Task and Sentiment Adaptation for Appraisal Tagging

Lin Tian
RMIT University, Australia
lin.tian2@student.rmit.edu.au

Xiuzhen Zhang*
RMIT University, Australia
xiuzhen.zhang@rmit.edu.au

Maria Myung-Hee Kim
Defence Science and Technology Group Australia
myung.kim@defence.gov.au

Jennifer Biggs
Defence Science and Technology Group Australia
jennifer.biggs@defence.gov.au

Abstract

Sentiment analysis and opinion mining of the opinion-bearing text are important tasks in NLP. The Appraisal framework in systemic functional linguistics is a theory for analysing the linguistic patterns for expressing emotion and opinion. Manual annotation of appraisals however, requires linguistic expertise, and is costly and time-consuming. In this paper, we study how to automatically identify and tag appraisal text segments. We formulate the problem as a sequence tagging problem and propose a novel approach, Adaptive Appraisal (A^2), which employs task and sentiment adapters on pre-trained language models for sequence appraisal tagging. Experiments on user comments, blogs and microblogs show that A^2 outperforms baseline models and achieves good performance for cross-domain and cross-lingual settings. Source code for A^2 is available at: https://github.com/ltian678/AA-code.git

1 Introduction

With the development of the Web technology, opinion-bearing user generated texts such as reviews, users comments, blogs and microblogs are widespread. Sentiment analysis and opinion mining on such texts are prominent NLP tasks that have attracted extensive research studies in the literature (Liu, 2022). On the other hand, the Appraisal framework (Martin and White, 2003) is a systemic functional linguistic theory describing how language is used by writers or speakers to express emotion and opinion. The Appraisal framework consists of three subsystems: 1) Attitude, which includes personal emotion, judgement and evaluation of entities; 2) Engagement, which regards one's own opinions or with respect to others; and 3) Graduation, which describes strength of the attitude and engagement expressed.

Appraisal annotated resources have been used for deeper sentiment and emotion analysis (Whitelaw et al., 2005) than simple sentiment classification, but building such resources manually requires significant linguistic expertise and is time-consuming (Read and Carroll, 2012; Kolhatkar et al., 2020). Automated appraisal tagging would be extremely beneficial to support annotation and analysis efforts by expert linguists. To our best knowledge, the only publicly available appraisal annotated corpus is the Simon Fraser University Opinion and Comments Corpus (SOCC) (Kolhatkar et al., 2020)[1], which is based on news comments. In creating the SOCC corpus, 663,173 user comments were collected, and expert linguists manually annotated 1,043 comments.

To the best of our knowledge, there has been no prior work leveraging machine learning for automatically tagging appraisals. We fill this gap in this paper. We especially target domains with voluminous opinionated texts but zero or very limited appraisal annotation resources, such as blogs and microblogs. Our research focuses on sequence tagging for cross-domain and cross-lingual texts with low resources. In the literature, various approaches for sequence tagging tasks have been reported, including transfer learning (Lee et al., 2018a), few-shot learning (Hofer et al., 2018) and multi-task learning (Changpinyo et al., 2018; Kann et al., 2018; Liu et al., 2018). However, none of the existing studies on sequence tagging consider zero-shot cross-domain or cross-lingual settings.

We propose a model for automatic Appraisal tagging. Our model A^2, namely Adaptive Appraisal, utilises joint task and sentiment adapters based on pre-trained language models for tagging appraisal segments in text sequences. Our model leverages the adapter-based transfer learning framework for cross-domain and cross-lingual appraisal tagging. Based on the pre-trained language model, we propose the task adapter for appraisal tagging across different domains. For instance, the language that

*Corresponding author

[1] https://github.com/sfu-discourse-lab/SOCC

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1960–1970

short Microblog posts use for expressing appraisals is very different from the language of long blog posts. It is therefore necessary to enable the adaptive ability from one domain to different domains. We further propose the sentiment adapter to capture the sentiment knowledge for appraisal tagging, capitalised on the strong correlation between appraisals and sentiments. Rather than employing appraisal annotations for sentiment analysis as in existing studies (Whitelaw et al., 2005), we propose to leverage the rich sentiment analysis resources for automatic appraisal tagging.

In this study, we seek to answer the following research questions.

- RQ1: Can we leverage the adapter-based framework for the within-domain appraisal tagging task?
- RQ2: Can the sentiment knowledge fused adapter improve the adapter-based framework for cross-domain appraisal tagging?
- RQ3: Can the sentiment knowledge fused adapter improve the adapter-based framework for the cross-lingual appraisal tagging task?

