Prior Knowledge Driven Domain Adaptation

Gourab Kundu KUNDU2@ILLINOIS.EDU
Ming-Wei Chang MCHANG21@ILLINOIS.EDU
Dan Roth DANR@ILLINOIS.EDU
Computer Science Department, University of Illinois at Urbana Champaign, IL 61801

Abstract

The performance of a natural language system trained on one domain often drops significantly when testing on another domain. Therefore, the problem of domain adaptation remains one of the most important natural language processing challenges. While many different domain adaptation frameworks have been proposed, they have ignored one natural resource – the prior knowledge on the new domain. In this paper, we propose a new adaptation framework called Prior knowledge Driven Adaptation (PDA), which takes advantage of the knowledge on the target domain to adapt the model. We empirically study the effects of incorporating prior knowledge in different ways. On the task of part-of-speech tagging, we show that prior knowledge results in 42% error reduction when adapting from news text to biomedical text. On the task of semantic role labeling, when adapting from one news domain to another news domain, prior knowledge gives 25% error reduction for instances of *be* verbs (unseen in training domain) and 9% error reduction over instances of all verbs.

1. Introduction

Domain adaptation is an important issue for statistical learning based systems. For example, in natural language processing (NLP) tasks, statistical models trained on labeled data of one domain perform well on the same domain, but their performance degrades severely when tested in a different domain. For example, all systems of CoNLL 2005 shared task (Carreras & Màrquez, 2005) on Semantic Role Labeling (SRL) show a performance degradation of almost 10% or more when tested on a different domain. For the task of part-of-speech (POS) tagging, performance drops almost 9% when systems trained on Wall Street Journal (WSJ) domain are tested on the Biomedical domain (Blitzer et al., 2006). Since labeling is expensive and time-consuming, there is a need for adapting the model trained on a large amount of labeled data of a domain (source domain) to a new domain (target domain) which may have very few or no labeled data.

One important resource that has largely been ignored in domain adaptation efforts is the prior knowledge on the target domain. Moreover, such prior knowledge is easy to collect and it is available for many domains. Prior knowledge may be available about the content of the domain like the vocabulary or structure of the sentences or styles of text. For example, transcribed text usually does not have any capitalization that is present in manually written text and this information is often known a priori. Prior knowledge may also be available about the annotation differences of the source and the target domain. For example, the names of all entities are annotated as proper nouns in WSJ domain (WSJWiki) whereas the names of genes are annotated as common nouns in Biomedical domain (BioIEWiki). As another example, in the CoNLL 2007 shared task of domain adaptation (Nivre et al., 2007) for dependency parsing, there were significant annotation differences across the source and target domain data. The participating teams did not have access to any labeled data in the target domain and so they could not learn a model to account for the annotations in the target domain. In the end, no team could improve the results substantially above the result obtained by the source domain model applied directly to the target domain data.

Over the years, many adaptation frameworks have been proposed in the literature. Some of them focus on how to use a small amount of labeled data from the

Presented at *the ICML 2011 Workshop on Combining Learning Strategies to Reduce Label Cost*, Bellevue, WA, USA, 2011.

target domain together with a large amount of labeled data from the source domain (Finkel & Manning, 2009; Chelba & Acero, 2004; Daumé III, 2007; Jiang & Zhai, 2007). Other adaptation techniques (Huang & Yates, 2009; Blitzer et al., 2006) focus on adapting their models from learning correspondence information from the unlabeled data of the source and target domain. However, these techniques do not have any information about the difference of annotations across the source and target domain. To our knowledge, very few existing adaptation algorithms use prior knowledge about the target domain to improve adaptation.

In this paper, we ask and propose solutions for two research questions:

- Is prior knowledge on the target domain useful for adaptation? (Section 4)
- Which method is most effective in incorporating the prior knowledge? (Section 5)

On the tasks of POS tagging and semantic role labeling, we show that our framework for incorporating prior knowledge (PDA) can lead to significant improvements.

2. Tasks and Datasets

We evaluate PDA on two natural language tasks: POS tagging and semantic role labeling. POS tagging assigns a part-of-speech tag for each word in a sentence. Semantic role labeling assigns semantic roles to different parts of a sentence for each predicate in the sentence. Both these tasks are critical for natural language understanding and so adapting them is very important.

