Instructions for ACL 2020 Proceedings

Anonymous ACL submission

1 Introduction

Semantic role labeling(SRL) is the task of identifying and classifying the semantic roles of each argument of each predicate in a sentence. For example, for the sentence "Tom bought earphones from Jerry last month." and the predicate verb bought, SRL yields the following outputs:
[ARG0 Tom] [V bought] [ARG1 earphones] [ARG2 from Jerry] [AM-TMP last month]
Here ARG0 represents the buyer, ARG1 represents the thing bought, ARG2 represents the entity bought from, V represents the verb and AM-TMP is an adjunct indicating the timing of the action.
Recent work about English semantic role labeling systems involve: treating SRL as a BIO tagging problem and using deep bidirectional LSTMs as an end-to-end system for SRL without syntactic input (Zhou and Xu, 2015). Zhou and Xu (2015) introduce two other features to decrease the ambiguity of the predicate by expanding their context. Tan et al. (2018) use self-attention to capture the dependency relationships between two tokens. All these end-to-end systems outperform the traditional models (Pradhan et al., 2013; Täckström et al., 2015). Many natural language understanding methods use distributed word representations such as GloVe and Word2Vec to make the algorithms have better performance. Peters et al. (2018) introduce EMLo which is a deep contextualized word representation and can be easily integrated into existing models. Besides, Peters et al. (2018) add ELMo in the deep BiLSTM architecture in the SRL task and achieve a new state-of-the-art on the OntoNotes benchmark (Pradhan et al., 2013). After the development of Bert (Devlin et al., 2019), it is popular to fine-tuning Bert on downstream tasks such as sentence classification, question answering for its outstanding performance. Shi and Lin (2019) show a simple Bert-based model can also have the state-of-art performance in the SRL task without adding external features. Following Shi and Lin (2019), we apply the Bert model and a linear layer over the label set in English SRL task only using the input sentence and the predicate as the features. For Chinese semantic role labeling systems, SVMs are used as the classifier of the semantic role label for each node in the syntactic parse tree for the sentence (Sun and Jurafsky, 2004; Sun et al., 2009). Wang et al. (2015) use RNN and BiLSTM in Chinese SRL because RNN and LSTM have the advantage to better capture the contextual information, which is beneficial to model long-range and bidirectional dependencies simultaneously. However, all the approaches mentioned above use the lexical and syntactic information as the features such as POS tag, head word and whether the constitutent is before or after the verb. It is natural to think: if using Bert model without external features in English SRL systems can outperform the previous approaches rely on lexical or syntactic features, can we use the similar idea to applying large language pretrained models in Chinese SRL systems to minimize effort on the feature engineering but also can achieve the state-of-art results? Our experiment result shows that using Bert model and a linear layer over the label set without external features in the Chinese SRL system can outperform the previous approaches that use syntactic features.

2 Model

In our experiment, the task of argument identification and classification in SRL is to predict a sequence by using the sentence-predicate pair as input features and the label set which use BIO tagging strategy to tag the semantic role label of arguments for each predicate in the sentence.The input sequence is designed as [[CLS]sentence[SEP]predicate[SEP]] so the predi-

cate can interact with the whole sentence through the attention mechanisms.Then the input sequence is fed to the encoder of Bert model. The contextual representation of the input sentence ([CLS]sentence[SEP]) from Bert is followed by the predicate indicator embeddings. The output of the Bert model is the hidden states of the last layer, which is denoted as s G = [g1, g2, ..., gn]. For the final prediction on each token gi, the hidden state of predicate gp is concatenated to the hidden state of the token gi, and then fed into a layer of feed forward network over the label set. The softmax function is used as the activation function for the output layer in the feed forward network to select the semantic role label that has the highest probability among all the semantic role labels as the semantic role assigned to the argument of the predicate. We use the code [1] to construct our model and the code [2] to fine-tune the model on our preprocessed datasets.

3 Experiments

We train and evaluate our model on the English SRL dataset and Chinese SRL dataset separately. For English semantic role labeling, the Universal Proposition Banks dataset is used. For Chinese semantic role labeling, the CPB 1.0 dataset are used. The following subsections will give a short introduction of the dataset used in the experiment, the data processing and the experimental setup.

