文章编号：1003-0077(2017)06-0019-09

基于 LSTM 和 N-gram 的 ESL 文章的语法错误自动纠正方法

谭咏梅，杨一枭，杨 林，刘姝雯

（北京邮电大学 计算机学院，北京 100876）

摘 要：针对英语文章语法错误自动纠正(Grammatical Error Correction，GEC)问题中的冠词和介词错误，该文提出一种基于 LSTM(Long Short-Term Memory，长短时记忆)的序列标注 GEC 方法；针对名词单复数错误、动词形式错误和主谓不一致错误，因其混淆集为开放集合，该文提出一种基于 ESL(English as Second Lauguage)和新闻语料的 N-gram 投票策略的 GEC 方法。该文方法在 2013 年 CoNLL 的 GEC 数据上实验的整体 F_1 值为 33.87%，超过第一名 UIUC 的 F_1 值 31.20%。其中，冠词错误纠正的 F_1 值为 38.05%，超过 UIUC 冠词错误纠正的 F_1 值 33.40%，介词错误的纠正 F_1 为 28.89%，超过 UIUC 的介词错误纠正 F_1 值 7.22%。

关键词：语法错误自动纠正；LSTM；N-gram 投票策略；ESL 语料

中图分类号：TP391 **文献标识码**：A

Grammatical Error Correction Using LSTM and N-gram

TAN Yongmei, YANG Yixiao, YANG Lin, LIU Shuwen

(School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: To deal with the incorrect usage of articles and prepositions in GEC (Grammatical Error Correction) area, this paper proposes a sequence labeling method. As for incorrect usage of noun form, verb form and subject-verb agreement, this paper proposes an N-gram voting strategy based on corpus collected from ESL (English as Second Language) essays and news. The results show that the method in this paper on CoNLL (2013) corpus achieves an overall F_1 score of 33.87%, outperforming the top ranked UIUC's F_1 score (31.20%), and a 38.05% F_1 score for article errors and 28.89% for preposition errors, both exceeding UIUC's result (33.40% for article errors and 7.22% for preposition errors, respectively).

Key words: grammatical error correction; LSTM; N-gram voting strategy; ESL corpus

0 引言

英语是当今世界上最主要的国际通用语言，全球有超过 10 亿人将英语作为第二语言使用。语法(syntactic)错误是 ESL(English as Second Language)学习者最常犯的一种错误[1]。语法错误自动纠正(Grammatical Error Correction，GEC)指利用计算机对文章进行自动语法错误纠正。

冠词错误、介词错误、名词单复数错误、动词形式错误和主谓不一致错误是 ESL 学习者常犯的五类语法错误[2]。它们对文章智能评改系统的性能影响最大，因此本文重点关注并解决这五类错误。

这五类错误中，冠词和介词错误的变化形式有限，可将其看作序列标注问题，且长短时记忆(Long Short-Term Memory，LSTM)对于序列标注问题效果较好，因此，本文提出了一种基于 LSTM 的序列标注 GEC 方法。名词单复数错误、动词形式错误和主谓不一致错误变化形式多样，所对应的混淆集为开放集合。本文提出一种基于 ESL 和新闻语料的 N-gram投票策略 GEC 方法。

收稿日期：2017-07-26 **定稿日期**：2017-10-11

1 相关工作

GEC开始于20世纪80年代,Writer's Workbench主要使用规则进行语法错误纠正,随后出现了基于句法分析的Epistle系统,1993年微软的Word基于拓展短语结构语法(Augmented Phrase Structure Grammar)对文本进行语法错误纠正。

LSTM由Sepp Hochreiter和Jurgen Schmidhuber于1997年提出[3],通过设置输入门、输出门、遗忘门等,解决了循环神经网络(Recurrent Neural Network)的梯度消失和信息的长期依赖问题,在处理序列问题中,效果较为突出。

随着各种规模语料库的出现,基于语料库的统计方法成为有效的GEC方法。HOO在2011年、2012年连续举办了两年相关评测任务[4-5],CoNLL在2013年、2014年继续举办了相关评测任务[2,6]。

GEC方法主要可以分为基于规则和基于统计两类。基于规则的方法主要依赖语言学家编写的语法规则,可分为以下两类:

(1) 基于上下文无关规则驱动的方法。其主要依赖语言学家编写的语法规则进行错误检查[7],少量规则对系统不够实用,大量规则则会出现互相矛盾的问题。该方法局限性太大,错误检查范围有限。

