Label Embedding for Transfer Learning

Rasha Obeidat, Xiaoli Fern, Prasad Tadepalli
School of Electrical Engineering and Computer Science
Oregon State University
{obeidatr, xfern, tadepall}@eecs.oregonstate.edu

Abstract. Automatically tagging textual mentions with the concepts, types and entities that they represent are important tasks for which supervised learning has been found to be very effective. In this paper, we consider the problem of exploiting multiple sources of training data with variant ontologies. We present a new transfer learning approach based on embedding multiple label sets in a shared space, and using it to augment the training data.

Keywords— transfer learning, Label embedding.

I. INTRODUCTION

Automatically tagging textual mentions with ontological concepts, types, and entities that they represent is useful in many knowledge-intensive fields such as biology and medicine. This problem is studied under the names of Named Entity Recognition, Entity Linking, and Wikification**.** Supervised learning from annotated training data has been found to be an effective method to tackle this task. However, in most fields in general, and biology in particular, there are often multiple ontologies. For example, different ontologies such as the Cell Type Ontology, the Protein Ontology, the Sequence Ontology, and the Gene Ontology might overlap, but use different vocabulary, and provide complementary information [11]. Each ontology comes with its own annotated training data, which presents the problem of reconciling the different ontologies and effectively using the training data for the old (source) ontologies in training for a new (target) ontology.

The above problem is an instance of *Transfer learning*, *which* aims to leverage the training data from one or more source domains to improve the sample efficiency in a related target domain. *Domain Adaptation* is Transfer learning where the source and the target domains use the same label set but have different distributions [1]. Transfer learning where the label sets are variant across domains is far less studied. In many real world applications, the ontologies or label sets of different tasks could be (implicitly) overlapping and/or intricately related. For example, one biological application of natural language processing is to tag natural texts with proteins from a given protein ontology. In a related task, we might need to tag the text with genes based on a specific gene ontology. The two ontologies are clearly related and may provide useful information toward one another. For such tasks, we need a transfer learning approach that can be applied with variant ontologies/label sets, which will learn simultaneously from both domains and thus enhance the efficiency of learning.

Standard Domain adaptation techniques [3,4] are not directly applicable to this problem because they assume that the label sets are invariant. Recent work proposed a solution based on finding a mapping between the labels using *Canonical Correlation Analysis (CCA)*, and then reducing the problem to the standard domain adaptation setting [5].

We develop a method that embeds the source and target labels in a shared space and takes advantage of the shared space to transfer the knowledge. Instead of using the label embedding to produce a mapping between the source and target labels, we directly employ the label embeddings to augment the feature representation of the target examples by the predicted source label embeddings. After that, a model is trained on the target side. We conducted a preliminary study on the task of Named Entity Recognition in which we used a two dataset that use different but related annotation scheme. We ashow that our approach significantly outperforms several baselines.

II. PROBLEM SETUP

A domain $D_i = (X_i, P(X_i))$ consists of two components: the feature space X_i and the corresponding marginal distribution $P(X_i)$. Let $T_i = (Y_i, f_i(.))$ be the task i where Y_i is the label set of the domain i, and let $f_i(.) = X_i \rightarrow Y_i$ be a function that maps X_i to Y_i. The goal of transfer learning is to use the knowledge of f_s learned from source domain-task pair (D_s, T_s) to improve the learning of f_t on the target side (D_t, T_t).

In standard domain adaptation (aka transductive transfer learning [3,4,6]), the source and the target tasks are the same , i.e., $T_s = T_t$., while the domains differ (either $X_s = X_t$ or $P(X_s) = P(X_t)$). On the other hand, in the inductive transfer learning setting [7,5], which includes our work, the domains are the same or closely related, but the tasks differ, i.e, $T_s \neq T_t$..

III. TRANSFER LEARNING VIA LABEL EMBEDDING

In this section, we describe our approach to learn label embeddings and use them to transfer the learning across the domains. We follow the method presented in Kim et al. [5] to induce the label embeddings. Specifically, we use Canonical Correlation Analysis (CCA)[8] to project both source and target labels to a shared space where the correlation between the projected vectors is maximized. Then, we employ these embeddings to transfer the knowledge from the source domain to the target domain. The projection vectors then can be used to reduce the dimensionality of the variables by projecting them into k-dimensional space, where k is a parameter to be tuned.

To use the extracted embeddings in transferring the knowledge, we propose a method that works as follows: first, we train a model on the source domain, and use it to make predictions on the target domain. Then, we augment the feature space of each instance in the target domain with the label embedding corresponding to the predicted source label. Finally, a model is trained on the target domain.

A nice property of this method is that it can be applied regardless the type of relationships between the source and the target labels. It works with 1-to-1, n-to-1, and 1-to-n relationships. It is also applicable if the label types overlap.

IV. Experimental Setup

In this section, we describe our experimental setup and results on the task of Named Entity Recognition (NER).

