MuPAD: A Chinese Multi-Domain Predicate-Argument Dataset

Anonymous ACL submission

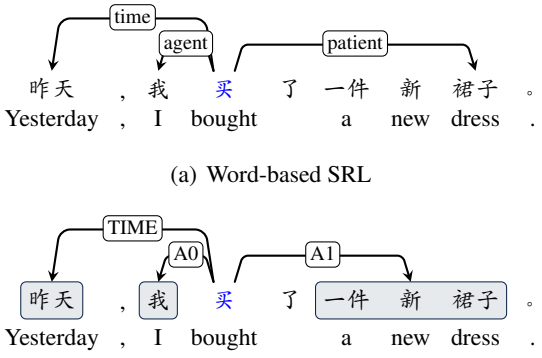
Abstract

During the past decade, neural network models have made tremendous progress on in-domain semantic role labeling (SRL). However, performance drops dramatically under the out-of-domain setting. In order to facilitate research on cross-domain SRL, this paper presents MuPAD, a multi-source predicate-argument dataset in Chinese. MuPAD consists of 30,897 sentences and 92,051 predicates which covers 6 different domains. MuPAD exhibits three important features. 1) Based on a frame-free annotation methodology, we avoid writing complex frames for new predicates. 2) We explicitly annotate omitted core arguments to recover more complete semantic structure, considering that content omission is ubiquitous in multi-domain Chinese texts. 3) We compile 53-page annotation guidelines and adopt strict double annotation for improving data quality. This paper describes in detail the annotation methodology and process of MuPAD, and presents in-depth data analysis. We present benchmark cross-domain results on MuPAD, using Chinese Proposition Bank as a heterogeneous dataset.

1 Introduction

As a fundamental NLP task, semantic role labeling (SRL), also known as shallow semantic parsing, aims to capture the major semantic information of a sentence based on predicate-argument structure. Basically, SRL tries to answer "who did what to whom where and when". Previous works have shown that SRL can help various downstream tasks including information extraction (Bastianelli et al., 2013), machine translation (Shi et al., 2016), reading comprehension (Zhang et al., 2020), plagiarism detection (Paul and Jamal, 2015), etc.

Figure 1 gives two examples of SRL results. According to the spanning range of semantic roles, there exist two typical representation forms, i.e., word-based and span-based SRL. This work adopts the word-based formulation, in which a semantic role corresponds to a single word. In contrast, span-based SRL, adopted by most previous datasets, takes a multi-word span as a semantic role. The direction of arcs is from predicates to arguments, and the labels indicate the types of semantic roles. For example, the arc from "买(bought)" to "裙子(dress)" with a label "patient" means that the semantic role between the predicate "买(bought)" and the argument "裙子(dress)" is "patient".



(a) Word-based SRL

(b) Span-based SRL

Figure 1: An example of two formulations, where the word-based SRL is adopted by CSB and MuPAD, and span-based is adopted by CPB and CNB.

Recently, neural-based Chinese SRL has achieved great progress thanks to the rise of deep learning methods (Marcheggiani et al., 2017; He et al., 2018; Cai et al., 2018), especially of powerful pre-trained language models (PLMs). However, despite the significant improvements, existing studies on Chinese SRL mainly focus on the in-domain setting, where both training and test data are from the same domain (Wang et al., 2015; Guo et al., 2016b; Xia et al., 2017). SRL performance drops dramatically when the domain of test data is different from that of the training data, known as the domain adaptation problem.

Meanwhile, with the rapid growth of user-generated web data, cross-domain SRL has be-

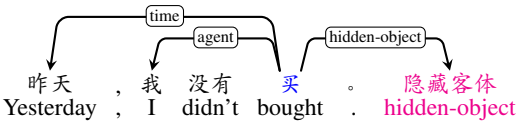


Figure 2: An example sentence with omission in MuPAD.

come an important and challenging task in realistic NLP systems (Jiang and Zhai, 2007; Ramponi and Plank, 2020). However, due to the scarcity of multi-domain labeled data, most current researches on SRL still focus on in-domain scenario, and make very limited progress in the domain adaptation task.

There are three existing publicly available Chinese SRL datasets, including Chinese Proposition Bank (CPB) (Xue and Palmer, 2005), Chinese NomBank (CNB) (Xue, 2006a), and Chinese SemBank (CSB) (Xia et al., 2017). All the above datasets are in-domain data which mainly contain canonical news texts and magazine/textbook articles, inapplicable to cross-domain research.

In order to facilitate future study on cross-domain SRL, this paper presents MuPAD, a multi-source predicate-argument dataset in Chinese, consisting of 30,897 sentences and 92,051 predicates, from 6 different domains. Overall, MuPAD has the following important features.

(1) Following CSB instead of CPB and CNB, we adopt a frame-free annotation methodology, considering that it requires a very high level of linguistic background for defining frames for new predicates, and a lot of new predicates may appear in non-canonical texts.

