ISCAS at SemEval-2022 Task 10: An Extraction-Validation Pipeline for Structured Sentiment Analysis

Xinyu Lu[*], **Mengjie Ren**[*], **Yaojie Lu**, **Hongyu Lin**
Chinese Information Processing Laboratory
Institute of Software, Chinese Academy of Sciences, Beijing, China
181303216@yzu.edu.cn, mjren@bupt.edu.cn,
{yaojie2017,hongyu}@iscas.ac.cn

Abstract

ISCAS participated in both sub-tasks in SemEval-2022 Task 10: Structured Sentiment competition. We design an extraction-validation pipeline architecture to tackle both monolingual and cross-lingual sub-tasks. Experimental results show the multilingual effectiveness and cross-lingual robustness of our system. Our system is openly released on: https://github.com/luxinyu1/SemEval2022-Task10/.

1 Introduction

Aspect-based Sentiment Analysis (ABSA) aims to detect the fine-grained sentiment tendency lying underneath texts. After decades of development, this area has formed a large family of tasks. Nevertheless, many of them are too simple or overlap with each other. Meanwhile, the popular evaluation resources are limited both in number and linguistic diversity. SemEval-2022 Task 10 (Barnes et al., 2022) is proposed to unify different sub-tasks in ABSA and introduces new metrics, new datasets on different languages to better evaluate methods in this area. Task 10 challenges its participants to extract opinion quadruple (*holder*, *target*, *expression*, *polarity*) from texts across English, Spanish, Basque, Catalan and Norwegian in monolingual (Sub-task 1) or cross-lingual (Sub-task 2) manners. Figure 1 provides two aspect-level annotations in a same sentence.

We applied an extraction-validation pipeline system and participated in both sub-task. Our system ranked at 10^{th} in 32 teams on the monolingual task without using extra data, and achieved competitive performance on the cross-lingual task. Besides, the proposed pipeline can be employed universally in monolingual and cross-lingual scenarios.

[*]This work was finished during their internship at ISCAS-CIP Lab.

2 Background

2.1 Task Definition

Task 10 is formalized as detecting all opinion tuples $O = O_i, ..., On$ in given text s. Concretely, each opinion O_i is a quadruple (h, t, e, p), denoting a ***holder*** who expresses a ***polarity*** $\in \{\text{Positive}, \text{Neutral}, \text{Negative}\}$ towards a ***target*** through a sentiment ***expression***. It's worthy to note that, *h*, *t*, *e* can be empty in this task. Following Cai et al. (2021), tuples with empty values are regard as implicit opinions in this system description paper. For the example in Figure 1, the quadruple "(–, them, don't believe negative)" is an implicit opinion.

2.2 Related Work

Aspect-based Sentiment Analysis Recently, there has been a large body of work focusing on different sub-tasks of ABSA. Generally we divide these sub-tasks into two categories: atomic and compound. Atomic ones take single element (e.g., t, e, or p) as the output and most of them can be treated as a sequence tagging problem (Li and Lam, 2017; Xu et al., 2018; Li et al., 2018; Wu et al., 2020b; Pouran Ben Veyseh et al., 2020). The compound ones need to find pairs (e.g., (t, p)), triplets (e.g., (t, e, p)) or even quadruples (e.g., (h, t, e, p)) from the inputs. Some works (Peng et al., 2020; Xu et al., 2021) use pipeline architecture to extract the elements separately and then make combinations; meanwhile some works use Seq2Seq models (Yan et al., 2021; Zhang et al., 2021) or unified tagging schemes (Mitchell et al., 2013; Zhang et al., 2015) to solve these sub-tasks in an end-to-end manner.

