy score does not consider class size.

$$N = TN + TP + FN + FP$$

$$S = \frac{TP + FN}{N}$$

$$P = \frac{TP + FP}{N}$$

$$MCC = \frac{TP/N - S \cdot P}{\sqrt{PS(1-S)(1-P)}}$$

5 Results

We noticed that the most successful performance is done by two separate m-BERT models. Both Token Classification + Text Classification m-BERT models (with 1e-4 and 2e-4 learning rates) showed two best accuracy and F1-scores. This architecture was also the most time- and resource-consuming, so this strategy can be artificially chosen the best in order to find the best model with "pure" best offline result. However, if we talk about some practical issues and goals, we would rather consider the joint m-BERT model's performance the best as it is the third accurate model and, which is also important, is lighter than two Token Classification and Text Classification models. The joint XLM-Roberta's scores are also high and comparable with others. The more detailed results are shown in Table 4 below.

Model	Epochs	Intents			Slots	
Joint m-BERT		F1	Accuracy	MCC	F1	MCC
	5	0.81	0.79	0.75	0.81	0.78
	9	0.85	0.84	0.82	0.82	0.8
	15	**0.91**	**0.93**	**0.94**	**0.9**	**0.87**
Joint XLM-Roberta	5	0.69	0.72	0.7	0.51	0.52
	9	0.79	0.83	0.84	0.77	0.75
	15	**0.94**	**0.95**	**0.95**	**0.89**	**0.87**
Token Classification + Text Classification models, learning rate (LR) = 1e-4	5	**0.96**	**0.97**	**0.96**	**0.96**	**0.95**
Token Classification + Text Classification models, learning rate (LR) = 2e-4	5	**0.97**	**0.98**	**0.96**	**0.96**	**0.95**

Table 4. Results after fine-tuning models on English dataset.

After transfer learning, Token Classification + Text Classification models with learning rate 1e-4 had slightly better results than the same pair of models with learning rate 2e-4. The difference in two models' performances is notable on Thai data's results. XLM-Roberta outperformed m-BERT's results which is quite often to happen in common tasks like, for example, in (Lee 2019) and (Conneau 2020). Joint m-BERT's performance is also very comparable and quite high. The more

detailed results are shown in Table 5 below.

Model	Epochs	Intents			Slots	
Joint m-BERT		F1	Accuracy	MCC	F1	MCC
	5	0.81	0.79	0.75	0.81	0.78
	9	0.85	0.84	0.82	0.82	0.8
	15	**0.91**	**0.93**	**0.94**	**0.9**	**0.87**
Joint XLM-Roberta	5	0.69	0.72	0.7	0.51	0.52
	9	0.79	0.83	0.84	0.77	0.75
	15	**0.94**	**0.95**	**0.95**	**0.89**	**0.87**
Token Classification + Text Classification models, learning rate (LR) = 1e-4	5	**0.96**	**0.97**	**0.96**	**0.96**	**0.95**
Token Classification + Text Classification models, learning rate (LR) = 2e-4	5	**0.97**	**0.98**	**0.96**	**0.96**	**0.95**

Table 5. Results after transfer-learning on Spanish and Thai data..

It is to notice that in all cases transfer-learned models on Thai data performed relatively worse than models fine-tuned on Spanish data. It could have happened for many reasons, one of which is suggested below. Table 6 also contains possible solutions of our supposed problems.

№	Possible reasons	Possible solutions
1	Difference in sizes of train datasets: there are more utterances in Spanish than in Thai for fine-tune models.	1. Enlarge Thai dataset using • augmentation algorithms • crowd-sourcing methods • do minor classes reduplication, try to balance the classes 2. Fine-tune model for Thai data on more epochs and search more appropriate models' hyperparameters
2	Difference of language families: the more similar from linguistical prospective the better fine-tuned model's results	1. try to approve or disapprove this suggestion and: • improve fine-tuning model's embeddings, and/or • augmentate Thai data with Thai/other typologically closer languages

Table 6. Possible reasons why transfer learning on Spanish data was more successful than on Thai data and possible solutions of the problems.

More detailed (intent-by-intent and slot-by-slot) analysis is given below.

5.0.1 Intent Classification

As for intent classification performance on Spanish data, we can see considerable success at classifying major classes for all models. It is also to mention that all models tend to mix up one-domain classes which is much better as if in most cases mixed classes from other domains. Best results at Intent Classification were gained by joint XLM-Roberta and m-BERT models for Intent Classification, though the former two models were able to detect minor classes more accurate than XLM-Roberta (see classification for such minor Intent classes as, for example, "alarm/modify alarm"or "alarm/snooze alarm"). Though there was not too many examples in minor classes, all models quite successfully performed in Intent Classification task. Figures 5-7 depict confusion matrices of Intent Classification in Spanish data. True labels - vertically, predicted labels - horizontally. In all Figures 5-7 we can notice the prevalence of correctly predicted classes for each true class, which cannot be considered as good models' performance.

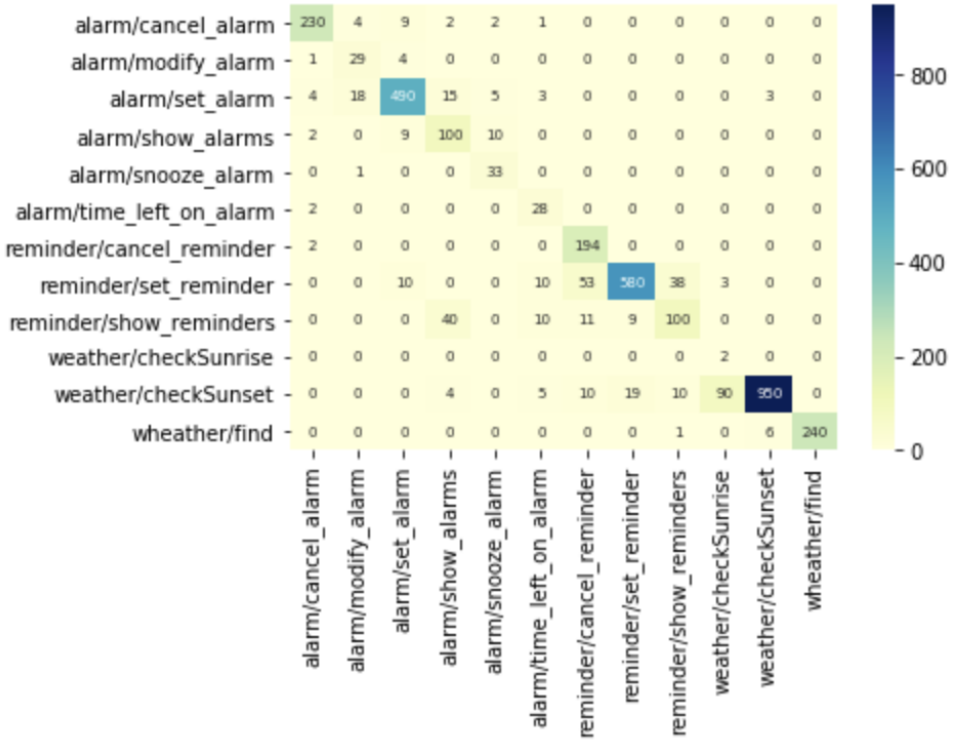

Figure 5. Confusion matrix for true and predicted Intent types by XLM-Roberta for Spanish.

