HW-TSC at SemEval-2022 Task 3: A Unified Approach Fine-tuned on Multilingual Pretrained Model for PreTENS

Yinglu Li, Min Zhang, Xiaosong Qiao, Minghan Wang
Hao Yang, Shimin Tao, Ying Qin
Huawei Translation Services Center, Beijing, China
{liyinglu, zhangmin186, qiaoxiaosong, wangminghan,
yanghao30, taoshimin, qinying}@huawei.com

Abstract

In the paper, we describe a unified system for task 3 of SemEval-2022. The task aims to recognize the semantic structures of sentences by providing two nominal arguments and to evaluate the degree of taxonomic relations. We utilise the strategy that adding language prefix tag in the training set, which is effective for the model. We split the training set to avoid the translation information to be learnt by the model. For the task, we propose a unified model fine-tuned on the multilingual pretrained model, XLM-RoBERTa. The model performs well in subtask 1 (the binary classification subtask). In order to verify whether our model could also perform better in subtask 2 (the regression subtask), the ranking score is transformed into classification labels by an up-sampling strategy. With the ensemble strategy, the performance of our model can also be improved. As a result, the model obtained the second place for subtask 1 and subtask 2 in the competition evaluation.

1 Introduction

As we all know, the proposal of BERT(Devlin et al., 2018; Vaswani et al., 2017), which is based on masked language modeling, is a huge milestone in the history of natural language understanding (Peters et al., 2018; Schuster et al., 2019). Compared with various language representation models, BERT successfully pushes the GLUE score at 7.7 points absolute improvement(Devlin et al., 2018). Soon, different types of language models such as XLNet(You et al., 2019), RoBERTa(Liu et al., 2019),mBert (Devlin et al., 2018; Radford and Narasimhan, 2018) and XLM (Lample and Conneau, 2019) are also proposed. Compared with some strong monolingual models introduced above, XLM-RoBERTa is more competitive on the GLUE and XNLI benchmarks(Conneau et al., 2019). There are lots of pretrained models proposed in recent years, which are capable of learning the implicit knowledge. Some works have proved that the neural language model has learnt the implicit linguistic knowledge and this knowledge can significantly affect the predictions through fine-tuning(Miaschi et al., 2020; Puccetti et al., 2021). It has been believed that the pretrained models trained on Wikipedia and other datasets already have some implicit knowledge. Implicit knowledge could be classified into two categories: the connection between two objects, and the implicit logic and syntax behind the sentence. It means, we need to build a model which is capable of verifying the rationality and reliability of sentences, testing whether the latent knowledge can be expressed explicitly.

Presupposed taxonomy is a kind of concept in computational linguistics. Two arguments could have several different taxonomy relationships. For example, the sentence "I like piano, but not the instrument" is a classic pattern which contains two arguments. In the transition sentence, "piano" and "instrument" have a taxonomic relation, so the conclusion could be drawn that the sentence is implausible and unacceptable. Similarly, there are a set of sentences that could have such contradiction in the competition. The goal of the SemEval-2022 Task 3 (Zamparelli et al., 2022) is to recognize the presupposed taxonomy relation between two nouns, which could be a complex linguistic problem. In order to obtain the skill, the model needs to understand the implicit meaning of the sentences as well as whether the presupposed taxonomy relation exists in one pair of entities.

ID	Sentence	Labels
5155	*I like **ham**, but not **fish**.*	1
2560	*I like **restaurants**, and **clerks** too.*	1
3711	*I like **jewlry** more than **skirts**.*	1
4481	*I like **scientists** more than **geneticists**.*	0
3104	*I do not like **seafood**, I prefer **salmon**.*	0

Table 1: Dataset Samples for subtask 1

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 291 - 297
July 14-15, 2022 ©2022 Association for Computational Linguistics

2 Task Description

Along the lines of ideas above, there are several representative pretained-models in the competition. To be specific, we tried commonly used pretrained models including Bert, Roberta-large, and multilingual language models like XLMR. At the same time, we suspect that the former model can be generalized to subtask 2. The final ranking result supports our hypothesis strongly.

Subtask 1 is a binary classification task aiming at predicting the acceptability of sentences (A(1) VS UA(0)). There are two parameters that have the presupposed taxonomy relation, followed by the label "0"(the semantic relation is not acceptable) or "1" (the semantic relation is acceptable). Note that label "0" stands for the contradictory sentences, and label "1" stands for the plausible sentence. Some samples are shown in table 1, in which most of them obey a similar pattern: "I like A, but not B" or "I do not like A, I prefer B...", etc. The bold portion shows the key entities in the sentence. It is obvious that the taxonomies relation between two entities stands for some implicit information, which should be learnt by our model.