(2) 基于简单统计的规则驱动的方法在提取规则的时候考虑了上下文[8],可以有效避免规则的错误使用,但错误纠正范围仍然有限。

基于统计的GEC方法[9-12],即使用机器学习的方法对英文写作中的错误进行纠正时,纠正的性能依赖于语料库的构建。本文在进行识别和纠正时,将新闻语料、ESL语料和纠正后的中国学生写的英文文章语料(PIGAI语料①)混合使用。

在本文的N-gram方法中,使用大量的新闻语料进行N-gram的频次统计,以用于对名词、动词、主谓错误等的识别和纠正;在神经网络模型中,对不存在语法错误的语料进行人工错误生成,以平衡语料之间的差异并补充用于模型训练的语料。

对于冠词和介词的纠正,传统的GEC方法使用N-gram或者基于规则的方法对语法错误进行纠正。单纯的使用固定窗口大小的上下文信息进行纠正,信息使用并不充分,且当窗口大小变大时,难以对模型进行训练。LSTM网络模型可以学习到决定介词或者冠词使用的长期依赖信息,并且可以避免传统循环神经网络中可能发生的梯度消失等问题。

因此,本文将冠词和介词错误看作一项特殊的序列标注任务,提出一种基于LSTM的序列标注GEC方法。在训练时,使用ESL语料和补充语料,对特定冠词或介词进行标注。针对名词单复数错误、动词形式错误和主谓不一致错误,其混淆集为开放集合。提出一种基于ESL和新闻语料的N-gram投票策略的GEC方法。

2 基于LSTM和N-gram的ESL文章的GEC方法

基于LSTM和N-gram的ESL文章的GEC方法系统架构如图1所示。

针对冠词和介词错误,将其看作一项特殊的序列标注任务,该文提出一种基于LSTM的序列标注GEC方法。首先,对于已有词性标注的训练语料进行预处理,将冠词词性用一个特殊标记"ART"代替,将介词词性用一个特殊标记"TO"代替,把上述标记与冠词或介词的位置进行对换。然后,使用LSTM进行模型训练。最后,将训练得到的模型用于测试数据。

针对名词单复数错误、动词形式错误和主谓不一致错误,混淆集为开放集合,提出基于ESL和新闻语料的N-gram投票策略的GEC方法。

2.1 N-gram搜索服务及知识库

2.1.1 N-gram搜索服务

语法错误的纠正策略基于N-gram的频次统计,因此需首先建立N-gram搜索服务。使用的N-gram② 来源为约12 000个新闻网站2006年的所有新闻,统计信息如表1所示。