Pre-trained Language Models Pre-trained Language Models (PLMs) are deep neural networks pre-trained on large-scale corpora. Unlike traditional static word embedding methods, PLMs aim to learn dynamic contextual embedding of words in sentences from the unlabeled text. Recent re-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1305 - 1312
July 14-15, 2022 ©2022 Association for Computational Linguistics

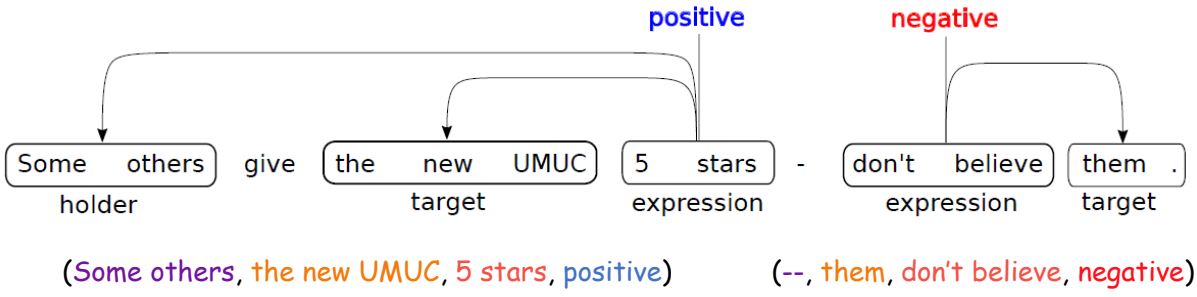

Figure 1: An example of Semeval 2022 Task 10. This figure is modified based on Figure 1 in (Barnes et al., 2021), the original sentiment graph representation is linearized to quadruple representation in this task. "–" indicates that this element in quadruple is empty.

search shows PLMs perform well in various syntactic tasks, such as POS tagging.

BERT (Devlin et al., 2019) is a typical language representation model based on the Transformer encoder architecture. It is pre-trained on two unsupervised tasks: Mask Language Modeling (MLM) and Next Sentence Prediction (NSP). mBERT[1] is a multilingual version of BERT pre-trained on the wiki dumps of 104 languages.

RoBERTa (Liu et al., 2019) removes the NSP task, which has no prominent effect in BERT pre-training and further improves BERT with dynamic masking, deeper network, longer input sequence, and larger training corpora. By virtue of these robust optimizations, RoBERTa significantly outperforms BERT on many tasks. XLM-RoBERTa (Conneau et al., 2020) extends RoBERTa architecture to the multilingual scenario by scalable pre-training on filtered CommonCrawl data containing 100 languages.

SKEP (Tian et al., 2020) incorporates sentiment knowledge into PLMs through sentiment masking and three sentiment pre-training objectives. It provides a unified contextual representation for downstream sentiment tasks.

NB-BERT (Kummervold et al., 2021) is a Norwegian instance of BERT in low-resource language. To alleviate the shortage of pre-training Norwegian corpora, OCR is conditionally used to mine good texts from digital copies.

3 System Overview

To tackle this task, we design a pipeline system that decouples this complex problem into a two-step pipeline with an extraction stage and validation stage. In the extraction stage, we first extract *target-expression-polarity* using an extended grid tagging schema, and then extract *holder* with a question answering system. In the validation stage, we employ a neural validator to determine the extracted results whether are valid in texts. Figure 2 illustrates the overall architecture of our system.

3.1 *Target-expression-polarity* co-extraction

Target-expression-polarity co-extraction aims to extract (t, e, p) triplets from text s (Peng et al., 2020). However, existing works (Peng et al., 2020; Wu et al., 2020a) usually assume that all opinions are expressed explicitly and pay little attention to implicit opinion extraction. In our system, we extend Grid Tagging Scheme (GTS) (Wu et al., 2020a) to adapt both implicit and explicit opinion extraction.

Original tagging space in GTS is an upper triangular grid, whose length and width is the tokenized sequence length l. Specifically, for $i, j \in [0, l]$, cell (i, j) contains the tag for token-pair (t_i, t_j) in the grid tagging. We integrate two new labels $\{\mathrm{IA}, \mathrm{IO}\}$ into the original tagging scheme and end up with a label set containing eight labels: $\mathcal{Y} = \{\mathrm{A}, \mathrm{O}, \mathrm{IA}, \mathrm{IO}, \mathrm{Pos}, \mathrm{Neu}, \mathrm{Neg}, \mathrm{N}\}$. The grid representation of implicit opinions can thus be implemented by filling IA or IO label in the cell of token-pair (t_0, t_0) while not interfering with the representation of explict opinions. We believe this strategy is also reasonable under the perspective of sentence embedding in pretrained encoders, owing to that hidden-state of `[CLS]` (or `<s>` in RoBERTa) token which later fed into the token-level classifier, is often used as the semantic representation of the whole sentence.

