SOOCHOW UNIVERSITY

汉语文本纠错标注规范

Annotation Guidelines for Chinese Textual Error Correction

参与编者：章岳 李嘉诚
hillzhang1999@qq.com
指导老师: 李正华

2021 年 6 月 7 日版本
自 2021 年 2 月开始编制

苏州大学人类语言技术研究所
Human Language Technology Research Center, Soochow University

April 26, 2022

目录

1 前言

汉语文本纠错 (Chinese Textual Error Correction, CTEC) 任务, 指的是利用自然语言处理技术自动识别并纠正中文文本中所包含的语法错误、拼写错误、语序错误、标点错误等, 是中文自然语言处理的一项重要任务。下面就是汉语文本纠错的一个示例，我们用蓝色标注了原句子中出错的地方，用红色标注了纠正的结果。

错误句子: 那些空气污染也没有助于人生的身体建康。
正确句子: 那些空气污染也无助于人的身体健康。

例 1: 汉语文本纠错示例

随着深度学习的发展，基于神经机器翻译的文本纠错模型获得了较大的成功。但是，由于参数量倍增，基于神经机器翻译的文本纠错模型需要更大规模的高质量数据集进行训练。在英文上，已有一些学者进行了构建文本纠错数据集的尝试，包括人工标注及数据增强，并取得了成效。但在中文上，类似的工作还不多。数据的稀缺严重制约了汉语文本纠错模型的性能，因此，构建较高质量的大规模汉语文本纠错数据集非常有必要。

目前，汉语文本纠错数据集主要有 NLPCC2018-task2 所公开的 Lang-8 数据集 (约 120 万句)[3] 和北京语言大学的 HSK 数据集 (约 15 万句)[4]。当前公开的较大规模平行语料 (如 Lang-8 数据集) 都以类似网络众包的形式得到，并没有对标注规范进行严格地制定，也没有对标注者进行筛选和培训，故而数据质量不高，语料存在诸多不合规的地方。

为了帮助标注人员更高效准确地标注汉语文本纠错数据，我们编写了这份标注规范。本规范将：(1) 定义汉语常见文本错误分类体系，并给出相应的例句；(2) 介绍各类汉语文本错误的纠正方法，并给出纠错示例；(3) 介绍汉语文本纠错语料的标注准则和技巧。

现阶段，本规范主要针对的是汉语文本中出现的明显错误，而尽量不修改文本的表达、辞藻。在将来，我们还可能进行更深入的探索，尝试对汉语文本进行润色或升华，将不通顺、不优美的句子变得更加通顺、更上层次。

2 标注目的

我们标注汉语文本纠错数据集的主要目的是：

- 为基于机器学习、深度学习的汉语文本纠错模型提供较大规模、较高质量的错误句子-正确句子平行语料，改善目前已有的训练数据质量较低的现象，从而提升中文语法纠错系统的性能。
- 提供一份能够较好地评价已有汉语文本纠错系统的测试集，未来能够在此基础上组织公开评测。
- 构建汉语文本错误分类体系，为全面评价汉语文本纠错系统、定制化纠错、汉语文本纠错数据增强和清洗等打下基础。

3 标注提醒

3.1 标注准则

参考之前已有的文本纠错数据集建设工作 [5][6]，我们在这里简单地规定汉语文本纠错数据标注的准则：

(1) 标注者标注的数据绝大多数为纯中文数据，但如果遇到了中英文混搭的句子，那么对于中文需要纠正所有类型的错误，而对于英文只纠正拼写层面的错误。

(2) 如果遇到了句子中有表达情绪的标点，如“!!!”、“???”等，不需要做过多地修改。

(3) 标注者如果遇到了无法理解句义的句子，不应该自作主张推测句义，而是应该将其标注为“病句”，并跳过该任务。例如：“雨天适合你，跟你绑定，这个在类似绑定了链接，就在网店。”但是，如果句子的含义比较清晰，仅仅是表述有些混乱，则应该尽可能进行纠错，例如：“其实呃陈逸飞去世这个事情呢就是也蛮就是说我觉得蛮典型的。”

(4) 标注者对于句子中出现的谐音、方言等，不需要进行纠正，但对于此类问题应结合上下文和语境仔细分析。

(5) 对于句子中出现的结巴、重复现象，如果不是写作者有意为之，则需要进行纠正。

3.2 标注技巧

(1) 规范中给出了常见汉语文本错误及其纠正方式，在实际标注过程中要学会套用规范中的例子，用最合适的方式标注。

(2) 对于规范中没有明确给出，但有相近例子的错误类型及纠正，应学会灵活地举一反三。

(3) 标注过程中，对于拼写类型和词语误用类型的错误，要善于利用已有的词典资源，先充分理解，再标注。千万不要贸然标注，影响质量。可以利用的资源如：百度汉语 (hanyu.baidu.com)，在线新华字典 (https://zidian.aies.cn/) 等。

(4) 实际标注时，可能会遇到人名、地名、机构名等命名实体类型的拼写错误，对于这一类错误，要学会多查百度、谷歌等搜索引擎。

(5) 标注之前，请首先浏览一遍完整的句子，对句义有一个大体的了解。标注时应强调直觉，不应该对句义做过多地联想和揣测。但也不应完全忽视句义的作用，例如：“随着电影市场回暖，花木兰的票房越来越高。”如果我们忽略了句义，就无法得知句子中的“花木兰”指的是一部电影，从而就无法进行如下的标点符号纠正：“随着电影市场回暖，《花木兰》的票房越来越高。”

(6) 有时候，一个错误的句子可以对应多个正确的修改方案，例如“他弟弟已经写完那篇文章。”，既可以改成“他弟弟已经写完了那篇文章。”，也可以改成“他弟弟已经写完那篇文章了。”对于此类错误句子，标注者应该同时给出多个有把握的正确答案。

(7) 在句法中，汉语把句子成分划分为主要成分 (主语、谓语、宾语) 和附加成分 (定语、状语、补语)。以“他弟弟已经写完那篇文章”为例，“弟弟”是主语，“写”是谓语，“文章”是宾语，“他”和“那篇”分别作主语和宾语的定语，“已经”是谓语“写”的状语，“完”是谓语“写”的补语。分析句子时，先查主要成分，然后查附加成分，最后检查主要成分同附加成分之间的关系。这样，可以看出来句子是不是有问题。