表1 N-gram详细信息

窗口大小	文章数/万	句子数/亿	单词数/亿
1-5	1 460	1.26	34

为了提高其查询效率,使用开源搜索引擎solr③对其建立倒排索引,提供搜索服务。

2.1.2 知识库

冠词和介词的变化形式有限,都处于封闭集合内。针对冠词和介词建立有限混淆集。

① www.pigai.org

② http://webscope.sandbox.yahoo.com/catalog.php?datatype=l

③ https://lucene.apache.org/solr/

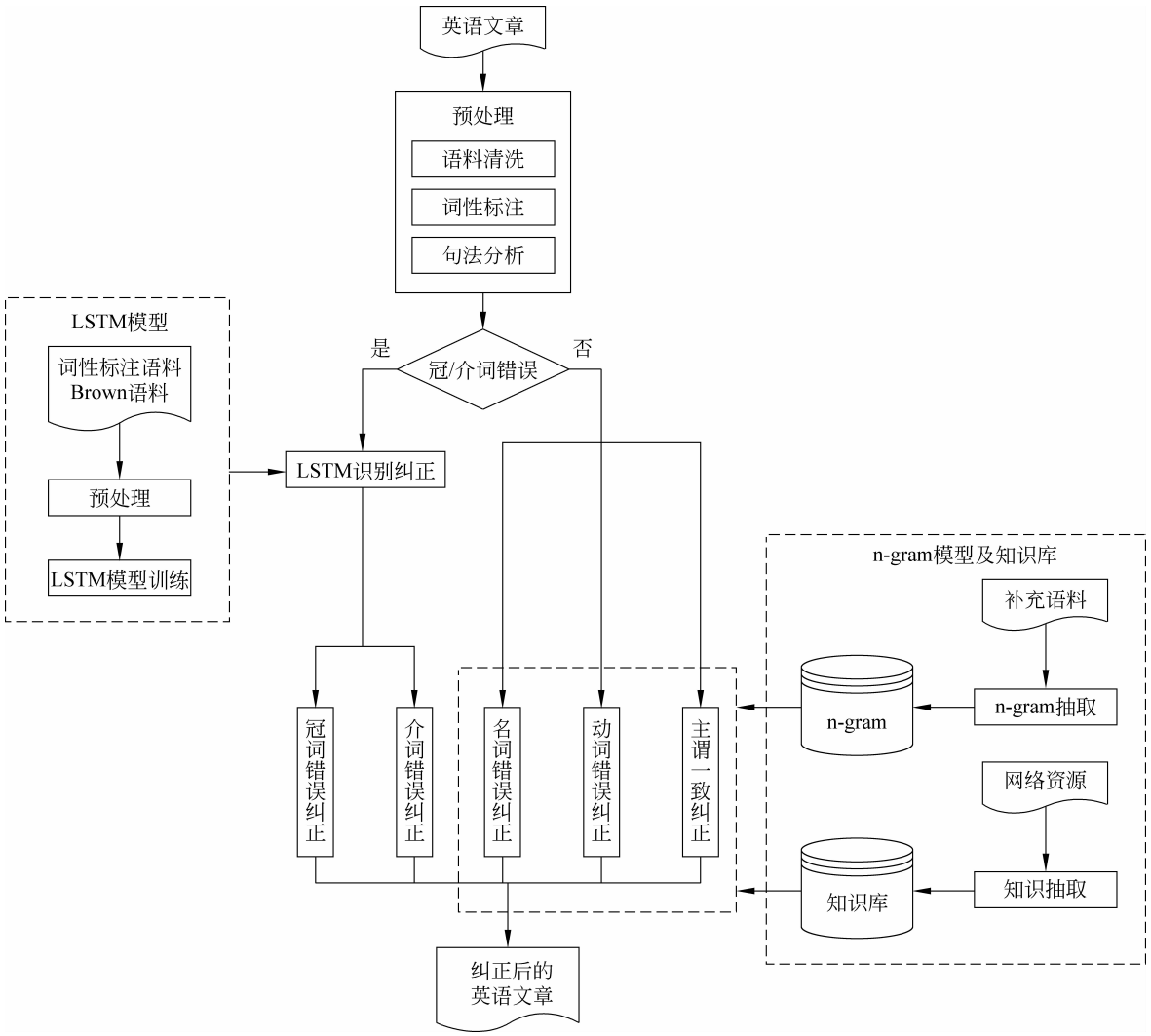

图 1　系统架构图

名词及动词不像冠词和介词那样变化形式有限，其变化形式是开放集合。因此针对名词错误、动词形式、主谓不一致错误分别建立变化表。

冠词混淆集　冠词混淆集包含三种情况：the，a/an，null。null 代表不使用冠词。

介词混淆集　介词混淆集包含常见的 17 个介词：on，about，into，with，as，at，by，or，from，in，of，over，to，among，between，under，within。

名词单复数变化表　名词单复数变化表包括：名词单数、名词复数，如表 2 所示。

表 2　名词单复数变化表

类型	单数	复数
规则	word，...	words，...
不规则	person，...	people，...
单复数一致	fish，...	fish，...

动词形式变化表　动词形式变化表主要包括：动词原形、过去式、过去分词、现在分词，如表 3 所示。

表 3　动词形式变化表

类型	动词原形、过去式、过去分词、现在分词
规则	work，worked，worked，working...
不规则	write，wrote，written，writing have，had，had，having...

动词单复数变化表　动词单复数变化取决于其主语单复数形式。动词单复数变化表主要包括：动词单数，动词复数，如表 4 所示。

表 4　动词单复数变化表

类型	单数、复数
规则	plays，play carries，carry...
不规则	has，have is/am，are...

2.2 移动窗口及 N-gram 投票策略

对混淆集为开放集的 GEC 方法，基于移动窗口[10]及 N-gram 投票策略。

2.2.1 移动窗口

移动窗口（Moving Window）定义如式（1）所示。

$$MW_{i,k}(w)=\{w_{i-j},\cdots,w_{i-j+(k-1)},j=0,k-1\} \tag{1}$$

w_i为句中第 i 个单词，k 代表窗口大小，j 为窗口内第一个单词与 w_i的距离。如表 5 所示。

表 5 移动窗口

窗口大小	N-grams
$k=5,j=[0,4]$	as it reduces the **chances** it reduces the **chances** of reduces the **chances** of wrong the **chances** of wrong doings **chances** of wrong doings that
$k=4,j=[0,3]$	it reduces the **chances** reduces the **chances** of the **chances** of wrong **chances** of wrong doings
$k=3,j=[0,2]$	reduces the **chances** the **chances** of **chances** of wrong
$k=2,j=[0,1]$	the **chances** **chances** of

