SPDB Innovation Lab at SemEval-2022 Task 10: A Novel End-to-End Structured Sentiment Analysis Model based on the ERNIE-M

Yalong Jia, Zhenghui Ou, Yang Yang
Shanghai Pudong Development Bank
{jiayl2, ouzh, yangy103}@spdb.com.cn

Abstract

Sentiment analysis is a classical problem of natural language processing. SemEval 2022 sets a problem on the structured sentiment analysis in task 10, which is also a study-worthy topic in research area. In this paper, we propose a method which can predict structured sentiment information on multiple languages with limited data. The ERNIE-M pretrained language model is employed as a lingual feature extractor which works well on multiple language processing, followed by a graph parser as a opinion extractor. The method can predict structured sentiment information with high interpretability. We apply data augmentation as the given datasets are so small. Furthermore, we use K-fold cross-validation and DeBERTaV3 pretrained model as extra English embedding generator to train multiple models as our ensemble strategies. Experimental results show that the proposed model has considerable performance on both monolingual and cross-lingual tasks.

1 Introduction

Sentiment analysis (Liu, 2012) is widely used in many aspects of computer science nowadays, such as human computer interaction, lingual feature extraction, etc. Affective computing techniques enable us to explore the sentiment message, which depicts the preference, emotion or even idea of people, behind the sentence itself.

Sentiment analysis task can be classified into text level, sentence level, entity level and opinion tuple level. The goal of text or sentence level sentiment analysis is to predict sentiments of given documents or sentences. On the other hand, entity level sentiment analysis needs to consider sentiments between each entity in given sentence. Furthermore, a sentence may contain multiple entities with different sentiments (positive or negative). Therefore, in comparison to entity level sentiment analysis, opinion tuple level sentiment analysis requires additional extraction of relations between entities and opinions.

Structured sentiment analysis (Barnes et al., 2022) is a task to predict a sentiment graph for given sentences. It can be theoretically cast as an information extraction problem in which one attempts to find all of the opinion tuples $O = O_1, ..., O_n$ in a text. As we can see in Figure 1, each opinion O_i is a tuple (h, t, e, p) where h is a holder who expresses a polarity p towards a target t through a sentiment expression e, implicitly defining pairwise relationships between elements of the same tuple.

The structure of the paper is as follows. Section 2 briefly reviews recent works on similar tasks; Section 3 describes our model structure in detail. Section 4 shows the analysis of the given datasets; Section 5 introduces our experimental setting and results. And finally in Section 6, we make a conclusion and give the ideas about future works.

2 Background

Dividing structured sentiment analysis into multiple subtasks is a traditional approach, by first identifying holders, targets and expressions through Named Entity Recognition (NER) module, then predicting relations among the entities. (Peng et al., 2020; Li et al., 2019) are baselines with good performance. On the other hand, the end-to-end sentiment analysis is a straight-forward approach, which directly extracts target and expression without splitting them into sub-tasks. (He et al., 2019) presents an interactive multi-task learning network(IMN) implemented by a series of a multi-layer CNN modules. Recently, (Barnes et al., 2021) cast the structured sentiment problem as the dependency graph parsing problem, and proposed a method that outperforms the SOTA(state-of-the-art) baselines on

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1401 - 1405

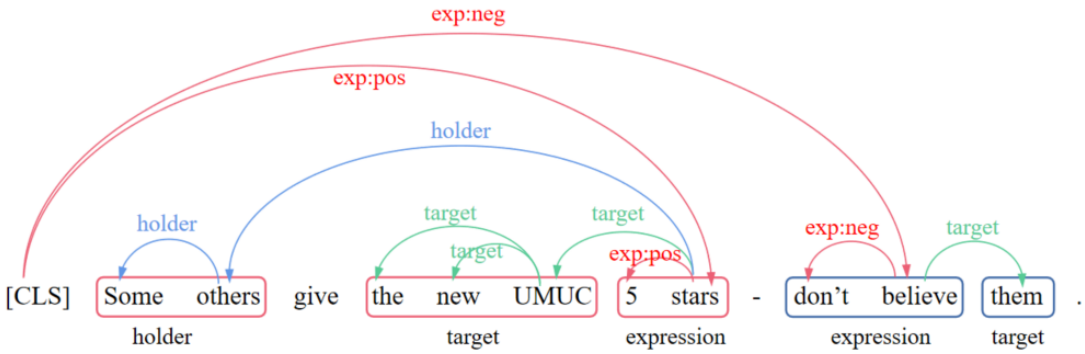

Figure 1: A structured sentiment graph is composed of a holder, target, sentiment expression, their relationships and a polarity attribute. Holders and targets can be none.

extensive experimental results.

