

语料库研究方法综述

司开源

(云南经贸外事职业学院, 云南 昆明 651705)

摘要: 语料库 (Corpus) 指为了某一研究目的收集的, 用电脑可读形式保存的语言文本集合。语料库应用于词典编纂, 语言教学, 传统语言研究, 自然语言处理中基于统计或实例研究等方面。本文主要以中文文本校对语料库为案例, 对语料库研究方法进行综述。

关键词: 语料库; 综述; 中文文本校对

中文文本的自动校对技术在图书、报纸、网络媒体等领域具有非常重要的应用价值。并且随着计算机在新闻及图书出版领域的广泛应用, 各种电子出版物以惊人的速度涌现。尤其是随着当下互联网的越来越普及, 各种电子杂志、网页更是层出不穷。每时每刻都会有大量的中文文本数据被上传到互联网上, 这些文本数据最初往往只由几种输入法输入到互联网上, 比如拼音输入法、五笔输入法、OCR 输入法、ASR 输入法等。目前学术界和工业界主要通过收集真实数据的方法来获得标记语料库, 但是这种方法需要消耗大量的人力物力, 并且难以大量获取。为解决这一问题, 本文对中文文本校对语料库的自动生成技术展开研究。

一、语料库相关研究

(一) 语料库的定义

“语料库”一词来自于拉丁文 “Body”, 它通常指的是口语或笔语中任何语段或篇章。对于“语料库”的定义, 语言学家们众说纷纭, 语言学界内尚无统一的定论。早期研究这方面的国外语言学家们给出的若干定义, 且在不同时期均得到业界的认可。例如, 辛克莱 (Sinclair) 将语料库的含义归纳为“为了用作语言的样本, 通过一定的特殊规则挑选、分类、排序的语言素材的集合”。谢家成 (2003) 给出的语料库定义语料库是指存在计算机里的大量语言材料和定位检索管理软件的结合。综合来看, 语料库是基于语言研究基础上, 采用电子方式对语言材料进行长久的储存, 主要是收集人们日常口语与书面语的代表性语句的汇总集, 以此指代特殊情况下语言含义。语料库的构建是依据语言学标准选择并且排序的语言运用材料汇集, 将其作为语言的样本。

(二) 语料库的分类

现如今语料库发展迅速, 为适应不同需求而产生不同种类的语料库, 因此语料库的分类纷繁复杂。

从专业与否的角度, 语料库可分为通用语料库和专业语料库。通用语料库的特点是容量大, 题材广, 实用度高。常用的通用语料库包含美国国家语料库, 美国当代英语语料库等; 专业语料库是某个特定领域的语言的反映。例如, 如果我们研究英语中商务谈判和签订商务协议方面的内容, 我们则无法再通用语料库中查询到相对专业性强的知识, 而需要选择应用专业的商务英语语料库。

平行语料库与类比语料库。平行语料库是指由原文文本, 及其对应的译文文本构成的双语或多语语料库。平行语料库根据翻译方向的不同可划分为单向平行语料库、双向平行语料库和多向平行语料库; 类比语料库是指同一种语言的不同形式的变体或者不同的语言的文本组成的两个或两个以上的语料库, 其可划分为单语类比语料库和双语或多语类比语料库。例如, ECC (English Comparable Corpus) 等等。

随着近些年语料库在语言教学中的应用, 语料库家族逐渐产生了一种学习者语料库, 即由学习者自身习得的作品组成, 例如, 中国英语学习者语料库、中国学生英语口语语料库等等。

(三) 语料库建设步骤

通过分析、对比国内外大型及各种专业语料库的研究现状, 各种语料库建设的包含的步骤基本相同, 语料库建设的一般步骤主要包括语料的收集、语料筛选、语料标注赋码以及语料库的检索等, 下面分别介绍如下:

