Domain Adaptation of SRL Systems for Biological Processes

Dheeraj Rajagopal[♠*] **Nidhi Vyas**[♠*] **Aditya Siddhant**[♠] **Anirudha Rayasam**[♠]
Niket Tandon[♣] **Eduard Hovy**[♠]
[♠]Carnegie Mellon University
[♣]Allen Institute for Artificial Intelligence
{dheeraj,nkvyas,asiddhan,arayasam}@andrew.cmu.edu
nikett@allenai.org, hovy@cmu.edu

Abstract

Domain adaptation remains one of the most challenging aspects in the wide-spread use of Semantic Role Labeling (SRL) systems. Current state-of-the-art methods are typically trained on large-scale datasets, but their performances do not directly transfer to low-resource domain-specific settings. In this paper, we propose two approaches for domain adaptation in biological domain that involve pre-training LSTM-CRF based on existing large-scale datasets and adapting it for a low-resource corpus of biological processes. Our first approach defines a mapping between the source labels and the target labels, and the other approach modifies the final CRF layer in sequence-labeling neural network architecture. We perform our experiments on ProcessBank (Berant et al., 2014) dataset which contains less than 200 paragraphs on biological processes. We improve over the previous state-of-the-art system on this dataset by 21 F1 points. We also show that, by incorporating event-event relationship in ProcessBank, we are able to achieve an additional 2.6 F1 gain, giving us possible insights into how to improve SRL systems for biological process using richer annotations.

1 Introduction

Semantic Role Labeling (SRL) is shallow semantic representation of a sentence, that allows us to capture the roles of arguments that anchor around an event. Despite significant recent progress in Deep SRL systems (He et al., 2017; Tan et al., 2017), there has been limited work in adapting such systems to low resource domain-specific scenarios where the label space of both domains are completely different. Additionally, existing domain adaptation for SRL requires an overhead of annotating the new corpus using guidelines similar to the source dataset, and every domain-specific corpora might not necessarily adhere to the same label structure and similar annotation guidelines.

We present two different domain adaptation strategies that rely on training the model on a large corpora (source dataset) and fine-tuning on a low-resource domain-specific corpus (target dataset), more specifically biological processes domain. The first approach uses mappings from the source label space to the target label space. For this, we present DeepSRL-CRF, which incorporates a CRF layer over the DeepSRL model (He et al., 2017) with an intermediate step of mapping labels from source to target domain. For the second approach, we use a CNN-LSTM-CRF model to pre-train the neural network weights on the source domain, and adapt the final CRF layer of the network based on the target label space. We then fine-tune the model on the target dataset.

For empirical evaluation, we explore the challenge of SRL in ProcessBank dataset, where the target domain (biological processes) is drastically different compared to the source domain (news). Both of our approaches are effective for adapting SRL systems for biological processes. Compared to the previous best system, we get an improvement of about 24 F1 points when we use label-mapping approach, and about 21 F1 point improvement when we adapt the final CRF layer. Our contributions can be summarized as follows:

1. Two different approaches for domain adaptation of SRL for biological processes, with our code and models publicly available [1]

2. An in-depth analysis for each of the domain adaptation strategies, both perform significantly better in low-resource SRL for biological processes

3. Analysis of the model performance when the

[*]Both authors equally contributed to the paper.

[1]https://github.com/dheerajrajagopal/SciQA

Proceedings of the BioNLP 2019 workshop, pages 80–87
Florence, Italy, August 1, 2019.

target corpus is annotated with event-event relationships to the SRL corpus

2 Models

To label the event-argument relationships, we propose two models inspired from the current state-of-art on the SRL and NER literature. Since our downstream task lends itself to a low-resource setting, we hypothesize that an LSTM-CRF architecture would be better suited for the role-labeling task.

DeepSRL-CRF : We introduce DeepSRL-CRF, that is inspired from DeepSRL (He et al., 2017). The DeepSRL-CRF model uses a stacked BiLSTM network structure as its representation layer with a CRF layer on top. The overall model uses stacked BiLSTMs using an interleaved structure, as proposed in Zhou and Xu (2015). As described in the original model, we use gated highway connections (Zhang et al., 2016; Srivastava et al., 2015) to prevent over-fitting.