We list the meanings of labels in our extended GTS separately in Table 1 and provide a tagging example for the extended GTS in Figure 3.

The decoding algorithm and inference steps we exploit are identical to the original paper.

[1] https://github.com/google-research/bert

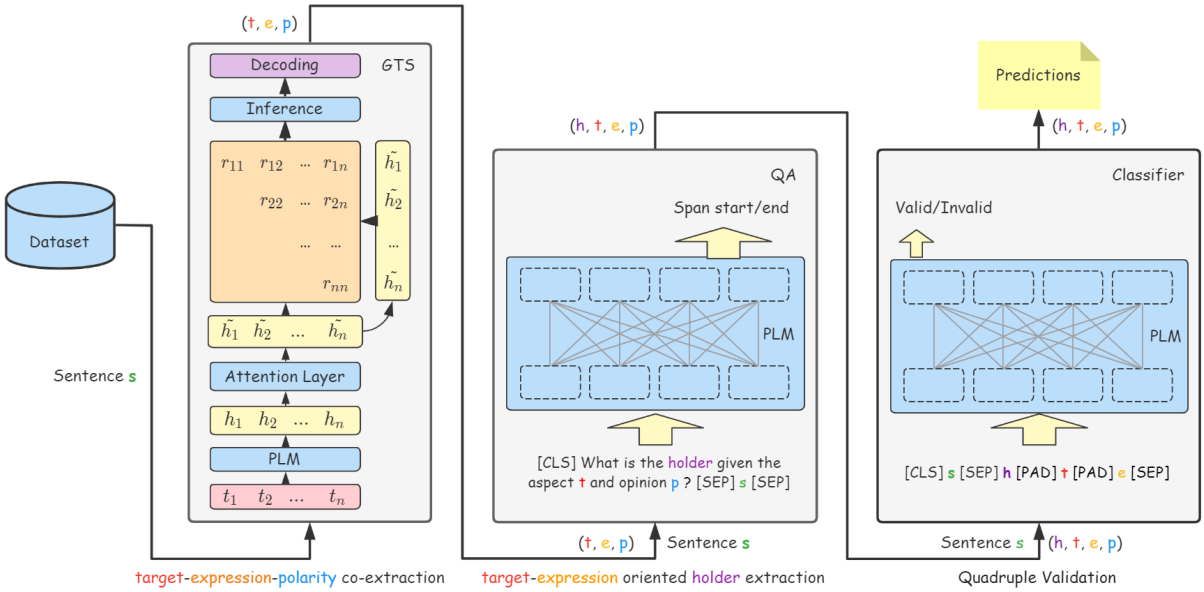

Figure 2: The overview of our system in SemEval-2022 Task 10. Best viewed in color.

Tags	Meanings of tags in cell (i, j)
A O	t_i and t_j belong to the same *target* term. t_i and t_j belong to the same sentiment *expression* term.
IA IO	$i = j = 0$, indicating an implicit *target* term. $i = j = 0$, indicating an implicit *expression* term.
Pos Neu Neg	t_i and t_j respectively belong to an *target* term and an *expression* term, and they form Positive/Neutral/Negative opinion pair relation.
N	No relation between t_i and t_j

Table 1: The meanings of tags in our extended GTS. Cell (i, j) contains the tag for token-pair (t_i, t_j).

