arXiv:2301.01579v1 [cs.CL] 4 Jan 2023

从众包序列标注中学习歧义

Xiaolei Lu [1]

[1] 香港城市大学

摘要

大多数众包学习方法将标注者之间的分歧视为噪声标注，而专家之间的相互分歧往往是自然语言固有歧义和不确定性的良好指标。本文提出了一个名为“从众包序列标注中学习歧义（LA-SCA）”的框架，用于探索可靠标注者之间的相互分歧，并有效保留易混淆标签信息。首先，我们构建了一个层次化贝叶斯模型，从众包群体中推断真实标注，并将可靠性相似的标注者归为一组。通过对标注者所在群体的规模、标注者的可靠性以及各序列中元素的无歧义性之间的关系进行建模，我们计算可靠标注者针对歧义元素的相互分歧，从而获得易混淆标签信息，并将其融入代价敏感序列标注。词性标注和命名实体识别任务上的实验结果表明，我们提出的框架在从众包群体中推断真实标注和预测未知序列方面表现出竞争力；对层次聚类结果的解读还有助于发现可靠性相似的标注者的标注模式。

引言

序列标注是指为观测到的序列数据分配标签序列，广泛应用于自然语言处理（NLP）任务，包括词性（POS）标注、组块分析和命名实体识别（NER）。许多下游自然语言处理应用（例如关系抽取和机器翻译）都能从这些基础自然语言处理任务的序列标签分配中获益。

传统的序列标注模型，如隐马尔可夫模型（HMM）和条件随机场（CRF），需要人工设计特征；要在特定数据集上取得良好效果，这些特征必须经过精心设计。在过去十年中，深度序列模型提升了序列标注的性能。例如，Bi-LSTM-CRF（Huang、Xu 和 Yu，2015）以及 Transformer（Vaswani 等，2017）。然而，这些序列标注模型需要大量带有精确标注的训练数据，而制作这类数据既耗时又费力。

近年来，成熟的商业众包平台（如 Amazon Mechanical Turk 和 CrowdFlower（Finin et al. 2010））蓬勃发展，成为获取大型带标签数据集的有效工具。众包利用群体智慧的贡献，但由于标注者的专业水平各不相同，众包标签的质量仍无法得到保证。因此，众包学习的重点主要在于估计标注者的可靠性，并基于估计得到的真实标注构建预测模型。例如，Snow 等人（Snow et al. 2008）采用偏差校正来整合非专家标注。Raykar 等人（Raykar et al. 2010）提出联合估计逻辑回归分类器的系数和标注者的专业水平。

许多有效模型，如 HMM-Crowd（Nguyen 等，2017年）和 Sembler（Wu、Fan 和 Yu，2012），将众包扩展到序列标注，从而更好地聚合众包序列标注。但这些方法是在只有一个真实标注的假设下衡量众包标签的质量。因此，标注者之间的分歧只能被视为噪声标注。然而，自然语言处理领域的研究表明，专家之间的相互分歧可能是语言固有歧义和不确定性的良好指标（Plank、Hovy 和 Søgaard，2014b）。显然，对于语言学上的难例，并没有明确的答案。如图1所示，“like”可以标注为连词或形容词。此外，专家之间的相互分歧还可能揭示易混淆标签信息，这些信息与难例在标签对上的分布有关。图2展示了词性标注任务中的标签混淆矩阵，其中“形容词”（形容词）和“名词”（名词）更容易混淆。将易混淆标签信息巧妙地纳入监督学习，可以增强分类器的鲁棒性（Plank、Hovy 和 Søgaard，2014b）。然而，现有的众包序列模型并未考虑标注者之间的相互分歧。

为探究可靠标注者之间的相互分歧，并有效保留易混淆标签信息，本文提出了一种名为“从众包序列标注中学习模糊性”（LA-SCA）的框架。我们的贡献总结如下：

首先，我们构建一个层次贝叶斯模型，将标注者分为不同的簇。通过对描述同一簇中标注者可靠性的混淆矩阵施加层次先验，该模型允许属于同一簇的标注者具有不同但相近的可靠性，旨在保留可靠标注者之间的相互分歧

Twitter data: …	How	to	make	Daily	is	out	…	
Annotations: …	ADV	**ADP**	VERB	**NOUN**	VERB	**ADJ**	…	
	ADV	**PRT**	VERB	**ADJ**	VERB	**ADV**		

Twitter data: …	like	one	of	your	French	fry	girls	…	
Annotations: …	**CONJ**	NOUN	ADP	DET	**ADJ**	NOUN	NOUN	…	
	ADJ	NOUN	ADP	DET	**NOUN**	NOUN	NOUN		

图1：由不同专家对 Twitter 词性标注数据进行双重标注的两个示例。

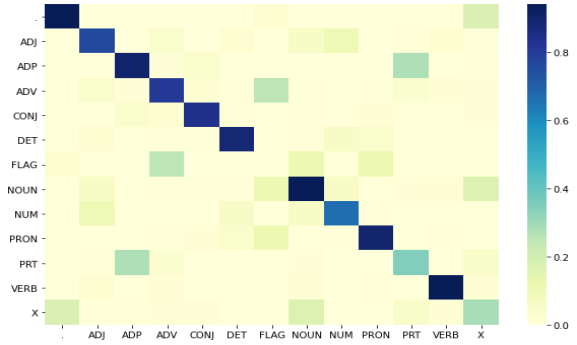


图2：从 500 条 Twitter 词性标注数据的两份金标准标注中得出的标签混淆矩阵。

标注者。

其次，我们构建一个低秩模型，用于刻画每个序列中标注者参与的群组规模、标注者可靠性与元素无歧义性之间的关系。随后，可以获得可靠标注者在歧义元素上的相互分歧，并据此计算标签混淆矩阵。

第三，将代价敏感机制与序列标注相结合，以促使包含真实标注的、更加易混淆的标签序列获得较低的汉明损失，从而提高序列模型的鲁棒性。

相关工作

隐马尔可夫模型（HMM）（Juang and Rabiner 1991; Qiao et al. 2015）和条件随机场（CRF）（Lafferty, McCallum, and Pereira 2001; Sarawagi and Cohen 2004）构成了序列标注中最常见的生成式-判别式模型组合。随着深度学习模型取得巨大成功，深度学习与图模型的结合日益受到关注（Chu et al. 2016）。例如，Bi-LSTM-CRF（Huang, Xu, and Yu 2015）结合双向 LSTM 网络和条件随机场层，旨在高效编码过去与未来的输入特征。此外，Transformer（Vaswani et al. 2017）引入注意力机制来学习长距离依赖，在效率和性能方面均取得了显著提升。然而，传统模型和深度学习模型都需要大量带有精确标注的训练数据，这不仅成本高昂，而且耗时极长。

将半监督学习引入序列标注模型（例如半监督条件随机场和半监督支持向量机），可以在一定程度上减轻序列标注的负担，但这种学习机制仍需要精确的标注。

众包为收集大规模标注数据集提供了一种有效工具。现有的众包学习模型可分为两类：包装模型和联合模型。前者使用从众包群体中推断出的真实标签进行后续的分类器学习，而联合模型则同时估计标注者的可靠性并学习预测模型。Dawid & Skene（DS）（Dawid 和 Skene, 1979年）聚合模型及其变体（例如 GLAD（Whitehill 等, 2009年））探讨了对真实标签、标注者的可靠性及相应标注之间关系进行建模的不同方法，然后采用期望最大化（EM）方法估计真实标签和标注者的可靠性。Se mbler（Wu、Fan 和 Yu，2012）和 HMM-Crowd（Nguyen 等，2017）模型旨在聚合多个标注，以学习序列模型。Simpson 和 Gurevych（2018年）进一步考虑了序列标注中的标签依赖，并采用贝叶斯方法对众包序列标签进行建模。需要指出的是，上述众包学习模型假设只有一个真实标签，且未考虑标注者之间的相互分歧。因此，这些模型无法捕捉自然语言处理任务中固有的语义歧义，也无法保留易混淆的标签信息。