POS Tagging Adaptation In POS tagging adaptation, we use the WSJ domain as the source domain and the Biomedical domain as the target domain. Following the experimental setting from (Jiang & Zhai, 2007), we use 6166 WSJ sentences from Sections 00 and 01 of Penn Treebank as the training data and 2730 PubMed sentences from the Oncology section of PennBioIE corpus as the target domain data[1]. The learning algorithm we used is Structured SVM (Tsochantaridis et al., 2004) with the L2-Loss functions implemented in the **J-LIS** package (Chang et al., 2010b). The features used are the current word, previous and next word, prefixes and suffixes of upto length 3, capitalization, hyphenation and presence of number in the current word.

[1] http://bioie.ldc.upenn.edu/publications/latest_release/data/

SRL adaptation For semantic role labeling (SRL) (Carreras & Màrquez, 2005) adaptation, we use the WSJ domain as the source domain and the Ontonotes news section as the target domain. We use Section $02 - 21$ from Penn TreeBank labeled with PropBank as training data and Section 00 of voa (Voice of America) of Ontonotes release 4.0 as the testing data. The baseline SRL model is our implementation of (Punyakanok et al., 2008) which was the top performing system in the CoNLL 2005 shared task (Carreras & Màrquez, 2005). Due to space constraints, we only give a short description of the system. The first phase of the system is *Pruning*. In this phase, we use the XuePalmer heuristic (Xue & Palmer, 2004) for pruning away unlikely arguments. The second phase is *Argument Identification* which utilises a binary classifier to decide if a candidate argument supplied by the pruning phase is an argument or not. The third phase is *Argument Classification*. In this phase, a multi-class classifier is used to predict the types of the argument candidates. The final phase is *Inference* where a global consistent labeling of argument candidates is produced subject to some linguistics and structural constraints such as *arguments cannot overlap* or *each verb can take at most one core argument*. For more details, interested readers are referred to (Punyakanok et al., 2008).

In the training data of WSJ, semantic role annotation is available for all verbs except *be* verbs e.g. *am, is, are* etc. However, in Ontonotes, semantic roles are also annotated for these verbs. Since some semantic roles (core arguments $A0 - A5$) depend on the verb instead of being verb independent, the model trained over WSJ performs poorly on Ontonotes since it did not see any labeled data for *be* verbs.

The performances of our baseline models for POS tagging and SRL are shown respectively in Table 1 and Table 2. In both tasks, our systems achieve state-of-the-art results when tested on the same domain (WSJ). Table 1 and Table 2 also show respectively the performance of the baseline model of (Jiang & Zhai, 2007) in Bio domain and the performance of (Punyakanok et al., 2008) in WSJ domain. In SRL data, predicates are given for each sentence and the system has to predict semantic roles for each predicate. Therefore the system predicts some semantic roles for *be* verbs in Ontonotes and the F1 score is not 0 inspite of not observing any labeled data with *be* verbs.

Table 1. Results (F1) of Baseline model of POS Tagging

System	WSJ	Bio
Baseline	96.8	86.2
(Jiang & Zhai, 2007)	N/A	86.3

Table 2. Results (F1) of Baseline model of SRL. PRY08 represents the system in (Punyakanok et al., 2008).

System	all verbs WSJ	all verbs Ontonotes	be verbs Ontonotes
Baseline	76.4	58.6	15.5
PRY08	76.3	N/A	N/A

3. Prior Knowledge

On the task of POS tagging, certain differences exist among the source and target domain annotations. From the *annotation manual* (WSJWiki) for WSJ, we find that hyphenated compounds are treated as either adjective or adverb. Nevertheless from the *annotation wiki* (BioIEWiki) for Bio domain, hyphenated compounds are gene names and so are labeled as common noun. The following knowledge is directly extracted from the annotation wiki for Bio domain:

1. Hyphenated words should be tagged as common noun (NN).
2. Digit-letter combinations should be tagged as common noun (NN).
3. Hyphen should be tagged as HYPH.

The BioIE wiki also says:

> There are few proper nouns in these files, mostly names of individual persons or organizations. Proper names of persons, organizations, places, and species names should be tagged as NNP. Conversely, trademarked drugs or other substances that are capitalized are tagged as NN. This rule also includes gene names and symbols, abbreviations for diseases, and other pieces of biochemical jargon.

Motivated by this, we also add the following knowledge:

1. If any word unseen in source domain is followed by the word *gene*, it should be tagged as common noun (NN).
2. If any word does not appear with proper noun tag (NNP) in training data, predict the tag NN instead of NNP for that word.
3. If any word does not appear with proper noun plural tag (NNPS) in training data, predict the tag NNS instead of NNPS for that word.