To summarise, our contributions are twofold: (1) we propose an adapter-based framework for appraisal tagging; and (2) we propose task and sentiment adapters to further enhance the framework for cross-domain and cross-lingual generalisation ability.

2 Related Work

Our work is related to sequence tagging – such as named entiy recognition, semantic role labeling, where token sequences in the input text are tagged with class labels. Our work is especially related to the task of sequence tagging with low resources for training. In the literature, to address the issue of low resources, sequence taggers based on transfer learning (Lee et al., 2018a), few-shot learning (Hofer et al., 2018) and multi-task learning (Changpinyo et al., 2018; Kann et al., 2018; Liu et al., 2018) have been reported. Our research falls under transfer learning.

Various transfer learning strategies and techniques have been proposed for NLP tasks addressing the issue of scarce labelled data. Early transfer learning algorithms have addressed the target domain data scarcity problem and boosted the model's generalisation ability via learning domain-agnostic knowledge for transfer (Kim et al., 2015; Lee et al., 2018b). Modern pre-trained language models (e.g. BERT) have achieved the state-of-the-art performance for a range of Natural Language Processing (NLP) tasks via transfer learning by fine-tuning parameters for different tasks (Devlin et al., 2019; Radford et al., 2019; Liu et al., 2019; Li et al., 2019; Yin et al., 2020). Apart from cross-domain, cross-lingual transfer learning has also been investigated, in particular for part-of-speech tagging and dependency parsing (Ruder et al., 2019). Algorithms have been proposed for transfer learning tasks of low resource languages (Kim et al., 2017; Schuster et al., 2019). All these transfer learning frameworks require fine-tuning parameters of the full model to achieve knowledge transfer, which limits the capacity for models to adapt to many target domains.

Adapter-based transfer is a recently proposed parameter-efficient transfer learning mechanism for adapting a pre-trained model to a target task without fine-tuning all parameters. Adapter modules was originated from computer vision, to control the convolutions and adapt models to multiple domains (Rebuffi et al., 2017). Then, in NLP application, adapters have been widely used for quick adaption in combination with existing large language models to new tasks (Houlsby et al., 2019) and avoiding catastrophic forgetting issues (McCloskey and Cohen, 1989). Üstün et al. (2020) generated adapter parameters from language embeddings for multilingual dependency parsing. Pfeiffer et al. (2021) combined the information stored in multiple adapters for most robust transfer learning between monolingual tasks. As adapter modules have been proved effective for efficient transfer learning with large language models, we propose task and sentiment adapters for the cross-domain and cross-lingual appraisal tagging tasks.

3 The Appraisal Framework

Following the Systemic Functional Linguistics (SFL) theory (Eggins, 2004), the Appraisal framework is a theory describing the linguistic patterns for authors to express emotion and opinion. The Appraisal framework consists of three semantic systems including Attitude, Graduation and Engagement. Attitude is divided into three sub-systems: *Affect*, *Judgement* and *Appreciation*. *Affect* deals with a person's emotional reactions (e.g. happy,

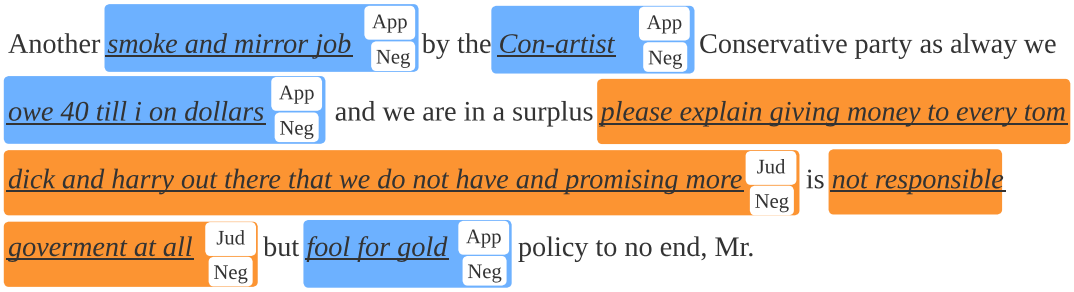

Figure 1: Sample appraisal annotations from SOCC dataset ("App" = Appreciation, "Jud" = Judgement, "Neg" = Negative)

confident), *Judgement* deals with assessing people's behaviour (e.g. powerful, truthful), and *Appreciation* deals with constructing the value of things (e.g. fascinating, exciting). In this work, we focus on the three sub-systems of Attitude and their Polarity (*Positive*, *Negative* and *Neutral*). Figure 1 shows an example for appraisal annotations from the SOCC dataset (Kolhatkar et al., 2020); each span is labelled with an attitude and its polarity.

