Modeling Sequential Annotations for Sequence Labeling With Crowds

Xiaolei Lu and Tommy W. S. Chow, Fellow, IEEE

Abstract—Crowd sequential annotations can be an efficient and cost-effective way to build large datasets for sequence labeling. Different from tagging independent instances, for crowd sequential annotations, the quality of label sequence relies on the expertise level of annotators capturing internal dependencies for each token in the sequence. In this article, we propose modeling sequential annotation for sequence labeling with crowds (SA-SLC). First, a conditional probabilistic model is developed to jointly model sequential data and annotators’ expertise, in which categorical distribution is introduced to estimate the reliability of each annotator in capturing local and nonlocal label dependencies for sequential annotation. To accelerate the marginalization of the proposed model, a valid label sequence inference (VLSE) method is proposed to derive the valid ground-truth label sequences from crowd sequential annotations. VLSE derives possible ground-truth labels from the tokenwise level and further prunes subpaths in the forward inference for label sequence decoding. VLSE reduces the number of candidate label sequences and improves the quality of possible ground-truth label sequences. The experimental results on several sequence labeling tasks of Natural Language Processing show the effectiveness of the proposed model.

Index Terms—Crowdsourcing, labeling consistency, nonlocal label dependency, sequential annotations.

I. INTRODUCTION

SEQUENCE labeling, has been widely applied in Natural Language Processing and Computational Biology. Given a sequence of observations, sequence labeling assigns an appropriate label for each observation. For example, part-of-speech (POS) tagging can be formulated as a sequence labeling task, which predicts the POS of each word for the sentence sequence and produces a POS sequence of equal length.

There have been many statistical models proposed with promising results for sequence labeling (e.g., structured SVM (S-SVM) [1] and graphical models). Similar to other supervised learning methods, sequence labeling requires a large number of training sequences with complete annotations, which is costly and laborious to produce. As semisupervised learning can perform learning tasks from a small amount of labeled data and large amounts of unlabeled data, semisupervised sequence labeling has exhibited advantages over a full manual annotation. However, semisupervised-based methods still need exact annotations.

In recent years, crowdsourcing has been widely adopted to obtain large labeled datasets in a cheap and efficient way. Crowdsourcing is the process of obtaining a large amount of labeled data from a group of ordinary people. It does not require domain experts but relies on utilizing the contribution of the group’s intelligence [2]. But the quality of labels obtained from crowdsourcing cannot be guaranteed because the expertise level of annotators varies. Therefore, learning from crowd labels still mainly focuses on estimating the reliability of an individual annotator and building classifiers to predict new input data in the aggregation process. For example, Snow et al. [3] used bias correction to combine nonexpert annotation. Raykar et al. [4] proposed to jointly estimate the coefficients of a logistic regression classifier and the annotators’ expertise. Yan et al. [5] further considered the case that the reliability of an annotator is not consistent across all the input data.

Crowdsourcing has been exploited in sequence labeling. Early work extended the methods (e.g., majority voting and weighted voting [6]) developed for independent instances with crowds to handle crowd sequential annotations, which aims to infer the ground-truth label for each token of the sequence. To better model crowd sequential annotation, studies on traditional sequence labeling models (e.g., hidden Markov models (HMMs) [7] and conditional random fields (CRFs) [8]) have been conducted aiming to include additional modeling of the quality of crowd label sequences. These methods generally treat each token as an independent instance and assume that the label quality for each token by annotators is only a function of their level of expertise. However, there exists internal label dependency among tokens in a sequence, which is different from instances that are independent of each other. For example, given the sentence “This inconsistency with observation sent Albert Einstein back to the drawing board and, on 25 November 1915, Einstein presented the updated field equations to the Prussian Academy of Sciences” for named entity recognition, the label dependency constrains all the mentions of “Einstein” located in different tokens to share the same label with “PERSON” [9]. Therefore, for crowd sequential annotations, the quality of label sequence relies on annotators’ expertise in capturing internal dependencies for each token in a given sequence.
In this article, to improve the estimation of the quality of label sequence of crowd sequential annotations, we propose a modeling sequential annotation method for sequence labeling with crowds (SA-SLC). Through modeling annotator’s expertise in capturing internal label dependency for sequential annotation, SA-SLC enhances the quality of measuring the entire label sequence and then assigns different weights to crowd label sequences for classifier learning. As a result, the negative effect of unreliable annotations in the optimization can be significantly reduced. Our contributions can be summarized as follows.

First, a conditional probabilistic model is developed to jointly model sequential data and annotators’ reliability. By decomposing internal label dependency into local and nonlocal label dependencies, we introduce categorical distribution to estimate annotators’ expertise in capturing internal label dependency, which results in improving the estimation of the quality of label sequences in the aggregation process. The concept of local and nonlocal label dependencies will be elaborated on in the later section of the article. Expectation maximization (EM) is employed for parameter estimation, where in the M step, a weighted sequential model is defined to learn parameters for predicting new sequences.

Second, to accelerate the marginalization of the proposed model, valid label sequence inference (VLSE) is proposed to derive the valid ground-truth label sequences from crowd sequential annotations. VLSE first estimates the possible ground-truth labels in the tokenwise level and prunes certain subpaths that violate the label constraint in the forward inference for label sequence decoding. As a result, the number of candidate label sequences for marginalization is substantially reduced and the efficiency of parameter estimation can be improved.

To demonstrate the effectiveness of the proposed model, we conduct experiments on the datasets collected from news articles and biomedical papers. The results show that our proposed model performs stably with varying annotators’ expertise and in most cases achieves better performance than the state-of-the-art methods.