(8) 即使构成句子的各种成分是完整无缺的，但是还要进一步检查成分之间在语义上是否配得上。例如：动词谓语“唱”可以同宾语“歌”结合动宾词组“唱歌”，而同“舞”就配不上，不能说“唱舞”。哪个词可以跟哪个词搭配，哪个词不能跟哪个词搭配，很多情况没有规律可循。即使下很大的功夫，词语搭配不当的现象仍是不可避免的。因此，这也是检查句子时不能忽略的方面。

(9) 一个句子如果在语法上找不出问题，那就看看在使用概念进行判断、推理时，是否讲得通。试看下边一个病句，“安娜买了香蕉、西瓜、草莓和很多水果”。显而易见，这是由于弄错了概念之间的关系造成的错误。“香蕉、葡萄、草莓”是种概念，“水果”是属概念。这句话误把属种关系的概念当作并列关系的概念。应改“水果”为表示种概念“苹果”、“桔子”一类的词。

4 汉语文本错误理论

4.1 文本错误的定义

语言是人类最重要的交际工具。人们以语言为转告信息或者自己的感情、想法。这样的语言是句子组成的。造句时需要一定的规则，如果说出来的话和出来的句子违背汉语的组合规律或者违背客观事物的事理，有碍交际，这样的句子就含有文本错误。 所谓文本错误，包括但不限于句子在词法、句法或者逻辑上的毛病。违背词法、句法或逻辑规则的句子就是通常所说的病句。正确的句子，是人们根据交流思想的需要，选用同所要表达的语义一致的词语，按照一定的结构规则组合而成的，是语义内容和结构形式的统一体。

4.2 文本错误的产生原因

产生文本错误的原因主要有以下三点 [7]：(1) 思维不清晰。语言是思维的工具，又是思维的成果，语言的准确首先取决于思维的清晰。尽管清晰的思维未必有准确的语言表达，但含混不清的语言，必然反映思维的混乱。(2) 认识不清楚。语言反映的复杂内容涉及到众多领域，各种门类及各个方面，倘若对所要表达的内容若明若暗，似懂非懂，

就不可能有准确的语言表达。(3) 组合不规范。只有按照语言规则组合起来的句子，才能恰当的表达一个概念，合理地造成一个判断和推理，否则就只能是一堆杂乱无章的堆积物。此外，还可能有修辞方面的原因。

4.3 文本错误的三个层面

文本错误在实际语言场景下，可以细分为以下这三个层面：

(1) 语法层面:语法层面的语病指的是句中词语和标点本身选择出现了错误，或者句中词语之间无法建立清晰完整的结构关系。其中包括：词语拼写出错、标点错用、词语冗余、词语丢失、语序出错等。

(2) 语义层面:语义层面的语病指的是句中词语和标点本身选择没有错误，但与客观对象联系起来，便出现了不相符合的矛盾，也就是违反了符号与符号之间在语义上的选择。其中包括：语义搭配不当、语义重复、语义歧义、语义含混、语义矛盾等。例如以下这个例子：

错误句子:妈妈看着我送的礼物，欣慰地笑着，无言以对。

正确句子:妈妈看着我送的礼物，欣慰地笑着，喜不自禁。

例 2: 语义层面错误

在上面的例子中，“无言以对”指的是没有话来应付，多用于理屈词穷时。这句话在语法上没有任何问题，但却是一种典型的语义搭配不当。

(3) 语用层面:语用层面的语病指的是说话人在言语交际中使用了符号关系正确的句子，却不自觉地违反了人际规范，社会规约或者不合时间、空间、不看对象，这类性质的错误就叫语用错误。例如以下这个例子：

错误句子:学生：“喂！王老师，快去给我修改作文！”

正确句子:学生：“王老师，请您帮我修改作文。”

例 3: 语用层面错误

在上面的例子中，我们可以得知说话者的身份是一名学生，而受话者的身份是一名老师，那么原句中说话的语气应该是尊敬的。

在实际标注过程中，我们希望标注者能够纠正 100% 的语法层面错误，对于较为明显的语义层面错误也需要进行纠正。除此之外，较为复杂的语义层面错误和语用层面错误不需要纠正。

5 错误类型详解

我们参考了北大版新汉语水平考试 (HSK) 攻略对汉语文本错误类型进行划分，并给定了一些错误类型的实例以供参考。

5.1 标点级别错误

标点符号的使用十分简单，根据写作需要和实际情况使用即可。但实际写作场景中，许多人在此方面却存在着不少问题。这一类错误的纠正不难，但需要标注人员细心观察。如果对标点符号的使用方法不确定，可以参考网络资源，例如：https://wenku.baidu.com/view/f8e1a838bb1aa8114431b90d6c85ec3a87c28b38.html。

5.1.1 标点冗余

标点冗余主要指的是在不必要的地方插入了标点。如下所示：

错误句子:所以一些人说 ，： “读书一点用处都没有。”
正确句子:所以一些人说: “读书一点用处都没有。”

例 4: 标点冗余

错误句子:他非常爱他的父、母亲。
正确句子:他非常爱他的父母亲。

例 5: 标点冗余

下面的例子中，“和”“及”等连词本身的功能就是连接并列成分，故中间不能用顿号，更不能用逗号，否则会造成功能重复。

错误句子:在这个寒冷的冬天，火把就是孩子心里的温暖，和希望。
正确句子:在这个寒冷的冬天，火把就是孩子心里的温暖和希望。

例 6: 标点冗余

下面的例子中，破折号的作用是解释说明，和“叫做”功能重复。

错误句子:社会上有一种人叫做—— “空想主义者”。
正确句子:社会上有一种人叫做 “空想主义者”。

例 7: 标点冗余

下面的例子中，省略号和“等”的功能重复，应删去任意一个。

错误句子:车辆人为损毁、违规占道、私人改装……等行为的出现，使我们感到痛心。

正确句子:车辆人为损毁、违规占道、私人改装等行为的出现，使我们感到痛心。

例 8: 标点冗余

标点冗余错误比较依赖语感来判断，额外插入的标点往往不符合汉语的停顿习惯。

5.1.2 标点丢失

标点丢失主要指的是在句中、句末漏写了本应存在的标点。这一类错误往往是因为书写不规范、粗心等原因造成的。如下所示：

错误句子:人为了生存不管是干净的空气还是污染的空气都要呼吸

正确句子:人为了生存，不管是干净的空气还是污染的空气，都要呼吸。

例 9: 标点缺失

此外，还有一种非常常见的标点漏用，是书名号、引号的漏用。一旦出现书名、篇名、报刊名、文件名、戏曲名、歌曲名、图画名等名称，需要在两侧加上书名号。如果要表示人说的话、讽刺、比喻、强调等，需要加引号。如下所示：

