BioTrHMM: Transfer Learning-Based Biomedical Named Entity Recognition Algorithm (Postprint)

Authors: Gao Bingtao, Zhang Yang, Liu Bin

Date: 2018-05-20T00:00:00+00:00

Abstract

Traditional biomedical named entity recognition methods require large amounts of annotated data from the target domain, yet annotated data is costly. To reduce the demand for target domain annotated data in biomedical named entity recognition, the problem of named entity recognition in biomedical texts is formulated as a transfer learning-based Hidden Markov Model problem. The target domain dataset requiring named entity recognition does not need extensive data annotation; instead, identification and classification for the target domain are achieved through transfer learning methods. Using data from related domains as auxiliary datasets, the data gravity method is employed to evaluate the contribution degree of auxiliary dataset samples in target domain learning, with weights calculated on both auxiliary and target domain datasets for transfer learning. Based on the weight learning model, a transfer learning-based Hidden Markov Model algorithm called BioTrHMM is constructed. Experiments on the GENIA corpus dataset demonstrate that the BioTrHMM algorithm exhibits superior performance compared to traditional Hidden Markov Model algorithms; it achieves favorable named entity recognition performance with only a small amount of target domain annotated data.

Full Text

Preamble

BioTrHMM: A Biomedical Named Entity Recognition Algorithm Based on Transfer Learning

Gao Bingtao, Zhang Yang†, Liu Bin
(College of Information Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China)

Abstract: Traditional biomedical named entity recognition methods require large amounts of labeled data in the target domain, yet data annotation is prohibitively expensive. To reduce the dependency on target-domain annotations, this paper frames biomedical named entity recognition as a Hidden Markov Model problem enhanced with transfer learning. The proposed approach eliminates the need for extensive data labeling in the target domain, enabling model training through knowledge transfer from related domains. Using relevant domain data as auxiliary datasets, we employ a data gravitation method to evaluate the contribution of auxiliary samples to target-domain learning, computing weights for transfer learning across auxiliary and target datasets. Building upon this weighted learning framework, we develop BioTrHMM, a transfer learning-based Hidden Markov Model algorithm. Experiments on the GENIA corpus demonstrate that BioTrHMM outperforms conventional HMM algorithms while requiring only minimal labeled data from the target domain to achieve strong named entity recognition performance.

Keywords: Transfer learning; Hidden Markov Model; Named entity recognition; Text mining

0 Introduction

Traditional biomedical named entity recognition methods typically require large annotated datasets to ensure robust classification performance. However, in practice, annotated data for domains of interest are often scarce, and manual annotation incurs substantial costs. Transfer learning addresses this challenge by leveraging knowledge from related domains to assist target-domain learning, thereby significantly reducing the demand for annotated target-domain data and mitigating annotation expenses.

This paper employs instance-based transfer learning to migrate knowledge from auxiliary datasets, addressing the target domain's learning challenges while minimizing reliance on annotated target data. The algorithm must resolve two key issues: (a) obtaining high-performance predictive models with limited target-domain annotations, and (b) enabling cross-domain knowledge transfer to support target task learning. We utilize a data gravitation approach to assess the contribution of auxiliary dataset samples to the target learning problem, assign weights accordingly, and propose a sample-based transfer learning method. By modifying the learning algorithm and classification methodology of Hidden Markov Models, we introduce a weighted HMM algorithm called BioTrHMM.

Hidden Markov Models (HMM) have been widely applied in traditional biomedical named entity recognition, including HMM-based classifiers using word similarity smoothing techniques, the PowerBioNE biomedical named entity recognition system, and various studies demonstrating HMM effectiveness in biomedical NER. These conventional approaches demand large annotated training sets to achieve satisfactory performance. In contrast, our BioTrHMM algorithm re-

quires only a small amount of target-domain annotated data, performing transfer learning on related but differently annotated auxiliary datasets to construct predictive models that identify named entities in target domain datasets. Experiments on the GENIA corpus demonstrate that BioTrHMM achieves superior predictive performance compared to traditional HMM while substantially reducing manual annotation overhead.