In this multilingual task, three languages (French, Italian and English) are included. Table 1 only presents the English samples and three datasets express completely consistent arguments. The same id indicates the same meaning in a different language.

On the contrary, subtask 2 is a regression task aiming at predicting the degree of Acceptance in a seven Likert-scale. The only difference between the dataset of subtask 1 and subtask 2 is the label. In subtask 2, the label(score) is a float number between 1 and 7, which indicates the acceptable degree of the sentence. Some English data samples are shown in table 2. Note that the sentence with higher score is more reasonable than the sentence with lower one. For instance, the sentence "I like seafood, but not crabs" sounds like a correct sentence in daily life. But the sentence "I like beef, an interesting type of caviar" is a contradictory sentence without any doubt, because the beef is not a caviar.

ID	Sentence	Scores
261	*I like **beef**, an interesting type of **caviar**.*	1.09
440	*I like **trees** more than **grass**.*	5.64
207	*I like **shrubs**, an interesting type of **fir**.*	2.67
60	*I like **trees** but not **birches**.*	1.83
436	*I like **oaks** more than **grass**.*	5.83
104	*I like **seafood**, but not **crabs**.*	6.42

Table 2: Dataset Samples for subtask 2

In addition, Table 3 provides the size of the train and test set. Obviously, the size of dataset for subtask 2 is much smaller than that for subtask 1. So it becomes important to expand the size of dataset of subtask 2. There are two main subtasks in PreTENS (Presupposed Taxonomies: Evaluating Neural Network Semantics).

Task	**Type**	**Language**	**Train size**	**Test size**
Subtask 1	Classification	En	5840	14560
		Fr	5840	14560
		It	5840	14560
Subtask 2	Regression	En	526	1009
		Fr	526	1009
		It	526	1009

Table 3: Task Dataset Description

3 System

3.1 Data Process for subtask 1

Based on a suitable single model and adaptive finetuning models which have learnt enough implicit information, it becomes possible to express the explicit knowledge(taxonomies) for a model.

The baseline model provided by competition reaches accuracy at 0.8, which is an amazing result. We also find the training set is extremely unbalanced in the proportion of positive and negative samples for some patterns, as we can see in Table 9. So we tried to adjust the proportion of those unbalanced patterns. However, the accuracy decreased a lot after the adjustment. This result shows that the model is actually learning the proportion of labels in the dataset instead of the implicit information. In order to solve the problem in the competition, we need to choose some models which are capable of learning implicit knowledge.

3.1.1 Language Tags

Considering the three languages in our competition, we apply the prefix tag to indicate which language the sentence is in. The XLM-RoBERTa is a multilingual model so the input data consists of mixed sentences in three languages.

As we can see from the figure, angle brackets and language abbreviation is used as the uniform prefixes. Specifically, <fr> stands for French, <en> stands for English and <it> stands for Italian. The

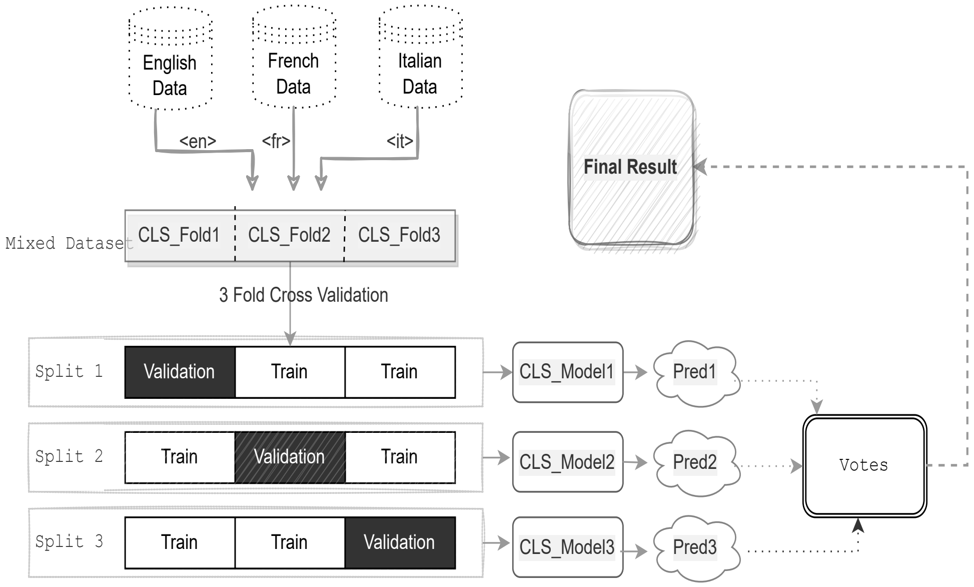

Figure 1: Ensemble classification models

Raw Sentence	**Language Tag**	**New Sentence**
I like jewelry more than skirts	en	***<en>****I like jewelry more than skirts.*
J' aime les bijoux plus que les jupes .	fr	***<fr>****J' aime les bijoux plus que les jupes .*
Amo i gioielli più delle gonne .	it	***<it>****Amo i gioielli più delle gonne .*