(2) We explicitly annotate omitted core arguments with two special labels, i.e., "hidden-subject" and "hidden-object", considering that it is ubiquitous that people try to avoid repetition by omitting previous content in context, especially in non-canonical Chinese texts. Figure 2 gives an example.

(3) We adopt strict double annotation for each sentence in order to improve quality. If two annotators submit inconsistent results, a senior annotator determines the final answer. We also compile 53-page annotation guidelines to be studied and referenced by the annotators.

Finally, we employ multi-task learning (MTL) framework over the popular and competitive model (Cai et al., 2018) for SRL and conduct preliminary cross-domain SRL experiments on our dataset. Moreover, we use CPB2.0 as a heterogeneous dataset and BERT to further boost the performance of cross-domain SRL.

2 Related Work

SRL Corpora. Large-scale annotated data is a prerequisite to developing high-performance SRL systems (Fürstenau and Lapata, 2009; Xia et al., 2020). The most representative ones in English are FrameNet (Baker et al., 1998), PropBank (Kingsbury and Palmer, 2002), and NomBank (Meyers et al., 2004). FrameNet is a large-scale manually annotated semantic lexicon resource, which describes the semantic frames underlying the meanings of words. PropBank and NomBank are built by adding predicate-argument structures to the constituents of syntactic parser trees in the Penn Treebank (Marcus et al., 1993). Following FrameNet, PropBank and NomBank also use frames to guide argument and semantic role annotations. However, the development of frames is both time-consuming and labor-intensive, which requires annotators to be equipped with a strong linguistic background. Moreover, the texts of the two datasets are mainly from the news domain (Wall Street Journal). There is one widely used out-of-domain data (Brown) of PropBank (Carreras and Màrquez, 2005) that only has 426 sentences, confining the progress of English cross-domain SRL.

For Chinese, CPB (Xue and Palmer, 2005), CNB (Xue, 2006a), and CSB (Xia et al., 2017) are the most commonly used SRL corpora. CPB and CNB annotate predicate-argument relations to the sentences in the Chinese Treebank (Marcus et al., 1993), which also need pre-defined frames. Differently, CSB uses non-predicate-specific roles, which are defined for predicates with similar syntactic and semantic regularities. The texts of CPB and CNB are mainly from Xinhua newswire (mainland China), Hong Kong news, and Sinorama Magazine (Taiwan) (Hajic et al., 2009), and CSB texts are from online articles and news. These corpora mainly belong to the same news domain, which can hardly be used in Chinese SRL domain adaptation. Therefore, we annotate a large-scale Chinese multi-domain SRL dataset, aiming to promote the development of Chinese SRL domain adaptation.

Domain adaptation. Domain adaptation has been an important and challenging research topic in NLP (Daumé III, 2007; Ganin and Lempitsky, 2015; Guo et al., 2016a; Kim et al., 2016, 2017; Clark et al., 2018; Zhao et al., 2018). Kim

et al. (2016) propose a neural shared-private model for the cross-domain slot sequence tagging task, which utilizes separate BiLSTM encoders to obtain domain-invariant and domain-specific representations, achieving significant improvements on all domains. Jia et al. (2019) improve the domain adaptation power of multi-task learning by utilizing domain and task embeddings to generate domain-related and task-related model parameters, which has been proven effective for unsupervised cross-domain NER. Li et al. (2019b) present a tri-training model for cross-domain transfer learning, in which they generate additional high-quality auto-labeled training examples from unlabeled data with iterative training. To facilitate cross-domain Chinese dependency parsing research, Li et al. (2019a) present one large-scale cross-domain dataset. Based on our newly annotated multi-domain Chinese SRL data, we conduct preliminary experiments, providing a strong baseline for future work.

3 Data Annotation

In this section, we describe the details of the data annotation procedure. We compile a detailed and systematic annotation guideline for annotators' reference. Compared with previous guidelines that almost adopt span-based argument form, our guideline proposes to use word-based argument form, which is free from distinguishing the argument boundary. Moreover, our guideline does not require pre-defined frames for predicates, which are difficult and time-consuming to build and update. During the annotation process, we adopt strict double annotation for all the sentences to guarantee quality. Finally, we construct the first Chinese high-quality, large-scale, multi-domain open predicate-argument dataset (MuPAD) covering six domains, i.e., news, product blog, product comment, web fiction, legal, and medical domains.

Annotation guideline. After several months of investigation on previous SRL guidelines, we compile a frame-free guideline that adopts word-based argument and explicitly annotates omitted core arguments. We design two coarse-grained categories according to the importance of the semantic role and further divide them into 24 fine-grained labels to capture the semantic relationships between predicates and arguments, as shown in Table 1. It is worth mentioning that in order to understand the sentence more completely, we for the first time define "hidden-subject" and "hidden-object" labels to explicitly annotate the omitted core arguments of a predicate in a sentence. Finally, we compile a 53-page guideline with detailed illustrations of each label and corresponding concrete examples. The current version of our guideline is attached in our supplementary material. We also gradually improve our guideline according to the feedback during the annotation.