为了探究标注者之间的相互分歧，Plank、Hovy 和 Søgaard（2014a）根据双重金标准标注推导出一个标签混淆矩阵，并表明，对易混淆标签信息敏感的词性标注分类器更加稳健。Sober´on 等人（2013）提出了 CrowdTruth 方法，用于建模语义解读中的歧义，并将标注者对模糊案例给出的所有可靠标注视为高质量标注。Dumitrache、Aroyo 和 Welty（2019）使用能够感知歧义的真实标签来训练开放域关系抽取分类器，结果表明，在标注质量和效率方面，能够感知歧义的众包群体优于专家。然而，基于 CrowdTruth 的模型只为模糊实例保留多个真实标签，却忽视了有助于稳健分类器学习的易混淆标签信息。

近年来，研究者提出了多标签众包方法，从众包群体中识别多标签任务的多个真实标签。与单标签众包中发现标注者之间的相互分歧不同，大多数多标签众包方法假设由一位标注者指定多个真实标签。例如，Zhang 和 Wu（2018）通过结合标签之间的相关性，扩展了生成式单标签众包方法；Li 等人（2018）则进一步利用邻居标注以及各标注者节省精力的标注行为，共同估计标注者专业水平和多标签分类器。此外，还有一些研究在假设可靠标注者共享相同标签相关性的前提下，探索多标签众包共识（Yu 等人，2020；Tu 等人，2020），但这种做法无法保留可靠标注者之间的相互分歧。

提出的框架

所提出的 LA-SCA 框架包含三个部分：通过对众包群体进行层次化建模，推断真实标签和可靠标注者；利用低秩模型，根据可靠标注者对歧义元素的相互分歧获取易混淆标签信息；在代价敏感序列标注中纳入标签混淆信息。具体内容如下：

群体标注的层次建模

令 $Y=\{y_{i1},y_{i2},...,y_{iL}\}_{i=1}^{N}$ 表示 L 位标注者针对 N 个实例提供的众包标注。每位标注者 l 均属于一个簇 $c\in\{1,2,...C\}$，并由混淆矩阵 $\Psi^{c_l}\in[0,1]^{T\times T}$ 表征，其中 T 表示 Y 的可能标签集大小。

我们假设同一簇中的标注者具有相似的可靠性，但他们给出的标注可能不同。例如，可靠性较低的标注者在标注特定实例时会给出不同的标注，而可靠的标注者对模糊实例则可能持有不同意见。为保留同一簇中标注者之间的分歧，我们对混淆矩阵 Ψ^{c_l} 的每一行采用如下层次化先验：

$$\eta_t^c\sim\text{Exponential}(\lambda_t),\eta_t^c>0, \tag{1}$$

$$\boldsymbol{\beta}_t^c\sim\text{Dirichlet}(\boldsymbol{\alpha}_t),0<\boldsymbol{\beta}_{tj}^c<1,\sum_{j=1}^{T}\boldsymbol{\beta}_{tj}^c=1, \tag{2}$$

$$\Psi_t^{c_l}\sim\text{Dirichlet}(\eta_t^c\boldsymbol{\beta}_t^c), \tag{3}$$

$$y_{il}\sim\text{Multinomial}(\Psi_{z_i}^{c_l}), \tag{4}$$

其中，c_l 表示标注者 l 属于簇 c，而 z_i 是 i_{th} 个实例的真实标签。η_t^c 和 $\boldsymbol{\beta}_t^c$ 分别可理解为 $\Psi_t^{c_l}$ 的精确率和均值。

此外，簇分配 c_l 和真实标签 z_i服从如下多项分布：

$$c_l\sim\text{Multinomial}(\boldsymbol{\nu}), \tag{5}$$

$$z_i\sim\text{Multinomial}(\boldsymbol{\gamma}), \tag{6}$$

其中，$\boldsymbol{\nu}$ 和 $\boldsymbol{\gamma}$ 分别从狄利克雷($\boldsymbol{\epsilon_\nu}$)和狄利克雷($\boldsymbol{\epsilon_\gamma}$)中采样。

我们采用折叠吉布斯抽样（Griffiths and Steyvers 2004年；Lakkaraju 等 2015）来估计关于隐变量 z_i 和 c_l 的条件分布（更多计算细节见技术附录 A）。令 $\boldsymbol{c}$ 和 $\boldsymbol{z}$ 分别表示簇分配和真实标签，$\boldsymbol{c}^{-l}$ 表示将标注者 l 排除在簇分配之外，而 $\boldsymbol{z}^{-i}$ 则排除实例 i_{th}。给定其余变量时，标注者 l的簇分配条件分布计算如下：

$$\begin{aligned}&p(c_l=c|\boldsymbol{c}^{-l},\boldsymbol{z},Y,\eta,\boldsymbol{\beta})\\&\propto p(c_l=c|\boldsymbol{c}^{-l})\times p(Y^l|Y^{-l},\boldsymbol{z},\boldsymbol{c},\eta,\boldsymbol{\beta})\\&\propto(n_c^{-l}+\boldsymbol{\epsilon_\nu}/C)\times\prod_t\frac{\Gamma(\eta_t^c)}{\Gamma(n_{lt}+\eta_t^c)}\prod_s\frac{\Gamma(n_{lts}+\eta_t^c\boldsymbol{\beta}_{ts}^c)}{\Gamma(\eta_t^c\boldsymbol{\beta}_{ts}^c)},\end{aligned} \tag{7}$$

其中，n_c^{-l}表示被分配到簇 c的标注者数量（不包括 l）。n_{lt} 表示由 l 标注且真实标签为 t的实例数量。n_{lts} 表示由 l 标注为 s 且真实标签为 t的实例数量。

类似地，$p(z_i=t|\boldsymbol{z}^{-i},Y,\boldsymbol{c},\eta,\boldsymbol{\beta})$表示为

$$\begin{aligned}&p(z_i=t|\boldsymbol{z}^{-i},Y,\boldsymbol{c},\eta,\boldsymbol{\beta})\\&\propto p(z_i=t|\boldsymbol{z}^{-i})\times p(y_i|Y^{-i},z_i=t,\boldsymbol{z}^{-i},\boldsymbol{c},\eta,\boldsymbol{\beta})\\&\propto(n_t^{-i}+\boldsymbol{\epsilon_\gamma}/T)\times\prod_c\prod_{c_l=c}\frac{\prod_s(n_{lts}^{-i}+\eta_t^c\boldsymbol{\beta}_{ts}^c)^{I(y_{il}=s)}}{n_{lt}^{-i}+\eta_t^c},\end{aligned} \tag{8}$$

其中，n_t^i 表示真实标签为 t的实例数（不包括 i）。n_{lt}^i 表示由 l 标注且真实标签为 t的实例数（不包括 i）。n_{lts} 表示由 l 标注为标签s 且真实标签为 t的实例数（不包括 i）。

由于似然函数 $p(Y|\boldsymbol{z},\boldsymbol{c},\eta,\boldsymbol{\beta})$的指数先验与狄利克雷先验不共轭，我们采用 Metropolis-Hastings（MH）算法（Chib and Greenberg 1995）估计每个簇的条件后验分布$p(\boldsymbol{\beta}_{tj}^c|\boldsymbol{\beta}_{t(\sim j)}^c,\eta_t^c,Y^c)$和 $p(\eta_t^c|\boldsymbol{\beta}_t^c,Y^c)$，并选用对称提议分布（即均匀分布）来模拟候选样本（算法详情见技术附录 C 的算法 1 和 2）。

