arXiv:2301.00374v1 [cs.CL] 1 Jan 2023

使用遗传算法优化可读性

Jorge Martinez-Gil

Software Competence Center Hagenberg GmbH

Softwarepark 32a, 4232 Hagenberg, Austria

jorge.martinez-gil@scch.at

摘要

本研究介绍了 ORUGA，这是一种旨在自动优化任何英语文本可读性的方法。该方法的核心思想是，某些因素会影响文本的可读性，其中一些因素可以量化（词数、音节数、副词的有无等）。这些因素的特性使我们能够实施遗传学习策略，用最合适的同义词替换部分现有单词，从而便于优化。此外，本研究还通过多目标优化技术，力求保留原文的内容和形式。这样既不会显著改变文本的句法结构，也不会明显扭曲原始信息的语义内容。我们对大量且类型多样的文本进行了全面研究，结果证实，该方法能够在所有情况下优化可读性，同时不会显著改变文本的形式或含义。此方法的源代码可在 https://github.com/jorge-martinez-gil/oruga 获取。

关键词：文本可读性、文本优化、遗传算法

1. 引言

可读性是一项用于说明文本易于阅读程度的衡量指标。它对应于目标受众读者所应具备的读写能力水平。因此，在用户获取信息时，可读性被视为促进用户体验的最关键因素之一。这一点至关重要，因为它是信息生产者与消费者之间建立信任关系的关键。需要注意的是，复杂度、易读性或排版等因素有助于提升文本的可读性。然而，并非所有因素都可以量化，也无法通过自动化技术进行优化。在本文中，我们仅关注具有可量化性质的因素，这些因素始终围绕与待优化文本相关的基础或高级统计数据展开。

因此，文本可读性是指根据文本自身的特点，阅读和理解一段给定文本有多容易。这些特点通常可以通过一些指标来衡量，例如一个句子中的音节数量。用于生成可读性

评分的单词多样性可用于衡量这一指标（Collins-Thompson，2014）。因此，ORUGA 旨在通过优化文本的可读性，帮助读者理解文本。

在本研究的语境中，应明确区分可读性与质量。质量强调语法、拼写和文体等基本要素。然而，所谓文本可读，是指尽可能清晰地呈现文本信息，并使其更好地契合受众需求。因此，文本可读性与文本难度的概念部分重叠；文本难度也是人类语言的一个重要方面，并影响着大多数阅读书面信息者的日常生活。

我们的研究基于这样一个前提：可以通过考量与文本相关的指标，再使用特定的数学公式计算可读性。由于不同的可读性分数采用不同的数学公式计算，因此可以设计一种全局优化策略，将文本中的许多词语替换为同义词。这种优化旨在根据用户的需求，最大化或最小化此类数学公式得出的结果。在实践中，可以通过遗传算法实现这一策略，本文将对此进行阐述。

因此，我们的研究从计算机科学领域提供了一种视角，重点关注可读性评估方法所使用的基础文本表示和度量指标。由此，本研究对研究现状最重要的贡献如下：

- 我们首次提出了一种能够自动优化任意文本可读性的技术，即无需对文本进行实质性改动，便可自动提高或降低文本的可读性程度。

- 我们研究了目前可用于优化文本可读性的最佳同义词来源。具体而言，我们研究了 WordNet、word2vec 和网络爬取，并建立了一种分类方法，用于优化最多十种不同类型的文本。

- 我们提出了另一种基于多目标优化的方法，其目标是确保只替换原文中所需的最少数量的单词，从而不影响文本结构。

- 最后，我们研究了多种策略，用于衡量（进而优化）原文与生成文本之间的语义距离。通过这种方式，可以控制对原文含义的影响。

本研究的结构如下：第2节介绍了提高文本可读性的最先进的方法和工具。第3节阐述了我们方案的技术细节；这些技术细节基于一种遗传切分策略的设计，该策略使我们能够在消耗合理资源的同时探索广阔的搜索空间。第4节说明了如何利用多目标优化来最大限度地减少对原文形式的影响

技术。第5节介绍了如何确保原文与所得文本之间的距离尽可能小，从而保留原文的精髓。需要指出的是，第3、4、5节介绍了不同实验的设计、获得的原始结果及其后续分析。最后，我们总结了本研究带来的主要启示。