Table 4: Adding Language Tags for sentences with the same id

strategy helps to provide some linguistic information to the model artificially. Obviously, there are only linguistic differences between sentences with the same id, and their essential meanings are consistent. Imagining the situation without artificially added labels, there might be some difficulties in identifying sentence pairs with the same id but in different languages.

3.1.2 Dataset Split

We also apply the three-fold cross-validation in the competition. After further fine-tuning, the model has achieved good results on subtask 1. In the task description chapter, we mentioned that the same meaning is expressed in different languages in the dataset, which can be told by ids. In other words, sentences with same meanings in different languages share the common id.

We find that sentences with the same meaning might appear in the training set and the test set respectively in a multilingual language model. So, the model might learn the meaning of translation which should be avoided in our tasks. Therefore, in this section, we divide the data and put sentences with the same ID into the same set to prevent the model from learning the translation information. With such treatment, it can be ensured that the model learns implicit knowledge instead of cheating with translation.

3.2 A Unified Model for Subtask 2

The Situation becomes difficult in subtask 2. Considering the high similarity between datasets used by subtask 2 and subtask 1, and the size of the latter is much smaller than the former, it is important to use the data augmentation. In order to get good performance with the model in subtask 1, we make some efforts to transform the data provided in subtask 2 by using the method used above.

In other words, our method can be easily transferred from subtask 1 to subtask 2, which not only reduces the workload and training time but also makes the prediction more reliable because of the extension of datasets.

From table 1 to 3 which describes the sample and size of datasets, the label in subtask 2 are real values from 1 to 7. The label is a score used to assess the reasonableness of the sentence. Com-