$p(\boldsymbol{\beta}_{tj}^c|\boldsymbol{\beta}_{t(\sim j)}^c,\eta_t^c,Y^c)$ is given as

$$\begin{aligned}&p(\boldsymbol{\beta}_{tj}^c|\boldsymbol{\beta}_{t(\sim j)}^c,\eta_t^c,Y^c)\propto p(Y^{cj}|Y^{-cj},\boldsymbol{\beta}_t^c,\eta_t^c)\times p(\boldsymbol{\beta}_{tj}^c|\boldsymbol{\beta}_{t(\sim j)}^c)\\&\propto\prod_{c_l=c}\left\{\frac{\Gamma(n_{ltj}+\eta_t^c\boldsymbol{\beta}_{tj}^c)}{\Gamma(\eta_t^c\boldsymbol{\beta}_{tj}^c)}\frac{\Gamma(n_{ltt}+\eta_t^c\boldsymbol{\beta}_{tt}^c)}{\Gamma(\eta_t^c\boldsymbol{\beta}_{tt}^c)}\right\}\\&\times\left(\frac{\boldsymbol{\beta}_{tj}^c}{1-u_{tj}^c}\right)^{\lambda_t\boldsymbol{\alpha}_{tj}-1}\left(1-\frac{\boldsymbol{\beta}_{tj}^c}{1-u_{tj}^c}\right)^{\lambda_t\boldsymbol{\alpha}_{tt}-1},\end{aligned} \tag{9}$$

其中 $\boldsymbol{\beta}_t^c=[\boldsymbol{\beta}_{t1}^c,...,\boldsymbol{\beta}_{tj}^c,...,\boldsymbol{\beta}_{tT}^c]$。$u_{tj}^c=1-\sum_{s=1,\ s\neq t,\ s\neq j}^{T}\boldsymbol{\beta}_{ts}^c$。$\boldsymbol{\beta}_{tt}^c$ 的条件后验分布通过 $\boldsymbol{\beta}_{tt}^c=1-\sum_{j=1,\ j\neq t}^{T}\boldsymbol{\beta}_{tj}^c$ 得到。$p(\boldsymbol{\beta}_{tj}^c|\boldsymbol{\beta}_{t(\sim j)}^c,\ \eta_t^c,\ Y^c)$ 的推导细节见技术附录 B。

$p(\eta_t^c|\boldsymbol{\beta}_t^c,Y^c)$ 定义为

$$\begin{aligned}p(\eta_t^c|\boldsymbol{\beta}_t^c,Y^c)\propto&\prod_{c_l=c}\frac{\Gamma(\eta_t^c)}{\Gamma(n_{lt}+\eta_t^c)}\\&\times\prod_j\frac{\Gamma(n_{ltj}+\eta_t^c\boldsymbol{\beta}_{tj}^c)}{\Gamma(\eta_t^c\boldsymbol{\beta}_t^c)}\times\lambda_t e^{-(\lambda_t\eta_t^c)}.\end{aligned} \tag{10}$$

(Sabetpour 等，2021年)

通过迭代估计 z_i、c_l、$\boldsymbol{\beta}_t^c$ 和 η_t^c 直至收敛，可以将可靠性相似的标注者归入同一个簇。由于专家之间的相互分歧能够揭示语言歧义案例，为了识别包含可靠标注者的簇，我们根据估计得到的真实标注 $\boldsymbol{z}$计算每个簇的共享混淆矩阵。每个实例的一个真实标注

不会影响共享混淆矩阵的估计，因为模糊案例在整个数据集中所占比例很小（Plank、Hovy 和 Søgaard 2014a）。簇 c的共享混淆矩阵 $M^c \in \mathcal{R}^{T\times T}$中的元素定义为

$$M^c_{t,s} = \frac{\sum_i I(z_i = t)\prod_{l\in c} I(y_{il} = s)}{\sum_i I(z_i = t)}, \tag{11}$$

并通过 $\arg\max_c \sum_{t=1}^{T} M^c_{t,t} \Big/ T$ 获得高可靠性簇。

通过低秩模型识别歧义

基于已识别出的可靠标注者，为了估计序列中每个元素的歧义程度，我们假设在高可靠性簇中，标注者对歧义元素的判断会形成若干小组，而对于无歧义元素，在该序列标注上更可靠的标注者会与其他标注者保持一致。受用于描述小组规模、标注者可靠性与任务清晰度之间关系的定量公式（Tian 和 Zhu 2012）启发，我们为长度为 N^s 的 s_{th} 序列构建一个 $L \times N^s$ 矩阵A。在该矩阵中，每个条目 $A(l, N^s_j)$ 表示标注者 l 参与标注 s_{th} 序列中 j_{th} 元素时所属小组的规模。我们将 $A(l, N^s_j)$ 定义为

$$A(l, N^s_j) = \omega^s_l \times \mu^s_j, \tag{12}$$

其中， ω^s_l 表示标注者 l 在标注s_{th} 序列时的可靠性，而 μ^s_j 表示 j_{th} 元素的无歧义程度。

直观而言，如果标注者更可靠，或元素的歧义性更低，则组的规模会更大。因此，我们采用秩一分解来刻画 $A(l,\ N^s_j)$、ω^s_l和 μ^s_j之间的关系。 s_{th}序列中每个元素的无歧义程度计算如下:

$$A^s = U\Lambda V^T, \tag{13}$$

$$\boldsymbol{\omega}^s = U_{\cdot 1}\sqrt{\Lambda_{11}}, \tag{14}$$

$$\boldsymbol{\mu}^s = V_{\cdot 1}\sqrt{\Lambda_{11}}, \tag{15}$$

其中 $\boldsymbol{\omega}^s = [\omega^s_1, \omega^s_2, ..., \omega^s_L]$ 和 $\boldsymbol{\mu}^s = [\mu_1, \mu_2, ..., \mu_{N_s}]$。

识别歧 biguity:

a. 识别歧义元素。我们对整个序列数据的无歧义程度估计值进行排序，并选择适当的百分比 p，将处于最低 p 范围内的元素识别为模糊案例。

b. 计算标注者之间的相互分歧。对于识别出的歧义案例，可靠标注者之间的分歧会提供多个可能的真实标签。令 $\{y_{jt}\}_{t=1}^{t=L'}$ $(L' <= L)$ 表示标注者为 s_{th} 序列中的 j_{th} 歧义元素所分配的标签集合，则 y_{jt} 的得分可定义为

$$S(y_{jt}) = \frac{\sum_{l=1}^{L} I(y_{jl} = y_{jt})\omega^s_l}{\sum_{l=1}^{L} I(y_{jl} = y_{jt})}. \tag{16}$$

实际上，歧义实例的金标准标注数量有限（Plank、Hovy 和 Søgaard 2014b）。我们按 $S(y_{jt})$降序为每个歧义元素选出排名前两位的标签，并将它们与分层建模中推断出的真实标注结合起来。

c. 获取易混淆标签信息。标签混淆矩阵 $CF \in \mathcal{R}^{T\times T}$ 用于表示标签对之间的混淆程度，其中条目 $CF(i,\ j)$ 定义为 $p(z(x) = i,\ z(x) = j)$ 与 $p(z(x) = j,\ z(x) = i)$ 的均值，其中 $p(z(x) = i,\ z(x) = j)$ 的计算方式如下

$$p(z(x) = i, z(x) = j) = \frac{\sum_k I(z(x_k) = i, z(x_k) = j)}{\sum_k I(z(x_k) = i)}, \tag{17}$$

其中， k 表示整个序列数据集中的 k_{th} 元素，而 $p(z(x) = j,\ z(x) = i)$的计算方式类似。

代价敏感序列标注

给定 $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}_{i=1}^{N}$ 序列数据集，其中 $\boldsymbol{z}_i$ 是通过层次贝叶斯建模推断得到的真实标签。传统的训练准则是最大化条件对数线性模型的似然；该准则不会区分真实标注与所有错误输出，而是通过归一化对它们施加相同的惩罚。为了改进序列标注，我们采用代价敏感机制，在训练中纳入易混淆标签信息；与真实标注越易混淆的标签序列，其代价越低。代价敏感序列标注的目标定义为

$$L_{\mathrm{CS}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \frac{\exp\left\{\boldsymbol{\theta}^T \boldsymbol{f}(\boldsymbol{x}_i, \boldsymbol{z}_i)\right\}}{\sum_{\boldsymbol{z}_j} \mathrm{cost}(\boldsymbol{z}_j, \boldsymbol{z}_i)\exp\left\{\boldsymbol{\theta}^T \boldsymbol{f}(\boldsymbol{x}_i, \boldsymbol{z}_j)\right\}}, \tag{18}$$

