Domain adaptation for semantic role labeling of clinical text

Yaoyun Zhang1,*, Buzhou Tang1,2,*, Min Jiang1, Jingqi Wang1, Hua Xu1

ABSTRACT

Objective Semantic role labeling (SRL), which extracts a shallow semantic relation representation from different surface textual forms of free text sentences, is important for understanding natural language. Few studies in SRL have been conducted in the medical domain, primarily due to lack of annotated clinical SRL corpora, which are time-consuming and costly to build. The goal of this study is to investigate domain adaptation techniques for clinical SRL leveraging resources built from newswire and biomedical literature to improve performance and save annotation costs.

Materials and Methods Multisource Integrated Platform for Answering Clinical Questions (MiPACQ), a manually annotated SRL clinical corpus, was used as the target domain dataset. PropBank and NomBank from newswire and BioProp from biomedical literature were used as source domain datasets. Three state-of-the-art domain adaptation algorithms were employed: instance pruning, transfer self-training, and feature augmentation. The SRL performance using different domain adaptation algorithms was evaluated by using 10-fold cross-validation on the MiPACQ corpus. Learning curves for the different methods were generated to assess the effect of sample size.

Results and Conclusion When all three source domain corpora were used, the feature augmentation algorithm achieved statistically significant higher F-measure (83.18%), compared to the baseline with MiPACQ dataset alone (F-measure, 81.53%), indicating that domain adaptation algorithms may improve SRL performance on clinical text. To achieve a comparable performance to the baseline method that used 90% of MiPACQ training samples, the feature augmentation algorithm required <50% of training samples in MiPACQ, demonstrating that annotation costs of clinical SRL can be reduced significantly by leveraging existing SRL resources from other domains.

Keywords: semantic role labeling, shallow semantic parsing, clinical natural language processing, domain adaptation, transfer learning

INTRODUCTION

Natural language processing (NLP) technologies are important for unlocking information embedded in narrative reports in electronic health record systems. Although various NLP systems have been developed to support a wide range of computerized medical applications, such as biosurveillance and clinical decision support, extracting semantically meaningful information from clinical text remains a challenge. Semantic role labeling (SRL) (also known as shallow semantic parsing), which extracts semantic relations between predicates and their arguments from different surface textual forms, is an important method for the extraction of semantic information. State-of-the-art SRL systems have been developed and applied to information extraction in open domains and various biomedical subdomains. However, very few SRL studies have been conducted in the clinical domain, probably due to the lack of large-scale annotated corpora. The creation of such clinical SRL corpora would be both time-consuming and expensive.

In this study, we approach SRL on clinical narratives as a domain adaptation problem. The goal is to adapt existing the SRL corpora of newswire text and biomedical literature to the clinical domain. By transferring knowledge from existing corpora in other domains to the clinical domain, we aim to improve the performance of clinical SRL and reduce the cost of developing one de novo. We used three existing SRL corpora outside the clinical domain and evaluated three state-of-the-art domain adaptation algorithms on the task of SRL for clinical text. Our results showed that domain adaptation strategies were effective for improving the performance or reducing the annotation cost of SRL on clinical text. To the best of our knowledge, this is the first work that has introduced domain adaptation algorithms for clinical SRL.

BACKGROUND

The task of SRL is to label semantic relations in a sentence as predicate argument structures (PASs) to represent propositions. The definition of PAS originated from the predicate logic for proposition representation in semantics theory. There is a large body of work on extracting semantic relations in biomedical text. Many are based on the sublanguage theory by Harris, which describes the properties of language in closed domains. Typically, in a closed domain such as medicine, there are a limited number of primary semantic types and a set of constraints that can determine how different semantic types of the arguments can be linked to form semantic predications. Linguistic String Project (LSP) and Medical Language Extraction and Encoding System (MedLEE), which use sublanguage grammar, are two early NLP systems for the extraction of semantic relations in the medical domain. SemRep is another biomedical semantic relation extraction system, which extracts semantic predications defined in the Unified Medical Language System Semantic Network from biomedical literature. Recently, Cohen et al. examined the syntactic alternations in the argument structure of domain-specific verbs and associated nominalizations in the PennBioIE corpus, and found that even in a semantically restricted domain, syntactic variations are common and diverse. Currently, many sublanguage-based clinical NLP systems often recognize semantic relations by manually extracted patterns using rule-based methods. SRL, however, focuses on unifying variations in the surface syntactic forms of semantic relations based on annotated corpora. It is inspired by previous research into semantic frames and the link between semantic roles and syntactic realization. Although current SRL approaches are primarily developed in open domains (thus, types of semantic roles or...
arguments may not be sufficient or appropriate for the medical domain, they can be refined and extended to the medical domain, thus providing alternative or complementary approaches for clinical semantic relation extraction.

In SRL, a predicate usually refers to a word indicating a relation or an attribute, and arguments refer to syntactic constituents acting as different semantic roles to the predicate. The core arguments represent the essential arguments of a predicate, whereas adjunctive arguments express general properties of a predicate such as time and location. For example, in Figure 1, in the sentence “We resected the left carotid stenosis 5 months ago,” “resected” is the predicate, “We” comprises the role of agent (initiator/executor in the relation), indicating the surgeon performing the action “resect”; “the left carotid stenosis” comprises the role of patient (receiver in the relation), indicating the entity removed by the action “resect,” and “5 months ago” is the time. The roles of agent and patient are the core arguments and time is the adjunctive argument of the predicate.