窗口大小 k 的选择和 j 的取值范围直接影响着 GEC 的效果，针对不同的错误类型，选择不同的 k，j 值。

2.2.2 N-gram 投票策略

本策略模拟现实生活中的投票表决机制，含语法错误候选的 N-gram 片段代表一个可能具有投票权利的候选人。由于语料库有限，N-gram 片段的频次可能出现非常稀疏的情况。本策略设置一个最小有效频次，只有当查询到的频次高于最小有效频次时，此 N-gram 片段才具有投票权利。

在现实生活中，不同的人针对不同领域所投的票的重要性是不一样的，例如：领域专家的投票重要性高于普通人。本策略使用 N-gram 片段长度模拟领域专家的专业程度，N-gram 越长所投票的重要性越高。

最后，针对投票结果，得到纠正后的结果。具体算法如图 2 所示。

n-gram 投票策略

输入：n-grams 片段集合 Nset[window][candidates]，最小有效频次集合 Fset[window]，权重集合 Wset[window]，n-grams 搜索服务 Ngram。

输出：投票结果集合 Rset[candidates]。

```
from window=3 to 5:
    maxFreq=0,vote=null
    for candidate in Nset[window]:
        freq=Ngram.getFreq(candidate)
        if (freq>=Fset[window]andfreq>maxFreq):
            maxFreq=freq
            vote=candidate
    Rset[vote]+=Wset[window]
return Rset.Max
```

图 2 N-gram 投票策略

Fset 和 Wset 为参数。Fset 为最小有效频次，只有当查询到的频次大于 Fset 时才可参加投票。Wset 用于调整不同长度 N-gram 片段投票的权重，N-gram 片段长度越长其权重越大。

此算法基于语料库，由于一方面语料库规模有限不可能包含所有的片段，另一方面语料库中存在噪音数据。所以，设置最小有效频次 Fset，只有当查询出的 N-gram 片段频次大于此频次，才能说明语料库中包含相关语料，此 N-gram 片段具有投票权利。依据实验对比，将 Fset 设置为 100。

具有投票权利的 N-gram 片段的频次代表改为相应结果的概率。假设修改冠词错误时，“have an apple”的频次为 2，“have the apple”的频次为 1，“have apple”的频次为 1。那么，根据语料库改为“have an apple”的概率将大于“have the apple”及“have apple”。而投票策略是要在“have an apple”、“have the apple”及“have apple”中选出一个作为投票对象，本策略选择概率大的作为投票对象。

2.3 基于 LSTM 的标注纠正策略

该文对混淆集为固定集合的语法错误使用基于 LSTM 的标注纠正策略。

2.3.1 LSTM 模型原理

在进行序列数据的标注时，当前单词的标注信息一般依赖于上下文信息。传统的序列标注方法依赖统计或者融合的方法[13-14]，而循环神经网络通过建立隐藏层的序列关系，可以很好的提取序列信息[3]。其中，LSTM 通过设置门限单元和 Cell，可以有效避免传统循环神经网络在训练时可能会出现的

梯度消失和梯度爆炸等问题[15]。

相比单向的 LSTM 模型仅仅累积当前时刻之前的信息，双向的 LSTM 可以累积当前时刻的上下文信息，使得模型可以综合上下文信息进行序列的标注。

2.3.2　基于 LSTM 的标注纠正策略

基于 LSTM 的标注模型如图 3 所示。模型首先将单词转换成单词向量作为模型的输入，每个时刻输入序列中相应位置的单词向量。在训练的过程中，词向量作为参数进行更新。模型使用词向量作为 LSTM 单元的输入，并在每个时刻，输出相应的标注向量。其中，标注集合为所有的词性集合和所有介词或者冠词的混淆集的并集。该标注向量的维度和标注集合总数一致，并通过 softmax 选择概率最大的标注进行标记。

标注模型依赖 BPTT(back propagation through time)算法[3]，使用随机梯度下降的方式进行监督训练。

本文针对混淆集固定的语法错误，即冠词和介词进行标注纠正。在标注之前，将序列中的冠词或者介词使用统一的标识进行表示。在标注时，统一的标识被标注成具体的介词或者冠词，实现语法中冠词或介词错误的纠正。

例如，在进行冠词纠错时：

原句：

"*Debate on* **the** *legislation*, *which faces* ***a*** *veto threat from president Bush*, *is to continue today*."

首先，将其处理为：

"*Debate on* ***ART*** *legislation*, *which faces* ***ART*** *veto threat from president Bush*, *is to continue today*."

