

基于CRF的中文语法错误诊断系统的实现与应用

李斌, 王浩畅

引用本文

李斌, 王浩畅. 基于CRF的中文语法错误诊断系统的实现与应用[J]. 计算机科学, 2024, 51(6A): 230900073-6.

LI Bin, WANG Haochang. Implementation and Application of Chinese Grammatical Error Diagnosis System Based on CRF [J]. Computer Science, 2024, 51(6A): 230900073-6.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[Dilithium算法的FPGA高效扩展性优化](#)

FPGA Efficient Scalability Optimization of Dilithium

计算机科学, 2024, 51(6A): 230800138-9. <https://doi.org/10.11896/jsjcx.230800138>

[基于本体驱动的航空情报表格信息结构化研究](#)

Ontology-driven Study on Information Structuring of Aeronautical Information Tables

计算机科学, 2024, 51(6A): 230800150-7. <https://doi.org/10.11896/jsjcx.230800150>

[一种基于异构图神经网络和文本语义增强的实体关系抽取方法](#)

Method for Entity Relation Extraction Based on Heterogeneous Graph Neural Networks and TextSemantic Enhancement

计算机科学, 2024, 51(6A): 230700071-5. <https://doi.org/10.11896/jsjcx.230700071>

[融合BERT模型与词汇增强的中医命名实体识别模型](#)

TCM Named Entity Recognition Model Combining BERT Model and Lexical Enhancement

计算机科学, 2024, 51(6A): 230900030-6. <https://doi.org/10.11896/jsjcx.230900030>

[基于Transformer的司法文书命名实体识别方法](#)

Named Entity Recognition Approach of Judicial Documents Based on Transformer

计算机科学, 2024, 51(6A): 230500164-9. <https://doi.org/10.11896/jsjcx.230500164>

基于 CRF 的中文语法错误诊断系统的实现与应用

李 斌¹ 王浩畅²

1 湖南科技大学计算机科学与工程学院 湖南 湘潭 411201

2 东北石油大学计算机与信息技术学院 黑龙江 大庆 163318

摘 要 随着中国国际影响力的提高和汉语国际地位的提升,将中文作为第二语言学习的外国人数量逐年增加,中文已成为世界上最为流行的语言之一。基于此,中文语法错误诊断的研究备受关注。首先,从中文语法错误诊断的定义出发,总结目前的研究现状。其次,通过对各种中文语法错误诊断方法的分析,构建了基于条件随机场的中文语法错误诊断系统,探究中文语法自动检错系统及其具体应用流程,以帮助中文学习者提高学习效率。在 CGED2016 数据集上的实验结果表明,该系统在检测层和识别层上的性能较好,在位置层上还需要改进。

关键词: 中文语法错误诊断;序列标注;条件随机场;自然语言处理

中图分类号 TP391.1

Implementation and Application of Chinese Grammatical Error Diagnosis System Based on CRF

LI Bin¹ and WANG Haochang²

1 School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, Hunan 411201, China

2 School of Computer and Information Technology, Northeast Petroleum University, Daqing, Heilongjiang 163318, China

Abstract With the improvement of China's international influence and the worldwide status of Chinese, the number of foreigners who learn Chinese as a second language increases year by year, and Chinese has become one of the most popular languages in the world. Based on this, the research of Chinese grammatical error diagnosis has attracted much attention. This paper first summarizes the current research status from the definition of Chinese grammatical error diagnosis. Secondly, through the analysis of various Chinese grammatical error diagnosis methods, a Chinese grammatical error diagnosis system based on conditional random field (CRF) is constructed to explore the Chinese grammar automatic error detection system and its specific application process, so as to assist Chinese learners in improving their learning efficiency. Experimental results on the CGED2016 dataset show that the system performs well in the detection and identification levels and needs to be improved in the position level.

Keywords Chinese grammatical error diagnosis, Sequence annotation, Conditional random field, Natural language processing

1 引言

随着中国国际地位的显著提升以及“一带一路”国家战略的推动,中文学习已经成为全球的趋势。然而,由于中文语法结构与英文有明显差异,其复杂性和不确定性导致外国学习者在学习和交流中常出现各种语法错误。因此,中文语法错误诊断(Chinese Grammatical Error Diagnosis, CGED)的重要性日益凸显。一个准确的中文语法错误诊断系统能够帮助外国学习者及时发现语法错误,从而使其更好地学习中文,并深入了解中国文化。

中文语法错误诊断系统的任务是开发一个即时识别语言输入者输入的中文语句中语法错误的计算机辅助系统,是面向教育应用的自然语言处理研讨会(NLP Techniques for Educational Application Workshop, NLP-TEA Workshop)的评测任务之一,其目的在于推动中文语法诊断、纠错等相关研究,助力外国学习者更好地学习中文。CGED自2014年起已成功举办七届,主要将语法错误类型定义为以下4种:冗余

