Chinese Opinion Role Labeling with Corpus Translation: A Pivot Study

Ranran Zhen[1], Rui Wang[2], Guohong Fu[3]*, Chengguo Lv[1], Meishan Zhang[4]
[1]School of Computer Science and Technology, Heilongjiang University, China
[2]Vipshop (China) Co., Ltd.
[3]Institute of Artificial Intelligence, Soochow University, China
[4]School of New Media and Communication, Tianjin University, China
{zenrran, mason.zms}@gmail.com
{mars198356, ghfu}@hotmail.com
2004085@hlju.edu.cn

Abstract

Opinion Role Labeling (ORL), aiming to identify the key roles of opinion, has received increasing interest. Unlike most of the previous works focusing on the English language, in this paper, we present the first work of Chinese ORL. We construct a Chinese dataset by manually translating and projecting annotations from a standard English MPQA dataset. Then, we investigate the effectiveness of cross-lingual transfer methods, including model transfer and corpus translation. We exploit multilingual BERT with Contextual Parameter Generator and Adapter methods to examine the potentials of unsupervised cross-lingual learning and our experiments and analyses for both bilingual and multilingual transfers establish a foundation for the future research of this task[1].

1 Introduction

Fine-grained opinion mining has been a crucial task in natural language processing (NLP) for a long time (Kim and Hovy, 2006a; Breck et al., 2007; Wilson et al., 2009; Qiu et al., 2011; Irsoy and Cardie, 2014; Liu et al., 2015; Wiegand et al., 2016) and it aims to discover useful structural information of user opinions from unstructured text, which is the relation between expression and entities such as *Who expressed what kind of sentiment towards what?*. The EXPRESSION conveys attitudes including sentiments, agreements, beliefs, or intentions (e.g., voiced his condolences in Figure 1); the entities consist of the HOLDER who expresses the opinion (e.g., Chen.) and the TARGET which the opinion is expressed to (e.g., the families) (Breck et al., 2007; Yang and Cardie, 2012; Katiyar and Cardie, 2016a). Here we focus on the opinion role labeling (ORL) task which is to identify opinion holders and targets (Marasović and Frank, 2017; Zhang et al., 2019a) when the expressions are given.

[Chen]$_{\text{holder}}$ [voiced his condolences]$_{\text{expression}}$ to [the families]$_{\text{target}}$.

Figure 1: An example of fine-grained opinion mining.

Most of the previous researches focus on the English ORL, benefiting from the benchmark MPQA dataset (Wiebe et al., 2005) which includes span-based annotations of opinion expressions, holders and targets. The task is commonly solved by sequence labeling models with the BIO conversion scheme (Kim and Hovy, 2006b; Choi et al., 2006; Yang and Cardie, 2013; Johansson and Moschitti, 2013). Recently, neural BiLSTM-CRF models have achieved state-of-the-art performance on this task (Katiyar and Cardie, 2016b; Marasović and Frank, 2017; Zhang et al., 2019a). However, the studies on other languages are relatively rare due to the scarcity of annotated datasets. To our best knowledge, there is only one exception by Almeida et al. (2015a), which has annotated a small-scale dataset for the Portuguese language.

Unsupervised cross-lingual transfer (Xu et al., 2018) is one promising way to address the low resource problem for ORL. Under the neural setting, there are two representative categories of methods: model transfer (McDonald et al., 2013; Swayamdipta et al., 2016; Daza and Frank, 2019) and corpus translation (Zhang et al., 2019b). The model transfer trains a model on a resource-rich language by using only language-independent features such as multilingual BERT (Devlin et al., 2018; Pires et al., 2019) and then apply it to the target language. The corpus translation approach firstly obtains parallel corpora through either human or machine translation and then projects the annotations from the source language to the target side.

In this work, we present the first study of the

*Corresponding author.

[1]We release the code and way of obtaining Chinese dataset at https://github.com/zenRRan/ChineseORL-with-Corpus-Translation.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10139–10149
November 7–11, 2021.

Chinese ORL. First, we construct a benchmark corpus by manually translating the English MPQA corpus, which involves auto-translation (i.e., automatic sentence translation and opinion aligning) and human refinement. Second, we investigate the performance of unsupervised cross-lingual transfer for Chinese ORL based on the annotated corpus. We investigate the Contextual Parameter Generator Networks (PGN) in multilingual BERT with Adapter (known as parameter efficient in learning) method (Üstün et al., 2020) (we call it PGN-Adapter) and discover the complementarity of the model transfer and corpus translation methods.

We conduct experiments on the newly constructed Chinese dataset to evaluate our methods, together with the English MPQA corpus (Wiebe et al., 2005) and the Portuguese dataset (Almeida et al., 2015a) for cross-lingual transfer. We observe that for the unsupervised cross-lingual transfer from the English corpus, the translation-based method is better than the model transfer, and their combination leads to further improvements. Although the scale of the Portuguese corpus is much smaller, adding it into the multilingual transfer still outperforms the bilingual counterpart.