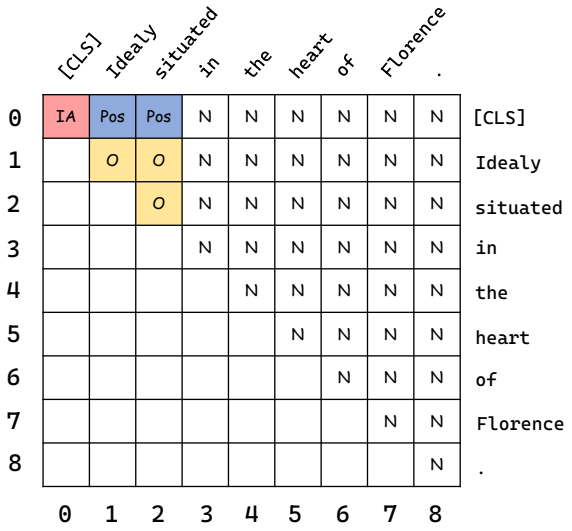

Figure 3: The extended Grid Tagging Scheme on opinion triplet $(-, \text{Idealy situated}, \text{Positive})$ in sentence "Ideally situated in the heart of Florence.". "–" indicates this element in triplet is empty.

Model Ensemble We ensemble the different GTS models using a variety of backbones as the final predictor. Specifically, we perform an unweighted average of predicted distributions $p_{ij} \in \mathbb{R}^d$ from each model on token-pair (t_i, t_j) and get $\bar{p_{ij}}$. The final predicted label index is $\text{argmax}(\bar{p_{ij}})$.

3.2 *Target-expression* oriented *holder* extraction

After obtaining (t, e, p) triplets from the previous step, we further predict *holder* for each given triplet extracted from text s, i.e., *target-expression* oriented holder extraction. We cast this problem as a Question Answering (QA) task, where the context is text s and the answer is the holder span.

Query Construction For holder extraction, we construct the query q for the QA system with the (t, e, p) triplet. Under the multilingual setting of this task, we design different question templates in different languages. The details of the question templates are shown in Table 2.

Encoding and Inference We adopted the same setting as Devlin et al. (2019) to handle the QA task. The input query message q and text s are presented as a single packed sequence:

$$\boldsymbol{x} = \begin{cases} [\texttt{[CLS]}; q; \texttt{[SEP]}; s; \texttt{[SEP]}] & \text{if BERT} \\ [\texttt{<s>}; q; \texttt{</s></s>}; s; \texttt{</s>}] & \text{if RoBERTa} \end{cases} \quad (1)$$

Language	Question Template
English	What is the holder given the aspect t and the opinion e ?
Spanish	¿Cuál es el titular de la opinión dado el aspecto t y la opinión e ?
Basque	Zein da helburu t eta e iritzia emanda iritzia duenak ?
Catalan	Quin és el titular de l'opinió donat l'aspecte t i l'opinió e ?
Norwegian	Hva er meningshaveren gitt aspektet t og meningen e ?

Table 2: The question templates we make to get query message in different languages. t denotes the target term and e denotes the expression term. When t or e is empty, the string "empty" are given to the templates as the term.

Then the context-aware representations of $\boldsymbol{x}$ are fed to a feed-forward linear layer to detect the span-start and span-end position. Note that we treat the special symbol `[CLS]` (or <s>) as the impossible answers for implicit opinions that without corresponding *holders*.

In detail, we feed the tokenized input sequence $\boldsymbol{x}$ into the encoder of PLMs. The last hidden-states $\mathbf{H}^x \in \mathbb{R}^{l \times d}$ can be represented by:

$$\mathbf{H}^x = \begin{cases} [\boldsymbol{h}_{\texttt{[CLS]}}; \boldsymbol{h}_q; \boldsymbol{h}_{\texttt{[SEP]}}; \boldsymbol{h}_s; \boldsymbol{h}_{\texttt{[SEP]}}] & \text{if BERT} \\ [\boldsymbol{h}_{\texttt{<s>}}; \boldsymbol{h}_q; \boldsymbol{h}_{\texttt{</s></s>}}; \boldsymbol{h}_s; \boldsymbol{h}_{\texttt{</s>}}] & \text{if RoBERTa} \end{cases} \tag{2}$$