2. 研究现状

我们先来看一个可读性的正式定义。例如，Chall 和 Dale（1995年）将可读性定义为“影响读者成功理解给定文本的所有要素的总和”。读者的成功程度，是衡量读者以最佳速度阅读文本时理解得有多好的指标。与此同时，Mc Laughlin（1969年）将可读性定义为“某些人认为阅读材料具有说服力且易于理解的程度”。除了这些定义之外，我们尤其关注可读性中可量化的方面，也就是能够客观测量的内容。否则，就难以借助计算机对其进行优化。让我们看看文献对这些可量化方面有何论述。

2.1. 科学文献中的文本可读性

可读性指标通常采用简单特征来计算文本的可读性程度。一些常用特征包括句子数或词数、不同词的比例、音节总数、文本中不同词所占的比例、数字数量、含有较多音节的词数等。乍看之下，这些度量指标似乎过于简单，但它们之所以广泛应用，主要有两个重要原因：与人工调查等替代方法相比，它们的使用成本更低、速度更快；而且据专家称，它们通常能给出极为可靠且符合实际的结果。

提高可读性的目标，是增加读者理解文本所表达的思想和观点的可能性，从而尽量减少误解，并促进信息处理，同时无需耗费太多精力和能量。基于这一目标，我们可以找到许多关于如何提高文本可读性的建议。然而，这些都是人工编写的规范，需要人工操作员通过手动修改文本将其付诸实践。例如，对于某项指标，最好缩短句子；对于另一项指标，最好用简单的单词替换复杂的单词，等等。

这正是我们对研究现状的贡献所在。以某项指标为目标自动优化文本，意味着我们无需手动采取任何会改变文本的操作。遗传算法会自动找到处理方法。

2.2. 为什么文本可读性很重要?

在许多情境和人群中，可读性至关重要，尤其是在需要向受众传达书面信息时。例如，

• 教师需要先确认文本的可读性，再决定它是否适合学生。这在语言学习中尤为重要。通过本文介绍的方法，可以针对特定的学习者群体优化文本。

• 在广告领域，可读性有助于广告商与潜在消费者建立信任关系。使用可读性高的文本来推广商品或服务通常并非好主意，因为信息可能无法传达给种群中的重要部分。在搜索引擎优化领域，这一点更为重要，因为许多搜索引擎会在响应用户搜索时将可读性指标用作排名因素。

• 对于从事网站（Pantula & Kuppusamy, 2022）、新闻（Qin, 2021）乃至教育材料（Ante, 2022）工作的专业人士来说，可读性也很重要。在一些国家，法律甚至要求政府机构提供达到特定可读性水平的文本信息，以便覆盖整个种群。

2.3. 可读性指标

有多种指标可以量化文本的可读性（Meade & Smith, 1991）。大多数指标都是为英语设计的（Maqsood et al., 2022），不过也有研究探讨其他语言的可读性（Madrazo Azpiazu & Pera, 2020）。不一一详述，我们可以按时间顺序列举一些在处理英语时较为流行或曾经流行的指标。

文本可读性不仅取决于文本的特征，也取决于有意理解该文本的个体的教育背景。下面在使用公式测量可读性时，我们会看到这一点。我们将看到，可读性指标以文本为输入，并计算出一个数值分数，该分数通常对应于理解文本所需的教育程度。

2.3.1. Dale-Chall可读性

Dale-Chall可读性公式（Dale 和 Chall，1948年）需要一份包含 3,000 个单词的词表，这些单词是美国四年级学生能够可靠理解的，如公式 1 所示。

$$DCRF = 0.1579\left(\frac{difficult\ words}{total\ words} \times 100\right) + 0.0496\left(\frac{total\ words}{total\ sentences}\right) \tag{1}$$

2.3.2. SMOG可读性

SMOG可读性等级（Mc Laughlin，1969年）可以通过一个最初用于检查健康信息的公式来评估，如公式2所示。它对应于理解文本所需的受教育年限。

$$SMOG = 1.0430\sqrt{number\ of\ polysyllables \times \frac{30}{number\ of\ sentences}} + 3.1291 \tag{2}$$

2.3.3. ARI可读性

ARI评估阅读文本所需的美国年级水平（Senter & Smith，1967）。在某些方面，它与其他公式相似。它的不同之处在于，它不统计音节，而是统计字符：字符越多，单词就越复杂。它还会统计句子，如公式3所示。这使它有别于其他一些公式。

$$ARI = 4.71\left(\frac{total\ Characters}{total\ Words}\right) + 0.5\left(\frac{total\ Words}{total\ Sentences}\right) - 21.43 \tag{3}$$