To summarize, in this paper, we have the following contributions:

- We manually translate and annotate a Chinese fine-grained ORL corpus for research purposes, especially for the cross-lingual ORL study.
- We conduct cross-lingual ORL (to Chinese) through unsupervised model transfer and corpus translation with PGN-Adapter, setting up strong baselines for future research.
- We perform extensive experiments and analyses to demonstrate the pros and cons of the different approaches.

2 Related Work

Fine-Grained Opinion Mining There have been a number of studies in fine-grained opinion mining (Wilson et al., 2009; Qiu et al., 2011; Wiegand et al., 2016). Kim and Hovy (2006a) exploit a semantic role labeller to extract opinion holders and topics. Choi et al. (2005) and Breck et al. (2007) model the task by sequence labeling with CRF to discover opinion holders and recognize opinion expressions, respectively. Yang and Cardie (2012)'s semi-Markov CRF model outperforms the standard CRF, and Irsoy and Cardie (2014) and Liu et al. (2015) use recurrent neural network for opinion mining. Johansson and Moschitti (2013) and Katiyar and Cardie (2016a) propose joint models for opinion expressions, holders and targets.

Opinion Role Labeling As for ORL, Marasović and Frank (2018) exploit multi-task learning about how to use SRL information to improve ORL scores. Zhang et al. (2019a) utilize semantic role labeling to enhance ORL, where three different integrating approaches are compared. Bo et al. (2020) propose a dependency-based graph convolutional networks to enhance ORL with syntax information. All these studies focus on the English ORL by using supervised models, assuming that a training corpus is already available. In this work, we investigate Chinese ORL, building a benchmark dataset for Chinese manually and then studying unsupervised cross-lingual transferring for the task.

Cross-Lingual Transfer Learning Cross-lingual transfer learning has been extensively applied in NLP, including sentiment classification (Zhou et al., 2016), POS tagging (Täckström et al., 2013; Wisniewski et al., 2014; Kim et al., 2017), named entity recognition (Zirikly and Hagiwara, 2015), semantic role labeling (Fei et al., 2020), and dependency parsing (McDonald et al., 2011; Tiedemann et al., 2014; Guo et al., 2016; Zhang et al., 2019b). Unsupervised cross-lingual transferring has received great interest (Duong et al., 2015; Xu et al., 2018), which is our major focus. The work of Zhang et al. (2019b) is mostly related to our study, which applies model transferring and corpus translation to dependency parsing. Our work focuses on ORL, applying the two approaches for the Chinese language.

PGN-Adapter PGN-Adapter is used for merging different languages to same space by Adapter and PGN methods with BERT. About the Adapter method which a pre-trained network added between the transformer encoder layer, there are many studies on using adapter modules (Rebuffi et al., 2018; Stickland and Murray, 2019; Houlsby et al., 2019). PGN is first proposed by Platanios et al. (2018) for universal neural machine translation task. And Üstün et al. (2020) integrated that two methods above in dependency parsing which inspired us to merge this idea into our ORL task.

The transfer method doesn't work in Eger et al.

Raw	The president had sidelined Masire after accusing him .
Translated	这位总统在指控马西尔之后退居二线。
Revision	在指控马西尔之后，总统把他排挤到了一边。
Raw	Russian guards seize 87 kg of heroin on Tajik-Afghan border .
Translated	俄罗斯警卫在塔吉克-阿富汗边境抓获87公斤海洛因。
Revision	俄罗斯士兵在塔吉克-阿富汗边境缴获87公斤海洛因。

Table 1: Two examples of manual revisions for automatic translations, where the first example indicates a translation error, and the second example indicates an improper translation.

(2018), because 1) Google's MT system is much better in 2020 than in 2018. Lower translation quality causes more problems for the argument mining; while high-performance MT system enhances the translation-based approach. 2) Projection strategy is also different from ours (Section 5.1). We choose to project the non-cross labels only, in order to ensure the mapping quality. 3) In addition, the more advanced methods like PGN, Adapter and BERT also play a significant role in cross-lingual tasks.

3 The Construction of Chinese Dataset

We manually construct a Chinese ORL dataset to facilitate our research. In order to reduce the overall cost, we exploit corpus translation to assist the construction process, converting the English MPQA corpus (Wiebe et al., 2005; Wilson, 2008) into Chinese. The conversion contains the following four steps by order: (1) sentence translation, (2) manual revision, (3) opinion projection, and (4) manual correction. The first and third steps formalize into automatic corpus translation, which has been used as one approach for unsupervised cross-lingual transfer, and the second and fourth steps are used to ensure the final quality. The whole construction is conducted at the sentence-level.

Sentence Translation Neural machine translation (NMT) has achieved state-of-the-art performances for a range of language pairs (Vaswani et al., 2017). In particular, the state-of-the-art NMT can reach a BLEU score over 45 (Li et al., 2019). Thus it is applicable to use NMT for automatic sentence translation. Here we first translate all the English sentences of the MPQA dataset into Chinese by using the google translator[2] automatically.

