学界 | 在 Twitter 中使用图像和文本命名实体识别

2017-11-17 机器海岸线

选自 arXiv

作者：Diego Esteves, Rafael Peres, Jens Lehmann, and Giulio Napolitano 等

机器海岸线编译

参与：方建勇

Named Entity Recognition in Twitter using Images and Text

Diego Esteves[1], Rafael Peres[2], Jens Lehmann[1,3], and Giulio Napolitano[1,3]

[1] University of Bonn, Germany
{surname}@cs.uni-bonn.de,
[2] Federal University of Rio de Janeiro, Brazil
rafaelperes@ufrj.br
[3] Fraunhofer IAIS, Bonn, Germany
giulio.napolitano@iais.fraunhofer.de

论文链接：https://arxiv.org/pdf/1710.11027

摘要：命名实体识别（NER）是信息提取的重要子任务，旨在寻找和识别命名实体。尽管取得了近期的成绩，但我们仍然面临着正确检测和分类实体的局限性，突显短而嘈杂的文本，如 Twitter。大多数 NER 方法中一个重要的消极方面是高度依赖于手工特征和领域特定的知识，这是实现最新结果所必需的。因此，设计处理这种语言复杂背景的模型仍然具有挑战性。在本文中，我们提出了一种不依赖任何特定的语言资源或编码规则的多层次体系结构。与传统方法不同，我们使用从图像和文本中提取的特征对命名实体进行分类。在 Ritter 数据集上针对 Twitter 的最先进的 NER 的实验测试呈现竞争结果（0.59 F-measure），表明这种方法可能导致更好的 NER 模型。

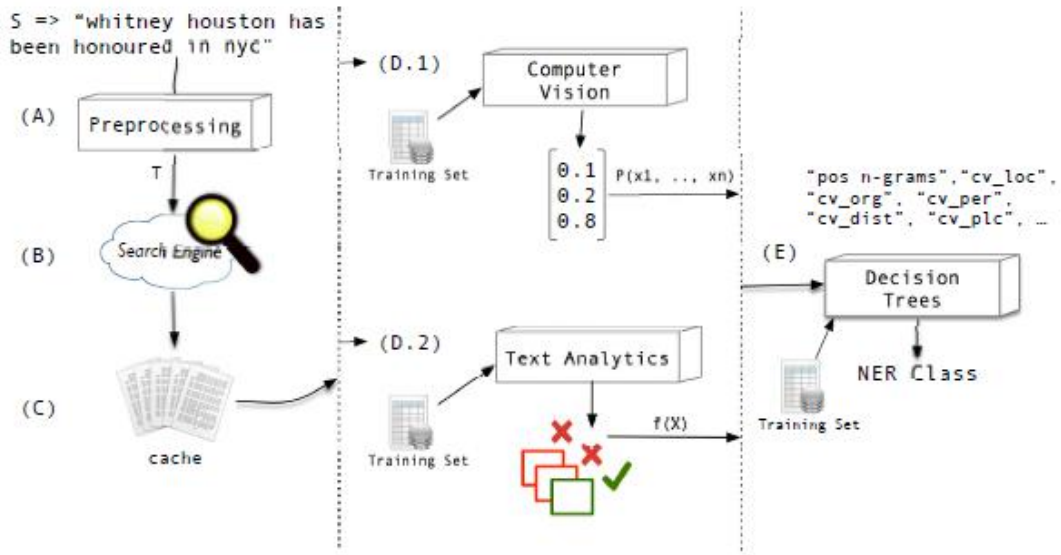

图 1：该方法概述：将计算机视觉和机器学习结合在一个通用的 NER 架构中。

NER	Images Candidates (number of trained models)
LOC	Building, Suburb, Street, City, Country, Mountain, Highway, Forest, Coast and Map (10)
ORG	Company Logo (1)
PER	Human Face (1)

表 1：NER 类和在给定图像中待检测的各个对象。 对于 LOC 来说，由于候选对象的多样性，我们训练了更多的模型。

NER Class	Precision	Recall	F-measure
Person (PER)	0.86	0.53	0.66
Location (LOC)	0.70	0.40	0.51
Organisation (ORG)	0.90	0.46	0.61
None	0.99	1.0	0.99
Average (PLO)	0.82	0.46	**0.59**

表 2：我们在 Ritter 数据集中的方法的性能测量：4 倍交叉验证。

NER System	Description	Precision	Recall	F-measure
Ritter et al., 2011 [19]	LabeledLDA-Freebase	0.73	0.49	0.59
Bontcheva et al., 2013 [3]	Gazetteer/JAPE	0.77	0.83	0.80
Bontcheva et al., 2013 [3]	Stanford-twitter	0.54	0.45	0.49
Etter et al., 2013 [6]	SVM-HMM	0.65	**0.49**	0.54
our approach	Cluster (images and texts) + DT	**0.82**	0.46	**0.59**

表 3：用于短文本（Ritter 数据集）的最新 NER 的性能测量（PER，ORG 和 LOC 类别），不依赖手工制定的规则和地名录的方法以灰色突出显示。（ Etter 等人，2013 年使用 10 个类别进行训练）。

✂----------