Automatic SRL was first introduced by Gildea and Jurafsky in 2002. Since then, SRL has attracted attention owing to its usefulness for multiple NLP tasks and applications, such as information extraction and question answering. With public availability of large-scale annotated corpora such as PropBank and promotion by CoNLL shared tasks, many data-driven approaches have been developed for SRL in open domains such as newswire. This approach is standardized and divides SRL into several successive tasks. Argument identification (AI) finds all syntactic constituents with semantic roles, that is, arguments in the sentence. Argument classification (AC) determines the specific semantic role of an argument. Global inference refines the output of preceding tasks using global constraints. These tasks can be conquered individually or as a combined task.

Other approaches include resolving syntactic parsing and SRL jointly by integrating them into a single model, or by using Markov Logic Networks as the learning framework. In the last few years, efforts have focused on building SRL systems for biomedical literature. Wattarajujeet al. developed PASBio, which analyzes and annotates the PASs of over 30 verbs for molecular events. Kogan et al. annotated PAS in medical case reports. The LSAT system, developed by Shah and Bork, used SRL to identify information about gene transcripts. Bethard et al. extracted information about protein movement by using an SRL system, in which 34 verbs and four semantic roles focusing specifically on protein movement were defined and annotated in their corpus. Barnickel et al. presented a neural network–based SRL system for relation extraction with emphasis on improving processing speed. Paek et al. semantically analyzed abstracts from randomized trials with SRL; however, the predicate set only contained five verbs. The BIOSMILE system by Tsi et al. was built on the BioProp corpus, in which PASs of 30 verbs were annotated following the PropBank guideline. Interestingly, their results showed that the SRL system trained on PropBank alone did not perform well on BioProp.

All the previously described SRL systems were built on annotated corpora of biomedical literature, facilitating literature-based information extraction applications. However, very few studies have been conducted to investigate SRL techniques for clinical text from electronic health records. For example, Wang et al. analyzed and annotated PASs of 30 predicates in operative notes following the PropBank style, but they did not build an automatic SRL system. The first clinical SRL system was reported by Albright et al., who created an annotated corpus, Multisource Integrated Platform for Answering Clinical Questions (MiPACQ), containing multiple syntactic and semantic layers including SRL information. The SRL dataset in MiPACQ contains 1137 predicates and the SRL performance on that dataset was 79.91% by adopting an existing SRL method developed for newswire. The primary limitation for clinical SRL research is apparently the lack of annotated SRL corpora in the medical domain. It is time-consuming and expensive to create large annotated clinical corpora, because this often requires manual review by domain experts such as physicians. Moreover, medicine consists of different subdomains (e.g., internal medicine, pathology, and radiology) and the languages of different subdomains can be distinct, which makes it challenging to transfer machine learning–based models from one subdomain to another. For example, the MiPACQ dataset was a collection of Mayo Clinic’s clinical notes (CNs) and pathology notes regarding colon cancer. When the SRL model built from MiPACQ was tested on two other CN datasets of different genres and note styles, namely radiology notes from the Strategic Health IT Advanced Research Projects and colon cancer clinical and pathology notes from Temporal Histories of Your Medical Events, the performance dropped significantly. Therefore, to construct high-performance SRL systems for each subdomain, we may have to create annotated corpora for every specific subdomain, which would require substantial effort and resources.

To address this limitation in clinical SRL development, we propose an investigation of domain adaptation techniques for SRL. The task of domain adaption is to adapt a classifier that is trained on a source domain to a new target domain. This improves performance and reduces dependency of the classifier on large annotated
datasets in the target domain.\textsuperscript{41,42} Transfer learning algorithms are often employed for domain adaptation.\textsuperscript{43} Given a source domain, $D_s$, and its learning task, $T_s$, a target domain, $D_t$, and its learning task, $T_t$, transfer learning aims to improve the learning of the target predictive model in $D_t$ by using the knowledge from $D_s$ and $T_s$.\textsuperscript{28} Most commonly used transfer learning algorithms can be categorized into instance-level and feature-level approaches. Let $L_s$ denote the labeled dataset of $D_s$, $L_t$ denote the labeled dataset of $D_t$, and $U_s$ denote the unlabeled dataset of $D_s$. Here, instance-level transfer learning algorithms aim to select or weight instances in $L_s$ for use in the target domain.\textsuperscript{43} This does not depend on the machine learning algorithm used for building classifiers, but requires a small gap between the source and target domains. Feature-level transfer learning algorithms aim to find a new feature representation that reduces the difference between features in $D_s$ and $D_t$ and highlights the similarity between them. This has a moderate dependency on machine learning algorithms, but is more tolerant of domain gap. Transfer learning algorithms have been effective in solving the problem of data scarcity in $D_t$ for several key bioinformatics areas, such as sequence classification and gene expression data analysis.\textsuperscript{44} Dahlmeier and Ng\textsuperscript{45} addressed SRL on BioProp by using domain adaptation algorithms for the first time, with PropBank as the source domain dataset. Their results demonstrated that the cost of developing an SRL system for interpreting molecular events could be significantly reduced. Ferraro et al.\textsuperscript{46} also showed improved performance of POS tagging on clinical narratives using the feature-level transfer learning algorithm Easy Adapt. More recently, Laippala et al.\textsuperscript{47} investigated the use of “source only,” “target only,” and “source + target” in statistical parsing of Finnish; however, no domain adaptation algorithm was employed in their work.