It should be noted that the linguistic Appraisal framework in this study is different from the Appraisal theory of emotion (Ellsworth and Smith, 1988). The Appraisal theory of emotion (Ellsworth and Smith, 1988) describes that emotions are the result of the way in which people appraise or evaluate events and situations in terms of their relevance and significance to their goals, needs and values. According to this theory, emotions are generated by the way in which people appraise events and situations, and the specific emotion that is generated depends on the type of appraisal that is made. The functional linguistic Appraisal framework (Oteíza, 2017) on the other hand, describes that emotions are the result of the way in which people use language to evaluate and interpret events and situations. The specific emotion that is generated depends on the specific linguistic patterns that are used.

4 Problem Statement

We frame our appraisal tagging task as a BIO sequence tagging problem (Ramshaw and Marcus, 1995), where segments are tagged with the Attitude labels of Affect, Appreciation and Judgement, and Polarity labels of Positive, Negative and Neutral. Let $S = (w_1, w_2, ..., w_l)$ be an input sentence, where w_i is the i-th token and l is the sequence length. The objective of the proposed model A^2 is to identify a set of attitude tags $T_{att} = (\text{B-Attitude}, \text{I-Attitude}, O)$ and a set of polarity tags $T_{pol} = (\text{B-Polarity}, \text{I-Polarity}, O)$ for each $w_i \in S$.

5 Methodology

The architecture of adaptive appraisal (A^2) model is shown in Figure 2. It comprises two modules: (1) a language model for automatic BIO tagging with two target tasks, and (2) a sentiment adapter for learning sentiment specific knowledge and a task adapter for generating task-specific representations.

5.1 The Adaptive Appraisal (A^2) Model

The model predicts Appraisal Attitude and Polarity labels simultaneously. For the input sequence $S = (w_1, w_2, ..., w_l)$, we first feed to adapter-based transformer to generate token embeddings.

$$v = \text{Adapter-Transformer}([\text{CLS}] \oplus w_1 \oplus ... \oplus w_l)$$
$$v = [CLS], Emb[w_1], ..., Emb[w_l]$$

Each token embedding(e.g. $Emb[w_1]$) is presented by a d dimensional vector, where BERT-based encoder $d = 768$. Then we get a list of token representations with sequence length l and dimension d, denoted as $v \in \mathbb{R}^{l \times d}$. For two different objectives, we pass the token representations to two separate multilayer perceptron (MLP), denoted as Φ_{att} and Φ_{pol}.

$$\begin{aligned} \Phi_{att} &= \sigma(XU) \\ \Phi_{pol} &= \sigma(XU) \\ Z &= \Phi_{att}\,(Emb[w_1], ..., Emb[w_l]) \\ Z' &= \Phi_{pol}\,(Emb[w_1], ..., Emb[w_l]) \end{aligned}$$