---

1 Problem Definition

The challenge of requiring extensive annotated samples for biomedical text named entity recognition, coupled with the high cost of manual annotation, motivates our transfer learning approach. For a target dataset D_t, we utilize a related but distinct domain dataset D_s as an auxiliary dataset, transforming the problem into a transfer learning scenario for HMM.

Given the training dataset $D = D_s \cup D_t$, where $V = (v_1, v_2, ..., v_m)$ represents the observation sequence and $I = (i_1, i_2, ..., i_m)$ represents the corresponding part-of-speech state sequence, our objective is to complete knowledge transfer from D_s by assigning weights to samples, yielding $D'_t = \{(\text{sample}, w)\}$ where w is the sample weight. We construct an HMM model $\lambda = (A, B, \pi)$ on D'_t such that for a given test observation sequence $V_{\text{test}} \in D_t$ (with V_{test} drawn from the same distribution as D_t), the model $f : V \to I$ identifies the part-of-speech state sequence corresponding to the observation sequence. Specifically, for a given observation sequence V, the model classifies the sequence' s part-of-speech states to produce the corresponding state sequence I, outputting observation samples whose part-of-speech states are entity types, thereby completing named entity recognition.

---

2 BioTrHMM Algorithm

We propose BioTrHMM, a transfer learning-based Hidden Markov Model algorithm that achieves strong performance on target datasets using minimal target-domain data. By evaluating the contribution of auxiliary dataset samples to the target learning problem through data gravitation, we assign weights to samples for knowledge transfer and modify the HMM learning algorithm to obtain a transfer learning-enabled model. Our technical approach comprises four main steps: dataset construction, knowledge transfer learning, model training, and prediction/evaluation, as illustrated in [Figure 1: see original paper]. Sections 2 and 3 of [Figure 1: see original paper] constitute the core BioTrHMM algorithm, detailed below.

2.1 Instance-Based Data Transfer

Transfer learning addresses cross-domain knowledge acquisition challenges and has proven effective in named entity recognition and software defect prediction. Instance-based transfer learning methods such as nearest neighbor (NN) filter and transfer naive Bayes (TNB) assign different weights to samples based on their contribution to model construction—samples more similar to target data receive greater weights. We adopt the data gravitation model from prior work to evaluate similarity between auxiliary and target dataset samples, computing weights for each sample in the dataset.

To assign weights to data samples, we assess the contribution of auxiliary data to the target learning problem through similarity between auxiliary and target samples. We compute similarity using both word similarity and edit distance between samples in D_s and D_t.

2.1.1 Similarity Computation Definition 1 (Word Similarity): The maximum length of identical substrings between two different word strings, defined as:

$$\text{Similarity} = \frac{l_{\max}}{l}$$

where $l_{\max}$ denotes the length of the longest common substring, and l denotes the length of the longer word string.

Edit Distance: The minimum number of edit operations required to transform one string into another, where permitted operations include character substitution, insertion, and deletion. For strings c and d with lengths y and e respectively, the edit distance $ED(y, e)$ is computed as:

$$ED(0, 0) = 0$$

$$ED(0, e) = ED(y, 0) = y$$

$$ED(y, e) = \min \begin{cases} ED(y-1, e) + 1 \\ ED(y, e-1) + 1 \\ ED(y-1, e-1) + \begin{cases} 0 & \text{if } c_y = d_e \\ 1 & \text{otherwise} \end{cases} \end{cases}$$