(R)、缺词(M)、选词错误(S)、词序错误(W);对系统性能从3个层面进行评估:检测水平(Detection Level,即识别句子是否有错误)、识别水平(Identification Level,即识别错误句子的具体错误类型)和位置水平(Position level,即识别错误的位置和对应类型)。该任务能够自动识别出语句所包含的错误类型和相应的错误位置,为学习者在使用中文进行写作时提供辅助性帮助,促进汉语语法规则的正确运用,提高汉语的使用效率和语法能力;同时,还可以使更多的人学习汉语以提升自身价值和竞争力,更好地传播中国文化,彰显文化自信,助推“一带一路”发展。

2 研究现状

中国国际地位的提升带动全球汉语学习热潮。同期纵比,英文语法错误诊断的准确度已达到较高水平,而中文语法错误诊断的性能仍然较低。为更好地帮助汉语走出去,满足中文学习者的需求,中文语法错误诊断系统的研发刻不容缓。

2014年,Yu等^[1]首次组织中文语法错误诊断共享任务,

基金项目:国家自然科学基金(61402099);黑龙江省自然科学基金(LH2021F004)

This work was supported by the National Natural Science Foundation of China(61402099) and Natural Science Foundation of Heilongjiang Province of China(LH2021F004).

通信作者:李斌(libin31209@163.com)

重点解决中文语法中常出现的 4 种语法错误。Cheng 等^[2]使用 CRF^[4]和支持向量机(Support Vector Machine, SVM)模型来检测和解决词序错误。2015 年, Yeh 等^[4]采用基于统计词和词性模式的特征提取方法, 再应用 CRF 模型来检测语法错误, 确定句子中的错误类型。2016 年, Zheng 等^[5]将该任务当成序列标注问题, 采用长短期记忆网络(Long Short-Term Memory, LSTM)^[6]模型进行句子语法错误标注。2017 年, Yang 等^[7]同样将任务视为序列标注问题, 提出 LSTM-CRF 模型, 采用 3 种集成策略来提高系统性能, 并在大赛中夺得冠军。此后, 该领域的研究者开始选用双向长短期记忆网络(Bi-LSTM)模型作为分析中文语法错误诊断任务的基本模型。

2018 年, 中文语法错误诊断任务有了飞跃式发展, 并在同年首次加入了中文语法纠错任务。Fu 等^[8]将该任务看作一个机器翻译任务, 使用分阶段的方法, 针对特定的错误设计特定的模板。Li 等^[9]将 BiLSTM-CRF 模型、神经机器翻译(Neural Machine Translation, NMT)模型和统计机器翻译(Statistical Machine Translation, SMT)模型组成一个混合模型, 以检测和纠正语法错误。Fu 等^[10]构建了一个基于 Bi-LSTM-CRF 模型的系统, 并整合一部分新特征, 如分词、高斯 ePMI、词性标注(POS)和点互信息(PMI), 以此来提高模型性能, 并在第五届 CGED 测评中获得冠军。

2019 年, BERT (Bidirectional Encoder Representations from Transformers)^[11]语言表征模型的使用, 在自然语言处理的大部分任务中产生了当时最好的效果, 促使研究人员也开始使用 BERT 处理 CGED 任务。2020 年, 第六届 CGED 测评中, Wang 等^[12]提出 ResBERT 检测模型, 在 BERT 模型基础上融入残差网络(ResNet), 以此来提高性能, 使模型可以更好地检测语法错误位置, 并在大赛中获得综合排名第一的成绩, 其多项核心指标保持领先。Luo 等^[13]提出使用句法依赖树和在多任务学习框架下将序列标记和序列到序列(Sequence-to-Sequence, Seq2seq)^[14]模型相结合两种方法, 其都是基于 BERT 模型。Cao 等^[15]提出了一种新的模型 BSGED, 该模型以较少的先验特征获得更好的结果, 大大减少了特征工程的工作量, 其是基于 BERT, Bi-LSTM 和 CRF 模型。

综上所述, 针对 CGED 任务的研究, 由最初的基于规则和基于统计的传统方法, 转化为基于深度学习的序列标注问题, 到后来将其看作机器翻译任务, 再到现在基于 BERT 的中文文本自动检错算法, 经历了逐步深化、不断创新的过程。从最新的研究成果来看, 序列标注任务仍是主流思路, BERT 框架也是采用此方法, 研究者们都是在预训练模型上进行微调, 将其转化为序列标注任务, 以此来提高模型性能。

