基于深度学习的中文至拼音首字母自动转化方法

胡升泽[1] 蔡伟柯[2] 何春辉[1*]
(1、国防科技大学信息系统工程重点实验室,湖南 长沙 410073 2、国防科技大学教研保障中心,湖南 长沙 410073)

摘 要:随着智能技术的发展,许多智能产品的搜索系统中都集成了拼音首字母搜索功能,它可以极大提升用户的体验效果。但因为中文里面普遍存在多音字的使用,这给中文至拼音首字母的自动转换工作带来了极大的挑战。为了改善现有转换方法对中文多音字首字母自动转换准确率较低的问题,提出了一种基于深度学习的中文至拼音首字母转换方法。相关实验结果表明,在结合单音字首字母映射表进行微调的 Bi-LSTM-CRF 模型在独立测试数据集上达到了 99.7%的平均准确率。

关键词:深度学习;拼音首字母;Bi-LSTM-CRF;搜索引擎
中图分类号:TP391 文献标识码:A 文章编号:2096-4390(2020)02-0098-02

随着搜索引擎和智能技术的快速发展,很多系统都集成了中文首字母快速检索功能。较常见的有 KTV 点歌系统中歌曲名称的搜索以及智能电视中电视剧或电影名称的搜索等。它不同于传统的搜索引擎,为了提升用户的体验效果,它通常会简化用户的输入操作,只需用户按顺序输入检索内容的首字母,无需输入检索条件的全部内容,这样可以降低用户的输入难度,从而提升用户的检索体验。

这种基于首字母构建的快速检索系统虽然可以大大提升用户的体验效果。但它也面临着一个亟待解决的核心问题,即如何高效、准确的完成中文至拼音首字母的自动转换。众所周知,中文是一种很特殊的语言,它除了常见的单音字之外,还包含很多的多音字。对于单音字而言,汉字至拼音首字母的自动转换比较简单,但是对于多音字的汉字至拼音首字母的自动转换是一个较复杂的任务,它需要依赖上下文语义信息才能正确的完成自动转换。华逢兆采用汉字的分级结构实现了汉字转化为拼音首字母的功能[1]。这种方法虽然可以在大部分情况下完成汉字至拼音首字母的转换任务,但是它的转换准确率还有待进一步提升,尤其是面临多音字的正确转换显得捉襟见肘。

近来,随着硬件水平的提升使得深度学习算法在文本挖掘领域得到了广泛的应用[2]。因此,文章引入了深度学习算法来提升中文至拼音首字母的自动转换性能。在数据标注阶段,将需要转化的中文和它所对应的拼音首字母进行编码形成序列映射。最后用这些标注过的数据来完成深度学习模型的训练。

1 中文至拼音首字母自动转换算法

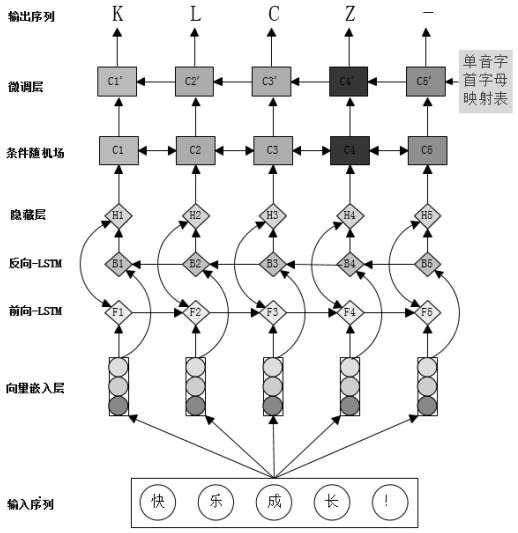
由类型来分,中文至拼音首字母的自动转换可以归为自然语言处理[3]中的序列标注任务。考虑到 Bi-LSTM-CRF(双向长短时记忆条件随机场)[4] 序列标注模型在很多任务上都取得了优秀的表现。因此,文章采用了这种深度学习网络结构来构建中文至拼音首字母的自动转换算法。其结构如图所示.

由图可知,自动转换算法一共包含了 8 个层次。首先是输入序列层,实现中文字符串的输入。接下来是字符向量嵌入层,用来完成中文字符的向量化表示。核心部分是双向长短时记忆网络层,它利用前向 - LSTM 层和反向 - LSTM 层来获取上下文的特征。其次通过隐藏层来实现数据转换。再次再利用条件随机场层给出最佳的序列预测结果。最后再结合单音字首字母映射表对预测结果中的单音字首字母进行微调并输出最终的首字母序列标注结果。

2 数据预处理

2.1 数据集获取

为了验证算法的性能,利用开源的网络爬虫工具 WebMagic① 从豆瓣电影②网站中爬取到 5 万部中文电视剧或电影名称。此外,还融合了搜狗实验室对外公开的精简版③新闻数据集包含的全部中文新闻标题共同作为模型训练和测试的原始语料。



Bi-LSTM-CRF 中文至拼音首字母自动转换算法的结构图

2.2 数据的标注

汉字至拼音首字母标注需要将输入的中文汉字字符序列对应的转换为这些字符所对应的拼音首字母序列的形式。根据中文的相关拼音发音标准,约定整个标注数据中只包含 24 类不同的字符标签。这些标签分别为 3 个单韵母和 20 个声母以及 1 个非中文的统一映射符。像电视剧名称“《快乐成长》”,就将它对应的字符序列标注为“- KLCZ- ”。因为整个数据集较大,其中将 80%作为训练数据集,15%作为验证数据集,5%作为独立测试集。在数据标注阶段,文章借助了中文到拼音开源的自动转换工具 HanLP 并结合人工校正的方式来完成数据的标注。最后,使用上述标注方式得到的标注数据来完成深度学习模型的训练、验证和测试。