Manual Revision Next, we let several native speakers check the translation quality, and make revisions to the imperfect translations. There can be two types of revisions. On the one hand, the translated sentences may have errors, and human intervention is required to correct these issues. On the other hand, the automatic sentences may not match the style of native speakers, and we let our annotators rewrite these sentences. Table 1 shows two examples of the two conditions, respectively.

Opinion Projection Third, we project all opinions (expressions, holders and targets) from the English sentence into its Chinese translation. Before the projection, we use the Stanford Segmentor tool for word segmentation[3]. The overall projection is supported by automatic word alignments, which can be produced by using a word-alignment tool. Here we exploit the fast-align tool[4] (Dyer et al., 2013) to calculate the alignment probabilities.

Figure 2 shows an example to illustrate the projection process. Concretely, given an English-Chinese sentence pair $(e_1 \cdots e_n, c_1 \cdots c_m)$ and its English-to-Chinese alignment probabilities $a(c_j|e_i)$, the projection is performed as follows:

(1) We incrementally obtain the text spans in the Chinese sentences for the opinion expressions as well as their holders and targets in the English sentence.

(2) For each word e_i in the English sentence, we find its corresponding word c_{p_i} in the Chinese sentence by using $p_i = \arg\max_j a(c_j|e_i)$, resulting in a set of one-one mapping word pairs: $M = \{(e_1, c_{p_1}), \cdots, (e_n, c_{p_n})\}$.

(3) For each span $e_{i,j}$ (i.e., expression, holder or target) in the English sentence, we find its corresponding span $c_{i',j'}$ in the Chinese sentence by maximizing the covered word-pair set M with the least span length.

(4) We remove the projected span when $(j' - i') \geq 2 * (j - i + 1)$ which is regarded as low-quality. If one expression is removed, its holder and target are removed as well.

Manual Correction The last step is to perform another checking manually to ensure the quality of automatic opinion projection. There could be several types of errors, including word boundary errors and miss-alignments. And as for the continuity and fluency in the sentence, we do some trade-offs as shown in Table 2.

By using the above four steps, we can obtain a benchmark dataset for Chinese ORL, the argument comparison about Chinese and English can

[2]https://translate.google.com/

[3]https://nlp.stanford.edu/software/segmenter.html

[4]https://github.com/clab/fast_align

Problem	English	Chinese
Continuity	But $\text{anyone}_{\text{holder}}$ who wants to speak the language of violence ... ... rein in militants $\text{keen to fight}_{\text{expression}}$ President Mugabe 's rule .	但 任何[any] 想 在 ... 上 讲 暴力 语言 的 $\text{人[person]}_{\text{holder}}$ 都 是 在 玩火 。 ... 控制 那些 $\text{持续[keep] 与 穆加贝 总统 的 统治 作 斗争[to fight]}_{\text{expression}}$ 的 激进分子
Fluency	With a $\text{mandate}_{\text{expression}}$ to cut costs , Goldin implemented ...	为了[to] 降低[cut] 成本[cost] ， 高银[Goldin] 实施[implement] ...

Table 2: There are some cases of conflict problems between Chinese and English. Two main items are continuity and fluency in some sentences. The first and second cases are about one English word or continuous phrase translate into two discontinuous Chinese words, we will only annotate the core continuous part. For semantic fluency after translation, some English words will not be translated, as shown in the last case.

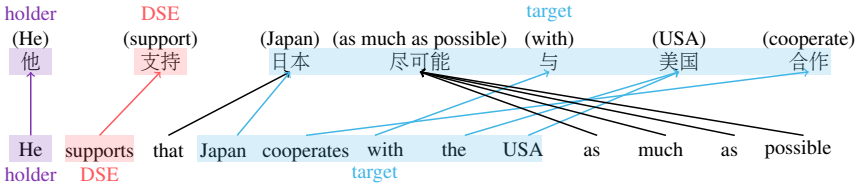


Figure 2: An example to illustrate automatic opinion projection. Noticed that DSE is Direct Subjective Expression which is all about expression descriptions in this paper.

Section	Length			Distance	
	E	H	T	E-H	E-T
Chinese					
Train	1.27	2.07	4.56	3.89	3.92
Dev	1.55	1.91	3.87	4.17	3.94
Test	1.27	2.03	4.76	3.62	3.50
English					
Train	1.41	2.34	5.48	3.08	4.42
Dev	1.42	2.41	5.12	3.26	4.56
Test	1.42	2.29	5.61	3.23	3.91

Table 3: This is an argument comparison of Chinese and English on word-level length and distance. E, H and T indicate Expression, Holder and Target argument.

be seen in Table 3. *Noticeably, when we apply the corpus translation approach for unsupervised cross-lingual transferring, only the first and third steps are required, and all human interventions would be removed for full automation.* For corpus translation of other language pairs, one just need to replace English by the desired source language, and Chinese by the desired target language.