3.1 English Semantic Role Labeling Dataset

Based on the different ways on arguments annotation, there are two kinds of semantic role labeling dataset, one is span-based which annotates the syntactic spans of arguments which is called span based dataset such as PropBank(Kingsbury and Palmer, 2003), the other is dependency-based which annotates the syntactic heads of arguments as CoNLL 2008 and 2009 shared tasks (Surdeanu et al., 2008; Hajic et al., 2009) propose.For English semantic role labeling, we use the Universal Proposition Banks[3] as dataset.The English Universal Proposition Bank(English UPB) is released by projecting the rolesets and its arguments under the frame of PropBank on top of the constituent analysis in the original English Web Treebank to the dependency trees from the Universal Dependencies analysis.We use English UPB as the dataset to train and evaluate our model because the Universal Dependencies provide cross-linguistically consistent treebank annotation for many languages within a dependency-based lexicalist framework(Nivre et al., 2016), which is helpful for us to conduct the following experiments to compare the performance of our model on different languages.The English UPB dataset is dependency-based.The training dataset has 12543 sentences, the development set has 2002 sentences and the test set has 2077 sentences.We use half of the original training dataset which has 6274 sentences in the experiment as training the model on the full original training dataset will cost a large amount of the time. We use the original development set and test set for validation and evaluation.

In the data preprocessing, we modify the code[4] which is used to preprocess the dataset of CoNLL 2009 shared task (Hajic et al., 2009) in order to make it can be used to process the English UPB dataset which follows the CoNLL-U Format[5]. We use the BIO tagging strategy to tag the arguments and their semantic roles.In the English UPB dataset, there are many possible parts of speech for the predicate such as verb, noun, adjective and so on. For example, in the sentence "This killing of a respected cleric will be causing us trouble for years to come.", "killing" and "come" both are the predicates of this sentence, "killing" is a noun and "come" is a verb. Following the referenced code, we only tag the syntactic head of the arguments only if the predicate is a verb. In other words, in our training, development and test dataset, all the predicates are verbs. Besides, one sentence might have many different predicates and the arguments of each predicate will be tagged with their corresponding semantic roles.Therefore, a sentence has n predicates will be processed in n times. For instance, the sentence "DPA: Iraqi authorities announced that they had busted up 3 terrorist cells operating in Baghdad." has the following four BIO tagging sequences:

[1] https://github.com/huggingface/transformers/blob/master/src/transformers/models/bert/modeling_bert.py

[2] https://github.com/angel-daza/bert4srl/blob/master/train.py

[3] The dataset can be found here: https://github.com/System-T/UniversalPropositions/tree/master/UP_English-EWT

[4] https://github.com/angel-daza/bert4srl/blob/master/pre_processing/CoNLL_Annotations.py

[5] https://universaldependencies.org/format.html

{”seq_words”: [”DPA”, ”:”, ”Iraqi”, ”authorities”, ”announced”, ”that”, ”they”, ”had”, ”busted”, ”up”, ”3”, ”terrorist”, ”cells”, ”operating”, ”in”, ”Baghdad”, ”.”], ”BIO”: [”O”, ”O”, ”O”, ”B–ARG0”, ”B–V”, ”O”, ”O”, ”O”, ”B–ARG1”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”]}

{”seq_words”: [”DPA”, ”:”, ”Iraqi”, ”authorities”, ”announced”, ”that”, ”they”, ”had”, ”busted”, ”up”, ”3”, ”terrorist”, ”cells”, ”operating”, ”in”, ”Baghdad”, ”.”], ”BIO”: [”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”B–V”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”]}

{”seq_words”: [”DPA”, ”:”, ”Iraqi”, ”authorities”, ”announced”, ”that”, ”they”, ”had”, ”busted”, ”up”, ”3”, ”terrorist”, ”cells”, ”operating”, ”in”, ”Baghdad”, ”.”], ”BIO”: [”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”B–ARG0”, ”O”, ”B–V”, ”O”, ”O”, ”O”, ”B–ARG1”, ”O”, ”O”, ”O”, ”O”]}

