SPDB Innovation Lab at SemEval-2022 Task 3: Recognize Appropriate Taxonomic Relations Between Two Nominal Arguments with ERNIE-M Model

Yue Zhou, Bowei Wei, Jianyu Liu, Yang Yang
Shanghai Pudong Development Bank
{zhouy93, weibw1, liujy27, yangy103}@spdb.com.cn

Abstract

Synonyms and antonym practices are the most common practices in our early childhood. It correlated our known words to a better place deep in our intuition. At the beginning of life for a machine, we would like to treat the machine as a baby and build a similar training for it as well to present a qualified performance. In this paper, we present an ensemble model for sentence logistics classification, which outperforms the state-of-art methods. Our approach essentially builds on two models including ERNIE-M and DeBERTaV3. With cross-validation and random seed tuning, we select the top performance models for the last soft ensemble and make them vote for the final answer, achieving the top 6 performance.

1 Introduction

Synonym, antonym and their relations from unstructured text are fundamental problems in information classification field. These problems can be decomposed into three subtasks: word extraction using regrex, relation extraction (Zelenko et al., 2003), (Bunescu and Mooney, 2005), and classifying the logistics between them. However, an end-to-end model, i.e. ERNIE-M model (Ouyang et al., 2020), is proposed to solve the three tasks.

Presupposed Taxonomies - Evaluating Neural-network Semantics (PreTENS) (Zamparelli et al., 2022) is a task to predict the acceptability of simple sentences containing constructions whose two arguments are presupposed to be or not to be in an ordered taxonomic relation. In this paper, we first present a simple approach with the ERNIE-M model to solve the task. Although the ERNIE-M model performs unexpectedly impressive, the model has poor robustness. Hence, the additional pre-trained model is introduced to solve the robustness problem. The latest model DeBERTaV3 (He et al., 2021) has outstanding performance on cross-linguistic tasks, which outperforms BERT and DeBERTa on many tasks. The proposed model consists of two parts: the basic ERNIE-M model and the pre-trained model DeBERTaV3. The DeBERTaV3 model shares the same pre-trained data with ERNIE-M called XNLI (Conneau et al., 2018), which can improve the performance and robustness as well. The DeBERTaV3 model is trained independently, which has significant improvement for English but somehow brought no improvement for other languages. Based on the above conclusion, we employ the DeBERTaV3 model for English-task only.

To better understand the effectiveness of the proposed model, we started a bunch of analyses. The first problem is the data-set limitation. Two additional datasets were imported, i.e., the translated dataset from Google translation which is translated from three languages, and the XNLI dataset. However, larger datasets don't lead to better performance. We compared the performance of the ERNIE-M model on four sets of data: the given data, the given data with translated data, the given data with XNLI augmentation, and the given data with both the translated data and XNLI data. We do the same experiments with the DeBERTaV3 model as well. The results show that the combination of ERNIE-M with all the three datasets and DeBERTaV3 with the given English data perform the best.

2 Related Work

Multilingual model ERNIE-M proposes a new training method that encourages the model to align the representation of multiple languages with monolingual corpora, to overcome the constraint that the parallel corpus size places on the model performance. There are two models in ERNIE-M which are Cross-Attention masked language mod-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 266 - 270

eling (CAMLM) and Back-Translation masked language modeling (BTMLM).

Cross-Attention masked language modeling (CAMLM) is to align cross-language semantic representations on parallel corpora. Then, the multilingual representation is enhanced with transferability learned from parallel corpora.

Back-Translation masked language modeling (BTMLM) is trained to generate pseudo-parallel sentences from monolingual sentences. The generated pairs are then used as the input of the model to further align the cross-lingual semantics, thus enhancing the multilingual representation.

DeBERTaV3 presents a new pre-trained language model, which improves the original DeBERTa model by replacing mask language modeling (MLM) with replaced token detection (RTD), a more sample-efficient pre-training task. They all come from an important field, multilingual models.

Since the related paper was published at the end of 2021, there are no similar tasks have been done and published.

3 Our Approach

In this section, we first introduce the methods to solving the multi-language problem and then present our work about improving the performance on uni-language. To extenuate over-fitting for a specific language, our team uses a multi-language ensemble learning strategy that includes a pre-trained language model and a multilingual language model. Based on the approach above, it makes the learned representation generalizable across languages and improves the performance in finding the suitable taxonomic relations in two nominal arguments.

3.1 Multilingual Language Model Training

Our key idea of solving multilingual language tasks is to learn the language invariant feature space shared among multiple languages. We tried multilingual masked language modeling (MMLM), translation language modeling (TLM), and cross-attention masked language modeling (CAMLM) have been tried. However, the scale of the parallel corpus is quite limited, which limits the performance of the model.