错误句子:妈妈说：你已经是个大孩子了。

正确句子:妈妈说：“你已经是个大孩子了”。

例 10: 标点缺失

5.1.3 标点误用

标点误用在日常中文写作中非常常见。比较典型的标点误用有以下几种情况：

(1) 实心点代替句号。受英文的影响，许多人会将中文句号（。）写成英文句号（.）。

(2) 逗号代替顿号。不少人没有掌握顿号的用法，不知道逗号和顿号使用的区别。

(3) 一逗到底。写作的过程中，部分人没有根据句意为句子打上合理的标点，而是全部使用逗号。这种情况会造成句子结构混乱，表达层次不明。如下所示：

错误句子: 这个时候，妈妈做了几道菜，我帮妈妈小心翼翼地放在桌子上，妈妈， 弟弟和我，我们三个人静静地坐在火堆旁等着爸爸回家吃饭，每次爸爸都要在这个时候回来的，可到现在爸爸还没回来。

正确句子: 这个时候，妈妈做了几道菜，我帮妈妈小心翼翼地放在桌子上。妈妈、 弟弟和我，我们三个人静静地坐在火堆旁等着爸爸回家吃饭。每次爸爸都要在这个时候回来的，可到现在爸爸还没回来。

例 11: 一逗到底和逗号、顿号混淆

(4) 疑问句、感叹句后面跟句号。如下所示：

错误句子:那我们一定要参加这个活动吗。

正确句子:那我们一定要参加这个活动吗?

例 12: 疑问句、感叹句后面跟句号

(5) 书名号和引号混淆。书名号是用于标明书名、篇名、报刊名、文件名、戏曲名、歌曲名、图画名等的标点符号，亦用于歌曲、电影、电视剧等与书面媒介紧密相关的文艺作品；双引号则表示引语、表示特定称谓、表示特殊含义（也表示否定和讽刺）、表示着重论述的对象、用于话语之中。二者比较容易混淆，如下所示：

错误句子: 我非常爱看“复仇者联盟”。

正确句子:我非常爱看《复仇者联盟》。

例 13: 书名号和引号混淆

5.2 拼写级别错误

拼写错误指的是人们在写作过程中，由于粗心或者知识匮乏，导致书写了错别字或者错误单词。这一类错误的纠正需要标注者充分利用已有的词典资源或者百科资源，对于有疑惑的地方要多查词典。

书写者犯下拼写错误，往往是在以下几种情况：(1) 想要写的字和实际写的字同音或者音近；(2) 想要写的字和实际写的字形近。

5.2.1 字音混淆错误

目前，绝大多数中国互联网用户都在使用拼音输入法，因此中文拼写错误主要是音同或者音近的错误类型。

例如，下面的例子中，“宿舍”和“舒舍”发音相近，导致了书写者的拼写错误。

错误句子:我们舒舍有四个人。

正确句子:我们宿舍有四个人。

例 14: 音近导致的拼写错误

5.2.2 字形混淆错误

除了拼音输入法以外，一些用户由于使用五笔输入法或其他字形输入法（如中国台湾的用户），也容易因为混淆字形而出现拼写错误。

例如，下面的例子中，“隘”和“溢”字形相近，导致了书写者的拼写错误。

错误句子:这座关溢非常雄伟。

正确句子:这座关隘非常雄伟。

例 15: 形近导致的拼写错误

5.2.3 词内部字符异位错误

实际标注过程中，标注者也可能会遇到词内部字符异位现象，例如:

错误句子:我非常爱吃阴冬功。

正确句子:我非常爱吃冬阴功。

例 16: 词内部字符异位错误

5.2.4 命名实体拼写错误

之前的拼写错误大都发生在常见的汉语单词上，除了它们以外，汉语中还存在着众多的命名实体词，如人名、机构名、地名以及其他所有以名称为标识的实体。这些词也非常容易产生拼写错误，需要标注者结合搜索引擎、百科资源等额外知识判断，如下所示:

错误句子:我们都是海南詹州人。

正确句子:我们都是海南儋州人。

例 17: 地名拼写错误

错误句子:唱过《中国人》的牛德华是著名的香港影星。

正确句子:唱过《中国人》的刘德华是著名的香港影星。

例 18: 人名拼写错误

对命名实体拼写错误进行标注时，标注人员一定要保持谨慎，只有百分百确认句子中的命名实体拼写出错的时候才能进行纠正，切勿过度纠错。

例如，下面的例子中，写作者的朋友可能本来就叫“牛德华”，因此千万不能将其修改成“刘德华”。标注者需要仔细对比例 18 和例 19，区分这两种情况。

错误句子:我有一位叫作牛德华的好朋友。
错误纠正:我有一位叫作刘德华的好朋友。
正确句子:我有一位叫作牛德华的好朋友。

例 19: 命名实体过度纠正现象

此外，尽管本规范主要关注中文语法纠错语料标注，但实际标注时，一些需要标注的句子中也可能存在少量英文等其他语言。如果发现其他语言存在拼写错误，标注者也可进行标注。

5.3 词语级别错误

词语级别错误往往指的是句子中单个词语或成语的使用出现了问题，但句子的句法结构并没有问题。这一类错误在汉语文本错误中属于最常见的一类，通常可以细分为以下三种错误：词语冗余、词语丢失和词语误用。

5.3.1 词语冗余

表示相同或相近意思的词语在句子中同时出现，会造成语义重复和句子累赘，这就是词语冗余，也称句子成分冗余。这一点主要涉及写作者对词语的理解和使用。

重复的两个词语常常紧挨着出现，因此读句子时，要注意相邻两个词的意思是否完全相同，若相同，则有可能犯词语冗余的毛病。下面例子中的“特别兴奋极了”，“特别”与“极了”重复，应删去一个。