即，使用统一的符号"ART"进行代替句子中所有的冠词作为输入。

标注模型输出为：

"*NN IN* ***the*** *NN*, *WDT VBZ* ***a*** *NN NN IN NNP NNP*. *VBZ TO VB NN*."即，冠词部分标注为具体的冠词。其余部位输出相应的词性标注。

2.3.3　人工错误生成

因新闻语料和 ESL 语料之间存在差异，故模型在训练时，对新闻语料进行人工错误补充，以减小语料之间的差异。根据语法错误的类型，随机选择句子中的动词或者名词进行形式的变化。例如，随机地将名词的单复数形式进行修改，将动词的时态进行修改等。

2.4　冠词错误识别与纠正

冠词错误主要包括：冠词误用，冠词冗余，冠词缺失。错误类型举例如下。

1) 冠词误用：

例　"It is also ***the*** advance of surveillance technology."

将"***the*** advance"改为"***an*** advance"。

2) 冠词冗余：

例　"It gives ***the*** police a better control of the criminal."

将"***the*** police"改为"police"。

3) 冠词缺失：

例　"Government had to uninstall all the devices in the end."

将"Government"改为"**The** *government*"。

本文将冠词错误纠正看作一项特殊的序列标注任务，涉及三个子模块：冠词错误预处理模块、冠词错误识别与纠正模块和冠词错误后处理模块。

因冠词缺失在冠词错误中的占比较小[2]，本文主要处理前面两种错误类型。

2.4.1　冠词错误预处理模块

将冠词词性用一个特殊标记"*ART*"代替，把词性与冠词的位置进行对换。这样，句子中所有出现冠词的地方都被替换为"*ART*"，而其对应的词性则被修改为此处应该出现的冠词。如"**A_DT** record_NN date_NN has_VBZ n't_RB been_VBN set_VBN ._."处理为"**ART_A** record_NN date_NN has_VBZ n't_RB been_VBN set_VBN ._."。

2.4.2　冠词错误识别与纠正模块

根据给定句子，判断句子中可能存在冠词使用错误的位置，对句子进行词性标注，然后识别出所有词性被标注为冠词(a，an，the)的地方。使用基于 LSTM 的序列标注方法进行冠词错误识别与纠正。

基于 LSTM 的序列标注 GEC 方法，系统架构如图 3 所示。其中，w_n 为输入的待纠正句子的第 n 个单词，t_n 为输出的纠正后句子的第 n 个单词。首先，将输入句子中的每个单词转换为词向量表示；然后，经过两层 LSTM 模型，得到标注结果。

2.4.3　冠词错误后处理模块

将上一步骤的结果中为特殊标记"ART"的单词与词性标记进行对换，再将词性标记删除，得到最终输出结果。

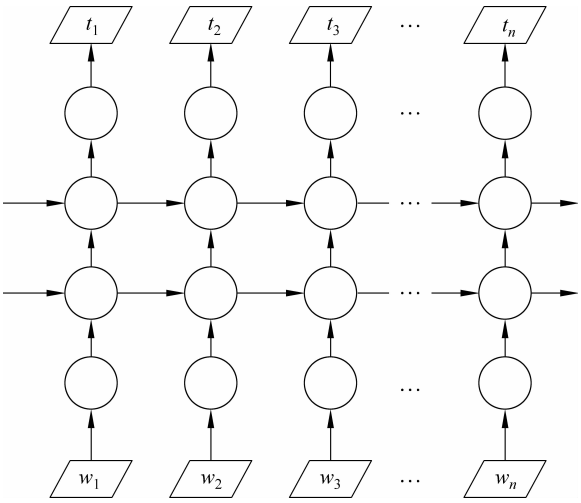

图 3 基于 LSTM 的冠词、介词错误识别与纠正方法

2.5 介词错误识别与纠正

介词错误主要包括：介词误用、介词冗余、介词缺失。错误类型举例如下。

1）介词误用：

例 "Pets are supposed to be chained when they are out ***on*** the streets."

将"***on*** the streets"改为"***in*** the streets"。

2）介词冗余：

例 "In that case, they would have no choice but to seek ***for*** the power of loyal police."

将"seek ***for***"改为"seek"。

3）介词缺失：

例 "Although we are not implanted ***with*** chips we are exposed under CCTV."