II. RELATED WORK

There are two main types of approaches for sequence labeling: 1) generative and 2) discriminative. HMMs [10], [11] and CRFs [12] form the most popular generative-discriminative pair for sequence labeling. Many variants derived from these basic graphical models have been applied to handle complex sequence labeling tasks. For example, Coupled HMM [13] is developed to model the interactions between two state (i.e., label) sequences to capture more information from the structure. Similarly, hierarchical CRFs [14] model multiple output sequences for joint sequence labeling tasks (e.g., POS tagging and Chunking). Furthermore, Sarawagi and Cohen [15] proposed semi-Markov CRFs (semi-CRFs) to predict and segment the output sequence, which is more appropriate for the tasks of chunking and named entity recognition compared to CRFs. In recent years, the models combining deep learning and graphical models achieve competitive results compared to traditional models, such as CRF-CNN [16] and Bi-LSTM-CRF [17].

Early work on sequence labeling has obtained promising results. But the collection of a large number of training datasets with exact annotations is still a highly challenging issue. First, it is financially expensive to employ annotators with a high level of expertise who can provide precise labelings. Furthermore, it is prohibitively time consuming to manually label complex structured outputs than traditional discrete values. For example, the annotation for named entity recognition and POS tagging requires labeling each word in all sentences level.

Semisupervised sequence labeling has received increasing attention as it requires only a small amount of labeled data. Semisupervised CRFs [18] minimize the conditional entropy on unlabeled training instances and then combines with the objective of CRFs, which improves the performance of supervised CRFs. Brefeld and Scheffer [19] proposed to incorporate the co-training principle into the support vector machine to minimize the number of errors for labeled data and the disagreement for the unlabeled data, which can outperform fully supervised SVM in specific tasks. It is worth noting that semisupervised sequence labeling partly lightens the burden of sequential annotations. But semisupervised sequence labeling still needs exact labelings.

Crowdsourcing provides an efficient and cost-effective way to obtain large amounts of labeled datasets. However, the label quality cannot be guaranteed as the expert level of annotators varies. The traditional way is to infer the ground-truth label from crowd annotations, for example, majority voting [20] and weighted voting [6]. Since the subsequent classifier learning heavily depends on the inferred ground-truth label, Raykar et al. [4] developed a probabilistic framework to jointly learn the classifier and ground truths, where annotators’ expertise and the ground truths are estimated iteratively. To improve this probabilistic framework, Raykar and Yu [21] used the spammer score to rank annotators, which aims to eliminate the spammers and infer the ground truths based on the reliable annotators in each iteration. Yan et al. [5] redefined annotator’s expertise by assuming that the reliability of annotators is not consistent across all the input data. The above crowdsourcing learning models perform on the single-instance level. There have been research works studying crowd sequential annotations. HMM-Crowd [22] is developed to aggregate crowd sequential annotations in the HMM to infer a best single sequential annotation, while BSC [23] is a fully Bayesian approach that further models the effect of local sequential dependencies on annotators’ reliability. Wu et al. [24] extended the probabilistic framework of Raykar et al. [4] to crowd sequential labeling by aggregating multiple annotations with the weighted CRF model. Rodrigues et al. [20] proposed CRF-MA that assumes the reliability of an annotator is consistent among all the labels, which models the reliability of an annotator as a multinomial random variable and then treats it as a latent variable in the CRF model.

It is worth noting that sequential annotation aims to provide a reasonable label sequence for the entire input sequence.
Unlike the instances that are independent of each other, there exists internal label dependency among the tokens in the sequence. In this article, different from the above crowdsourcing sequential models, the proposed model measures annotators’ expertise and view in capturing internal label dependency among the sequence, which aims to improve the estimation of the quality of label sequence from crowd sequential annotations and enhance the performance of predicting new sequences as a result.

III. PRELIMINARIES

A. Learning From Crowd Labels

Given the dataset $X = \{x_i\}_{i=1}^N$, let $Y = \{y_i^1, y_i^2, \ldots, y_i^K\}_{i=1}^N$ where $\{y_i^1, y_i^2, \ldots, y_i^K\}$ denote the crowd labels for the $i$th data by $K$ different annotators. $Z = \{z_i\}_{i=1}^N$ represents the ground-truth label set, which is unknown to the annotators. Given the training set with crowd labels, the goal of modeling is to estimate the annotators’ expertise and a classifier to predict labels for new instances.

Formulating a probabilistic model of the learning process is a principled way to explore crowd labels. Based on the graphical model as shown in Fig. 1, the joint conditional distribution can be defined as

$$p(Y, Z|X) = \prod_i p(z_i|x_i) \prod_k p(y_i^k|x_i, z_i)$$

where $p(y_i^k|x_i, z_i)$ measures the annotators’ expert level, which, in most cases, is expressed as

$$p(y_i^k|x_i, z_i) = p(y_i^k|z_i) = \left(1 - \eta^k\right)^{|z_i - y_i^k| - 1} \eta^k$$

where $\eta^k$ denotes the expert level of the $k$th annotator.

Generally, for the independent instance with crowd labels, it is reasonable to assume that $y_i^k$ rely on $z_i$ only as no internal annotation dependency needs to be considered.

B. Sequence Labeling

Sequence labeling assigns a label sequence to the sequential input data. CRFs exhibit promising performance on sequential tagging tasks. It directly models the conditional probability of the label sequence without the assumption on the dependencies among the observations.

