一种基于多模态主题模型的图像自动标注方法*

田 璟 郭 智 黄 宇 黄廷磊 付 琨
(中国科学院电子学研究所地理空间信息处理与应用系统技术重点实验室 北京 100190)

摘 要：大部分传统的图像自动标注方法需要训练数据中具有精准的标注词，然而这样的数据通常是需要人工标注的，因此获取成本较高，且存在一定的主观性。该文提出一种全新的图像自动标注方法，通过结合自然语言理解领域实体识别的技术，充分利用图像周边自带环绕文本，将图像视觉特征、环绕文本以及实体抽取所得到的能够描述图像中显著特征的词在概率主题模型中进行联合建模，学习到多种数据模态之间的关联关系，从而实现图像的自动标注。在 UIUC Pascal Sentence 数据集上的实验证明该方法比传统方法具有更好的图像标注预测以及检索性能。

关键词：图像标注；主题模型；隐狄利克雷分配；Gibbs 采样
中图分类号：TP391 TN911.73 **文献标识码**：A **国家标准学科分类代码**：520.2040

Automatic image annotation method based on multi-modal topic model

Tian Jing Guo Zhi Huang Yu Huang Tinglei Fu Kun
(Key Laboratory of Technology in Geo-spatial Information Processing and Application System, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Most of the traditional approaches for automatic image annotation generally demand on training data with exact labels. However, this kind of data usually comes from human generated tags, which are too subjective and often difficult to obtain. In this paper, we propose a novel image annotation model which can utilize the rich surrounding text of images by integrating entity extraction technology of Natural Language Processing. Then the image features, surrounding text and extracted entity words which we assume that can more directly describe the salient objects in the corresponding images are modeled in a uniform probabilistic topic model framework. The learned correlations among different data modalities can be used in image annotation afterwards. Experimental results on UIUC Pascal Sentence dataset show that our model outperforms the traditional topic model-based image annotation methods in both annotation and retrieval.

Keywords: image annotation; topic model; LDA; Gibbs sampling

1 引 言

近年来，随着数字技术的迅猛发展，视觉数据以前所未有的规模被创造和储存，逐渐成为了与文本数据一样常见的数据形态，这催生了对大规模图像进行有效的组织、索引和检索的需求。图像自动标注的目的是给图像分配语义相关的文本描述，属于图像语义检索的中间步骤。图像自动标注[1]中最重要的部分是建立图像的低层特征与高层语义之间的关联映射，关联关系越合理，自动标注获得的文本描述质量越高，对图像内容的表达更加语义相关，图像检索得到的结果就越正确。现有的图像自动方法以建立关联模型为主[2,4,6]。这类方法通过利用已标注好的数据集学习图像的视觉特征与文本关键词之间的关联，然后将这种关联应用于待标注图像。文献[2]针对图像标注任务提出了几个基于基本主题模型 LDA(latent dirichlet allocation)[3]的变形模型，其中 C-LDA(Correspondence LDA)作为其中性能较为突出的一个模型，是将 LDA 模型应用于图像标注的标志性工作，该模型认为标注同一图像的标注词之间存在一定的关系，且该关系可以反映图像中区域之间的相互关联关系，C-LDA 通过对这种关系进行建模，最终得到既可以很好地拟合 2 种类型数据之间联合分布，又可以建模其之间条件关系的模型。文

收稿日期：2014-12
* 基金项目：国家 863 计划项目(NO.2014AA7013033)资助课题

献[4]是 C-LDA 的一个变形，借鉴有监督的主题模型 S-LDA(Supervised LDA)[5]的思想，利用一个 softmax 函数对 S-LDA 做了扩展，将连续的图像特征映射到离散的图像类别上，使得模型可以同时做图像分类与标注。文献[6]所提出的 trmm-LDA(topic-regression multi-modal LDA)是 C-LDA 的另一种变形，通过在图像与文本 2 种模态的产生式过程中引入一个线性高斯回归模块，改进了 C-LDA 中 2 种模型之间的关联方式，使得两种数据模态之间的最优主题数可以根据自身的特点设置而无需保持一致，是一种更灵活的关联方式。然而，上述以建立关联模型为基础的方法都需要精确标注的训练集，而在实际中这样的数据是难于获取的，因而无法扩展到数据量比较大的应用场景中；而且，随着因特网的发展，图像数据的呈现形式趋于多样化，网络图像多伴有环绕文本，包括标题、时间信息、地理信息以及用户信息等多元信息，而在传统的图像标注方法中一般缺少对这类信息的建模，相对于低层图像特征，丰富的文本信息并没有得到充分的利用。