On the task of SRL, the model trained over WSJ is unaware of the semantic role pattern for *be* verbs, but *the frame file* for *be* verbs in Ontonotes (included in Ontonotes release 4.0) contains annotation decisions for different senses of the *be* verbs. The knowledge extracted from the frame file is listed below:

1. If *be* verb is immediately followed by another verb, it can not have any core argument. Example: John *is* eating.
2. If *be* verb is immediately followed by the word *like*, it can have core arguments of $A0$ and $A1$. Example: And I *'m* like why 's the door open?
3. Otherwise, it can only have core arguments of $A1$ and $A2$. Example: John *is* an idiot.

In both POS tagging and SRL, knowledge used is readily available and comes at no cost.

4. Is prior knowledge on the target domain useful for adaptation?

In this section, we try to answer the first research question about the usefulness of prior knowledge. We first review Constrained Conditional Model (**CCM**) (Chang et al., 2008), which is a popular framework for incorporating prior knowledge in statistical models.

Definition (CCM) *A CCM can be represented by two weight vectors, w and ρ, given a set of feature functions $\{\phi_i(\cdot)\}$ and a small set of constraints $c = \{C_j(\cdot)\}$. The score for an assignment $y \in \mathcal{Y}$ on an instance $x \in \mathcal{X}$ can then be obtained by*

$$f_c^w(x,y) = \sum_i w_i \phi_i(x,y) - \sum_j \rho_j C_j(x,y) \quad (1)$$

where each $C_j : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ is a Boolean function indicating whether the joint assignment (x,y) violates j-th constraint. A CCM then selects $y^* = \text{argmax}_y f_c^w(x,y)$ as its prediction.

In the CCM formulation, we directly use external knowledge to set the constraint penalty ρ to ∞.

Testing with CCM For testing with CCM, the best assignment of Equation (1) can be found via Integer Linear Programming (ILP). Although ILP is intractable in the limit, it can be quite successful in practice when applied to structured prediction problems, which are often sparse (Roth & Yih, 2007).

Training with CCM The standard approach to train a CCM is to simply use standard, linear model training algorithms. That is, we discard the constraints at training time but enforce them at testing time. This approach is referred to *Learning Plus Inference* (L+I) (Punyakanok et al., 2005).

Our baseline SRL system uses some domain independent prior knowledge c. The optimal output y for input x is found by optimizing the scoring function in Equation 1. For POS tagging, no prior knowledge is used and $|c| = 0$. In the following, we define prior knowledge on the target domain to be $c' = \{C'_k(.)\}$. For our first algorithm *PDA-KW* (Algorithm 1), we simply use c' together with c. The new scoring function is in Equation 2 where ρ' is set to ∞.

$$f^w_{c,c'}(x,y) = \sum_i w_i \phi_i(x,y) - \sum_j \rho_j C_j(x,y) - \sum_k \rho'_k C'_k(x,y) \quad (2)$$

Results for this case are listed in Table 3.

From Table 3, we see that prior knowledge improves 5% for POS tagging, 4% for all verbs and 19% for be verbs for SRL. Note that these results are obtained **without any retraining** whereas all other adaptation frameworks need to retrain their model either using labeled or unlabeled data from the target domain.

Algorithm 1 Inference with Prior Knowledge without *Retraining* (PDA-KW)

Require: Input: Source Model w, Domain Independent Prior Knowledge c, Prior Knowledge on the Target Domain c', Test Set T

1: **for** each sentence $x \in T$ **do**
2: predict $\hat{y} \leftarrow \arg\max_y f^w_{c,c'}(x,y)$
3: **end for**

5. Which method is most effective in incorporating the prior knowledge?

In this section, we try to answer the second research question about how to apply prior knowledge. We would like to see if we can take advantages of the prior knowledge and perform retraining for better adaptation. Our strategy is to embed constraints in self training. The insight is that our knowledge is very accurate but apply rarely (high precision but low recall). Therefore we want to add the sentences where the knowledge changed predictions on the target domain and learn a new model so that it can generalize even when the knowledge does not apply. The algorithm (PDA-ST) is given in Algorithm 2.

