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 98.2，F1 值为 9.82。

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：虫 (insect)、大学 (university)、游戏 (game)、政治家 (politician)、城市 (city) 等。表 1 展示的是其中一条已标注数据，而未标注数据则只有 URL 信息。

URL	标签
http://zhishi.me/baidubaike/resource/%FF%FE%F3%67%DE%5D	city

Table 1. 已标注样本示例

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 2 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。

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：

三元组结构文件存储	含义	形式	是否必有
./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 2. 所提供的四种特征信息

“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。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 3 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 3. 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 4 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 4. 本方法所使用的文本特征

3 制定规则模版