Transfer Learning for Bibliographic Information Extraction

Quang-Hong Vuong[1] and Takasu Atshuhiro[2]
[1]*Master Student, Hanoi University of Science and Technology, Dai Co Viet, Hanoi, Vietnam*
[2]*National Institute of Informatics, Hitotsubashi, Tokyo, Japan*

Keywords: Transfer Learning, Bibliographic Information Extraction, Conditional Random Fields, Page Layout Analysis, Digital Libraries.

Abstract: This paper discusses the problems of analyzing title page layouts and extracting bibliographic information from academic papers. Information extraction is an important task for easily using digital libraries. Sequence analyzers are usually used to extract information from pages. Because we often receive new layouts and the layouts also usually change, it is necessary to have a machenism for self-trainning a new analyzer to achieve a good extraction accuracy. This also makes the management becomes easier. For example, when the new layout is inputed, There is a problem of how we can learn automatically and efficiently to create a new analyzer. This paper focuses on learning a new sequence analyzer automatically by using transfer learning approach. We evaluated the efficiency by testing three academic journals. The results show that the proposed method is effective to self-train a new sequence analyer.

1 INTRODUCTION

Recently, the digitization of documents is very popular. However what we need is not just the digitization of books, we also want to create an information archive accessible from everywhere in the world. Digital libruaries (DLs) is a type of information storage. The researchers have built their institutional repositories that can be accessed from web. As it is known, bibliographic information about documents are indispensable for the efficient access to and ultilization of digital documents. Moreover, bibliographic information extraction is a key technology for realizing such information archives as intellectual legacies because it will enable the extraction of various kinds of metadata and will provide the users of such archives with full access to rich information sources.

For academic documents that we have studied here, we are interested in title, abstract, author .etc. These information can be used to identify records which are stored in different DLs. Many scientists have studied to extract bibliographic information from papers and documents. (Peng and McCallum, 2004) presented an empirical exploration of several factors, including variations on Gaussian, exponential and hyperbolic-L_1 priors for improved regularization. (Takasu, 2003) has proposed a method for extracting bibliographic attributes from reference strings captured using optical character recognition (OCR) and an extended hidden Markov model. (I. G. Councill and Kan, 2008) used conditional random field (CRF) model to label the token sequences in the reference strings. He also used a heuristic model to identify reference strings from a plain text file and to retrieve the citation contexts. (Takasu and Ohta, 2014) have proposed a method to detect layout changes and how they learn to use a new sequence analyzer efficiently. Although there were many results, it remains an active research area, with several competitions having been held [1].

In addition, for accurate information extraction, the scientists have proposed different rule-based methods that can exploit both logical structure and page layout. However, most of them can not learn automatically when we received a new page layout. Therefore, we studied and proposed a method, which can learn automatically a new page layout by using transfer learning approach.

Transfer learning has been known as an approach that addresses the problem of how to utilize much of the labeled data in the source domain to solve related but different problems in a target domain, even when the training and testing problems have different dis-

[1]http://www.icdar2013.org/program/competitions

Vuong Q. and Atsuhiro T..
Transfer Learning for Bibliographic Information Extraction.
DOI: 10.5220/0005283003740379
In *Proceedings of the International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2015), pages 374-379
ISBN: 978-989-758-076-5

tributions or features (S. J. Pan and Yang, 2013). To cater for the various situations involving the source and target domains and tasks, we can identify three transfer-learning categories, namely inductive transfer learning, transductive transfer learning, and unsupervised transfer learning (Quang-Hong and Takasu, 2014).

In this paper, we focused on how to use transfer learning for bibliographic information extraction to train a new analyzer automatically. We evaluated the efficiency and the correctness by testing three journals data set.

In summary, our main contributions are as follows.

- We propose a new method that can learn a new analyzer automatically.
- We implemented to prove the efficiency and the correctness of our method.

The remainder of the paper is organized as follows. In Section 2, we discuss related work. In Section 3, we present our proposed method. We describe our experiments, present our experimental results, and discuss them in Section 4. Finally, Section 5 concludes the paper and suggests some future work.

2 RELATED WORK

2.1 Semi-supervised Conditional Random Field

Semi-supervised approach is used as the base learner of our transfer learning method (F. Jiao and Schuurmans, 2006). The main contribution of semi-supervised approach is ultilization the unlabeled data to improve the accuracy. It is also a easy approach to use and discover the latent components of unlabeled data to train. Therefore, we use semi-supervised CRF as base learner to train a new analyzer that fits for unlabeled data. In following, we present more details about semi-supervise CRF.