In this study, we explored both instance-level and feature-level domain adaptation algorithms for SRL on clinical narratives. We used PropBank, NomBank, and BioProp as source domain datasets and MiPACQ as the target domain dataset. Dahlmeier and Ng\textsuperscript{45} previously conducted a domain adaptation study on biomedical literature using BioProp as the target dataset and PropBank as the source dataset, and they obtained promising results on molecular event interpretation. Our study design is similar to the work of Dahlmeier and Ng.\textsuperscript{45} However, we focus on clinical text instead of biomedical literature on molecular events. Previous studies have shown that clinical reports and biomedical literature are two very different sublanguages in terms of semantic relation types and complexity.\textsuperscript{25} Their differences were also demonstrated in various other NLP tasks, such as word sense disambiguation and medical term identification.\textsuperscript{48,49} Furthermore, BioProp only contains semantic roles of 30 verb predicates with 1962 PASs; MiPACQ has 722 verb predicates and 415 nominal predicates with 12 575 PASs. Therefore, it is important to assess domain adaptation methods on the larger clinical SRL corpus, in addition to biomedical literature. Moreover, we investigated the effect of additional external corpora (i.e., BioProp and NomBank) and their combinations on clinical SRL, which was not reported in Dahlmeier and Ng’s work. To the best of our knowledge, this is the first work that has introduced domain adaptation algorithms for clinical SRL. Our evaluation showed that domain adaptation algorithms can improve performance or reduce annotation costs for clinical SRL.

**METHODS**

**Datasets**

We used four annotated SRL datasets in our study, three as source domain datasets and one as the target dataset, as described in the following.

**Source domain datasets**

The PropBank corpus\textsuperscript{18} is the most widely used corpus for developing SRL systems.\textsuperscript{50} The corpus is built from news articles of the *Wall Street Journal* and is available through the Linguistic Data Consortium (http://www.ldc.upenn.edu). Semantic roles of verb predicates are annotated in this corpus. The PropBank corpus has 25 sections, denoted as sections 00–24. We used the standard training set of sections 2–21 as the source domain dataset.

The NomBank corpus\textsuperscript{16} contains annotated semantic roles of nominal predicates. Similar to PropBank, it is built from news articles of the *Wall Street Journal*, based on the Penn TreeBank. It is available online at http://nlp.cs.nyu.edu/meyers/NomBank.html. Following PropBank, the standard training set of sections 2–21 was used as the source domain dataset.\textsuperscript{50}

The BioProp corpus\textsuperscript{17} is annotated based on the GENIA Treebank.\textsuperscript{51} The GENIA Treebank facilitates information extraction from biomedical literatures about proteins. It is available for downloading from the GENIA project web site (http://www.nactem.ac.uk/genia/genia-corpus/treebank). Specifically, BioProp was created from 500 MEDLINE abstracts. The articles were selected based on the keywords “human,” “blood cells,” and “transcription factor.” Semantic roles of verb predicates are annotated in BioProp.

**Target domain dataset**

MiPACQ is built from randomly selected CNs and pathology notes of Mayo Clinic related to colon cancer.\textsuperscript{13} Annotations of layered linguistic information including part of speech tagging, PAS for SRL, named entities, and semantic information from Unified Medical Language Systems are available for building NLP components. The predicate–argument semantic annotations follow PropBank guidelines. Both verb and nominal predicates are annotated in MiPACQ.

Table 1 displays the statistics of the four corpora. MiPACQ contains 722 verb predicates with 9780 PAS and 415 nominal predicates with 2795 PASs. PropBank and BioProp (PB) have only verb predicates; NomBank has only nominal predicates. Among the 722 verb predicates in MiPACQ, 644 are common with PropBank and 15 are common with BioProp. Among the 415 nominal predicates, 265 are common with NomBank. As displayed in Table 1, among the three source domain datasets, the sizes of PropBank/NomBank are significantly larger than BioProp (~50 times more PASs). Moreover, the size of BioProp is much smaller than MiPACQ, with a ratio of 1:38 for predicates.

**Domain Adaptation Algorithms**

The three transfer-learning algorithms employed in this study are described in detail in the following.

**Instance pruning.** Instance pruning (InstancePrune) trains a classifier on $L_s$ and uses this classifier to predict class labels for $L_t$.\textsuperscript{52} The top $p$ instances that are predicted wrongly, ranked by prediction confidence, are removed from $L_s$. The intuition here is that instances that are very different from the target domain will affect the prediction ability of the classifier model. The remaining instances, $L_s'$, are added to $L_t$ as training data.

**Transfer self-training.** Transfer self-training (TransferSelf) borrows the idea of self-training from the framework of semisupervised learning into transfer learning.\textsuperscript{53} It iteratively trains the classifier by transferring a subset of $L_s'$ with high similarity to instances in $D_t$ from $L_s$ to enrich $L_t$ as the training data.\textsuperscript{54} First, a classifier trained using $L_t$
is applied to $L_S$. For each category in $L_S$, the top $n$ correctly classified instances ranked by prediction confidence are selected and added into $L_T$ as the training data. The classifier is then retrained on the enriched training data and applied to $L_S - L_T$ again to select more instances. The iteration terminates when the prediction confidence of instance in $L_S - L_T$ is less than a specified threshold or the maximum allowed iterations is exceeded. The final classifier is obtained by training on the combination of $L_T$ and $L_S$.