{”seq_words”: [”DPA”, ”:”, ”Iraqi”, ”authorities”, ”announced”, ”that”, ”they”, ”had”, ”busted”, ”up”, ”3”, ”terrorist”, ”cells”, ”operating”, ”in”, ”Baghdad”, ”.”], ”BIO”: [”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”B–ARG0”, ”B–V”, ”O”, ”B–ARGM–LOC”, ”O”]}

3.2 Chinese Semantic Role Labeling Dataset

We train and evaluate our model on both span based SRL dataset and dependency based SRL dataset in Chinese. For the span based SRL, we use the Chinese Universal Propositions Banks dataset(Chinese UPB)[6]; for the dependency based SRL, the Chinese Proposition Bank 1.0(Xue and Palmer, 2003)(CPB 1.0) dataset is used.

We follow the original split of the train, dev and test set of the Chinese UPB datset. The train set has 3997 sentences, the dev set has 500 sentences and the test set has 500 sentences.We modify the code[7]to preprocess Chinese UPB dataset.Similar to preprocess the English UPB dataset,we use the BIO tagging strategy to tag the semantic role label of each argument.Each sentence that has n predicate will be processes n times. Here is an example sentence-BIO sequence pair in the training set: (The English translation of the sentence is : It seems simple, just choose one of two to make a decision, but in fact they represent the relatives and friends around you try to give you different opinions, but in the end, it is you who decides.)

”seq_words”: [”看似”, ”簡單”, ”,”, ”只”, ”是”, ”二”, ”選”, ”一”, ”做”, ”決擇”, ”,”, ”但”, ”其實”, ” 他們”, ”代表”, ”的”, ”是”, ”你”, ”周遭”, ”的”, ”親朋”, ”好友”, ”,”, ”試”, ”著”, ”給”, ”你”, ”不同”, ”的”, ”意見”, ”,”, ”但”, ”追根究底”, ”,”, ”最後”, ”決定”, ”的”, ”還是”, ”自己”, ”.”], ”BIO”: [”O”, ”B-AM-TMP”, ”O”, ”B-AM-ADV”, ”B-V”, ”O”, ”O”, ”O”, ”B-A2”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”, ”O”]

As the original CPB 1.0 dataset need to be purchased from LDC which might lead to a lot of time consuming as the grant has to be permitted by the university, we use part of CPB 1.0 dataset which is free to be downloaded and used from the link[8].This dataset is part of the original CPB 1.0 dataset and has already been splitted into train set, dev set and test set.The train set has 17840 sentences, the dev set has 1116 sentences.The dataset has already been processed.Each word is segmented in the sentence and assigned its part-of-speech.They use the BIOES tagging strategy to annotate the semantic role label of the arguments of the predicate in the sentence. For each sentence, there is only one predicate. Here is

[6]The dataset can be download here:https://github.com/System-T/UniversalPropositions/tree/master/UP_Chinese

[7]https://github.com/angel-daza/bert4srl/blob/master/pre_processing/CoNLL_Annotations.py

[8]https://github.com/Nrgeup/chinese_semantic_role_labeling/tree/master/src/data

an example of the token in the dataset: 党政(party and government)/NN/I-ARG0.
In the data processing, we use the last 2000 sentences in the train set of part of CPB 1.0 datatset as the test set to be used in the experiment. This is because in the test set of part of CPB 1.0 dataset, the semantic role labels of the arguments of the predicate in the sentences have not been annotated.In our experiment, the train set has 15840 sentences, the dev set has 1116 sentences and the test set has 2000 sentences. We modify the code[9][10] to convert the sentences in the dataset into the format of "sentence-BIOES sequence". Each sentence is only processes once as there are only one predicate in one sentence. Here is an example sentence and its BIOES tagging sequence in the train set:(The English translation of the sentence is Song Ho-kyung conveyed the cordial greetings from the DPRK leader to the Chinese leader, expressing deep gratitude to the Chinese party and government leaders and people for mourning the death of Chairman Kim Il-sung on behalf of the DPRK side.)