where l is the length of the tokenized sentence, and d is the dimension of PLMs. The final linear span prediction network takes $\mathbf{H}^x$ as the input and outputs two probabilities $p_s, p_e \in \mathbb{R}^l$ for span-start and span-end prediction:

$$p^s, p^e \propto \text{softmax}(\text{Linear}(\mathbf{H}^x)) \tag{3}$$

For model learning, the whole parameters in the QA model are optimized by maximizing the likelihood of span-start and span-end positions:

$$\mathcal{L}_{\mathcal{QA}} = -\frac{1}{N} \sum_{i=1}^{N} \left[\log \left(p^s_{y^s_i} \right) + \log \left(p^e_{y^e_i} \right) \right] \tag{4}$$

where N is the number of spans in a single batch, y^s_i and y^e_i are ground-truth span-start and span-end positions respectively.

3.3 Quadruple Validation

To reduce the errors accumulated in previous steps, we design a binary classifier that determines if a combination of *holder*, *target*, and *expression* is valid in text s. The valid triplets predicted by this sub-system are kept along with their corresponding *polarity*, while the others are discarded.

Encoding and Inference We utilize the pre-trained transformers to obtain the representation of text and triplets. Since BERT-like models are more sensitive to sentence-pair input, we concatenate h, t, e with a special symbol `[PAD]` and treat them together as sentence `B`. Concretely, we build the sequence pack in the form of:

$$\boldsymbol{x} = [[\texttt{CLS}]; s; [\texttt{SEP}]; h\,\texttt{[PAD]}\,t\,\texttt{[PAD]}\,e; [\texttt{SEP}]] \tag{5}$$

Under the circumstances of implicit opinion, the empty h, t, e terms are replaced with a special token `[EMP]`.

The validator network takes the representation of $\boldsymbol{x}_{\texttt{[CLS]}}$ as the input and returns the binary validation result. We implement the validator with a linear feed-forward layer.

Span Manipulation Considering that the combinations from sub-spans of the golden *holder*, *target*, and *opinion* terms are also treated as weighted correct predictions, we perform span manipulation to build a more robust classifier. For each ground-truth *holder*, *target*, *expression* term in triplet, we enumerate all the sub-spans and the original term in their corresponding set $\mathbf{H}, \mathbf{T}, \mathbf{E}$, the final triplet candidate pool is the Cartesian product of the three set: $\mathbf{H} \times \mathbf{T} \times \mathbf{E}$.

Finally, for each golden triplet, we randomly select at most k triplets (must include the original one) from the candidate pool as positive samples.

Negative Sampling We further design several rules to mine the negative (i.e., invalid) samples from raw datasets and manipulated golden triplets, including:

1.1. If a golden triplet has a holder, remove the holder and keep other elements.

1.2. If a golden triplet doesn't have a holder, use a holder dictionary to mine pseudo holders from text, packaging the mined holders (if there exist any) with the golden triplet.

2. If a text has multiple golden triplets, exchange the holder / target / expression terms in one with the other.

3. Randomly sample triplets.

Rules $1 \rightarrow 3$ are sequentially executed until q samples have been harvested, where q is in positive correlation with the number of positive samples. Meanwhile, we remove all the weighted true and true samples from the mined pseudo negative samples.

4 Experimental setup

4.1 Data Splits

Monolingual Sub-task This sub-task contains 7 different datasets (Agerri et al., 2013; Wiebe et al., 2005; Toprak et al., 2010; Barnes et al., 2018; Øvrelid et al., 2020) across 5 languages. We leveraged the origin splits provided by the organizer and did not include any extra data. The details of data splits are shown in Table 3.

Dataset	Splits		
	Train	Dev	Test
OpeNER$_{\text{en}}$	1,744	249	499
OpeNER$_{\text{es}}$	1,438	206	410
NoReC$_{\text{Fine}}$	8,634	1,531	1,272
MPQA	5,873	2,063	2,112
DS$_{\text{unis}}$	2,253	232	318
MultiB$_{\text{ca}}$	1,174	167	335
MultiB$_{\text{eu}}$	1,063	152	305

Table 3: Data splits.