Let $\mathbf{X}$ be a random variable over data sequences to be labeled, and $\mathbf{Y}$ be a random variable over corresponding label sequences. All components, Y_i, of $\mathbf{Y}$ are assumed to range over a finite label alphabet $\mathcal{Y}$. Assume we have a set of labeled samples, $\mathcal{D}^l = \left(\left(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}\right), ..., \left(\mathbf{x}^{(N)}, \mathbf{y}^{(N)}\right)\right)$ and unlabeled samples $\mathcal{D}^u = \left(\mathbf{x}^{(N+1)}, ..., \mathbf{x}^{(N+M)}\right)$. We would like to build a CRF model

$$
\begin{aligned}
p_\theta(\mathbf{y}|\mathbf{x}) &= \frac{1}{Z_\theta((\mathbf{x}))} exp(\sum_{k=1}^{K} \theta_k f_k(\mathbf{x}, \mathbf{y})) \\
&= \frac{1}{Z_\theta((\mathbf{x}))} \exp(< \theta, f(\mathbf{x}, \mathbf{y}) >)
\end{aligned} \tag{1}
$$

over sequential input and output data $\mathbf{x}$, $\mathbf{y}$, where parameter vector $\theta = (\theta_1, ..., \theta_K)^T, f(\mathbf{x}, \mathbf{y}) = (f_1(\mathbf{x}, \mathbf{y}), ..., f_K(\mathbf{x}, \mathbf{y}))^T$ and

$$
Z_\theta((\mathbf{x})) = \sum_{\mathbf{y}} \exp(\sum_{k=1}^{K} \theta_k f_k(\mathbf{x}, \mathbf{y})) \tag{2}
$$

(F. Jiao and Schuurmans, 2006) proposed a semi-supervised learning algorithm which exploits a form of *entropy regularization* on unlabeled data. For semi-supervised CRF, they proposed to maximize the following objective

$$
\begin{aligned}
RL(\theta) = &\sum_{i=1}^{N} \log p_\theta(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}) - U(\theta) \\
&+ \gamma \sum_{i=N+1}^{M} \sum_{\mathbf{y}} p_\theta(\mathbf{y}|\mathbf{x}^{(i)}) \log(\mathbf{y}|\mathbf{x}^{(i)})
\end{aligned} \tag{3}
$$

Here, γ is a tradeoff parameter that controls the influence of the unlabeled data. It determines the impact of unlabeled data set. Because our target is to learn a new analyzer that is closest to new data set, we set it is large enough.

2.2 Unilateral Transfer Adaboost Method

(Quang-Hong and Takasu, 2014) presented the $UnilateralTransferAdaBoost$ method, which extends $TransferAdaBoost$ (W. Dai and Yu, 2007) in terms of transfer learning. The algorithm aims to boost the accuracy of a weak learner by carefully adjusting the weights of training instances and learns a classifier accordingly. The main idea of $Unilateral - TrAdaBoost$ is that, at each iteration, the effect of training instances that are misclassified is reduced by multiplying its weight by $\beta^{|h_t(x_i) - c(x_i)|}$, where $h_t(x_i) : X \to Y$ is the hypothesis that, at the t^{th} iteration, $\beta \in (0, 1]$. Therefore, in the next round, those misclassified diff-distribution training instances that are dissimilar to the same-distribution training instances will affect the learning process less than in the current round. The decision function is then

$$
h_f(\mathbf{x}) = arg \max_k \sum_{t=\lceil \frac{N}{2} \rceil}^{N} \boldsymbol{u}_{k,t}^T \mathbf{x} \log \beta_t \tag{4}
$$

where $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$, ε_t is the error for hypothesis h_t, $\mathbf{u}_{k,t}$ is weight vector associated with class k and t^{th} hypothesis, and N is the maximum number of iterations for the $Unilateral - TrAdaBoost$ algorithm.

They proposed a new strategy to update the weight vector. They only updated the weight of misclassified samples from different distribution. Therefore, we can use this strategy to semi-supervised methods. More details present in the next section.

3 PROPOSED METHOD

To enable transfer learning, we use the unlabeled data set that have the new page layout to run a role in building the classification model. We call these data *target-domain data*. Moreover, these target-domain data do not have label. Therefore, we can not use them to train a classifier. The labeled training data, whose distribution may differ from the target-domain data, perhaps because they are out-dated, are called $source-domaindata$. The classifiers learned from these data cannot classify the target-domain data well due to different domains.