其中， $\boldsymbol{f}(\boldsymbol{x}_i,\ \boldsymbol{z}_i)$表示特征函数。$\mathrm{cost}(\boldsymbol{z}_j,\ \boldsymbol{z}_i)$用于衡量易混淆标签信息对损失的影响。加权汉明损失定义如下，用于描述$\mathrm{cost}(\boldsymbol{z}_j,\ \boldsymbol{z}_i)$:

$$\mathrm{cost}(\boldsymbol{z}_j, \boldsymbol{z}_i) = \frac{1}{K_i}\sum_{k=1}^{K_i} (1 - p(\boldsymbol{z}_{jk}, \boldsymbol{z}_{ik})) * (\boldsymbol{z}_{jk} \oplus \boldsymbol{z}_{ik}), \tag{19}$$

其中， K_i 是 i_{th} 序列中的词元数量。 $\oplus$ 是异或布尔运算符， $p(\boldsymbol{z}_{jk},\ \boldsymbol{z}_{ik})$由标签混淆矩阵得出。

实验

我们针对英语词性标注和命名实体识别开展了实验。词性分析中存在许多标注者意见不一的难例，因此一直备受争议（Plank、Hovy 和 Søgaard，2014b）；而在命名实体识别中，命名实体的定义和划分仍存在争议。接下来的章节将展示定量结果，以考察我们的框架在推断真实标签、预测未知序列以及保留易混淆标签信息方面的有效性。

数据集

目前已发表的数据集无法同时满足众包标注和多重金标准标注的需求。我们采用了以下多重金标准标注和众包标注数据集：

词性标注：大多数词性标注数据集只包含一个金标准标注，因此无法识别难例。因此，我们使用了一项研究成本敏感型词性标注器的工作（Dumitrache、Aroyo 和 Welty 2019）中的三个 Twitter词性标注数据集，包括 500 条带有双重金标准标注的推文（为简便起见，记为 T-DGA）、RITTER-TEST 数据集（118 条推文）和 INHOUSE 数据集（200 条推文）。我们将 T-DGA 用作训练数据（双重金标准标注保证了难例的存在），并将 RITTER-TEST 和 INHOUSE 用作测试数据集[1]。

命名实体识别：CoNLL-2003 共享命名实体识别数据集（Sang 和 De Meulder 2003）是自然语言处理领域用于序列标注的最常见基准之一，包含四类实体：人物（PER）、地点（LOC）、组织（ORG）和其他类别（MISC）。Rodrigues 等（Rodrigues、Pereira和Ribeiro 2014）将 CoNLL-2003 中的 400 篇文章发布到亚马逊Mechanical Turk，以收集众包标注。共有 47 位标注者，每篇文章的平均标注者人数为 4.9。本文对这些众包标注数据进行预处理后，选取了 3000 个句子级序列，并使用 CoNLL 2003 测试数据[2]。

基线方法

我们使用以下六种模型从众包序列数据中学习，作为基线方法。

MVtoken（Sang 和 De Meulder 2003年）：通过在词元级选择得票较多的标签，获得真实标注序列。

DS（Dawid 和 Skene 1979）：采用期望最大化算法在词元级为每张投票赋予权重。

MACE（Hovy 等 2013）：通过引入一个二元潜变量来表示每位标注者是否以及何时在敷衍标注，该模型能够识别哪些标注者值得信赖，并生成真实标签。

Sembler（Wu、Fan 和 Yu 2012）：该模型将实例级的众包课程学习扩展到序列级，并联合估计标注者的可靠性和序列模型。

HMM-Crowd（Nguyen 等人，2017）：该模型基于隐马尔可夫模型，并通过纳入标注者标签质量和众包变量的参数，进一步对“众包组件”进行建模。

HC-CLL：为验证成本敏感序列标注的有效性，我们还通过最大化条件对数似然来训练序列预测模型。

实验设置

合成众包标注：由于 T-DGA 没有真实的众包标注，我们通过控制标注的精确率来模拟可靠性各不相同的标注者。在实际应用中，标注者的数量有限，因此我们将标注者总数设为 15，并安排三种不同的分配方案：$[5,5,5]$, $[8,4,3]$ 和 $[3,4,8]$。在每种方案中，设置三个不同的精确率范围：$[0.9,0.7]$、$[0.7,0.4]$ 和$[0.4,0.1]$，以表示从高到低的不同可靠性水平。

LA-SCA 框架：根据贝叶斯信息准则，在范围 $[2,5]$ 内为标注者选择最佳簇数。λ_t 设为 2。为确认众包标注优于随机标注，将 $\boldsymbol{\alpha}_t$ 的对角元素设为 0.7，非对角元素设为 0.3。此外，我们选择 $p = 10\%$ 来识别歧义元素。

实验结果

与基线方法比较 我们评估了所提出框架在推断训练数据的真实标签和预测测试集方面的有效性。

词性标注任务：

对于 implicit 词性标注，将三种不同的众包标注$[8,4,3]$, $[5,5,5]$、$[3,4,8]$ 分别记为 ca_1,、ca_2 和 ca_3,。表1展示了在 T-DGA 数据集上推断真实标签的准确率（HC-CLL 在推断真实标签方面与 LA-SCA 相同）。可以看出，大多数众包模型在高质量标注比例增加时表现更好。由于 Gold 1 和 Gold 2 对某些词元的标签分配不同，各比较模型在这两种金标准标注下的表现也有所不同。对于低质量标注的情况（即 ca_3），所提出的分层贝叶斯模型能够有效识别可靠性较高的标注者，从而有助于指导真实标签的估计并提升性能。DS 和 HMM-Crowd 也取得了有竞争力的结果，因为它们通过迭代估计标注者的可靠性和真实标签，减轻了低质量标注带来的负面影响。

表1：推断 T-DGA 数据集真实标签的准确率（%）。

模型	金标准1			金标准2		
	ca1	ca2	ca3	ca1	ca2	ca3
MVtoken	91.82	90.84	83.70	81.94	81.81	73.20
DS	93.34	90.97	91.99	83.53	**81.90**	82.14
MACE	89.80	84.79	84.56	80.76	76.26	75.89
Sembler	93.05	89.58	85.78	83.36	80.22	76.99
HMM-Crowd	**93.40**	91.59	90.38	**83.72**	82.06	81.12
LA-SCA	93.30	**92.59**	**92.22**	83.47	81.73	**83.71**

表 2 报告了各对比方法在 RITTER-TEST 和内部数据集上的 F1 分数。总体而言，从质量更高的真实标签中学习的模型能够取得更好的预测性能。对于将推断出的真实标签输入序列模型的包装模型（即 MVtoken、DS 和 MACE），其预测性能很大程度上取决于估计出的真实标签的质量。因此，在 ca_3 设置中，包装模型（即 MVtoken、DS 和 MACE）的 F1 分数低于联合模型（即 Sembler 和 HMM-Crowd）。所提出的层次化

[1] RITTER-TEST 和 INHOUSE 都只有一个金标准标注。[2] CoNLL 2003 测试集只有一个金标准标注。

贝叶斯模型 HC-CLL 能够有效识别可靠性高的簇，从而在处理低质量标注时保持稳定的性能。与 HC-CLL 相比，LA-SCA 在 ca_1 和 ca_2 设置下表现更好，因为低质量的众包标注（即 ca_3）无法提供有效的易混淆标签信息，这更有可能在代价敏感序列标注中引入大量噪声，进而降低预测性能。

表2：各模型在 RITTER-TEST 和内部数据集上的表现（%）。

模型	RITTER- TEST			INHOUSE		
	ca_1	ca_2	ca_3	ca_1	ca_2	ca_3
MVtoken	59.35	58.72	58.72	53.15	52.29	48.03
DS	**67.58**	60.43	58.69	54.33	48.49	48.14
MACE	59.89	58.04	60.05	47.92	53.33	49.12
Sembler	59.82	61.74	60.53	49.65	50.22	49.93
HMM-Crowd	61.10	61.30	61.79	49.98	49.12	52.40
HC-CLL	61.18	65.12	**62.57**	52.97	54.55	**52.88**
LA-SCA	66.20	**67.32**	61.57	**55.65**	**57.25**	50.84