For manual revisions of the translated sentences and corrections of final ORL annotations, we recruit three volunteers, which are all native Chinese speaking and fluent in English. For translation part, it is based on the result of machine translation and then we made a few minor corrections. Two students modify it first by themselves, and then they proofread it together to select the better one between the two cases they translated. If they have a conflict, the third one can make suggestions and get a final version. As long as the sentence's meaning is correct and conforms to Chinese habits, so we did not calculate translation's kappa. The kappa scores of word alignment are much higher, so we omit it in the paper. Since the corpus is not built from scratch, it is constructed by translation. Thus, the selected sentences are directly sourced from the English MPQA corpus. The manual efforts of translation, as well as the alignment, are highly straightforward with little ambiguities. For word alignment part, it is similar to the translation part.

4 Model

Opinion role labeling aims to discover the opinion arguments given opinion expressions. The task can be modeled as a sequence labeling problem (Zhang et al., 2019a). We adopt the BMESO scheme to convert spans of opinion arguments into a sequence of word-level boundary tags, where B, M and E denote the beginning, middle and ending words of an argument, respectively, S denotes the word itself is an argument, and O denotes the rest of the words. Formally, assuming that the input sentence is sent $= w_1, ..., w_n$, and a given opinion expression is expr $= w_b, ..., w_e (1 \leq b \leq e \leq$n), our task is to assign a sequence of boundary tags $t_1, ..., t_n$. We exploit a BiLSTM-CRF framework based on PGN-Adapter to implement our model. Figure 3 shows an overview of our model.

PGN-Adapter This model (Üstün et al., 2020) is based on the traditional BERT architecture (Devlin et al., 2018). Let the whole input sentence $w_1, \cdots, w_n$ which is decomposed into the word-piece sequences $c_1, \cdots, c_m$ and get the input representation $r_i, \cdots, r_m$ by summing each c_i and position embedding p_i. Then, each r_i is passed to a stacked self-attention layers (SelfAttn) to generate

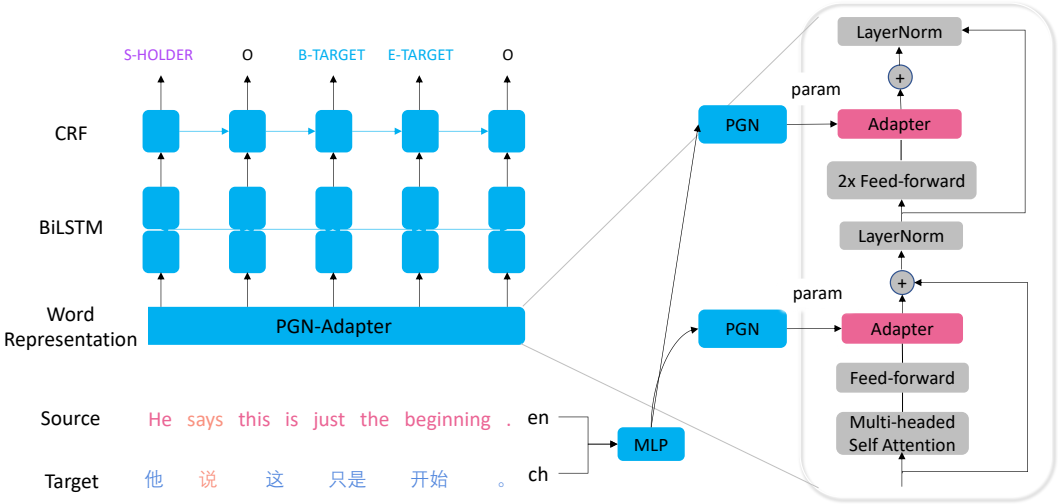


Figure 3: The model architecture of ORL by PGN-Adapter method with the expression "says" given.

the final encoder representation $\boldsymbol{e}_i$:

$$\begin{aligned} \boldsymbol{r}_i &= \boldsymbol{c}_i + \boldsymbol{p}_i \\ \boldsymbol{e}_i &= \text{SelfAttn}(\boldsymbol{r}_i; \boldsymbol{\theta}^{\text{ada}}) \end{aligned} \tag{1}$$

where $\boldsymbol{\theta}^{\text{ada}}$ denotes the adapter modules. Following Houlsby et al. (2019), in this module, two adapters with two feedforward projections and a GELU nonlinearity are merged into each transformer layer as shown in Figure 3.

To obtain the amount of sharing cross languages, we generate the trainable parameters of the adapter module by the PGN method. The weights of the adapters are following:

$$\boldsymbol{\theta}^{\text{ada}} = \boldsymbol{W}^{\text{ada}} \cdot \boldsymbol{I}_{\text{e}} \tag{2}$$

where $\boldsymbol{W}^{\text{ada}}$ is the parameters in adapter modules. The parameters in the BERT model are frozen except the adapter part, thus our model can be much more parameter-efficient than fine-tuning, and meanwhile, our preliminary results show that the method does not hurt the performance. $\boldsymbol{I}_{\boldsymbol{e}}$ is a language embedding by a multi-layer perceptron MLP^{lang}:

$$\boldsymbol{I}_e = \text{MLP}^{\text{lang}}(\boldsymbol{I}_t) \tag{3}$$

where $\boldsymbol{I}_t$ is a typological feature vector from the URIEL language typology database (Littell et al., 2017). Following Üstün et al. (2020), we set our language embeddings from syntactic, phonological and phonetic inventory feature with k-nearest neighbors approach.

