MuCPAD: A Multi-Domain Chinese Predicate-Argument Dataset

Yahui Liu, Haoping Yang, Chen Gong,* Qingrong Xia, Zhenghua Li, Min Zhang
Institute of Artificial Intelligence, School of Computer Science and Technology,
Soochow University, China
{yahuiliu.nlp,hpyang3,gongchen.nlp}@foxmail.com,
kirosummer.nlp@gmail.com,
{zhli13,minzhang}@suda.edu.cn

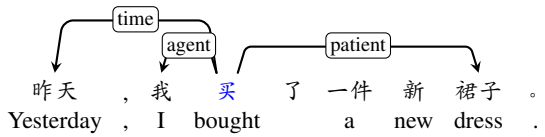
Abstract

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 MuCPAD, a multi-domain Chinese predicate-argument dataset, which consists of 30,897 sentences and 92,051 predicates from six different domains. MuCPAD 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 omission of content words is ubiquitous in multi-domain Chinese texts. 3) We compile 53 pages of annotation guidelines and adopt strict double annotation for improving data quality. This paper describes in detail the annotation methodology and annotation process of MuCPAD, and presents in-depth data analysis. We also give benchmark results on cross-domain SRL based on MuCPAD.

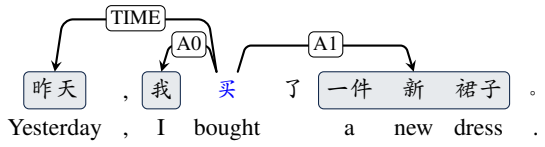
1 Introduction

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" (Màrquez et al., 2008). Previous works have shown that SRL can help various downstream tasks, including information extraction (Bastianelli et al., 2013), plagiarism detection (Paul and Jamal, 2015), machine translation (Shi et al., 2016), reading comprehension (Zhang et al., 2020), etc.

Figure 1 gives two examples of SRL structure. According to the definition of semantic roles, there exist two typical representation forms, i.e., the word-based and the span-based. This work adopts the word-based form, in which an argument corresponds to a single word. In contrast, span-based SRL, adopted by most previous datasets, takes a word span as an argument. 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 representation adopted in MuCPAD



(b) Span-based SRL representation adopted in CPB and CNB

Figure 1: Examples of two SRL formulations.

Recently, Chinese SRL research has achieved tremendous 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) (Shi and Lin, 2019; Conia and Navigli, 2020; Paolini et al., 2021). However, existing studies on Chinese SRL mainly focus on the in-domain setting, where 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-

*Corresponding author.

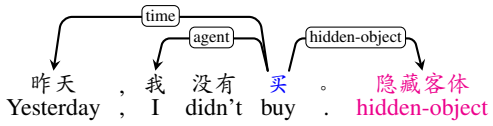


Figure 2: An example sentence with an omitted core argument in MuCPAD.

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, recent research on SRL makes very limited progress in the domain adaptation scenario.

As far as we know, there are three publicly available Chinese SRL datasets, i.e., Chinese Proposition Bank (CPB) (Xue and Palmer, 2005), Chinese NomBank (CNB) (Xue, 2006a), and Chinese SemBank (CSB) (Xia et al., 2017). All these datasets mainly contain canonical texts from newspapers or magazine/textbook articles.

In order to facilitate research on cross-domain SRL, this paper presents MuCPAD, a multi-domain Chinese predicate-argument dataset, consisting of 30,897 sentences and 92,051 predicates, from 6 different domains. Overall, MuCPAD 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 to define new frames for new predicates or new senses, and a lot of new predicates or new senses may appear in multi-domain texts.

(2) As shown in Figure 2, we explicitly annotate omitted core arguments with two special labels, i.e., "hidden-subject" and "hidden-object", in order to capture richer semantics expressed by predicates. It is ubiquitous that people try to avoid repetition by omitting previous content in context, especially in non-canonical Chinese texts.