In this task, 7 small datasets involving 5 different languages are given, which means our method needs to support multiple languages with limited data. Therefore, in this work, we use ERNIE-M pretrained model as word embedding generator, as it shows SOTA performances in various of NLP tasks in multiple languages. Because of the outstanding performance of (Barnes et al., 2021), we employ a similar network structure to extract the sentiment information.

3 Model structure

The model structure is similar to the head-final model structure in (Barnes et al., 2021), while we use a pretrained model in this work. A bert-style (ERNIE-M (Ouyang et al., 2020)) pretrained model takes sentences starting with "[CLS]" token as input. We connect "[CLS]" token with the last word of the expression, which is the root node of the tuple (h, t, e). The connection type is related to the sentiment polarity, "exp:pos" for positive polarity, "exp:neu" for neutral polarity, and "exp:neg" for negative corresponding to polarity, such that we are able to predict sentence polarity based on the connection type. As shown in Figure 1, a connection from "[CLS]" to "stars" and a connection from "[CLS]" to "believe" are established, with a connection type of "exp:pos" and "exp:neg" respectively. We describe the model structure in detail below.

For a given sentence $\overrightarrow{x} = (x_1, x_2, ..., x_n)$, where $x_i (1 \leq i \leq n)$ represents a single word. In this work, we use ERNIE-M pretrained model as a text feature extractor, which takes subword tokens as the model input. We apply subword-based tokenization on the input words.

As shown in Figure 2, we apply subword-based tokenization on the input sentence, getting $\overrightarrow{t} = (t_1, t_2, ..., t_m)$ for any $t_j (1 \leq j \leq m)$ representing a subword token. For instance, word "restful" will be split into "rest" and "##ful", where "##" indicates that the token is not the start of a word. And the process can be easily inversed by these special characters, such as we can restore ["Great", "and", "rest", "##ful", "place", "to", "stay", "."] to its original status ["Great", "and", "restful", "place", "to", "stay" "."]. After that, we input the subword tokens into the ERNIE-M model and get the embedding of the subword tokens. Then, we apply average pooling on the subword embeddings($\overrightarrow{v}$ in equation 1) which belong to a same original word to get the word representation $\overrightarrow{c} = (c_1, c_2, ..., c_n)$.

After obtaining the word representations of the sentence, we perform a position-wise feed-forward networks to obtain the representation of the heads and dependents, where heads represent head nodes, and the dependents represent follower nodes. Then we use Bilinear Attention Network (Kim et al., 2018) to calculate the pairwise correlation between each two words in the sentence.

$$\overrightarrow{v} = (v_1, ..., v_m) = \text{ERNIE-M}(t_1, ..., t_m) \quad (1)$$

$$h_i^{head} = FFN^{head}(c_i) \quad (2)$$

$$h_j^{dep} = FFN^{dep}(c_j) \quad (3)$$

$$score_{i,j} = Bilinear(h_i^{head}, h_j^{dep}) \quad (4)$$

In equation 4, we obtain a score matrix that indicates the relationship between each word in the input sentence pair-wisely, by passing head and

		ALL		Valid						Polarity		
		#	avg	#	avg	Opinion	Holder	Target	Exp	pos	neu	neg
Darmstadt_unis	train	2,253	19.99	681	21.25	806	63	806	806	340	102	364
	dev	232	18.09	82	20.21	98	9	98	98	29	15	54
MPQA	train	5,873	23.39	1,254	29.83	1,706	1,425	1,481	1,706	671	337	698
	dev	2,036	23.22	416	31.33	570	406	494	570	231	124	215
MultiBooked_ca	train	1,174	15.62	1,002	16.14	1,989	169	1,705	1,989	1,273	0	716
	dev	167	13.37	140	14.21	258	15	211	258	151	0	107
MultiBooked_eu	train	1,063	10.52	899	10.77	1,679	205	1,277	1,679	1,401	0	278
	dev	152	10.70	120	10.51	203	33	152	203	167	0	36
NoReC_fine	train	8,634	16.71	4,555	19.55	8,448	898	6,778	8,448	5,695	0	2,753
	dev	1,531	16.92	821	19.12	1,432	120	1,152	1,432	988	0	444
OpeNER_en	train	1,744	14.72	1,400	14.99	2,884	266	2679	2,884	2,101	0	783
	dev	249	14.22	198	14.98	400	49	371	400	284	0	116
OpeNER_es	train	1,438	17.13	1,252	17.58	3,042	176	2748	3,042	2,472	0	570
	dev	206	17.08	174	17.71	387	23	363	387	317	0	70