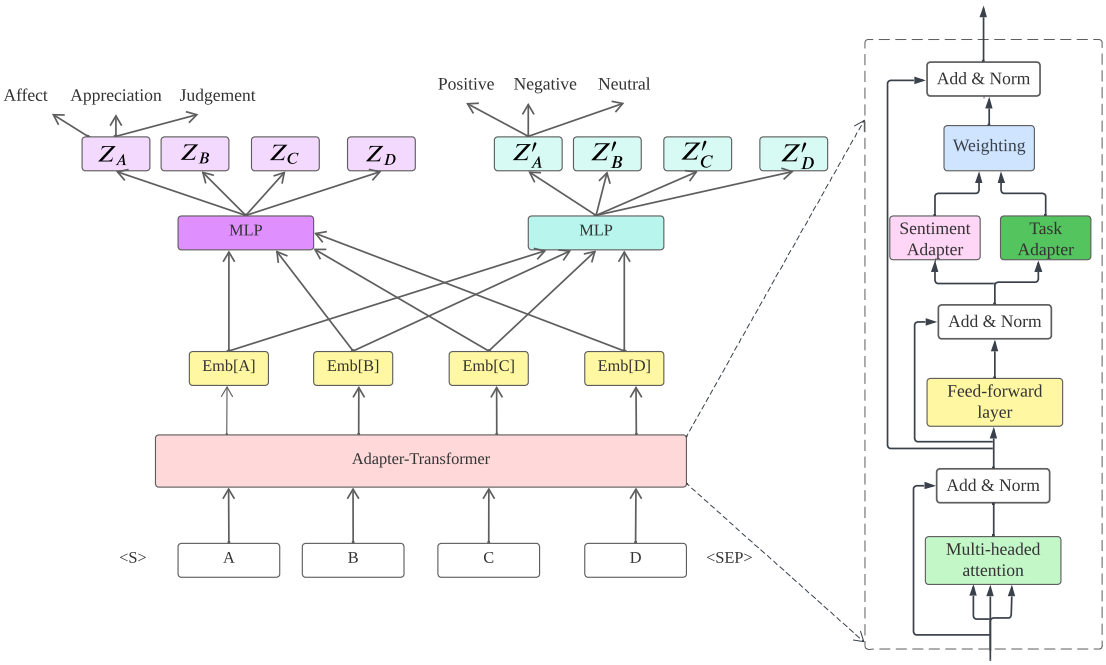

Figure 2: Overall architecture of Adapter-based Automatic Appraisal Framework ("A, B, C, D" are the tokens in the input sequence S)

where σ is ReLU (Agarap, 2018) activation function. U is the linear projection along the sequence length. Normalisation and biases are omitted for brevity. Due to the imbalance on our appraisal data, the dice loss (Li et al., 2020) has been adopted for token tagging tasks. To perform joint training with data D_{train} based on pre-trained language model on both the Attitude and Polarity labels, we minimise the overall loss:

$$\mathcal{L}^d_{att} = 1 - \frac{2p_d y_d + \gamma}{p_d^2 + y_d^2 + \gamma}$$

$$\mathcal{L}^d_{pl} = 1 - \frac{2p_d y_d + \gamma}{p_d^2 + y_d^2 + \gamma}$$

$$\mathcal{L} = \frac{1}{|\mathcal{D}_{train}|} \sum_{d \in \{D_{train}\}} \left(\mathcal{L}^d_{att} + \mathcal{L}^d_{pl} \right)$$

where λ is the L_2 regularisation parameters and Θ represents the parameters set. Following the dice loss setting, we set $\gamma = 1$, and p_d is the possibility of the data d belongs to the prediction y_d after the softmax function. $\mathcal{L}^d_{\text{att}}$ and $\mathcal{L}^d_{\text{pl}}$ are the loss functions for Attitude and Polarity labels, respectively. The reason for using the dice loss is to mitigate the impact of our imbalanced Appraisal labelled data.

5.2 Sentiment and Task Adapters

To develop a sentiment adapter and a task adapter, we followed an efficient adapter architecture recently proposed by Pfeiffer et al. (2021). They defined the adaptor structure by simply combining down and up projection with a residual connection. To examine whether sentiment knowledge can boost the performance of the appraisal tagging task, we propose employing a sentiment adapter on pre-trained language models. Furthermore, to capture task-specific knowledge, we propose employing a task adapter on pre-trained language models. The task adapter is trained with our appraisal training data D_{train}.

We denote sentiment-specific adaptive parameters Ω and task-specific adaptive parameters Ψ. In our architecture, we allocate the sentiment-specific adapter in parallel with our task-specific adapter after the feed-forward layer, followed by a ReLU activation at each layer l:

$$\Omega_l \left(\mathbf{h}_l, \mathbf{r}_l \right) = \mathbf{U}_l \left(\text{ReLU} \left(\mathbf{D}_l \left(\mathbf{h}_l \right) \right) \right) + \mathbf{r}_l$$

$$\Psi_l \left(\mathbf{h}_l, \mathbf{r}_l \right) = \mathbf{U}_l \left(\text{ReLU} \left(\mathbf{D}_l \left(\mathbf{h}_l \right) \right) \right) + \mathbf{r}_l$$

where $\mathbf{h}_l$ is the hidden states passing through the transformer architecture and $\mathbf{r}_l$ represents the residual at layer l. $\mathbf{D}_l$ is the projection presentation at layer l. To combine two adapters, we further introduce our simple yet effective adapter weight layer Θ, for each transformer layer l, the function

denoted as:

$$\Theta_l = \alpha\Omega_l\left(\mathbf{h}_l, \mathbf{r}_l\right) + \left(1-\alpha\right)\Psi_l\left(\mathbf{h}_l, \mathbf{r}_l\right)$$

where α is the learned parameter and α can be different across transformer layers.

The sentiment adapter is fine-tuned with the sentiment datasets [2] at the sentence-level with negative log-likehood loss with $\mathcal{D}_{senti}$ as sentiment training set, as following:

$$\mathcal{L}_{senti} = -\frac{1}{|\mathcal{D}_{senti}|}\sum_{n=1}^{N}\log\left(p\left(y^n \mid x^n\right)\right)$$

5.3 Cross-lingual Inference

Follow the adapter-based cross-lingual framework MAD-X (Pfeiffer et al., 2020), we include a target language adapter when we transfer from English→Chinese and Chinese→English tasks. For instance, if we transfer from English→Chinese, we will plug in the Chinese adapter before the parallel task and sentiment adapters. Note that for cross-lingual tests, we also swap the base encoder from BERT to XLM-R to handle multi-lingual token embedding.