(3) We adopt strict double annotation for all sentences in order to improve quality. If two annotators submit inconsistent results, a senior annotator determines the final answer. We also compile 53 pages of annotation guidelines to be studied and referred to by the annotators.

Based on our newly annotated MuCPAD, we conduct preliminary cross-domain SRL experiments and analysis. We enhance the basic SRL model by exploiting CPB2.0 as a heterogeneous dataset under the multi-task learning (MTL) framework, and by utilizing powerful contextualized word representations from pretrained language models (PLMs).

We release MuCPAD along with our annotation guidelines for research usage at https://github.com/SUDA-LA/MuCPAD.

2 Related Work

English SRL Data. Large-scale annotated data is a prerequisite to develop 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 and uses semantic frames to represent meanings of words. A frame corresponds to a sense of a word, and defines the specific meanings of its core roles (i.e., "A0-A5"). In other words, labels for core semantic roles have predicate-sense-specific meanings.

PropBank and NomBank are built by adding predicate-argument structures to the constituents of syntactic parser trees in Penn Treebank (Marcus et al., 1993). Their semantic roles are naturally span-based, instead of word-based. PropBank considers verbal predicates, while NomBank supplements nominal predicates. Following FrameNet, PropBank and NomBank use frames to represent semantic meanings of predicates. However, the development of frames is both time-consuming and labor-intensive, and requires annotators to be equipped with strong linguistic background.

The texts of PropBank and NomBank are mainly from the news domain, i.e., Wall Street Journal, except 426 sentences from the Brown corpus, which is usually used as an out-of-domain section of PropBank.

It is also noteworthy that there are PropBank-style SRL data for other languages, such as Portuguese (Duran and Aluísio, 2011, 2012), Arabic (Pradhan et al., 2012), Finnish (Haverinen et al., 2015), and Turkish (Sahin and Adali, 2018).

Chinese SRL Data. CPB (Xue and Palmer, 2005), CNB (Xue, 2006a), and CSB (Xia et al., 2017) are the three publicly available SRL datasets in Chinese. CPB and CNB, corresponding to PropBank and NomBank in English respectively, add predicate-argument structure of verbal predicates and nominal predicates into Penn Chinese Tree-

	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: Semantic role labels adopted in our guidelines. Predicates in the example sentences are marked by "[]".

bank (Marcus et al., 1993). The semantic roles are based on pre-defined frames as well. Moreover, sentences in CPB and CNB mainly come from canonical texts, such Xinhua newswire, Hong Kong news, and Sinorama Magazine (Hajic et al., 2009).

In contrast, CSB uses general-purpose role labels, such as "agent" and "patient", and the sentences are mainly from canonical texts such as online articles and news as well.

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., 2017; Clark et al., 2018; Zhao et al., 2018).

Kim et al. (2016) proposed 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) proposed parameter generation networks for cross-domain NER. They idea is dynamically generate parameters of network modules (such as BiLSTMs) according to predicted domain distribution.

To facilitate cross-domain Chinese dependency parsing research, Li et al. (2019a) proposed a large-scale multi-domain dataset for Chinese dependency parsing. They organized the NLPCC-2019 shared task on cross-domain dependency parsing (Peng et al., 2019). Li et al. (2019b) rank the first place in the shared task, based on a tri-training approach.

However, possibly due to the lack of multi-domain data, research on cross-domain SRL is scarce so far. We hope our newly annotated MuCPAD can promote future research in this direction.

3 Data Annotation

This section describe the annotation methodology and annotation process of MuCPAD in detail.

Annotation guidelines. After an extensive survey of previous works on SRL data annotation, we compile 53 pages of annotation guidelines. We adopt 24 fine-grained general-purpose role labels to capture the semantic relationships between predicates and arguments, as shown in Table 1, most of which are borrowed from the guidelines of CSB. In particular, we introduce two special labels, i.e., "hidden-subject" and "hidden-object", to explicitly annotate omitted core arguments. Our guidelines illustrate each each label in detail using concrete examples, and is gradually improved according to feedback of annotators during the course of the annotation project.