Given the input sequence $x = \{x_1, x_2, \ldots, x_L\}$ and the label sequence $y = \{y_1, y_2, \ldots, y_L\}$, linear-chain CRFs define the conditional probability $p(y|x)$ as

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{d=1}^D \theta_d F_d(x, y)\right)$$

where $F_d(x, y) = \sum_{i=1}^L f_d(y, y_{i-1}, x, i)$ are feature functions. $Z(x) = \sum_y \exp\left(\sum_{d=1}^D \theta_d f_d(y, y_{i-1}, x, i)\right)$ is the partition function and $\theta$ is a weight vector.

Given the training datasets $X = \{x_i\}_{i=1}^N$ and $Y = \{y_i\}_{i=1}^N$, parameter $\theta$ can be estimated by maximizing $\sum_{i=1}^N \ln p(y_i|x_i)$ in which optimization algorithms, such as conjugate gradient and L-BFGS [25], can be used.

IV. MODELING SEQUENTIAL ANNOTATION FOR SEQUENCE LABELING WITH CROWDS

In this section, we describe the formulation of SA-SLC and the proposed VLSE for the efficient marginalization of SA-SLC. Also, EM for parameter estimation is presented.

A. Formulation

Given $N$ sequential dataset $X = \{x_i\}_{i=1}^N$ where $x_i = \{x_i^j\}_{j=1}^{L_i}$ and $Y = \{y_i^1, y_i^2, \ldots, y_i^K\}_{i=1}^N$ is the crowd label sequences tagged by $K$ different annotators. The joint conditional probability $p(Y|X)$ is defined as

$$p(Y|X) = \prod_i \prod_k p(y_i^k|x_i)$$

$$= \prod_i \prod_k \sum_{z_i} p(y_i^k, z_i|x_i)$$

$$= \prod_i \prod_k \sum_{z_i} p(y_i^k|z_i, x_i)p(z_i|x_i)$$

where $z_i$ denotes one possible label sequence for the $i$th input sequence. $p(y_i^k|z_i, x_i)$ is defined as

$$p(y_i^k|z_i, x_i) = \prod_{j=1}^{L_i} p(y_i^k|z_i, x_{ij})$$

where $L_i$ denotes the length of the $i$th sequence.

Different from labeling the instances that are assumed to be independent of each other, internal label dependency among the tokens in the sequence should be considered for sequential annotation [24]. When inferring the output label of a given token, a good annotator will utilize local and nonlocal label dependencies and then provide the most reasonable label, where $p(y_{ij}|z_{ij}, x_{ij}) \neq p(y_{ij}|z_{ij})$. In this article, we formulate $p(y_{ij}|z_{ij}, x_{ij})$ with parametric modeling to measure the expert level of annotators in capturing local and nonlocal label dependencies.

First, to model the annotator’s expertise in capturing local label dependency, $p(y_{ij}|z_{ij}, x_{ij})$ is defined as

$$p(y_{ij}|z_{ij}, x_{ij}) = p(y_{ij}|y_{ij-1}, z_{ij})$$

where $y_{ij}$ depends on $z_{ij}$ and the annotator’s previous annotation $y_{ij-1}$. For example, in a task of named entity recognition, as shown in Fig. 2, a good annotator will not assign “I” to the word “carrier” if the previous annotation for “national” is “O.” When the word “Bosnian” is tagged with “B-ORG,” “Association” is more likely to be annotated with “I-ORG.”
For the $k$th annotator, $y^k_{ij}$ is assumed to follow categorical distribution given $(y_{ij-1}, z_{ij})$, which is expressed as:

$$y^k_{ij} \sim \text{Cat}(\alpha^k, \alpha^{k+1}, \ldots, \alpha^M)$$  \hspace{1cm} (7)

where $M$ is the number of all possible labels. The expert level of capturing local dependency of $k$th annotator is modeled by an $M \times M \times M$ confusion matrix $\alpha^k$. $\alpha^k(i, j, h)$ is defined as

$$\alpha^k(i, j, h) = p(y_i = h | z_i = j, y_{i-1} = i)$$  \hspace{1cm} (8)

where the ground truth is $j$, previous annotation is $i$, and the assigned label is $h$.

In sequence labeling, encoding nonlocal label dependency further improves label consistency of the entire sequence [9], [26]. For sequential annotation, the annotator naturally refers to the same (or similar) mention in previous annotation when deciding the label of the current token. As shown in Fig. 2, the annotation for the word “EgyptAir” (or “Bosnian”) can be a good indicator for labeling the subsequent occurrence “EgyptAir” (or “Bosnian”). In this article, we focus on modeling the nonlocal label dependency between the same mentions, which is expressed as

$$p(y_j | z_j, x_j) = p(y_j | z_j, [x_j, (x', y')])$$

$$= p(y_j | y'_j, z_j)$$  \hspace{1cm} (9)

where $x'$ is the same mention of $x_j$ in the previous sequence. Then, $y'_j$ has the pseudolabel $y'_j$, which equals the label $y'$ assigned to $x'$. Since $x_j$ may have several same mentions, we choose the one whose position is the closest to $x_j$.

Similarly, $y^k_{ij}$ follows categorical distribution, given $(y'_{ij}, z_{ij})$, then the expert level of capturing nonlocal label dependency of the $k$th annotator is modeled by an $M \times M \times M$ confusion matrix $\beta^k$. $\beta^k(i, j, h)$ is defined as

$$\beta^k(i, j, h) = p(y_i = h | z_i = j, y'_i = i)$$  \hspace{1cm} (10)

where the ground truth is $j$, previous same mention is $i$, and the assigned label is $h$.