3 实验分析

(转下页)

通讯简介:何春辉(1991-),男,汉族,湖南永州人,硕士,高级算法工程师。

电网规划用电网运行数据自动收集系统设计与实现

李友平 何 飓 胡成恩 孙文兵 储国良
(国网安徽省电力有限公司安庆供电公司,安徽 安庆 246003)

摘 要:随着电网的不断发展,电网系统各类运行数据量急剧增长、数据类型多样,对数据存储、处理、价值挖掘提出了更高要求。针对电力生产数据日益增多及数据结构多种多样发展趋势,怎样才能从大量数据中寻找其中的规律,发现隐藏在数据中有价值的信息,成为当前数据分析工作的难题。同时随着公司信息化建设的稳步推进,在智能电网标准体系建设等方面,对公司信息化建设提出新的要求。为进一步加强主网监控对公司主营业务的提升作用,有效支撑企业信息化及智能电网建设,满足移动信息化应用的稳定性、扩展性、可维护性、安全性等需要,有必要全面开展公司电网规划用电网运行数据自动收集系统建设,利用大数据挖掘技术,发现其潜在生产应用价值。

关键词:电网运行 数据自动收集 数据挖掘
中图分类号:TM73 文献标识码:A 文章编号:2096-4390(2020)02-0099-02

1 背景

随着"互联网+"和全球能源互联网战略的发展与应用,近些年来,海量的数据成为生活中最为普通却又极为昂贵的财富。电力生产中产生的数据信息涉及到发电、输电、变电、配电、用电、调度各环节,是跨单位、跨专业、跨业务数据分析与挖掘,以及数据可视化的大数据表现。数据也从结构化数据分析向多类型数据分析的转变、从抽样数据分析向全量数据分析的转变、从小批量数据分析向海量数据分析的转变、从单一业务数据分析向跨业务数据分析的转变从准实时数据分析向实时数据分析的转变促使公司积极通过改进海量数据同步技术,提升集中服务组件的数据传输承载能力,支持未来实时性更高、单次数据传输量更大、数据类型更丰富的数据集成需求。

2 安全设计

GB/T 22239-2008《信息安全技术信息系统安全等级保护基本要求》将信息系统安全等级由低到高划分为五级,该应用部署在公司内网中,按要求该项目属于二级。具体要求为:应能够防护系统免受来自外部小型组织的、拥有少量资源的威胁源发起的恶意攻击、一般的自然灾难、以及其他相当危害程度的威胁所造成的重要资源损害,能够发现重要的安全漏洞和安全事件,在系统遭到损害后,能够在一段时间内恢复部分功能。

根据国家电网公司"分区分域、安全接入、动态感知、全面防护"的安全策略,需要对该项目的物理安全、边界安全、应用安全、数据安全、主机安全、网络安全及终端安全进行安全防护设计。

同时对于整个系统,需要完整的权限控制,防止某些人恶意攻击系统,修改原始记录,同时对于数据库中的数据需要定时备份,防止系统数据丢失。此外,系统要求用户在登陆时需要 (转下页)

3.1 评测指标

在文章的实验评测环节,采用平均准确率来评估模型的性能。准确率的定义如下:对于一个输入的中文序列,如果拼音首字母自动转换方法能将它映射成一个完全正确的首字母序列,意味转换成功,只要转换结果中包含一个错误首字母意味转换失败。对于平均准确率的计算,需要统计所有参与评测的样本总数中转换成功的数量,并用它除去参与评测的样本总数。它的计算公式如下:

$$\text{平均准确率}=\frac{\text{所有评测样本中正确转换的样本总数}}{\text{评测样本总数}}\times 100 \quad (1)$$

3.2 实验结果

为了充分的验证模型性能,采用独立测试数据集对条件随机场、Bi-LSTM-CRF 以及结合单音字首字母映射表进行微调的 Bi-LSTM-CRF 这 3 种不同的模型进行了实验对比,并结合平均准确率指标对不同模型的性能进行评估。相关的实验结果如表所示。

不同模型在独立测试数据集上的实验结果

模型名称	平均准确率
条件随机场(CRF)	94.1%
Bi-LSTM-CRF	99.3%
微调的Bi-LSTM-CRF	99.7%

根据表的实验结果可知,不同模型之间存在一定的差距。CRF 的平均准确率为 94.1%,Bi-LSTM-CRF 模型取得了 99.3% 的平均准确率,但是在结合单音字首字母映射表进行微调后,微调的 Bi-LSTM-CRF 模型的平均准确率高达 99.7%。

4 结论

在中文至拼音首字母自动转换任务上,文章提出了基于深度学习的中文至拼音首字母自动转换方法,实验结果表明这种方法可以有效的提升多音字的转换准确率,且在融入单音字首字母映射表后可以有效提升中文至拼音首字母转换模型的性能。

注释

①https://www.oschina.net/p/webmagic.
②https://movie.douban.com/.
③https://www.sogou.com/labs/resource/cs.php.

参考文献

[1]华逢兆.PB 中实现将汉字转换为拼音首字母的方法[J].科技风,2015(11):215-216.
[2]崔嘉乐,姜明洋,裴志利,等.基于深度学习的文本挖掘研究[J].内蒙古民族大学学报(自然汉文版),2016,31(5):403-407.
[3]宗成庆.统计自然语言处理[M].北京:清华大学出版社,2013.
[4]姚茂建,李晗静,吕会华,等.基于 BI-LSTM-CRF 神经网络的序列标注中文分词方法[J].现代电子技术,2019,42(1):103-107.