CNN-LSTM-CRF : We adapt the state-of-art sequence-labeling model by Ma and Hovy (2016). This is an end-to-end model, which uses a BiLSTM, Convolutional Neural Network (CNN) and CRF to capture both word- and character-level representations. The model first uses a CNN to capture character-level representation. These embeddings are concatenated with the word-level embeddings and fed into a BiLSTM to capture the contextual information at word-level. Here, we adapt this model to additionally concatenate 100-dimensional predicate indicators for every word before feeding the result into a BiLSTM. The output vectors from the BiLSTM are fed into the CRF layer, which jointly decodes the best sequence. The model uses dropout layers for both CNN and BiLSTM to prevent overfitting.

3 Domain Adaptation

Label Mapping : In our first approach, we perform domain by mapping each label from the target label-space to the source label-space by aligning it to the closest label from the source dataset. Since we used the CoNLL-2005 and CoNLL-2012 datasets for pre-training, we used the PropBank labels to map each relation in ProcessBank according to the PropBank annotation guidelines. Although there is human intervention in the pipeline, it is time-efficient since this process has to be done only once for a target dataset. We asked three independent annotators to perform the mapping of these labels, and the majority voted mapping was used as the final mapping scheme. In case of no majority vote, the annotators discussed to reach a consensus. We had an inter-annotator agreement of 0.8. The entire process for ProcessBank dataset took approximately two hours. The mapping for individual relationships are given in Table 1. The network architecture did not change throughout the training process for both source and target domains. The final CRF layer of the neural network maintains the same dimensions as the source domain.

PropBank	ProcessBank
ARG0	Agent
ARGM-LOC	Location
ARG2	Theme
ARG3	Source
ARG4	Destination
ARG1	Result
ARGM-MNR	Other

Table 1: Label Mapping: PropBank to ProcessBank

Adapting the CRF Layer : In the second approach, we maintain the network weights for the BiLSTM layers constant from the pre-training and we learn the transition and emission probabilities from scratch in the target domain dataset. More specifically, we first train the entire model on CoNLL-2005 and CoNLL-2012 SRL data. Next, we replace the final CRF layer with the label-space dimensions in our target domain, and keep the remaining weights in the model as is. Finally, we start fine-tuning the entire model by training it on the target data. Contrary to the previous approach, this approach does not require any manual intervention.

Event Interactions : The ProcessBank dataset is also annotated with event-event interactions. In our model, we also study whether event-event structure is important in predicting the event-argument structure. We leverage this additional event-event interaction annotations, and add them to the input to predict the event-argument role-labels. From an annotation perspective, this experiment helps us analyze whether the event-event structure labels are the bottle-neck for better SRL performance - especially in domain specific settings.

4 Experiments

Experimental Setup : For evaluation, we use the CoNLL-2005 (Carreras and Màrquez, 2005) and CoNLL-2012 (Pradhan et al., 2013) datasets as our primary large-scale datasets with the standard splits. For the domain adaptation scenario, we use the ProcessBank dataset (Berant et al., 2014)[2]. We used 134 annotated paragraphs for training, 19 for development and 50 for testing. Each passage in the ProcessBank dataset describes a *process*, defined by a directed graph (T, A, E_{tt}, E_{ta}), where nodes T denote event triggers and A denote their corresponding arguments. E_{tt} represents labeled edges event-event relations and E_{ta} describe event-argument relations. The edges E_{ta} are annotated with semantic roles AGENT, THEME, SOURCE, DESTINATION, LOCATION, RESULT and OTHER. Each E_{tt} edge between event and another event is annotated with the relations CAUSE, ENABLE and PREVENT. Our experiments primarily focus on the prediction of the event-argument structures E_{ta} since the source datasets that we use for domain adaption do not contain any event-event relationship annotation.

Baselines : In our first set of baselines, we compare our models on the CoNLL-2005 and CoNLL-2012 tasks. We use the previous state-of-the-art SRL system from He et al. (2018) as our baseline. [3]. Since our model is based on LSTM-CRF hybrid architecture, we implement two other baselines for our approach. We use a standard BiLSTM-CRF model (Huang et al., 2015), and a model based on the structured attention proposed in Liu and Lapata (2017) which uses CRF style structure in the intermediate layer. For a fair comparison, we augmented this structured attention based network with a CRF layer on top. We use 300D GLoVe embeddings (Pennington et al., 2014) across all models. For domain adaptation, we use the original system from Berant et al. (2014) as the baseline. It uses the approach in Punyakanok et al. (2008), where for each trigger, a set of argument candidates are first determined, and then a binary classifier uses argument identification features to prune this set with high recall.