2.3.4. Flesch Kincaid可读性

如公式4所示，Flesch Kincaid可读性评分是一种基于年级水平的指标，在保险业中十分常用。年级水平让人们更容易理解文本（Kincaid等人，1975）。Flesch Kincaid年级水平（FKGL）介于8至10之间，意味着公众应能轻松读懂该文本。FKGL至今仍是应用最广泛的公式。

$$FKGL = 206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right) \tag{4}$$

过去十年间，人们提出了多种自然语言处理（NLP）技术，用于确定文本的可读性。因此，与经典方法使用公式测量有限文本特征的做法不同，这些新方法尝试衡量句子和单词的理解难度，甚至句法的复杂度（Martinc等人，2021）。一些基于可读性预测因素（如衔接和连贯性）的技术也受到了广泛关注。然而，迄今为止，这些方法在预测文本可读性方面仍未能胜过本文讨论的经典技术（Todirascu等人，2016）。

2.4. 语义相似度

语义相似度测量领域（Martinez-Gil，2022）是多个不同研究社群（信息检索、数据库集成、自然语言处理等）中最活跃的领域之一（Rus等人，2013）。这是因为它对工业界和学术界所采用的诸多框架、方法和工具都有重大影响（Navigli & Martelli，2019）。近几年，围绕这一主题的文献数量激增（Chandrasekaran & Mago，2021）。然而，大多数研究都侧重于确定单词（Zhu & Iglesias，2017）、短语或文档之间的相似度（Martinez-Gil & Chaves-Gonzalez，2019；Martinez-Gil & Chaves-Gonzalez，2020）。

因此，能够有效剔除与给定单词最相似单词的可能性很少得到讨论。为此，人们手动编纂了同义词库，也有一些词嵌入技术能够很好地完成这项工作。最具代表性的

解包括 WordNet（Pedersen 等人，2004）、word2vec（Mikolov 等，2013年）和 BERT（Devlin 等人，2019）。不过，必须考虑到，基于词嵌入计算的技术对资源的需求非常高（Martinez-Gil & Chaves-Gonzalez，2022年）。

2.5. 相较于研究现状的贡献

近几年来，依据对文本结构和所用单词的形式化分析来确定文本可读性，一直是文献中反复探讨的主题。因此，人们提出了许多度量指标来衡量文本可读性。这些度量指标的共同点是，通常分数越高，文本就越难理解。换言之，理解文本所需的教育程度越高。因此，许多传播专业人士经常借助工具来判断文本是否适合特定受众。然而，这些工具都无法自动完成专业人士的工作，因此我们的贡献至关重要。

据我们所知，这是首次有人提出一种方法，能够在不显著改变文本内容或形式的前提下，自动提升文本的可读性。此外，我们使用主题和可读性水平各不相同的大量文本，对这一创新方法进行了全面测试。我们还提供了首个实现的源代码，供有兴趣试用或改进该方法的人使用。

3. 第一部分：功能性解决方案的设计与实现

接下来，我们将说明如何运用基于遗传算法的方法来优化文本可读性。本节将概述技术预备知识，讨论实现，展示一个说明性示例，介绍我们的方法在实践中的运作方式，并利用真实数据和用例开展实验研究，帮助我们了解该方法在实践中的表现。

3.1. 技术预备知识

我们的假设是，可以找到一组同义词来替换原文中的某些单词，从而优化可读性公式返回的值。例如，如果可读性公式会因单词较长而给予奖励或惩罚，我们就必须找到长度较短的同义词。实际上，我们只需弄清楚哪种公式最能代表特定场景中的可读性，并将其指定为适应度函数。遗传算法会决定如何进行。因此，我们面对的是一个经典的优化问题。

我们打算只处理词汇，因为它是该语言中最关键的部分之一。人们普遍认为，词汇是语言的核心部分，因为没有词汇，就不可能传达任何信息（Wilkins，1972）。为了避免歪曲原始信息，我们不处理专有名词、介词或其他停用词。

虽然可以通过对给定文本中的单词范围进行暴力搜索来解决这个问题 $w_0, w_1, ..., w_n$，但在处理大型文本时，遗传算法方法的扩展性非常好。在这种情况下，暴力搜索的成本会高得令人难以承受。我们也可以处理其他方面，例如结构，但这样做可能会再次歪曲原文的本质。