Based on the above description, $p(y^k_j | z_j, x_j)$ is defined as

$$p(y^k_j | z_j, x_j) = \prod_{j=1}^{L_j} p(y^k_{ij} | z_{ij}, x_{ij})$$

$$= \prod_{j=1}^{L_j} p(y_{ij} | y'_{ij}, z_{ij})^{\exists x_{ij} \in x_t(0, j-1)}$$

$$\times p(y_{ij} | y_{ij-1}, z_{ij})^{1-\exists x_{ij} \in x_t(0, j-1)}$$  \hspace{1cm} (11)

where $\exists x_{ij} \in x_t(0, j-1)$ denotes that there exists at least one $x_{ij}$ in the sequence $x_t(0, j-1)$. $\llbracket P \rrbracket = 1$ if $P$ is true and 0, otherwise.

To improve the efficiency of parameter estimation and provide reasonable modeling of label dependency, in (11), the proposed SA-SLC ignores local label dependency when there is a previous occurrence of $x_{ij}$. Utilizing local and nonlocal label dependencies simultaneously requires a four-dimensional confusion matrix with size $M \times M \times M \times M$ that encodes label information from previous same mention, previous annotation, the ground truth, and the assigned label, which greatly increases the size of parameters to be estimated.

**B. Decoding: Valid Label Sequence Inference**

The joint conditional probability, as shown in (4), marginalizes all possible ground-truth label sequences out. But finding the joint conditional probability may not be practical because it is computationally expensive, especially the number of these sequences is exponential to the length of the sequence. For example, in the task of named entity recognition, for a sentence with ten words, the total number of possible label sequences can be $9^{10}$. Besides, for sequence annotation, enumerating all possible label sequences will generate a large number of invalid sequences, for example, possible label sequences for the sequence in Fig. 3 could be “O O O I-ORG O I-ORG I-ORG” and “O O O B-PER I-ORG O I-ORG,” which makes no sense for sequence analysis.

In this section, we describe a newly developed method, called VLSE, to derive the valid ground-truth label sequences.
from crowd sequential annotations. It estimates the possible ground-truth label sequences at the tokenwise level by exploring the label consistency among annotators. VLSE then decodes the valid label sequences for the entire input sequence.

First, let \( |l_i|_{r=1}^R \) denote the set of unique labels collected from \( K \) annotations, where \( R \) is the total number of unique labels. We define the degree of label consistency \( LC_{ij} \) for the \( j \)th element in the \( i \)th sequence as
\[
LC_{ij} = \frac{\max_l \left( \{ n_{lr} \}_{r=1}^R \right)}{R} \tag{12}
\]
where \( n_{lr} \) represents occurrence times of the label \( l_r \) in \( K \) annotations.

To derive the possible ground-truth labels \( L_{ij} \) for the \( j \)th element in the \( i \)th sequence, we consider three different cases, which are expressed as
\[
L_{ij} = \begin{cases} 
\arg \max_l \left( \{ n_{lr} \}_{r=1}^R \right), & \text{if } LC_{ij} \geq T_1 \\
\{ l_{r} \}_{r=1}^R, & \text{if } T_2 < LC_{ij} < T_1 \\
\{ l_{m} \}_{m=1}^M, & \text{if } LC_{ij} \leq T_2 
\end{cases} \tag{13}
\]
where \( M \) is the number of all possible labels. By specifying appropriate threshold \( T_1 \) and \( T_2 \), three cases of obtaining the set of possible true labels \( L_{ij} \) are as follows: 1) if \( LC_{ij} \) is high, the label with majority vote is chosen; 2) if \( LC_{ij} \) is moderate, the set of all labels assigned by the annotators to \( x_{ij} \) is considered; and 3) if \( LC_{ij} \) is low, all available labels in the tagging scheme are considered.

Generally, differences remain among \( K \) annotations. Annotators are more likely to reach an agreement when tagging simple data, where \( LC_{ij} \) can be \( K \). For the element in the sequence with much lower \( LC \), all possible labels will be considered as the ground-truth labels. For example, as shown in Fig. 4, there are five annotators tagging the sequence—“Dimitris Kontogiannis, Athens Newsroom +301 3311812-4.” The possible ground-truth labels for each token can be derived based on (13), which are presented in Fig. 5.

To derive the valid label sequences, based on the determined possible ground-truth labels for each element, VLSE further prunes some subpaths that violates the constraints created from prior knowledge. For example, for a task of named entity with “PERSON, ORGANIZATION, and LOCATION,” the constraints can be “I-” cannot follow “O,” next to “B-ORG” could be “B-,” “O,” and “I-ORG” (same to all “B-”), “I-PER” only follow “B-PER” or “I-PER” (same to “I-”). Let \( P_t(s_i) \) denote the possible subpaths from the previous token \( t-1 \) to the state \( s_i \) (i.e., label) in the current token \( t \), which is defined as
\[
P_t(s_i) = \left\{ P_{t-1}(s_j) \right\}_{j=1}^T \tag{14}
\]
where \( \{ s_j \}_{j=1}^T \) is the set of possible states in the token \( t-1 \) and the transition \( s_j \to s_i \) does not violate the constraints.

By incorporating the constraints into the forward inference for label sequence derivation, VLSE further reduces the number of candidate label sequences in the marginalization, which results in improving the quality of possible ground-truth label sequences. For example, as shown in Fig. 6, without pruning the number of possible ground-truth label sequences is 729. Based on the constraints described above for the task of named entity recognition with “PERSON, ORGANIZATION, and LOCATION,” the subpaths with the red-dotted line will be pruned. As a result, some invalid label sequences, such as “O B-PER I-LOC I-ORG B-ORG O I-PER O” and “O B-PER I-ORG I-ORG B-ORG I-ORG I-ORG I-PER O,” can be excluded. It is worth noting that the number of possible ground-truth label sequences is decreased to 315, which is about half the size of the original number without pruning. This pruning process will be highly significant in obtaining the valid possible ground-truth label sequences.
C. Parameter Estimation

The objective function is defined as

$$\log p(Y|X; \theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} \log \sum_{z_i} p(y^k_i|z_i, x_i)p(z_i|x_i; \theta)$$

(15)

where $\theta$ represents the weight vector for the feature function.