Dataset. We used *CoNLL 2003*[1] NER benchmark dataset as a source domain and a small dataset called *TAC-KBP2015*[2] NER dataset as a target. CoNLL2003 defines four entity types: Person (PER), Organization (ORG), Location (LOC), and Miscellaneous (MICS). TAC-KBP2015 defines six entity types: Person (PER), Title (TTL), Organization (ORG), Geopolitical Entities (GPE), Location (LOC), and Facilities (FAC). Our approach doesn't need any prior knowledge of the matching types between *CoNLL 2003* and *TAC-KBP2015*.

Evaluation. We follow CoNLL exact match evaluation protocol for the NER task [9]. In particular, we calculate the recall, the precision, and the F_1*-score* for each entity type, and then micro-average the recalls, the precisions, and the F_1*-scores*.

Features and Training. We employ the standard set of features used by Stanford NLP group to train their NER[3]. The feature set includes: word features, orthographic features, feature conjunctions and others. We also train our model using Stanford NER system[4]. It provides a general implementation of Conditional Random Field [10]. We use label embeddings of size 5 in all of our experiments.

Baselines.To investigate the effectiveness of our method *AugmntTr*, we compare it to two other baselines:

- *TargetOnly*: train a model on the target dataset.
- *Pred*: use the output of source predictor as an additional feature to train a model on the target dataset.

V. Results and discussion

In this section, we present the experimental results of all approaches under study. The results are summarized in Table 1. it shows that our method *AugmntTr* produces about 7% and 9% F_1*-score* improvement over *TargetOnly* and *Pred* methods. This illustrates the ability of CCA to discover the relationship between label types in CoNLL2003 and TAC-KBP2015 datasets. Augmenting the feature space of TAC-KBP2015 dataset with the label embedding of CoNLL2003 labels transfers the knowledge from CoNLL2003 to TAC-KBP2015 via these embeddings.

TABLE I. Micro-averaged and macro-averaged recall, precision and F1 -scores of the methods TargetOnly, Pred, and AugmntTr on the task of Named Entity Recognition.

Baseline	Avg-R	Avg-P	Avg-F_1
TargetOnly	0.618	0.753	0.679
Pred	0.576	0.756	0.654
AugmntTr	**0.745**	**0.746**	**0.745**

Conclusion

We present an approach to transfer the learning with different label sets between the source and the target domains. Our approach makes use of label embeddings induced by CCA. We augment the feature space of the target data with embeddings of the predicted source labels, and then, train a model on the target domain. We find that CCA is able to produce high quality label embeddings that are capable of transferring the knowledge across domains, this explains the superiority of our approach over the baselines.

Acknowledgments

We gratefully acknowledge the support of DARPA and AFRL under the contract number FA8750-13- 2-0033.

References

[1] S. J. Pan and Q. Yang, "A survey on transfer learning," Knowledge and Data Engineering, IEEE Transactions on, vol. 22, no. 10, pp. 1345–1359, 2010.

[2] G. Schweikert, G. R¨atsch, C. Widmer, and B. Sch¨olkopf, "An empirical analysis of domain adaptation algorithms for genomic sequence analysis,"in Advances in Neural Information Processing Systems, 2009, pp.1433–1440.

[3] H. Daum´e III, "Frustratingly easy domain adaptation," arXiv preprint arXiv:0907.1815, 2009.

[4] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics, 2006, pp. 120–128.

[5] Y.-B. Kim, K. Stratos, R. Sarikaya, and M. Jeong, "New transfer learning techniques for disparate label sets," ACL. Association for Computational Linguistics, 2015.

[6] J. Jiang and C. Zhai, "Instance weighting for domain adaptation in nlp," in ACL, vol. 7, 2007, pp. 264–271.

[7] S. J. Pan, Z. Toh, and J. Su, "Transfer joint embedding for cross-domain named entity recognition," ACM Transactions on Information Systems (TOIS), vol. 31, no. 2, p. 7, 2013.

[8] H. Hotelling, "Relations between two sets of variates," Biometrika, vol. 28, no. 3/4, pp. 321–377, 1936.

[9] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," Lingvisticae Investigationes, vol. 30, no. 1, pp. 3–26, 2007.

[10] R. Leaman, G. Gonzalez et al., "Banner: an executable survey of advances in biomedical named entity recognition." in Pacific Symposium on Biocomputing, vol. 13. Citeseer, 2008, pp. 652–663.

[11] C.-T. Tsai and D. Roth, "Concept grounding to multiple knowledge bases via indirect supervision," Transactions of the Association for Computational Linguistics, vol. 4, pp. 141–154, 2016.

[1] http://www.cnts.ua.ac.be/conll2003/ner/

[2] http://www.nist.gov/tac/2015/KBP/

[3] http://nlp.stanford.edu/projects/biNER/en.prop

[4] http://nlp.stanford.edu/software/CRF-NER.shtml