Data selection. For source domain data, we choose all the 10.3K sentences with 16.5K predicates from Chinese SemBank (Xia et al., 2017) and randomly select 6.7K sentences with 24.5 predicates from CoNLL-2009 Chinese dataset (Hajic et al., 2009), all of which are canonical newswire resources.

For target domain data, we select 51K predicates (14K sentences) in total from five specific domains, including product blog (PB), product comment (PC), web fiction (ZX), legal (LAW), and medical (MED) domains. Both PB and PC are non-canonical data from Taobao[1], where PB is from Taobao headline website, and PC is comments on products written by users. ZX is selected from a popular Chinese fantasy novel called "Zhuxian" (ZX, known as "Jade Dynasty"). LAW is extracted from the China artificial intelligence law challenge 2018[2]. MED is crawled from the medical sector of People's Daily Online and Sina.com.

We give predicates to the annotators during the annotation process. For source domain data, we directly choose their predicates in Chinese SemBank and CoNLL-2009 Chinese dataset. For target domains, which have already been annotated with syntactic information (Li et al., 2019a), we choose their predicates according to their syntactic information and a dictionary extracted from Chinese frames[3].

Quality Control. We employ 23 undergraduate students as part-time annotators and select five annotators who have a lot of experience in data annotation as expert annotators to handle annotation inconsistency issues. Before the formal annotation, each annotator is trained for several hours to be familiar with the guideline and our annotation tool. During the annotation process, we adopt a strict double annotation workflow to guarantee the quality of the annotated data. Specifically, each task is

[1] http://www.taobao.com
[2] http://cail.cipsc.org.cn:2018/
[3] https://verbs.colorado.edu/chinese/cpb/html_frames

	Label	Example	Argument
Core roles	agent	我[打]他(I [hit] him)	我(I)
	co-agent	我和他[讨论] (I [discuss] with him)	我(I)、他(him)
	expe (experiencer)	天气真[好] (The weather [is] really good)	天气(weather)
	hidden-subject	[吃]饭了吗？([Ate]？)	隐藏主体(hidden-subject)
	patient	他被[打]了(He was [hit])	他(he)
	pred-patient (predicate-patient)	他[喜欢]打篮球(He [likes] playing basketball)	打(play)
	dative	[给]他书([Give] him a book)	他(him)
	relative	这[是]他的书(This [is] his book)	书(book)
	hidden-object	你[吃]了吗？(Have you [eaten]?)	隐藏客体（hidden-object)
	subj-obj (subject-object)	温度计[伸入]水中(The thermometer is dipped into the water)	温度计(thermometer)
Non-core roles	tool (instument)	用钢笔[写]字([Write] with pen)	钢笔(pen)
	material	用颜料[刷]墙([Brush] the wall with pigment)	颜料(pigment)
	manner	按计划[执行] ([Perform] according to plan)	计划(plan)
	loc (location)	在学校[学习] ([Study] at school)	学校(school)
	beg-loc (begin location)	从学校[出发] ([Start] from school)	学校(school)
	end-loc (end location)	[流]入大海([Flow] to the ocean)	大海(ocean)
	dir (direction)	向西[流] ([Flow] to the west)	西(west)
	time	星期天去[打篮球] ([Play] basketball on Sunday)	星期天(Sunday)
	beg-tm (begin time)	比赛七点开始[进行] (The game [starts] at seven o 'clock)	七点(seven o'clock)
	end-tm (end time)	会议[开]到三点(The meeting [runs] until three o'clock)	三点(three o'clock)
	range	在数学上[有]天赋([Have] an aptitude for mathematics)	数学(mathematics)
	cause	我因为爱你才[撒谎] (I [lied] because I love you)	爱(love)
	quantity	我[跑]了一圈(I [ran] a lap)	一圈(a lap)
	separated	我们[见]过面(We have [met])	面(met)

Table 1: The table of semantic role labels in our guideline, where "[]" denotes the predicate.

randomly assigned to two annotators to annotate independently. If the submissions from two annotators are the same, the consistent answer is taken as the final answer. Otherwise, the task is assigned to a third expert annotator to decide the final answer by comparing and analyzing the inconsistent submissions. All the annotators are paid for their work. The average salary is 28 RMB per hour.

Annotation tool. We build a browser-based annotation tool to support and manage the double annotation workflow of the Chinese SRL annotation task. Given a task, the annotation tool highlights the predicate with a red circle, and the annotators are asked to annotate all the arguments of the marked predicate. If the marked predicate is not common verbs, nominal verbs, or adjectives, as defined in our guideline, the annotators are asked to click a checkbox to inform us.

4 Analysis on MuPAD

In this section, we analyze the MuPAD dataset from different perspectives to gain more insights.