6 Datasets

For the cross-domain and cross-lingual appraisal tagging task, we conducted experiments on two datasets: the SOCC and POST datasets. The SOCC dataset comprises of 10,399 opinion news articles and 663,173 comments from the Canadian daily newspaper, *The Glove and Mail*. In addition to the raw text, the corpus includes specific annotations from multiple perspectives: negation, appraisal, constructiveness and toxicity. Among them, we used the appraisal annotations for our experiments. The POST dataset comprises of Twitter and Blog posts annotated with appraisals developed in-house following the same Appraisal Framework (Martin and White, 2003); it contains not only English text but also Chinese text. In the POST dataset, various spans of appraisals were annotated, as evaluation occurs at all levels of languages (words, phrases, clauses or entire sentence). In our experiments, we used the word and phrase level of labels. When there are phrases with more than one appraisal labels overlapping, we opted the longest span label.The data statistics of these two datasets are shown in Table 1 and Table 2.

[2] https://github.com/cardiffnlp/tweeteval

	Affect	Judgement	Appreciation	#Total
Negative	175	2,342	2,350	4,867
Positive	46	469	1,173	1,688
Neutral	5	10	53	68
#Total	226	2,821	3,576	6,623

Table 1: Statistics of Appraisal labels in the SOCC dataset

	Affect	Judgement	Appreciation	#Total
Blogs				
Negative	548	758	333	1,639
Positive	505	397	256	1,158
Neutral	124	53	24	201
Tweets-English				
Negative	117	335	138	590
Positive	112	155	82	349
Neutral	50	25	28	103
Tweets-Chinese				
Negative	29	144	78	251
Positive	50	80	97	227
Neutral	10	12	14	36
#Total	1,545	1,959	1,050	4,554

Table 2: Statistics of Appraisal labels in the POST dataset

7 Experiments and Results

7.1 Experiment Setup

A^2 is an adapter-based framework adding adapter modules to pre-trained language models; A^2(BERT) and A^2(RoBERTa) are based on language models, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), respectively. We compared the performance of A^2 framework models with following baseline models:

- Conditional Random Fields (CRF) (Lample et al., 2016) is a baseline that is widely used, feature-based model for a sequence tagging task.
- GLoVe+FFN is a baseline where tokens are encoded by the max and mean of GloVe (Pennington et al., 2014) [3] embeddings and followed by a feedforward neural network (FFN) for sequence labelling.
- BERT (Devlin et al., 2019) is a baseline based on the off-the-shelf BERT token embeddings, where we include further pre-training (BERT-PT) (Gururangan et al., 2020) and fine-tuning (BERT-FT) on a sentiment dataset.

[3] https://github.com/stanfordnlp/GloVe

	#Total	Avg. Appraisal span length
News Comments	6,623	6.11
Blogs	2,998	12.02
Tweets	1,042	4.46
Tweets(CHN)	514	3.01

Table 3: Statistics of Appraisal expressions in different domains. Span length is counted by # of tokens.

- RoBERTa (Liu et al., 2019) is a baseline similar to BERT, where we include further pre-training (RoBERTa-PT) and fine-tuning (RoBERTa-FT) on a sentiment dataset.
- M-BERT is a baseline based on multilingual BERT.
- XLM-R (Conneau et al., 2019) is a baseline based on the multilingual language model XLM-R.

7.2 Implementation Details

We implemented our models in PyTorch using the HuggingFace library[4] and their pretrained BERT[5] and RoBERTa[6] models. Adapters in language models are implemented with the AdapterHub [7] package. The 100-dimension GloVe word embedding is applied for the GloVE+FFN model.

For the input sequence, we set maximum token length= 384 and dropout rate $= [0.5, 0.6]$ for token embeddings. Learning rate is tuned in the range between $[1e^{-5}, 5e^{-5}]$ for BERT and $[1e^{-6}, 5e^{-6}]$ for RoBERTa based on the development set. All models use the Adam optimiser (Kingma and Ba, 2014), and our experiments are run using one A100 GPU with 40GB Memory.

7.3 Results

We evaluated A^2 for each attitude and polarity category. Each result is an average of three runs with different random seeds. For pre-training, we followed the procedures in (Gururangan et al., 2020) with masked LM loss. Table 4 and Table 5 present the in-domain test results on the SOCC and POST datasets, respectively.