Table 1: Statistic of the given datasets, including the number of samples and average word count of all samples and valid samples(where opinions are not empty), as well as number of opinion, holder, target, expression and polarity in each dataset.

dependent embedding into a bilinear attention network, followed by a softmax layer. We apply cross-entropy loss as the loss function during model training.

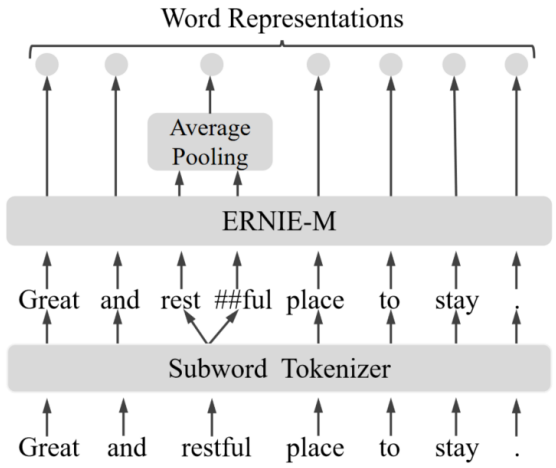

Figure 2: The process of generating the word representations.

4 Data

SemEval-2022 task 10 provides a total of 7 datasets: NoRec_fine(Øvrelid et al., 2020), MultiBooked_eu(Barnes et al., 2018), MultiBooked_ca(Barnes et al., 2018), OpeNER_es(Agerri et al., 2013), OpeNER_en(Agerri et al., 2013), MPQA(Wiebe et al., 2005), Darmstadt_unis(Toprak et al., 2010), involving different languages (en, ca, eu, no, es) and a couple of domains. MultiBooked_ca (Catalan), MultiBooked_eu (Basque), Opener_en(English), and Opener_es(Spanish) belong to hotel reviews domain. NoRec is a Norwegian dataset in literature, movies, video games, restaurants, music and theater domains. Darmstadt_unis is an English dataset in university domain, and MPQA is an English dataset in news domain. Our data analysis on the given datasets is shown in Table 1.

We can see that there are three different types of sentiment(positive, neutral, negative) in the opinions of Darmstadt_unis and MPQA, while others only contain Positive and Negative sentiments. In addition, we find that the number of expressions is always larger than or equal to the number of holders and targets for all datasets, which means there is at least one expression in each opinion. We can draw a conclusion that in a quadruple (h, t, e, p), both h and t may be missing, but e and p are not. Therefore, the expression(more precisely, the last word of expression) is defined as the root node of a (h, t, e) triplet.

5 Experiment and result

5.1 Data preprocessing

We process the labels of the original data into the format that the model required. Taking the sentence shown in Figure 1 as an example, the preprocessed result is shown in Table 2. We truncated the sentence due to the length of the sentence.

5.2 Experiments

In model validation period, we introduce two strategies, i.e., data augmentation and focal loss, which

	[CLS]	Some	others	give	the	new	UMUC	5	stars	.
[CLS]	-	-	-	-	-	-	-	-	exp:pos	-
Some	-	-	-	-	-	-	-	-	-	-
others	-	holder	-	-	-	-	-	-	-	-
give	-	-	-	-	-	-	-	-	-	-
the	-	-	-	-	-	-	-	-	-	-
new	-	-	-	-	-	-	-	-	-	-
UMUC	-	-	-	-	target	target	-	-	-	-
5	-	-	-	-	-	-	-	-	-	-
stars	-	-	holder	-	-	-	target	exp:pos	-	-
.	-	-	-	-	-	-	-	-	-	-

Table 2: The processed label for the model, where "-" indicates that there is no relationship between two words. The words on the y-axis are heads, and the words on the x-axis are dependents.