3 中文语法错误诊断方法

目前常用的查错方法有如下 3 种: 基于规则的方法、基于统计的方法、基于深度学习的方法。

3.1 基于规则的方法

基于规则的方法是以人为制定的语言学规则为基础, 将待测文本置于此规则下与之进行对比, 对待测文本中与预设规则相符的字符串诊断为正确, 否则诊断为存在错误。该方法的优点是实现比较简单, 且检错的准确率较高; 缺点是准确率受规则的限制, 语法规则无法覆盖到所有的错误类型, 不够灵活且维护费用高。

3.2 基于统计的方法

基于统计的方法是对庞大的语料库进行统计分析, 从中提取语言规律的信息, 并构建语言模型, 将待测文本的字符串与之相匹配, 然后通过设置阈值来诊断待测文本是否存在错误^[16]。常使用朴素贝叶斯、N 元语法(N-gram)、SVM、CRF 等算法模型。该方法对语料库的构建要求很高, 并且规模需要足够大, 训练语料库的领域类型和大小直接影响准确率的高低, 还需对模型不断地调参优化等; 但是它的错效率和可维护性较高, 且扩展性较强。

3.3 基于深度学习的方法

基于深度学习的方法首先构建一个模型, 然后对语料集进行训练, 提取特征, 再使用此模型进行测试, 检测语法错误。基于该方法的主流错误诊断有以下两种。一是把错误诊断任务看成序列标注任务, 构建 LSTM 模型、循环神经网络(Recurrent Neural Network, RNN)^[17]模型、BERT 模型等, 然后再组合 CRF 模型等。此方法在识别和定位语法错误中效果甚佳, 但其过大的级联效应容易出现错误的扩大化和累积效应, 影响系统的性能。二是把错误诊断任务看成翻译任务, 使用机器翻译模型, 将包含语法错误的句子翻译成正确的句子。NMT 和 SMT 等机器翻译模型相继提出与使用。此方法不需要手动提取特征, 但是十分依赖训练数据, 前期对模型的训练时间较长, 耗费大量计算资源, 且对数据集的数量与质量要求很高。

目前, 中文语法错误诊断系统构建时并不是单纯使用某一种方法, 而是将多种方法进行融合。尽管基于深度学习的方法在中文语法错误诊断方面处于主导地位, 但是其面临较高的数据训练成本与低可维护性的问题, 其应用面临一定的挑战性。鉴于此, 本文将采用 PyLTP 进行数据预处理。由于数据集的多样性, PyLTP 的个性化分词特点更适用于这种丰富主题文本。此外, 通过分析现有中文语法错误诊断方法, 本系统把中文语法错误诊断任务看成是一个序列标注问题。针对序列标注问题, CRF 作为主流方法, 其灵活性和可解释性可以适应复杂的情况。因此, 将主要基于 CRF 模型开展中文语法错误诊断研究。首先, 对句子进行分词、词性标注、依存句法分析等数据预处理; 其次, 确定特征模板, 并使用 CRF++ 工具包训练序列标注模型; 最后, 对序列标注模型进行测试、评估和应用。

4 相关工作

4.1 PyLTP 系统

语言技术平台(Language Technology Platform, LTP)是哈工大社会计算与信息检索研究中心历时多年研制的一整套开放中文自然语言处理系统^[18], 旨在为用户提供高效精准的中文自然语言处理服务。其主要包括词法、句法、语义等多项中文处理核心技术, 曾多次在国内外技术评测中获得优异成绩。本系统采用的则是针对 Python 语言的封装包, 使用 PyLTP 系统对中文文本进行分词、词性标注、依存句法分析、语义角色标注、语义依存分析^[19]。

4.2 序列标注模型

将中文语法错误诊断任务视为序列标注任务。序列标注问题常见的 3 种模型是隐马尔可夫(Hidden Markov Model, HMM)^[20]、最大熵马尔可夫(Maximum Entropy Markov Model, MEMM)^[21]和 CRF 模型。其中, CRF 模型由 Lafferty

等提出,是一种无向图概率模型,在序列标注任务中效果显著,有效地解决了另两种模型出现的标注偏差以及标签之间的依赖关系信息问题。从问世以来,CRF 在分词、词性标注和命名实体识别等序列标注任务中取得了瞩目的成果,对特征的融合能力较强,且更多考虑的是整个句子的局部特征的线性加权组合。

4.3 特征工程

特征的选择和优化是影响结果的关键因素,特征选择的好坏,直接决定了模型性能的高低^[22]。本系统选择了 5 种特征输入到模型中。字符:采用字嵌入解决序列标注问题,将使用字符向量作为输入特征。词性标注:词性是最常被使用的特征之一,根据词语的词汇意义和语法特点进行分类,B-POS 表示开始字符的词性标注,I-POS 表示中间和结束字符的词性标注。分词信息:B-Segmentation 表示所属分词的开始字符的分词信息,I-Segmentation 表示所属分词中间或结束字符的分词信息。父节点:是当前单词进行依存句法分析后所对应的节点,B-head 表示所属分词开始字符的父节点,I-head 表示所属分词中间和结束字符的父节点。依存关系:表示当前分词进行依存句法分析后和对应父节点之间的关系,B-relation 表示所属单词开始字符的依存关系,I-relation 表示所属单词中间和结束字符的依存关系。