命名实体识别任务：

在命名实体识别标注任务中，类别“O”占总类别的很大比例，因此我们使用 F1 分数而非准确率来报告 CoNLL 2003 命名实体识别任务训练数据的真实标签推断性能。如表3所示，所提出的层次贝叶斯模型取得了最高的 F1 分数，DS模型也取得了有竞争力的结果。表3还展示了预测测试数据标签的性能。由于众包训练数据有限，各比较方法的整体性能远低于已报告的结果（Rodrigues、Pereira和Ribeiro 2014年）。尽管如此，所提出的 LA-SCA 框架仍优于基线方法，但优势较小。由于在命名实体识别任务中存在一些容易混淆且需要相互区分的标签，代价敏感学习机制不可避免地会产生标签噪声，因此直接最大化对数似然也可以与代价敏感最大化相媲美。

表3：CoNLL 2003 命名实体识别任务上的评估（%）。

模型	推断真实标签	预测
MVtoken	63.17	38.52
DS	65.32	39.21
MACE	60.07	37.10
Sembler	63.25	38.87
HMM-Crowd	63.44	39.31
HC-CLL	**67.54**	40.56
LA-SCA	**67.54**	**41.56**

识别模糊案例在本节中，我们研究 LA-SCA 在识别模糊案例和保留易混淆标签信息方面的表现。我们展示了 T-DGA 数据集（ca_1 设置）上的结果，因为该数据集提供了比较基准。

首先，我们使用以下指标衡量识别模糊

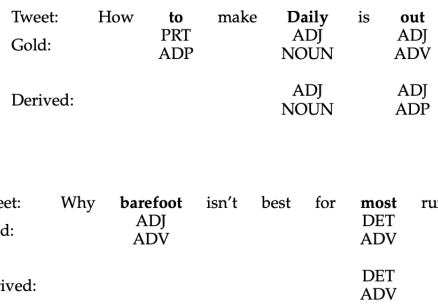


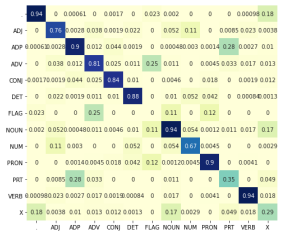
图 3：T-DGA 中两个关于歧义案例的示例，包含金标准标注和派生标注。

案例的性能：

$$\mathrm{acc}_1 = \frac{\#\text{correctly identified ambiguous cases}}{\#\text{all ambiguous cases}}, \quad (20)$$

$$\mathrm{acc}_2 = \frac{\#\text{correctly double annotated ambiguous cases}}{\#\text{all ambiguous cases}}, \quad (21)$$

并且我们得到 $\mathrm{acc}_1 = 725/931 = 0.779$ 和 $\mathrm{acc}_2 = 614/931 = 0.660$。由此可以得出，LA-SCA 成功识别了 T-DGA 中的大多数模糊案例。我们进一步展示了 T-DGA 中两个关于模糊案例的示例，包括金标准标注和推导出的标注，如图 3 所示。LA-SCA 成功识别了标签易混淆对 [“形容词”、“名词”] 和 [“限定词”、“副词”]所对应的模糊案例。



(a) 金标准矩阵。



(b) 派生矩阵（ca_1）。



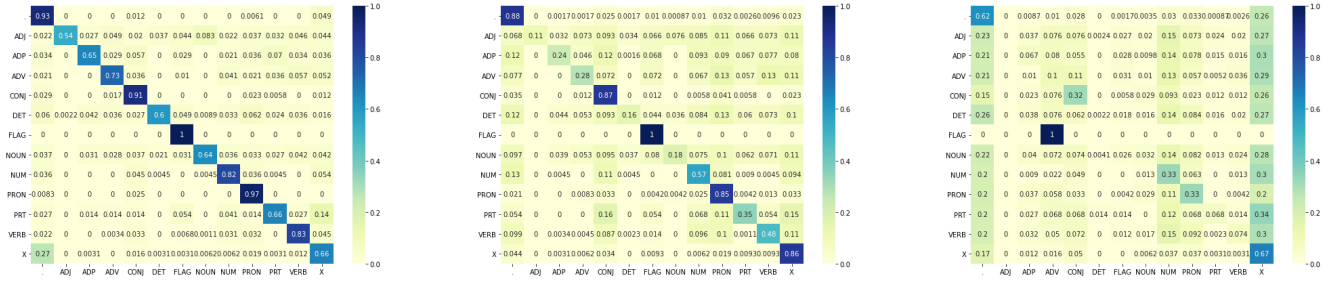
(c) 派生矩阵（ca_2）。



(d) 派生矩阵（ca_3）。

图4：T-DGA 上金标准与派生标签混淆矩阵的比较。

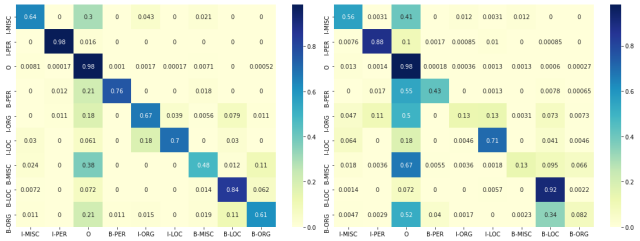
图4分别展示了金标准标签混淆矩阵，以及三种设置下推导出的混淆矩阵。由于



(a) 簇 1　　(b) 簇 2　　(c) 簇 3

图 5：词性标注任务中三个簇的共享混淆矩阵。

众包群体包含噪声标签信息，因此某些标签的一致性低于金标准，这可能会产生错误的易混淆标签信息。例如，在 Twitter 推文“FollowerSale is mostTrusted company to buy”中，“to”的金标准标注是 [“VERB”]，而推导出的易混淆标签集是 [“VERB”、“PRON”]。不过，推导出的矩阵也保留了一些与金标准相似的易混淆标签信息。例如，形容词 [“ADJ”]与名词 [“NOUN”],以及 [“X”]类别之间的一致性较低，后者更容易与标点符号 [“.”] 和名词 [“NOUN”]混淆。



(a) 簇 1。　　(b) 簇 2。

图 6：命名实体识别任务中各簇共享的混淆矩阵。

解读簇 对众包标注进行聚类，有助于发现可靠性相似的标注者所共有的模式。这里，我们演示如何解读在词性标注和命名实体识别任务中估计得到的各簇共享混淆矩阵。

词性标注任务：

我们选择 ca_1 设置，并展示三个簇共享的混淆矩阵。通过查看图 5 中三个共享混淆矩阵的对角线元素，可以看出，所开发的贝叶斯分层模型能够很好地区分可靠性不同的标注者。簇 1 中的标注者可靠性较高，平均成功识别值超过 0.7；而可靠性较低的标注者则被归入第三个簇，其平均成功识别值低于 0.3。

命名实体识别任务：

我们展示了从亚马逊土耳其机器人收集的众包标注的聚类结果。总体而言，这些众包标注的质量较好。从图 6 可以看出，在标注“I-PER”“O”“I-LOC”和“B-ORG”时，这两个簇都具有较高的可靠性。簇 1 的标注更可靠，其平均对角线值为 0.742，而簇 2 的平均对角线值为 0.535。

讨论

LA-SCA 评估的主要问题在于，它需要在同时包含众包标注和多个金标准标注的数据集上进行。对于 T-DGA，我们必须模拟众包标注，并用精确率从全局视角衡量标注者的可靠性。这会导致一种情况：在精确率计算中，简单案例更有可能被错误标注，

$[0.9, 0.7]$，这与可靠标注者的决策不符。在众包标注的 CoNLL-2003 命名实体识别任务中，每篇文章分配的标注者数量有限，而且众包标注质量较高，这在一定程度上阻碍了对标注多样性的探索。

尽管数据集不完整在一定程度上限制了 LA-SCA 的适用性，但所提出的层次贝叶斯建模在从可靠性较低的真实众包标注和合成众包中推断真实标签方面展现出了竞争力。由于代价敏感机制预期标签混淆矩阵在任务（如命名实体识别）中是稀疏的，即只有少数标签之间更容易混淆，因此如何显著提升对未知序列的预测效果仍有待探索。