Formally, let $\mathbf{X}_t$ be the target-domain data, $\mathbf{X}_s$ be the source-domain data, $\mathbf{X} = \mathbf{X}_t \cup \mathbf{X}_s$ be the domain-data, and $\mathbf{Y} = \{y_i\}$ be the set of category labels. The training data set $\mathbf{T}$ contained labeled set $\mathbf{T}_s$, and unlabeled set $\mathbf{T}_t$. $\mathbf{T}_s$ represents the source-domain data that $\mathbf{T}_s = \{(\mathbf{x}_s^i, \mathbf{y}_s^i)\}$, where $\mathbf{x}_s^i \in \mathbf{X}_s (i = 1, \ldots, N)$. $\mathbf{T}_t$ represents the target-domain data that $\mathbf{T}_t = \{\mathbf{x}_t^i\}$, where $\mathbf{x}_t^i \in \mathbf{X}_t (i = 1, \ldots, M)$. N and M are the sizes of $\mathbf{T}_s$ and $\mathbf{T}_s$, respectively. The combined training set $\mathbf{T} = \{(\mathbf{x}_i, y_i)\}$ is defined as follows

$$\mathbf{x}_i = \begin{cases} \mathbf{x}_i^s, i = 1, \ldots, N; \\ \mathbf{x}_i^t, i = N+1, \ldots, N+M; \end{cases}$$

Here, $\mathbf{T}_s$ corresponds to some labeled data from a source-domain data that we try to reuse as much as we can; however we do not know which part of $\mathbf{T}_s$ is useful to us. What we only have is a unlabeled data set $\mathbf{T}_t$ from target-domain data, and then use these data to find out the useful part of $\mathbf{T}_d$. The problem that we are trying to solve is: given an unlabeled data set from target-domain data $\mathbf{T}_t$, a labeled data set from source-domain data $\mathbf{T}_s$, the objective is to train an analyzer to label each token with its type of bibliographic component.

We now present our method, $Transfer-CRF$, which extends $Unilateral-TrAdaBoost$ (Quang-Hong and Takasu, 2014) for CRF. However, Unilateral-TrAdaBoost is similar to most traditional machine learning methods which need a few labeled data to train. Therefore, it can not be used to learn a new model automatically. In our extension to Transfer-CRF, Transfer-CRF applied Unilateral-TrAdaBoost's learning strategy to filter only consistency samples to build a good model. Thus, in our extension, we use a mechanism to choose useful samples.

A formal description is presented in Algorithm 1. we can see that at each iteration, if a training sample from source-domain data is mistakenly predicted, it may conflict with the target-domain data. Therefore it will reduce its effect by remove from trainning set or reduce its weight in training phase (here we remove it from trainning set). The algorithm stoped when the number of iteration is larger than a number that is inputed by user or we can not remove any sample from source-domain data. The output of algorithm contain the labeled data that is consistent with unlabeled data. Therefore, we can use them to train a better analyzer.

Algorithm 1: Transfer-CRF.

1 . **procedure** $TRANSFER-CRF(\mathbf{T}_t, \mathbf{T}_s, N)$
2 . **Input:** Given two data set $\mathbf{T}_t$, $\mathbf{T}_s$ to train, and number of iteration K
3 . **for** $i \leftarrow 1, K$ **do**
4 . Call **semi-supervised CRF**, providing the combined training set T. Return with a hypothesis $h_i : \mathbf{X} \rightarrow \mathbf{Y}$.
5 . **if** $h_i(\mathbf{x}_i \in \mathbf{X}_s) \neq y_i$
6 . Remove $\mathbf{x}_i$ from $\mathbf{T}_s$
7 . **end if**
8 . **if** can not remove any sample
9 . break **for**
10. **end if**
11. **end for**
12. **Output:** a new analyzer h_f that is the last hypothesis

4 EXPERIMENTS

This section examines the efficiency and the effectiveness by evaluating the accuracy in labeling the unlabeled data that have new layout.

4.1 Dataset

For this experiment, we used the same three journals as in our previous study (M. Ohta and Takasu, 2010), as follows:

- Journal of Information Processing by the Information Processing Society of Japan (IPSJ): We used papers published in 2003 in this experiment. This dataset contains 479 papers, most of them has been written in Japanese.
- English IEICE Transactions by the Institute of Electronics, Information and Communication Engineers in Japan (IEICE-E): We used papers pub-

Table 1: Feature templates of CRF for bibliographic component labeling (M. Ohta and Takasu, 2010).