5 系统实现与应用

该文构建了基于 CRF 的中文语法错误诊断模型,其主要分为训练和测试两个部分,其结构如图 1 所示。

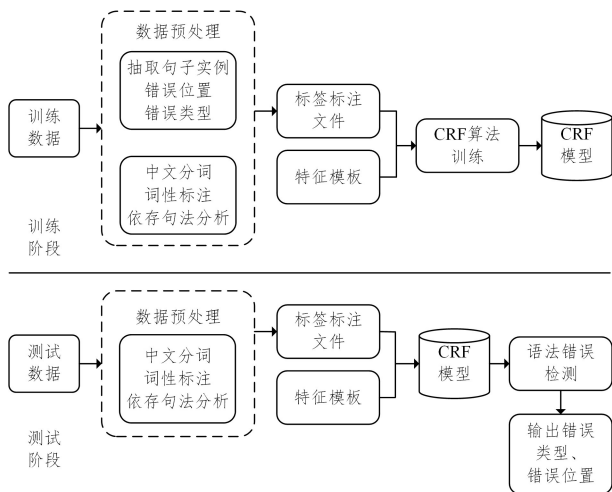


图 1 基于 CRF 的中文语法错误诊断研究框架

Fig. 1 Research framework of Chinese grammatical error diagnosis based on CRF

5.1 数据预处理

5.1.1 数据准备

本系统采用的是 CGED2016 年的 HSK 中文简体语料库,将其解压后可以看到内部含有 Training, Test 和 Evaluation 3 个文件夹。Training 文件夹下数据集中的所有语句都被用来训练中文语法错误诊断系统,每个语句中的语法错误注释和相应的纠正都是以 SGML 格式表示,如图 2 所示,语句中可能含有一个或多个错误,其错误位置和错误类型都已被标记出来。该数据集提供了共包含了 24 797 个语法错误的 10 071 条训练语句,其中包含 R(5 538)、M(6 623)、S(10 949)、W(1 687)。

```
<DOC>
<TEXT id=" 20200209550523150152_2x1" >
若父母主动和孩子平起平坐,倾心交谈,互相充分理解对方,会能够大幅减少互相抱怨的情况。
</TEXT>
<CORRECTION>
若父母主动和孩子平起平坐,倾心交谈,互相充分理解对方,会大幅减少互相抱怨的情况。
</CORRECTION>
<ERROR start_off=" 29" end_off=" 30" type=" R" ></ERROR>
</DOC>
```

图 2 SGML 格式的 HSK 训练语料库

Fig. 2 HSK training corpus in SGML format

5.1.2 数据处理

由于 CGED 任务涉及到识别错误边界问题,如将语句拆分成词将会在词与相应错误区间的端点之间存在偏差(见表 1),因此本系统将在字符层面解决该问题。除了错位区间问题外,还存在不同类型的错误区间可能与其他错误区间重叠(见表 2)。为避免这种重叠问题,分别处理这 4 种类型的错误。假设这 4 种类型的错误可能会相互影响,因此,对训练数据进行预处理时,通过删除最少的间隔数。最终,删除了一小部分可以接受的错误间隔。

表 1 位置偏差实例

Table 1 Example of position deviation

| | | |
|--|-----------------------------|-------|
| 人 ₁ 们 ₂ 越 ₃ 来 ₄ 重 ₅ 视 ₆ 自 ₇ 己 ₈ 的 ₉ 健 ₁₀ 康 ₁₁ ,12 希 ₁₃ 望 ₁₄ 没 ₁₅ 有 ₁₆ 污 ₁₇ 染 ₁₈ 的 ₁₉ 环 ₂₀ 保 ₂₁ 20-21 里 ₂₂ 生 ₂₃ 活 ₂₄ 。25 | | |
| 错误区间 | 15,15 | 21,21 |
| 错误类型 | M | S |
| 纠正 | 人们越来越重视自己的健康,希望在没有污染的环境里生活。 | |

表 2 错误区间重合实例

Table 2 Example of error interval coincidence

| | | |
|--|-----------------------|-----|
| 他 ₁ 们 ₂ 不 ₃ 知 ₄ 道 ₅ 吸 ₆ 烟 ₇ 会 ₈ 未 ₉ 成 ₁₀ 年 ₁₁ 人 ₁₂ 造 ₁₃ 成 ₁₄ 的 ₁₅ 各 ₁₆ 种 ₁₇ 伤 ₁₈ 害 ₁₉ 。20 | | |
| 错误区间 | 8,12 | 9,9 |
| 错误类型 | W | M |
| 纠正 | 他们不知道吸烟对未成年人会造成的各种伤害。 | |