2.1.2 Weight Computation Since part-of-speech categories for named entity data belong to NN (noun, singular or mass), we compute similarity between NN-labeled samples in D_s and entity-type samples (or NN-labeled samples) in D_t. Let $p_{\text{Similarity}}$ and $p_{\text{EditDistance}}$ denote similarity between a sample in D_s and the p-th entity-type sample or NN-labeled sample in D_t. The weight W for target samples can be calculated using either:

$$W \propto \frac{m_1 m_2}{r^2} \cdot \text{Similarity}$$

or

$$W \propto \frac{m_1 m_2}{r^2} \cdot \frac{1}{\text{EditDistance}}$$

where m_1 and m_2 represent the masses of the two objects, K is a constant, and r is the distance. The final weight for each NN-labeled sample in D_s is:

$$W_{\max} = \max_{p=1,2,\ldots,m} W_p$$

where m is the total number of entity-type samples and NN-labeled samples in D_t.

Through this weighting scheme, we obtain the weighted dataset D'_t for model construction.

2.2 BioTrHMM Learning Algorithm

Our base model is a Hidden Markov Model with parameters: state transition probability matrix $A = [a_{ij}]_{n\times n}$, observation probability matrix $B = [b_j(k)]_{n\times m}$, and initial state probability vector $\pi = (\pi_i)$. We modify these parameters for the transfer learning scenario.

Traditional parameter learning computes transition probabilities from state transition counts. In our transfer learning context, parameter A is computed as:

$$a_{ij} = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}}$$

where w_{ij} represents the weight of transitions from state i to state j, and $\sum_{j=1}^{n} w_{ij}$ is the sum of all transition weights from state i.

The observation probability matrix B is computed as:

$$b_j(k) = \frac{N_{jk}}{N_j}$$

where N_{jk} is the count of observing v_k in state j, and N_j is the total count of all possible observations in state j.

The initial state probability vector π is computed as:

$$\pi_i = \frac{w_i}{\sum_{i=1}^{n} w_i}$$

where w_i represents the initial weight for state i.

We learn the parameters A, B, and π on D'_t through statistical methods to obtain model λ.

2.3 BioTrHMM Classification Algorithm

After learning model λ on D'_t, we modify the Viterbi algorithm for classification. Using sample weights from D'_t, we compute a weighted state transition matrix to replace the count-based transition matrix in the original Viterbi algorithm. Given model λ, Viterbi efficiently finds the state sequence corresponding to an observation sequence, yielding the part-of-speech sequence for a given text sequence. By analyzing this state sequence, we extract observations corresponding to named entity types, completing the named entity recognition task.

We define two variables: $\delta_t(i)$ represents the maximum probability of all single paths ending in state i at time t, and $\psi_t(i)$ represents the $t-1$-th node of the path achieving this maximum probability. The modified algorithm proceeds as follows:

a) Initialization:

$$\delta_1(i) = \pi_i b_i(v_1) = \frac{w_i}{\sum_{i=1}^{n} w_i} \cdot b_i(v_1), \quad i = 1, 2, ..., n$$

$$\psi_1(i) = 0, \quad i = 1, 2, ..., n$$

b) Recursion: For $t = 2, 3, ..., T$:

$$\delta_t(i) = \max_{1 \le j \le n} \left[\delta_{t-1}(j) a_{ji}\right] b_i(v_t) = \max_{1 \le j \le n} \left[\delta_{t-1}(j) \cdot \frac{w_{ji}}{\sum_{i=1}^{n} w_{ji}}\right] \cdot b_i(v_t), \quad i = 1, 2, ..., n$$

$$\psi_t(i) = \arg\max_{1 \le j \le n} \left[\delta_{t-1}(j) a_{ji}\right] = \arg\max_{1 \le j \le n} \left[\delta_{t-1}(j) \cdot \frac{w_{ji}}{\sum_{i=1}^{n} w_{ji}}\right], \quad i = 1, 2, ..., n$$

c) Termination:

$$P^* = \max_{1 \le i \le n} \delta_T(i)$$

$$i_T^* = \arg\max_{1 \le i \le n} \delta_T(i)$$

d) Optimal path backtracking: For $t = T-1, T-2, ..., 1$:

$$i_t^* = \psi_{t+1}(i_{t+1}^*)$$

The resulting optimal path corresponds to the part-of-speech state sequence. Based on this state sequence, we output observation samples corresponding to named entity types, completing the named entity recognition task.