Annotation consistency. As aforementioned, each task is assigned to two annotators. If the two submissions are inconsistent, a third expert annotator is asked to handle the inconsistency and decide the final results. The first two rows in Table 2 show the predicate- and argument-wise (pred- and arg-wise) annotation consistency ratios in all domains.

The pred-wise consistency ratio is defined as $\frac{\#\texttt{Pred}_{\texttt{annoA}\cap\texttt{annoB}}}{\#\texttt{Pred}_{\texttt{annoA}\cup\texttt{annoB}}}$, where the denominator is the total number of predicates submitted by all annotator pairs, and the numerator is the number of predicates with consistent arguments from annotator pairs. We can see that the pred-wise annotation consistency ratios in most domains are lower than 60%. Even the highest pred-wise consistency ratio, which is achieved in the PC domain, is only 71.23%. This means that more than a quarter of the annotation tasks need to be further checked by a third expert annotator, demonstrating the difficulty of the SRL annotation task and the importance of performing strict double annotation to guarantee data quality.

In addition, it is worth noting that the pred-wise consistency ratio in PC domain is much higher than that in the other five domains. We believe this is related to the average number of arguments per predicate. For further investigation, we calculate the average number of arguments and find the number of average arguments per predicate is the lowest in PC domain. Therefore, it is relatively easier for the annotators to recognize the arguments in PC domain.

The arg-wise consistency ratio is defined as $\frac{\#\texttt{Arg}_{\texttt{annoA}\cap\texttt{annoB}}}{\#\texttt{Arg}_{\texttt{annoA}\cup\texttt{annoB}}}$, where the denominator is the total number of arguments submitted by all annotator pairs, and the numerator is the number of arguments that receive the same arcs and labels from two annotators. As shown in Table 2, the arg-wise consistency ratios in most of the domains are lower than 75%, excepting that PC achieves the highest arg-wise consistency ratio of 83.63%.

		Source	PB	PC	ZX	LAW	MED	AVG
Consistency ratio	pred-wise	48.86	57.58	**71.23**	48.86	47.18	50.57	54.05
	arg-wise	74.48	74.67	**83.63**	74.65	71.49	73.24	75.36
Accuracy	overall	85.86	82.08	**89.40**	83.80	84.57	85.78	85.25
	agent	**93.50**	82.19	88.16	91.45	85.54	86.09	87.82
	time	**91.17**	88.98	85.02	83.69	85.65	86.87	86.90
	hidden-subject	88.79	90.61	**94.45**	79.65	85.54	90.45	88.25
	patient	87.02	88.36	**90.20**	86.32	85.66	89.26	87.80
	loc	**85.96**	79.12	83.71	84.32	84.25	79.30	82.78
	pred-patient	**85.12**	84.28	84.08	83.97	84.11	84.58	84.36
	expe	81.58	85.27	**90.89**	86.39	84.20	87.67	86.00

Table 2: Analysis on consistency ratio and accuracy. "AVG" represents the average accuracy of a label in six domains. Boldface indicates the maximum value of each row, underline represents the minimum value of each row.

Annotation accuracy. In the third row of Table 2, we present the argument-wise (arg-wise) annotation accuracy. The annotation accuracies in all domains are more than 80%, indicating that our guideline is reasonable, which ensures the quality of annotation data.

To gain more insights into the accuracy regarding different labels, we select the labels with high proportions of 5 core and 2 non-core roles according to the proportion of labels in source domain data and calculate their accuracies in Table 2. As we can see, "hidden-subject" achieves the highest average accuracy, demonstrating the omitted subject is easy to recognize. The lowest average accuracy is 82.78% on "loc", probably because it is a non-core label with the lowest proportion of all the labels and is prone to be missed by the annotators.

Label distribution. Figure 2 illustrates the label distribution in all the domains. We choose 2 core labels and 2 non-core labels with high proportions in source domain data. Besides, we also choose "loc" and "hidden-object" to analyze due to their characteristic distributions in different domains. From Figure 2, we find that these target domains have different characteristics. PC and PB have more omitted core arguments (i.e., "hidden-subject" and "hidden-object"), which shows the texts in these two domains are more non-canonical.

In addition, LAW has more additional semantic roles than other domains, such as "time", "loc", etc. This is due to the fact that the elements (e.g., time and location) of legal cases in LAW should be described in detail.

Besides, there are more "agent" roles and fewer omitted arguments in ZX, showing that ZX is more canonical. Furthermore, on the whole, the label distribution of ZX and the source domain is the most similar, indicating that the knowledge of the source domain can be easier to transfer to ZX.

Looking into the distribution of hidden-subject and hidden-object labels in all the domains, we find that hidden labels exist in all the domains, especially in non-canonical texts like PC and PB, demonstrating the necessity of annotating hidden labels. In addition, "hidden-subject" takes a higher proportion than "hidden-object" in all domains, reflecting that the subject role of the predicate in Chinese sentences is often omitted.