"seq_words": ["宋浩京", "", "了", "朝", "人", "", "中", "人", "的", "切", "候", "，", "代表", "朝方", "", "中", "党政", " 人", "和", "人民", "哀悼", "金日成", "主席", " 逝世", "表示", "深切", "意", "。"], "BIO": ["O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-ARG0", "I-ARG0", "I-ARG0", "I-ARG0", "E-ARG0", "rel", "B-ARG1", "I-ARG1", "E-ARG1", "O", "O", "O", "O"]

3.3 Experimental Set up

In our experiment, Bert-based-Chinese model[11] is trained,evaluated and tested on the Chinese SRL dataset.The Bert-base-multilingual-cased model(MBert) is trained,evaluated and tested on both Chinese SRL dataset and English SRL dataset.The learning rate is 2e-5.

[9] https://github.com/angel-daza/bert4srl/blob/master/pre_processing/CoNLL_Annotations.py

[10] https://github.com/Nrgeup/chinese_semantic_role_labeling/blob/master/src/data/1.build_dict.py

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

4 Experiment results and discussion

4.1 Train MBert on English UPB dataset

Table 1 shows the performance of MBert on English UPB dataset. The overall F1 score is 80.93. ARG0, ARG1, ARGM-MOD, ARGM-NEG, ARGM-TMP and R-ARG0 have the F1 score greater than 80, this can demonstrate that the model can classify these semantic role labels accurately. The F1 score of ARG3 is 0 because the model mainly misclassifies ARG3 as ARG2, which is shown in figure 1. The F1 score of ARGM-EXT is 29.63 as the ARGM-EXT is mostly misclassified as ARGM-ADV or ARGM-MNR by the model. The F1 score of ARGM-COM is 0 as the number of ARGM-COM is only 42 in the train set, which makes the classifier can not learn the weight to classify ARGM-COM accurately. For the continuation roles (C), the F1 score of C-ARG0, C-ARG1-DSP, C-ARG2, C-ARG3 and C-ARGM-LOC is 0 because the number of these continuation roles in the train set is smaller than 26. The F1 score of C-ARG1 is 16.67 and the model mainly misclassifies C-ARG1 as ARG2 and ARG1. For the reference roles (R), the model performs well in classifying R-ARG0, R-ARG1 but cannot classify R-ARG2, R-ARGM-ADV, R-ARGM-DIR, R-ARGM-LOC, R-ARGM-MNR and R-ARGM-TMP accurately as these reference roles has much lower proportion in the train set compared to the proportion of R-ARG0 and R-ARG1 in the train set.

4.2 Train MBert on Chinese UPB dataset

Table 2 shows the performance of MBert on Chinese UPB dataset. The overall F1 score is 55.30. Besides, the F1 score of each semantic role label class is lower than 62.00. This can illustrate that the model can not classify the semantic role label in the Chinese UPB dataset accurately. However, the table 3 shows that the overall F1 score of MBert on CPB 1.0 dataset is 83.19. The same model mechanism differs in F1 score of 21.19 on the two datasets let us speculate that the model can not perform well on the Chinese UPB dataset is because the Chinese UPB dataset has the lower quality than other UPB dataset.

4.3 Train MBert on CPB 1.0 dataset

The table 3 shows the performance of MBert on the CPB 1.0 dataset. The F1 score of ARG0 is 80.74 and the F1 score of ARG1 is 90.86, this can demon-

SRL class	P	R	F1	Number
ARG0	86.50	90.37	88.39	5823
ARG1	85.61	87.50	86.54	10093
ARG1-DSP	0.00	0.00	0.00	4
ARG2	76.30	77.19	76.74	3753
ARG3	0	0	0	201
ARG4	78.95	27.27	40.54	98
ARGM-ADV	65.84	62.15	63.94	1479
ARGM-CAU	52.00	63.41	57.14	227
ARGM-COM	0	0	0	42
ARGM-DIR	32.84	52.38	40.37	164
ARGM-DIS	79.14	70.97	74.83	594
ARGM-EXT	44.44	22.22	29.63	107
ARGM-GOL	0.00	0.00	0.00	81
ARGM-LOC	54.34	66.11	59.65	573
ARGM-MNR	48.89	60.00	53.88	553
ARGM-MOD	94.99	97.18	96.07	1463
ARGM-NEG	90.43	96.59	93.41	539
ARGM-PRD	0	0	0	107
ARGM-PRP	60.27	65.67	62.86	275
ARGM-PRR	52.81	68.12	59.49	296
ARGM-TMP	79.50	88.96	83.96	1725
C-ARG0	0.00	0.00	0.00	5
C-ARG1	38.46	10.64	16.67	175
C-ARG1-DSP	0.00	0.00	0.00	4
C-ARG2	0.00	0.00	0.00	25
C-ARG3	0.00	0.00	0.00	2
C-ARGM-LOC	0.00	0.00	0.00	0
R-ARG0	87.14	95.31	91.04	327
R-ARG1	74.00	80.43	77.08	283
R-ARG2	0.00	0.00	0.00	17
R-ARGM-ADV	0.00	0.00	0.00	17
R-ARGM-DIR	0.00	0.00	0.00	1
R-ARGM-LOC	33.33	44.44	38.10	23
R-ARGM-MNR	0.00	0.00	0.00	7
R-ARGM-TMP	0.00	0.00	0.00	17
Overall	80.47	81.40	80.93	