However, we found that using the transferability learned from parallel corpora to enhance the model's learning of large-scale monolingual corpora to enhance multilingual semantic representation can achieve a good effect. ERNIE-M does this by making the predictions of tokens depending on tokens in another language, but not on other tokens in this language. Therefore, we choose ERNIE-M as the baseline model for this task and explore on this basis to improve the prediction effect.

In the process of using multilingual language models, we mainly adopt random search to fine-tune the ERNIE-M model and data augmentation methods are used for model training. Cross-lingual natural language inference (XNLI) dataset is used and the English training set is translated to Italian (E2I set). Firstly, the English training set is combined with the French and E2I set. Then, the model is fine-tuned with the combined training set. Finally, the augmented task training set in three languages is adopted for fine-tune process.

3.2 Cross-validation

To improve the robustness of our model, our team apply cross-validation for training. Firstly, by using different random seeds, we divided the training set which included all three languages ten times. Through this process, we obtained 10 folds of data, which contain 15768 training samples and 1751 validation samples in each fold. During the fine-tuning process, we used random search to optimize hyper-parameters like epochs, learning rate, and batch size. By using F1-Score as our evaluation metric, the best model at all the ten-fold of training is saved. Finally, by making predictions on the test set, we save the mean of the probability of all ten best-saved models. This result is our final output of ERNIE-M. Cross-validation process is shown in Figure 1.

3.3 Pre-trained Language Model

To enhance the effect in a single language subtask, we consider using an enhanced mask decoder and a disentangled attention mechanism to improve the effect. DeBERTaV3 meets our needs by using Electra-style pre-training and gradient unwrapping embedding sharing. We have tried to use DeBERTaV3 for training in each single language subtask respectively.

3.4 Ensemble

By using the multilingual language model and pre-trained language model respectively, we have two groups of validation set results for each language. We adopt the mean of the best-saved models from ERNIE-M and DeBERTaV3 after making predictions on the validation set. After comparing the

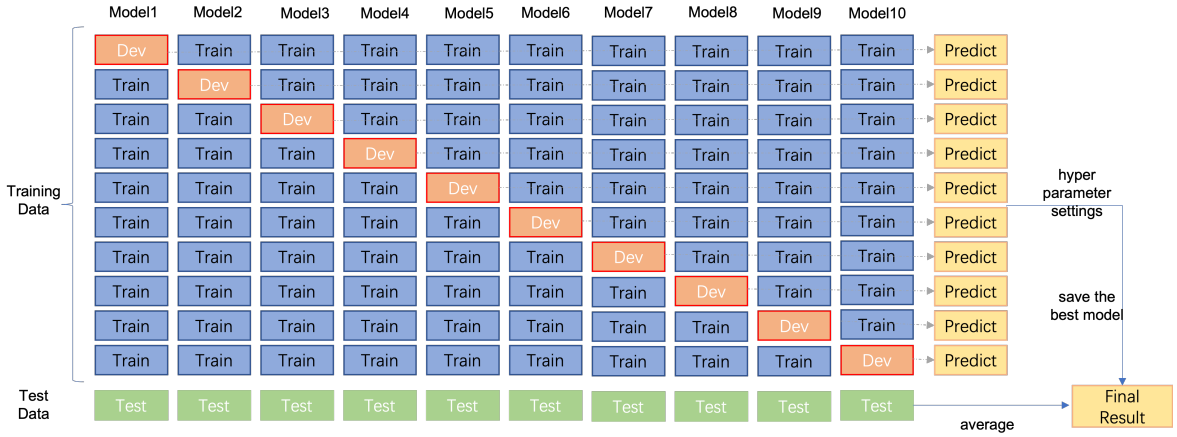

Figure 1: The process of 10-fold cross-validation and ensemble. The training set which includes all three languages is divided randomly 10 times by setting different random seeds. In each division, the training set is divided into 10 parts, of which 9 parts are respectively used as the training set and the remaining 1 part is used as the validation set. And finally, the average of all saved best models predicted on the test set is the final results.

combination result, we finally used different strategies in different languages. For the English subtask, we retain the strategy of merging the two types of models. For French and Italian subtasks, the result from cross-validation of the multilingual language model is used directly.

3.5 Data Augmentations

As the total number of labeled data in each language is only 5840, it's liable to overfit the training data even with pre-trained models. The overfitting phenomenon may be more significant than expected because the data is generated programmatically through manually verified templates. To increase the size of training data, we use the following data augmentation methods: 1) translate English data into French and Italian by using Baidu translate 2) translate English data into French and Italian by using Google translate 3) translate French and Italian data into English by using Google translate. We find that the augmentation can help delay the overfitting occurrence slightly, especially for large models.

4 Experiments

In this section, we first describe the dataset and our data preprocessing steps, and then we present the details of the experimental setup for subtask1.