错误句子:终于看到了大熊猫，儿子显得特别兴奋极了。
正确句子 1:终于看到了大熊猫，儿子显得特别兴奋。
正确句子 2:终于看到了大熊猫，儿子显得兴奋极了。

例 20: 词语冗余错误

下面例子中的“多达到数百个”，“达”和“到”重复，应删去“到”。

错误句子:中国古代地域辽阔，民族众多，历史上形成的传统节日多达到数百个。
正确句子:中国古代地域辽阔，民族众多，历史上形成的传统节日多达数百个。

例 21: 词语冗余错误

下面例子中的“历史的史册”，“历史”与“史册”重复，应删去“历史的”。

错误句子:布达拉宫维修工程将作为历史文化遗产保护和发展史上的一个丰碑而载入历史的史册。

正确句子:布达拉宫维修工程将作为历史文化遗产保护和发展史上的一个丰碑而载入史册。

例 22: 词语冗余错误

下面例子中，数词前的“约”和数词后的“左右”都表示“约数、大概”，语义重复，应删去一个。

错误句子:中国每年举办的达到一定规模的展会活动项目约有 3000 个左右。

正确句子 1:中国每年举办的达到一定规模的展会活动项目有 3000 个左右。

正确句子 2:中国每年举办的达到一定规模的展会活动项目约有 3000 个。

例 23: 数量词前后成分重复

下面例子中，“十分酷似”中，“酷”表示程度很高，没有必要再使用一个表示程度高的“十分”，应删去。

错误句子:如果我们要列举香港十分酷似动物的山峰，那么狮子山必定位居榜首。

正确句子:如果我们要列举香港酷似动物的山峰，那么狮子山必定位居榜首。

例 24: 表示程度的词语重复

一些句子似乎看不出有冗余的成分，但仔细分析句子中的某个词，会发现这个词所包含的意义还是与其他词互相交叉。标注者在平时标注时，应对这些词有所关注。

下面的例子中，“练就”这个词本来就有“练出来、练成功”的意义，因此之后的“出来”是冗余成分，应删去。

错误句子:因为经常与外国人接触，他练就出来了外交家的口才，很快同留学生成了好朋友。

正确句子:因为经常与外国人接触，他练就了外交家的口才，很快同留学生成了好朋友。

例 25: 词语冗余错误

下面的例子中，深爱”这个词表示“爱的很深”，本来就有程度很高的意思，因此之前的“十分”是冗余成分，应删去。

错误句子:他十分深爱他的祖国。

正确句子:他深爱他的祖国。

例 26: 词语冗余错误

下面的例子中，“堪”本身就表示“可以”，因此“可以”和“堪”语义重复。应删去“可以”。

错误句子:在钟的王国里，可以堪称世界古钟之王的是北京大钟寺里的华严钟。

正确句子:在钟的王国里，堪称世界古钟之王的是北京大钟寺里的华严钟。

例 27: 词语冗余错误

下面的例子中，“非常”和“奇缺”语义有重复，“奇缺”就是非常紧缺的意思。

错误句子:这是一种非常奇缺的材料，使用的时候一定不能浪费。

正确句子:这是一种奇缺的材料，使用的时候一定不能浪费。

例 28: 词语冗余错误

5.3.2 词语丢失

句子是由词或者词组构成的。现代汉语中的句子一般有六大成分，即主语、谓语、宾语、定语、状语、补语等。一个句子要表达一个完整的意思，它的结构也必须是完整的。“所谓结构完整，并不是说句子必须具备通常所说的六大成分，而是说句子应当由表达完整意思所必需的成分来构成。”必要的句子成分若是缺失，就会造成词语丢失现象的出现。

句子中的词语丢失，最常见的是主语、谓语和宾语等主要成分缺失，有时也有其他一些成分缺失的情况，如介词和关联词。

(1) 主语丢失：主语即是用来执行句子的动作或行为的主体。主语残缺，意味着整个句子“群龙无首”，也就让接受者无法明确主体是谁。在日常的对话或表达中常常存在主语省略的情况，如有明确的实际对话环境、承接上文主语或主语不言自明等。若非特殊情况而进行主语的省略，便可视为主语丢失。

下面的例子中，“经过”和“成为”应该有一个共同的主语，“他”放在“经过”这一谓语的后面就造成了主语的缺失，因此应把“他”提到“经过”的前面，充当两个谓语动词的主语。

错误句子:最终经过他的不懈努力，成为了一个地位很高的长官。

正确句子:最终经过不懈努力，他成为了一个地位很高的长官。

例 29: 主语丢失错误

下面的例子中，介词“从”导致主语缺失，应把“我”提到“从”之前。

错误句子:从我懂事开始就一直在为一个字奔波。

正确句子:我从懂事开始就一直在为一个字奔波。

例 30: 主语丢失错误

下面的例子中，“双赢”的主语应该有两个，句子中只给出了一个，这也属于一种主语丢失的情况。

错误句子:公司善待员工，员工也会尽全力地服务于公司，那么这个公司能达到双赢。

正确句子:公司善待员工，员工也会尽全力地服务于公司，那么这个公司和员工能达到双赢。

例 31: 主语丢失错误

(2) 谓语丢失：谓语紧跟主语之后，一般由动词充当，来表示主语的状态。在很多情况下，主语和宾语都可以省略，但谓语的作用更加重要。如上阵杀敌的时候，主将只需要说一个“打”字，士兵们就明白是让自己打敌人，所以谓语丢失的状况并不多见。但在梳理的过程中，仍然出现了一些谓语丢失的现象。

下面的例子中，在“重大”前需要添加一个谓语动词“做出”。

错误句子:在重大决定前，更要把握好机会。

正确句子:在做出重大决定前，更要把握好机会。

例 32: 谓语丢失错误

下面的例子中，“按照法规”是条件状语，缺乏修饰的谓语，应在“法规”后添加一个动词“行事”。

错误句子:只有严格按照法规，我们才能保证自己做的是正确的。

正确句子:只有严格按照法规行事，我们才能保证自己做的是正确的。

例 33: 谓语丢失错误

(3) 宾语丢失：简单而言，宾语是谓语的接受者。宾语和主语一样，在一定的情况下能够省略。但在不该省略的地方却省略了宾语，就会造成宾语丢失现象的出现。宾语的丢失，往往和谓语动词有密切的关系。

下面的例子中，在两个分句中，“有”都是充当谓语动词，而“惊天动地”和“执着奋进”则都是形容词充当定语，所以造成了“有”缺乏宾语、两个形容词缺乏修饰中心语情况的出现。可在“惊天动地”后加“的事业”，在“执着奋进”后加“的精神”