As shown in the first group of both Table 4 and Table 5, compared with CRF and GloVe+FNN models, both BERT- and RoBERTa-based models yield better performance based on the average F1 scores on both datasets. This indicates that existing pre-trained language models have better capability to handle the appraisal tagging task. Moreover within the off-the-shelf language models, RoBERTa generally performs better than BERT.

The second group of both Table 4 and Table 5 shows the results of the continued pre-training and sequential fine-tuning language models. Comparing the performance of the first group of off-the-shelf language models, we found that our models with both strategies boost the performance further on all the Attitude (*Affect*, *Appreciation* and *Judgement*) labels and Polarity (*Positive*, *Negative* and *Neutral*) labels. The improved performance on the polarity labels especially implies that sentiment knowledge is useful for the appraisal tagging task. On the question of which strategy works better for this task, we observed that the continued pre-training strategy generally can bring better performance compared with the sequential fine-tuning strategy on both datasets in most of the cases except on the average F1 scores (e.g. 65.24% vs. 65.48%).

The main results of the appraisal tagging performance on our A^2 models are presented in the last group of both Table 4 and Table 5. As shown at the overall F1 score, both our A^2 models yield strong performance on these two datasets compared with all the other models. Note that we only fine-tuned the adapter parameters, which only apply to 15% of the overall number of parameters in the transformer-based language models.

7.4 Ablation Study

We compared the "jointly adap" approach of our A^2 framework (A^2) with the "sequentially adapt" approach in the literature (Pfeiffer et al., 2020), where the task-specific adapter is stacked on top of the sentiment-fused adapter (A^2_{seq}). As shown in the first group of Table 6, the joint adapters with weighting layer (A^2) gives better performance compared with stacking both adapters (A^2_{seq}) for the appraisal tagging task.

Furthermore, we conducted an ablation study to compare the performance of our A^2 framework against one without the task adapter (w/o T-Adpt) and one without the sentiment adapter (w/o S-Adpt). As presented in the second group of Table 6, the model with both sentiment and task adapters (A^2) performs the best and the sentiment adapter

[4] https://github.com/huggingface
[5] https://huggingface.co/bert-base-cased
[6] https://huggingface.co/roberta-base
[7] https://adapterhub.ml/.

Model	Avg. F1	Affect	Appreciation	Judgement	Positive	Negative	Neutral
CRF	42.67	35.69	38.24	40.77	44.26	52.25	41.55
GloVe+FFN	51.23	45.15	48.35	51.66	52.09	54.01	52.52
BERT	58.07	50.37	59.95	60.59	63.73	61.56	52.27
RoBERTa	59.47	53.66	60.56	62.54	60.63	63.77	55.64
BERT-PT	65.24	60.14	63.24	71.17	67.75	68.04	61.09
BERT-FT	65.48	61.48	65.15	70.05	66.68	69.37	60.14
RoBERTa-PT	67.10	61.66	66.28	73.34	68.04	71.85	61.44
RoBERTa-FT	66.93	61.90	66.57	69.01	67.41	70.79	65.89
A^2(BERT)	68.20	65.58	68.82	71.25	67.75	72.02	63.78
A^2(RoBERTa)	69.81	67.71	69.90	73.41	68.89	73.34	65.50

Table 4: F1 scores of all models on the SOCC test set ("PT" = pre-training, "FT" = fine-tuning)

Model	Avg. F1	Affect	Appreciation	Judgement	Positive	Negative	Neutral
CRF	47.51	43.56	43.03	44.79	49.97	44.88	44.66
GloVe+FFN	61.92	53.70	59.59	63.56	64.31	68.44	60.54
BERT	68.45	62.56	67.69	71.23	74.45	69.28	65.50
RoBERTa	69.91	63.43	69.81	70.49	75.97	72.21	67.56
BERT-PT	71.68	68.07	70.07	76.42	75.21	73.38	66.90
BERT-FT	69.17	66.16	69.76	72.25	71.30	70.68	64.55
RoBERTa-PT	73.01	69.98	71.72	78.63	74.11	75.32	68.35
RoBERTa-FT	71.72	69.50	72.65	77.11	73.28	74.32	63.46
A^2(BERT)	72.45	71.25	69.38	71.83	77.01	77.85	67.36
A^2(RoBERTa)	74.86	74.45	71.62	74.50	78.18	79.45	70.95

Table 5: F1 scores of all models on the POST English test set ("PT" = pre-training, "FT" = fine-tuning)

brings major improvement to model performance on the POST test set, by delivering $+8.23$ and $+8.2$ F1 scores on Polarity labels, *Positive* and *Negative*, respectively.

The experiment results indicate that the joint adapter approach can fully leverage both sentiment-specific and task-specific information through our adapter architecture and sentiment knowledge can greatly enrich token semantics.