Difficulty in annotation. To further analyze the difficulty in annotation, we calculate the proportion of the arguments with the same arcs but different labels from two annotators among all the arguments with the same arcs. We find that the confusion pattern "agent, expe" accounts for the largest proportion of 22.23%, which means the label "agent" is prone to be confused with "expe". The reason for the confusion may be the part-of-speech (POS) of predicates. The part-of-speech (POS) of the predicates can affect the argument labels. However, the POS for some predicates is subtle and vague in Chinese, confusing the argument labels. Taking the sentence "纽扣(Buttons) 一天(a day) 坏(getting broken) 一个(one)" as an example, "坏" may be misunderstood as an adjective and thus the argument "纽扣" is incorrectly annotated as "expe". Actually, "坏" acts as a verb in this sentence and the correct label of "纽扣" is "agent". The second confusion pattern is "patient, pred-patient", with a proportion of 12.6%, due to the misunderstanding of the POS of the augment. It is also difficult for annotators to distinguish "agent" and "patient". For example, in the sentence "新(new)

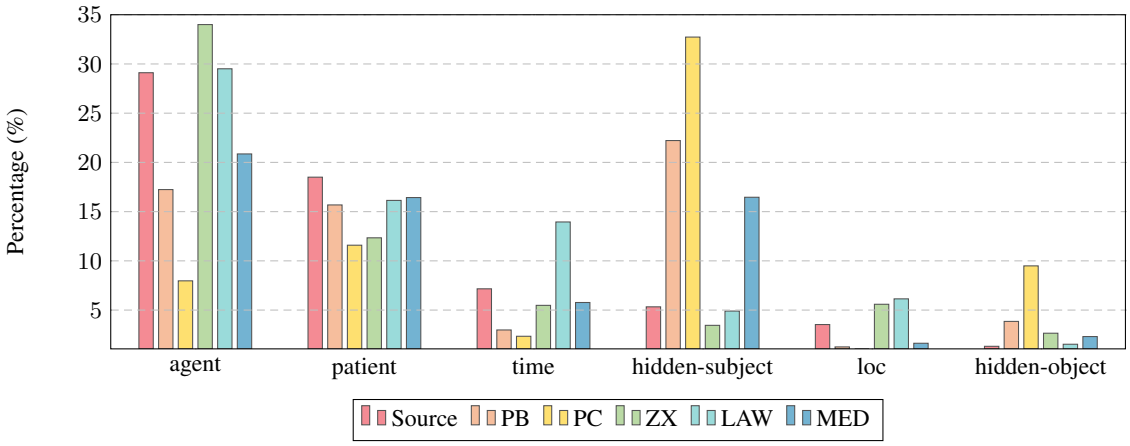


Figure 3: Label distribution in different domains.

衣服(cloths) 被(was) 弄脏了(soiled)", the preposition "被" is omitted. As a result, the label of "新(new) 衣服(cloths)", which is "patient", may be confused with "agent" due to the omission.

5 Approach

Based on our newly annotated multi-domain Chinese SRL data MuPAD, we conduct preliminary experiments, aiming to provide strong benchmark results for future work. Specifically, we present a MTL approach to improve the performance of multi-domain SRL by learning from multiple heterogeneous datasets simultaneously. We also enhance the model with the contextualized word representation BERT for further improvement.

Following previous works (Marcheggiani et al., 2017; He et al., 2018; Cai et al., 2018), we treat SRL parsing as a sequence labeling problem and aim to find the tag sequence $\hat{\boldsymbol{y}}$ with the highest score:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{w})}{\operatorname{argmax}}\, score(\boldsymbol{w}, \boldsymbol{y}) \quad (1)$$

where $y_i \in \mathcal{L}$ is the label of the i-th word w_i. $\mathcal{L}$ is the semantic role label set, which contains 24 semantic role labels and an extra "None" label to indicate there is no semantic relationship between the concerned word and the predicate. $\mathcal{Y}(\boldsymbol{w})$ represents all the possible label sequence.

5.1 Basic Biaffine Parser

In this work, we adopt the widely used biaffine parser (Dozat and Manning, 2017) as our baseline, and build the MTL framework based on it. Figure 4 shows the model architecture of the biaffine parser.

First, the biaffine parser takes the concatenation of the pre-trained word embedding $\mathbf{e}_i^{pre}$, the randomly initialized word embedding $\mathbf{e}_i^{r}$, the word representation $\mathbf{r}_i^{c}$ and the predicate embedding $\mathbf{e}_i^{p}$ as the input, i.e., $\mathbf{x}_i = \mathbf{e}_i^{pre} \oplus \mathbf{e}_i^{r} \oplus \mathbf{r}_i^{c} \oplus \mathbf{e}_i^{p}$, where $\mathbf{r}_i^{c}$ represents the CNN output of the word constituent characters. Then, the input representation $\mathbf{x}_i$ is fed into a three-layer BiLSTM to obtain its context-aware representation $\mathbf{h}_i$. After that, two separate MLPs are applied over $\mathbf{h}_i$ to get two lower-dimensional representation $\mathbf{h}_i^{pred}$ (as predicate) and $\mathbf{h}_i^{arg}$ (as candidate arguments). Finally, the score of the label between the predicate and the argument is computed via a biaffine operation. During training, we adopt the local cross-entropy loss.