Further, word-level representations $\boldsymbol{x}_1 \cdots \boldsymbol{x}_n$ can be derived by averaged pooling over the covered word pieces of each word.

BiLSTM-CRF Following the front outputs, we apply bi-directional LSTMs (BiLSTM) and conditional random fields (CRF) (Lafferty et al., 2001) to get high-level features and compute the probability of each candidate output tag sequence $y = y_1, \cdots, y_n$. The concrete calculation method is defined as follows:

$$\begin{aligned} \boldsymbol{h}_1 ... \boldsymbol{h}_n &= \text{BiLSTM}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \\ \boldsymbol{o}_i &= \boldsymbol{W}\boldsymbol{h}_i, i \in [1, n] \\ \text{Score}(y) &= \sum_{i \in [1,n]} \boldsymbol{o}_i \left[y_i\right] + \boldsymbol{T}\left[y_i, y_{i-1}\right] \\ p(y|\text{sent}, \text{expr}) &= \frac{e^{\text{Score}(y)}}{\sum_{y'=y'_1,\cdots,y'_n} e^{\text{Score}(y')}} \end{aligned} \tag{4}$$

where y' traversing all candidate outputs, $\boldsymbol{W}$ and $\boldsymbol{T}$ are parameters. We use the Viterbi algorithm based on $\text{Score}(y)$ to search for the ORL tag sequence of the maximum score.

Training Objective We exploit sentence-level cross-entropy loss for model training, which can be described as:

$$\mathcal{L} = -\log p(g = g_1, ..., g_n|\text{sent}, \text{expr}) \tag{5}$$

where $g = g_1, ..., g_n$ denotes the gold-standard tag sequence of a given sentence-expression pair.

Section	#sent	Holder	Target
English			
Train	4846	2438	2533
Dev	2298	1196	1259
Test	1435	779	802
Chinese			
Train	4846	2417	2457
Dev	2298	1196	1259
Test	1435	759	779
Train^{en}_{auto}	4846	1849	1822
$\text{Train}^{en}_{half-auto}$	4846	1805	1722
Portuguese			
ALL	1226	661	769

Table 4: Data statistics of different languages, where #sent is the sentence number. Holder and Target indicate how many labeled themselves there are in the corresponding datasets.

Model Transfer We use multilingual word representations to achieve the cross-lingual transfer of the model. In particular, we use the pre-trained multilingual BERT-Base (cased version)[5] (Devlin et al., 2018). All pretrained parameters inside the BERT are frozen during ORL training.

5 Experiments

5.1 Datasets

English Dataset. We use the widely-adopted ORL benchmark dataset, the MPQA version 2.0 (Wiebe et al., 2005; Wilson, 2008) to evaluate our models. We focus on identifying expression-holder and expression-target relations with expressions given. We split the whole corpus into fixed training, development and testing sections.

Chinese Dataset. We construct the Chinese dataset as described in Section 3, and the basic statistics of the corpora are shown in Table 4 where we should point out that the reason for the number of auto-annotated opinion-arguments reduction is that we removed the cross labels. The Chinese dataset consists of three parts: 1) manually translated and word-aligned from the English dataset; 2) manually translated but automatically word-aligned ($\text{Train}^{en}_{half-auto}$); and 3) automatically translated and word-aligned (Train^{en}_{auto}). *The data splitting method is directly mapped from the corresponding English divisions for a fair investigation.*

Portuguese Dataset. We use the Portuguese ORL dataset released by Almeida et al. (2015b) for the cross-lingual transfer as well. The whole dataset is used to help Chinese ORL.

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

5.2 Evaluation Metrics

As usual, we use precision (P), recall (R) and (Exact) F1-score to evaluate our methods. Following (Marasović and Frank, 2017), we exploit two additional soft evaluation metrics, binary and proportional overlapping scores. In detail, Binary F1 treats an entity as correct if it contains an overlapping region with the gold-standard, and the proportional F1 assigns a partial score proportional to the ratio of the overlapping region.

5.3 Settings

The implementation of models of our experiments are Pytorch with version 1.4 and GPU device with V100 (32G). There are several hyper-parameters contained in the model. We set the output hidden size of BiLSTM to 200 and the layer number of BiLSTM to 3. To prevent overfitting, we set the dropout rate to 0.33. As for PGN-Adapter, we set the language embedding size in [16, 32, 50, 64], adapter size in [128, 256, 512], language embedding dropout rate is 0.1.

We exploit online training to learn model parameters, and use the Adam optimizer (Kingma and Ba, 2014) with learning rate 0.002. And the adapter module learning rate is 5e-6 by AdamW optimizer (Loshchilov and Hutter, 2017). The mini-batch size is set to 32, and the parameters of the model are updated every 4 mini-batches. We use gradient clipping by max norm 1.0. By default, we set the maximum epoch number to 40 to early stop, and evaluate development and test datasets every 160 steps. At last, we select the final model when the development's result is the best one.