The set of parameters $\{\alpha, \beta, \theta\}$ can be estimated by maximizing the above likelihood function. Since the ground-truth label sequence is missing, we employ the EM [27] to solve likelihood maximization, which is detailed as follows.

**Initialization:** Let $\alpha^k_{ij}$ denote the $j$th row of the $i$th layer of $\alpha^k$, we sample $\alpha^k_{ij}$ from a common Dirichlet prior with $\alpha^k \sim \text{Dir}(1)$. Similarly, $\beta^k_{ij}$ is sampled from the Dirichlet distribution with $\beta^k_{ij} \sim \text{Dir}(1)$. Furthermore, we randomly select an annotator and use the corresponding labelings as pseudo ground truths to initialize $\theta$.

**E-Step:** Compute $\tilde{p}(z_i)$ for each valid ground-truth label sequence $z_i$

$$\tilde{p}(z_i) = p(z_i| x_i, y_i) = \frac{\prod_{k=1}^{K} p(y^k_i|z_i, x_i)p(z_i|x_i)}{\sum_{z_i} \sum_{k=1}^{K} p(y^k_i|z_i, x_i)p(z_i|x_i; \theta)}.$$  

(16)

**M-Step:** Maximize $\sum_{i=1}^{N} \sum_{k=1}^{K} E_{\tilde{p}(z_i)}[\log p(y^k_i, z_i|x_i)]$. The expectation over $\log p(y^k_i, z_i|x_i)$ can be transformed into

$$\sum_{i=1}^{N} \sum_{k=1}^{K} E_{\tilde{p}(z_i)}[\log p(y^k_i|z_i, x_i)] = \sum_{i=1}^{N} E_{\tilde{p}(z_i)}[\log p(z_i|x_i)]$$

$$+ \sum_{i=1}^{N} \sum_{k=1}^{K} E_{\tilde{p}(z_i)}[\log p(y^k_i|x_i, z_i)].$$  

(17)

The parameter $\theta$ is obtained by maximizing $\sum_{i=1}^{N} E_{\tilde{p}(z_i)}[\log p(z_i|x_i)]$, which is expressed as

$$\sum_{i=1}^{N} E_{\tilde{p}(z_i)}[\log p(z_i|x_i)] = \sum_{i=1}^{N} \sum_{z_i} \tilde{p}(z_i) \log p(z_i|x_i)$$

$$= \sum_{i=1}^{N} \sum_{z_i} \tilde{p}(z_i)$$

$$\times \left[ \sum_{d=1}^{D} \theta_d F_d(x_i, z_i) - \log(Z(x_i)) \right].$$  

(18)

where the parameter optimization is similar to the basic CRF model that is described in Section III-B.

Then, maximum-likelihood parameter estimation for categorical distribution is employed to obtain the parameters $\{\alpha, \beta\}$.

V. EXPERIMENTS

In this section, we evaluate the performance of SA-SLC on the NLP datasets related with news and biomedicine by comparison with baselines and parameters’ sensitivity analysis (i.e., the number of annotators and annotators’ level of expertise). Apart from the four NLP datasets containing both gold and simulated crowdsourced label sequences, we also apply SA-SLC to the crowd dataset collected from Amazon’s Mechanical Turk (AMT) to investigate its effectiveness in a real-world application.

A. Datasets

**CoNLL 2003 [28]:** The dataset, consisting of over 21 000 sentences, is used for named entity recognition, which includes four types of named entities: 1) persons; 2) organizations; 3) locations; and 4) names of miscellaneous.

**BC2GM [29]:** This dataset consists of 20 000 sentences collected from biomedical publication abstracts and the annotations have two entity types (i.e., Gene or Protein).

**NCBI-Disease [30]:** The NCBI Disease corpus consists of 793 PubMed abstracts and is used to extract disease mentions. The abstract is fully annotated at the mention of diseases.

**JNLPBA [31]:** The Joint workshop on NLP in Biomedicine and its Applications corpus consists of 2404 publication abstracts, which is annotated for five entity types: 1) DNA; 2) RNA; 3) cell line; 4) cell type; and 5) protein.

The above datasets only contain gold label sequences. To simulate crowd labels from multiple annotators, we choose precision to present annotators’ reliability as accuracy is not the best measure for imbalanced datasets. Then, crowd labels are generated by controlling the average precision $p$ among $K$ annotators, where a correct named entity should exactly match the gold entity. The details of the four datasets are summarized in Table I.

| Dataset     | #sentences or abstracts | #labels |
|-------------|------------------------|--------|
| CoNLL 2003  | 6000                   | 3453   | 9     |
| BC2GM       | 8000                   | 5038   | 3     |
| NCBI-Disease| 400                    | 200    | 3     |
| JNLPBA      | 1000                   | 500    | 11    |

B. Baselines

We select multiple baselines that can be grouped into two types: 1) wrapper methods and 2) joint models. Wrapper methods directly infer the ground truths from crowds and then input these labels to a sequence labeling model, while joint models estimate the annotator’s expertise and the classifier simultaneously [32]. The baselines are described as follows.