对 SGML 格式的训练集进行预处理,抽取出句子实例、出错类型和出错位置,使用“O”表示正确的字符,“B-X”表示错误类型 X 的开始位置,“I-X”表示错误类型 X 的中间和最后位置。根据特征工程的模式,对句子实例做出如下预处理,以“别只能想自己,想你周围的人”为例说明。使用 PyLTP 分词系统对该语句进行中文分词、词性标注和依存句法分析,生成如图 3 所示的格式,从左至右,每一列的意义为字符、字符的词性标注、字符的分词信息、字符的父节点、字符的依存关系和字符的错误类型;并对 HSK 训练集的所有数据进行同样的处理。最后将处理后的数据进行保存。

```
别 B-d B-别 B-想 B-ADV O
只 B-d B-只 B-能 B-ADV O
能 B-v B-能 B-想 B-ADV B-R
想 B-v B-想 B-Root B-HED O
自 B-r B-自己 B-想 B-VOB O
己 I-r I-自己 I-想 I-VOB O
, B-wp B-, B-想 B-WP O
想 B-v B-想 B-想 B-COO B-M
你 B-r B-你 B-周围 B-ATT B-M
周 B-nd B-周围 B-人 B-ATT O
围 I-nd I-周围 I-人 I-ATT O
的 B-u B-的 B-周围 B-RAD O
人 B-n B-人 B-想 B-VOB O
。 B-wp B-。 B-想 B-WP O
```

图 3 预处理之后的训练数据

Fig. 3 Pre-processed training data

5.2 序列标注模型训练

使用 CRF++ 进行模型训练时,前提是要构建特征模板。特征模板是深入扩展特征选择,可将不同特征进行组合以获得目标词的一维新特征体系,其中的特征项可表示特征文件中任意项与其前后项的对应关系^[23]。因此,为了使模型

发挥较好的性能,需要选择合适的特征模板的构造方法。本系统利用启发式的方法,根据窗口大小和实验结果,定义和选取了 32 组特征项,如表 3 所列。选取训练文件及特征模板后,对模型进行训练,再通过多次调参和迭代最终生成模型。

表 3 特征模板

Table 3 Feature template

| 特征模板 | 特征模板 |
|---------------------------------|--|
| U01: %x[0,0]/%x[0,1] | U017: %x[0,1]/%x[1,1] |
| U02: %x[0,0]/%x[-1,0] | U018: %x[0,0]/%x[1,0]/%x[1,1] |
| U03: %x[0,0]/%x[-1,1] | U019: %x[0,1]/%x[1,0]/%x[1,1] |
| U04: %x[0,1]/%x[-1,0] | U020: %x[0,0]/%x[2,0] |
| U05: %x[0,1]/%x[-1,1] | U021: %x[0,0]/%x[2,1] |
| U06: %x[0,0]/%x[-1,0]/%x[-1,1] | U022: %x[0,1]/%x[2,0] |
| U07: %x[0,1]/%x[-1,0]/%x[-1,1] | U023: %x[0,1]/%x[2,1] |
| U08: %x[0,0]/%x[-2,0] | U024: %x[0,0]/%x[2,0]/%x[2,1] |
| U09: %x[0,0]/%x[-2,1] | U025: %x[0,1]/%x[2,0]/%x[2,1] |
| U010: %x[0,1]/%x[-2,0] | U026: %x[0,0]/%x[0,2] |
| U011: %x[0,1]/%x[-2,1] | U027: %x[0,0]/%x[-1,2] |
| U012: %x[0,0]/%x[-2,0]/%x[-2,1] | U028: %x[0,2]/%x[-1,2] |
| U013: %x[0,1]/%x[-2,0]/%x[-2,1] | U029: %x[0,0]/%x[-1,0]/%x[-1,1]/%x[-1,2] |
| U014: %x[0,0]/%x[1,0] | U030: %x[0,1]/%x[-1,0]/%x[-1,1]/%x[-1,2] |
| U015: %x[0,0]/%x[1,1] | U31: %x[0,0]/%x[-2,2] |
| U016: %x[0,1]/%x[1,0] | U32: %x[0,1]/%x[-2,2] |

5.3 序列标注模型测试

模型训练完之后,需对该模型性能进行测试。实验中所用的测试数据是 CGED2016 中的测试集,Test 文件夹下共有 3011 个句子,其中包含 1539 个正确的句子,每个错误句子至少有一处语法错误,共有 3695 条语法错误的的数据,其错误类型分布如表 4 所列。