结论

本文提出了一种名为从众包序列标注中学习歧义（LA-SCA）的框架，用于探索可靠标注者之间的相互分歧，并有效保留易混淆标签信息，以改进鲁棒的序列分类器学习。实验结果表明，LA-SCA 在从众包群体推断真实标注和预测测试集方面取得了具有竞争力的性能。此外，所识别出的簇有助于解释具有相似可靠性的标注者的标注模式，从而帮助任务设计者改进标注指南。

A. 估计关于隐变量 z_i 和 c_l的条件分布

a. $p(c_l = c|\boldsymbol{c}^{-l}$、 $\boldsymbol{z}$、 Y、 η、 $\boldsymbol{\beta})$ 的推导如下:

$$p(c_l = c|\boldsymbol{c}^{-l}, \boldsymbol{z}, Y, \eta, \boldsymbol{\beta}) \propto p(c_l = c|\boldsymbol{c}^{-l}) \times p(Y^l|Y^{-l}, \boldsymbol{z}, \boldsymbol{c}, \eta, \boldsymbol{\beta}), \tag{22}$$

其中， $p(c_l = c|\boldsymbol{c}^{-l})$的计算方式为

$$\begin{aligned}
&p(c_l = c|\boldsymbol{c}^{-l}) \\
&= p(c_1, c_2, .., c_l = c)/p(c_1, c_2, ..., c_{l-1}) \\
&= \frac{\Gamma(\epsilon_\nu)}{\Gamma(L+\epsilon_\nu)} \prod_{c=1}^{C} \frac{\Gamma(n_c+\epsilon_\nu/C)}{\Gamma(\epsilon_\nu/C)} \Big/ \frac{\Gamma(\epsilon_\nu)}{\Gamma(L+\epsilon_\nu-1)} \prod_{c=1}^{C} \frac{\Gamma(n_c^{-l}+\epsilon_\nu/C)}{\Gamma(\epsilon_\nu/C)} \\
&= \frac{\Gamma(L+\epsilon_\nu-1)}{\Gamma(L+\epsilon_\nu)} \frac{\Gamma(n_c+\epsilon_\nu/C)}{\Gamma(n_c^{-l}+\epsilon_\nu/C)} \\
&= \frac{n_c^{-l}+\epsilon_\nu/C}{L+\epsilon_\nu-1}.
\end{aligned} \tag{23}$$

似然 $p(Y|\boldsymbol{z}, \boldsymbol{c}, \eta, \boldsymbol{\beta}$通过对变量 Ψ^{cl}积分消去后得到:

$$\begin{aligned}
&p(Y|\boldsymbol{z}, \boldsymbol{c}, \eta, \boldsymbol{\beta}) \\
&= \prod_c \prod_{l\in c} \prod_t \int p(Y^{c_l}|\Psi_t^{c_l}) p(\Psi_t^{c_l}|\eta^c \boldsymbol{\beta}_t^c) d(\eta^c \boldsymbol{\beta}_t^c) \\
&= \prod_c \prod_{l\in c} \prod_t \left[\frac{\Gamma(\eta_t^c)}{\Gamma(n_{lt}+\eta_t^c)} \prod_s \frac{\Gamma(n_{lts}+\eta_t^c \boldsymbol{\beta}_{ts}^c)}{\Gamma(\eta_t^c \boldsymbol{\beta}_{ts}^c)} \right].
\end{aligned} \tag{3}$$

然后，通过贝叶斯法则得到条件分布:

$$p(Y^l|Y^{-l}, \boldsymbol{z}, \boldsymbol{c}, \eta, \boldsymbol{\beta}) = \prod_t \left[\frac{\Gamma(\eta_t^c)}{\Gamma(n_{lt}+\eta_t^c)} \prod_s \frac{\Gamma(n_{lts}+\eta_t^c \boldsymbol{\beta}_{ts}^c)}{\Gamma(\eta_t^c \boldsymbol{\beta}_{ts}^c)} \right]. \tag{4}$$

b. $p(z_i = t|\boldsymbol{z}^{-i}, Y, \boldsymbol{c}, \eta, \boldsymbol{\beta})$的推导如下:

$$\begin{aligned}
&p(z_i = t|\boldsymbol{z}^{-i}, Y, \boldsymbol{c}, \eta, \boldsymbol{\beta}) \\
&\propto p(z_i = t|\boldsymbol{z}^{-i}) \times p(y_i|Y^{-i}, z_i = t, \boldsymbol{z}^{-i}, \boldsymbol{c}, \eta, \boldsymbol{\beta}),
\end{aligned} \tag{5}$$

其中， $p(z_i|\boldsymbol{z}^{-i})$的计算方式为

$$\begin{aligned}
&p(z_i = t|\boldsymbol{z}^{-i}) \\
&= p(z_1, z_2, .., z_i = t)/p(z_1, z_2, ..., z_{i-1}) \\
&= \frac{\Gamma(\epsilon_\gamma)}{\Gamma(N+\epsilon_\gamma)} \prod_{t=1}^{T} \frac{\Gamma(n_t+\epsilon_\gamma/T)}{\Gamma(\epsilon_\gamma/T)} \Big/ \frac{\Gamma(\epsilon_\gamma)}{\Gamma(N+\epsilon_\gamma-1)} \prod_{t=1}^{T} \frac{\Gamma(n_t^{-i}+\epsilon_\gamma/T)}{\Gamma(\epsilon_\gamma/T)} \\
&= \frac{\Gamma(N+\epsilon_\gamma-1)}{\Gamma(N+\epsilon_\gamma)} \frac{\Gamma(n_t+\epsilon_\gamma/T)}{\Gamma(n_t^{-i}+\epsilon_\gamma/T)} \\
&= \frac{n_t^{-i}+\epsilon_\gamma/T}{N+\epsilon_\gamma-1},
\end{aligned} \tag{6}$$

并且 $p(y_i|Y^{-i}, \ z_i = t, \ \boldsymbol{z}^{-i}, \ \boldsymbol{c}, \ \eta, \ \boldsymbol{\beta})$ 可表示为

$$\begin{aligned}
&p(y_i|Y^{-i}, z_i = t, \boldsymbol{z}^{-i}, \boldsymbol{c}, \eta, \boldsymbol{\beta}) \\
&= \prod_c \prod_{l\in c} \left[\frac{\Gamma(\eta_t^c)}{\Gamma(n_{lt}+\eta_t^c)} \prod_s \frac{\Gamma(n_{lts}+\eta_t^c \boldsymbol{\beta}_{ts}^c)}{\Gamma(\eta_t^c \boldsymbol{\beta}_{ts}^c)} \Big/ \frac{\Gamma(\eta_t^c)}{\Gamma(n_{lt}^{-i})+\eta_t^c)} \prod_s \frac{\Gamma(n_{lts}^{-i})+\eta_t^c \boldsymbol{\beta}_{ts}^c)}{\Gamma(\eta_t^c \boldsymbol{\beta}_{ts}^c)} \right] \\
&= \prod_c \prod_{l\in c} \frac{\prod_s (n_{lts}^{-i} + \eta_t^c \boldsymbol{\beta}_{ts}^c)^{I(y_{il}=s)}}{n_{lt}^{-i} + \eta_t^c}.
\end{aligned} \tag{7}$$

B

a. $p(\boldsymbol{\beta}_{tj}^c|\boldsymbol{\beta}_{t(\sim j)}^c, \eta_t^c, Y^c)$ 的推导如下:

$$p(\boldsymbol{\beta}_{tj}^c|\boldsymbol{\beta}_{t(\sim j)}^c, \eta_t^c, Y^c) \propto p(Y^{cj}|Y^{-cj}, \boldsymbol{\beta}_t^c, \eta_t^c) \times p(\boldsymbol{\beta}_{tj}^c|\boldsymbol{\beta}_{t(\sim j)}^c). \tag{8}$$

首先， $\boldsymbol{\beta}_t^c$服从狄利克雷分布:

$$\boldsymbol{\beta}_t^c|\boldsymbol{\alpha}_t \sim \text{Dirichlet}(\boldsymbol{\alpha}_t), 0 < \boldsymbol{\beta}_{tj}^c < 1, \sum_{j=1}^{T} \boldsymbol{\beta}_{tj}^c = 1, \tag{9}$$