As for intent classification performance on Thai data, we can see quite successful

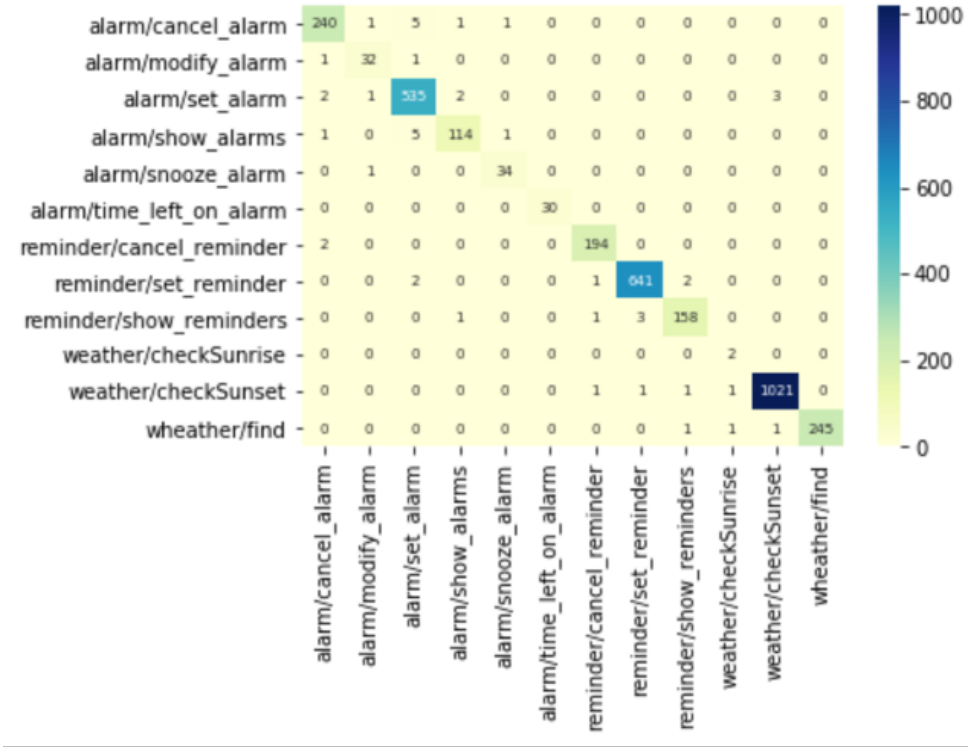

Figure 6. Confusion matrix for true and predicted Intent types by m-BERT for Intent Classification, LR=2e-4, for Spanish.

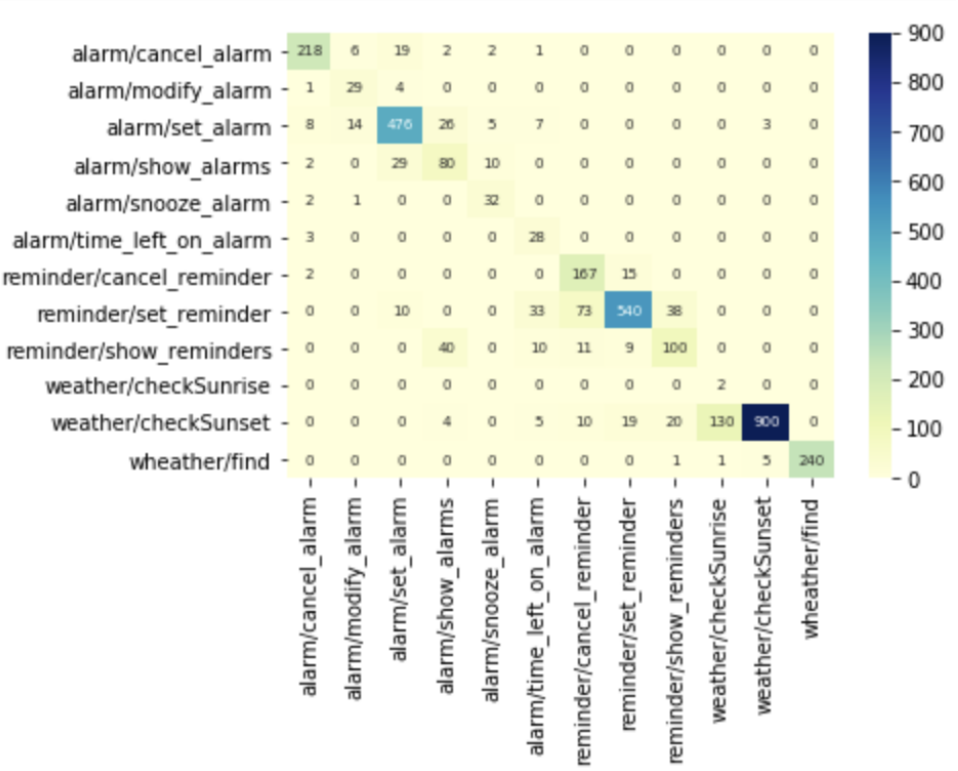

Figure 7. Confusion matrix for true and predicted Intent types by joint m-BERT for Spanish.

results at classifying major classes for all models. All models tend to mix up one-domain classes which is much better as if in most cases mixed classes from other

domains. However, as in Thai dataset there were less utterances of each Intent class, not all the classes, especially small, were not as accurately detected as larger ones. Figures 8-10 depict confusion matrices of Intent Classification in Thai data. True labels - vertically, predicted labels - horizontally