3 Experiments and Results

3.1 Experimental Setup

To evaluate BioTrHMM' s predictive performance for biomedical named entity recognition, we compare it against traditional HMM. The most widely used biomedical annotation corpus is GENIA v3.02, containing 2,000 annotated abstracts from MEDLINE (approximately 360,000 words) with 36 part-of-speech categories, including five biomedical entity types. We conducted experiments using GENIA v3.02 data (http://www.nactem.ac.uk/genia/genia-corpus), focusing on protein named entity recognition. Evaluation metrics include precision, recall, and F1-score. The entity label distribution in GENIA v3.02 is shown in .

In our experiments, D_t represents the target set containing protein entity labels and other part-of-speech labels, while D_s represents the auxiliary set where protein entity labels are processed as NN type. The auxiliary set label distribution is shown in .

We define three parameters: α represents the size of target set D_t, β represents the size of auxiliary set D_s, and γ represents the proportion of the corpus used (when using the entire GENIA v3.02 corpus, $\gamma = 1$). We ensure result validity through ten-fold cross-validation for each experiment.

3.2 Experimental Results

We conducted experiments from multiple perspectives to validate BioTrHMM' s performance.

3.2.1 Experiments on Target Set Size By transferring knowledge from D_s, BioTrHMM demonstrates significantly improved recognition performance over HMM. We tested both algorithms across different α values, with results shown in [Figure 2: see original paper]. The results indicate that for the same target set size, BioTrHMM substantially outperforms traditional HMM. Through auxiliary set learning, BioTrHMM acquires more relevant knowledge for the target task, yielding a more robust model. While both algorithms share the same target set, BioTrHMM leverages differently distributed auxiliary data to dramatically improve performance without increasing annotation costs.

3.2.2 Experiments on Auxiliary-to-Target Ratio To investigate the impact of auxiliary set size, we conducted experiments with fixed training set sizes while varying the ratio between D_s and D_t. We set the size ratios to 2:1, 3:1, 4:1, and 5:1. compares results when $\gamma = 1$. Although BioTrHMM' s performance decreases as auxiliary set size reduces, it remains comparable to traditional HMM. This demonstrates that even with smaller auxiliary sets, BioTrHMM maintains competitive predictive effectiveness.

3.2.3 Experiments on Overall Dataset Size To further validate algorithm effectiveness, we conducted experiments on different dataset sizes, testing with $\gamma = 1$, $\gamma = 0.8$, and $\gamma = 0.6$. Results in and show that target dataset scale significantly impacts traditional HMM performance. In contrast, BioTrHMM maintains strong predictive performance even as dataset size decreases. Crucially, BioTrHMM effectively reduces the required amount of annotated target data, lowering manual annotation costs while preserving classification performance.

Overall results demonstrate that BioTrHMM achieves superior performance using only one-third or less of the annotated target data required by traditional HMM. The choice of similarity metric has minimal impact on BioTrHMM, as both word similarity and edit distance-based implementations outperform conventional HMM.

---

4 Conclusion

This paper addresses the challenge of requiring extensive annotated target samples for traditional biomedical named entity recognition by proposing BioTrHMM, a transfer learning-based Hidden Markov Model algorithm. By computing similarity between auxiliary and target dataset samples, BioTrHMM assigns weights based on contribution to the target task, enabling effective cross-domain knowledge transfer. The algorithm modifies HMM' s learning algorithm to train model parameters on weighted datasets, establishing a predictive model under transfer learning conditions. Experimental results demonstrate that BioTrHMM achieves better predictive performance with significantly less target-domain annotated data. The proposed method is applicable not only to biomedical text named entity recognition but also generalizable to named entity recognition in broader text mining contexts.