Type	Feature	Description
Unigram	$< i(0) >$	Current line ID
	$< x(0) >$	Current line abscissa
	$< y(0) >$	Current line ordinate
	$< w(0) >$	Current line width
	$< h(0) >$	Current line height
	$< g(0) >$	Gap between current and preceding lines
	$< cw(0) >$	Median of character widths in the current line
	$< ch(0) >$	Median of character heights in the current line
	$< \#c(0) >$	Number of characters in the current line
	$< ec(0) >$	Proportion of alphanumerics in the current line
	$< kc(0) >$	Proportion of kanji in the current line
	$< jc(0) >$	Proportion of hiragana and katakana in the current line
	$< s(0) >$	Proportion of symbols in the current line
	$< kw(0) >$	Presence of predefined keywords in the current line
Bigram	$< y(-1), y(0) >$	Previous and current labels

lished in 2003. This dataset contains 473 papers written in English.

- Japanese IEICE Transactions by the Institute of Electronics, Information and Communication Engineers in Japan (IEICE-J): We used papers published between 2000 and 2005. This dataset contains 964 papers, most of them has been written in Japanese.

As in (M. Ohta and Takasu, 2010), we used the following labels for the bibliographic components:

- Title: We used separate labels for Japanese and English titles because Japanese papers contained titles in both languages.
- Authors: We used separate labels for author names in Japanese and English as in the title.
- Abstract: As for the title and authors, we used separate labels for Japanese and English abstracts.
- Keywords: Only Japanese keywords are marked up in the IEICE-J.
- Other: Title pages usually contain paragraphs such as introductory paragraphs that are not classified into any of the above bibliographic components. We assigned the label other to the tokens in these paragraphs.

Note that different journals have different bibliographic components in their title pages.

Because we used the chain-model CRF, the tokens must be serialized. We therefore used lines extracted via OCR as tokens and serialized them according to the order generated by the OCR system. We labeled each token for training and evaluation manually.

4.2 Features of the CRF

As in (M. Ohta and Takasu, 2010), the data set has 15 features including 14 unigram features, i.e., the feature function $f_k(y_{i1}, y_i, \mathbf{x})$ is independently calculated with the previous label y_{i1}. Another feature is bigram feature, i.e., the feature function $f_k(y_{i1}, y_i, \mathbf{x})$ is dependently calculated with the previous label y_{i1}. Table 1 summarizes the set of feature templates. Their values were calculated automatically from the token and label sequences.

An example of the bigram feature template $< y(-1), y(0) >$ is:

$$f_k(y_{i1}, y_i, \mathbf{x}) = \begin{cases} 1 & if\, y_{i-1} = title and y_i = author \\ 0 & otherwise \end{cases}$$

An example of the unigram feature is:

$$f_k(y_{i1}, y_i, \mathbf{x}) = \begin{cases} 1 & if\, y_i = author \\ 0 & otherwise \end{cases}$$

The bigram features present label structure of chain-CRF with the corresponding parameter θ_k showing how likely a label follows another label.

4.3 Comparison of Methods

Our experiments are implemented in the following three cases:

- We use an journal to train and the remaining journals to test.
- For each journal, we use a small number of sample to train and the rest to test.

- We use an journal as the labeled data, and another journal as the unlabeled data to train an analyzer, then we use the analyzer to label the unlabeled data.

We measured the accuracy of a learned CRF from three cases and compare them. The accuracy was measured by (5) (*precision*) and $\overline{F_1}$ [2] that is the average of all $F_1(y_i)$, where

$$\frac{\#number\ of\ correct}{\#total\ predicted} \tag{5}$$

$$F_1(y_i) = \frac{2precision(y_i).recall(y_i)}{precision(y_i)+recall(y_i)} \tag{6}$$

where $precision(y_i)$ and $recall(y_i)$ have been defined as follows:

$$precision(y_i) = \frac{a_{y_i}}{N_{y_i}}$$

$$recall(y_i) = \frac{a_{y_i}}{M_{y_i}}$$

where a_{y_i} is number of correct y_i, N_{y_i} is number of y_i in the predicted results, M_{y_i} is number of y_i in the sample.

4.4 Results and Discussion

In the first experiment, we use chain-CRF to learn a new analyzer. The training corpus is the labeled datas which comes from the source-domain, and testing corpus is the new unlabeled datas from the target-domain. The experiment results in the table 2 show that this approach is inefficient. The accuracy on both measurement (5) and $\overline{F_1}$ is significantly decreased.