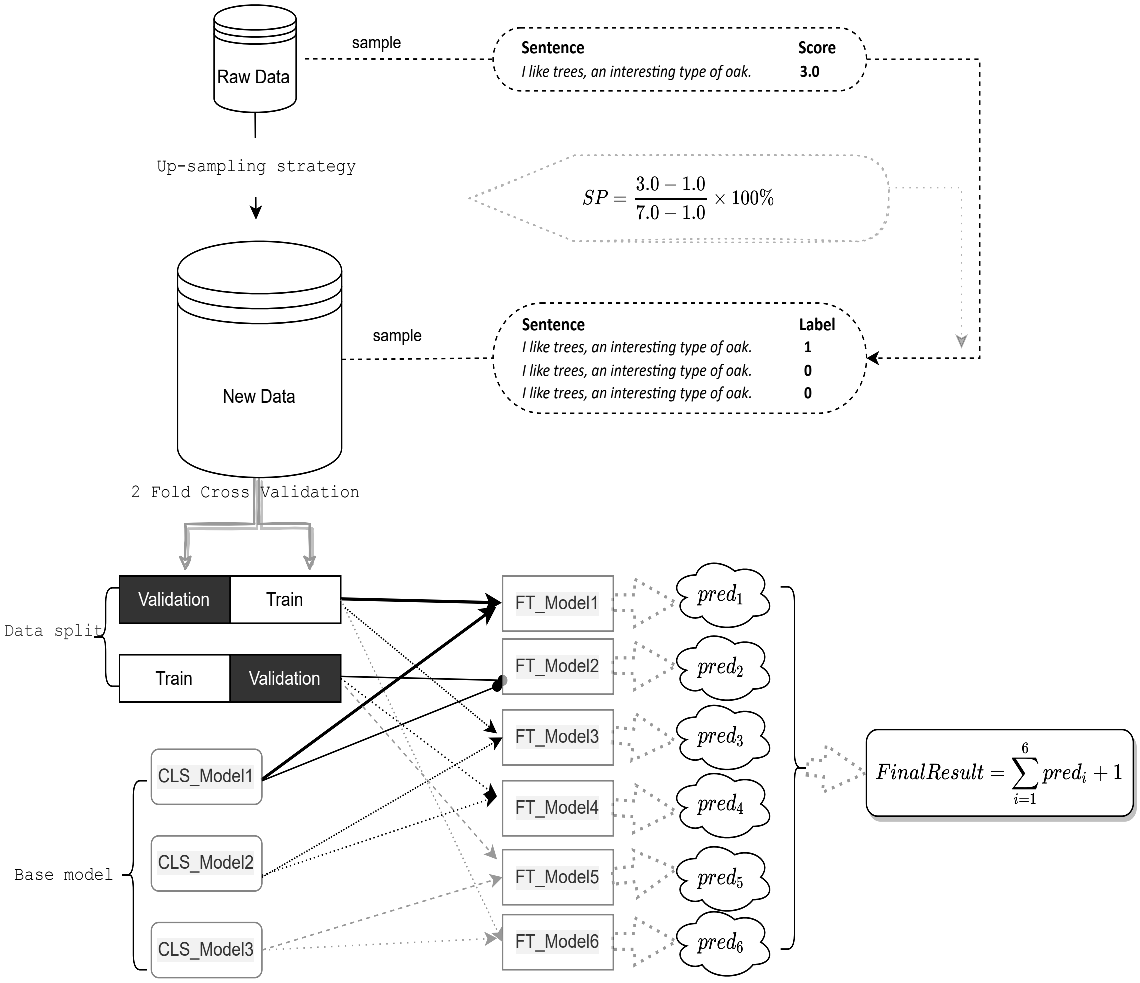

Figure 2: Ensemble regression models

paring tags in two subtasks, it is easy to imagine if sampling with the principle of score as probability, the meaning of the original score could be involved in the new label.

For each sentence from raw dataset, the score would be transformed to label according to the sampling possibility.

Sampling Probability(sp) is defined in the Eq 1, which depends on the ub(upper-bound of score) and lb(lower-bound of score).

$$sp = \frac{score - 1.0}{ub - lb} \times 100\% \qquad (1)$$

The up-sampling strategy is described in figure 2, which uses the expansion coefficient equalling 3 as an example. As for the sentence "I like trees, an interesting type of oak.", the sampling probability could be calculated according to the formula 1, which equals 1/3. Consequently, the raw sentence is duplicated to three same sentences with the label (1, 0, 0), which comes from the sampling strategy.

In data processing, we choose 10 as the expansion coefficient while converting the original score into sampling probability. 10 times the size of datasets, an expansion dataset is obtained in this way.

We use the new expansion dataset to train the model from subtask 1. At the same time, since there are three models obtained from subtask 1, and we split the new dataset into two copies, obviously $3 \times 2 = 6$ models have been used for the ensemble.

Considering the lack of regression data, we make some attempts to up-sampling the regression data. More specifically, the difference between classification data and regression data is the label. The former could be an integer while the latter could be a float between 1 and 7.

Apart from the label, the sentence have the similar pattern and entity. So it might be effective to reuse the model in subtask 1. Each original sentence is duplicated to ten same sentences with the label which might be 0 or 1 depending on the possi-

Model	Train Set	Dev Set	Dev Acc	Global Rank
CLS_Model1	CLS_fold1, 2	CLS_fold3	0.9426	88.4185
CLS_Model2	CLS_fold2, 3	CLS_fold1	0.9350	-
CLS_Model3	CLS_fold1, 3	CLS_fold2	0.9140	-
Ensemble	-	-	-	92.7968

Table 5: Ensemble Strategy for subtask 1

Model	EN Macro	EN F1	FR F1 Macro	FR F1	IT F1 Macro	IT F1
CLS_Model1	89.0199	89.0200	89.1187	89.1190	87.1169	87.1170
Ensemble	93.0410	92.5830	93.0116	92.5470	92.3388	91.8020

Table 6: Detailed Results of ensemble models for subtask 1

bility. Obviously, the possibility and the regression score is linearly and positively correlated. After obtaining an expanded dataset which has 10 times the size of original dataset of subtask 2, our model could learn some new knowledge in training.

3.3 Ensemble

Through the step described in section 3.1.2, we obtained six models for subtask 1. According to the requirements described in the competition, subtask 2 would be evaluated by the Spearman correlation coefficient, which is a metric used to express distribution trends. In other words, when a distribution is appropriately scaled, the magnitude of the Spearman correlation coefficient would remain the same. This inspires us to directly superimpose the results of the six models for ensemble. To keep the final result size between 1 and 7, we shift the score of the result by 1.

