基于半监督学习的实体类型预测方法说明

许家铭, 郑孙聪, 徐博, 田冠华

中国科学院自动化研究所. 100190, 北京, 中国
{jiaming.xu,suncong.zheng,boxu,guanhua.tian}@ia.ac.cn

Abstract. 本说明文档介绍了一种基于半监督学习的实体类型预测方法，数据由 JIST2015 评测提供。本方法首先将实体词所链接的非结构摘要文本和结构化属性文本信息统一以向量空间模型进行向量化表示，并基于此训练线性支持向量机模型对多个实体类别进行打分。然后通过制定较为简单的规则对类别得分进行修正，并基于此得分设定阈值将未标注数据中置信度较高的样本加入到已标注数据中进行模型再学习。最终得到的实体类型预测模型具有如下几个优点：(1). 未使用任何外部语料数据; (2). 制定规则非常简单; (3). 模型训练速度很快。最终在训练集的交叉验证评价指标精度值为 98.06%，F1 值为 9.81，且模型训练平均耗时为 0.15 秒。

Keywords: 实体类型预测，半监督学习，支持向量机，JIST2015

1 数据说明

JIST2015 评测共发布 1,897 条 URLs，其中有已标注数据 1,397 条，未标注数据 500 条，实体类型共有 10 种，如：insect (虫)、university (大学)、game (游戏)、politician (政治家)、city (城市)、scene (景点) 等，而未标注数据则只有 URL 信息。

为了进行实体类型预测，数据中每条 URL 样本至多可获得四种信息，分别为：非结构化摘要 (Abstracts) 信息、结构化属性信息框 (Infobox data)、外部链接网址 (External URLs) 以及相关内容链接 (Related URLs)。如表 1 所示，非结构化摘要信息以文本形式必然存在，而属性信息框则是结构化的词特征且不一定为每一个实体词所有。另外，官方数据提供的外部链接网址和相关内容链接只提供了网址，而相关文本信息则需要用户自己基于此网址进行爬取，且此网址并非每个实体词的必有项。

三元组结构文件存储	含义	形式	必有项
./zhishime_abstracts_zh_enc.nt	Abstracts	文本	是
./zhishime_infobox_properties_zh_enc.nt	Infobox data	属性词	否
./zhishime_external_links_zh_enc.nt	External URLs	网址	否
./zhishime_related_pages_zh_enc.nt	Related URLs	网址	否

Table 1. 数据所提供的四种特征信息

2 文本特征抽取

本小节以标注数据中的如下 URL 做为示例介绍文本特征抽取过程：

“http://zhishi.me/baidubaike/resource/%FF%FE%4E%53%0F%59%61%21”

其标注的实体类型为 game (游戏)。

2.1 文本特征预处理

本方法对所有文本信息进行预处理，其过程如下：(1). 英文字母全部小写；(2). 过滤掉所有字符；(3). 中文分词；(4). 停用词过滤。

2.2 非结构化摘要文本特征

上述示例 URL 所对应的摘要信息如下：

“《华夏Ⅱ》是一款由深圳网域公司开发的 3D 网络游戏。该游戏以神、人、魔、幽冥四界之间的纷争和有情世间的恩怨情仇为主线，再现了《山海经》、《淮南子》等古书中记载的传奇故事，充分体验到华夏民族的创造力和想象力。”

可以看出摘要首句包含着重要的实体类型信息，因而在保留摘要全部文本特征的情况下，额外抽取其首句并进行分词处理。为了保证首句信息的完全保留，本方法文本预处理在进行正常中文分词的情况下，同时对首句进行 1-gram 分词以防止中文分词造成词项不匹配的语义鸿沟问题。如以下所示：