1. 语料的收集

语料的收集是指通过各种方式获取所需要的语料, 即选择合适的语料, 做预处理, 为语料的标注做好先期准备。语料中信息资源的搜集是创建语料库的首要前提, 是成功与否的基石, 语料收集工作的优劣, 对于语料库整体质量起到关键作用。传统的键盘输入和扫描输入, 费时费力, 极易出错不说, 需要消耗大量的精力去人工校对。现代计算机技术和网络资源使语料库语料的收集变得方便快捷, 文献资源、网络资源, 包括新闻、电子邮件等, 语料的大量收集在使用信息采集软件之后可以轻易做到。

2. 语料筛选

语料筛选是对所收集到的语料进行鉴别和选择, 根据语料的类型及平衡度, 决定语料的取舍, 筛选出真实、适用、能满足语料库需求的语料的过程。筛选语料可采用标题筛选、摘记筛选、标记筛选等方法。首先看语料来源, 必须是正规可信的信息源; 其次看标题, 淘汰掉无关的、重复的语料; 再看正文, 筛选出能突出主题思想、比较典型、从实际出发的语料。

3. 语料标注

语料标注就是对语料库中的原始语料进行加工, 把各种表示语言特征的赋码标注在相应的语言成分上, 以便于计算机的读取。现在中文信息处理技术已经比较成熟了, 一些工作可以进行完全或部分自动处理, 比如分词等。所以使用已有的技术可以进行部分功能的自动标注。

4. 语料库检索

语料库检索就是在已建成的语料库中, 结合已知条件搜索需要的语料, 最终得出相关语料的过程。目前常用的检索条件, 如文献类型、题目、著者等。先统计语料库的 N-gram 汉字串频次,

并将统计结果分别按频次大小及汉字串 Unicode 编码进行倒排序索引,对用户输入的检索请求,先检索汉字串频次统计结果,向用户反馈一个备选汉字串集合;然后让用户参与检索优化过程,选择正确性较高的汉字串;最后在语料库中检索用户选定的检索词。

二、校对语料库构建方法研究

在获取拼写校对标记数据时,大多数研究者采取了收集真实标记数据的方法。香港应用科技研究所通过收集香港小学生作文中的错误句子来构建语料库,共收集到 6890 个长度合理的至少包含 1 个错误的句子。Wu 等人收集小学一年级学生作文中的错误句子共 700 多条。国际计算语言学会中文语言处理小组(SIGHAN)为促进中文文本拼写校对任务的研究,历年来多次公布标记数据,但标记数据总数量仅一万余条。

同时也有少数研究者提出自动构建拼写校对语料库的方法,比如王鼎民等提出了一种混合的自动构建中文拼写校对语料库的方法,他们的方法主要是用来生成误用语音相似字符和误用字形相似字符这两种形式的错误,该方法并未考虑到不同的输入法会产生不同形式的错误,本文在构建拼写校对语料库时,为校对对不同输入法生成的文本构建对应的拼写校对语料库。

在获得语法校对标记数据时,多数研究者也主要是通过收集真实数据的方式,比如面向教育应用的自然语言处理工作坊从 2014 年到 2018 年共举办了 5 次中文语法错误发现评测,其中在 2014 年到 2016 年公布的是繁体中文语法校对数据,2016 年到 2018 年公布的是简体中文语法校对数据,据统计,NLPTEA 在 2017 年公布了 10449 个评测句子,其中包含了 26448 个错误点;2018 年只公布了 1067 个错误点。计算自然语言学习会议(CoNLL)14 年公布了一些人工收集的中文和英文纠错数据,包括中英文拼写错误的标记数据和中英文语法错误的标记数据。