**MVtoken [20]:** The ground-truth label sequence is obtained by choosing the label with more votes at the token level.

**David–Skene (DS) [6]:** The EM algorithm is employed to assign weight to each vote at the token level.

**MACE [33]:** By including a binary latent variable that denotes if and when each annotator is spamming, the model can identify which annotators are trustworthy and produce the true label.

**Sembler [24]:** The model focuses on crowd sequential labeling, which is the extension of the crowdsourcing problem on the instances that are independent of each other.
### TABLE II

| Model          | CoNLL 2003        | BC2GM            |
|---------------|-------------------|------------------|
| MVtoken       | 63.09 ± 4.70      | 59.06 ± 5.81     |
| DS            | 70.02 ± 4.35      | 58.33 ± 1.17     |
| MACE          | 64.21 ± 6.67      | 61.98 ± 0.71     |
| Sembler       | 63.71 ± 4.95      | 63.59 ± 1.16     |
| HMM-Crowd     | 69.17 ± 4.39      | 61.65 ± 1.25     |
| SA-SLC        | 69.74 ± 5.13      | 64.71 ± 1.07     |

*●/o denotes whether the performance of SA-SLC is statistically superior/inferior to the comparing methods (pairwise t-test at 5% significance level).*

### TABLE III

| Model          | NCBI-disease      | JNLPBA           |
|---------------|-------------------|------------------|
| MVtoken       | 73.75 ± 2.81      | 53.76 ± 1.84     |
| DS            | 72.79 ± 3.19      | 51.54 ± 0.95     |
| MACE          | 72.41 ± 2.55      | 56.26 ± 1.96     |
| Sembler       | 74.79 ± 3.25      | 54.27 ± 2.06     |
| HMM-Crowd     | 74.08 ± 3.16      | 54.18 ± 2.06     |
| SA-SLC        | 75.59 ± 2.78      | 58.14 ± 0.94     |

*●/o denotes whether the performance of SA-SLC is statistically superior/inferior to the comparing methods (pairwise t-test at 5% significance level).*

**HMM-Crowd [22]:** Based on HMMs, HMM-Crowd further models the “crowd component” by including additional parameters for measuring the reliability of annotators.

In the experiments, MVtoken, DS, and MACE are in the category of wrapper methods. To be fair, we choose CRFs as the subsequent sequence labeling model for these methods.

### C. Experimental Results

In this section, we follow the evaluation measure employed by CoNLL 2003 with:

\[
F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]  

Where precision is the percentage of named entities detected by the learning model while recall is the percentage of named entities present in the dataset that are detected by the learning model, and only a named entity that exactly matches the gold can be counted as the correct one.

First, we choose the number of annotators and annotating precision in the medium level to approximate the real setting of crowd annotations in AMT [20], where \( K = 5 \) and \( p = 0.5 \). Tables II and III report the mean and standard deviation of the \( F1 \) score with 10-fold cross-validation on the combination of training and development datasets. Furthermore, a pairwise \( t \)-test at a 5% significance level is recorded. We can observe that the proposed SA-SLC achieves the best performance in BC2GM, NCBI-disease, and JNLPBA datasets, and it significantly outperforms MACE, MVtoken, and Sembler in all selected datasets.

Then, to predict label sequences of test data, we repeat sampling crowd annotations three times with fixed \( p \) and record the average performance. Figs. 7–10 summarize the \( F1 \) score of each comparing method on a testing set of four datasets, where \( p \) varies from 0.9 to 0.1. The following observations can be made.

1) For the CoNLL 2003 dataset, the performance of each comparing method is significantly positively correlated with \( p \). When \( p \) is lower than 0.5, the reliability of crowd annotations is greatly reduced, which degrades the performance of the subsequent learning and prediction. It is worth noting that MVtoken and DS outperform the...
TABLE IV

CoNLL2003: the selected vector from $\alpha$ parameter matrix

| Ground-truth = “I-ORG” | “B-ORG” | “I-ORG” | “B-PER” | “I-PER” | “B-LOC” | “I-LOC” | “B-MISC” | “I-MISC” | “O” |
|-----------------------|---------|---------|---------|---------|---------|---------|---------|---------|-----|
| Previous annotation = “B-ORG” | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | $\alpha_7$ | $\alpha_8$ | $\alpha_9$ |

TABLE V

CoNLL2003: the selected vector from $\beta$ parameter matrix

| Ground-truth = “B-ORG” | “B-ORG” | “I-ORG” | “B-PER” | “I-PER” | “B-LOC” | “I-LOC” | “B-MISC” | “I-MISC” | “O” |
|-----------------------|---------|---------|---------|---------|---------|---------|---------|---------|-----|
| Same mention annotation = “B-ORG” | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $\beta_7$ | $\beta_8$ | $\beta_9$ |

Fig. 10. Performance of comparing methods on the JNLPBA dataset.

other comparing models in most cases but joint models (Sembler, HMM-Crowd, and SA-SLC) perform stably with the lower $p$. There is a certain degree of randomness in simulating annotators with different expertise. Medium level configuration (e.g., $p = 0.5$ and $p = 0.7$) may benefit the simple learning mechanism of MV and DS, which results in improving the quality of inferred ground truths.

2) As shown in Fig. 8, SA-SLC outperforms the other comparing methods on the BC2GM dataset, the joint models (i.e., SA-SLC, Sembler, and HMM-Crowd) for crowdsourced learning always achieves better results than the wrapper methods (e.g., MVtoken and DS). But there is no positive correlation between $p$ and $F_1$ score for most of the comparing methods. The major reason is that annotators have not many choices as the size of the label set for BC2GM is small and increasing annotators' precision does not significantly improve the quality of inferred ground truths.