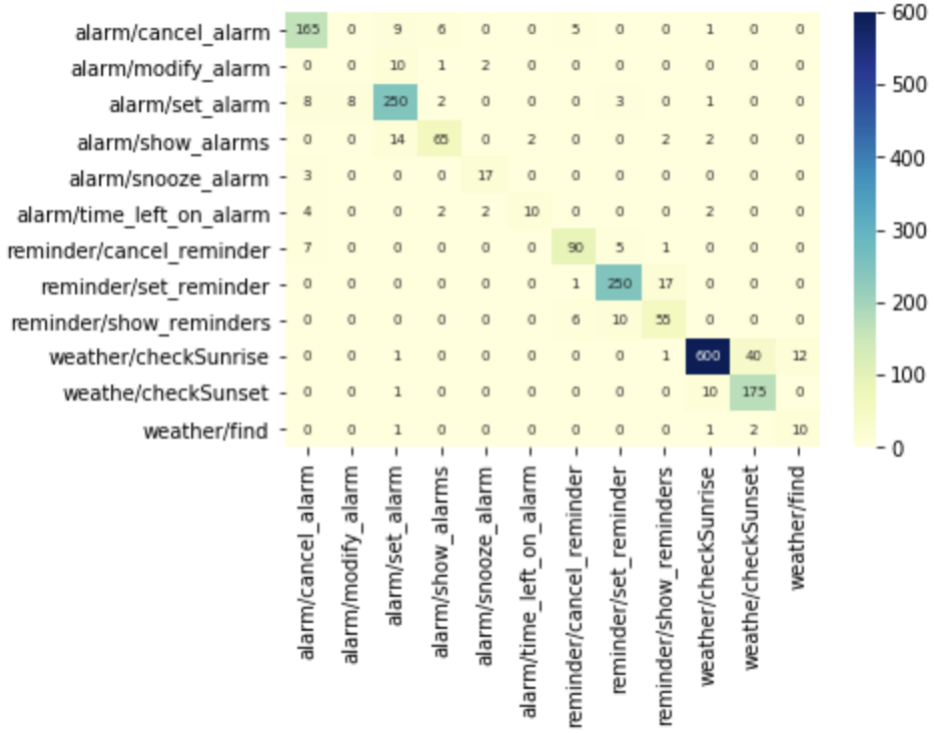

Figure 8. Confusion matrix for true and predicted Intent types by XLM-Roberta for Thai.

Though, as it was mentioned, all models' results on Thai data are quite lower than on Spanish one, we can also notice that, as for models fine-tuned on Spanish, all models succeeded in defining major classes and did some mix-ups mostly inside same domain. To sum up models' intent classification performances' analysis, we would say that all models succeeded in this task and gained quite high and competitive results, but there were a few mentioned tendencies which are better to be corrected in future improvements – it would be useful to reduce in- and (which is more important) outside-the-domain classification mix-ups lower misclassification of minor classes by, for example, minor classes data augmentation.

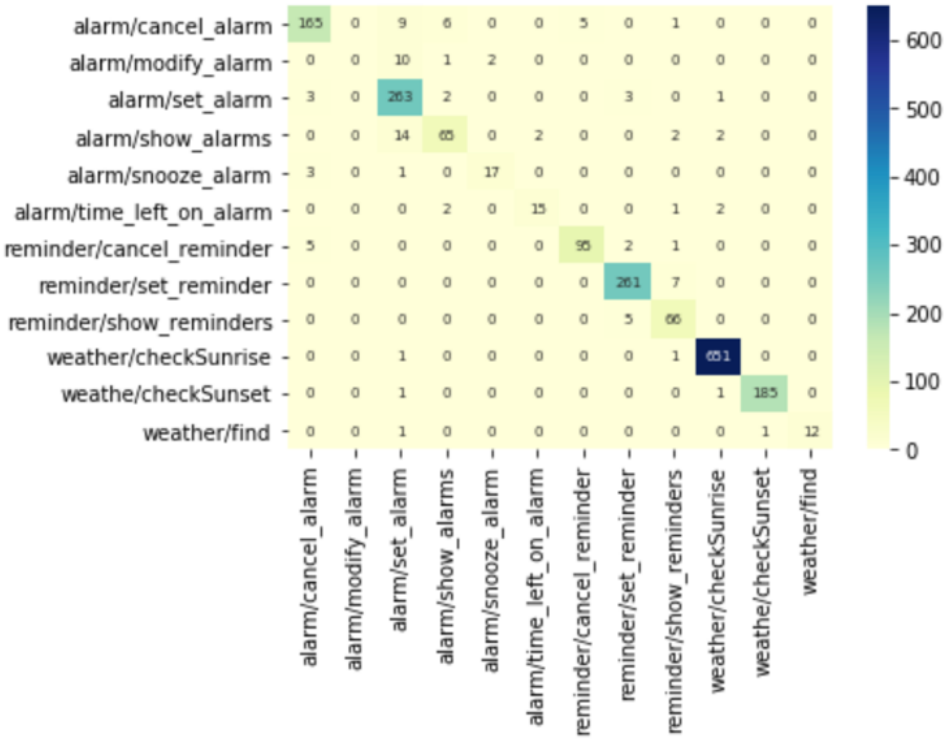

Figure 9. Confusion matrix for true and predicted Intent types by m-BERT for Intent Classification, LR=2e-4, for Thai.

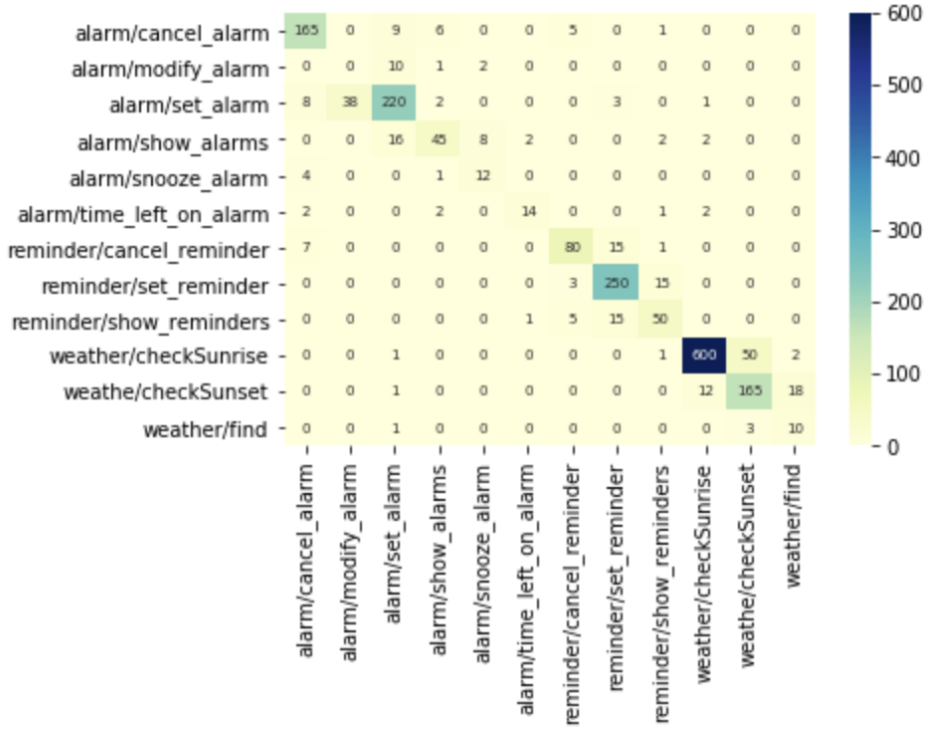

Figure 10. Confusion matrix for true and predicted Intent types by joint m-BERT for Thai.