Tao 等人提出基于 RNN 的 Sequence to Sequence 模型生成英文语法纠错语料库,基本思路和本文生成语法校对语料库的思路较为一致,都是先使用公布的数据训练模型,然后再使用训练好的模型生成具有语法错误的句子,但是也有以下三点不同:一是本文仅使用基于翻译的方法生成多字符、少字符以及字符替换类型的错误句子,在生成别字错误时,使用了构建拼写校对语料库的相关方法,之所以没使用翻译的方法来生成别字错误,是因为实验证明使用构建拼写校对语料库的相关方法生成的别字错误更接近真实的情况;二是使用的模型不同。

Jared 等人提出基于互译的方法生成英文语法校对语料库,比如先将英语翻译为日语或者法语等其他语言,在将其翻译回英文,这样翻译回英文的句子就可能出现语法错误,如此便生成了英文语法校对语料库。但是这种方法显然可控性较差,当翻译系统精度很高时,语法错误就会很难发生。

三、文献评述

综上所述,学术界在获取拼写校对标记数据时,大多数研究者采取了收集真实标记数据的方法,也有部分学者提出自动构建

拼写校对语料库的方法,还有部分学习提出基于 RNN 的 Sequence to Sequence 模型生成英文语法纠错语料库,此外,也有少数学者主张采用互译的方法生成英文语法校对语料库。中文语法错误包括了别字错误(拼写错误)、多字、少字和词序颠倒四种情况。无论是语法校对还是拼写校对,当前可用于中文文本校对的训练数据不足,深度学习模型难以得到充分训练,导致校对效果不佳。

参考文献:

- [1] 马雪芹.基于语料库中文书评作者身份建构不礼貌话语研究[J].浙江万里学院学报,2020,33(02):55-59.
- [2] 徐琳宏,丁堃,陈娜,李冰.中文文献引文情感语料库构建[J].情报学报,2020,39(01):25-37.
- [3] 罗永宏,林鹤凯.中文网络评论情感语料库的构建研究[J].福建电脑,2019,35(10):40-41.
- [4] 邓雪纯.基于语料库的留学生中文写作中模糊限制语语用研究[J].海外英语,2019(05):206-208.
- [5] 盛晨,孔芳,周国栋.中文篇章零元素语料库构建[J].北京大学学报(自然科学版),2019,55(01):15-21.
- [6] 郭勃,冯旭鹏,刘利军,黄青松.基于平行语料库的双语协同中文关系抽取[J].计算机应用,2017,37(04):1051-1055.
- [7] 靖纯.基于新闻语料库的中文自动校对改进方案探讨[J].中国传媒科技,2016(06):15-17.
- [8] 余一骄,刘芹.大规模中文语料库检索技术研究[J].计算机科学,2015,42(02):217-223.
- [9] 张美云,张娜.基于大数据语料库的国际中文教学新模式研究[J].创新创业理论与实践,2020,3(18):130-132.
- [10] 周运会,吴世雄.国外语料库隐喻研究综述[J].外语学刊,2015(1):7.
- [11] 董颖,刘日升,那春光.我国英语语料库计量研究综述[J].图书馆学刊,2011(01):126-129.
- [12] 赵琼.基于语料库的译者文体研究综述[J].现代语文(语言研究),2018(008):131-135.
- [13] 郝玉荣,罗桂保.国内语料库翻译研究现状综述[J].广东学院学报,2010,21.
- [14] 任敬辉.国内基于语料库的翻译研究综述(2007-2013)[J].长春师范大学学报,2015,34(6):5.
- [15] 方秀才.基于语料库的英语教学与研究综述:成就与不足——根据 22 种语言学类 CSSCI 来源期刊近 30 年的统计分析[J].外语电化教学,2012(3):6.
- [16] 何常丽.语料库语言学研究综述[J].渤海大学学报(哲学社会科学版),2009,18(3):72-74.
- [17] 王龙吟,何安平.课堂教学语料库在英语教学研究中的应用——英语课堂会话语篇分析系列论文综述[J].中小学外语教学,2005,28(1):5.
- [18] 常芳玲.国内外基于语料库的历时语言研究文献综述[J].山东外语教学,2020,41(6):11.