本文提出了一种全新的针对带有环绕文本图像的自动标注方法。首先，利用实体抽取方法对环绕文本做预处理，得到能够描述图像中具有显著性的目标的“实体词”。其次，进一步利用改进的主题模型对包括视觉特征、“实体词”和环绕文本中的其他词在内的 3 种不同的数据模态进行联合建模，学习其之间隐含的关联关系。最后，学习到的隐含关联关系可以用来对新图像的标注词进行预测，实现图像的自动标注。

2 基于多模态实体主题模型的图像标注

本文提出的多模态主题模型与传统方法的不同之处在于环绕文本部分的处理。通过结合自然语言理解领域的实体识别方法抽取得到的“实体词”，具有能够描述图像中显著内容的特点，与环绕文本中的其他词相比，往往能更直接地指出图像中的场景、物体以及动作等对标注具有明显促进作用的特定目标，因此称之为“实体词”，与传统自然语言理解中的“实体”[7]有所不同，这里的实体词针对图像任务做了定义上的扩展，并使用专门的实体抽取工具箱——UW Twitter NLP[8]作为实体抽取的工具。建模的方式采取了主题模型建模，建模的对象包括原始环绕文本、图像视觉词以及抽取得到的实体词。考虑到模型能够同时对文本、实体、图像联合建模，因此称之为多模态-实体主题模型(multi-modal entity LDA)，简称 MME-LDA。

2.1 多模态联合建模

MME-LDA 模型的概率图模型如图 1 所示。与原始的 LDA 相比，该模型共有 3 条并行的路径，从左往右依次代表原始环绕文本、抽取得到的自定义实体词、视觉词共 3 类数据的产生式过程，它们分别拥有不用的词典以及词典先验，但是共用一个主题分配向量以及该向量对应的先验。

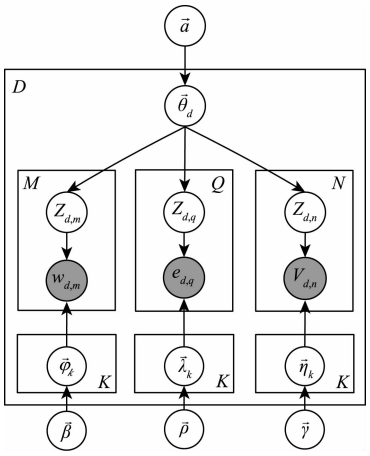

图 1 MME-LDA 概率图模型

在模型中，每一个文档都由 3 种数据组成，假设整个文档集中文档的数目为 D 篇，对其中的任意一篇文档 d，均有 M_d 个普通文本词，Q_d 个实体词，N_d 个视觉词组成，3 种词的词典维度分别为 W，E 和 V，主题的维度为 K。具体到其中每一个普通文本词 $w_{d,m}$，实体词 $e_{d,q}$，视觉词 $v_{d,n}$ 的生成，其产生式过程如下：

1)词典生成过程。生成普通词词典 $\vec{\varphi}_k \sim Dir(\vec{\beta})$，$\vec{\varphi}_k$ 是 W 维的列向量，代表词典中的每个词发生的概率，$\vec{\beta}$ 是 Dirichlet 分布的参数，为 W 维向量；生成实体词词典 $\vec{\lambda}_k \sim Dir(\vec{\rho})$，$\vec{\lambda}_k$ 是 E 维的列向量，代表词典中的每个词发生的概率，$\vec{\rho}$ 是 Dirichlet 分布的参数，为 E 维向量；生成视觉词词典 $\vec{\eta}_k \sim Dir(\vec{\gamma})$，$\vec{\eta}_k$ 是 V 维的列向量，代表词典中的每个词发生的概率，$\vec{\gamma}$ 是 Dirichlet 分布的参数，为 V 维向量。

2)选择该文档的主题分配情况 $\vec{\theta}_d \sim Dir(\vec{\alpha})$，$\vec{\theta}_d$ 是 K 维的列向量，代表的是各个主题发生的概率，$\vec{\alpha}$ 是 Dirichlet 分布的参数，为 K 维向量。