表 4 HSK 测试集错误类型分布

Table 4 HSK test set error type distribution

| 数据集 | # 错误 | R | M | S | W |
|-----|----------------|-----------------|-----------------|------------------|----------------|
| HSK | 3695 (100%) | 802 (21.71%) | 991 (26.82%) | 1620 (43.84%) | 282 (7.63%) |

首先对测试数据抽取句子实例,利用 PyLTP 分词系统对该语句进行中文分词、词性标注和依存句法分析等预处理,生成“单个字符+字符词性标注+字符分词信息+字符父节点+字符依存关系”五列的格式,不同列之间通过空行作为分隔符,并将测试语料库预处理结果进行保存并命名为 Test_Input.txt。然后将处理好的数据通过已经训练好的模型进行测试,经过重定向将结果进行保存并命名为 Test_Output.txt。与 Test_Input.txt 文件相比,Test_Output.txt 文件多了最后一列,即对应字符预测的错误类型。对于测试数据中的语句,如该语句不包含语法错误,则系统输出“correct”;如该语句包含语法错误,则输出“start_off,end_off,error_type”,其中 start_off 是语法错误发生时字符的开始位置,end_off 是语法错误发生时字符的结束位置,error_type 是 R,S,M 和 W 错误类型中的一种。最后将系统所有的输出结果保存为 Test_Predict.txt,以便评估该系统的性能。

5.4 序列标注模型性能评估

表 5 所列为用于评估系统性能的混淆矩阵。在这个矩阵中,TP(真正)是存在语法错误的语句被系统准确识别出含有语法错误语句的数量,FP(假正)是不存在语法错误的语句被系统识别为含有语法错误语句的数量,TN(真负)是不存在语

法错误的语句被系统准确识别出不含有语法错误语句的数量,FN(假负)是存在语法错误的语句被系统识别为不含有语法错误语句的数量。

表 5 混淆矩阵

Table 5 Confusion matrix

| 混淆矩阵 | | 系统结果 | |
|------|----------|--------------------|--------------------|
| | | Positive(错误的) | Negative(正确的) |
| 黄金标准 | Positive | TP(True Positive) | FN(False Positive) |
| | Negative | FP(False Positive) | TN(True Negative) |

系统主要从检测(Detection)、类型识别(Identification)和位置(Position) 3 个水平来判断模型性能,并通过准确率 P(Precision)、召回率 R(Recall)和 F-Measure(F1)这 3 个评测指标来对 3 个水平进行衡量,得到评测的结果,从而来评估该模型的性能。其评测指标计算公式如下:

$$P = \frac{TP}{TP + FP} \quad (1)$$

$$R = \frac{TP}{TP + FN} \quad (2)$$

$$F1 = \frac{2PR}{P + R} \quad (3)$$

将本系统评测结果与 CGED2016 共享任务^[24]中每支参赛团队运行最好的官方数据评测结果进行对比,3 个水平下的性能评估结果对比如表 6 所列。从表中可以看到,在检测层中,本系统的准确率(P)仅落后于 SKY;但召回率(R)和 F-Measure(F1)均优于 SKY,基于深度学习的方法在 R 得分上表现优异,但其 P 得分却不尽人意;HIT 的 F1 得分最高,但其 P 得分低于本系统。综合可知,本系统的 P,R 和 F1 得分相对来说较高,能更高效地识别句子的错误。在识别层中,SKY 的 P 得分最高,但 R 和 F1 得分均落后于本系统;HIT 的 F1 得分最高,但 P 低于本系统。虽然本系统 R 得分稍低,但综合来看,本系统识别错误句子的错误类型的准确率较好。在位置层中,各团队之间的系统性能差异很

大,从 0.0034 到 0.3696,本系统有待加强。

从实验的结果来看,传统的方法在自然语言处理任务中相比现流行的神经网络方法依然具有一定的优势,仍被广泛应用于 CGED 任务中。CRF 作为一种具有灵活特征空间的序列模型,被 CYUT, HIT 和 SKY 团队在它们各自的系统中选择使用。本系统也是基于 CRF 模型,通过精心设计的模板

一样可以在任务中和神经网络保持同样的性能;此外, CYUT 系统也只基于 CRF 模型,除了位置层性能比本系统高外,检测层和识别层的性能都低于本系统。使用 CRF 模型在汉语语法错误的自动检测方面具有良好的性能,尤其是错误率和准确率;但对于中文语法位置错误,被系统检测到的错误总体数量偏少,影响了系统整体评价。