5.4 Models

In order to make a better analysis of the experiment, we selected three models: BERT, Adapter, and PGN-Adapter.

BERT We use multilingual BERT-Cased to be the baseline model.

Adapter The Adapter modules in BERT can capture language-specific information automatically. It also is a baseline to further verify the validity of the next model.

PGN-Adapter Compared with Adapter, PGN-Adapter model integrates PGN method in Adapter, and it incorporates richer language information from an external given language type. To the best of our knowledge, this model is likely to be better at

Method	Model	Holder	Target	Overall	Holder	Target	Overall	Holder	Target	Overall
		Exact F1			Proportional F1			Binary F1		
w/o Human Translation										
ModelTrans	BERT	63.64	35.49	50.51	70.76	55.39	63.56	73.89	63.32	68.95
	Adapter	66.09	37.66	52.76	70.71	56.12	63.77	72.97	62.36	67.98
	PGN-Adapter	68.61	41.92	56.13	74.05	58.31	66.66	76.02	64.10	70.43
CorpusTrans	BERT	67.27	37.14	53.17	74.02	56.42	65.74	76.47	63.08	70.17
Combine	BERT	68.45	38.25	55.01	74.35	51.03	63.93	76.32	55.45	67.02
	Adapter	**71.81**	44.43	**58.95**	75.90	60.41	68.60	77.22	66.16	72.02
	PGN-Adapter	71.66	**45.47**	58.76	**77.49**	**64.38**	**71.02**	**79.37**	**70.57**	**75.03**
w/ Human Translation										
HumanTrans	BERT	71.79	50.20	61.15	77.36	**67.14**	72.28	79.20	**73.29**	76.26
ModelTrans	BERT	72.63	47.98	60.86	78.34	65.89	72.37	80.42	72.64	76.69
	Adapter	72.99	50.43	62.07	78.55	66.72	72.83	80.48	73.25	**76.97**
	PGN-Adapter	73.57	50.04	62.27	79.05	66.67	73.08	80.79	72.27	76.69
CorpusTrans	BERT	71.58	47.49	59.89	77.40	63.10	70.45	79.57	69.41	74.63
Combine	BERT	73.22	48.48	61.08	78.36	65.33	71.95	80.33	71.67	76.08
	Adapter	71.72	48.89	61.02	77.48	64.28	71.27	79.82	69.67	75.05
	PGN-Adapter	**75.02**	**50.86**	**63.43**	**79.44**	66.81	**73.36**	**80.98**	72.16	76.74

Table 5: Experimental results of bilingual transfer from English to Chinese, where ModelTrans is model transfer, CorpusTrans is corpus translation and HumanTrans is human translation. Combine is the combination of ModelTrans and CorpusTrans.

achieving the best possible performance for cross-language tasks. The purpose of using this model is to better show the following model transfer and translation-based cross-lingual transfer methods.

Note that if the source and target languages are the same, only the first model, BERT, is used.

6 Experiment Results and Analysis

This section provides an overview of the English-Chinese transfer and multilingual transfer experiments on the basis of adding Portuguese to the Chinese target language. Further, we analyze the results in detail. In order to understand the performance of different roles (i.e., the HOLDER and TARGET), we measure the performance variance along with the span length of the arguments. As for the cross-lingual analysis, apart from the MT-based setting, we also add one more semi-automatic setting, i.e., manual translation with automatic alignment.

Just to be clear in advance, the following method CorpusTrans is using the fixed multilingual BERT embeddings only, no Adapter or PGN-Adapter method, as the Train, Dev and Test datasets are all in Chinese.

6.1 English-Chinese Transfer

Our experiments mainly focus on Chinese as the target language.

Table 5 shows the results of the English-Chinese transfer. We observe that 1) almost all experiments using the PGN-Adapter model achieve the best results in each group; 2) when manual translation and annotation are available, the model combining all the datasets performs the best: English, Chinese and automatic translation from the English corpus; 3) comparing model transfer and automatic translation, the latter outperforms the former by a large margin; and 4) if we combine the two approaches, we further improve the performance, although still inferior to the manual translation model. In short, from English to Chinese, the translation-based method is in favor of the model transfer, even with machine-translated data.

Notice that, the "auto" approaches could be easily adapted into other language pairs without large labor costs. Apart from Portuguese (Section 6.2), we will explore more low-resource target languages on the source side in the future.

6.2 Multilingual Transfer

In this section, we also conduct experiments by using multilingual transfer, thanks to the availability of the Portuguese dataset (Almeida et al., 2015a). For a fair comparison, we still focus on Chinese as the target language.