Cross-lingual Sub-task This sub-task uses a zero-shot setting in which models are trained on the resource that does not contain annotations in the target language. For each target language, we combine all the training sets of OpeNER$_*$, MPQA, and MultiB$_*$ in other languages.

4.2 Implementation and Hyperparameters

This section generally describes the system implementation details and the selection of parameters. The detailed settings can be found in Appendix A.

Monolingual Sub-task For extended GTS and QA part in our pipeline, we tune and select models based on SF$_1$ (Sentiment Graph F1 (Barnes et al., 2021)) scores on the development splits. For the validator part, the models are tuned based on the classification accuracy on the manipulated and sampled development datasets.

In order to maximize the advantages of our system, we test a number of high-performing PLMs and finally RoBERTa$_{\text{large}}$, XLM-RoBERTa$_{\text{large}}$, NB-BERT$_{\text{large}}$, SKEP-ERNIE$_{\text{large}}$ and ensemble model [BERT$_{\text{large}}$+SKEP-ERNIE$_{\text{large}}$] are adopted to the training on different datasets in extended GTS. The max sequence length is set to the max of training and development sets, and meanwhile, the number of hops is chosen in 2 and 3 for GPU memory limitation. The large extended GTS models are trained on a single A100 80G GPU.

For QA sub-system, we use several task-pre-trained PLMs as backbones, such as XLM-RoBERTa$_{\text{large}}$-SQuAD[2], distilBERT$_{\text{base}}$-SQuAD[3] and RoBERTa$_{\text{large}}$-SQuAD[4].

For the validation step, we add LaBSE[5], which is a PLM focusing on language-agnostic sentence embedding and mBERT to the model pools in GTS training.

We fine-tuned all models on the training data using linear learning rate scheduler and the warming up strategy with the learning rate of 3e-5/3e-6 and the batch size of 8~64. We set all random seeds to 1 for reproducibility.

Cross-lingual Sub-task We set all holder positions in tuples to empty instead of leveraging the QA sub-system to extract holders. This is because the QA sub-system requires extra enhancements to fitting the cross-lingual setting (Cui et al., 2019).

The cross-lingual backbone in extended GTS is XLM-RoBERTa$_{\text{large}}$, and LaBSE for the validator.

5 Results

In this section, we report the scores on the development and test datasets of two sub-tasks separately. We use SF_1 (Sentiment Graph F1), *SP* (Sentiment Graph Precision) and *SR* (Sentiment Graph Recall) to evaluate the performance of our system.

5.1 Monolingual Sub-task

Table 4 reports the results of the monolingual sub-task, which ranks 10^{th} in 32 teams. Table 5 shows the ablation analysis of different components on the development set of monolingual tasks. We can see that: 1) Grid-tagging-scheme based target-expression-polarity co-extraction achieves good performance in different languages. 2) The proposed validator can effectively filter out invalid

[2]https://huggingface.co/deepset/xlm-roberta-large-squad2/

[3]https://huggingface.co/distilbert-base-uncased-distilled-squad/

[4]https://huggingface.co/deepset/roberta-large-squad2/

[5]https://tfhub.dev/google/LaBSE/1/

	SF_1	SP	SR
OpeNER$_{\text{en}}$	0.710	0.788	0.646
OpeNER$_{\text{es}}$	0.669	0.735	0.614
NoReC$_{\text{Fine}}$	0.487	0.539	0.444
MPQA	0.269	0.369	0.211
DS$_{\text{unis}}$	0.416	0.480	0.366
MultiB$_{\text{ca}}$	0.658	0.720	0.605
MultiB$_{\text{eu}}$	0.651	0.705	0.605

Table 4: Sub-task 1 Results.