错误句子:想要有惊天动地，先要有执着奋进。

正确句子:想要有惊天动地的事业，先要有执着奋进的精神。

例 34: 宾语丢失错误

下面的例子中，句子的主干是“作家（主语）讲述（谓语）”，明显“讲述”的宾语缺乏，可在最后添加一个“的历程”。

错误句子:他讲述了自己在学生时代里成为一名优秀作家。

正确句子:他讲述了自己在学生时代里成为一名优秀作家的历程。

例 35: 宾语丢失错误

下面例子中，介词“对于”后面紧跟的都是作为定语出现的描述性内容，缺乏了宾语中心语，可在“劳动合同”后面添加“的事情”。

错误句子:对于一则新闻上报道“在上班时睡觉”可解除劳动合同，这样的做法太过苛刻。

正确句子:对于一则新闻上报道“在上班时睡觉”可解除劳动合同的事情，这样的做法太过苛刻。

例 36: 宾语丢失错误

(4) 其他丢失：词语丢失包括的种类很多，在调查分析的过程中，除了主谓宾的丢失也发现了其他类型的词语，但数量较少，故将其归为“其他丢失”一类。其中包括副词、介词、关联词、修饰语等的丢失。

例如，下面的例子属于介词丢失，应在“人性”前面添加一个“对”。

错误句子:当制度演变成了人性的束缚，它就可能引起社会的混乱。

正确句子:当制度演变成了对人性的束缚，它就可能引起社会的混乱。

例 37: 介词丢失错误

例如，下面的例子属于副词丢失，应该在“关羽”后面添加一个表示假设的副词“如果”。

错误句子:有时宽容也能害人，想当年关羽没有放过败走的曹操，历史可能改写。

正确句子:有时宽容也能害人，想当年关羽如果没有放过败走的曹操，历史可能改写。

例 38: 副词丢失错误

例如，下面的例子属于关联词残缺，表示选择的“不论... 还是...”结构，应在“黑暗的悬崖”前添加一个“还是”。

错误句子:不论前方是无尽的深渊，黑暗的悬崖，只要你迈出了朝着理想的第一步，那么你就已经成功了。

正确句子:不论前方是无尽的深渊，还是黑暗的悬崖，只要你迈出了朝着理想的第一步，那么你就已经成功了。

例 39: 关联词丢失错误

5.3.3 词语误用

词语使用不当也是汉语文本错误中一个十分重要的类别，主要产生的原因是写作者对某个单词的词义、词性了解不足。词语误用的实际情况有很多，但是也有一些典型的错误类型。这里，我们主要根据词性来划分错误类型。

(1) 动词使用错误：动词的主要功能是充当谓语，其分为及物动词和不及物动词。动词使用错误的原因主要有：1. 和宾语搭配不当；2. 词性误用；3. 词义误用；4. 及物动词和不及物动词混淆。

例如，下面的例子属于动宾搭配不当。“绽放”形容花开时由花蕾花瓣紧闭展开的样子，与“光辉”搭配不当，可改为“散发”。

错误句子:这样一个年过八旬的老奶奶在她即将逝去的生命中仍然绽放着希望的光辉。

正确句子:这样一个年过八旬的老奶奶在她即将逝去的生命中仍然散发着希望的光辉。

例 40: 动宾搭配不当错误

例如，下面的例子属于动宾搭配不当。“任命”的宾语应该是人，不能是“思想”。

错误句子:我们的党中央，曾经在红军长征时期，任命毛泽东的思想为指导思想，许多的劳苦大众的命运都交付于他的手中。

正确句子:我们的党中央，曾经在红军长征时期，选择毛泽东的思想为指导思想，许多的劳苦大众的命运都交付于他的手中。

例 41: 动宾搭配不当错误

例如，下面例子中的“喜爱”是动词，这里应使用名词，可改为“喜好”。

错误句子:他的政策完全是根据个人喜爱而设。

正确句子:他的政策完全是根据个人喜好而设。

例 42: 词性误用导致的动词使用错误

例如，下面的例子中，能愿动词使用不当，应把“不可能”改为“不能”。

错误句子:信心对于我们来说是十分重要的，绝对不可能失去。

正确句子:信心对于我们来说是十分重要的，绝对不能失去。

例 43: 词义误用导致的动词使用错误

例如，下面例子中的“启程”是不及物动词，后面不能直接带宾语，因此应改为“从北京启程了”。

错误句子:中午，我们就要启程北京了

正确句子:中午，我们就要从北京启程了

例 44: 不及物和及物混淆导致的动词使用错误

(2) 名词使用错误：名词的主要作用是充当主语和宾语。名词使用错误的原因主要是：1. 词性误用；2. 词义误用。

例如，下面例子中，“效果”是名词，在这里不能独立地作为动词出现，可改为“生效”。

错误句子:所以，我们要约束自己的行为，让规则可以更好地效果。

正确句子:所以，我们要约束自己的行为，让规则可以更好地生效。

例 45: 词性误用导致的名词使用错误

例如，下面的例子中，名词“梦餍”多指噩梦，句子所表达的色彩为褒义，这里可改为“梦乡”。标注者在标注词语误用类型的错误时，不仅要关注词义，也要关注词语的感情色彩。

错误句子:当我们在甜蜜的梦餍中酣畅淋漓时。

正确句子:当我们在甜蜜的梦乡中酣畅淋漓时。

例 46: 词义误用导致的名词使用错误

(3) 副词使用错误：副词多是用来表示程度、情状、时间和频率、范围、否定、语气等的词，其作用以修饰动词和名词为主。副词位置的不当会造成语病，对副词本身的意思的理解和适用情况缺乏了解也会造成副词的使用错误。