Data selection. We select the data to annotate from six domains, i.e., news, product blogs (PB), product comments (PC), web fictions (ZX), legals (LAW), and medical (MED) domains. Table 2 shows the data statistics.

News consists of the sentences in Chinese SemBank (Xia et al., 2017) and CoNLL-2009 Chinese dataset (Hajic et al., 2009). Specifically, we choose all 10.3K sentences with 16.5K predicates

	News	PB	PC	ZX	LAW	MED
#Sent	16,974	11,317	17,074	5,891	11,156	5,624
#Pred	40,989	3,753	3,890	1,575	2,813	1,892

Table 2: Statistics of annotated data. "#Pred" and "#Sent" represent the number of predicates and sentences.

from Chinese SemBank (Xia et al., 2017) and randomly select 6.7K sentences with 24.5K predicates from CoNLL-2009 Chinese dataset (Hajic et al., 2009). 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 section of People's Daily Online[3] and Sina.com[4].

After selecting the sentences, we also need to select the concerned predicates in the sentences for annotators to annotate their corresponding arguments. For news domain, we directly choose the predicates in Chinese SemBank and CoNLL-2009 Chinese dataset. For other 5 domains, we choose the predicates according to several predefined rules which are carefully designed by considering both the dependency tree structures of the sentences and a frame dictionary extracted from the Chinese frames[5]. For example, the root words of dependency trees are considered as predicates; the head words with "subject" or "object" dependency labels are considered as predicates; all the words that can be matched in the frame dictionary are considered as predicates.

Quality Control. We employ 20 undergraduate students as annotators, and select 5 experienced annotators as expert annotators to handle annotation inconsistency issues. All the annotators are paid for their work, and the salary is determined by their annotation quantity and quality. The average salary is 28 RMB per hour.

Before real annotation, each annotator is trained for several hours to be familiar with our guidelines and our annotation tool. During the annotation process, we adopt a strict double annotation workflow to guarantee the annotation quality. Specifically, each task is randomly assigned to two annotators to annotate independently. If the submissions from the pair of 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.

Annotation tool. We build a browser-based annotation tool to support the double annotation workflow. For each annotation sentence, the annotation tool highlights the predicate in the sentence for the annotators to annotate all the arguments of the highlighted predicate. We also design a "not-predicate" checkbox in the annotation tool and ask annotators to click this checkbox to inform us when the highlighted texts are not pre-defined predicates in our guidelines.

4 Analysis on MuCPAD

In this section, we analyze the MuCPAD 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 major row in Table 3 shows the predicate- and argument-wise annotation consistency ratios (Marcus et al., 1994; Guo et al., 2018) in all domains.

The predicate-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 annotators, and the numerator is the number of predicates with consistent arguments from all annotator pairs. We can see that the predicate-wise annotation consistency ratios in most domains are lower than 60%. Even the highest predicate-wise consistency ratio, which is achieved in 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 predicate-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

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

		News	PB	PC	ZX	LAW	MED	Average
Consistency ratio	predicate-wise	48.86	57.58	**71.23**	48.86	47.18	50.57	54.05
	argument-wise	74.48	74.67	**83.63**	74.65	71.49	73.24	75.36
argument-wise annotation 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 3: Analysis on consistency ratio and accuracy. "AVG" represents the average accuracy of a label in six domains, which is obtained by averaging the values of the six domains. Boldface indicates the maximum value of each row, underline represents the minimum value of each row.