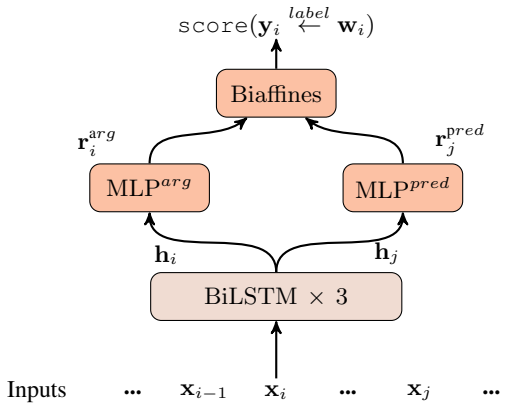


Figure 4: The basic architecture of Biaffine Parser.

To obtain multi-domain results on the basic biaffine parser, we train the parser on source domain data and make predictions on target domain data using the source parser.

5.2 Combining Multi-Domain Data with MTL

The goal of MTL is to improve model performance by learning the underlying common knowledge from multiple related tasks (Collobert and We-

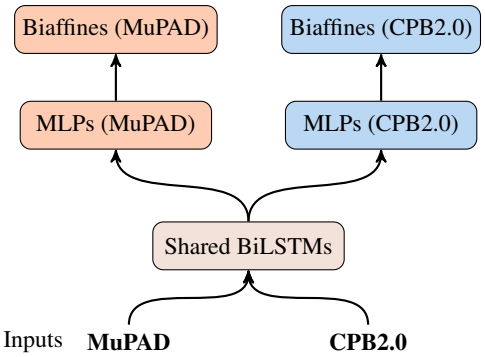


Figure 5: The framework of MTL.

ston, 2008; Guo et al., 2016a; Li et al., 2019a). MTL framework can be naturally used for multi-domain learning by utilizing multiple heterogeneous datasets and considering each dataset as a single task to improve the performance in target-domain dataset.

As shown in Figure 5, we extend the original biaffine parser to the MTL framework to accommodate multi-domain SRL. In this work, we treat our MuPAD as the target data, and treat CPB2.0 as the auxiliary data to provide additional semantic information to gain improvements in MuPAD. Specifically, the SRL parsing on MuPAD data and CPB2.0 data are considered as two separate tasks. They share the same word/predicate embeddings and BiLSTM parameters. Over the shared BiLSTMs, two separate MLPs and biaffines are employed for MuPAD and CPB2.0 SRL parsing respectively.

5.3 Enhancing with BERT

Recently proposed pre-trained language model BERT (Devlin et al., 2019) has shown its great power in learning and capturing contextualized representations and has proved to be beneficial in a variety of NLP tasks, such as information retrieval (Yang et al., 2019b), question answering (Yang et al., 2019a), and word segmentation (Huang et al., 2020). In this work, we use BERT to extract the contextualized representations for words and treat them as additional features to augment the input representation, i.e., $\mathbf{x}_i = \mathbf{e}_i^{pre} \oplus \mathbf{e}_i^{r} \oplus \mathbf{r}_i^{c} \oplus \mathbf{e}_i^{p} \oplus \mathbf{e}_i^{BERT}$.

6 Experiments

Data. For source domain data, we re-annotate 6,682 sentences randomly selected from CoNLL-2009 (Hajic et al., 2009) and all the sentences from Chinese SemBank (Xia et al., 2017) following our newly compiled guideline. For target domain data, we use our annotated MuPAD which contains 5 domains, including PB, PC, ZX, LAW, and MED. The data statistics for source and target domain data are shown in Table 3. For the auxiliary data used to provide additional semantic information to improve multi-domain SRL parsing, we randomly select 13,170 sentences with 72,616 predicates from CPB2.0 (Xue, 2006b), which is newswire data.

Evaluation metric. We adopt the standard precision ($\frac{\#\texttt{Arg}_{\texttt{correct}}}{\#\texttt{Arg}_{\texttt{pred}}}$), recall ($\frac{\#\texttt{Arg}_{\texttt{correct}}}{\#\texttt{Arg}_{\texttt{gold}}}$) and F1 score ($\frac{2PR}{P+R}$) for SRL evaluation.

Settings. We implement the SRL model of Cai et al. (2018) with PyTorch and follow their hyperparameter settings, such as the dimensions of embeddings, learning rate, and dropout ratios. We use bert-base-chinese[4] to obtain contextualized representations for words, and the dimension of the BERT representation is 768. During training, early stopping is triggered if the peak performance in dev data does not increase in 50 consecutive iterations.

6.1 Multi-Domain Results of Baseline

The first row of Table 4 presents the results in the source/target domain dev/test data using the base biaffine parser trained on the source data.