表 6 实验结果对比

Table 6 Comparison of experimental results

| 团队 | 实现方法 | Detection | | | Identification | | | Position | | |
|---------------|------------|-----------|--------|--------|----------------|--------|--------|----------|--------|--------|
| | | P | R | F1 | P | R | F1 | P | R | F1 |
| 本系统 | CRF | 0.6493 | 0.6277 | 0.6383 | 0.5758 | 0.4432 | 0.5009 | 0.1797 | 0.1624 | 0.1706 |
| CYUT-Run2 | CRF | 0.5951 | 0.6440 | 0.6186 | 0.5238 | 0.4509 | 0.4846 | 0.2034 | 0.2225 | 0.2125 |
| SKY-Run2 | CRF+N-garm | 0.8746 | 0.3505 | 0.5005 | 0.8821 | 0.2972 | 0.4446 | 0.7054 | 0.2217 | 0.3373 |
| SKY-Run3 | CRF+N-garm | 0.8652 | 0.3750 | 0.5232 | 0.8744 | 0.3185 | 0.4669 | 0.7144 | 0.2430 | 0.3627 |
| HIT-Run3 | CRF+LSTM | 0.6071 | 0.7296 | 0.6628 | 0.5002 | 0.5447 | 0.5215 | 0.3695 | 0.3697 | 0.3696 |
| NCYU-Run2 | RNN+LSTM | 0.4964 | 0.9755 | 0.6580 | 0.2588 | 0.5263 | 0.3470 | 0.0158 | 0.0217 | 0.0183 |
| PKU-Run2 | Bi-LSTM | 0.4945 | 0.8254 | 0.6185 | 0.2765 | 0.3594 | 0.3125 | 0.0595 | 0.0923 | 0.0724 |
| YUN-HPCC-Run2 | CNN/LSTM | 0.4886 | 0.7113 | 0.5793 | 0.2681 | 0.4565 | 0.3378 | 0.0022 | 0.0070 | 0.0034 |

5.5 序列标注模型性能应用

将本系统应用于智能对外汉语学习系统^[25]中。智能对外汉语学习系统是针对对外汉语学习者开发的,主要以对话形式,结合聊天机器人来帮助学习者模拟真实情境的对话训练。聊天过程中,本系统通过对学习者输入的语句进行分词、词性标注和句子结构的解析提取特征,然后再通过事先构建好的 CRF 模型对语句的语法错误进行检测,并将检错结果及时返回到系统界面。学习者将文本输入系统进行聊天时,系统会自动检测文本是否存在语法错误。若存在,学习者将被提示是否选择对语法错误进行修改,如果选择修改,基于语法学习的目标,系统将提示学习者文本语法错误的具体位置以及错误的类型,并引导学习者自主进行修改和完善;如果忽视修改,基于对话训练的目标,系统将根据输入的文本进行语句相似度分析,以反馈恰当的对话内容,保障对话的完整性与流畅性^[26]。在此过程中,通过创设聊天情境,提高了学习者的中文综合运用能力和中文实际情境操作能力。

因此,本系统不仅可以进行中文语法错误类别检测,还能识别出具体的错误位置。通过跟踪定位、即时反馈语法错误,学习者能形成对错误的直观认识,增强语言纠错意识,带动大脑中语言知识的迁移和强化,有利于培养学习者自我查错能力和自我评价能力。同时,学习者能够利用本系统主动参与写作活动,学会和掌握语法应用及写作技巧,减少重复出错率,增强运用中文写作的信心,从而全面提升自主写作能力。

结束语 本文提出了一种基于 CRF 的中文语法错误诊断模型,总的来说,CRF 在自动检测汉语语法错误方面具有很大的潜力,没有严格的独立性假设,特征设计灵活。在数据预处理阶段,系统在字符层面决定处理语法纠错问题,与词层面相比,避免了词和错误位置间隔偏差的问题,提高了系统模型的性能;系统运用 PyLTP 工具包对数据进行处理,很大程度上简化了文本的自然语言处理过程,提高了运行效率;系统使用的特征中除了使用基本的字符和词性特征外,又集成了分词信息、父节点和依存关系到 CRF 模型中,极大提高了系统的准确率。

但本系统仅能进行中文语法错误的诊断,还未形成对应的纠错方案。因此,在未来的工作中,将加强语料库的构建,

获取和收集更多的数据集,以拓宽模型运用范围,提高模型性能;会结合目前比较流行的神经网络模型,实现识别语法错误的同时能够直接纠正错误;会根据输入文本数据,更高效地整理出中文语法的固定规律,从而进一步优化和完善模型。

放眼当下,精准把握时代发展契机,越来越多的教育者参与到中文语法错误诊断模型的研究中,同时中文语法错误诊断大赛也在持续进行,笔者相信中文语法错误诊断留有限作为的空间,其不仅有利于提升中华文化自信,更好地为我国文化走出去提供实践路径,同时为外国学习者学习中文提供了智能工具,激发各国民众学习中文的热情,促进中外友好发展,为一带一路创造更多发展活力。