Table 1: the performance of MBert on English UPB dataset and the number of each semantic role label class in the English UPB dataset.

pred.\gold	ARG0	ARG1	ARG1-DSP	ARG2	ARG3	ARG4	ADV	CAU	COM	DIR	DIS	EXT	GOL	LOC	MNR	MOD	NEG	PRD	PRP	PRR	TMP	C-ARG0	C-ARG1	C-DSP	C-ARG2	C-LOC	R-ARG0	R-ARG1	R-ARG2	R-ADV	R-DIR	R-LOC	R-MNR	R-TMP
ARG0	1352	58	0	11	1	0	2	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
ARG1	34	2367	0	87	7	5	9	0	2	1	5	0	1	2	3	0	0	1	1	18	3	0	8	1	0	0	0	3	1	0	0	0	0	0
ARG1-DSP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG2	4	63	0	792	27	15	7	1	3	6	1	0	8	5	5	0	0	6	7	0	1	0	13	0	0	0	0	0	0	0	0	0	0	0
ARG3	0	0	0	0	0	0	0	0	0	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG4	0	0	0	2	0	15	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ADV	0	2	0	4	1	0	266	3	0	2	14	11	0	0	10	0	0	1	0	0	7	1	1	0	0	0	0	0	0	0	0	0	0	0
CAU	0	0	0	0	1	0	4	26	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
COM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DIR	0	4	0	11	2	6	0	0	0	22	0	0	0	4	0	0	0	1	0	0	1	0	4	0	0	0	0	0	0	0	0	0	0	0
DIS	1	1	1	0	0	0	12	1	0	0	110	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EXT	0	1	0	2	0	0	5	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GOL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LOC	0	5	0	15	2	11	4	1	0	3	0	0	2	119	4	0	0	1	0	0	3	0	0	0	0	0	0	0	0	0	0	3	0	0
MNR	1	3	0	6	4	0	13	0	1	0	0	9	0	5	66	0	0	3	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
MOD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	379	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NEG	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	170	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRP	0	2	0	7	2	1	3	0	0	0	0	0	3	0	0	0	0	0	44	47	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRR	0	40	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TMP	0	4	0	7	0	0	13	0	0	0	0	1	0	0	3	0	2	1	1	0	411	0	0	0	0	0	0	0	0	0	0	0	0	2
C-ARG0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C-ARG1	0	1	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	1	0	0	0	0	0	0	0	0	0
C-ARG1-DSP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C-ARG2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C-ARG3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C-LOC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R-ARG0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	61	1	0	0	0	0	1	0
R-ARG1	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0
R-ARG2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R-ADV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R-DIR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R-LOC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	6	0
R-MNR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R-TMP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 1: Confusion matrix of MBert trained on English UPB dataset

SRL Class	P	R	F1
ARG0	66.67	56.42	61.12
ARG1	70.00	55.07	61.64
ARG2	59.93	45.06	51.45
ARG4	0	0	0
ARGM-ADV	73.33	22.00	33.85
ARGM-CAU	0	0	0
ARGM-DIR	0	0	0
ARGM-EXT	0	0	0
ARGM-LOC	32.26	19.23	24.10
ARGM-MNR	56.25	16.36	25.35
ARGM-NEG	72.22	54.17	61.90
ARGM-TMP	46.47	37.97	41.79
C-A1	0	0	0
R-A0	0	0	0
R-A1	0	0	0
Overall	64.04	48.66	55.30