Adding the English and Portuguese, Table 6 displays the experimental results. The first two train corpora (4 lines) represent the model transfer and (automatic) translation-based methods with two languages, respectively. The observations are similar to the bilingual settings. When we combine the two methods, only a few indicators have improved. When we compare the results with additional Portuguese data in Table 5, we find that the model transfer benefits from the second source language, but in the corpus translation part, performance is noticeably decreasing, due to the low quality of the

Train Method	Model	Holder	Target	Overall	Holder	Target	Overall	Holder	Target	Overall
		Exact F1			Proportional F1			Binary F1		
ModelTrans	BERT	64.54	37.04	51.20	71.39	57.38	64.62	74.17	66.21	70.30
	Adapter	65.25	43.39	54.52	71.88	61.04	66.57	74.16	68.37	71.32
	PGN-Adapter	70.78	**44.39**	57.58	77.29	**65.20**	**71.24**	79.45	**73.15**	**76.30**
CorpusTrans	BERT	66.99	36.48	53.04	74.42	52.98	64.60	77.01	58.94	68.73
Combine	BERT	68.97	39.49	55.14	74.21	56.71	65.94	75.72	62.27	69.40
	Adapter	**72.89**	43.01	58.68	76.82	60.44	68.99	78.19	66.33	72.54
	PGN-Adapter	71.99	43.62	**59.06**	**78.46**	57.66	68.98	**80.73**	62.27	72.30

Table 6: Results of multilingual transfer from English and Portuguese to Chinese without human translation.

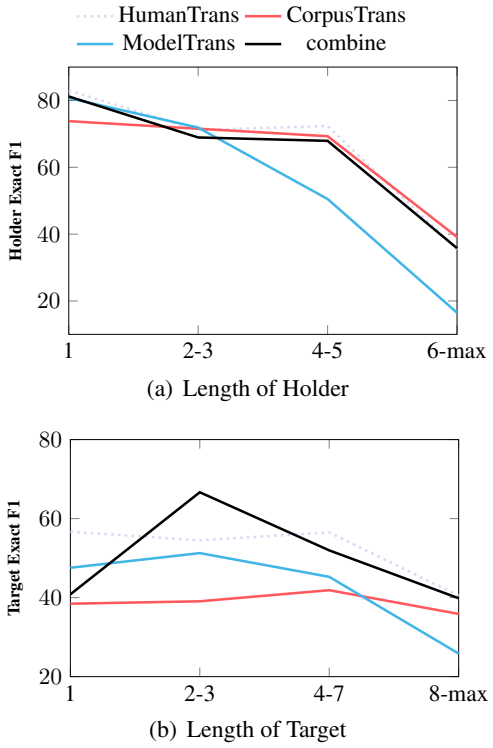
Method	Holder	Target	Overall
w/o Human Translation			
AutoAlign	71.66	45.47	58.76
w/ Human Translation			
AutoAlign	70.93	41.95	57.71
HumanAlign	73.57	50.04	62.27

Table 7: Results of the PGN-Adapter model on English-Chinese transfer with different settings, where AutoAlign is automatic alignment and HumanAlign is human alignment.

machine-translated data. The reason behind this may be that according to the MT community, the English-Chinese MT system achieves a 45+ BLEU score (Li et al., 2019), while Portuguese-Chinese MT is only around 20 (Liu et al., 2018).

6.3 Influence of the Span Length

Figure 4 illustrates the performance change along with the span length of the arguments, holder (up) and target (down), respectively. For the holder, the general tendency goes down, long spans with worse performance. However, it is worth pointing out that the $CorpusTrans$ method performs well at the long span holders, even higher than the $HumanTrans$ method which is created by human translators. We also see the $ModelTrans$ makes a worse score in the long span, but adding the $CorpusTrans$ obtains a similar performance with $HumanTrans$. That is to say, model transfer and translation-based model together are very helpful for both long and short $HOLDER$s. For the target, the best performances are all achieved for the middle-length spans, suggesting that the average length of the target is 4.8 words that is longer than 2.0 words about the holder. We speculate that short spans may not contain enough semantics, while the longer span's boundaries are not trivial to recognize correctly. As for $Combine$, we see the score in the middle even higher than the manual translation ($HumanTrans$), due to the mutual enhancement of the two methods.



(a) Length of Holder

(b) Length of Target

Figure 4: The influence of the span length of arguments. HumanTrans is human translation, CorpusTrans is corpus translation, ModelTrans is model transfer and Combine is to merge 3 methods together.

6.4 Influence of the Automatic Alignment

In addition to the machine translation setting and manual translation setting and alignment setting, we also explore the setting of manual translation with *automatic* alignment. Table 7 lists the results for comparison. We observe that the automatic alignment setting (with human translation) performs the worst among the three configurations. This might seem to be unexpected at the first glance, since the translation quality is still much better than the machine-translated ones. We speculate that, since human translation and machine translation behave quite differently, MT systems rely more on word alignment, while humans usually translate the sentence as a whole. The automatic alignment fails

to transfer the annotations from the source side to the target.

7 Conclusions

We presented the first work of Chinese ORL. First, we manually constructed a Chinese dataset with the help of corpus translation. Then, we investigated unsupervised cross-lingual transfer for Chinese ORL. We studied two different approaches, model transfer and corpus translation, respectively. Experiments and analyses were performed based on the annotated dataset. Results showed that unsupervised cross-lingual transfer is an effective method for Chinese ORL, and in addition, multi-source transfer further improves the results which are promising for future exploration of such cross-lingual transfer to other low-resource languages.