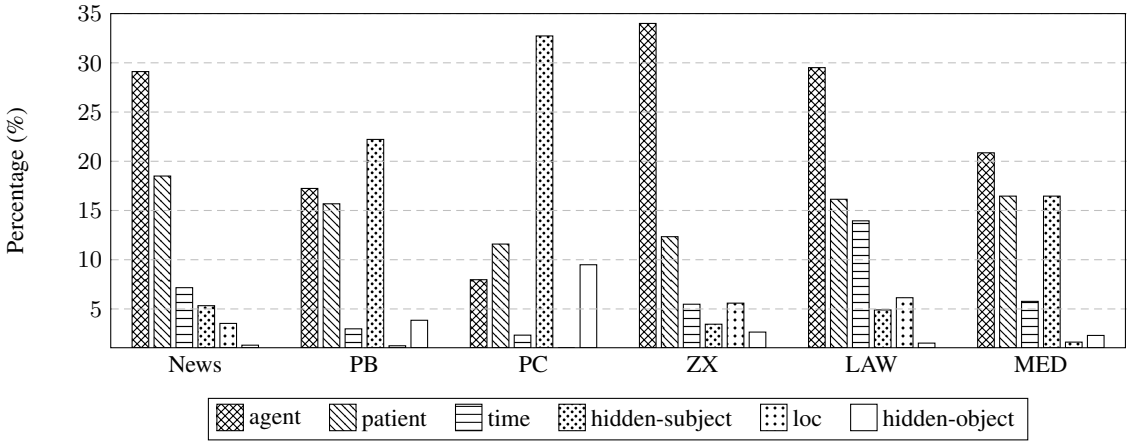


Figure 3: Label distribution in different domains.

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 argument-wise consistency ratio is defined as $\frac{\#\mathtt{Arg}_{\mathtt{annoA}\cap\mathtt{annoB}}}{\#\mathtt{Arg}_{\mathtt{annoA}\cup\mathtt{annoB}}}$, 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 the annotator pairs. As shown in Table 3, the argument-wise consistency ratios in most of the domains are lower than 75%, except that PC achieves the highest argument-wise consistency ratio of 83.63%.

Annotation accuracy. In the second major row of Table 3, we present the argument-wise annotation accuracy. Argument-wise annotation accuracy is defined as $\frac{\sum_{i=1}^{n}\#\mathtt{Arg}_{\mathtt{correct}_i}}{\#\mathtt{Arg}_{\mathtt{gold}}*2}$, where the numerator is the sum of the number of correctly annotated arguments submitted by all annotators, and the denominator is the total number of all gold arguments. The reason for "$*2$" in the denominator is that each task is annotated twice since it is assigned to two annotators for double annotation. The annotation accuracies in all the domains are more than 80%, indicating that our guidelines are reasonable, which ensures the quality of annotation data.

To gain more insights into the accuracy regarding different labels, we calculate the accuracy of 5 core labels and 2 non-core labels with high proportions for further analysis, which is shown in the third major row of Table 3. 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 labels and is prone to be missed by the annotators.

Label distribution. Figure 3 illustrates the label distribution in the 6 domains. The labels in Figure

3 are sorted in descending order by their proportion in News data. We choose 2 core labels and 2 non-core labels with the highest proportions in News data. Besides, we also analyze "hidden-subject" and "hidden-object" since their distributions in different domains are specific. As shown in Figure 3, the label distributions are vary across different domains.

For PC, both "hidden-subject" and "hidden-object" account for the largest proportion compared with that for all the other 5 domains, which means PC contains the most omitted core arguments. The reason is that PC is user-generated comments on a concerned product, and people tend to directly write the comment of the concerned product and omit the product name and the personal pronoun. For example, in the sentence "很 (very) 喜欢 (like), 买 (buy) 了 好 (very) 多 (much)", both the hidden-subjects and the hidden-objects of the predicates "喜欢 (like)" and "买 (buy)" are omitted. The omitted core arguments in PB also take relatively large proportion, similar to the reason for PC.

For ZX, it has the most "agent" roles and the least "hidden-subject" roles compared with other domains. This is owing to the genre of ZX texts, which are extracted from a popular Chinese fantasy novel with a lot of fictional characters. In order to make the story more understandable by readers, the names of the fictional characters are often explicitly written in the sentences, leading to more "agent" roles and fewer "hidden-subject" roles.

For LAW, it has more "time" roles and "loc" roles than other domains, since the elements (i.e., time and location) of the cases are usually frequently occurred to provide more accurate information.