其具体形式如下:

$$(\boldsymbol{\beta}_{t1}^c, ..., \boldsymbol{\beta}_{tj}^c, ..., \boldsymbol{\beta}_{tT}^c) \sim \text{Dirichlet}\left(\lambda_t \boldsymbol{\alpha}_{t1}, ..., \lambda_t \boldsymbol{\alpha}_{tj}, ..., \lambda_t \boldsymbol{\alpha}_{tT}\right). \tag{10}$$

根据聚合性质，我们得到

$$(\boldsymbol{\beta}_{tj}^c, u_{tj}^c, \boldsymbol{\beta}_{tt}^c = 1 - \boldsymbol{\beta}_{tj}^c - u_{tj}^c,) \sim \text{Dirichlet}\left(\lambda_t \boldsymbol{\alpha}_{tj}, a_j, \lambda_t \boldsymbol{\alpha}_{tt}\right), \tag{11}$$

其中

$$u_{tj}^c = 1 - \sum_{s=1, s\neq t, s\neq j}^{T} \boldsymbol{\beta}_{ts}^c, \tag{12}$$

$$a_j = \sum_{s=1, s\neq t, s\neq j}^{T} \lambda_t \boldsymbol{\alpha}_{ts}. \tag{13}$$

联合概率 $p(\boldsymbol{\beta}_{tj}^c, \ u_{tj}^c)$ 如下

$$\begin{aligned}
p(\boldsymbol{\beta}_{tj}^c, u_{tj}^c) &= \prod_{l\in c} \frac{\Gamma(\sum_{j=1}^{T} \lambda_t \boldsymbol{\alpha}_{tj})}{\Gamma(\lambda_t \boldsymbol{\alpha}_{tj}) \Gamma(a_j) \Gamma(\lambda_t \boldsymbol{\alpha}_{tt})} \\
&\times (\boldsymbol{\beta}_{tj}^c)^{\lambda_t \boldsymbol{\alpha}_{tj}-1} (u_{tj}^c)^{a_j-1} (1 - \boldsymbol{\beta}_{tj}^c - u_{tj}^c)^{\lambda_t \boldsymbol{\alpha}_{tt}-1}.
\end{aligned} \tag{14}$$

边缘概率 $p(u_{tj}^c|\eta_t^c, Y^c)$ 计算为

$$\begin{aligned}
p(u_{tj}^c) =& \frac{\Gamma(\sum_{j=1}^{T} \lambda_t \boldsymbol{\alpha}_{tj})}{\Gamma(a_j) \Gamma(\lambda_t \boldsymbol{\alpha}_{tt} + \lambda_t \boldsymbol{\alpha}_{tj})} \\
&\times (u_{tj}^c)^{a_j-1} (1 - u_{tj}^c)^{\lambda_t \boldsymbol{\alpha}_{tt} + \lambda_t \boldsymbol{\alpha}_{tj}-1}.
\end{aligned} \tag{15}$$

则条件分布 $p(\boldsymbol{\beta}_{tj}^c|\boldsymbol{\beta}_{t(\sim j)}^c)$ 为

$$p(\boldsymbol{\beta}_{tj}^c|\boldsymbol{\beta}_{t(\sim j)}^c) \propto \left(\frac{\boldsymbol{\beta}_{tj}^c}{1 - u_{tj}^c} \right)^{\lambda_t \boldsymbol{\alpha}_{tj}-1} \left(1 - \frac{\boldsymbol{\beta}_{tj}^c}{1 - u_{tj}^c} \right)^{\lambda_t \boldsymbol{\alpha}_{tt}-1}. \tag{16}$$

根据附录 A 中的公式 (3)，可以很容易地得到

$$p(Y^{cj}|Y^{-cj}, \boldsymbol{\beta}_t^c, \eta_t^c) = \prod_{l\in c} \left\{ \frac{\Gamma(n_{ltj}+\eta_t^c \boldsymbol{\beta}_{tj}^c)}{\Gamma(\eta_t^c \boldsymbol{\beta}_{tj}^c)} \frac{\Gamma(n_{ltt}+\eta_t^c \boldsymbol{\beta}_{tt}^c)}{\Gamma(\eta_t^c \boldsymbol{\beta}_{tt}^c)} \right\}. \tag{17}$$

最终我们得到

$$\begin{aligned}
&p(\boldsymbol{\beta}_{tj}^c|\boldsymbol{\beta}_{t(\sim j)}^c, \eta_t^c, Y^c) \propto p(Y^{cj}|Y^{-cj}, \boldsymbol{\beta}_t^c, \eta_t^c) \times p(\boldsymbol{\beta}_{tj}^c|\boldsymbol{\beta}_{t(\sim j)}^c) \\
&\propto \prod_{l\in c} \left\{ \frac{\Gamma(n_{ltj} + \eta_t^c \boldsymbol{\beta}_{tj}^c)}{\Gamma(\eta_t^c \boldsymbol{\beta}_{tj}^c)} \frac{\Gamma(n_{ltt} + \eta_t^c \boldsymbol{\beta}_{tt}^c)}{\Gamma(\eta_t^c \boldsymbol{\beta}_{tt}^c)} \right\} \\
&\times \left(\frac{\boldsymbol{\beta}_{tj}^c}{1 - u_{tj}^c} \right)^{\lambda_t \boldsymbol{\alpha}_{tj}-1} \left(1 - \frac{\boldsymbol{\beta}_{tj}^c}{1 - u_{tj}^c} \right)^{\lambda_t \boldsymbol{\alpha}_{tt}-1}.
\end{aligned} \tag{18}$$

b. $p(\eta_t^c|\boldsymbol{\beta}_t^c, Y^c)$ 的推导如下：
联合后验分布 $p(\eta_t^c, \boldsymbol{\beta}_t^c|Y^c)$ 给定为

$$p(\eta_t^c, \boldsymbol{\beta}_t^c|Y^c) \propto \prod_{l\in c} \frac{\Gamma(\eta_t^c)}{\Gamma(n_{lt}+\eta_t^c)} \prod_j \frac{\Gamma(n_{ltj}+\eta_t^c\boldsymbol{\beta}_{tj}^c)}{\Gamma(\eta_t^c\boldsymbol{\beta}_t^c)} \times \prod_j (\boldsymbol{\beta}_{tj}^c)^{\lambda_t\boldsymbol{\alpha}_{tj}-1} \times \lambda e^{-(\lambda\eta_t^c)}. \tag{19}$$

则条件后验分布 $p(\eta_t^c|\boldsymbol{\beta}_t^c, Y^c)$ 可得

$$p(\eta_t^c|\boldsymbol{\beta}_t^c, Y^c) \propto \prod_{l\in c} \frac{\Gamma(\eta_t^c)}{\Gamma(n_{lt}+\eta_t^c)} \prod_j \frac{\Gamma(n_{ltj}+\eta_t^c\boldsymbol{\beta}_{tj}^c)}{\Gamma(\eta_t^c\boldsymbol{\beta}_t^c)} \times \lambda e^{-(\lambda\eta_t^c)}. \tag{20}$$

C. 模拟$\boldsymbol{\beta}_{tj}^c$ 和 η_t^c的 Metropolis-Hastings 算法

算法 1：模拟 $\boldsymbol{\beta}_{tj}^c$ 的 Metropolis-Hastings 算法。

1：初始化 $\boldsymbol{\beta}_{tj}^{c0}$
2：**对于** $i = 1,2,\ldots$ **执行**3：提议：$\boldsymbol{\beta}_{tj}^{c\text{cand}} \sim U(0,$最小值$\{2\boldsymbol{\beta}_{tj}^{c_i-1}, 1\})$
4：接受概率：$\alpha(\boldsymbol{\beta}_{tj}^{c\text{cand}}|\boldsymbol{\beta}_{tj}^{c_i-1}) =$最小$\left\{1, \frac{p(\boldsymbol{\beta}_{tj}^{c\text{cand}}|\boldsymbol{\beta}_{t\,(\sim j)}^{c}, \eta_t^c, Y^c)\, p(\boldsymbol{\beta}_{tj}^{c_i-1}|\boldsymbol{\beta}_{t\,(\sim j)}^{c}, \eta_t^c, Y^c)}{}\right\}$
5：$u \sim U(0,1)$
6：**若** $u \le \alpha$ **则**
7：接受该提议：$\boldsymbol{\beta}_{tj}^{c_i} \leftarrow \boldsymbol{\beta}_{tj}^{c\text{cand}}$
8：**else**
9：拒绝该提议：$\boldsymbol{\beta}_{tj}^{c_i} \leftarrow \boldsymbol{\beta}_{tj}^{c_i-1}$
10：**结束 if**
11：**结束 for**