Acknowledgments

We thank all reviewers for their helpful comments. This work was supported by National Natural Science Foundation of China under grants 62076173 and 61672211.

Ethical Considerations

Our Chinese ORL dataset is sourced from the English MPQA dataset, an open-source corpus free for academic research. Under the market price according to the workload, the human efforts have been properly paid.

References

Mariana S. C. Almeida, Cláudia Pinto, Helena Figueira, Pedro Mendes, and André F. T. Martins. 2015a. Aligning opinions: Cross-lingual opinion mining with dependencies. 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 408–418, Beijing, China. Association for Computational Linguistics.

Mariana SC Almeida, Cláudia Pinto, Helena Figueira, Pedro Mendes, and André FT Martins. 2015b. Aligning opinions: Cross-lingual opinion mining with dependencies. 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 408–418.

Zhang Bo, Zhang Yue, Wang Rui, Li Zhenghua, and Zhang Min. 2020. Syntax-aware opinion role labeling with dependency graph convolutional networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Eric Breck, Yejin Choi, and Claire Cardie. 2007. Identifying expressions of opinion in context. In *IJCAI*, volume 7, pages 2683–2688.

Yejin Choi, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Sydney, Australia. Association for Computational Linguistics.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Angel Daza and Anette Frank. 2019. Translate and label! an encoder-decoder approach for cross-lingual semantic role labeling. *arXiv preprint arXiv:1908.11326*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Cross-lingual transfer for unsupervised dependency parsing without parallel data. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 113–122.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! *arXiv preprint arXiv:1807.08998*.

Hao Fei, Meishan Zhang, and Donghong Ji. 2020. Cross-lingual semantic role labeling with high-quality translated training corpus. *arXiv preprint arXiv:2004.06295*.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A distributed representation-based framework for cross-lingual transfer parsing. *Journal of Artificial Intelligence Research*, 55:995–1023.

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 *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Ozan Irsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 720–728.

Richard Johansson and Alessandro Moschitti. 2013. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3):473–509.

Arzoo Katiyar and Claire Cardie. 2016a. Investigating lstms for joint extraction of opinion entities and relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 919–929.

Arzoo Katiyar and Claire Cardie. 2016b. Investigating LSTMs for joint extraction of opinion entities and relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 919–929, Berlin, Germany. Association for Computational Linguistics.

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.

Soo-Min Kim and Eduard Hovy. 2006a. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8. Association for Computational Linguistics.

Soo-Min Kim and Eduard Hovy. 2006b. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

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

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is word segmentation necessary for deep learning of chinese representations? *arXiv preprint arXiv:1905.05526*.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443.

Siyou Liu, Longyue Wang, and Chao-Hong Liu. 2018. Chinese-Portuguese machine translation: A study on building parallel corpora from comparable texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ana Marasović and Anette Frank. 2017. Srl4orl: Improving opinion role labeling using multi-task learning with semantic role labeling. *arXiv preprint arXiv:1711.00768*.

Ana Marasović and Anette Frank. 2018. SRL4ORL: Improving opinion role labeling using multi-task learning with semantic role labeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 583–594, New Orleans, Louisiana. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the conference on empirical methods in natural language processing*, pages 62–72. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. *arXiv preprint arXiv:1808.08493*.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2018. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8119–8127.

Asa Cooper Stickland and Iain Murray. 2019. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference on Machine Learning*, pages 5986–5995. PMLR.

Swabha Swayamdipta, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016. Greedy, joint syntactic-semantic parsing with stack lstms. *arXiv preprint arXiv:1606.08954*.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Eighteenth Conference on Computational Natural Language Learning (CoNLL 2014)*.

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, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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.

Michael Wiegand, Christine Bocionek, and Josef Ruppenhofer. 2016. Opinion holder and target extraction on opinion compounds–a linguistic approach. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 800–810.

Theresa Wilson. 2008. Annotating subjective content in meetings. In *LREC 2008*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433.

Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. 2014. Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1779–1785.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. *arXiv preprint arXiv:1809.03633*.

Bishan Yang and Claire Cardie. 2012. Extracting opinion expressions with semi-markov conditional random fields. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1335–1345. Association for Computational linguistics.

Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1640–1649, Sofia, Bulgaria. Association for Computational Linguistics.

Meishan Zhang, Peili Liang, and Guohong Fu. 2019a. Enhancing opinion role labeling with semantic-aware word representations from semantic role labeling. 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 641–646.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2019b. Cross-lingual dependency parsing using code-mixed TreeBank. In *Proceedings of EMNLP-IJCNLP*, pages 997–1006.

Guangyou Zhou, Zhao Zeng, Jimmy Xiangji Huang, and Tingting He. 2016. Transfer learning for cross-lingual sentiment classification with weakly shared deep neural networks. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 245–254.

Ayah Zirikly and Masato Hagiwara. 2015. Cross-lingual transfer of named entity recognizers without parallel corpora. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 390–396.