5 Results

Semantic Role Labeling : Table 2 shows the SRL results[4] for the CoNLL-2005 and CoNLL-2012 datasets across all baseline models. From the table, it is evident that our DeepSRL-CRF model with ELMo embeddings performs slightly lesser than the current state-of-the-art SRL model DeepSRL with ELMo. We were able to perform significantly better than the other baselines – BiLSTM-CRF and Strucutured Attention model. Our DeepSRL-CRF model without ELMo performed significantly lower and the improvement was notably higher with ELMo.

Domain Adaptation : For all our domain adaptation experiments, we found that the DeepSRL and DeepSRL-CRF models reach similar F1 scores without any pre-training. Table 3 shows the results for the set of models that were trained for domain adaptation using label mapping. After pretraining it on the CoNLL 2005 and CoNLL-2012 dataset for 50 epochs, we fine-tuned the weights on the ProcessBank dataset without making any changes to the network. The results signify that the models that were effective for a large dataset, might not achieve similar gains when restricted to specific low-resource domains. The DeepSRL-CRF model, after incorporating event-event relationships, outperforms the previous system from Berant et al. (2014) by about 24 F1 points.

In our second domain adaptation approach, we test the CNN-LSTM-CRF model by learning the final CRF layer with transition and emission probabilities for the target label space. The CNN-LSTM-CRF model, without any pretraining achieves 40.62 F1 which is similar to previous performance from Berant et al. (2014). However, after pre-training it on CoNLL 2005 and CoNLL-2012 dataset for 50 epochs, the models outperforms by about 21.7 F1 points. Adapting the CRF layer, with transition and emission probabilities for the target domain data in its label space, shows impressive gains in the low-resource setting, specially when there is a limitation for using any human-intervention in the domain adaptation process. Although empirically effective, we believe that there is immense scope to understanding the impact of better initialization from a theoretical perspective. We also observe that pretraining

[2]For dataset statistics, we refer readers to Berant et al. (2014), Table 1. We use the same training and test split provided in the original dataset. We further split the training set into training and development set.

[3]Due to resource limitations, we were unable to run the same model for 500 epochs, so we report results from their paper for CoNLL-2005 and CoNLL-2012 datasets

[4]We use span-based precision, recall and F1 measure

Model	CoNLL-2005 (WSJ)			CoNLL-2012 (OntoNotes)		
	P	R	F1	P	R	F1
BiLSTM-CRF	80.9	79.4	80.3	80.0	77.8	78.9
Structured Attention	81.0	80.1	80.5	79.6	77.9	78.8
CNN-LSTM-CRF	82.1	82.7	82.4	81.7	83.0	82.3
DeepSRL	81.6	81.6	81.6	81.8	81.4	81.6
DeepSRL-ELMo	-	-	**87.4**	-	-	**85.5**
DeepSRLCRF	35.0	46.3	40.0	51.6	78.1	62.2
DeepSRLCRF-ELMo	84.7	83.6	84.1	84.4	85.8	85.1

Table 2: SRL results for CoNLL-2005 and CoNLL-2012 datasets. DeepSRL-ELMo resuls from He et al. (2018)

Model	Development			Test		
	P	R	F1	P	R	F1
Berant et al. (2014)	-	-	-	43.4	34.4	38.3
			CoNLL-2005			
DeepSRL	46.7	53.7	50.0	46.1	51.0	48.5
DeepSRL-ELMo	55.0	48.0	51.7	48.8	41.7	44.5
DeepSRLCRF	51.4	58.1	54.5	50.8	57.0	53.7
DeepSRLCRF-ELMo	53.5	66.2	59.2	49.1	63.2	55.3
+ Event relations	63.0	63.7	63.3	61.0	62.2	61.6
			CoNLL-2012			
DeepSRL	51.1	56.9	53.9	43.9	49.0	46.3
DeepSRL-ELMo	52.6	50.0	51.2	48.1	43.2	44.6
DeepSRLCRF	45.9	63.1	53.1	40.3	56.7	47.2
DeepSRLCRF-ELMo	44.6	**67.5**	53.7	36.9	62.1	46.3
+ Event relations	**65.0**	65.0	**65.0**	**62.1**	**63.0**	**62.6**

Table 3: SRL results for ProcessBank dataset - Domain adaptation using label mapping.