参考文献

- [1] YU L C, LEE L H, CHANG L P. Overview of grammatical error diagnosis for learning Chinese as a foreign language[C]// Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications. Nara, Japan: Asia Pacific Society for Computers in Education, 2014: 42-47.
- [2] CHENG S M, YU C H, CHEN H H. Chinese Word Ordering Errors Detection and Correction for Non-Native Chinese Language Learners[C]// Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. 2014: 279-289.
- [3] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]// Proceedings of the 18th International Conference on Machine Learning. 2001: 282-289.
- [4] YE H J F, YE H C K, YU K H, et al. Condition Random Fields-based Grammatical Error Detection for Chinese as Second Language[C]// Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications. Beijing, China: Association for Computational Linguistics and Asian Federation of Natural Language Processing, 2015: 105-110.
- [5] ZHENG B, CHE W, GUO J, et al. Chinese Grammatical Error Diagnosis with Long Short-Term Memory Networks[C]// Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications. Osaka, Japan: Natural

- Language Processing Techniques for Educational Applications, 2016;49-56.
- [6] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. *Neural Computation*, 1997, 9(8):1735-1780.
- [7] YANG Y, XIE P, TAO J, et al. Alibaba at IJCNLP-2017 Task 1: Embedding Grammatical Features into LSTMs for Chinese Grammatical Error Diagnosis Task[C]//Proceedings of the IJCNLP: Shared Tasks, 2017:41-46.
- [8] FU K, HUANG J, DUAN Y. Youdao's Winning solution to the NLPCCC-2018 Task 2 challenge: A neural machine translation approach to Chinese grammatical error correction[C]//Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing. Cham: Springer, 2018: 341-350.
- [9] LI C, ZHOU J, BAO Z, et al. A Hybrid System for Chinese Grammatical Error Diagnosis and Correction[C]//Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018:60-69.
- [10] FU R, PEI Z, GONG J, et al. Chinese Grammatical Error Diagnosis using Statistical and Prior Knowledge driven Features with Probabilistic Ensemble Enhancement[C]//Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018:52-59.
- [11] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019:4171-4186.
- [12] WANG S, WANG B, GONG J, et al. Combining Resnet and Transformer for Chinese Grammatical Error Diagnosis[C]//Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications. Suzhou, China: Association for Computational Linguistics, 2020:36-43.
- [13] LUO Y, BAO Z, LI C, et al. Chinese Grammatical Error Diagnosis with Graph Convolution Network and Multi-task Learning [C]//Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications. Suzhou, China: Association for Computational Linguistics, 2020:44-48.
- [14] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2014:3104-3112.
- [15] CAO Y, HE L, RIDLEY R, et al. Integrating BERT and Score-based Feature Gates for Chinese Grammatical Error Diagnosis [C]//Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications. Suzhou, China: Association for Computational Linguistics, 2020:49-56.
- [16] CHEN C. Research on the Diagnosis and Repair of Short Text Error Based on Web[D]. Wuhan: Central China Normal University, 2018.
- [17] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent Neural Network Regularization[J]. arXiv:1409.2329, 2014.
- [18] ZHAO Y N. The Study of Automatic Function Information Extraction and Classification Approach for Chinese Patent [D]. Tianjin: Hebei University of Technology, 2016.
- [19] WANG S S. Research on Construction Technology of Question Bank Based on Knowledge Graph [D]. Chongqing: Chongqing University of Posts and Telecommunications, 2021.
- [20] LIN Q M, QI Z Z. Research on Speech Emotion Recognition Based on HMM and ANN Mixed Model[J]. *Computer Technology and Development*, 2018, 28(10):74-78.
- [21] LANG B, FAN Y N. Personalized Learning Behavior Evaluation Method Based on Deep Neural Network[J]. *Computer Technology and Development*, 2019, 29(7):6-10.
- [22] ZHANG Y P. Research on the Construction Technology of Military Equipment Knowledge Graph [D]. Xi'an: Xidian University, 2021.
- [23] HAN W Y. Question Answering Oriented Automatic Chinese Grammatical Error Diagnosis Method [D]. Harbin: Harbin Institute of Technology, 2015.
- [24] LEE L H, RAO G Q, YU L C, et al. Overview of NLP-TEA 2016 Shared Task for Chinese Grammatical Error Diagnosis [C]//Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications. Osaka, Japan: Natural Language Processing Techniques for Educational Applications, 2016:40-48.
- [25] LI B, WANG H C. Design and Research of Intelligent Foreign Chinese Learning System[J]. *Computer Technology and Development*, 2022, 32(3):15-20.
- [26] LI B. Research and Implementation of Teaching Chinese as Foreign Language System Based on Chatbot [D]. Daqing: Northeast Petroleum University, 2019.



LI Bin, born in 1993, master, lecturer. His main research interests include natural language processing and intelligence education.