7.5 Cross-domain Performance

This set of experiments aim to answer our second research question (RQ2), "*Can the sentiment knowledge fused adapter improve the adapter-based framework for the cross-domain appraisal tagging?*". As shown in Table 3, the average appraisal expression length varies in different domains; for instance, Twitter appraisals (avg. 4 tokens) have less number of tokens compared with News Comment appraisals (avg. 6 tokens) and Blog appraisals (avg. 12 tokens). As the SOCC and POST datasets contain corpus from three different domains (Tweet, Blog and News Comment) we got 6 sets of cross-domain performance results in total: Tweet→Blog, Tweet→News Comment, Blog→Tweet, Blog→News Comment, News Comment→Tweet, and News Comment→Blog. For instance, under Tweet→Blog setting, we will use Tweet domain text as training and zero-shot testing the performance on Blog domain text. Table 7 shows the average F1 scores of the cross-domain performance results in the zero-shot setting.

When comparing the cross-domain performance of the baseline models (BERT and RoBERT) against the fine-tuned models (BERT-FT and RoBERT-FT), we observe that directly fine-tuning language models with a sentiment dataset does not always yield performance improvement. For example, while the BERT-FT improves its performance (from 0.61 to 5.58 of F1 score) than the BERT on the most of the cross-domain settings (except one 'News Comment→Blog' setting), the RoBERTa-FT worsens its performance (from -0.55 to -1.83 of F1 score) than the RoBERTa on the most of the cross-domain settings (except one 'Tweet→News Comment setting). This result indicates that directly fine-tuning language models without considering domain specific vectors can hurt naive trans-

Model	Avg. F1	Affect	Appreciation	Judgement	Positive	Negative	Neutral
A^2	74.86	74.45	71.62	74.50	78.18	79.45	70.95
A^2_{seq}	72.73	70.70	68.07	77.65	75.94	75.80	68.24
w/o T-Adpt	69.37	66.85	65.65	72.38	71.97	72.80	66.56
w/o S-Adpt	69.09	68.82	66.50	73.50	69.95	71.25	64.49

Table 6: F1 scores of RoBERTa-based models on the POST datasets ("T-Adpt" = task adapter, "S-Adpt" = sentiment adapter)

Model	T		B		N	
	B	N	T	N	T	B
BERT	50.81	53.45	52.87	45.13	56.21	53.17
BERT-FT	51.43	54.72	53.04	50.71	56.82	51.61
RoBERTa	52.24	56.26	54.64	48.03	59.29	55.45
RoBERTa-FT	51.69	57.62	52.81	52.54	58.10	54.44
A^2(BERT)	55.61	61.22	63.14	57.45	67.33	60.89
A^2(RoBERTa)	59.79	62.67	64.61	61.55	68.12	63.34

Table 7: Cross-domain performance (F1 score, "T" = Tweet, "B" = Blog, "N" = News Comment)

Model	eng→chn		chn→eng	
MBERT	68.11	61.07	72.64	52.12
XLM-R	70.71	63.77	76.06	56.69
A^2(MBERT)	69.45	64.33	71.54	61.78
A^2(XLM-R)	72.64	67.91	75.44	62.67

Table 8: Cross-lingual performance on the POST dataset ("eng" = English, "chn" = Chinese)

fer learning. On the other hand, our A^2(BERT) and A^2(RoBERTa) models achieve better performance than the fine-tuned models (BERT-FT, RoBERTa-FT) on all the six cross-domain settings. Moreover, the A^2(RoBERTa) model achieves consistently better performance than the A^2(BERT) model on all the six cross-domain settings.

When comparing the cross-domain performance (shown in Table 4 and Table 5) against the in-domain performance (shown in Table 7), we can see that all language models show substantial drop in performance. For example, on News Comments, the performance of the BERT-FT drops from 65.48 to 54.72 and 50.71 when it is originally trained on Tweet domain but test on News Comment domain. This demonstrates the challenge of the cross-domain task setting, which may contain a catastrophic forgetting issue, conflicting signals and domain requirements. However, when we incorporate adapters in A^2 framework, we observe that the performance gap diminishes significantly.

7.6 Cross-lingual Performance

Table 8 shows the cross-lingual performance results, from English to Chinese with tweets data and vice versa. To answer RQ3, "*Can the sentiment knowledge fused adapter improve the adapter-based framework for the cross-lingual appraisal tagging?*", we first fine-tuned the multilingual BERT (MBERT) and XLM-RoBERTa (XLM-R) models using the labels in the source language and applied them to the target languages with subword embeddings frozen.Then, we compared with them under a simple set up of our A^2 framework with plugging in a target language specific adapter. Both A^2(MBERT) and A^2(XLM-R) models demonstrate performance gains (avg. 6.96% and 5.06% for MBERT and XLM-R based) on the target language.