Table 3. Comparsion of results (F1) for the Baseline Model versus PDA-KW

System	POS	SRL	
		All verbs	Be verbs
Baseline	86.2	58.6	15.5
PDA-KW	**91.8**	**62.1**	**34.5**

Algorithm 2 Self training with Prior Knowledge (PDA-ST)

Require: Input: Source Model w, Domain Independent Prior Kowledge c, Prior Knowledge on Target Domain c', Source Domain Labeled Data D_s, Target Domain Unlabeled Data D_u, Test Set from Target Domain D_t

1: $D_l \Leftarrow D_s$
2: **for** each sentence $x \in D_u$ **do**
3: $y_a \leftarrow \arg\max_y f^w_c(x,y)$
4: $y_b \leftarrow \arg\max_y f^w_{c,c'}(x,y)$
5: **if** $y_a \neq y_b$ **then**
6: $D_l \Leftarrow D_l \cup (x, y_b)$
7: **end if**
8: **end for**
9: $w_t \Leftarrow \text{train}(D_l)$
10: **for** each sentence $x \in D_t$ **do**
11: predict $\hat{y} \leftarrow \arg\max_y f^{w_t}_{c,c'}(x,y)$
12: **end for**

We did not have significant amount of unlabeled data for the POS tagging experiment. Therefore the test set D_t (without gold labels) was used as the unlabeled data D_u. For the SRL experiment, we had significant amount of unlabeled data from target domain. Specifically, we used Section 01 of voa portion of Ontonotes as unlabeled data D_u and Section 00 of voa portion of Ontonotes as test data D_t. We also experiment with the scenario where we perform self training on target domain unlabeled data without prior knowledge. In terms of Algorithm 2, in line 6, instead of using (x, y_b), we use (x, y_a). The resulting model is denoted as *Self-training* in Table 4.

For self-training with SRL, there is no labeled data in WSJ domain for *be* verbs. Therefore in Algorithm 2,

Table 4. Comparsion of results (F1) of different ways of incorporating knowledge

System	POS	SRL	
		All verbs	Be verbs
Baseline	86.2	58.6	15.5
Self-training	86.2	58.3	13.7
PDA-ST	**92.0**	**62.4**	**36.4**
PDA-KW	91.8	62.1	34.5

$D_s = \phi$. Since prior knowledge is available for semantic roles of *be* verbs, we train a new model only for the argument classification phase and only for *be* verbs. During testing of *Self-training* or *PDA-ST*, we use the baseline model of the argument identifier phase for all verbs including *be* verbs. Then we use the baseline model of the argument classification phase for all verbs other than *be* verbs and the newly learned model for *be* verbs. For learning the new model, all features of the baseline model for the argument classification phase are used with one additional feature which is the label predicted by the baseline model for that argument.

Since we assume that no labeled data is available in the target domain, in all experiments, we do **not** perform any parameter tuning. For self training of POS tagging, the C parameter is set to 1 and for SRL of *be* verbs, the C parameter is set to the value of 0.4. These values are the same as used by the baseline model.

From Table 4, we see that the performance of *Self-training* is the same as the baseline for POS tagging but is worse than the baseline for SRL and *PDA-ST* performs slightly better than *PDA-KW*.

In Table 5, we compare our final model of POS tagging with (Jiang & Zhai, 2007). In the first setting of (Jiang & Zhai, 2007), the authors learn a model over 300 labeled sentences from the target domain. Then they gradually remove sentences from the source domain training set where gold label differs from the label predicted by the target domain model. We compare with the best result they obtain in this setting (denoted as (Jiang & Zhai, 2007)-1 in Table 5). In the second setting, they use all (2730) labeled sentences from the target domain and learn a model over weighted combination of source domain and target domain labeled data. We compare with the best result they obtain in this setting (denoted as (Jiang & Zhai, 2007)-2 in Table 5).

From Table 5, we see that even without using any labeled data from the target domain, *PDA-ST* significantly outperforms (Jiang & Zhai, 2007) when little labeled data are available and recovers 72% of accuracy gain that (Jiang & Zhai, 2007) had after adding a lot of labeled data from the target domain and using sophisticated weighting schemes. For the task of SRL adaptation, we are not aware of any work that reported results of adaptation from WSJ to Ontonotes and we are unable to compare with others.

Table 5. Comparsion of results (F1) for POS Tagging with (Jiang & Zhai, 2007)

System	POS	Amount of Target Labeled Data
PDA-ST	**92.0**	0
(Jiang & Zhai, 2007)-1	87.2	300
(Jiang & Zhai, 2007)-2	**94.2**	2730

6. Related Work

Domain Adaptation It is widely believed that the drop in performance of statistical models on new domains is from the shift of the joint distribution of labels and examples, $P(Y, X)$. Many domain adaptation frameworks have been proposed to address different aspects of the distribution shift. Labeled adaptation frameworks, which focus on the shift of $P(Y|X)$, makes use of target labeled examples to find the shift of the distribution and improve the model (Chelba & Acero, 2004; Daumé III, 2007; Finkel & Manning, 2009). Unlabeled adaptation frameworks use the unlabeled examples from both domains to address the shift of $P(X)$ (Blitzer et al., 2006; Huang & Yates, 2009). Frameworks that address both $P(Y|X)$ and $P(X)$ distributions also exist (Jiang & Zhai, 2007; Chang et al., 2010a). Different from all these frameworks, this paper proposes to use easy-to-get constraints that describe the domain difference to adapt the models.