For MED, the number of "hidden-subject" label accounts for a large proportion among all the labels in MED, only fewer than that of "agent" label, mainly because the descriptions of symptoms in MED usually omit the subjects. For example, in the sentence "酒精(alcohol) 中毒(poisoning): 发生(occur) 昏迷(coma) 不能(cannot) 催吐(induce vomiting), the subjects of the predicates "中毒(poisoning)", "发生(poisoning)", "昏迷(coma)", "不能(cannot)" are all omitted.

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 the 6 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". This is possibly because the POS for some predicates is subtle and vague in Chinese, causing the confusion of the argument labels. Taking the sentence "纽扣(Buttons) 一天(a day) 坏(getting broken) 一个(one)" as an example, "坏(getting broken)" may be misunderstood as an adjective and thus the argument "纽扣(buttons)" is incorrectly annotated as "expe". Actually, "坏(getting broken)" acts as a verb in this sentence and the correct label of "纽扣(buttons)" 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 argument. It is also difficult for annotators to distinguish "agent" and "patient". For example, in the sentence "新(new) 衣服(cloths) 被(was) 弄脏了(soiled)", the preposition "被(was)" 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, we conduct preliminary experiments, aiming to provide benchmark results. Specifically, we present a simple basic SRL model and enhance the model with the contextualized word representations from BERT for further improvements. Besides, we also present a MTL framework to improve the SRL performance by learning from multiple heterogeneous datasets simultaneously (Conia et al., 2021).

In this work, we focus on the *predicate-given* setting, which means we do argument identification and classification according to the given predicates in one sentence.

Following previous works (Cai et al., 2018; Zhang et al., 2019), we treat the predicate-given SRL task as a word pair classification problem and try to find the predicate-argument structure $\hat{\boldsymbol{y}}$ with

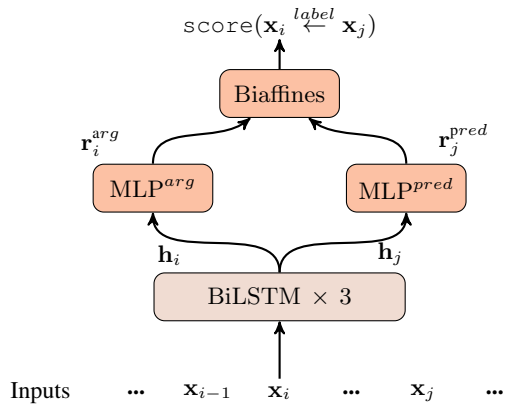


Figure 4: The architecture of our basic SRL model.

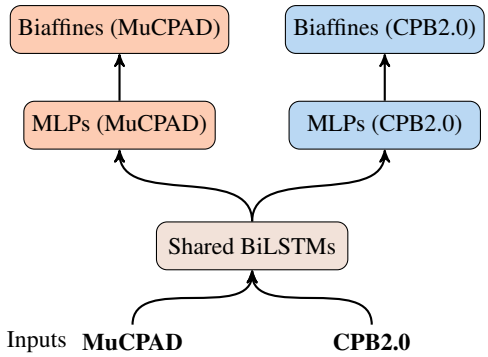


Figure 5: The framework of MTL.

the highest score:

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

where $\mathcal{Y}(\boldsymbol{x}) = \{(i, j, r) | i \in \mathcal{P}, 1 \leq j \leq n, r \in \mathcal{R}\}$ represents the set of all possible predicate-argument pairs. $\mathcal{P}$ is the set of given predicates, n is the number of sentence, and $\mathcal{L}$ 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 given predicate and the j-th word.

5.1 Basic SRL Model

Inspired by previous works (Cai et al., 2018; Zhang et al., 2019), we build a basic SRL model that utilizes the biaffine attention mechanism (Dozat and Manning, 2017) to score each candidate predicate-argument pair. Figure 4 shows the architecture of the basic model. During both training and evaluation, multiple predicates in the same sentence are handled simultaneously. First, the input sentence is encoded; then, scores between predicates and all other words are computed; finally, the roles of each predicate are determined via local classification.