We can see when we transfer from source to target language, existing multilingual language models perform poorly. For example, with XLM-R, when we test out eng→chn, there is a huge performance drop from 76.06% to 63.77% on Chinese data. Both A^2 models are slightly under-perform (avg. 0.85%) the base language models when Chinese as the source language, which may due to the sentiment adapter trained with English data only. Note that the performance on the source language does not decrease as we only replace the language-specific adapter at the inference time.

8 Illustration of a prediction error

To provide a qualitative analysis for our approach, we showcase an example of an annotated sentence from the SOCC dataset in Table 9. We present our A^2 prediction v.s. the human annotations. The original sentence is a comment towards an article discussing the aboriginal of Canada [8]. Based on

[8] https://www.theglobeandmail.com/opinion/too-many-first-nations-people-live-in-a-dream-palace/article6929035/

Sentence	The author does not seem to have much of a clue in spite of her elevated status.
A^2 prediction	The author does not seem to have much (Jud Neg) of a clue in spite of her elevated status. (App Neg)
Gold annotation	The author does not seem to have much of a clue (Jud Neg) in spite of her elevated status.

Table 9: Illustration of a prediction error

the human annotations, the whole span of "does not seem to have much of a clue" labelled with *Judgement* Attitude and *Negative* Polarity. Our A^2 framework can accurately predict the *Judgement* Attitude and *Negative* Polarity labels but missing three following tokens. It also falsely tags the text segment, "elevated status" as *Appreciation* Attitude and *Negative* Polarity.

9 Conclusion

We have proposed A^2, an adapter-based framework for automatically tagging Appraisal expressions. We have designed task and sentiment adapters of a small number of additional parameters to improve the capacity of pre-trained language models for quick adaptation for cross-domain and cross-language settings.

Limitations

Our system is based on the pre-trained language models and therefore assumes that GPU resources are available. The system is designed for tagging short opinion-bearing texts and the maximum text length is set to 384 tokens. Moroever, our system only performed cross-lingual tests from English to Chinese and Chinese to English for experiments. This is mainly constrained by the dataset availability. To our best knowledge, the POST dataset is the only available resource in Chinese with appraisal annotations. It is desirable to conduct more experiments on a broader set of languages to evaluate the generalisability of the A^2 model for cross-lingual adaptation.

Acknowledgements

This research is supported in part by the Australian Research Council Discovery Project DP200101441 and in part by the Defence Science and Technology Group, Australia.

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Soravit Changpinyo, Hexiang Hu, and Fei Sha. 2018. Multi-task learning for sequence tagging: An empirical study. In *COLING*.

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. *arXiv preprint arXiv:1911.02116*.

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.

Suzanne Eggins. 2004. *Introduction to systemic functional linguistics*. A&c Black.

Phoebe C Ellsworth and Craig A Smith. 1988. From appraisal to emotion: Differences among unpleasant feelings. *Motivation and emotion*, 12(3):271–302.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Maximilian Hofer, Andrey Kormilitzin, Paul Goldberg, and Alejo Nevado-Holgado. 2018. Few-shot learning for named entity recognition in medical text. *arXiv preprint arXiv:1811.05468*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *ICML*.

Katharina Kann, Johannes Bjerva, Isabelle Augenstein, Barbara Plank, and Anders Søgaard. 2018. Character-level supervision for low-resource pos tagging. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 1–11.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2832–2838.

Young-Bum Kim, Karl Stratos, Ruhi Sarikaya, and Minwoo Jeong. 2015. New transfer learning techniques for disparate label sets. 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 473–482.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2020. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4(2):155–190.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2018a. Transfer learning for named-entity recognition with neural networks. In *Language Resources and Evaluation Conference*.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2018b. Transfer learning for named-entity recognition with neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced nlp tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting bert for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:1910.00883*.

Bing Liu. 2022. *Sentiment analysis and opinion mining*. Springer Nature.

Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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*.

James R Martin and Peter R White. 2003. *The language of evaluation*, volume 2. Springer.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Teresa Oteíza. 2017. The appraisal framework and discourse analysis. In *The Routledge handbook of systemic functional linguistics*, pages 481–496. Routledge.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Jonathon Read and John Carroll. 2012. Annotating expressions of appraisal in english. *Language resources and evaluation*, 46(3):421–447.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. Udapter: Language adaptation for truly universal dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 625–631.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. In *ACL*.