例如，下面的例子中，副词“亦”的出现应该表示不同的人或物，而本句前后的主语都是“人生”，用“亦”导致了句意的逻辑混乱，可将其改为“就”。

错误句子:当人生的转折路口出现迷茫而不知所措时，不要害怕，人生亦是如此。

正确句子:当人生的转折路口出现迷茫而不知所措时，不要害怕，人生就是如此。

例 47: 副词使用错误

例如，下面的例子中，应将副词“多么”改为“那么”。

错误句子:即使自己的生命多么短暂，也要向着自己的伟大事业追去啊。

正确句子:即使自己的生命那么短暂，也要向着自己的伟大事业追去啊。

例 48: 副词使用错误

例如，下面的例子中，“下降将近一倍”使用不当，因为表示减少的时候不能使用倍数。应该改为“下降将近一半”。

错误句子:在中国的一些景点，淡季的票价比旺季的票价下降将近一倍。

正确句子:在中国的一些景点，淡季的票价比旺季的票价下降将近一半。

例 49: 副词使用错误

(4) 关联词使用错误：关联词是指能在复句中用来连接分句，并表明分句之间关系的连词、副词和短语。关联词的使用非常容易混淆。

例如，下面的例子中，关联词“只要”使用不当，应该改为“只有”。

错误句子:只要准确用词，才能恰到好处地揭示事物特征，恰如其分地表达思想感情。

正确句子:只有准确用词，才能恰到好处地揭示事物特征，恰如其分地表达思想感情。

例 50: 关联词使用错误

例如，下面的例子中，关联词“因此”使用不当，这里分句之间是转折关系，应使用表示转折的关联词“但是”或“然而”。

错误句子:日本、朝鲜和越南都使用过汉字，因此情况各有不同。

正确句子:日本、朝鲜和越南都使用过汉字，但是情况各有不同。

例 51: 关联词使用错误

(5) 形容词使用错误：形容词主要用来修饰名词，也可以修饰动词。但在未说明的情况下，其不能被作他用，否则将造成错误。

例如，下面的例子中，“便捷”是一个形容词，在此不能直接作为动词役使“人们”，应改为动词“方便”。

错误句子:而没有法规的约束，那么就算是再便捷人们，再对人们有益的东西也终将变成一个问题。

正确句子:而没有法规的约束，那么就算是再方便人们，再对人们有益的东西也终将变成一个问题。

例 52: 形容词使用错误

(6) 介词使用错误：介词的种类有很多，其主要附着在代词、名词或是动词的前面组成一个介词结构。如果写作者对介词表示的意义不清楚，很容易造成介词的使用错误。

例如，下面的例子中，介词“关于”使用不当，应该改为“对”。

错误句子:全国各大媒体关于曹操墓的开掘工作行了报道。

正确句子:全国各大媒体对曹操墓的开掘工作行了报道。

例 53: 介词使用错误

例如，下面的例子中，介词“以”使用不当，应该改为“为了”。

错误句子:以找到这些稳定的工作，我们有读书的必要。

正确句子:为了找到这些稳定的工作，我们有读书的必要。

例 54: 介词使用错误

5.4 句法级别错误

句法级别的汉语文本错误涉及到句子层面，而不仅仅是单个词语或者字的问题。句法级别的错误往往是由于句子违反了通用的句法结构，或是句子的逻辑违背了客观事理。

5.4.1 词序不当

词序就是语句中词语的排列顺序。正确的词序对词语在句子中的位置有固定的要求。一旦词语的顺序发生了变化，句子的意思在很大程度上也会发生相应的改变。因此，如果不按照所想要表达的意思的需要去组合句子，就有可能导致句子结构的混乱，从而造成句子成分之间关系的失调，影响句子意思的表达。

例如，下面的例子中，“明显”是动词中心语“加快”的修饰语，应放在“加快”前作状语。

错误句子:改革开放后，中国的经济增长速度加快明显起来。

正确句子:改革开放后，中国的经济增长速度明显加快起来。

例 55: 中心词和修饰语颠倒

例如，下面的例子中，状语位置不当，应将“在各种领域”提到“变得”前面。

错误句子:便利的生存条件不会让我们变得轻松在各种领域。

正确句子:便利的生存条件不会让我们在各种领域变得轻松。

例 56: 状语位置不当

下面的例子中，“大量”是定语，应放在名词中心语“外国游客”之前。这是定语误放在状语位置的例子。

错误句子:中国的名胜古迹大量吸引了外国游客。

正确句子:中国的名胜古迹吸引了大量外国游客。

例 57: 定语误放在状语位置

下面的例子中，“正确”应该是修饰动词“处理”的，做状语，应放在“处理”之前。这是状语误放在定语位置的例子。

错误句子:我们都知道，处理与朋友的正确关系是十分重要的。
正确句子:我们都知道，正确处理与朋友的关系是十分重要的。

例 58: 状语误放在定语位置

此外，当多个动词、定语、状语等并列时，还需要注意排列不当的问题。

通常来说，时间状语和地点状语并列时，时间状语应位于地点状语的前面。例如，下面的例子中，时间状语“2010 年”应该放在地点状语“在上海”的前面。

错误句子:在上海 2010 年召开了第 41 届世界博览会。
正确句子:2010 年在上海召开了第 41 届世界博览会。

例 59: 状语排列不当

注意多个定语之间的排列顺序。修饰同一个中心语的不同定语，在排列上常常遵循一定的规则。领属定语、指示词定语、数量定语和带“的”的定语要放在最前面，其他定语是大致遵循“时间-大小-长短-颜色-形状-质料-功能”的顺序。

例如，下面的例子中，“鲜艳的”是带“的”的定语，应放在颜色词“红”前面。

错误句子:过去，女孩子结婚的时候常穿红鲜艳的衣服。
正确句子:过去，女孩子结婚的时候常穿鲜艳的红衣服。

例 60: 定语排列不当

例如，下面的例子中，“木”表示桌子的质料，“圆”表示桌子的形状，表示形状的成分应放在表示质料的成分之前，应为“圆木桌子”。

错误句子:客厅里放着一张木圆桌子。
正确句子:客厅里放着一张圆木桌子。

例 61: 定语排列不当

5.4.2 逻辑不通

逻辑不通是指句子符合语法规范，但不符合事理逻辑。“逻辑不通”包括逻辑顺序不当、前后矛盾和主客颠倒等问题

(1) 逻辑顺序不当：逻辑顺序不当是一种十分常见的病句类型。从内容上看，主要是不符合时间上由先到后，范围上由窄到广，程度上由低到高的顺序；从句式上看，表现形式相当多，并列和复句是最常见的出现环境。因此标注者在看到并列短语和复句时，就应该提醒自己注意观察出现的几个成分之间是否符合逻辑顺序。