3) For the NCBI-disease dataset, SA-SLC always achieves the best performance. The $F_1$ score obtained by the most of comparing methods is positively correlated with $p$. Compared to MVtoken and Sembler, SA-SLC performs stably with varying $p$.

4) For the JNLPBA dataset, as shown in Fig. 10, SA-SLC achieves comparable results with MVtoken and always outperforms the other two joint models (i.e., Sembler and HMM-Crowd). When $p$ is lower than 0.5, SA-SLC obtains better results than the other comparing methods. Furthermore, most of the comparing methods achieve a higher $F_1$ score by increasing $p$.

Meanwhile, we observe that SA-SLC performs considerably better on the dataset with a small label set (e.g., BC2GM) than that with various labels (e.g., CoNLL 2003). Figs. 11–14 demonstrate the estimated $\alpha$ and $\beta$ of five annotators for the datasets CoNLL 2003 and BC2GM, where $p$ is selected as 0.5. Considering the size of the parameter matrix of $\alpha$ and $\beta$, we present some parts of it. For example, for the dataset CoNLL 2003, as shown in Table I, there are nine labels. We choose one vector, which is used to describe the annotator’s expertise in capturing local label dependency with the previous annotation and the ground truth are fixed with “B-ORG” and “I-ORG,” respectively. The selected vectors from $\alpha$ and $\beta$ parameter matrices for the datasets CoNLL 2003 and BC2GM are listed in Tables IV–VII, where each entry represents annotator’s expertise in capturing internal dependency.
We can see from Figs. 11 and 12 that by aggregating crowd annotations, the probability of including the ground-truth label for model learning can be improved. Furthermore, annotators have limited choices with the smaller label set and then the obtained crowd annotations contain much less wrong and invalid label information, which greatly help improve the performance of SA-SLC.

Generally, the performance of crowdsourced learning methods is significantly correlated with annotators’ expertise. Aggregating the crowd annotations or inferring the ground-truth label from crowd annotations is less likely to achieve good results with lower label quality of crowds (i.e., lower $p$). Different from the wrapper methods (e.g., MVTOKEN and DS) that infer the ground-truth label for the subsequent classifier learning, SA-SLC, Sembler, and HMM-Crowd estimate the label quality and build the classifier simultaneously, which greatly reduces the negative effect of unreliable annotation. Compared to Sembler and HMM-Crowd, SA-SLC improves the estimation of the quality of label sequence by considering the annotator’s expertise in capturing local and nonlocal label dependencies. Furthermore, SA-SLC incorporates rich and valid label information by exploring label consistency among annotators, which helps improve the performance of learning from crowd annotations with much lower $p$.

We further investigate the influence of a number of annotators. By increasing $K$ to 10 and 15, we record the $F_1$ score of each comparing method and present the results in Tables VIII and IX. For most of the comparing methods,
increasing the number of annotators can help improve the performance. For example, the performance of SA-SLC on the CoNLL 2003 dataset ($p = 0.1$) improves by 7.40% and 16.77% with $K = 10$ and $K = 15$, respectively. In most cases, the performance of crowdsourced learning methods is positively correlated with the expertise level with specified $K$. For example, the average increase rate over varying $p$ of SA-SLC and HMM-Crowd in the BC2GM dataset with $K = 10$ is 8.15% and 8.91%, respectively. Furthermore, SA-SLC always outperforms the other joint models (i.e., Sembler and HMM-Crowd) by varying $K$. It should be noted that in some cases, MVtoken achieves competitive results compared to SA-SLC. For the models that directly infer the ground-truth label from crowd annotations, increasing $K$ relaxes the limitations of label space for inference, which may improve the quality of the inferred ground-truth label.

**D. Real-World Application in Amazon’s Mechanical Turk**

Rodrigues et al. [20] selected 400 news articles from the CoNLL 2003 shared NER task and put them on AMT to collect crowd labels. There are a total of 47 annotators and the average number of annotators per article is 4.9. In this article, after preprocessing these crowd-labeled data, we select 3000 sentence-level sequences for training and use CoNLL 2003 development and test data. Table X shows the results of predicting CoNLL 2003 test data. We can see that joint models significantly outperform the wrapper methods and SA-SLC achieves the highest recall and $F_1$ score. Compared to the results of Task 2 reported by Nguyen et al. [22], HMM-Crowd, then CRF (or LSTM) achieve better results than the joint model HMM-Crowd. Training the powerful sequence labeling model (e.g., CRF and LSTM) on the aggregated labels obtained from HMM-Crowd may improve the performance.

In the NER task, capturing internal label dependency is important for an annotator to provide consistent and reliable annotations. SA-SLC effectively utilizes two confusion matrices to measure the annotator’s expertise in capturing local and nonlocal label dependencies. Table XI presents a sentence with crowd annotation from AMT. It seems that the third annotator provides consistent annotation but two “Shanghai” are expected to be labeled as named entity “B-LOC.” This wrong label assignment can be accurately characterized by the two confusion matrices of SA-SLC that reflect the annotator’s reliability in assigning label “B-LOC” and considering nonlocal label dependency.

**VI. Conclusion**

In this article, we proposed SA-SLC to model crowd sequential annotations for sequence labeling. The proposed SA-SLC builds a conditional probabilistic model to jointly model sequential data and annotators’ expertise. By modeling the expert level of the annotator in capturing internal label dependency for sequential annotation, SA-SLC improves the estimation of the quality of the label sequence from crowd annotations, which greatly reduces the negative effect of unreliable annotations in the optimization. Furthermore, a VLS method was designed to accelerate the marginalization of SA-SLC, which further improves the quality of possible ground-truth label sequences. We conducted the experiments on four sequential datasets with synthetic crowd annotations. By varying expertise levels and the number of annotators, in most cases, SA-SLC performs better than the other comparing methods. In the future, we will consider more appropriate (or complex) annotator’s behavior in the sequential annotation.