Both InstancePrune and TransferSelf are instance-level transfer learning algorithms. The differences between them include: (1) InstancePrune attempts to remove wrongly predicted source instances of high confidence; TransferSelf selects correctly predicted source instances of high confidence into the training set; and (2) the source instance selection in InstancePrune is conducted only once; TransferSelf adds source instances iteratively into the training set, thus leveraging selected source instances in previous iterations. In Dahlmeier and Ng,45 another instance level transfer learning algorithm named instance weighting52 was employed for domain adaptation. However, the experimental results were not promising.45 In our pilot study, we tried another instance weighting algorithm called TrAdaBoost55 and it did not perform well either. Hence, we did not employ instance weighting in our study.

Feature augmentation. For feature augmentation (FeatureAug), Daume´ III56 proposed a domain adaptation algorithm that maps feature vectors into a higher dimension. This algorithm is also called Easy Adapt, because it can be implemented simply with a few lines of Perl script.56 Denote $X_s$ and $X_t$ as the original feature vectors for source and target domain, respectively, then mapping is conducted as follows:

$$
X_s = \langle X_s, 0 \rangle
$$
$$
X_t = \langle X_t, 0 \rangle
$$

where $0$ is a zero vector of length $|X|$. By this transformation, the feature spaces of both $D_S$ and $D_T$ are augmented. Three versions of features are generated from each original feature vector, namely, “general,” “source-specific,” and “target-specific” versions. The intuition of this algorithm is to leverage the aggregation of the three feature space versions to learn an efficient feature representation for $D_T$. Common features between $D_S$ and $D_T$ are assigned with higher weights in instances of both domains; whereas features unique to $D_S$ or $D_T$ are assigned with higher weights only in instances of $D_S$ or $D_T$. A standard machine-learning algorithm will assign weights differently to features in each version. Effective features for $D_T$ will be emphasized from the general and target-specific versions.

Experiments

**System description**

Figure 2 shows the study design of the domain adaptation–based SRL. The SRL system can be viewed as consisting of the training stage and the testing stage. In the training stage, SRL is split into two subtasks: the AI and AC subtasks. First, a binary nonargument versus argument classifier is built as the argument identifier on the entire dataset for all predicates, instead of building one model per predicate. For AC, a multiclass classifier is built to assign semantic roles to arguments of all predicates. In the testing stage, for each predicate, the argument candidates first pass through the argument identifier. If one candidate is identified as an argument, it will go through the argument classifier that assigns the semantic role.
Features and machine-learning algorithm
Similar to the work of Tsai et al.\textsuperscript{17} and Dahlmeier and Ng,\textsuperscript{45} we adopted common features used in current state-of-the-art SRL systems. These features include seven baseline features from the original work of Gildea and Jurafsky\textsuperscript{31} and additional features taken from Pradhan et al.\textsuperscript{57} All were extracted from the syntactic parse tree, and are shown in Figure 3. The “Voice” feature is not used for nominal predicates.

We used the open source toolkit Liblinear\textsuperscript{58} for implementations of machine-learning algorithms. The logistic regression algorithm was applied to select confidence-based source domain training instances for InstancePrune and TransferSelf, because it outputs predictions with their probabilities, which can serve as the prediction confidence.\textsuperscript{52} The linear support vector machine algorithm was used to build SRL models owing to its high generalization ability for new data.\textsuperscript{59}

Experimental setup
Following the work of Dahlmeier and Ng,\textsuperscript{45} we used the gold-standard parsing annotations of PropBank, NomBank, BioProp, and MiPACQ in our SRL experiments. In addition to separately using each source domain dataset, the combination of PropBank and NomBank (PN), PB, and all three source datasets (PNB) are used as $D_s$ in our experiments to examine the influence of multiple sources on domain adaptation. Only the PASs with at least one argument were used. For each implemented method, all parameters were tuned for optimal performance.

Experiments and systematic analysis were conducted as discussed in the following.

1. Algorithms for domain adaptation: InstancePrune, TransferSelf, and FeatureAug were employed in this study, as described in the “Methods” section. To examine the effectiveness of these algorithms, three baseline methods were also developed for comparison: the “Source Only” method uses only $D_s$ to train a classifier; the “Target Only” method uses only $D_t$ to train a classifier; the “Source & Target” method directly combines both $D_s$ and $D_t$ to train a classifier.

2. Influence of sample size on domain adaptation: To determine the effect of sample size on SRL performance, classifiers were also trained using varying sample sizes of $D_s$ and $D_t$. We examined the performance of FeatureAug as the representative of the three domain adaptation algorithms, and used combinations of three sources as the $D_s$ because FeatureAug with combined sources showed optimal SRL performance in our study.

3. Domain adaptation for different predicate types: As described in the “Datasets” section, MiPACQ contains both verb and nominal predicates. PB have verb predicates only and NomBank has nominal predicates only. Among the 722 verb predicates in MiPACQ, 644 are common with PropBank and 15 are common with BioProp. Among the 415 nominal predicates, 265 are common with NomBank. The effects of domain adaptation on the performance of the common/uncommon predicates between $D_s$ and $D_t$, as well as the performance of the verb/nominal predicates were examined.