5.0.2 Slot Filling

For slot classification, models performed less accurate in some cases but still, as we can say, with flying colors. There is still an influence of major classes in

in-domain classification, but what is very important, the less examples in training dataset, the less probability of minor class prediction. We can notice that minor or smaller classes were tend to be mislabeled with major classes and, in most cases, they were mislabeled with major classes of the same superclass (for example, minor 'weather/attribute' class was mislabeled with 'weather/noun' class or, which is less expected and wanted, to major 'NoLabel' class) or of the same intent type which it is referred to. For all the models and XLM-Roberta especially, we can mention that the lack of utterances of minor classes was in a few cases a reason for misunderstanding of data patterns and misclassification.

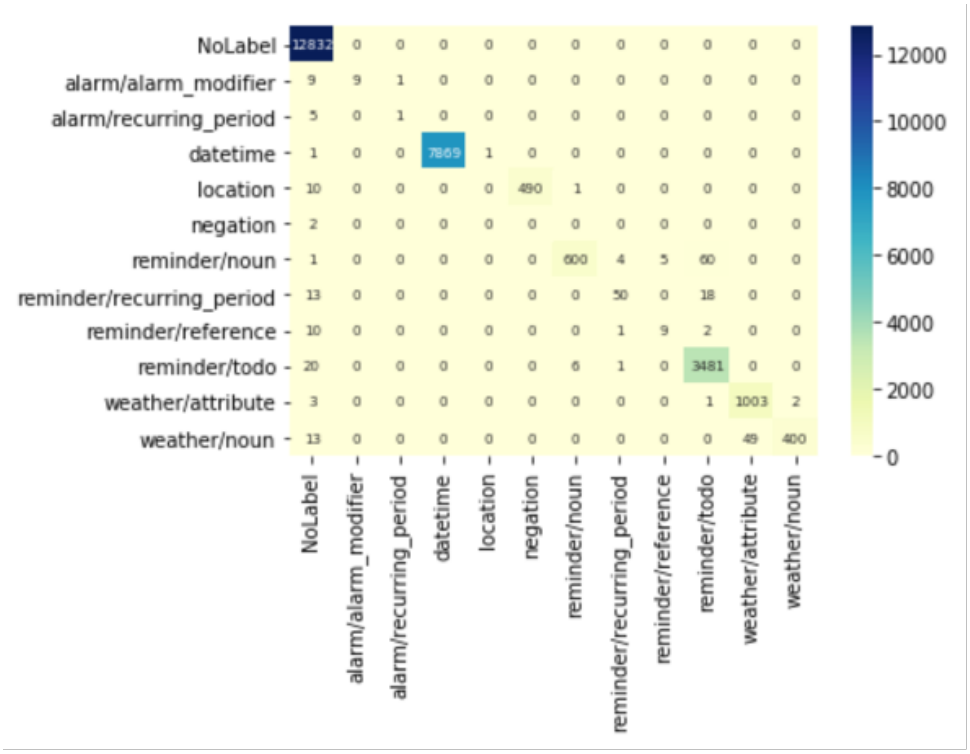

Figure 11. Confusion matrix for true and predicted Slot types by m-BERT for Intent Classification, LR=2e-4 for Spanish.

As for models' performances on Spanish data, we can notice that the results are slot-by-slot better than the Thai data's results below, but still most minor classes were mostly mislabeled and the problem connected with unbalanced data is more obvious for Slot Filling models' results than for previously analysed Intent Classification models' performances. Slot Filling models for Thai data did not receive enough data, but still the results are quite competitive. Minor classes (we need to say it is less than 20 or even 10 utterances for fine-tuning) were fully

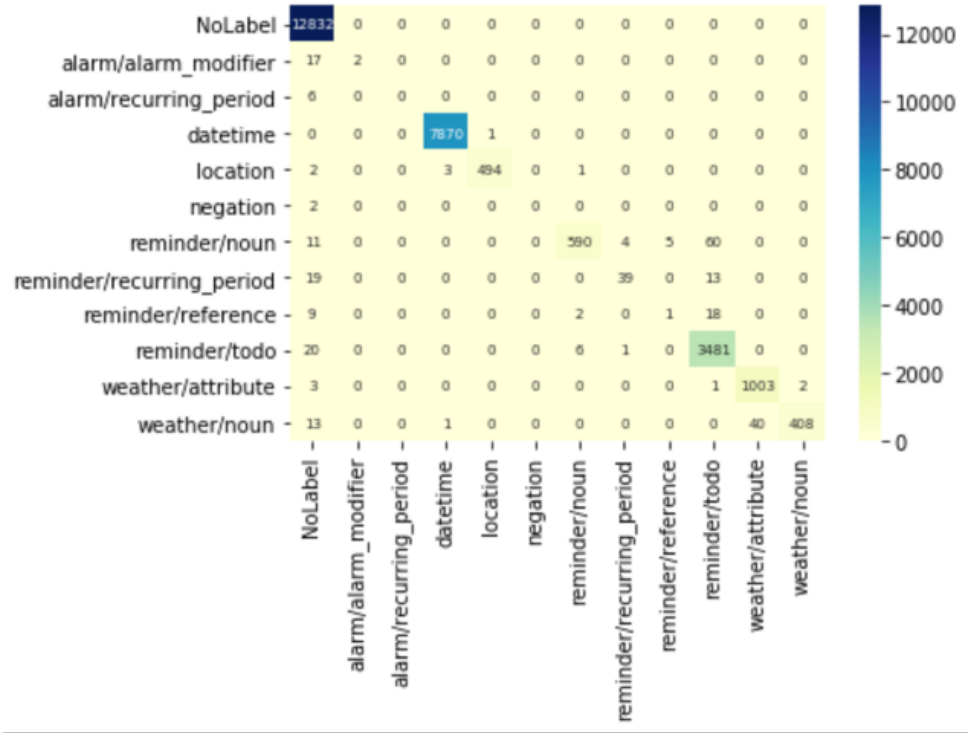

Figure 12. Confusion matrix for true and predicted Slot types by joint XLM-Roberta for Spanish.