将"implanted"改为"implanted (***with***)"。

将介词错误纠正看作一项特殊的序列标注任务，涉及三个子模块：介词错误预处理模块、介词错误识别与纠正模块和介词错误后处理模块。

因介词缺失在介词错误中的占比较小[2]，本文主要处理前面两种错误类型。

2.5.1 介词错误预处理模块

将介词词性用一个特殊标记"TO"代替，把词性与介词的位置进行对换。这样，句子中所有出现介词的地方都被替换为"TO"，而其对应的词性则被修改为此处应该出现的介词。如"Some_DT 0_CD institutions_NNS are_VBP part_NN **of_IN** the_DT pension_NN fund_NN ._."处理为"Some_DT 0_CD institutions_NNS are_VBP part_NN **TO_of** the_DT pension_NN fund_NN ._."。

2.5.2 介词错误识别与纠正模块

根据给定句子，判断句子中可能存在介词使用错误的位置，对句子进行词性标注，然后识别出所有词性被标注为介词的地方。

使用基于 LSTM 的序列标注方法进行介词错误纠正，系统架构如图 3 所示。

2.5.3 介词错误后处理模块

将上一步骤的结果中为特殊标记"TO"的单词与词性标记进行对换，再将词性标记删除，得到最终输出结果。

2.6 名词单复数错误纠正

名词单复数错误纠正模块基于名词单复数变化表及 N-gram 投票策略，主要针对名词单复数误用情况进行纠正。此模块具体纠正过程举例说明如下：

例 "This will, if not already, caused problems as there are very limited ***spaces*** for us ."

将"***spaces***"改为"***space***"。

1）对例句词性标注得到其词性序列，并提取词性标注为 NN 和 NNS 的单词得到错误候选集合 E={problems, spaces}；

2）使用名词单复数变化表得到相应的纠正候选集合。如：spaces 的纠正候选集合 C={space, spaces}；

3）基于纠正候选集合，使用大小为 3—5 的移动窗口获取 N-grams 片段集合。使用 N-gram 投票策略得到得票最高的纠正候选，在原句中进行替换。如果"space"为得票最高纠正候选，则纠正后的句子为"This will, if not already, caused problems as there are very limited ***space*** for us ."

2.7 动词及主谓不一致错误纠正

动词错误纠正模块主要针对动词形式误用情况及主谓不一致情况进行纠正。此模块依赖于动词形式变化表、动词单复数变化表及 N-gram 投票策略，具体纠正过程举例如下：

动词形式错误：

例 "The more people ***using*** it over us, the more power they will have on us."

将"***using***"改为"***use***"。

主谓不一致：

例 "Every move of us ***are*** easily tracked."

将"***are***"改为"***is***"。

1）对句子词性标注得到其词性序列。针对动词形式错误，提取词性标注为 VB、VBD、VBG、VBN 的单词作为其错误候选。针对主谓不一致错误，提取词性标注为 VBP、VBZ 的单词作为其错误候选。

2）根据错误候选及动词形式变化表/动词单复数变化表得到错误候选的纠正候选集合。

3）针对纠正候选，使用大小为 3—5 的移动窗口获取 N-grams 片段集合。使用 N-gram 投票策略得到得票最高的纠正候选，并在原句中进行替换。

3　实验

3.1　实验数据

实验数据来源于 CoNLL2013 的 GEC 评测任务，统计结果如表 6 所示。由于 CoNLL2013 语料没有正确的词性标注，且 CoNLL2013 训练语料较 PIGAI 词性标注语料[13]和 Brown 语料①规模较小。因此，使用 PIGAI 词性标注语料、Brown 语料和标注后的 CoNLL 语料等扩充 LSTM 训练语料标注时，用 Stanford 标注工具对其进行词性标注。其中，CoNLL2013 语料和 PIGAI 语料作为 ESL 语料，Brown 语料作为补充的新闻语料参与模型的训练。

表 6　CoNLL2013 的 GEC 评测任务数据统计

错误类型	训练集		测试集	
	数目/个	占比/%	数目/个	占比/%
冠词	6 658	14.8	690	19.9
介词	2 404	5.3	312	9.0
名词	3 779	8.4	396	11.4
主谓一致	1 453	3.2	122	3.5
动词形式	1 527	3.4	124	3.6
五种类型	1 5821	35.1	1 644	47.4
所有类型	4 5106	100.0	3 470	100.0

CoNLL2013 的 GEC 评测任务数据里标注了多种错误类型，但评测任务主要是针对冠词错误、介词错误、名词错误、主谓一致和动词形式错误这五种占比较高[2]的错误类型。