3)生成词。对每一个具体的词 $w_{d,m}$，$e_{d,q}$，$v_{d,n}$：

①依次按照多项式分布选择一个主题，$z_{d,m} \sim Mult(\vec{\theta}_d)$，$z_{d,q} \sim Mult(\vec{\theta}_d)$，$z_{d,n} \sim Mult(\vec{\theta}_d)$。

②依照选定主题 $z_{d,m}$，$z_{d,q}$，$z_{d,n}$ 条件下的多项式分布选择一个词 $w_{d,m} \sim Mult(\vec{\varphi}_{z_{d,m}})$，$e_{d,q} \sim Mult(\vec{\lambda}_{z_{d,q}})$，$v_{d,n} \sim Mult(\vec{\eta}_{z_{d,n}})$。

2.2 模型求解

由于模型参数较多，目标函数的形式复杂，使用变分方法进行求解具有一定的难度。因此，本文采用 Gibbs 采样[9]的方法来实现对模型的参数估计。在 MME-LDA 中，共有 3 个不同的变量要进行采样，分别为环绕文本主题 $z_{d,m}$，实体词主题 $z_{d,q}$，视觉主题 $z_{d,n}$。首先，需要对 3 个变量计算各自的更新公式；然后，在采样收敛后，利用前一步得到的统计量计算各类数据模态的词典；最后，在预测阶段，利用训练得到的词典、模型参数等先验知识实现对新图像的标注词预测。

文本词主题更新公式如下：

$$p(z_{d,m}\mid rest)=\frac{n_d^{(z_{d,m})}+\alpha}{\sum_{z_{d,m}=1}^{K}(n_d^{(z_{d,m})}+\alpha)}\cdot\frac{n_{z_{d,m}}^{(w_{d,m})}+\beta}{\sum_{w_{d,m}=1}^{W}(n_{z_{d,m}}^{(w_{d,m})}+\beta)} \tag{1}$$

式中：*rest* 代表排除了所要采样的主题之外的其他观测数据的当前主题分配情况，$\vec{z}_{-d,m}$ 代表除词项 $w_{d,m}$ 之外其余词项的主题分配情况。$n_d^{(z_{d,m})}$ 代表主题 $z_{d,m}$ 分配给文档 d 中的普通词的次数，$n_{z_{d,m}}^{(w_{d,m})}$ 代表词项 $w_{d,m}$ 分配给主题 $z_{d,m}$ 的次数。

同理，可得抽取得到的实体词的主题 $z_{d,q}$ 更新概率：

$$p(z_{d,q}\mid rest)=\frac{n_d^{(z_{d,q})}+\alpha}{\sum_{z_{d,q}=1}^{K}(n_d^{(z_{d,q})}+\alpha)}\cdot\frac{n_{z_d,q}^{(e_{d,q})}+\rho}{\sum_{e_{d,q}=1}^{E}(n_{z_d,q}^{(e_{d,q})}+\rho)} \tag{2}$$

视觉词的主题 $z_{d,n}$ 更新概率：

$$p(z_{d,n}\mid rest)=\frac{n_d^{(z_{d,n})}+\alpha}{\sum_{z_{d,n}=1}^{K}(n_d{}^{(z_{d,n})}+\alpha)}\cdot\frac{n_z^{(v_{d,n})}{}_{d,n}+\gamma}{\sum_{v_{d,n}=1}^{V}(n_{z_{d,n}}^{(v_{d,n})}+\gamma)} \tag{3}$$

Gibbs 采样收敛后，需要根据最后文档中所有词的主题分配情况来计算概率图模型中的隐变量，即文档中的主题分配情况 $\vec{\theta}$ 和 3 个模态各自的词典：$\vec{\varphi}$、$\vec{\lambda}$ 与 $\vec{\eta}$。在标注词预测阶段，需要用到的有普通词词典 $\vec{\varphi}$ 以及实体词词典 $\vec{\lambda}$，为与预测阶段作区分，此处统一写作 φ_{train}、λ_{train}。

$$\varphi_{train}=\frac{n_{z_{d,m}}^{(w_{d,m})}+\beta}{\sum_{w_{d,m}=1}^{W}(n_{z_{d,m}}^{(w_{d,m})}+\beta)} \tag{4}$$

$$\lambda_{train}=\frac{n^{(e_{d,q})}+\rho}{\sum_{e_{d,q}=1}^{E}(n^{(e_{d,q})}+\rho)} \tag{5}$$

在预测阶段，首先需要得到待标注新图像的主题分配情况：

$$\vec{\theta}_{new}=\frac{n_d{}^{(z_{d,n})}+\alpha}{\sum_{z_{d,n}=1}^{K}(n_d{}^{(z_{d,n})}+\alpha)} \tag{6}$$