Constrained Conditional Model Applying constraints on statistical models has been shown to be an effective way to improve the models (Roth & Yih, 2004; Clarke & Lapata, 2007). Constrained Conditional Models provide a general framework for combining constraints and statistical models (Chang et al., 2008). In this paper, we follow the same philosophy of combining domain knowledge and statistical models and use them to solve the domain adaptation problem. Our retraining algorithm is closely related to constrained-driven learning (Chang et al., 2007), which takes advantage of constraints to obtain better feedback to improve the statistical models. Nevertheless, we found that the use of constraints is critical when working on the domain adaptation problems.

7. Conclusion

In this paper, we introduce *PDA*, a new framework of adaptation based on prior knowledge of the target domain. We want to emphasize that prior knowledge about the domains is available for many domains and such knowledge is usually cheap to collect and can result in significant improvements. We show that using prior knowledge can give competitive results to using labeled data from the target domain. We also discover that augmenting prior knowledge with self training can give further improvements. In future, we want to apply *PDA* to other tasks. We also aim to find some theoretical justifications for self training with prior knowledge on a different domain.

Acknowledgements

This research is sponsored by the Army Research Laboratory (ARL) under agreement W911NF-09-2-0053 and by the Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181. Any opinions, findings, conclusions or recommendations are those of the authors and do not necessarily reflect the view of the ARL, the DARPA, AFRL, or the US government.

References

BioIEWiki. Pos tags in bioie wiki. http://bioie.ldc.upenn.edu/wiki/index.php/POS_tags.

Blitzer, John, McDonald, Ryan, and Pereira, Fernando. Domain adaptation with structural correspondence learning. In *EMNLP*, 2006.

Carreras, X. and Màrquez, L. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 2005.

Chang, M., Ratinov, L., and Roth, D. Guiding semi-supervision with constraint-driven learning. In *ACL*, 2007.

Chang, M., Ratinov, L., Rizzolo, N., and Roth, D. Learning and inference with constraints. In *AAAI*, 2008.

Chang, M., Connor, M., and Roth, D. The necessity of combining adaptation methods. In *EMNLP*, 2010a.

Chang, M., Goldwasser, D., Roth, D., and Srikumar, V. Structured output learning with indirect supervision. In *ICML*, 2010b.

Chelba, Ciprian and Acero, Alex. Adaptation of maximum entropy capitalizer: Little data can help a lot. In Lin, Dekang and Wu, Dekai (eds.), *EMNLP*, 2004.

Clarke, James and Lapata, Mirella. Modelling compression with discourse constraints. In *Proceedings of the 2007 Joint Conference of EMNLP-CoNLL*, 2007.

Daumé III, Hal. Frustratingly easy domain adaptation. In *ACL*, 2007.

Finkel, J. R. and Manning, C. D. Hierarchical bayesian domain adaptation. In *NAACL*, 2009.

Huang, Fei and Yates, Alexander. Distributional representations for handling sparsity in supervised sequence-labeling. In *ACL*, 2009.

Jiang, Jing and Zhai, ChengXiang. Instance weighting for domain adaptation in nlp. In *ACL*, 2007.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. The conll 2007 shared task on dependency parsing. In *CoNLL*, 2007.

Punyakanok, V., Roth, D., Yih, W., and Zimak, D. Learning and inference over constrained output. In *IJCAI*, 2005.

Punyakanok, V., Roth, D., and Yih, W. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 2008.

Roth, D. and Yih, W. A linear programming formulation for global inference in natural language tasks. In Ng, Hwee Tou and Riloff, Ellen (eds.), *CoNLL*, 2004.

Roth, D. and Yih, W. Global inference for entity and relation identification via a linear programming formulation. In Getoor, Lise and Taskar, Ben (eds.), *Introduction to Statistical Relational Learning*, 2007.

Tsochantaridis, I., Hofmann, T., Joachims, T., and Altun, Y. Support vector machine learning for interdependent and structured output spaces. In *ICML*, 2004.

WSJWiki. Penn treebank annotation guidelines. ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz.

Xue, N. and Palmer, M. Calibrating features for semantic role labeling. In *EMNLP*, 2004.