我们的思路是找出能够优化所需目标的单词组合（该组合将编码为一个个体）。适应度函数的选择很简单，其优势在于，解会自动揭示哪些类型的单词能够实现优化。

在这项工作中，我们使用三个主要的同义词来源：

• WordNet（Miller，1995年）是一部人工编纂的同义词词典。它或许是信息系统领域使用最广泛的词典之一，能够为每个候选词提供多个替换选项。这个库不会带来什么意外，不过也可能将某个单词替换为含义与原词相去甚远的同义词。无论如何，我们有办法解决这个问题。

• wordvec（Mikolov等，2013年）是一种根据相关文本语料库计算每个单词对应向量的方法。计算出各个向量后，可以通过计算过程找出与给定向量最相似的 N 向量。这样便能为相关单词获取合适的同义词。请注意，这一过程计算开销较大。

• 网络爬取（Mitchell，2018年）是指从一些通常较为专业的网站获取同义词，以便为每个候选词筛选出若干备选词。该方法可以获得许多高质量的候选词，但应负责任地使用，因为若发送数百乃至数千次请求，可能会给服务器端带来问题。因此，该方法适用于实验，但不应滥用。

3.2. 实现

这一新方法的实现基于一种采用遗传算法的经典优化方案。算法 1通过伪代码简要说明了整个过程：在经典遗传算法结构（Forrest，1996年）的基础上进行调整，采用能够实现选择、交叉和变异过程的不同算子，使给定种群逐步演化，最终达到预期目标。

请注意，在开始进化过程之前，必须对文本进行预处理，以识别出需要用同义词替换的候选词。我们建议，除专有名称、停用词和介词外，所有单词都作为候选词。之所以这样做，是因为这些词通常很难找到同义词。

算法 1 使用遗传算法优化可读性

```
过程 ORUGA
    种群 ←生成随机个体（种群）
    计算可读性评分（种群）
    当未达到停止条件时执行
        父代 ←个体选择（种群）
        子代 ←交叉（父代）
        子代 ←变异（子代）
        子代 ←计算可读性评分（子代）
        种群 ←更新种群（子代）
    结束循环
    优化后的文本 ←最优个体（种群）
    优化后的文本 ←必要时纠正错误 (优化后的文本)
    返回优化后的文本
```

关于遗传算法本身，解的参数细节将在下文讨论，但我们可以先了解如何以经典进化过程的形式实现它。这意味着随机选择一个个体种群，并计算其可读性评分。随后，我们进行选择、交叉和变异的迭代过程；最优个体会从一代传递到下一代，直到满足某个停止准则：达到可能的最高值（不太可能），或迭代次数耗尽（非常可能）。

最后，在进化过程结束时，我们会检查文本并纠正可能出现的语法错误，例如替换不符合其所在句子时态和形式的同义词。这样，我们就能得到修正后的可读性评分，该评分可能与自动计算得出的分数略有不同。作为回报，我们可以确保结果可用，或至少接近可用。

3.3. 示例说明

示例 1 展示了 ORUGA 如何处理从维基百科摘取的科学类文本。事实上，我们旨在观察 ORUGA 使用库 WordNet 中的同义词来尽量降低 FKGL可读性分数时的表现。请记住，FKGL（或其某种变体）可能是应用最广泛的指标，而对其进行优化能为多个领域带来优势。请注意，待处理文本的长度可以任意，但为了便于向读者展示，我们选择了（并且在本文中将始终选择）仅包含几个句子的文本。

EXAMPLE 1. SCIENCE

原文来源：维基百科

“海洋调节气候，并在水循环、碳循环和氮循环中发挥重要作用。自古以来，人类利用和研究海洋的相关记载便有据可查，其历史甚至可追溯至史前时期；而对海洋的现代科学研究则称为海洋学。海水中溶解量最多的固体是氯化钠。海水还含有镁、钙、钾和汞的盐类，以及许多其他元素，其中一些元素的浓度极低。盐度差异很大：海洋表层和大河河口附近的盐度较低，海洋深处的盐度较高；不过，溶解盐类的相对比例在各大洋之间变化不大。” FKGL 分数：14.06。

ORUGA——最小化 FKGL 分数——WordNet 库

海洋调节气候，并在 .....中发挥重要 作用，参与 H2O 循环、 C 循环和 N 循环。人类利用和研究海洋的历史可以追溯到 远古时代， 并且 早在 史前时期就已有相关记载，而对海洋的现代科学研究称为海洋学。盐水中含量最丰富的固体 物质 是 氯化钠， 即 Na 氯化物。海水 还 含有镁、钙、钾和汞等多种元素的盐，以及许多其他元素，其中一些元素的含量仅为 极低 浓度。盐度 变化 很大：在近海表层和大型河流入海口附近 较低， 而在海洋 深处 则较高；不过，各种 溶解 盐 的相对比例在各大洋之间 变化不大。FKGL 分数：11.85