正常分词：“华夏 款 深圳 网 域 公司 开发 d 网络 游戏”；

1-gram 分词：“华 夏 款 深 圳 网 域 公 司 开 发 网 络 游 戏”；

2.3 结构化属性词特征

上述示例URL所对应的属性词特征如下表 2 所示：

属性类型	游戏画面	游戏特征	中文名	游戏类型	游戏平台	开发商
属性值	3D	奇幻游戏	华夏Ⅱ	角色扮演	网络游戏	深圳网域

Table 2. 示例 URL 所对应的属性集合

可以看出，属性词特征中也具有重要的实体类型信息。考虑到属性类型繁多、较不规范，本方法直接将 URL 所包含的属性类型和属性值直接做为词特征进行统计模型训练。

2.4 特征集合

基于以上文本特征提取过程，可得到如表 3 中所述的几种文本特征。在实验部分对这几种特征进行组合对比。

特征	描述	特征	描述
L1	摘要文本信息	L2	摘要文本信息+属性词特征
L3	摘要文本首句（中文分词）	L4	摘要文本首句（1-gram分词）

Table 3. 本方法所使用的文本特征

3 制定规则模版

如上章节示例的非结构化摘要文本信息分析，首句文本中包含了实体类型的重要信息，而且存在“[实体词] 是 [类型关键词]”的子句式结构。由于仅根据 URL 无法准确识别实体词，因而本方法利用“是 [类型关键词]”的子句式结构进行模版匹配，并定义规则函数 $f(x, \mathbf{w}_i)$ 如下：

规则 1：对于某一实体 x，判断其对应摘要文本信息首句中的前 k 个子句中是否包含所属词“是”，同时其后存在“类型关键词”，且该“类型关键词”距离该子句末尾的字距离小于 l，则认为模版匹配，即 $f(x, \mathbf{w}_i) = 1$，否则 $f(x, \mathbf{w}_i) = 0$。

基于此规则并生成向量化特征 $Q \in \mathbb{R}^M$，每个元素值为：

$$Q_i = \begin{cases} 0.1, \ if \ f(x, \mathbf{w}_i) = 1; \\ 0, \ if \ f(x, \mathbf{w}_i) = 0; \end{cases} \quad i \in \{1, 2, ..., M\} \tag{1}$$

其中，M 为实体类型数，$\mathbf{w}_i$ 为第 i 种实体此类关键词集合。本方法分别定义每一类的关键词集合表 4 所示。

实体类型	关键词集合	实体类型	关键词集合
insect	insect, 虫	university	university, 大学
game	game, 游戏	politician	politician, 政治家, 政客
city	city, 城, 镇	song	song, 曲
novel	novel, 小说	scene	scene, 景色, 景点, 村
cartoon	cartoon, 卡通, 漫画, 动画	actor	actor, 演员

Table 4. 实体词类型关键词集合

标注实体类型的数据中与地理位置相关类型有 city、university 和 scene，较易混淆。然而从占地面积上讲，“平方米”不会用来形容 city 类。如训练语料中共出现“平方米”关键词 14 次，其中 university 类 6 项，scene 类 8 项，city 类 0 项。因而定义规则模版如下：

规则 2：对于某一实体 x，判断其对应摘要文本及属性文本信息中是否包含面积关键词“平方米”。若包含，则置 M 维零值向量 V 中 city 类所属位置的元素值 V_{city} 置为 -0.1。

4 自训练半监督学习

4.1 基于支持向量机模型估计

利用向量空间模型将 URL 所对应的文本特征中抽取的词特征进行向量化表示，并通过 TF-IDF 进行权重赋值，得到特征向量 X。本方法引入线性支持向量机 (SVM) 模型进行训练：$f(X) = \text{sgn}(W^T X)$。基于此模型对未标注数据进行多类别估计得到多类别预测的概率化向量 $Y \in \mathbb{R}^M$。

4.2 后处理修正

此处将支持向量机得到的概率化向量Y和通过规则模版匹配得到的向量化特征 Q 及 V 进行线性组合对多实体类别预测得分进行修正：

$$\hat{Y} = Y + \lambda Q + V \tag{2}$$

4.3 再训练学习

在得到修正后概率预测得分 $\hat{Y}$ 之后通过设定最大类别概率得分与第二大类别概率得分相差的置信度阈值 γ 将未标注数据中置信度较高的样本加入到已标注数据中进行模型再学习，即再通过步骤 4.1 和 4.2 进一步得到最终的概率化得分中最大概率值所对应的类型做为预测的最终结果。