3.2　评价方法

CoNLL2013 评价标准为 F_1[2]，定义如式(2)所示。

$$F_1 = \frac{2 * P * R}{P + R} \tag{2}$$

其中 P 与 R 分别表示准确率和召回率，定义如式(3)、式(4)所示。

$$P = \frac{N_{correct}}{N_{predicted}} \tag{3}$$

$$R = \frac{N_{correct}}{N_{target}} \tag{4}$$

$N_{correct}$ 指系统修改正确的错误的数目，$N_{predicted}$ 指系统修改的错误的数目，N_{target} 指语料本身存在的错误的数目。

3.3　实验结果及分析

基于 LSTM 和 N-gram 的 ESL 文章的 GEC 方法在 CoNLL2013 的 GEC 评测数据上的实验结果如表 7 到表 9，并与基于语料库的英语文章语法错误检查及纠正方法[16]和 2013 年评测第一名 UIUC[12]进行比较。

如表 7 所示，针对冠词错误的纠正，本文的方法的 F_1 值比 UIUC 方法高 5%，比 Corpus GEC 方法高 5%。针对介词错误的纠正，本文方法的 F_1 值比 UIUC 方法高 21%，比 Corpus GEC 方法高 13%。表明基于 LSTM 的序列标注 GEC 方法对冠词和介词语法错误纠正任务有效。这是由于词向量包含丰富的上下文信息，而使用 LSTM 更好地学习到了决定冠词或者介词使用的长期的依赖信息，所以结果较好。

表 7　冠词及介词错误纠正结果

错误类型	方法比较	P	R	F_1
冠词错误	UIUC	0.478 4	0.256 5	0.334 0
	Corpus GEC	0.434 9	0.271 2	0.334 5
	LSTM GEC	0.331 5	0.446 4	***0.380 5***
介词错误	UIUC	0.265 3	0.041 8	0.072 2
	Corpus GEC	0.134 6	0.188 3	0.157 0
	LSTM GEC	0.192 1	0.582 0	***0.288 9***

如表 8 所示在仅使用 N-gram＋vote 投票策略对名词及动词错误纠正时，F_1 值与 UIUC 方法都还存在一定的差距。这是由于 N-gram ＋ vote 策略基于的新闻语料与所需纠正的 ESL 文章具有差异性，会将大量正确句子改为错误句子。名词及动词变化表不能涵盖所有的名词及动词的变化形式，导致纠正名词及动词时还具有一定的局限性。

① http://www.nltk.org/nltk_data/

表8 名词及动词错误纠正结果

错误类型	方法比较	P	R	F_1
冠词错误	UIUC	0.522 3	0.383 8	0.442 5
	N-gram+vote	0.326 6	0.573 2	0.416 1
动词错误	UIUC	0.389 4	0.178 9	0.245 1
	N-gram+vote	0.160 2	0.219 5	0.185 2

如表9所示,对于全部五种类型错误的纠正,本文方法均优于UIUC方法,在2013年CoNLL的GEC数据上总的F_1值为33.87%,超过第一名UIUC总的F_1值31.20%。总的实验结果表明基于LSTM和N-gram的ESL文章的语法错误自动纠正方法是有效的。

表9 所有类型错误纠正结果

错误类型	本文方法			UIUC
	P	R	F_1	F_1
冠词	0.331 5	0.187 5	***0.239 5***	0.18
+介词	0.261 4	0.297 6	***0.278 3***	0.19
+名词	0.279 0	0.435 8	***0.340 2***	0.29
+主谓一致	0.275 3	0.460 1	***0.344 5***	0.30
+动词形式	0.265 2	0.468 7	***0.338 7***	0.31

4 结束语

针对冠词和介词错误,本文提出一种基于LSTM的序列标注GEC方法。针对名词单复数错误、动词形式错误和主谓不一致错误,本文提出一种基于N-gram投票策略的GEC方法。在2013年CoNLL的GEC评测数据上,针对冠词错误纠正F_1为38.05%,介词错误的纠正F_1为28.89%,所有五种类型错误的总F_1为33.87%,均高于评测第一名UIUC。实验结果表明,本文方法对冠词及介词错误的纠正是有效的,但仍有一些问题存在。例如,在介词缺失和冠词缺失时如何进行纠正;在纠正名词单复数错误及动词错误时,如何避免将正确句子改为错误句子;及动词形式中如果出现被动语态错误时该怎么纠正等。这些问题仍需进一步研究解决。

参考文献

[1] Kukich K. Techniques for automatically correcting words in text[J]. ACM Computing survey (CSUR), 1912,24(4),377-439.

[2] Ng H T, Wu S M, Wu Y, et al. The CoNLL-2013 Shared Task on Grammatical Error Correction[C]// Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, 2013: 1-12.

[3] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation 9,8(1997),1735-1780.