**References**

[1] Y. Yin, D. Xu, X. Wang, and M. Bai, “Online state-based structured SVM combined with incremental PCA for robust visual tracking,” IEEE Trans. Cybern., vol. 45, no. 9, pp. 1988–2000, Sep. 2015.

[2] J. Zhang, X. Wu, and V. S. Sheng, “Learning from crowdsourced labeled data: A survey,” Artif. Intell. Rev., vol. 46, no. 4, pp. 543–576, 2016.

[3] R. Snow, B. O’Connor, D. Jurafsky, and A. Y. Ng, “Cheap and fast—But is it good?: Evaluating non-expert annotations for natural language tasks,” in Proc. Conf. Empir. Methods Nat. Lang. Process., 2008, pp. 254–263.

[4] V. C. Raykar et al., “Learning from crowds,” J. Mach. Learn. Res., vol. 11, pp. 1297–1322, Aug. 2010.

[5] Y. Yan et al., “Modeling annotator expertise: Learning when everybody knows a bit of something,” in Proc. 13th Int. Conf. Artif. Intell. Stat., 2010, pp. 932–939.

[6] A. P. Dawid and A. M. Skene, “Maximum likelihood estimation of observer error-rates using the EM algorithm,” J. Roy. Stat. Soc. C, Appl. Stat., vol. 28, no. 1, pp. 20–28, 1979.
[7] B. H. Juang and L. R. Rabiner, “Hidden Markov models for speech recognition,” Technometrics, vol. 33, no. 3, pp. 251–272, 1991.

[8] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in Proc. 18th Int. Conf. Mach. Learn., 2001, pp. 282–289.

[9] V. Krishnan and C. D. Manning, “An effective two-stage model for exploiting non-local dependencies in named entity recognition,” in Proc. 21st Int. Conf. Comput. Languag. 44th Annua. Meeting Assoc. Comput. Languag., 2006, pp. 1121–1128.

[10] J. Cheng, J. H. Park, J. Cao, and W. Qi, “Hidden Markov model-based nonfragile state estimation of switched neural network with probabilistic quantized outputs,” IEEE Trans. Cybern., vol. 50, no. 5, pp. 1900–1909, May 2020.

[11] S. Dong, Z.-G. Wu, Y.-J. Pan, H. Su, and Y. Liu, “Hidden-Markov-model-based asynchronous filter design of nonlinear Markov jump systems in continuous-time domain,” IEEE Trans. Cybern., vol. 49, no. 6, pp. 2297–2306, Jun. 2019.

[12] Z. Zhong, J. Li, D. A. Clausi, and A. Wong, “Generative adversarial networks and conditional random fields for hyperspectral image classification,” IEEE Trans. Cybern., vol. 50, no. 7, pp. 3318–3329, Jul. 2020.

[13] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, “A coupled HMM for audio-visual speech recognition,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., vol. 2. Orlando, FL, USA, 2002, pp. 2013–2016.

[14] Y. Zhu, N. M. Nayak, and A. K. Roy-Chowdury, “Context-aware activity modeling using hierarchical conditional random fields,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 7, pp. 1360–1372, Jul. 2015.

[15] S. Sarawagi and W. W. Cohen, “Semantic Markov random fields for information extraction,” in Advances in Neural Information Processing Systems. Vancouver, BC, Canada: Curran Assoc., Inc., 2005, pp. 1185–1192.

[16] X. Chu, W. Ouyang, H. Li, and X. Wang, “CRF-CNN: Modeling structured information in human pose estimation,” in Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran, 2016, pp. 316–324.

[17] Z. Huang, W. Xu, and K. Yu, “Bidirectional LSTM-CRF models for sequence tagging,” in Proc. Int. Joint Conf. Artificial Intell., vol. 50, no. 7, pp. 3318–3329, Jul. 2020.

[18] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, “Introduction to the bio-entity recognition task at JNLPBA,” in Proc. Int. Joint Workshop Nat. Lang. Process. Biomed. Appl., 2004, pp. 70–75.

[19] O. Lan, X. Huang, B. Y. Lin, H. Jiang, L. Liu, and X. Ren, “Learning to contextually aggregate multi-source supervision for sequence labeling,” 2019. [Online]. Available: arXiv:1910.04289.

[20] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy, “Learning whom to trust with MACE,” in Proc. Conf. North Amer. Ch. Assoc. Comput. Languag. Human Lang. Technol., 2013, pp. 1120–1130.

Xiaolei Lu received the B.S. degree in electrical engineering from the China University of Mining and Technology, Xuzhou, China, in 2014, and the M.S. degree in management science and engineering from Tongji University, Shanghai, China, in 2017. She is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong.

Her research interests include natural language processing and probabilistic graphical models.

Tommy W. S. Chow (Fellow, IEEE) received the B.Sc. (1st Hons.) and Ph.D. degrees from the Department of Electrical and Electronic Engineering, University of Sunderland, Sunderland, U.K., in 1984 and 1988, respectively.

He is currently a Professor with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong. He has authored and coauthored over 160 technical Journal articles related to his research, five book chapters, and one book. His main research areas include neural networks, machine learning, pattern recognition, and fault diagnosis.

Prof. Chow received the Best Paper Award in 2002 IEEE Industrial Electronics Society Annual meeting in Seville, Spain. He currently serves for the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS and Neural Processing Letters as an Associate Editor.