The input vector is the concatenation of the pre-trained word embedding $\mathbf{e}_i^{pre}$, the randomly initialized word embedding $\mathbf{e}_i^{r}$, the character-based word representation $\mathbf{r}_i^{c}$, and the predicate indicator embedding $\mathbf{e}_i^{p}$.

$$\mathbf{x}_i = \mathbf{e}_i^{pre} \oplus \mathbf{e}_i^{r} \oplus \mathbf{r}_i^{c} \oplus \mathbf{e}_i^{p}$$

where $\mathbf{r}_i^{c}$ is produced by CNN, and the Boolean predicate indicator is true only for words that are given predicates.

A three-layer BiLSTM is applied to obtain context-aware representation of each word, i.e., $\mathbf{h}_i$.

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 argument).

Biaffines are used to compute scores of labels between a predicate and a word.

During training, we adopt the local cross-entropy loss. To obtain cross-domain results on the basic SRL model, we train the model on source domain data and make predictions on target domain data.

5.2 Enhancing with BERT

Recently proposed PLMs, such as BERT (Devlin et al., 2019), have shown the great power in learning and capturing contextualized representations and have proven 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 extract the fixed contextualized representations from BERT for words and treat them as additional features to augment the input representation, i.e., $\mathbf{x}_i = \mathbf{e}_i^{pre} \oplus \mathbf{e}_i^{r} \oplus \mathbf{r}_i^{c} \oplus \mathbf{e}_i^{p} \oplus \mathbf{e}_i^{BERT}$.

5.3 Utilizing Heterogeneous Data with MTL

MTL is a commonly used method to improve the model performance by learning the underlying knowledge from multiple related tasks or datasets (Collobert and Weston, 2008; Guo et al., 2016a; Li et al., 2019a). In this work, we design a MTL framework to utilize heterogeneous SRL datasets to boost the SRL model performance.

As shown in Figure 5, we extend the basic SRL model to the MTL framework. Specifically, the SRL parsing on MuCPAD 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 MuC-

(#Pred / #Sent)	News	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 4: Statistics of MuCPAD and CPB2.0. "#Pred" and "#Sent" represent the number of predicates and sentences.

	News		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
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
Baseline+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+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 5: Models' performance on the data of all domains.

PAD and CPB2.0 SRL parsing respectively.

6 Experiments

Data. Our experiments mainly focus on zero-shot single-source domain adaptation, that is, we have labeled training data for the source domain, and do not have labeled training data for the target domain. Specifically, we use the News domain of MuCPAD as the source domain, and the other five domains as target domains. The data statistics for source and target domains are shown in Table 4. For the auxiliary data used in the MTL framework, we randomly select 13,170 sentences with 72,616 predicates from CPB2.0 (Xue, 2006b), which belongs to the same newswire genre with the source domain 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 basic SRL model and MTL framework with PyTorch[6] and mainly follow the hyperparameters of Cai et al. (2018), such as the dimensions of embeddings, learning rate, and dropout ratios. We use bert-base-chinese[7] to obtain contextualized representations for words, and the dimension of the BERT representations is 768. During training, early stopping is triggered if the peak performance in dev data does not increase in 50 consecutive iterations.

Results of the basic model. The first row of Table 5 presents the results in the source/target domain dev/test data using the basic SRL model 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 different domains. Second, we find that the basic SRL model performs better on ZX and LAW compared with the other three target domains data, i.e., PB, PC, and MED. The probable reason is that ZX and LAW are novel and legal case, respectively, which are more canonical in text. 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.

Results with BERT. The second row of Table 5 shows the results of the baseline with BERT representations. We can see that the results of "Baseline+BERT" consistently increase by large margins compared with the corresponding baseline models without BERT (as shown in the first major row of table 5), demonstrating the great power of BERT in contextualized representation.

Results with heterogeneous CPB2.0. As shown in the third row of Table 5, benefiting from the additional semantic information provided by the auxiliary CPB2.0 data using the MTL framework, the SRL performance in all domains are improved compared with the baseline model. This indicates that the MTL framework is effective in capturing and learning the underlying common knowledge

[6]https://pytorch.org/

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

from heterogeneous data.