	origin	**+aug**	**+aug+focal**
NoRec	0.4798	0.4939	-
Multib_ca	0.7182	0.7281	-
Multib_eu	0.6781	0.7070	-
OpeNER_en	0.7271	0.7321	-
OpeNER_es	0.6758	0.7202	-
MPQA	0.3375	0.3564	0.3582
Dm_unis	0.3981	0.4412	0.4073

Table 3: The results on the dev dataset, origin means that there is no extra strategy applied, +aug means a mixed training set merged by other training sets and itself, and +aug+focal means that the strategy of focal loss is applied based on +aug.

dataset	**score**
NoRec_fine	0.497
Multib_ca	0.678
Multib_eu	0.723
OpeNER_en	0.745
OpeNER_es	0.735
MPQA	0.375
Dm_unis	0.380

dataset	**score**
EN-ES	0.620
EN-CA	0.543
EN-EU	0.527

Table 4: The left table is the result of monolingual, and the right table is the result of cross-lingual.

are described in detail below.

Data augmentation As shown in Table 1, the dataset is small. The largest dataset given is the NoRec dataset with 8,634 pieces of data, and only 4,555 pieces of data are left after removing the samples with empty opinions. Therefore, we use all datasets with the same polarity types for training with the help of ERNIE-M model, which supports multiple languages. For instance, we find that the Darmstadt_unis and MPQA have three categories of sentiment polarity, while the others have two categories of sentiment polarity. Hence, we merge Darmstadt_unis and MPQA into one dataset, and merge the others into another dataset. The results are shown in Table 3, and the evaluation method[1] is based on (Barnes et al., 2021). There is an average improvement of 2.36% after training the model with merged dataset, and a most significant improvement on the Opener_es at 4.44%.

[1]https://github.com/jerbarnes/semeval22_structured_sentiment

Focal loss As shown in Table 2, we can see the imbalance of the labels. The relationship between most words are none. Therefore, we introduce focal loss strategy, which reduces the loss weight of the "none" and increases the loss weight of the other labels. Because of the baseline scores on the MPQA and Darmstadt_unis are relatively lower, we test focal loss strategy on these two datasets. However, as we can see in Table 3, the performance is not good on Darmstadt_unis. Therefore, we do not use the focal loss strategy in the Darmstadt_unis solution.

5.3 Ensemble

We apply K-fold cross-validation and ensemble on the results of different pretrained models.

K-fold cross-validation For each dataset, we merge the original training set and the validation set, and perform K-fold segmentation after shuffling. The selection of K is related to the proportion of the original dataset training set and validation set. Notice that the proportion of the training set and the validation set after the segmentation should be approximated to their original propor-

tion. We repeat this operation K times and obtain K different models. Finally, we ensemble K different models to get the final model.

Usage of other pretrained models For English datasets (Darmstadt_unis, MPQA, Opener_en), we train additional model by replacing the ERNIE-M model with the DeBERTaV3(He et al., 2021) model as the lingual feature extractor, and keep the other parts unchanged. Then we perform K-fold cross-validation strategy on both models and ensemble them in the same way. The results are shown in Table 4.

6 Conclusion

In order to solve the issues of small dataset and cross language in SemEval-2022 task 10, we introduce a multilingual pretrained model ERNIE-M as a lingual feature extractor to the given baseline model. Furthermore, we use multiple strategies such as data augmentation, K-fold cross-validation, focal loss and ensemble to improve the model performance. In the future, we can take other techniques to do further optimization, such as different data augmentation technieques, fine-tuning the pretrained model with the in-domain dataset, and changing the label from word level to subword level to fit the subword representation of the pretrained model.

References

Rodrigo Agerri, Montse Cuadros, Sean Gaines, and German Rigau. 2013. OpeNER: Open polarity enhanced named entity recognition. In *Sociedad Española para el Procesamiento del Lenguaje Natural*, volume 51, pages 215–218.

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. *arXiv preprint arXiv:2105.14504*.

Jeremy Barnes, Oberländer Laura Ana Maria Kutuzov, Andrey and, Enrica Troiano, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, Erik Velldal, and Stephan Oepen. 2022. SemEval-2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Seattle. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:1906.06906*.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. *Advances in Neural Information Processing Systems*, 31.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6714–6721.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-m: enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. *arXiv preprint arXiv:2012.15674*.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5025–5033, Marseille, France. European Language Resources Association.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8600–8607.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584, Uppsala, Sweden. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.