Evaluation
Precision ($P$), recall ($R$), and $F_1$ measure were used as evaluation metrics for AI and combined SRL tasks. Precision measures the percentage of correct predictions of positive labels made by a classifier. Recall measures the percentage of positive labels in the gold standard that were correctly predicted by the classifier. $F_1$ measure is the harmonic mean of precision and recall. During the process of AC, the
Boundaries of candidate arguments are already identified by the AI step. Therefore, the accuracy of the classifier was used for evaluation, which is defined as the percentage of correct predictions with reference to the total number of candidate arguments correctly recognized in the AI step.

Ten-fold cross-validation was employed for performance evaluation. Nine folds of MiPACQ were merged with DS as the training set and one fold was used for testing. In experiments evaluating the influence of source domain sample size, nine folds of MiPACQ were merged with an increasing percentage of PropBank to generate the training set. In experiments evaluating the effect of target domain sample size, an increasing percentage of the nine-fold MiPACQ was added to the entire PropBank as the training set.

RESULTS
Table 2 lists the results of the implemented methods using both individual and combined source domain corpora as DS.

<table>
<thead>
<tr>
<th>Data</th>
<th>Methods</th>
<th>AI</th>
<th>AC</th>
<th>AI + AC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIPACQ</td>
<td>Target Only</td>
<td>93.24</td>
<td>92.81</td>
<td>93.02</td>
</tr>
<tr>
<td>PropBank + MIPACQ</td>
<td>Source Only</td>
<td>86.54</td>
<td>76.36</td>
<td>81.13</td>
</tr>
<tr>
<td></td>
<td>Source &amp; Target</td>
<td>93.99</td>
<td>90.36</td>
<td>92.14</td>
</tr>
<tr>
<td></td>
<td>InstancePrune</td>
<td>94.22</td>
<td>92.70</td>
<td>93.45</td>
</tr>
<tr>
<td></td>
<td>TransferSelf</td>
<td>93.23</td>
<td>92.77</td>
<td>93.00</td>
</tr>
<tr>
<td></td>
<td>FeatureAug</td>
<td>94.08</td>
<td>93.82</td>
<td>93.95</td>
</tr>
<tr>
<td>NomBank + MIPACQ</td>
<td>Source Only</td>
<td>84.20</td>
<td>11.03</td>
<td>19.50</td>
</tr>
<tr>
<td></td>
<td>Source &amp; Target</td>
<td>95.59</td>
<td>85.46</td>
<td>90.24</td>
</tr>
<tr>
<td></td>
<td>InstancePrune</td>
<td>94.53</td>
<td>90.28</td>
<td>92.35</td>
</tr>
<tr>
<td></td>
<td>TransferSelf</td>
<td>93.21</td>
<td>92.49</td>
<td>92.65</td>
</tr>
<tr>
<td></td>
<td>FeatureAug</td>
<td>93.31</td>
<td>92.70</td>
<td>93.00</td>
</tr>
<tr>
<td>BioProp + MIPACQ</td>
<td>Source Only</td>
<td>53.48</td>
<td>30.62</td>
<td>38.95</td>
</tr>
<tr>
<td></td>
<td>Source &amp; Target</td>
<td>93.43</td>
<td>92.07</td>
<td>92.74</td>
</tr>
<tr>
<td></td>
<td>InstancePrune</td>
<td>93.43</td>
<td>92.68</td>
<td>93.05</td>
</tr>
<tr>
<td></td>
<td>TransferSelf</td>
<td>93.41</td>
<td>92.77</td>
<td>93.09</td>
</tr>
<tr>
<td></td>
<td>FeatureAug</td>
<td>93.13</td>
<td>92.83</td>
<td>92.98</td>
</tr>
<tr>
<td>PN + MIPACQ</td>
<td>Source Only</td>
<td>91.04</td>
<td>78.42</td>
<td>84.26</td>
</tr>
<tr>
<td></td>
<td>Source &amp; Target</td>
<td>95.37</td>
<td>90.39</td>
<td>92.62</td>
</tr>
<tr>
<td></td>
<td>InstancePrune</td>
<td>95.13</td>
<td>92.68</td>
<td>93.89</td>
</tr>
<tr>
<td></td>
<td>TransferSelf</td>
<td>93.15</td>
<td>92.42</td>
<td>92.78</td>
</tr>
<tr>
<td></td>
<td>FeatureAug</td>
<td>94.50</td>
<td>93.89</td>
<td>94.19</td>
</tr>
<tr>
<td>PB + MIPACQ</td>
<td>Source Only</td>
<td>90.33</td>
<td>75.68</td>
<td>82.36</td>
</tr>
<tr>
<td></td>
<td>Source &amp; Target</td>
<td>94.38</td>
<td>89.44</td>
<td>91.84</td>
</tr>
<tr>
<td></td>
<td>InstancePrune</td>
<td>93.75</td>
<td>91.71</td>
<td>92.72</td>
</tr>
<tr>
<td></td>
<td>TransferSelf</td>
<td>92.37</td>
<td>91.23</td>
<td>91.80</td>
</tr>
<tr>
<td></td>
<td>FeatureAug</td>
<td>94.06</td>
<td>93.54</td>
<td>93.80</td>
</tr>
<tr>
<td>PNB + MIPACQ</td>
<td>Source Only</td>
<td>91.30</td>
<td>78.25</td>
<td>84.27</td>
</tr>
<tr>
<td></td>
<td>Source &amp; Target</td>
<td>95.37</td>
<td>90.27</td>
<td>92.75</td>
</tr>
<tr>
<td></td>
<td>InstancePrune</td>
<td>94.70</td>
<td>92.57</td>
<td>93.62</td>
</tr>
<tr>
<td></td>
<td>TransferSelf</td>
<td>93.20</td>
<td>92.42</td>
<td>92.81</td>
</tr>
<tr>
<td></td>
<td>FeatureAug</td>
<td>94.43</td>
<td>93.85</td>
<td>94.14</td>
</tr>
</tbody>
</table>