4.1 Dataset

Our dataset comes from two parts.

The first part is the trial dataset released by organizers, which is composed of English, French and Italian. Each language contains 5838 sentences. Because the trail dataset provided by organizers is only 5838 in each language, to increase the amount of data and make the model better, we use Google translator and Baidu translator to translate the English dataset into French and Italian again. The use of two different translators also increases the diversity of data.

The other part is that we use the public dataset – XNLI. We use XNLI dataset because it is often used in similar cross-language tasks. The XNLI dataset contains a total of 15 languages, and each language contains 7500 pairs of data. We used the English and French datasets in this competition. Because the XNLI dataset itself does not contain Italian datasets, we translated the English dataset into Italian and then used the three languages in ERNIE-M model training.

4.2 Experiment Settings

In this task, we mainly use the ERNIE-M model and DeBERTaV3 model. The ERNIE-M model is composed of 24 layers, 1024 hidden, and 16 heads. In terms of parameter selection, we set a set of parameters, as Table 2 shows.

We set up 10000 times of ERNIE-M model training, in which the specific values of the above parameters are randomly selected according to the table1 at each training. And in each training process, the training method of 10 folds cross-validation is used.

The DeBERTaV3 model is composed of 12 layers and a hidden size of 768. It has only 86M

Parameter	Value1	Value2	Value3	Value4
batch size	8	16	32	
lr decay	0.8	0.85	0.9	0.95
rdrop	1.0	3.0	5.0	7.0
epoch	2	4		
learning rate	2e-5	3e-5	4e-5	5e-5
dropout	0.1	0.2		

Table 1: Parameter Setting. We set different specific values of different parameters according to some previous experience. Because the appropriate value is not fixed in different tasks, we choose to use the random combination of various values of the above parameters for model training, so as to find the most appropriate parameter value and obtain the optimal model result.

Training Methods	Language	Precesion	Recall	F1
ERNIE-M	English	0.8240	0.9547	0.8846
ERNIE-M	French	0.8185	0.9402	0.8751
ERNIE-M	Italian	0.8163	0.9307	0.8698
Ensemble Model	English	0.9266	0.9605	0.9432
Ensemble Model	French	0.8125	0.9489	0.8754
Ensemble Model	Italian	0.8081	0.9467	0.8719

Table 2: Results of different models. We select the best model from a large number of randomly generated parameter training models and compare it with the final best ensemble result. And we can see that the performance of the three languages has been improved.

backbone parameters with a vocabulary containing 128K tokens which introduces 98M parameters in the Embedding layer. And we set batch size to 8, learning rate to 2e-5, and epoch to 3.

4.3 Main Results

The best single model on the development set is the ERNIE-M LARGE. And the model that uses DeBERTaV3 doesn't perform well in French and Italian, so we just use the results on English data. The best ensemble model on the test set is trained on both the XNLI dataset and the trial dataset. The ensemble model obtained English test set F1 scores of 94.325, French test set F1 scores of 86.792, and Italian test set F1 scores of 88.807. The ensemble model achieves the F1 score of 94.325 in English data, the F1 score of 86.792 in French data, and the F1 score of 88.807 in the Italian data. The results are shown in Table 2.

For the comparative analysis of the results of using only ERNIE-M as the baseline model and the ensemble model, we can see that the improvement of the ensemble model in English is relatively obvious, but the improvement in Italian and French is very weak. We think this is due to the following reasons: Firstly, Italian is not included in the original XNLI dataset. In this task, we translate English into Italian. So to a certain extent, the understanding of English by the ERNIE-M model is increased. Secondly, because DeBERTaV3 performs well in English, we only use its results in English, So the results for Italian and French did not get a big boost. This also shows that using the ensemble model can indeed improve the prediction. In the future, we will explore ensemble models that can improve predictions in Italian and French.

5 Conclusion

To solve the problem of judging whether the meaning of a sentence is self-consistent in multilingual language tasks, that is, the problem raised in task 3, we propose an ensemble model using ERNIE-M and DeBERTaV3, and regard this problem as a binary classification problem. Furthermore, to solve the issue of the small dataset, we use various strategies, such as K-ford cross-validation, translating the dataset using different translators, and introducing an external dataset - XNLI, a dataset commonly used in multilingual problems. In future efforts, we plan to further improve our model from these aspects. The first is to enrich the data, especially Italian and French, to help the model learn better. The second is that we could train more models on standard fine-tuning, multi-step fine-tuning,

multi-task learning, or adversarial training. Then try to ensemble different models to gain a better performance.

References

Razvan C. Bunescu and Raymond J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, page 724–731, USA. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

P. He, J. Gao, and W. Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv e-prints*.

X. Ouyang, S. Wang, C. Pang, Y. Sun, and H. Wang. 2020. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora.

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*.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3(3):1083–1106.