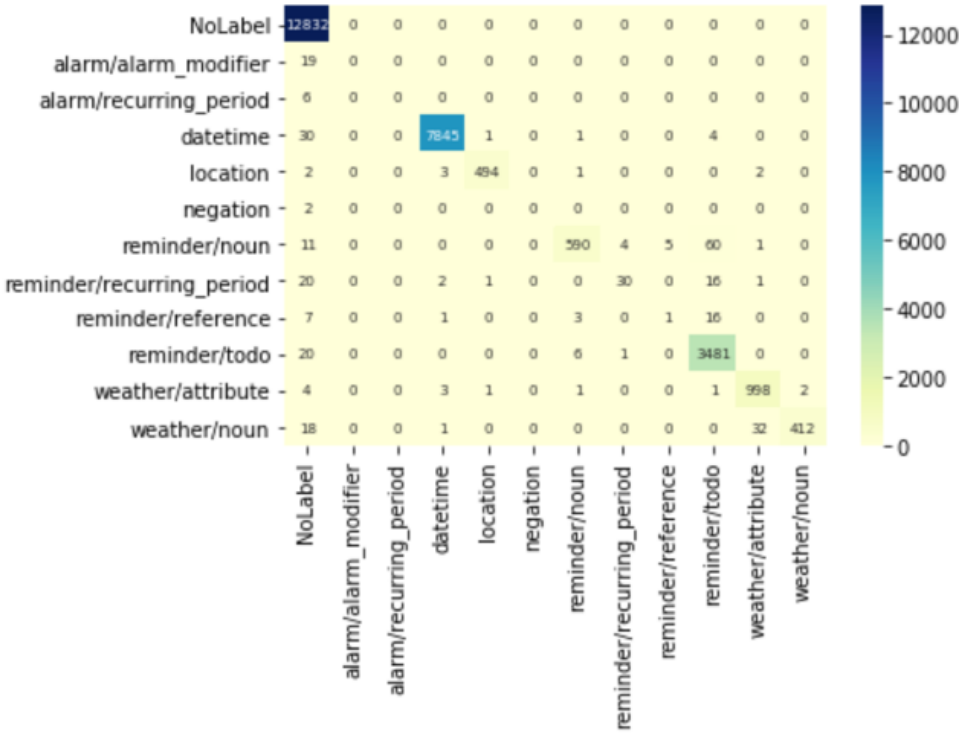

Figure 13. Confusion matrix for true and predicted Slot types by joint BERT for Spanish.

or mostly missclassified, but in classes with enough data for fine-tuning all used metrics are high.

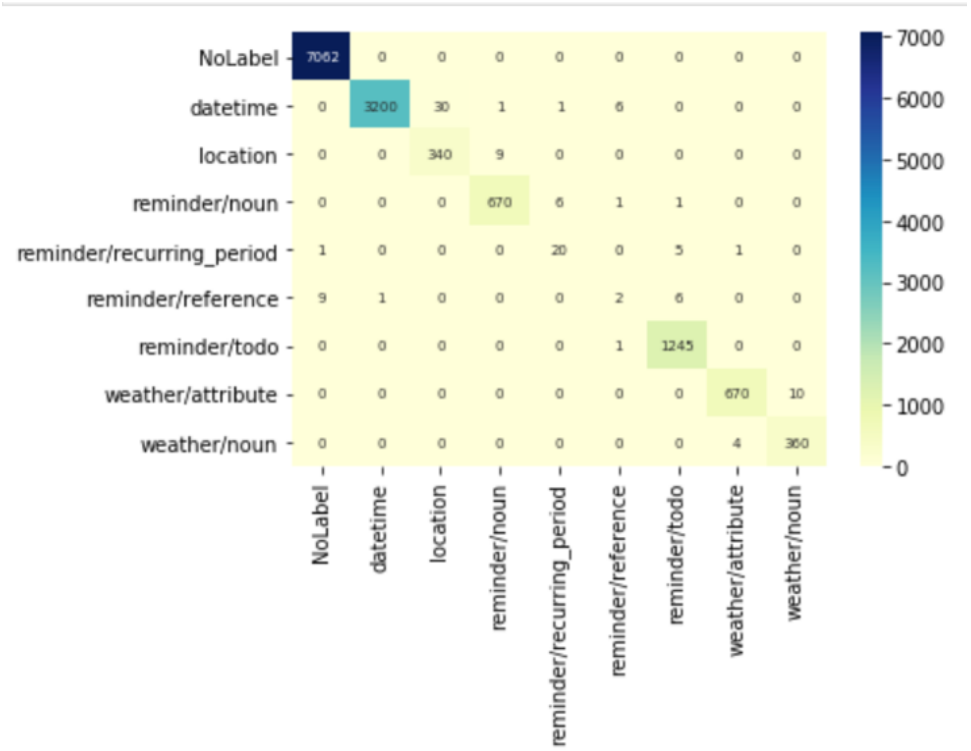

Figure 14. Confusion matrix for true and predicted Slot types by m-BERT for Intent Classification, LR=2e-4 for Thai.