*Statistically significant with p-value < 0.05 by the Wilcoxon signed-rank test.
FeatureAug with PropBank achieved the highest combined performance compared to the Target Only baseline. Among the three, however, the use of domain adaptation algorithms increased the performance of ARG0, ARG1, ARG2, ARG3-5, ARGM-ADJ, ARGM-ADV, ARGM-LOC, ARGM-MNR, ARGM-MOD, ARGM-NEG, ARGM-TMP, and ARGM-Others by 83.18% when FeatureAug algorithm was applied to the MiPACQ samples used for training when PropBank, NomBank, and BioProp were combined as DS. The Source Only baseline is a horizontal line. For other methods, increasing the sample size of the target domain (MiPACQ) leads to a consistent performance enhancement. However, the domain adaptation method (FeatureAug) clearly shows better performance than baselines of Target Only and Source & Target. Similarly, Figure 5 shows the learning curves obtained by increasing the sample size of the PNB sample size for training. Without domain adaptation, increasing the sample size of DS progressively decreased the performance. Nevertheless, a monotone increasing curve is clear when augmented with the domain adaptation algorithm (FeatureAug).

Table 4 displays the SRL performance of optimized domain adaptation methods for each source, for overlapping versus nonoverlapping predicates or verb versus nominal predicates, respectively. The scores listed in parentheses in the last column of Table 4 indicate the improvements in F-measures between the baseline (Target Only) and domain adaptation methods, which are calculated in the same way as those in Table 3. As illustrated in Table 4, the employed domain adaptation algorithms improved SRL on not only overlapping predicates (PropBank 1.48%, NomBank 1.66%, and BioProp 0.83%), but also nonoverlapping predicates (PropBank 1.14%, NomBank 1.44%, and BioProp 0.47%). Our results also suggested that the employed domain adaptation algorithms improved SRL performance on not only verb predicates, but also nominal predicates (Table 4). For example, when

\[
\text{F}_{\text{DomainAdapt}} = \frac{F(g, p)_{\text{DomainAdapt}} - F(g, p)_{\text{MiPACQ}}}{F(g, p)_{\text{MiPACQ}}}
\]

For example, FeatureAug with PropBank increases the F1 of ARG0 by 1.20% over using MiPACQ only. As shown in the table, FeatureAug with PropBank increased the performance of each argument. For NomBank, FeatureAug decreased the performance of ARG0, ARG3-5, ARGM-MOD, and ARGM-NEG, but increased the performance of the other arguments. For BioProp, InstancePrune increased performance on most arguments, but decreased the performance of ARG0, ARG3-5, ARGM-MNR, and ARGM-MOD.

Table 3 lists the performance of core arguments and adjunctive arguments with the highest frequencies. The remaining adjunctive arguments are listed in “ARGM-Others.” The scores listed in the parenthesis stand for the F1 value improvement by domain adaptation, which measures the extent of increase over using the target domain dataset only. Denoting the F1 values of using MiPACQ only and after domain adaptation as \( F_{\text{MiPACQ}} \) and \( F_{\text{DomainAdapt}} \), respectively, the F1 value improvement by domain adaptation is calculated by

\[
\text{Table 3: Combined SRL Performance for each Argument using MiPACQ only, FeatureAug with PropBank, FeatureAug with NomBank, and InstancePrune with BioProp (%)}
\]

<table>
<thead>
<tr>
<th>Argument</th>
<th>MiPACQ</th>
<th>PropBank_FeatureAug</th>
<th>NomBank_FeatureAug</th>
<th>BioProp_InstancePrune</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
</tr>
<tr>
<td>ARG0</td>
<td>87.55</td>
<td>87.32</td>
<td>87.44</td>
<td>88.25</td>
</tr>
<tr>
<td>ARG1</td>
<td>84.06</td>
<td>87.48</td>
<td>85.74</td>
<td>84.82</td>
</tr>
<tr>
<td>ARG2</td>
<td>79.28</td>
<td>75.80</td>
<td>77.50</td>
<td>80.42</td>
</tr>
<tr>
<td>ARG3-5</td>
<td>77.07</td>
<td>53.02</td>
<td>62.82</td>
<td>80.00</td>
</tr>
<tr>
<td>ARGM-ADJ</td>
<td>55.19</td>
<td>57.32</td>
<td>56.23</td>
<td>55.77</td>
</tr>
<tr>
<td>ARGM-ADV</td>
<td>66.15</td>
<td>69.03</td>
<td>67.56</td>
<td>68.69</td>
</tr>
<tr>
<td>ARGM-LOC</td>
<td>77.22</td>
<td>78.49</td>
<td>77.85</td>
<td>78.89</td>
</tr>
<tr>
<td>ARGM-MNR</td>
<td>74.51</td>
<td>69.30</td>
<td>71.81</td>
<td>75.95</td>
</tr>
<tr>
<td>ARGM-MOD</td>
<td>98.87</td>
<td>86.72</td>
<td>92.40</td>
<td>99.27</td>
</tr>
<tr>
<td>ARGM-NEG</td>
<td>91.07</td>
<td>89.47</td>
<td>90.27</td>
<td>95.51</td>
</tr>
<tr>
<td>ARGM-TMP</td>
<td>88.52</td>
<td>87.63</td>
<td>88.07</td>
<td>90.22</td>
</tr>
<tr>
<td>ARGM-Others</td>
<td>63.83</td>
<td>55.64</td>
<td>59.45</td>
<td>66.20</td>
</tr>
</tbody>
</table>
FeatureAug was used for PropBank (containing verb predicates only), the $F_1$ value of nominal predicates in MiPACQ was also improved by 1.16%.