例如，下面的例子中，按照动作的发生顺序，应该先总结，再提高。

错误句子:我们要注意多多提高总结自己。

正确句子:我们要注意多多总结提高自己。

例 62: 动词排列不当

例如，下面的例子中，按照动作的发生顺序，应该先来到池塘边，再跳下水。

错误句子:他跳下水，来到池塘边。

正确句子:他来到池塘边，跳下水。

例 63: 动词短语排列不当

(2) 主客体颠倒：主客体颠倒也是一类常见的逻辑错误。所谓“主客颠倒”是指一个句子中所陈述的主体和客体相互颠倒，是“语序不当”的一种。在遣词造句时，主宾关系往往通过介词“对、对于、和、与”等来反映，介词常引出对象。介词前是主体，介词后表示客体。即只能是人对物，或主观对客观，若颠倒过来就是背理的。

例如，下面的例子中，主体应该是“我”，而客体应该是“报纸”。

错误句子:在那个时候，报纸与我接触的机会是很少的。

正确句子:在那个时候，我与报纸接触的机会是很少的。

例 64: 主客体颠倒

例如，下面的例子中，我们比较一前一后两件事情，通常以后者为主体，前者为客体。

错误句子:去年的学习情绪和今年比较起来大不相同。

正确句子:今年的学习情绪和去年比较起来大不相同。

例 65: 主客体颠倒

(3) 前后矛盾：前后矛盾指的是句子前后叙述的内容不符，也包括一些因果逻辑错乱的问题。

例如，下面的例子中，“断了翅膀的小鸟”出现的直接结果是不会飞，而不是“失去了方向”。

错误句子:没有梦想、计划的人就好像是断了翅膀的小鸟，失去了方向。

正确句子:没有梦想、计划的人就好像是迷途的小鸟，失去了方向。

例 66: 前后矛盾

例如，下面的例子中，“正在”表示事情正在发生，“了”表示事情已经完成，相互矛盾，应删去一个。

错误句子:北京市正在采用了立法的方式保护老祖宗留下来的历史文化遗产。

正确句子 1:北京市采用了立法的方式保护老祖宗留下来的历史文化遗产。

正确句子 2:北京市正在采用立法的方式保护老祖宗留下来的历史文化遗产。

例 67: 前后矛盾

5.4.3 句式杂糅

句式杂糅主要包括两种类型。

第一种类型是格式糅合，指的是将两个意思相同或相近的格式放在一个句子里使用。人们在写句子时，本来使用了一种格式，但写的过程中由于句子内容等其他因素的干扰，换用了另一种格式，造成两种格式杂糅。

例如，下面的例子中，句式“原因是……”和句式“是……的结果”杂糅，应选用其中一个。

错误句子:形成沼泽的原因是水体沼泽化的结果。

正确句子 1:形成沼泽是水体沼泽化的结果。

正确句子 2:形成沼泽的原因是水体沼泽化。

例 68: 句式杂糅

例如，下面的例子中，“为了……”和“以……为目的”格式杂糅。应删除一个。

错误句子:昆虫鸣叫是为了吸引异性同类，或对其他动物进行警告为目的。

正确句子 1:昆虫鸣叫是为了吸引异性同类，或对其他动物进行警告。

正确句子 2:昆虫鸣叫是以吸引异性同类，或对其他动物进行警告为目的。

例 69: 句式杂糅

针对格式杂糅问题，标注者应该注意两点。首先，标注时看到句子里出现常用格式，要格外留心，注意观察是否出现两个格式的叠用，尤其是一些前后呼应的格式，如“从……出发”或“以……为基础”等，要检查前后用词是否成套，此处最常出现词语偷换现象。其次，平时应注意常用格式的积累，尤其是一些意义相同的、可能造成句式杂糅的格式，可以先做总结。此类句式杂糅的类型可以参考https://wenku.baidu.com/view/dcbb9804fbd6195f312b3169a45177232f60e437.html。

第二种类型是句子糅合，指的是将表意不同的两个句子糅合在一个句子里，此时常常是将前句的结尾用做后句的开头，造成句子结构混乱。

例如，下面的例子中，“大家要遵守交通规则”和“遵守交通规则是每一个市民的责任”这两个句子杂糅。

错误句子:大家要遵守交通规则是每一个市民的责任。

正确句子:大家要遵守交通规则，这是每一个市民的责任。

例 70: 句式杂糅

5.5 其它特殊错误

还有一些特殊类型的错误，在此进行记录。(有一些错误暂时不知道该如何分类。)

5.5.1 照应错误

照应错误：下面的例子中，生命的“短”可以照应“悲哀”，而“长”则不能照应。

错误句子:蜗牛感到自己很悲哀，不只是因为它生命的长短。

正确句子:蜗牛感到自己很悲哀，不只是因为它生命很短。

例 71: 照应错误

5.5.2 歧义错误

歧义问题：下面的例子中，一方面可理解为导师建议他对论文的观点加以修改，但是他没有听从；另一方面可理解为导师认为论文不需要修改。本句改为“他没有根据导师的建议对论文的观点加以修改，影响了论文的水平”。

错误句子:他没有听从导师的建议，对论文的观点加以修改，影响了论文的水平。

正确句子:他没有根据导师的建议对论文的观点加以修改，影响了论文的水平。

例 72: 歧义问题

5.5.3 语气不协调

语气不协调：下面的例子中，“慨叹”和后面的疑问句同时出现，导致语气的不协调。

错误句子:生活在这个万物滋长的社会中，我不禁慨叹社会的安详宁静仅仅是因为制度的限制吗?

正确句子:生活在这个万物滋长的社会中，我不禁疑惑社会的安详宁静仅仅是因为制度的限制吗?