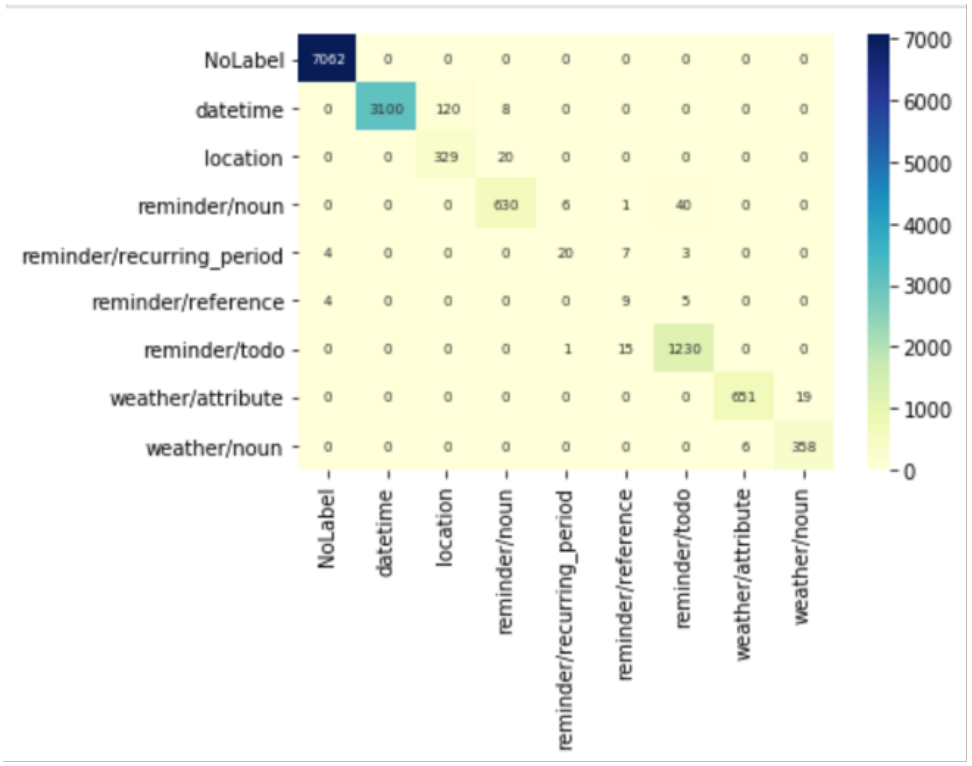

Figure 15. Confusion matrix for true and predicted Slot types by joint XLM-Roberta for Thai.

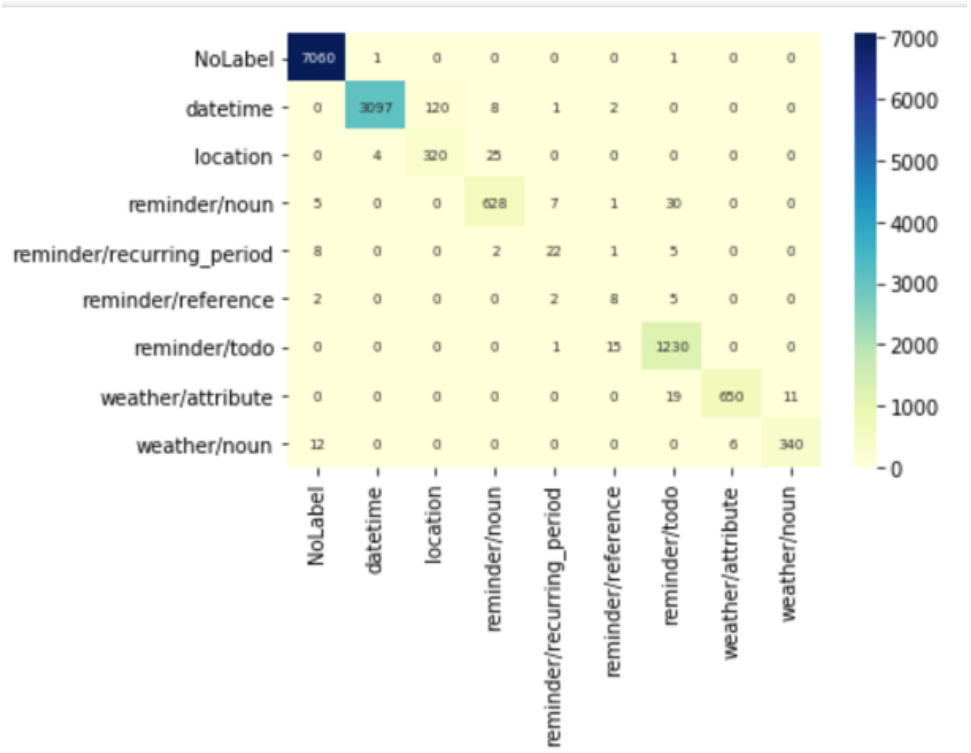

Figure 16. Confusion matrix for true and predicted Slot types by joint BERT for Thai.

6 Future prospects

6.0.1 CRF

As it was made in (Zhou and Xu 2015) and (Pang et al. 2019) for similar tasks with sequence labelling the use of Conditional Random Field (CRF) might have improved the performance of slot filling. The addition of a CRF layer on top of the joint m-BERT or joint XLM-Roberta models might not have had significant or even any positive influence on models' results like, for example, in (Chen et al. 2019) and (Hardalov 2020) respectively, but the performances of joint m-BERT + CRF or XLM-Roberta + CRF might have still been comparable and almost as high as the same models without CRF. As it was said before, we did not conduct such experiments due to computational limits as both models are quite heavy and long for converge.

6.0.2 More epochs and batch sizes

As in many works it was done, like in (Gupta et al. 2015) and (Schuster et al. 2019), the more epochs a model is trained the better results might be gained (except overfitting). In Schuster (2019) the best results were achieved after 282 epochs and a batch size of 582, in (Chen et al. 2019) after 30 epochs the joint BERT model gained high accuracy and F1-scores. So, for XLM-Roberta especially, the increase of epochs for model training may improve model's performance. Finding the most appropriate batch size can also be useful in order to boost models' work.

6.0.3 Unfreeze layers

In (Jaejun 2019) it was shown that the less pre-trained model's layers are frozen, the better model's performance is likely to be, so this conclusion could have been followed and it could have helped improve the conducted experiments. But due to recourse and energy limits, the method of (Jaejun 2019) could not have been conducted as the more layers we unfreeze, the longer and harder it is

to compute.

7 Conclusion

We proposed joint m-BERT-base, joint XLM-Roberta-base and two separate Token Classification + Text Classification models for solving such Dialog Oriented Task as simultaneous Intent Classification and Slot Filling in lower-recourse data in Thai and Spanish languages. Best results were shown by two separate models, though it took us much more time and it was considerably more resource-consuming than the use of the mentioned joint-fashioned models. The second best result was gained by XLM-Roberta, though some minor classes were fully mislabelled by major classes (mostly) of the same domain. All models' results on Thai data are quite lower than on Spanish one, as we consider, due to genealogical relationship of Spanish-and-English and Thai-and-English respectively or small Thai dataset size, which could be augmented or crow-sourced. Future prospects of improvement current work results could be done with unfreezing more layers in pretrained models or increase of epochs for model training. Source code can be found here: https://github.com/MatyashDare/CourseWork2021.

8 References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, Jimmy Lin. (2019). Rethinking complex neural network architectures for document classification. 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 4046–4051.

Alan Akbik, Duncan Blythe, Roland Vollgraf (2018). Contextual string embeddings for sequence labeling. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1638–1649.

Bob Carpenter (2009). Coding Chunkers as Taggers: IO, BIO, BMEWO, and BMEWO+. posted on October 14, 2009 at 2:47 pm and is filed under Carp's Blog, LingPipe in Use, LingPipe News. Qian Chen, Zhu Zhuo, Wen Wang (2019). BERT for Joint Intent Classification and Slot Filling. CoRR, abs/1902.10909.