**DISCUSSION**

In this study, we leverage existing annotation corpora from newswire and biomedical literature to improve the performance of clinical SRL by using domain adaptation algorithms. Our results showed that domain adaptation algorithms such as FeatureAug could improve the SRL task in the clinical domain by utilizing existing open domain corpora such as PropBank. In addition, we demonstrated that combining multiple sources from external domains further improved clinical SRL systems. To the best of our knowledge, this is the first study that has compared different domain adaptation algorithms for SRL in the medical domain.
The performance enhancement and cost reduction by domain adaptation for the clinical domain SRL task are illustrated more explicitly in Figure 4. It is not surprising that increasing the MiPACQ dataset for training constantly enhances the performance. An F₁ value of 81.67% was be achieved using 50% of MiPACQ dataset with domain adaptation (FeatureAug on the PNB corpora). This was higher than the baseline method (81.53%), which used 90% MiPACQ dataset for training (via 10-fold cross-validation), indicating a 40% reduction in annotation cost.

The SRL performance of the Source Only baseline indicates the direct influence of source domain datasets on MiPACQ without any domain adaptation algorithm. As illustrated in Table 2, the Source Only performance of all three source domain datasets was much lower than the Target Only performance. Specifically, the recall of Source Only with NomBank, which shared 265 nominal predicates with MiPACQ, was extremely low, with only 11.03% for AL and 6.65% for the combined task. On the other hand, BioProp achieved better recalls than NomBank using the Source Only method, with recalls of 30.62 and 16.25% for AL and the combined task, respectively. A possible explanation for the low recall with NomBank is that the high overlap of nominal predicates in NomBank compared to MiPACQ only contains 2795 (22.22%) PASs of noun predicates, in contrast to 9780 (77.78%) PASs of verb predicates.

We noticed that when a single source domain dataset is employed, domain adaptation algorithms performed differently for different source domain datasets. For example, although the dataset size of BioProp is much smaller than PropBank/NomBank, it achieved better performance than PropBank/NomBank using InstancePrune (Combined task F₁: BioProp 81.95%, PropBank 81.26%, NomBank 80.52%); whereas PropBank/NomBank outperformed BioProp using the FeatureAug algorithm (Combined task F₁: BioProp 81.63%, PropBank 82.83%, NomBank 81.85%). This is probably attributable to the essential difference among the D₂, PropBank/NomBank are built from the general English domain, while BioProp is built from biomedical literature. Based on our manual analysis of PASs in BioProp and MiPACQ, instance-level domain adaptation algorithms work better for BioProp, probably because the instances of BioProp selected by the InstancePrune algorithm have similar PASs, with a portion of instances in the target domain (MiPACQ). Feature-level algorithms have a better tolerance for domain gap, which is probably why FeatureAug worked best for PN. In addition, PropBank and NomBank are much larger and have higher dimensions of features, which provide the potential to contribute more useful features, as determined by the feature weighting mechanism in FeatureAug. The impact of BioProp seems trivial (PNB + MiPACQ versus PN + MiPACQ) or even negative (PB + MiPACQ versus PropBank + MiPACQ) when multiple source domain datasets are combined directly for domain adaptation. It is necessary to further investigate how to select source data instances efficiently when multiple source domain datasets are used. These findings may provide valuable insights for selecting source domain datasets and domain adaptation algorithms.

To identify which knowledge was transferred from the source domains to improve the performance, we examined the SRL results of individual instances. We found that most of the improvement was obtained by syntactic structural information learned from the source domains. As illustrated in Figure 6(a), due to the complex syntactic structure, the MiPACQ-only baseline failed to recognize “Because her dementia is progressive and, therefore, a terminal illness” as the ARGM-CAU (cause) argument of the predicate “favor” in the argument recognition stage. This argument was recognized correctly with the
effort of domain adaptation. Taking the sentence in Figure 6(b) as another example, the prepositional phrase “with his home psychiatrist” should be a core argument ARG1 (entity contacted) of the nominal predicate “contact.” However, the MiPACQ-only baseline labeled its role as an adjunctive argument, “ARGM-ADJ” (adjective). Although the PropBank dataset only annotated PAS for verb predicates, the similar syntactic structure for the verb “contact” still transferred successfully and correctly labeled the semantic role as ARG1. This explains the reason why the performance of nominal predicates in MiPACQ was also improved by PropBank.