进而得到，预测普通词条件概率 $p(\vec{w}\mid I)$，其中，$\vec{w}$ 为 W 维的向量：

$$p(\vec{w}\mid I)=\vec{\theta}_{new}\cdot\vec{\omega}_{train}=\frac{n_d{}^{(z_{d,n})}+\alpha}{\sum_{z_{d,n}=1}^{K}(n_d{}^{(z_{d,n})}+\alpha)}\cdot\frac{n_{z_{d,m}}^{(w_{d,m})}+\beta}{\sum_{w_{d,m}=1}^{W}(n_{z_{d,m}}^{(w_{d,m})}+\beta)} \tag{7}$$

以及预测实体词条件概率 $p(\vec{e}\mid I)$，其中，$\vec{e}$ 为 E 维的向量：

$$p(\vec{e}\mid I)=\vec{\theta}_{new}\cdot\lambda_{train}=\frac{n_d{}^{(z_{d,n})}+\alpha}{\sum_{z_{d,n}=1}^{K}(n_d{}^{(z_{d,n})}+\alpha)}\cdot\frac{n^{(e_{d,q})}+\rho}{\sum_{e_{d,q}=1}^{E}(n^{(e_{d,q})}+\rho)} \tag{8}$$

最后，将得到的预测词按照所得条件概率排序，实现对新图像的标注。

3 实 验

实验将 MME-LDA 模型与对比方法在同一个数据集上做了对比，验证了本文所提出方法的有效性。对比方法包括 C-LDA[2]，Text-LDA。其中，Text-LDA 是本文为方便实验效果的对比而专门设计的对比方法，该方法沿用了 MM-LDA[10] 的建模方式，对文本与图像分别进行建模，且文本部分仅作分词、去停用词等基本处理。

3.1 数据集及参数设置

采用文献[11]中的 UIUC Pascal Sentence 数据集进行实验，该数据集是 Pascal-VOC 数据集的一个子集，最初用于生成图像的句子描述。其中共有 1 000 幅图像，每幅图像带有 5 个相应的句子描述。为保证实验结果的客观性，该文采用了 5 折交叉验证，随机选取数据中的 800 幅做训练图像，剩余的 200 幅做测试图像，共重复 5 次，并取平均值。

图像标注的性能通过比较测试集的自动标注与原始标注进行评估。采用与文献[6]相同的方法，在混淆度、准确率与正确率等几个定量评价指标上，与对比方法C-LDA和 Text-LDA 做了对比。

3.2 混淆度评估

实验计算了在不同主题数目和不同视觉词维度的设定下，标注词预测概率 $p(\vec{w}\mid I)$ 和 $p(\vec{e}\mid I)$ 的混淆度。混淆度是语言模型中常用的一种评价指标，该值表征了语言模型的泛化能力，通常是混淆度的值越低越好。混淆度的计算有两步：首先计算预测词的似然，然后求全部似然的几何平均值的倒数。其定义如式(9)所示。

$$Perp=\exp\left(-\frac{\sum_{d=1}^{D}\sum_{w=1}^{W_d}\log p(w\mid I)}{\sum_d W_d}\right) \tag{9}$$

式中：D 为全部文档的数目，W_d 代表文档 d 中词项的维度。

图 2 给出了全部模型的实验结果。其中，MME-LDA 的预测结果有两部分：普通词与实体词，分别为图中的 MME-LDA-text 与 MME-LDA-entity。首先，与其他 2 个对比模型相比，随着主题数目的变化，MME-LDA 的预测结果(包括 MME-LDA-text 与 MME-LDA-entity)在总体上具有更低的混淆度；其次，相比 2 个对比模型，在相同的视觉词设定下，如图 2(a) MME-LDA 模型的混淆度曲线相对趋于平缓，而对比模型的混淆度曲线有明显的上升趋势，这表明随着主题数目的增长，对比模型存在过拟合的问题；最后，观察图 2(a)，(b)，(c) 3 个不同的视觉词数目设定下的混淆度对比情况，视觉词数目从 1 000 依次降至 600、200 的过程中，全部模型的标注预测混淆度在整体上均有所降低，且过拟合的问题也均随之减轻。