Model	Test		
	P	R	F1
Berant et al. (2014)	43.4	34.4	38.3
No pre-training	40.6	40.6	40.6
		CoNLL-2005	
BiLSTM-CRF	44.7	42.3	43.4
CNN-LSTM-CRF	56.8	55.5	56.1
+Event relations	55.3	53.4	54.4
		CoNLL-2012	
BiLSTM-CRF	42.8	41.0	42.3
CNN-LSTM-CRF	**59.7**	**60.2**	**60.0**
+ Event relations	58.8	57.7	58.3

Table 4: Results for ProcessBank - Domain adaption by replacing the CRF layer

on CoNLL-2012 dataset was more effective compared to pre-training on CoNLL-2005 dataset for this model. The former has about 35000 more training data instances than later.

Which domain adaptation technique works best? Our results show that the DeepSRL-CRF model based on label mapping approach perform the best overall (improvement of 24 F1 points) assuming we have event-event relationship annotations. In a setting where there are multiple datasets of different domains, training different network for each of the datasets might be cumbersome. We believe that the domain adaptation based on label mapping would suit such situations better. However, in the cases where there is no explicit label mapping possible or no readily available event-event interaction annotations in target domains, resorting to replacing the CRF layer would be the most effective for domain adaption gains. Our CNN-LSTM-CRF model achieves an improvement of 21 F1 points by replacing the CRF layer without event-event annotations. One of the drawbacks of this system is that it cannot be trained end-to-end. Given that there is limited overhead in modifying the architecture, we believe this wouldn't be a bottleneck for NLP systems. If end-to-end training is a hard constraint, we resort to our DeepSRL-CRF model. In terms of generalization capability and performance, pre-training on the CoNLL-2012 dataset and fine-tuning on the ProcessBank dataset with explicit label mapping with additional event-event relations gives us the

best results. [5]

6 Related Work

Domain adaptation leverages on large-scale datasets to help improve the performance on other smaller and similar tasks. From the SRL perspective, one of the earliest work from Daume III and Marcu (2006) showed simple but effective ways for 'transferring the learning' from a source to a target domain. Building on strong feature-rich models, Dahlmeier and Ng (2010) analyzed various features and techniques that are used for domain adaptation and conducted an extensive study for biological SRL task. Later, Lim et al. (2014) proposed a model that uses structured learning for domain adaptation. Although effective, these methods rely on hand-annotated features. Recently, there have been neural-network based attempts at Domain adaptation in SRL. Do et al. (2015) combined the knowledge from a neural language model and external linguistic resource for domain adaptation for biomedical corpora. Our work closely aligns to this work from a modeling stand-point. Our target domain is biological process descriptions from high-school biology without restrictions of PropBank style annotations.

Our work builds on multiple existing works, especially the dataset from Berant et al. (2014), using the thematic roles defined in Palmer et al. (2005). Our approach is inspired by the recent success in including structured representations in deep neural networks (He et al., 2017; Ma and Hovy, 2016) for structured prediction tasks. Our primary motivation is to improve the system performance for low-resource domain-specific event-argument labeling tasks, particularly biological processes. Argument labeling, specifically, SRL as been used for biomedical domain previously. E.g. Shah and Bork (2006) applied SRL in the LSAT system to identify sentences with gene transcripts, and Bethard et al. (2008) applied SRL to extract information about protein movement. However, developing annotated SRL data for each task can be expensive.

7 Conclusion

In this work, we present two new approaches to adapt deep learning models trained on large scale datasets, to smaller domain-specific biological process dataset. We present a LSTM-CRF based architectures which perform on-par with the state-of-the-art models for SRL but significantly better than them in low-resource domain-specific settings. We show significant improvement of approximately 24 F1 points over current best model for role-labeling on the ProcessBank - notably different in nature compared to CoNLL dataset.

[5] Please refer to the supplemental material 9 for a detailed discussion on results

References

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D Manning. 2014. Modeling biological processes for reading comprehension. In *EMNLP*, pages 1499–1510.

Steven Bethard, Zhiyong Lu, James H Martin, and Lawrence Hunter. 2008. Semantic role labeling for protein transport predicates. *BMC bioinformatics*, 9(1):277.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. pages 152–164. Association for Computational Linguistics.

Daniel Dahlmeier and Hwee Tou Ng. 2010. Domain adaptation for semantic role labeling in the biomedical domain. *Bioinformatics*, 26(8):1098–1104.

Hal Daume III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126.

Quynh Ngoc Thi Do, Steven Bethard, and Marie-Francine Moens. 2015. Domain adaptation in semantic role labeling using a neural language model and linguistic resources. *IEEE/ACM Trans. Audio, Speech & Language Processing*, pages 1812–1823.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, pages 249–256.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034.

Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2018. Jointly predicting predicates and arguments in neural semantic role labeling. *arXiv preprint arXiv:1805.04787*.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and whats next. In *ACL*, volume 1, pages 473–483.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Soojong Lim, Changki Lee, Pum-Mo Ryu, Hyunki Kim, Sang Kyu Park, and Dongyul Ra. 2014. Domain-adaptation technique for semantic role labeling with structural learning. *ETRI Journal*, 36(3):429–438.

Yang Liu and Mirella Lapata. 2017. Learning structured text representations. *CoRR*, abs/1705.09207.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *ACL*, volume 1, pages 1064–1074.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

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

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 *CoNLL*, pages 143–152.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287.

Parantu K Shah and Peer Bork. 2006. Lsat: learning about alternative transcripts in medline. *Bioinformatics*, 22(7):857–865.

Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *Advances in neural information processing systems*, pages 2377–2385.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2017. Deep semantic role labeling with self-attention. *arXiv preprint arXiv:1712.01586*.

Yu Zhang, Guoguo Chen, Dong Yu, Kaisheng Yaco, Sanjeev Khudanpur, and James Glass. 2016. Highway long short-term memory rnns for distant speech recognition. In *ICASSP*.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. In *ACL*, volume 1, pages 1127–1137.

8 Appendices

8.1 Parameter Settings

CNN-LSTM-CRF : The words that are absent in GloVe embeddings are replaced with <UNK> and intialized randomly. The character-embeddings are intialized with uniform samples as proposed in He et al. (2015). Weight matrices are initialized using Glorot initialization (Glorot and Bengio, 2010). Bias vectors are initialized to zero except the bias vector of Bi-LSTM (b_f) which is initialized to 1. Parameter optimization is performed using Adam optimizer with batch size of 32 and learning rate of 0.0001. We use a non-variational dropout of 0.5 on CNN and BiLSTM layers. We use a hidden size of 512, and use 5 layers for the BiLSTM. For character embeddings, we use a hidden size of 30. The CNN's use 30 filters.

DeepSRL-CRF : We maintain most of the experimental settings similar to He et al. (2017). We convert all tokens to lower-case, initialize with the embeddings. We use the Adadelta with $\epsilon = 1e^{-6}$ and $\rho = 0.95$ with mini-batch size 64. The dropout probability was set to 0.1 and gradient clipping at 1. The models are trained for 50 epochs (compared to 500 epochs in the original DeepSRL model) and use the best model from 50 epochs for pretraining. We do not add any constraints for decoding and we use the viterbi decoding to get our output tags.

9 Supplemental Material

9.1 Additional Discussion

DeepSRL-CRF: The DeepSRL-CRF model achieves comparable but slightly lower performance compared to the current state-of-the-art in the CoNLL-2005 and CoNLL-2012 SRL datasets. We observed that these performances did not directly translate to the ProcessBank dataset. In the limited-resource domain of ProcessBank, the final CRF layer had a more pronounced performance improvements. Adding CRF layer to DeepSRL model improves performance by atleast 4 F1 points when pre-trained using CoNLL-2005 and 1 F1 point when pre-training using CoNLL-2012 dataset. Adding ELMo embeddings to the DeepSRL and DeepSRL-CRF models did not result in performance gains in ProcessBank except for one experimental setup (DeepSRL-CRF pre-trained on CoNLL-2005). Across both

datasets, we acheived our best results when we incorporated event-event relations in the SRL annotation. Although a performance improvement is expected, the best results for domain adaptation was achieved after adding the event relations. The tags that gain most from the event relationships are *Agent*, *Destination*, *Source* and *Location*. The improvements primarily come from the gain in precision with a slight drop in recall. We believe that the reason for this improvement is the artifact of the dataset's event-event relationships tend to correlate often with these entities given the nature of these biological processes. Across CoNLL-2005 and CoNLL-2012, it did not make a considerable difference as to which dataset we used for pre-training. Although CoNLL-2012 has slighly better performance (shown in table 3, there could be additional hyper-parameter tuning that could lead to slightly different results between the two datasets.