Although knowledge transferred from source domains can be adapted to the clinical domain, the unique characteristics of clinical text require domain specific resources and solutions for further SRL improvement. One type of salient attribute is the clinical lexicons and semantic relations between them. In the phrase “an advanced oropharynx cancer treated with radiation therapy and chemotherapy,” “an advanced oropharynx cancer” is annotated as ARG2 (illness or injury) in the gold standard. However, it is labeled as ARG1 by our SRL system. Additionally, in the phrase “erectile dysfunction,” “erectile” is annotated as ARG1 (job, project) of “dysfunction,” which is mistakenly labeled as ARGM-ADJ (adjective). Clinical domain knowledge needs to be employed to precisely interpret these semantic relations. Another unique characteristic of clinical text is the high frequency of fragments; that is, grammatically incomplete sentences. Figure 6(c) illustrates the syntactic parsing result of the fragment sentence “Sigmoid, mass at 22 cm, endoscopic biopsy (AE46-395890; 09/06/65): Invasive, grade 2 of 4 adenocarcinoma identified.” In this sentence, there is no semantic relation between “biopsy” and the temporal phrase “09/06/65,” which is difficult to identify even when using the current open-domain state-of-the-art features. Clinical domain features like specific syntactic patterns of CNs need to be further explored.

State-of-the-art SRL systems usually employ a rich feature set and/or a global inference phase to further refine the output with global constraints. To verify the effects of global inference and domain adaptation, we developed a rule-based global inference module following important constraints defined in Punyakanok et al. The global inference phase improved SRL performance: the baseline Target Only method was improved from 81.53% to 82.12%. We then integrated the best domain adaption method FeatureAug with PNB with the improved SRL system (with global inference). Our results showed that FeatureAug with PNB further improved the SRL performance, with a combined $F_1$ of 83.88% (compared with the improved baseline of 82.12%). This indicates that domain adaptation and global inference...
are complementary. For optimized SRL systems with a global inference phase, domain adaptation methods may further improve the performance. The performance of integrating the global inference constraints with each of the methods implemented in our study can be found in Supplementary Appendix Table S1. Common features used in current state-of-the-art open domain SRL systems were adopted in our study.\textsuperscript{3,57} The previous work of domain adaptation on the BioProp dataset (Dahlmeier and Ng, 2010)\textsuperscript{45} also employed a similar feature set to our work. In this study, we built our SRL system by following the study in Dahlmeier and Ng.\textsuperscript{45} The current performance of our SRL system on PropBank is not state-of-the-art. Using the same datasets as in Punyakanok et al.,\textsuperscript{51} our SRL system achieved an $F_1$ of 82.77%, which is lower than the state-of-the-art performance of 86.81% $F_1$ in Punyakanok et al.\textsuperscript{51},\textsuperscript{67} probably due to the different feature sets, machine learning algorithms, and global constraints used in the study. We can further optimize the SRL performance by feature engineering\textsuperscript{57} in our future work. It is notable that domain adaptation made another contribution to significantly reduce the data annotation cost of the target dataset to achieve a comparable performance. As illustrated in Figure 4, it required $<50\%$ of training samples in the target dataset to achieve a comparable performance to the target-only baseline using 90\% of the target dataset.

One potential limitation of our work is the coverage in the target domain dataset. MiPACQ is built from Mayo Clinical ON and Mayo Clinical pathology notes related to colon cancer, which may contain limited clinical findings. In addition, clinical text consists of diverse narrative types, such as discharge summaries and clinic visit notes. Therefore, the SRL systems built on MiPACQ may need further adaptation for use in other clinical subdomains. In this study, we employed common features used in current state-of-the-art open domain SRL systems for the SRL task. However, the contribution of each feature type needs to be further examined for the SRL tasks in the clinical domain. Furthermore, instead of directly combining multiple source domain datasets, we plan to investigate more sophisticated multisource domain adaptation algorithms, such as weighting on source datasets with different distributions,\textsuperscript{62,63} which may allow us to more effectively employ multisource datasets.

**CONCLUSIONS**

Our study investigates domain adaptation techniques for SRL in clinical text. Three state-of-the-art domain adaptation algorithms were employed and evaluated by using existing SRL resources built from newswire, biomedical literature, and clinical text. Experimental results showed that domain adaptation significantly improved the SRL performance on clinical text, indicating its potential to reduce annotation costs when building clinical SRL systems.

**FUNDING**

This study was supported in part by National Institute of General Medical Sciences grant 1R01GM102282, National Library of Medicine grant R01LM010681, National Cancer Institute grant R01CA141307, and Cancer Prevention and Research Institute of Texas grant R1307.

**COMPETING INTERESTS**

None.

**CONTRIBUTORS**

The work presented here was carried out in collaboration among all authors. Y.Z., B.T., and H.X. designed methods and experiments. M.J. and J.W. carried out the experiments. Y.Z., B.T., and H.X. analyzed the data, interpreted the results, and drafted the article. All authors have been attributed to, seen, and approved the manuscript.

**ACKNOWLEDGEMENTS**

We would like to thank the anonymous reviewers for their constructive comments.

**SUPPLEMENTARY MATERIAL**


**REFERENCES**


AUTHOR AFFILIATIONS

1University of Texas School of Biomedical Informatics at Houston, Houston, TX, USA
2Department of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China

*Y.Z. and B.T. contributed equally to this article