On the one hand, comparing the improvements brought by MTL in all domains, we find that MED data obtains the largest gains of 9.65%/8.56% F1 in dev/test, respectively. The main reason is that the MED data belongs to the same newswire domain as the auxiliary CPB2.0 data. On the other hand, the improvement in LAW is the smallest. This can be explained by the difference in label distribution between LAW and CPB2.0. For example, as mentioned in Section 4, the labels "time" and "loc" in LAW account for the largest proportion (13.95% and 6.14% respectively) compared with other domains. However, the proportions of "time" and "loc" in CPB2.0 data are only 6.10% and 3.40% respectively (about half of that in LAW). Therefore, CPB2.0 cannot provide much more valid information to increase the performance of these labels.

Results with BERT and heterogeneous CPB2.0. Finally, when utilizing both BERT representations and the heterogeneous CPB2.0 data on our baseline, the enhanced model gives the best or comparable results in 5 of the 6 domains, with an average increase of 0.5% F1, showing that the MTL framework is effective in utilizing heterogeneous data and can complement the information obtained from BERT representations.

7 Conclusions

This paper presents a multi-domain Chinese predicate-argument dataset, named MuCPAD, which consists of 30,897 sentences with 92,051 predicates and covers 6 different domains. In particular, we adopt a frame-free annotation methodology, which does not require high-level linguistic background for defining frames for large amounts of new predicates or new senses in multi-domain data. Besides, considering that omission of content words is ubiquitous 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 MuCPAD from different perspectives. Finally, we conduct preliminary cross-domain experiments and analysis on MuCPAD.

Acknowledgements

The authors would like to thank the anonymous reviewers for the helpful comments. We are greatly grateful to all our annotators for their hard work in data annotation. This work is supported by National Natural Science Foundation of China (Grant No. 62176173, 61876116).

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.

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.

Simone Conia, Andrea Bacciu, and Roberto Navigli. 2021. Unifying cross-lingual semantic role labeling with heterogeneous linguistic resources. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Simone Conia and Roberto Navigli. 2020. Bridging the gap in multilingual semantic role labeling: a language-agnostic approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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*.

Magali Sanches Duran and Sandra Maria Aluísio. 2011. Propbank-br: a brazilian portuguese corpus annotated with semantic role labels. In *STIL*.

Magali Sanches Duran and Sandra Maria Aluísio. 2012. Propbank-br: a brazilian treebank annotated with semantic role labels. In *LREC*.

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.

L. Guo, L. I. Zhenghua, X. Peng, and M. Zhang. 2018. Annotation guideline of chinese dependency treebank from multi-domain and multi-source texts. *Journal of Chinese Information Processing*.

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.

Katri Haverinen, Jenna Kanerva, Samuel Kohonen, Anna Missilä, Stina Ojala, Timo Viljanen, Veronika Laippala, and Filip Ginter. 2015. The finnish proposition bank. *Language Resources and Evaluation*, 49:907–926.

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.

Mitch Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology*, pages 114–119, Plainsboro, New Jerey, USA. Association for Computational Linguistics.

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

Lluís Màrquez, Xavier Carreras, Kenneth C Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: an introduction to the special issue.

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.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *arXiv preprint arXiv:2101.05779*.

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

Xue Peng, Zhenghua Li, Min Zhang, Rui Wang, Yue Zhang, and Luo Si. 2019. Overview of the nlpcc 2019 shared task: Cross-domain dependency parsing. In *Proceedings of NLPCC-2019*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, Jeju Island, Korea. Association for Computational Linguistics.

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

Gözde Gül Sahin and Esref Adali. 2018. Annotation of semantic roles for the turkish proposition bank. *Language Resources and Evaluation*, 52:673–706.

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.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

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.

Yue Zhang, Rui Wang, and Luo Si. 2019. Syntax-enhanced self-attention-based semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 616–626.

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.