Table 2: the performance of MBert on Chinese UPB dataset.

pred.\gold	ARG0	ARG1	ARG2	ARG3	ADV	BNF	CND	DIR	DIS	EXT	LOC	MNR	PRP	TPC	TMP
ARG0	3532	204	323	10	18	0	0	0	1	0	23	6	2	8	12
ARG1	140	9032	158	17	4	4	0	0	0	9	9	2	31	0	2
ARG2	9	76	421	2	0	0	0	30	0	1	3	0	0	0	0
ARG3	0	24	6	0	0	0	0	2	0	0	0	0	0	3	0
ADV	24	7	2	0	886	0	3	9	39	0	2	16	2	2	17
BNF	0	1	3	0	0	34	0	0	0	0	0	0	11	0	0
CND	0	0	0	0	21	0	58	0	0	0	0	0	0	0	0
DIR	12	1	3	0	1	0	0	19	0	0	0	0	0	7	0
DIS	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
EXT	0	5	0	0	0	0	0	0	0	25	0	0	0	0	0
LOC	4	5	160	7	2	0	0	0	0	0	677	0	0	6	13
MNR	0	0	13	0	13	0	0	0	0	0	0	261	1	0	0
PRP	0	0	0	0	0	0	0	0	0	0	0	0	191	0	0
TPC	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TMP	20	29	4	0	11	0	0	0	0	0	0	0	7	0	994

Figure 2: Confusion matrix of MBert trained on CPB 1.0 dataset

strate that our model can perform well in classifying the ARG0 and ARG1 in the CPB 1.0 dataset which mainly due to ARG0 and ARG1 have the largest proportion in the dataset. The model has the F1 score of 45.07 in classifying ARG2 and mostly misclassify ARG2 as ARG0, which is shown in the figure 2. The F1 score of ARG3 is 0 and the model mostly misclassifies ARG3 as ARG1. The model can classify ARGM-ADV, ARGM-BNF, ARGM-CND, ARGM-EXT, ARGM-LOC, ARGM-MNR, ARGM-PRP and ARGM-TMP accurately. However, the model cannot perform well in classifying

SRL class	P	R	F1	Number
ARG0	76.75	85.17	80.74	25760
ARG1	90.50	91.22	90.86	18691
ARG2	64.37	34.68	45.07	2734
ARG3	0	0	0	446
ARGM-ADV	73.77	80.25	76.88	7679
ARGM-BNF	69.39	87.18	77.27	231
ARGM-CND	56.86	53.21	54.98	277
ARGM-DIR	37.25	31.67	34.23	339
ARGM-DIS	100.00	3.90	7.50	155
ARGM-EXT	80.65	64.10	71.43	181
ARGM-LOC	72.56	90.99	80.74	3342
ARGM-MNR	59.59	80.31	68.41	2046
ARGM-PRP	80.93	64.75	71.94	632
ARGM-TPC	0	0	0	248
ARGM-TMP	85.99	84.45	85.21	5592
Overall	82.77	83.62	83.19	

Table 3: the performance of MBert on CPB 1.0 dataset and the number of each semantic role label in the CPB 1.0 dataset.

ARGM-DIR, ARGM-DIS and ARGM-TPC. The F1 score of ARGM-DIR is 34.23 and the model mostly misclassify ARGM-DIR as ARG2. The F1 score of ARGM-DIS is 7.5 and the model mostly misclassify ARGM-DIS as ARGM-ADV. The F1 score of ARGM-TPC is 0 and the model mostly misclassify ARGM-TPC as ARG0 and ARGM-DIR.

4.4 Compare the performance of MBert and Bert-based-Chinese on CPB 1.0 dataset

The table 3 shows the overall F1 score of MBert on the CPB 1.0 dataset is 83.19 and the table 4 shows the overall F1 score of Bert-based-Chinese on the CPB 1.0 dataset is 82.67. This can demonstrate that MBert outperforms Bert-based-Chinese slightly on the CPB 1.0 dataset although Bert-based-Chinese is trained on Chinese corpus and is supposed to generate more accurate word representation in Chinese than MBert which is trained on the multi-lingual corpus. However, the MBert applied in our experiment is trained for 10 epoches and Bert-based-Chinese is trained for 9 epoches. The training loss of them have not converged to zero. Therefore, Bert-based-Chinese still has the possibility to outperform MBert on the CPB 1.0 dataset if we can train the two models until the training loss converges to zero.