[4] Kilgarriff A. Helping our own: the HOO 2011 pilot shared task[C]//Proceedings of the European Workshop on Natural Language Generation. Association for Computational Linguistics,2011: 242-249.

[5] Dale R, Anisimoff I, Narroway G. HOO 2012: a report on the preposition and determiner error correction shared task[C]//Proceedings of the Workshop on Building Educational Applications Using Nlp. Association for Computational Linguistics,2012: 54-62.

[6] Ng H T, Wu S M, Briscoe T, et al. The CoNLL-2014 Shared Task on Grammatical Error Correction[C]// Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, 2014: 1-14.

[7] 吴伟成,周俊生,曲维光. 基于统计学习模型的句法分析方法综述[J]. 中文信息学报,2013,27(3): 9-19.

[8] 董喜双,关毅. 基于有监督学习的依存句法分析模型综述[J]. 智能计算机与应用,2013,3(2): 11-15.

[9] Rozovskaya A, Chang K W, Sammons M, et al. The University of Illinois System in the CoNLL-2013 Shared Task[C]//Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task,2013.

[10] Kao T H, Chang Y W, Chiu H W, et al. CoNLL-2013 Shared Task: Grammatical Error Correction NTHU System Description[C]//Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task,2013: 20-25.

[11] Felice M, Yuan Z, Øistein E. Andersen, et al. Grammatical error correction using hybrid systems and type filtering[C]//Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task,2014: 15-24.

[12] Rozovskaya A, Sammons M, Dan R. The UI system in the HOO 2012 shared task on error correction [C]//Proceedings of the Workshop on Building Educational Applications Using Nlp. Association for Computational Linguistics,2013: 272-280.

[13] 谭咏梅,吴坤. 面向英语文章的词性标注算法[J]. 北京邮电大学学报,2014,37(6): 120-124.

[14] 郭永辉,吴保民,王炳锡. 一种用于词性标注的相关投票融合策略[J]. 中文信息学报,2007,21(2): 9-13.

[15] Hochreiter Y, Bengio P Frasconi, J Schmidhuber.

Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-term Dependencies[M]. S. C. Kremer and J. F. Kolen, editors, A Field Guide to Dynamical Recurrent Neural Networks. WiLey-IEEE Press, 2001.

[16] 谭咏梅，王晓辉，杨一枭. 基于语料库的英语文章语法错误检查及纠正方法[J]. 北京邮电大学学报，2016，39(4)：92-97.

谭咏梅(1975—)，通信作者，博士，副教授，主要研究领域为计算语言学、自然语言处理、机器学习、知识工程。
E-mail：ymtan@bupt.edu.cn

杨一枭(1992—)，硕士，主要研究领域为自然语言处理、机器学习。
E-mail：yang1xiao1@163.com

杨林(1992—)，硕士，主要研究领域为自然语言处理、机器学习。
E-mail：wximo@live.com

热烈祝贺倪光南院士获得全国“最美科技工作者”称号

2018 年 6 月 14 日，新闻联播报道由中宣部、科技部和中国科协联合主办的“最美科技工作者”评选结果揭晓，评选出张弥曼、多吉、倪光南、严纯华、邹学校、李贺军、李兴钢、蔚保国、秦川、王杜娟等 10 位先进典型，他们有的矢志不移自主创新，将核心技术牢牢掌握在自己手里；有的“板凳甘坐十年冷”，用科研成果赢得世界同行尊重；有的扎根基层一线，为扶贫攻坚和人民幸福付出毕生精力；有的投入社会公益，几十年如一日开展科普宣传……他们以实际行动，生动诠释了中华民族伟大创造精神、伟大奋斗精神、伟大团结精神、伟大梦想精神的真谛，展示了新时代中国科技工作者的良好精神风貌。

发布仪式以“科技中国梦建功新时代”为主题，现场播放了反映 10 位“最美科技工作者”先进事迹的视频短片，采访讲述了他们的工作生活感悟，并向他们颁发了“最美科技工作者”证书。

其中，倪光南院士为中国中文信息学会经由中国科协“优秀科技工作者代表”推荐获此殊荣，下面简要介绍倪光南院士先进事迹：

倪光南，男，汉族，1939 年 8 月生，浙江镇海人。现任中国科学院计算技术研究所研究员，中国工程院院士。曾任中国中文信息学会第五、六届理事长。长期从事计算机领域研发工作，推广国产 CPU、国产软件等，主持研发“联想式汉卡”和联想系列微型机，首创在汉字输入中应用联想功能。曾获国家科技进步奖一等奖。曾任联想集团首任总工程师。第八届全国人大代表，第八、九届全国政协委员。