5 实验设计与结果

本方法选用 Matlab 版本 libSVM[1] 做为实体类型预测模型。采用开源的 Java 版 ansj[2] 分词工具进行中文分词，而摘要首句分别进行 ansj 和 1-gram 两种分词模式抽取 L3 特征和 L4 特征。用户词典大小为 20,970，将训练样本分成 3 组进行交叉验证。并基于验证集设定模型中的参数，如设定规则 1 中的参数 k 和 l 及公式 2 中的参数 λ 均为 3，而置信度阈值 γ 为 0.2。

5.1 评价指标

官方的评估标准为基于预测准确率 P 和召回 R 的调和均数 $F1$，定义如下：

$$F1 = \frac{2PR}{P+R}, \quad P = \sum_{i=1}^{10} \frac{a_i}{b_i}, \quad R = \sum_{i=1}^{10} \frac{a_i}{c_i} \tag{3}$$

其中，a_i 是第 i 类实体类型预测正确的 URLs 样本数，b_i 是模型预测为第 i 类的样本数，而 c_i 则是第 i 类实体类型真实所有的 URLs 样本数。另外，本文采用精度值(ACC)做为实体类型预测模型的另一评估指标：

$$ACC = \sum_{i=1}^{N} \frac{\delta(y_i, p_i)}{N} \tag{4}$$

其中，N 为所有预测样本总数，$\delta(y_i, p_i)$ 为指示函数，当第 i 个未标注样本的预测类型 p_i 同真实类型 y_i 相同时该指示函数为 1，否则为 0。

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html
[2] https://github.com/ansjsun/ansj_seg

5.2 实验结果

从表 5 和表 6 中可以看到，只使用 URL 所对应的摘要文本信息时（即 L1）的初训练效果为 95.77% (ACC)/9.58 (F1)，而使用了全部特征并进行半监督自训练后效果提升至 98.06% (ACC)/9.81 (F1)。由于 ACC 采用的是微平均，而 F1 采用的宏平均，且不同类别数目不一致，因而会出现不同 F1 值和 ACC 值结果不完全一致的情况，如在使用 L2 特征时再训练及后处理情况下所对应的 ACC 值略有不同，而对应的 F1 值却完全相同。另外，从结果数据中发现，在初训练精度性能较低时，如使用特征 L3、L4，进行再训练时会引入噪音导致模型性能变差。而在初训练精度性能较高时，如使用特征 L2+L3、L2+L3+L4，进行再训练可进一步提高实体类型预测性能。

本方法实验环境为 Matlab (R2015a), 操作系统为 64 位 Windows，内存为 6.00GB，处理器采用因特尔公司的 i5-3230M，且主频为 2.60GHz。在提取好文本特征并选择最优实验模式进行模型训练及预测的平均总耗时为 0.15 秒。

特征集合	初训练	后处理	再训练	后处理
L1	95.77±0.44	96.63±0.24	96.42±0.65	96.42±0.53
L2	97.13±1.43	97.85±0.74	97.49±0.75	97.42±0.77
L3	92.77±1.61	93.70±1.54	93.05±0.61	93.34±0.73
L4	91.55±0.53	92.62±0.64	91.91±0.24	92.12±0.53
L2+L3	97.70±0.65	97.96±0.43	97.99±0.62	97.99±0.44
L2+L4	97.56±0.65	97.92±0.32	97.98±0.12	97.98±0.25
L2+L3+L4	97.63±0.77	98.03±0.32	**98.06±0.37**	**98.06±0.37**

Table 5. 不同特征组合下的 ACC 值 (%)

特征集合	初训练	后处理	再训练	后处理
L1	9.58±0.03	9.66±0.02	9.65±0.02	9.65±0.02
L2	9.71±0.14	9.78±0.07	9.76±0.06	9.76±0.06
L3	9.26±0.18	9.35±0.17	9.32±0.19	9.33±0.17
L4	9.15±0.06	9.26±0.08	9.20±0.04	9.21±0.06
L2+L3	9.77±0.06	9.79±0.04	9.80±0.04	9.80±0.04
L2+L4	9.76±0.06	9.78±0.03	9.79±0.01	9.79±0.01
L2+L3+L4	9.77±0.07	9.80±0.03	**9.81±0.03**	**9.81±0.03**

Table 6. 不同特征组合下的 F1 值