Alice Coucke, Alaa Saade, Adrien Ball, The odore Bluche, Alexandre Caulier, David Leroy, Cle ment Doumouro, Thibault Gisselbrecht, Francesco Calt- agirone, Thibaut Lavril, Mae l Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private- by-design voice interfaces. CoRR, abs/1805.10190.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov (2020). Unsupervised Cross-lingual Representation Learning at Scale. arXiv:1911.02116v2 [cs.CL] 8 Apr 2020.

Suman Dowlagar, Radhika Mamidi. (2021) Multilingual Pre-Trained Transformers and Convolutional NN Classification Models for Technical Domain Identification.

arXiv:2101.09012v1 [cs.CL] 22 Jan 2021.

Suyog Gupta, Wei Zhang, Fei Wang. 2015. Model Accuracy and Runtime Tradeoff in Distributed Deep Learning: A Systematic Study. arXiv:1509.04210

Momchil Hardalov, Ivan Koychev, Preslav Nakov. 2020. Enriched Pre-trained Transformers for Joint Slot Filling and Intent Detection. arXiv:2004.14848v1 [cs.CL] 30 Apr 2020.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, Jason Mars. 2019. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019), pages 1311–1316.

Jaejun Lee, Raphael Tang, Jimmy Lin. 2019. What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning. rXiv:1911.03090v1 [cs.CL] 8 Nov 2019

Maddalen López de Lacalle, Xabier Saralegi, Iñaki San Vicente. (2020). Building a Task-oriented Dialog System for languages with no training data: the Case for Basque. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2796–2802.

Ramshaw and Marcus (1995). Text Chunking using Transformation-Based Learning. ACL Third Workshop on Very Large Corpora, June 1995, pp. 82-94.Na Pang, Li Qian, Weimin Lyu, Jin-Dong Yang. 2019. Transfer Learning for Scientific Data Chain Extraction in Small Chemical Corpus with BERT-CRF Model. arXiv:1905.05615 [cs.CL].

Telmo Pires, Eva Schlinger, Dan Garrette. 2019. How multilingual is Multilingual BERT? arXiv:1906.01502v1 [cs.CL] 4 Jun 2019.

Sebastian Schuster, Sonal Gupta, Rushin Shah, Mike Lewis (2019). Cross-Lingual Transfer Learning for Multilingual Task Oriented Dialog. arXiv: 1810.13327 [cs.CL] 1 Apr 2019.

Gokhan Tur Dilek Hakkani-Tu r Larry Heck. (2010). What is left to be understood in ATIS? IEEE Spoken Language Technology Workshop (2010), pages 19-24.

Jie Zhou, Wei Xu. End-to-end learning of semantic role labeling using recurrent neural networks. (2015) In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1127–1137.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692

9 Appendix

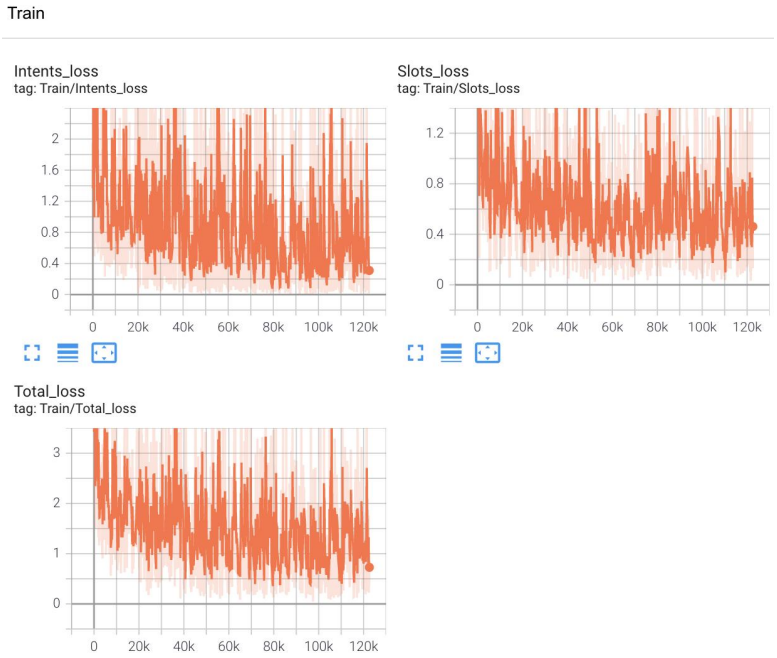

Figure 23. Train losses in 15 epochs during fine-tuning joint m-BERT.

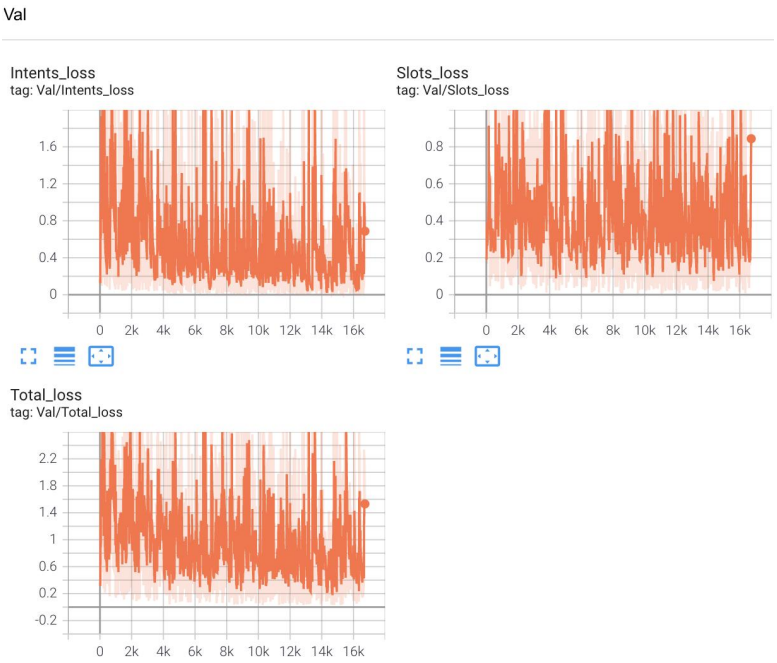

Figure 24. Validation losses in 15 epochs during fine-tuning joint m-BERT.