From section 4.1, we can see that IEICE-J and IPSJ have some similar characteristics (such as both written in Japanese). However, the accuracy is still small if we compare it with the other chain-CRFs whose training and testing corpus are drawn from the same domain. Similarly, we can also see that IEICE-J and IEICE-E have the same content but they are presented by other languages, some value of features such as the proportional(X) of several kinds of characters in the tokens are different. Thus the low accuracy is inevitable. In general, this approach can not be applied to learn a new analyzer automatically.

With the second experiment, we use a small number of samples to train and the rest to test. Table 3 show the precision and F1-accuracy in this case.

The table shows that chain-CRF is a good method which can reach a high F1-accuracy and precision. However, it needs the labeled data from same domain, and the F1-accuracy is trivial when the number of label is larger (IEICE-J has 5 labels, the two remain jounals have 4 labels). Therefore, chain-CRF can not be applied to learn a new analyzer automatically.

[2] http://en.wikipedia.org/wiki/F1_score

Table 2: The precision and F1-accuracy of chain-CRF when target-domain datas have new layout which are disimilar to source-domain datas.

Train set	Test set	Precision	$\overline{F_1}$
IEICE-J	IEICE-E	77.96%	47.68%
IPSJ	IEICE-E	71.39%	47.96%
IEICE-J	IPJP	76.44%	58.01%
IEICE-E	IPJP	63.57%	38.49%
IEICE-E	IEICE-J	75.80%	42.50%
IPSJ	IEICE-J	88.03%	68.14%

Table 3: The precision and the accuracy of chain-CRF when target-domain data and source-domain data are drawn from same domain.

Data set	Precision	$\overline{F_1}$
IEICE-E	94.17%	91.27%
IEICE-J	93.68%	79.10%
IPSJ	96.33%	90.78%

Table 4: The precision and the accuracy of Transfer-CRF when target-domain data has new layout and disimilar to source-domain data.

Train set	Test set	Precision	$\overline{F_1}$
IPSJ	IEICE-J	89.37%	73.96%
IEICE-E	IPJP	76.23%	50.26%
IEICE-J	IPSJ	81.04%	64.37%

Finally, we implement with both of labeled datas from source-domain data and unlabeled datas from target-domain datas. Table 4 shows the F1-accuracy and precision of Transfer-CRF with number of iterations is 5. Although the F1-accuracy and precision of Transfer-CRF is lower than chain-CRF, the corpus which is used by Transfer-CRF is the unlabeled data. Therefore, it can be used automatically to learn a new analyzer. Moreover, this result can be improved by increasing the number of iterations K or using the labeled data which has been already predicted and new unlabeled data to train a new model to improve the F1-accuracy.

5 CONCLUSIONS

In this study, we proposed a method that uses an existing data set to learn a new analyzer to label the unlabeled data that have new layout. The aim is to use information from existing data set that is sufficiently consistent with the unlabeled data set. With this method, we can learn a new analyzer automatically. Our method is combined Unilateral-

TrAdaBoost's strategy and semi-supervised CRF. In future work, we plan to investigate a new method that will detect new labels from new datasets.

REFERENCES

F. Jiao, S. Wang, C. L. R. G. and Schuurmans, D. (2006). Semi-supervised conditional random fields for improved sequence segmentation and labeling. In *International Committee on Computational Linguistics and the Association for Computational Linguistics*, pages 209–216.

I. G. Councill, C. L. G. and Kan, M. Y. (2008). Parscit: An open-source crf reference string parsing package. In *Language Resources and Evaluation Conference (LREC)*, page 8.

M. Ohta, R. I. and Takasu, A. Empirical evaluation of active sampling for crf- based analysis of pages. In *IEEE International Conference on Information Reuse and Integration (IRI 2010)*, pages 13–18.

Peng, F. and McCallum, A. (2004). Accurate information extraction from research pa- pers using conditional random fields. In *Human Language Technologies; Annual Conference on the North American Chapter of the Association for Computational Liguistics (NAACL HLT)*, pages 329–336.

Quang-Hong, V. and Takasu, A. (2014). Transfer learning for emotional polarity classification. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI2014)*, pages 94–101.

S. J. Pan, J. K. and Yang, Q. (2013). Transfer learning via dimensionality reduction. In *Proceedings of the conference on artificial intelligence*, pages 677–682.

Takasu, A. (2003). Bibliographic attribute extraction from erroneous references based on a statistical model. In *Joint Conference on Digital Libraries (JCDL 03)*, pages 49–60.

Takasu, A. and Ohta, M. (2014). Utilization of multiple sequence analyzers for bibliographic information extraction.

W. Dai, Q. Yang, G. X. and Yu, Y. (2007). Boosting for transfer learning. In *Proceedings of the international conference on machine learning*, pages 193–200.