	System	SF_1	SP	SR
OpeNER$_{\text{en}}$	Co-Extraction	0.686	0.710	0.664
	+ Holder Extraction	0.705	0.732	**0.681**
	+ Quadruple Validation	**0.717**	**0.786**	0.660
OpeNER$_{\text{es}}$	Co-Extraction	0.707	0.716	**0.698**
	+ Holder Extraction	0.707	0.716	**0.698**
	+ Quadruple Validation	**0.728**	**0.768**	0.692
NoReC$_{\text{Fine}}$	Co-Extraction	0.501	0.510	**0.492**
	+ Holder Extraction	0.501	0.510	**0.492**
	+ Quadruple Validation	**0.510**	**0.565**	0.465
MPQA	Co-Extraction	0.139	0.148	0.131
	+ Holder Extraction	0.345	0.362	**0.330**
	+ Quadruple Validation	**0.358**	**0.424**	0.309
DS$_{\text{unis}}$	Co-Extraction	0.370	0.453	0.313
	+ Holder Extraction	0.393	0.480	**0.333**
	+ Quadruple Validation	**0.398**	**0.493**	**0.333**
MultiB$_{\text{ca}}$	Co-Extraction	0.674	0.707	0.643
	+ Holder Extraction	0.677	0.711	**0.646**
	+ Quadruple Validation	**0.706**	**0.800**	0.631
MultiB$_{\text{eu}}$	Co-Extraction	0.567	0.553	0.581
	+ Holder Extraction	0.601	0.577	**0.627**
	+ Quadruple Validation	**0.625**	**0.665**	0.589

Table 5: Ablation analysis of our pipeline system on the dev sets in Sub-task 1.

triples and significantly improve the precision of the model.

5.2 Cross-lingual Sub-task

Table 6 shows the results on the cross-lingual sub-task. Compared to the monolingual sub-task, the experimental results shows that the proposed cross-lingual system still performs competitively without training on the target language.

6 Conclusion

In this paper, we propose a pipeline system for (*holder*, *target*, *expression*, *polarity*) quadruple extraction in ABSA, and adopt a verity of pre-trained language models in distinct parts of system. The evaluation results demonstrate the effectiveness and robustness of our system.

	SF_1	SP	SR
OpeNER$_{\text{es}}$	0.620	0.716	0.548
MultiB$_{\text{ca}}$	0.605	0.596	0.615
MultiB$_{\text{eu}}$	0.569	0.573	0.566

Table 6: Sub-task 2 Results.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3387–3402, Online. Association for Computational Linguistics.

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.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Cross-lingual machine reading comprehension. In *Proceedings of*

the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1586–1595, Hong Kong, China. Association for Computational Linguistics.

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

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a Norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4194–4200. International Joint Conferences on Artificial Intelligence Organization.

Xin Li and Wai Lam. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2886–2892, Copenhagen, Denmark. Association for Computational Linguistics.

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 preprint arXiv:1907.11692*.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654, Seattle, Washington, USA. Association for Computational Linguistics.

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.

Amir Pouran Ben Veyseh, Nasim Nouri, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2020. Introducing syntactic structures into target opinion word extraction with deep learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8947–8956, Online. Association for Computational Linguistics.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4067–4076, Online. Association for Computational Linguistics.

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.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020a. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2576–2585, Online. Association for Computational Linguistics.

Zhen Wu, Fei Zhao, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2020b. Latent opinions transfer network for target-oriented opinion words extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9298–9305.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double embeddings and CNN-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598, Melbourne, Australia. Association for Computational Linguistics.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4755–4766, Online. Association for Computational Linguistics.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*

(Volume 1: Long Papers), pages 2416–2429, Online. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2015. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 612–621, Lisbon, Portugal. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

A Experiment Details

Table 7 shows the detailed configurations of each sub-system in the two sub-tasks.