First, it is obvious that the performance in all the five target domains drops dramatically compared with the results on source data, with the gap of more than 18% in F1. This indicates that the model trained on source data has a challenge in making reliable predictions on target domain data due to the distributional mismatch between the different domains. Second, we find that the basic biaffine parser performs better on ZX and LAW compared with the other three target domain data, i.e., PB, PC, and MED. The probable reason is that ZX and LAW, which are novel and legal case, respectively, are more canonical in text. Therefore, ZX and LAW are more similar to the newswire source training data than the other three target domain data, leading to better performance in ZX and LAW dev/test. Third, PB has the lowest F1 score in both dev and test. This can be explained by the fact that PB is non-canonical data from Taobao headline website. The dissimilarity between the source training data and PB target data causes the low performance.

[4] https://huggingface.co/bert-base-chinese

(#Pred / #Sent)	Source	PB	PC	ZX	LAW	MED	CPB2.0
Train	32,790 / 13,022	–	–	–	–	–	72,616 / 13,170
Dev	4,098 / 1,875	3,796 / 1,255	5,658 / 1,295	1,784 / 492	3,718 / 778	1,874 / 478	–
Test	4,101 / 2,077	7,521 / 2,498	11,416 / 2,595	4,107 / 1,083	7,438 / 2,035	3,750 / 1,414	–

Table 3: Statistics of MuPAD and CPB2.0. '#Pred' and '#Sent' represent the number of predicates and sentences.

	Source		PB		PC		ZX		LAW		MED		AVG
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	
Baseline	69.55	68.40	44.24	44.59	46.12	47.37	50.98	49.26	47.05	49.77	44.88	46.70	50.74
MTL	**74.55**	**73.43**	**46.95**	**47.36**	**48.61**	**48.95**	**54.38**	**53.54**	**48.55**	**51.25**	**54.53**	**55.26**	**54.78**
Baseline+BERT	80.25	79.46	64.28	**66.23**	66.77	67.14	71.65	**71.77**	70.47	73.28	66.06	67.97	70.44
MTL+BERT	**81.05**	**80.85**	**64.35**	65.27	**67.75**	**68.13**	**72.44**	71.72	**70.53**	**73.79**	**66.38**	**69.07**	**70.94**

Table 4: Models' performance on the data of all domains.

6.2 Multi-Domain Results of MTL

As shown in the second row of Table 4, benefiting from the additional semantic information provided by the auxiliary CPB2.0 data using the MTL model, the SRL performance in all the domains are improved compared with the baseline model. This indicates that the MTL model is effective in capturing and learning the underlying common knowledge from multiple heterogeneous data to gain improvements for multi-domain SRL.

By comparing the improvements brought by MTL model in all the domains, we find that MED data and source data obtain the largest gains by 9.65%/8.56% and 5%/5.03% in dev/test, respectively. The main reason is that both MED and source data are canonical newswire data, similar to the auxiliary newswire CPB2.0 data. The improvement in LAW is the smallest. This can be explained by the difference in the label distribution in MED and CPB2.0. For example, as mentioned in Section 4, the label "time" and "loc" account for larger proportion (13.95% and 6.14% respectively) in LAW than in other 5 domains. However, in CPB2.0, the proportions of "time" and "loc" are only 6.27% and 3.50% (about half of that in LAW), and thus cannot provide much more additional information to increase the performance of these labels.

6.3 Combining BERT

The second major row of Table 4 shows the results of the baseline and MTL model combined with BERT representations. We can see that the results of both "Baseline+BERT" and "MTL+BERT" consistently increase by large margins compared with the corresponding models without BERT (as shown in the first major row of table 4), demonstrating the great power of BERT in contextualized representation. When comparing "MTL+BERT" with "Baseline+BERT", we find that the MTL model with BERT gives better or comparable results in 5 of the 6 domains, with an average increase of 0.5%, showing that the MTL model is effective in utilizing multiple heterogeneous data and can complement the information obtained from BERT representations.

7 Conclusion

This paper presents a Chinese multi-domain predicate-argument dataset, named as MuPAD, which consists of 30,897 sentences with 92,051 predicates and covers 6 different domains. Particularly, we adopt a frame-free annotation methodology, which does not require high-level linguistic background for defining frames for large amounts of new predicates in multi-domain data. Besides, considering the ubiquitousness of omission phenomena in Chinese, we explicitly annotate omitted core arguments with two special designed labels "hidden-subject" and "hidden-object" for better semantic understanding. To ensure annotation quality, we adopt strict double annotation and ask a third expert to handle annotation inconsistency. We also perform analysis on MuPAD from different perspectives. Finally, we conduct preliminary experiments with the MTL framework on our newly annotated MuPAD, hoping to provide benchmark results for future exploration on Chinese cross-domain SRL.

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of COLING-ACL*, pages 86–90.

Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2013. Textual inference and meaning representation in human robot interaction. In *Proceedings of JSSP*, pages 65–69.

Jiaxun Cai, Shexia He, Zuchao Li, and Hai Zhao. 2018. A full end-to-end semantic role labeler, syntactic-agnostic over syntactic-aware? In *Proceedings of COLING*, pages 2753–2765.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of CoNLL*, pages 152–164.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc V. Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of EMNLP*, pages 1914–1925.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of ICML*, pages 160–167.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*, pages 256–263.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of ICLR*.

Hagen Fürstenau and Mirella Lapata. 2009. Semi-supervised semantic role labeling. In *Proceedings of EACL*, pages 220–228.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of ICML*, pages 1180–1189.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2016a. A universal framework for inductive transfer parsing across multi-typed treebanks. In *Proceedings of COLING*, pages 12–22.

Jiang Guo, Wanxiang Che, Haifeng Wang, Ting Liu, and Jun Xu. 2016b. A unified architecture for semantic role labeling and relation classification. In *Proceedings of COLING*, pages 1264–1274.

Jan Hajic, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, M Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL*, pages 1–18.

Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai. 2018. Syntax for semantic role labeling, to be, or not to be. In *Proceedings of ACL*, pages 2061–2071.

Weipeng Huang, Xingyi Cheng, Kunlong Chen, Taifeng Wang, and Wei Chu. 2020. Towards fast and accurate neural Chinese word segmentation with multi-criteria learning. In *Proceedings of COLING*, pages 2062–2072.

Chen Jia, Liang Xiao, and Yue Zhang. 2019. Cross-domain NER using cross-domain language modeling. In *Proceedings of ACL*, pages 2464–2474.

Jing Jiang and Chengxiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of ACL*, pages 264–271.

Young-Bum Kim, Karl Stratos, and Dongchan Kim. 2017. Adversarial adaptation of synthetic or stale data. In *Proceedings of ACL*, pages 1297–1307.

Young-Bum Kim, Karl Stratos, and Ruhi Sarikaya. 2016. Frustratingly easy neural domain adaptation. In *Proceedings of COLING*, pages 387–396.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank.

Zhenghua Li, Xue Peng, Min Zhang, Rui Wang, and Luo Si. 2019a. Semi-supervised domain adaptation for dependency parsing. In *Proceedings of ACL*, pages 2386–2395.

Zuchao Li, Junru Zhou, Hai Zhao, and Rui Wang. 2019b. Cross-domain transfer learning for dependency parsing. In *Proceedings of NLPCC*, pages 835–844.

Diego Marcheggiani, Anton Frolov, and Ivan Titov. 2017. A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling. In *Proceedings of CoNLL*, pages 411–420.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguistics*, 19:313–330.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. Annotating noun argument structure for nombank. In *Proceedings of LREC*, volume 4, pages 803–806.

Merin Paul and Sangeetha Jamal. 2015. An improved srl based plagiarism detection technique using sentence ranking. *Procedia Computer Science*, 46:223–230.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp-a survey. In *Proceedings of COLING*, pages 6838–6855.

Chen Shi, Shujie Liu, Shuo Ren, Shi Feng, Mu Li, Ming Zhou, Xu Sun, and Houfeng Wang. 2016. Knowledge-based semantic embedding for machine translation. In *Proceedings of ACL*, pages 2245–2254.

Zhen Wang, Tingsong Jiang, Baobao Chang, and Zhifang Sui. 2015. Chinese semantic role labeling with bidirectional recurrent neural networks. In *Proceedings of EMNLP*, pages 1626–1631.

Qiaolin Xia, Lei Sha, Baobao Chang, and Zhifang Sui. 2017. A progressive learning approach to Chinese srl using heterogeneous data. In *Proceedings of ACL*, pages 2069–2077.

Qingrong Xia, Rui Wang, Zhenghua Li, Yue Zhang, and Min Zhang. 2020. Semantic role labeling with heterogeneous syntactic knowledge. In *Proceedings of COLING*, pages 2979–2990.

Nianwen Xue. 2006a. Annotating the predicate-argument structure of Chinese nominalizations. In *Proceedings of LREC*, pages 1382–1387.

Nianwen Xue. 2006b. Semantic role labeling of nominalized predicates in Chinese. In *Proceedings of NAACL-HLT*, pages 431–438.

Nianwen Xue and Martha Palmer. 2005. Automatic semantic role labeling for chinese verbs. In *Proceedings of IJCAI*, volume 5, pages 1160–1165.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019a. End-to-end open-domain question answering with bertserini. In *Proceedings of NAACL-HLT*, pages 72–77.

Wei Yang, Haotian Zhang, and Jimmy Lin. 2019b. Simple applications of bert for ad hoc document retrieval. volume abs/1903.10972.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware BERT for language understanding. In *Proceedings of AAAI*, pages 9628–9635.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. 2018. Adversarial multiple source domain adaptation. In *Proceedings of ICLR*, pages 8559–8570.