例 73: 语气不协调

5.5.4 事实型错误

事实型错误主要指的是句子中出现的某些内容与事实不符，例如领导人的姓名、职位、排序等，遇到这一类错误，标注者既要联系自己的知识储备，也要学会查询外部搜索引擎、知识库等。

例如，下面的例子中，印度的首都是新德里而非孟买，写作者错把孟买当成了印度首都，属于犯了事实型错误。

错误句子:去年暑假，我曾经在印度的首都孟买逗留。

正确句子:去年暑假，我曾经在印度的首都新德里逗留。

例 74: 事实型错误

6 标注错误分析

目前，中文领域已经有了一些针对文本错误的人工标注工作，例如 NLPCC2018 的测试数据集、HSK 数据集等。但是，由于没有统一的标注规范指导，上述数据集的标注存在着不少问题。在这里，我们以 NLPCC2018 的测试数据集中出现的几类典型标注问题为例，进行标注错误分析。

6.1 过度润色

过度润色指的是：原句子没有明显的错误，但标注者为了让它看起来更加通顺优美，而对其进行了过度的修改。尽管标注者的出发点是好的，但在目前阶段，我们的工作暂时不涉及润色，仅涉及纠错，所以我们标注工作的一条重要原则是：仅需要保证句子中没有错误即可。

例如，下面的例子中，标注者为了让句子更通顺，将“本来”修改为了“先前”，将“改成了”修改为“变更为”，并在“农历十月初三”后添加了“这一天”。尽管上述修改

的本意是好的，但原先的表达并没有明显的错误，因此并不需要进行修改。所以，对于该句的标注，我们仅需要保留添加“于”的操作即可。

错误句子:“开天节”本来在农历十月初三，但 1949 年改成了阳历 10 月 3 日。”

原有标注:“开天节”先前在农历十月初三这一天，但于1949 年变更为阳历 10 月 3 日。”

正确标注:“开天节”本来在农历十月初三，但于1949 年改成了阳历 10 月 3 日。”

例 75: 过度润色

同理，下面的例子中，原有标注将“吵嚷”修改为了“吵闹”，也违背了仅保证无错即可的原则。因此，该句不需要做任何改动。

错误句子:请勿在楼道里吵嚷。

原有标注:请勿在楼道里吵闹。

正确标注:请勿在楼道里吵嚷。

例 76: 过度润色

6.2 违背句子原意

标注者对于错误句子进行修改时，必须牢记：纠正错误时不能违背句子的原意。如果写作者的原意被纠错系统所更改，那么即使纠错系统的纠错能力再好，也会让写作者产生很差的使用体验，尤其是对于新闻等严谨的行业。所以，为了杜绝此现象，我们从标注数据集时就应该时刻注意。

例如，下面的例子中，写作者由于输入法等原因错把“的”写成了“大”，导致标注者产生了理解偏差，认为写作者的原意就是“大韩国”，因此修改的结果违背了写作者的本意。

错误句子:我是 2010 年上北京大学中文系大韩国留学生叫洪世美。

原有标注:我是 2010 北京大学中文系的来自大韩国的留学生，我叫洪世美。

正确标注:我是 2010 年上北京大学中文系的韩国留学生，我叫洪世美。

例 77: 违背句子原意

6.3 缺失成分

由于数据集的原始语料通常是从外国人撰写的作文中挑选，因此，某些句子可能会省略主语、宾语等，对于此类句子，标注者不应过分联想而强行增加具体内容。

例如，下面的例子中，原有标注加上了“体温”，但是在这里我们并不知道写作者的原意是什么，有可能是气温达到了 40 度，也有可能是体温达到了 40 度，还可能是水温达到了 40 度，所以我们不应过度联想。对于此类未知词语的省略问题，标注者应当在词语丢失的位置添加特定的占位符。

错误句子:从昨日已达到四十度。

原有标注:从昨日体温已达到四十度。

正确标注:从昨日起 [缺失成分]已达到四十度。

例 78: 未知词语省略

下面的例子中，原有标注默认出故障的主语是“电脑”，但实际上有很多东西都会故障，所以这里也不应该添加词语，而是应该用占位符标志 [实体] 代替。

错误句子:昨日晚上就故障了。

原有标注:昨日晚上电脑就出故障了。

正确标注:昨日晚上[缺失成分]就出故障了。

例 79: 未知词语省略

6.4 纠正不完全

在标注者完成了标注之后，需要对纠正结果进行仔细的检查，以确保纠正了原句子中 100% 的错误。

例如，下面这个例子中，“短片小说”应该改为“短篇小说”，但原有标注疏忽了。

错误句子:《倾城之恋》是张爱玲的短片小说中之一。

原有标注:《倾城之恋》是张爱玲的短片小说。

正确标注:《倾城之恋》是张爱玲的短篇小说。

例 80: 纠正不完全

6.5 纠正有误

纠正有误指的是：原句子中的某处地方本来是正确的，但是标注者对其进行了错误的修改。这种情况在实际标注中比较少，但仍然需要注意。标注者在每次完成标注后，一定要重新通读几遍标注结果，以确保不存在纠正有误或者纠正不完全的现象。

6.6 纠正无法理解的句子

在标注时，如果错误句子难以理解，则不应该强行纠正。

错误句子:只性格和体力这调查有点瓜葛。

原有标注:只有性格和体力这个调查有点关系。

例 81: 纠正无法理解的句子

7 修订记录

7.1 修订：2021.4.15

- 删除规范中所提到的“最小编辑距离原则”和“纠正从简原则”，希望能够尽可能保证标注结果的多样性。
- 对命名实体拼写错误（5.2.4 节）的例句进行更新，同时给定了何时纠正命名实体类型错误的准则。
- 标注未知词语省略问题时，占位符暂定为 [Noun] 或者 [Verb]，分别表示名词丢失和动词丢失。
- 更新了 4.3 节语用层面错误的例句，使其更加合理。

References

[1] Wang Y, Wang Y, Liu J, et al. A comprehensive survey of grammar error correction[J]. arXiv preprint arXiv:2005.06600, 2020.

[2] Naghshnejad M, Joshi T, Nair V N. Recent Trends in the Use of Deep Learning Models for Grammar Error Handling[J]. arXiv preprint arXiv:2009.02358, 2020.

[3] Zhao Y, Jiang N, Sun W, et al. Overview of the nlpcc 2018 shared task: Grammatical error correction[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2018: 439-445.

[4] Lee L H , Rao G , Yu L C , et al. Overview of the NLP-TEA 2016 Shared Task for Chinese Grammatical Error Diagnosis[C]// Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA'16). 2016.

[5] Rosen A, Hana J, Štindlová B, et al. Evaluating and automating the annotation of a learner corpus[J]. Language Resources and Evaluation, 2014, 48(1): 65-92.

[6] Zaghouani W, Mohit B, Habash N, et al. Large scale arabic error annotation: Guidelines and framework[J]. 2014.

[7] 李珠花. 新汉语水平考试六级阅读第一部分病句试题统计与分析. 硕士学位论文. 华中科技大学, 2012.

[8] 王芬. 现代汉语语病的三个平面分析. 硕士学位论文. 南京: 南京大学, 2005.