CNN-LSTM-CRF: The CNN-LSTM-CRF model on ProcessBank achieves 40.62 F1 without any pre-training. This result is comparable to the baseline, showing the importance of initialization of weights while training a neural network based model. However, we achieve substantial improvement of about 21.7 F1 with pre-training on CoNLL data and later adapting only the final CRF layer for the target label space. In contrast to DeepSRL-CRF, we notice that performance difference between pre-training on CoNLL-2005 and CoNLL-2012 is considerable (4 F1 points). We have to note that CoNLL-2012 dataset has about 35000 more training data instances than CoNLL-2005. We hypothesize that these additional training instances might have contributed to the final F1 score while training using CoNLL-2012 dataset. We also observe that pre-training improves the performance of tags that have less number of instances in the target domain (ProcessBank). One of the unique cases is shown in table 7, where *Source* tag prediction shows huge improvements (57.0 F1) after the model was pre-trained using the CoNLL data. However, we do not see the same trend for the *Other* tag. Further, as per table 5 and 6, the model particularly confused the *Other* tag with the *O* tag of the BIO scheme. In the original ProcessBank dataset, the tags that do not belong to the original proposed categories, were classified as one single *Other* category and this category had the least number of annotated examples. We believe that the combination of these factors made it challenging for the model to predict this particular category. According to table 5 and 6, the most frequent tags – *Theme* and *Agent* have high prediction accuracy. However, their spans are sometimes incorrectly identified. For instance the *Theme* tags are identified incorrectly as *O* or vice-versa. Overall *B* tags have higher precision than the *I* tags, and the model is able to better predict the start of a span than the end of a span.

From table 7, we also notice that annotating a dataset with event-event relationships does not consistently improve the performance which we observed in DeepSRL-CRF. These results also show that adding the CNN-layer of character embeddings to the BiLSTM-CRF model helps the model perform better across all the labels. emphasizing the relevance of these character embeddings.

%	Agt.	Dest	Loc	Oth.	Res.	Src.	The.	O
Agt.	**71.1**	1.0	0.0	0.0	0.0	1.0	6.2	20.6
Dest.	0.0	**53.9**	7.7	0.0	7.7	0.0	15.4	15.4
Loc	0.0	3.0	**45.5**	0.0	3.0	0.0	9.1	39.4
Oth.	0.0	25.0	25.0	0.0	0.0	0.0	0.0	**50.0**
Res.	0.0	0.0	2.2	0.0	31.1	0.0	24.4	**42.2**
Src.	0.0	15.8	0.0	0.0	0.0	**68.4**	15.8	0.0
The.	4.0	1.6	0.0	0.0	1.2	0.8	**85.9**	6.5

Table 5: Best performing CNN-LSTM-CRF model's breakdown of true (rows) and predicted (columns) *B* tags with BIO tagging scheme. (Agt.=Agent; Dest.=Destination; Loc.=Location; Oth.=Other; Res.=Result; Src.=Source; The.=Theme; O=*O* tag in BIO tagging)

%	Agt.	Dest	Loc	Oth.	Res.	Src.	The.	O
Agt.	**65.6**	1.1	0.0	0.0	0.0	0.0	7.1	26.2
Dest.	0.0	**43.0**	15.8	0.0	5.3	0.0	16.7	19.3
Loc	0.0	9.2	**48.7**	0.0	5.3	0.0	6.6	30.3
Oth.	0.0	16.7	16.7	0.0	0.0	0.0	0.0	**66.7**
Res.	0.0	0.0	0.8	0.0	**43.0**	0.0	20.3	35.9
Src.	0.0	6.5	0.0	0.0	0.0	**71.0**	22.6	0.0
The.	3.2	2.7	3.4	0.0	1.7	1.2	**73.2**	14.6

Table 6: Best performing CNN-LSTM-CRF model's breakdown of true (rows) and predicted (columns) *I* tags with BIO tagging scheme. (Agt.=Agent; Dest.=Destination; Loc.=Location; Oth.=Other; Res.=Result; Src.=Source; The.=Theme; O=*O*tag in BIO tagging)

		BiLSTM-CRF		CNN-LSTM-CRF		
	#Instances	PB only	+Pretrain. +Dom. adp.	PB only	Pretrain. +Dom. adp.	+Verb Relations
Agent	280	25.8	37.0	35.5	62.1	**63.3**
Destination	153	8.0	2.7	38.5	51.3	**53.1**
Location	109	4.8	1.8	26.1	**44.1**	38.8
Other	11	0.0	0.0	0.0	0.0	0.0
Result	173	2.8	12.0	11.1	**34.7**	25.0
Source	50	8.7	15.4	0.0	57.9	**59.1**
Theme	838	44.9	57.3	52.1	**67.2**	66.0

Table 7: F1 scores on different tags in ProcessBank with BiLSTM-CRF and CNN-LSTM-CRF model (PB=ProcessBank). Pre-training was done on CoNLL-2012 dataset