算法 2：模拟 η_t^c 的 Metropolis-Hastings 算法。

初始化 η_t^{c0}
2：**对于** $i = 1,$ 2、..... **执行**提议：$\eta_t^{c\text{cand}} \sim U(0,$最小$\{2\eta_t^{c_i-1}, 1\})$
4：接受概率：$\alpha(\eta_t^{c\text{cand}}|\eta_t^{c_i-1}) =$最小$\left\{1, \frac{p(\eta_t^{c\text{cand}}|\boldsymbol{\beta}_t^c,\ Y^c)\, p(}{}\right\}$
$\eta_t^{i-1}|\boldsymbol{\beta}_t^c,\ Y^c)$
$u \sim U(0,1)$
6：**若** $u \le \alpha$ **则**
接受提议：$\eta_t^{c_i} \leftarrow \eta_t^{c\text{cand}}$
8：**else**
拒绝提议：$\eta_t^{c_i} \leftarrow \eta_t^{c_i-1}$
10：**结束条件判断**
结束循环

参考文献

Chib, S.; and Greenberg, E. 1995. 理解 Metropolis-Hastings 算法。《美国统计学家》, 49(4): 327–335. Chu, X.; Ouyang, W.; Li, H.; and Wang, X. 2016. Crf-cnn：用于人体姿态估计中的结构化信息建模。*arXiv* 预印本 *arXiv:1611.00468*。

Dawid, A. P.；以及 Skene, A. M. 1979年。使用期望最大化算法进行观察者错误率的最大似然估计。《皇家统计学会期刊：*C*辑（应用统计）》，28(1)：20–28。Dumitrache, A.；Aroyo, L.；以及 Welty, C. 2019。带有歧义的众包框架消歧语料库。*arXiv*预印本 *arXiv:1904.06101*。Finin, T.；Murnane, W.；Karandikar, A.；Keller, N.；Martineau, J.；以及 Dredze, M. 2010。使用众包对 Twitter 数据中的命名实体进行标注。载于*NAACL HLT 2010*年利用亚马逊 *Mechanical Turk*创建语音与语言数据研讨会论文集，80–88。Griffiths, T. L.；以及 Steyvers, M. 2004。发现科学主题。《美国国家科学院院刊》，101（增刊1）：5228–5235。Hovy, D.；Berg-Kirkpatrick, T.；Vaswani, A.；以及 Hovy, E. 2013。使用 MACE 学习信任谁。载于*2013*年北美计算语言学协会会议论文集：人类语言技术，1120–1130。Huang, Z.；Xu, W.；以及 Yu, K. 2015。用于序列标注的双向LSTM-CRF模型。*arXiv*预印本 *arXiv:1508.01991*。Juang, B. H.；以及 Rabiner, L. R. 1991。用于语音识别的隐马尔可夫模型。《计量学》，33(3)：251–272。Lafferty, J.；McCallum, A.；以及 Pereira, F. C. 2001。条件随机场：用于分割和标注序列数据的概率模型。Lakkaraju, H.；Leskovec, J.；Kleinberg, J.；以及 Mullainathan, S. 2015。用于建模人类评价的贝叶斯框架。载于*2015*年*SIAM*国际数据挖掘会议论文集，181–189。SIAM。Li, S.-Y.；Jiang, Y.；Chawla, N. V.；以及 Zhou, Z.-H. 2018。来自众包的多标签学习。《*IEEE*知识与数据工程汇刊》，31(7)：1369–1382。Nguyen, A. T.；Wallace, B. C.；Li, J. J.；Nenkova, A.；以及 Lease, M. 2017。从众包标注中聚合和预测序列标签。载于会议论文集。计算语言学协会。会议，第2017卷，299。美国国立卫生研究院公共获取。Plank, B.；Hovy, D.；以及 Søgaard, A. 2014a。使用标注者间一致性损失学习词性标注器。载于第*14*届欧洲计算语言学协会分会会议论文集，742–751。Plank, B.；Hovy, D.；以及 Søgaard, A. 2014b。语言学上可争辩，还是纯粹错误？载于第*52*届计算语言学协会年会论文集（第*2*卷：短文），507–511。Qiao, M.；Bian, W.；Da Xu, R. Y.；以及 Tao, D. 2015。用于序列标注的多样化隐马尔可夫模型。《*IEEE*知识与数据工程汇刊》，27(11)：2947–2960。Raykar, V. C.；Yu, S.；Zhao, L. H.；Valadez, G. H.；Florin, C.；Bogoni, L.；以及 Moy, L. 2010。从众包中学习。《机器学习研究期刊》，11(4)。

Rodrigues, F.；Pereira, F.；以及 Ribeiro, B. 2014年。使用多名标注者进行序列标注。机器学习，95(2)：165–181。Sabetpour, N.；Kulkarni, A.；Xie, S.；以及 Li, Q. 2021年。从众包序列标签中发现真值。载于*2021*年*IEEE*国际数据挖掘会议（*ICDM*），539–548。IEEE。Sang, E. F.；以及 De Meulder, F. 2003年。CoNLL-2003共享任务简介：与语言无关的命名实体识别。*arXiv*预印本 *cs/0306050*。Sarawagi, S.；以及 Cohen, W. W. 2004年。用于信息抽取的半马尔可夫条件随机场。神经信息处理系统进展，17：1185–1192。Simpson, E.；以及 Gurevych, I. 2018年。一种使用众包进行序列标注的贝叶斯方法。*arXiv*预印本 *arXiv:1811.00780*。Snow, R.；O'Connor, B.；Jurafsky, D.；以及 Ng, A. Y. 2008年。便宜且快速——但效果好吗？评估自然语言任务中的非专家标注。载于*2008*年自然语言处理中的经验方法会议论文集，254–263。Sober´on, G.；Aroyo, L.；Welty, C.；Inel, O.；Lin, H.；以及 Overmeen, M. 2013。衡量众包真值：结合工作者行为过滤器的分歧度量。载于*CrowdSem 2013*研讨会，第2卷。Tian, Y.；以及 Zhu, J. 2012。从众包中学习：存在思想流派时的情形。载于第*18*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，226–234。Tu, J.；Yu, G.；Domeniconi, C.；Wang, J.；Xiao, G.；以及 Guo, M. 2020。通过联合矩阵分解实现多标签众包共识。知识与信息系统，62(4)：1341–1369。Vaswani, A.；Shazeer, N.；Parmar, N.；Uszkoreit, J.；Jones, L.；Gomez, A. N.；Kaiser, Ł.；以及 Polosukhin, I. 2017。注意力机制即一切。神经信息处理系统进展，30。Whitehill, J.；Wu, T.-f.；Bergsma, J.；Movellan, J.；以及 Ruvolo, P. 2009。谁的投票应该得到更多权重：对未知专业水平标注者的标签进行最优整合。神经信息处理系统进展，22：2035–2043。Wu, X.；Fan, W.；以及 Yu, Y. 2012。Sembler：通过众包序列标注集成提升质量。载于*AAAI*人工智能会议论文集，第26卷。Yu, G.；Tu, J.；Wang, J.；Domeniconi, C.；以及 Zhang, X. 2020。主动多标签众包共识。*IEEE*神经网络与学习系统汇刊，32(4)：1448–1459。Zhang, J.；以及 Wu, X. 2018。面向众包的多标签推断。载于第*24*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，2738–2747。