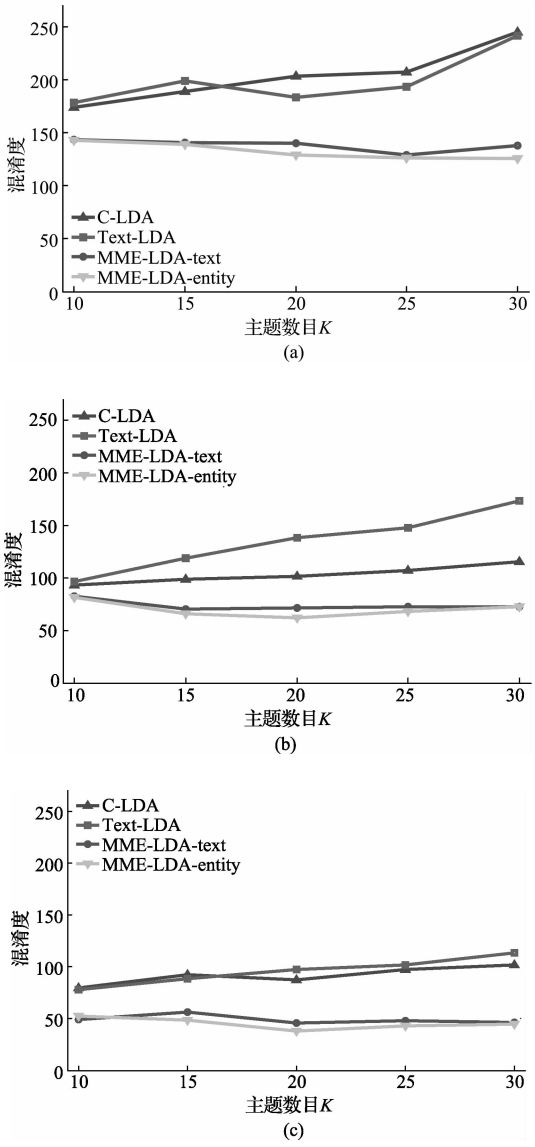

图 2　各方法在不同视觉词个数设定下的混淆度比较

3.3　准确率与召回率评估

图像自动标注作为图像检索的中间步骤，其另一个重要的指标为准确率与召回率。与文献[6]中一致，取前 5 个后验概率最大的关键词作为每幅图像的标注结果，并随机选取 3 个词“bus”，“sitting”，“water”为例，在测试集上以其为查询关键词进行图像检索，并将检索所得的图像分别按照后验概率 $p(\vec{w}\mid I)$ 和 $p(\vec{e}\mid I)$ 进行排序，评估它们在 MME-LDA 与对比模型中的表现，如图 3 所示。因为所选取的 3 个词均既为普通词又为实体词，所以 MME-LDA 的准确率-召回率曲线有 2 条，分别代表普通标注词 MME-LDA-text 与实体标注词 MME-LDA-entity。在图 3(a)、(b)、(c)中 MME-LDA-entity 和 MME-LDA-text 的准确率-召回率曲线与 C-LDA 与 Text-LDA 相比，在大部分情况下，在召回率相同时具有更高的准确率；在图 3(a)、(b)中，C-LDA 在总体上比 Text-LDA 表现更好，尤其在(a)中以“water”为检索词时，可见 C-LDA 因为建模了图像与文本之间的关联关系而在一定程度上提高了标注预测的准确率；而在(c)图中以“bus”为检索词时 C-LDA 与 Text-LDA 给出了相近的性能，其 2 条曲线几乎重叠，且总体上均低于 MME-LDA-entity 和 MME-LDA-text 的准确率，预示着 C-LDA 的建模方法中对图像、文本模态关联建模的方式与 MME-LDA 相比可能不够灵活，导致其性能相对较差。