For each semantic role label, we compare the F1 score performed by MBert and the F1 score per-

SRL	P	R	F1
ARG0	76.57	83.00	79.66
ARG1	90.41	90.60	90.51
ARG2	58.77	22.90	32.96
ARG3	0	0	0
ARG4	0	0	0
ARGM-ADV	78.44	82.07	80.21
ARGM-BNF	92.68	97.44	95.00
ARGM-CND	44.55	44.95	44.75
ARGM-DIR	31.91	25.00	28.04
ARGM-DIS	100.00	5.19	9.88
ARGM-EXT	93.10	69.23	79.41
ARGM-LOC	70.56	88.58	78.55
ARGM-MNR	70.03	79.08	74.28
ARGM-PRP	80.00	67.80	73.39
ARGM-TPC	16.22	23.08	19.05
ARGM-TMP	86.90	86.83	86.87
Overall	83.07	82.27	82.67

Table 4: The F1 score of each semantic role label performed by Bert-based-Chinese on CPB 1.0 dataset.

pred.\gold	ARG0	ARG1	ARG2	ARG3	ADV	BNF	CND	DIR	DIS	EXT	LOC	MNR	PRP	TPC	TMP
ARG0	3442	224	301	9	6	0	0	0	0	1	17	7	4	8	4
ARG1	153	8970	158	20	0	1	0	6	1	7	16	2	31	0	1
ARG2	12	65	278	0	12	0	0	26	0	4	3	0	3	1	0
ARG3	1	31	16	0	0	0	0	2	0	0	0	0	0	1	0
ADV	19	2	7	0	906	0	3	0	34	0	0	14	9	0	24
BNF	0	0	2	0	0	38	0	0	0	0	0	0	0	0	0
CND	4	0	0	0	34	0	49	0	2	0	0	0	0	0	0
DIR	4	2	5	0	2	0	0	15	0	0	0	0	0	10	0
DIS	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0
EXT	0	2	0	0	0	0	0	0	0	27	0	0	0	0	0
LOC	14	4	164	4	2	0	0	0	0	0	659	0	0	6	16
MNR	8	0	14	0	6	0	0	0	0	0	0	257	1	0	0
PRP	0	4	2	0	16	0	0	0	0	0	0	0	200	0	0
TPC	22	0	0	0	6	0	0	0	0	0	0	0	0	6	0
TMP	14	16	3	0	9	0	10	0	0	0	0	0	0	0	1022

Figure 3: confusion matrix of Bert-based-Chinese on CPB 1.0 dataset

formed by Bert-based-Chinese. We find that for most of the semantic role labels, their F1 scores performed by MBert are roughly equal to their F1 scores performed by Bert-based-Chinese. However, Bert-based-Chinese still has its unique advantage on classifying some semantic role labels compared to MBert as the F1 socres of some semantic role labels performed by Bert-based-Chinese are higher than the F1 scores of these semantic roles performed by MBert. The F1 score of ARGM-BNF performed by Bert-based-Chinese is 95.00, which is 17.73 higher than the F1 score of ARGM-BNF performed by MBert. The F1 score of ARGM-TPC performed by Bert-based-Chinese is 19.05 and the F1 score of ARGM-TPC performed by MBert is 0 because Mbert cannot classify even one AGRM-TPC accurately. Besides, for some semantic role labels, the F1 score of them performed by MBert is higher than the F1 core of them performed by Bert-based-Chinese. The F1 score of ARG2 performed by MBert is 45.07, which is 12.11 higher than the F1 score of ARG2 performed by Bert-based-Chinese. The F1 score of ARGM-CND performed by MBert is 54.98 and the F1 score of ARGM-CND performed by Bert-based-Chinese is 44.75 because Bert-based-Chinese model misclassifies ARGM-CND as ARGM-TMP which is shown in figure 3 while MBert can classify most of ARGM-CND accurately.