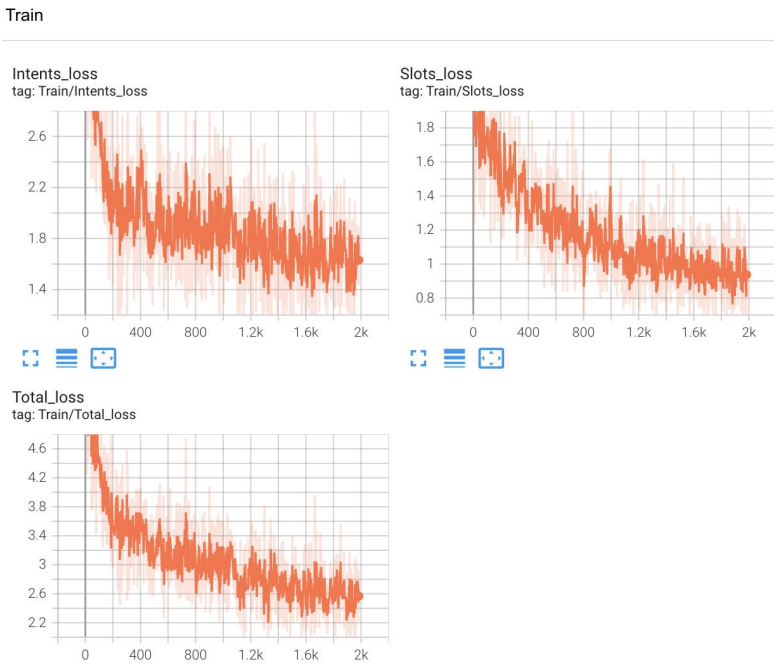

Figure 25. Train losses in 15 epochs during fine-tuning joint m-BERT on Spanish data.

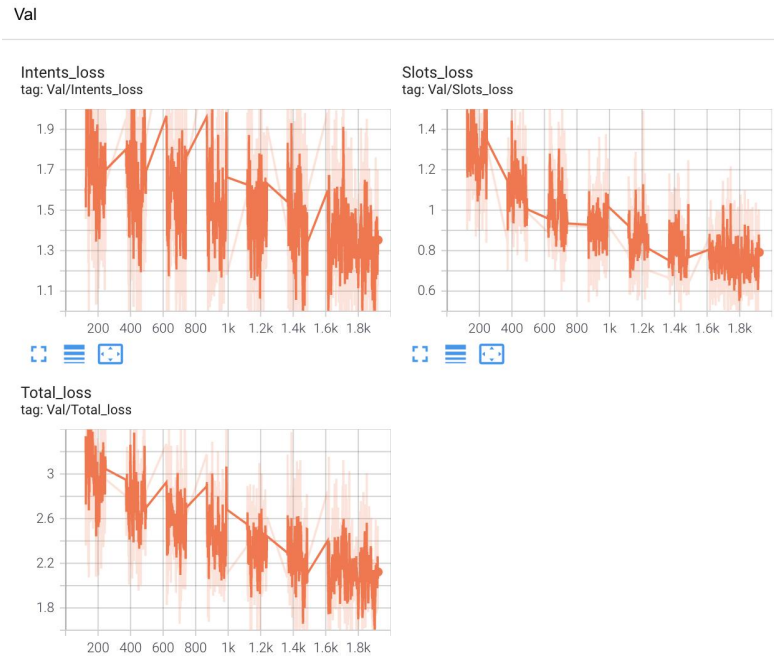

Figure 26. Validation losses in 15 epochs during fine-tuning joint m-BERT on Spanish data.

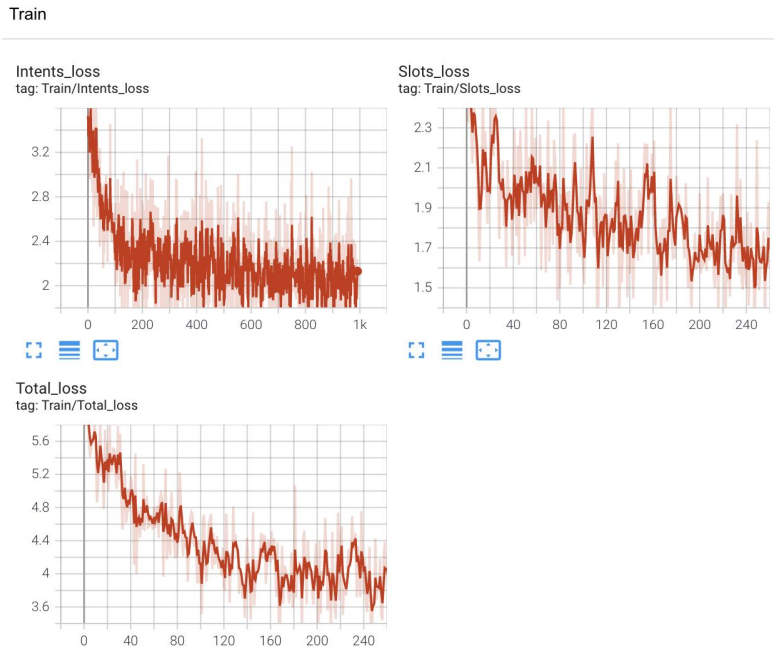

Figure 27. Train losses in 15 epochs during fine-tuning joint XLM-Roberta on Spanish data.

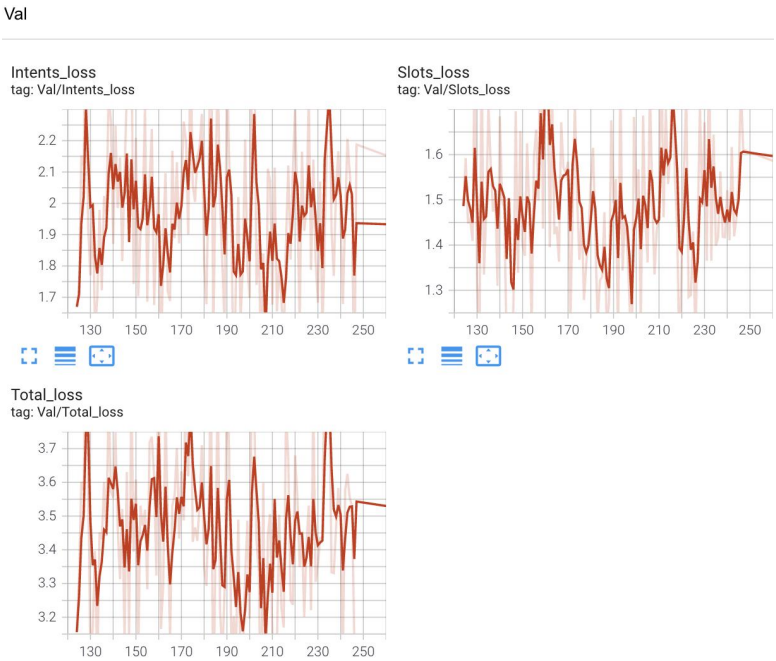

Figure 28. Validation losses in 15 epochs during fine-tuning joint XLM-Roberta on Spanish data.