This work focuses exclusively on HMM improvements. However, numerous studies have shown that Conditional Random Fields (CRF) outperform HMM for named entity recognition. Building upon these findings, future work will explore transfer learning for named entity recognition based on CRF to further enhance recognition performance.

References

[1] Horn H, Schoof E M, Kim J, et al. KinomeXplorer: an integrated platform for kinome biology studies [J]. Nature Methods, 2014, 11(6): 603–604.

[2] Srinivasagan K G, Suganthi S, Jeyashenbagavalli N. NER for Hindi language using association rules [C]// Proc of International Conference on Data Mining and Intelligent Computing. 2014.

[3] Gayen V, Sarkar K. An HMM based named entity recognition system for Indian languages: The JU System at ICON [R]. Published in ArXiv, 2013.

[4] Arnold A, Nallapati R, Cohen W W, et al. Exploiting feature hierarchy for transfer learning in named entity recognition [C]// Proc of the 46th Annual Meeting of the Association for Computational Linguistics. 2008: 245-253.

[5] Pan Jialin, Yang Qiang. A survey on transfer learning [J]. IEEE Trans on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[6] Liu Jie, Yu Kai, Zhang Yi, et al. Training conditional random field using transfer learning for gesture recognition [C]// Proc of IEEE International Conference on Data Mining. 2010.

[7] Magimai-Doss M, Rasipuram, Aradilla G, et al. Grapheme-based automatic speech recognition using KL-HMM [C]// Proc of InterSpeech. 2011.

[8] Cui Xiaodong, Huang Jing, Chien J T. Multi-view and multi-objective semi-supervised learning for HMM-based automatic speech recognition [J]. IEEE Trans on Audio, Speech, and Language Processing, 2012, 20(7): 1923-1935.

[9] Hu Hao, Zheng Wenchen, Yang Qiang. Cross-domain activity recognition via transfer learning [J]. Pervasive and Mobile Computing, 2011, 7(3): 344–358.

[10] Cook D, Feuz K D, Narayanan C, et al. Transfer learning for activity recognition: a survey [J]. Knowledge and Information Systems, 2010, 36: 537–556.

[11] Pan Jialin, Toh Zhiqiang, et al. Transfer joint embedding for cross-domain named entity recognition [J]. ACM Trans on Information Systems, 2013, 31(2): Article 7.

[12] Ma Ying, Luo Guangchun, Zeng Xue, et al. Transfer learning for cross-company software defect prediction [J]. Information and Software Technology, 2012, 54: 248-256.

[13] Rabiner L, Juang B. An introduction to hidden Markov models [J]. IEEE ASSP Magazine, 1986.

[14] Zhuang Fuzhen, Luo Ping, He Qing, et al. Transfer learning research progress [J]. Journal of Software, 2015, 26(1): 26-39.

[15] Bui Q C, Katrenko S, Sloot P M A. A hybrid approach to extract protein-protein interactions [J]. Bioinformatics, 2011, 27(2): 259-265.

[16] Huang Yuanhua, Xu Bosen, Zhou Xueya, et al. Systematic characterization and prediction of post-translational modification cross-talk [J]. Molecular & Cellular Proteomics, 2015, 14(3): 761-770.

[17] Zhang Yang, Li Jianliang, Hu Zhengguo. NewsGrouper: A software tool for automatically extracting important news [J]. Computer Engineering, 2002, 28(4): 83-84.

[18] Teixeira J, Sarmento L, Oliveira E. A bootstrapping approach for training a NER with conditional random fields [C]// Portuguese Conference on Artificial Intelligence. Berlin: Springer, 2011.

[19] Konkol M, Konopík M. CRF-based Czech named entity recognizer and consolidation of Czech NER research [J]. Berlin: Springer-Verlag, 2013.

Note: Figure translations are in progress. See original paper for figures.