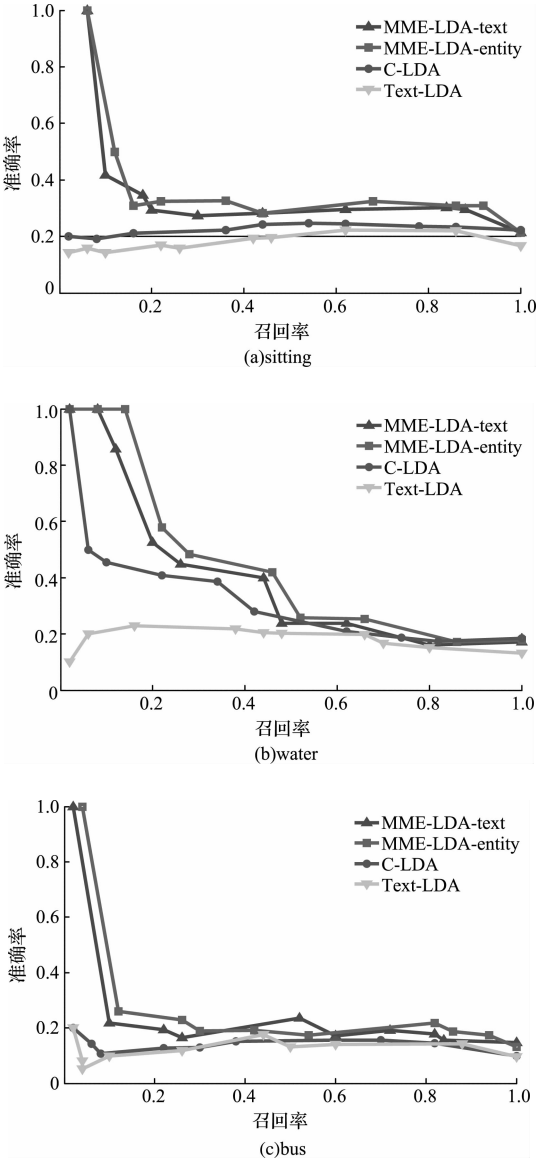

图 3　各方法在部分预测词上的准确率与召回率比较

综上可见，对多模态的建模尤其是对实体词的单独建模所学习到的隐含关联关系可以提升标注预测的能力，且表现为实体词与普通词的标注性能均有所提升，从而表明模型中实体词与普通词之间并非简单地并列，而是有相互促进的关系。

4 结 论

本文研究了使用文本分析辅助实现图像自动标注的问题，并提出了一种全新的多模态概率主题模型。通过融合自然语言理解领域的实体识别技术，该模型将文本、图像以及从文本中抽取到的实体进行了多模态的联合建模。实体的引入是对周边文本数据中有助于标注的信息的一种变相的语义加权处理，使得对图像中的显著对象进行建模成为可能。在 UIUC Pascal Sentence 数据集上的实验验证了所提方法的可行性与有效性。本文所提出的 MME-LDA 方法是对带有环绕文本的图像数据的一个尝试性工作，自然语言理解领域的其他方法对图像语义标注与理解的更多作用有待于进一步的挖掘。

参 考 文 献

[1] 张滢，齐美彬，周云，等. 基于特征提取和多示例学习的图像区域标注[J]. 电子测量与仪器学报，2014，28(8)：909-914.

[2] BLEI D M，JORDAN M I. Modeling annotated data [C]. Proceedings of the ACM Conference on Special Interest Group on Information Retrieval，2003：127-134.

[3] BLEI D M，NG A Y，JORDAN M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research，2003(3)：993-1022.

[4] WANG C，BLEI D M，FEI-FEI L. Simultaneous image classification and annotation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2009：1903-1910.

[5] BLEI D M，MCAULIFFE J D. Supervised topic models[J]. Advances in neural information processing systems，2007(7)：121-128.

[6] PUTTHIVIDHYA D，ATTIAS H T，NAGARAJAN S S. Topic regression multi-modal Latent Dirichlet Allocation for image annotation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2010：3408-3415.

[7] BLACK W J，RINALDI F，MOWATT D. FACILE：Description of the NE system used for MUC-7[C]. Proceedings of the Seventh Message Understanding Conference，1998.

[8] LEON DERCZYNSKI，ALAN RITTER，SAM CLARK. Twitter part-of-speech tagging for all：overcoming sparse and noisy data[C]. Proceedings of Recent Advances in Natural Language Processing，2013：198-206.

[9] NEAL R. Markov chain sampling methods for Dirichlet process mixture models[J]. Journal of computational and graphical statistics，2000，9(2)：249-265.

[10] RAMAGE D，HEYMANN P，MANNING C D，et al. Clustering the tagged web[C]. Proceedings of the ACM International Conference on Web Search and Data Mining，2009：54-63.

[11] FARHADI A，HEJRATI M，SADEGHI M A，et al. Every picture tells a story：generating sentences from images[C]. Proceedings of the European Conference on Computer Vision，2010：15-29.

作 者 简 介

田璟，1987 年出生，博士研究生。主要研究方向为数据挖掘、图像自动标注。

E-mail：tianjing0303@163.com