4.5 Compare the performance of MBert in Chinese SRL and English SRL

In order to compare the difference between Chinese SRL and English SRL, we use the performance of MBert trained on CPB 1.0 dataset as the result of Chinese SRL and the performance of MBert trained on English UPB dataset as the result of English SRL. Firstly, the F1 score of ARG0, ARG1, ARGM-ADV and ARGM-TMP in Chinese SRL is close to the F1 score of these semantic role labels in English SRL. Secondly, the F1 score of ARG2, ARGM-DIR and ARGM-DIS in Chinese is lower than the F1 score of these semantic role labels in English. The F1 score of ARG2 in Chinese is 45.07 and the F1 score of ARG2 in English is 76.74 because the model mostly misclassifies ARG2 as ARG0 in Chinese but can classify most of ARG2 accurately in English. The F1 score of ARGM-DIR in Chinese is 34.23 and the F1 score of ARGM-DIR in English is 40.37. The model mainly misclassifies ARGM-DIR as ARG2 in Chinese but can classify most of ARGM-DIR in English.The F1 score of ARGM-DIS in Chinese is 7.5 and the F1 score of ARGM-DIS in English is 74.83. Most of ARGM-DIS are classifies as ARGM-ADV by the model in Chinese but the model can classify most of ARGM-DIS accurately in English.Thirdly, the F1 score of ARGM-LOC, ARGM-MNR and ARGM-PRP in Chinese is higher than the F1 score of these semantic role labels in English which might because the number of these semantic role labels in Chinese SRL dataset is larger than the number of these semantic role labels in English SRL dataset, which is shown in table 5.The model is therefore benefited by the sufficient training data. The F1 score of ARG2 in Chinese is 45.07 and the F1 score of ARG2 in English is 76.74. ARG2 is mainly misclassified as ARG0 in Chinese but the model can classify most of ARG2 in English accurately. Finally, the model cannot perform well in classifying ARG3 in both Chinese and English. The F1 score of ARG3 in both Chinese and English is 0. The model mainly misclassifies ARG3 as ARG1 mostly in Chinese and mainly misclassifies ARG3 as ARG2 in English.

References

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

Jan Hajic, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, M Antònia Martí, Lluís

SRL class	Chinese F1	English F1	Number in Chinese	Number in English
ARG0	80.74	88.39	18691	5823
ARG1	90.86	86.54	25760	10093
ARG2	45.07	76.74	2734	3753
ARG3	0	0	446	201
ARGM-ADV	76.88	63.94	7679	227
ARGM-DIR	34.23	40.37	339	164
ARGM-DIS	7.5	74.83	155	594
ARGM-EXT	71.43	29.63	181	107
ARGM-LOC	80.74	53.69	3342	573
ARGM-MNR	68.41	53.88	2046	553
ARGM-PRP	71.94	62.86	632	275
ARGM-TMP	85.21	83.96	5592	1725

Table 5: The F1 score of common semantic role label performed by MBert in CPB 1.0 dataset and English UPB dataset and the number of these semantic role labels in CPB 1.0 dataset and English UPB dataset.

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 the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18.

Paul Kingsbury and Martha Palmer. 2003. Propbank: the next level of treebank. In *Proceedings of Treebanks and lexical Theories*, volume 3. Citeseer.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling.

Honglin Sun and Dan Jurafsky. 2004. Shallow semantic parsing of chinese. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 249–256.

Weiwei Sun, Zhifang Sui, Meng Wang, and Xin Wang. 2009. Chinese semantic role labeling with shallow parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1475–1483.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177.

Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015. Efficient inference and structured learning for semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:29–41.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep semantic role labeling with self-attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Zhen Wang, Tingsong Jiang, Baobao Chang, and Zhifang Sui. 2015. Chinese semantic role labeling with bidirectional recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1631, Lisbon, Portugal. Association for Computational Linguistics.

Nianwen Xue and Martha Palmer. 2003. Annotating the propositions in the Penn Chinese treebank. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 47–54, Sapporo, Japan. Association for Computational Linguistics.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. 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 1127–1137, Beijing, China. Association for Computational Linguistics.