For subtask 1, the dataset is split into three folds for cross-validation and then three models are obtained. Since the target of subtask 1 is to figure out whether the sentence is plausible or not, we use the voting strategy for ensemble. Specifically, three models could have three decisions about the label of one sentence. Naturally, the voting ensemble strategy could be applied in the stage. The final result would be the majority decision.

For subtask 2, the situation is different. It has been cleared that there are three models in the classification task. Based on each model, we use the strategy of up-sampling to expand our dataset and transform it into the classification one.

4 Results and Analysis

Table 5 and Table 6 illustrate the ensemble strategy and ensemble results for subtask 1. The accuracy is the metric to evaluate the result in subtask 1. The results of subtask 2 are illustrated in table 7 and table 8, and the metric is Spearman. Because of the relative measurement of the metric, there are some small shifts in our predictions actually bring no influence to the value of Spearman. Those results indicate that our model has an outstanding improvement compared to baselines.

Except for the experiments introduced above, there were still a lot of directions we tested, but not all of them were useful.

1) From the perspective of the pattern, we found the positive and negative patterns are unbalanced in the dataset. From the Table 9, there are some examples that appear in the dataset. The number of positive samples and negative samples is very unbalanced. So, we balanced the dataset and trained the model. However, this strategy was useless for our model, and it even decreased the accuracy of the model. We suspect that the test set is also an unbalanced dataset.

2) We tried to extract the entities of each sample, and concatenated the embeddings of entities and embeddings of "<CLS>". Eq 2 shows the change between origin embedding and the new embedding.

$$[EA, EB] -> [CLS, EA, EB] \qquad (2)$$

Then, we classified it with a binary classifier (Softmax). But there were no improvements in the performance.

Model	Base Model	Train Set	Dev Set	Dev Spearman
FT_Model1	CLS_Model1	Reg_fold1	Reg_fold2	0.6871
FT_Model2	CLS_Model1	Reg_fold2	Reg_fold1	0.7429
FT_Model3	CLS_Model2	Reg_fold1	Reg_fold2	0.7038
FT_Model4	CLS_Model2	Reg_fold2	Reg_fold1	0.7382
FT_Model5	CLS_Model3	Reg_fold1	Reg_fold2	0.7242
FT_Model6	CLS_Model3	Reg_fold2	Reg_fold1	0.6993
Ensemble	-	-	-	0.7582

Table 7: Ensemble Strategy for subtask 2

Model	Global Rank	RHO(IT)	RHO(FR)	RHO(EN)
FT_Model1	0.6871	0.7376	0.7986	0.6917
Ensemble	0.7572	0.7591	0.8050	0.7060

Table 8: Detailed Results of Ensemble Fine-tuned Model for subtask 2

3) We tried to extract all entities and capture all the related contents on Wikipedia and trained our language model based on the dataset. But it only brought limited improvements. Because Wikipedia is a very common dataset, the language model might be trained on the dataset before.

Pattern	Label 0/1
He trusts a , except his b.	0 / 12
He does not trust a , he prefers his b.	12 / 0
He likes a , an interesting type of b .	0 / 9

Table 9: Some patterns with unbalanced ratio of positive samples with negative samples.

5 Conclusion

In the experiment, we propose a solution for two subtasks of SemEval2022 Task 3. We illustrate the importance of data preprocessing. The data split and the use of the language tags both have a positive influence on the model performance. Considering the similarity of the dataset and the good performance of the model in subtask 1, we also explore the feasibility of a unified model. The unified model has great performance on both subtask 1 and subtask 2. In the future, there are some other directions that could be tried in the following explorations. We find that the model trained on a large dataset performs well, so, exploring large models might be an interesting direction.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, page arXiv:1810.04805.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. *arXiv e-prints*, page arXiv:1901.07291.

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

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv e-prints*, page arXiv:1802.05365.

Giovanni Puccetti, Alessio Miaschi, and Felice Dell'Orletta. 2021. How do BERT embeddings organize linguistic knowledge? In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep*

Learning Architectures, pages 48–57, Online. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pretraining.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *CoRR*, abs/1902.09492.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Yang You, Jing Li, Jonathan Hseu, Xiaodan Song, James Demmel, and Cho-Jui Hsieh. 2019. Reducing BERT pre-training time from 3 days to 76 minutes. *CoRR*, abs/1904.00962.

Roberto Zamparelli, Shammur A. Chowdhury, Dominique Brunato, Cristiano Chesi, Felice Dell'Orletta, Arid Hasan, and Giulia Venturi. 2022. Semeval-2022 task3 (pretens): Evaluating neural networks on presuppositional semantic knowledge. In *Proceeding of SEMEVAL 2022*.