Dataset	Subsystem	Backbone	Hyper-parameters
		Monolingual	
OpenNER$_{\text{en}}$	Co-Extraction	[BERT$_{\text{large}}$+SKEP-ERNIE$_{\text{large}}$]	n-hop=3,lr=3e-5, bs=8, msl=132, epochs=100
	Holder Extraction	RoBERTa$_{\text{large}}$-SQuAD	lr=3e-5, bs=16, msl=384, epochs=15
	Quadruple Validation	BERT$_{\text{large}}$	bs=16, lr=3e-6, msl=512, epochs=10
OpenNER$_{\text{es}}$	Co-Extraction	XLM-RoBERTa$_{\text{large}}$	n-hop=3,lr=3e-5, bs=8, msl=193, epochs=100
	Holder Extraction	XLM-RoBERTa$_{\text{large}}$-SQuAD	lr=3e-5, bs=32, msl=384, wus=100, epochs=15
	Quadruple Validation	LaBSE	bs=32, lr=3e-5, msl=512, epochs=10
NoReC$_{\text{Fine}}$	Co-Extraction	NB-BERT$_{\text{large}}$	n-hop=3,lr=3e-5, bs=16, msl=125, epochs=100
	Holder Extraction	XLM-RoBERTa$_{\text{large}}$-SQuAD	lr=3e-5, bs=32, msl=384, epochs=15
	Quadruple Validation	NB-BERT$_{\text{base}}$	bs=32, lr=3e-6, msl=512, epochs=10
MPQA	Co-Extraction	RoBERTa$_{\text{large}}$	n-hop=2,lr=3e-6, bs=16, msl=230, wus=2000, epochs=100
	Holder Extraction	BERT$_{\text{base}}$-distilled-SQuAD	lr=3e-5, bs=64, msl=384, epochs=15
	Quadruple Validation	SKEP-ERNIE$_{\text{large}}$	bs=64, lr=3e-6, msl=512, epochs=5
DS$_{\text{unis}}$	Co-Extraction	SKEP-ERNIE$_{\text{large}}$	n-hop=3,lr=3e-5, bs=8, msl=229, wus=500, epochs=100
	Holder Extraction	RoBERTa$_{\text{large}}$-SQuAD	lr=3e-5, bs=16, msl=384, wus=1000, epochs=20
	Quadruple Validation	SKEP-ERNIE$_{\text{large}}$	bs=64, lr=3e-5, msl=512, epochs=10
MultiB$_{\text{ca}}$	Co-Extraction	XLM-RoBERTa$_{\text{large}}$	n-hop=3,lr=3e-5, bs=8, msl=265, epochs=100
	Holder Extraction	XLM-RoBERTa$_{\text{large}}$-SQuAD	lr=3e-5, bs=32, msl=384, epochs=15
	Quadruple Validation	mBERT$_{\text{base}}$	bs=32, lr=3e-5, msl=512, epochs=10
MultiB$_{\text{eu}}$	Co-Extraction	XLM-RoBERTa$_{\text{large}}$	n-hop=3,lr=3e-5, bs=8, msl=132, epochs=100
	Holder Extraction	XLM-RoBERTa$_{\text{large}}$-SQuAD	lr=3e-5, bs=32, msl=384, epochs=15
	Quadruple Validation	LaBSE	bs=32, lr=3e-5, msl=512, epochs=10
		Cross-lingual	
OpenNER$_{\text{es}}$	Co-Extraction	XLM-RoBERTa$_{\text{large}}$	n-hop=3,lr=3e-6, bs=8, msl=265, epochs=100
	Quadruple Validation	LaBSE	bs=128, lr=3e-6, msl=512, wus=1000, epochs=10
MultiB$_{\text{ca}}$	Co-Extraction	XLM-RoBERTa$_{\text{large}}$	n-hop=3,lr=3e-6, bs=8, msl=193, epochs=100
	Quadruple Validation	LaBSE	bs=128, lr=3e-6, msl=512, wus=1000, epochs=10
MultiB$_{\text{eu}}$	Co-Extraction	XLM-RoBERTa$_{\text{large}}$	n-hop=3,lr=3e-6, bs=8, msl=152, epochs=100
	Quadruple Validation	LaBSE	bs=128, lr=3e-6, msl=512, wus=1000, epochs=10

Table 7: Detailed configurations of the subsystems. The abbreviation "bs" stands for batch size, "msl" for max sequence length, "wus" for number of warm-up steps. "[A+B]" represents an ensemble model using backbones A and B.