如上所示，ORUGA 可以优化文本输入：首先自动识别哪些单词可以替换为同义词，然后通过搜索能够改善待优化指标结果的同义词来完成处理。这样一来，对初始消息的影响很小，不过确实有些同义词可能会使文本含义产生轻微偏差，因此需要用户进行最终审核。不过，无需担心，因为本文稍后将讨论这一问题的解决方法。

现在来看示例2。这是一段关于奥地利历史的书面文本，也摘自维基百科。我们希望利用通过网络爬取自动获取的同义词，尽可能降低 FKGL 可读性评分。

EXAMPLE 2. HISTORY

原文来源：维基百科

“奥地利在第一个千年末由东部边区和匈牙利边区的残余部分发展而来。它最初是巴伐利亚的一个边境侯国，1156年成为神圣罗马帝国的公国，1453年则升格为大公国。16世纪，维也纳开始成为帝国的行政首都，奥地利也因此成为哈布斯堡君主国的核心地区。1806年神圣罗马帝国解体后，奥地利建立了自己的帝国，成为一个强国以及德意志邦联的主导成员。1866年奥普战争中的战败导致邦联终结，并为次年奥匈帝国的建立铺平了道路。” FKGL 分数：13.20

ORUGA——最小化 FKGL 分数——网络爬取

奥地利 在第一个千年末从东部边区和匈牙利边区的 废墟中 崛起 。它最初是巴伐利亚的一个 边疆伯国，后来发展成为 一个 公国，并于1156年成为神圣罗马帝国的一部分；随后 又 于1453年升格为 大公国。16世纪，维也纳 成为 指挥部所在地 ， 因此奥地利 成为 哈布斯堡君主国的核心地区。1806年神圣罗马帝国 解体后，奥地利 建立 了自己的帝国，并 成为强国及德意志邦联的主导 成员。1866年奥普战争中的战败导致邦联终结，并 为 次年 建立 奥匈帝国铺平了道路。 FKGL 分数：11.35

我们再次可以看到，遗传算法通过降低文本可读性评分，巧妙地让文本变得更容易被更多人理解。显然，一些单词是否合适可能存在争议，但本研究的首要目标，即优化可读性评分，已经实现。正如前文所述，我们将在本文后续部分概述最终成果。

最后，让我们看看相反的情况，也就是看看能否在不失文本本质的前提下，让文本变得更难读。在示例3中，我们有一段同样摘自维基百科、关于体育的文本。我们不想让尽可能多的人都能轻松读懂它，反而希望提高可读性难度。现在，我们将尝试使用另一种指标，例如 ARI。

EXAMPLE 3. SPORTS

原文来源：维基百科

“皇家马德里足球俱乐部（意为皇家马德里足球俱乐部），通常简称皇家马德里，是一家位于马德里的西班牙职业足球俱乐部。俱乐部于1902年以马德里足球俱乐部之名成立，自创立以来，主场球衣一直以白色为传统颜色。real 是西班牙语中的“皇家”之意；1920年，阿方索十三世国王授予俱乐部这一尊称，并将王冠加入其徽章。自1947年起，皇家马德里一直在位于马德里市中心的圣地亚哥·伯纳乌球场进行主场比赛。与大多数欧洲体育组织不同，皇家马德里的会员（会员）在俱乐部的整个历史中一直拥有并经营着俱乐部。” ARI分数：12.69。

ORUGA - 最大化 ARI 分数 - 网络爬取

皇家马德里足球俱乐部，意为皇家马德里足球俱乐部，通常简称皇家马德里，是一家西班牙职业足球企业，成立于马德里。俱乐部于1902年以马德里足球俱乐部之名创立，该企业一直始终身着一套白色主场球衣。俱乐部自创立以来便一直如此。其称号中的“Real”在西班牙语中意为“皇家”，并于由阿方索十三世国王授予该企业于1920年获得，同时还在队徽上加上了王冠。皇家马德里一直在熟悉的马德里市中心的圣地亚哥·伯纳乌 球场进行比赛，球场于1947年落成。与大多数欧洲体育机构不同，皇家马德里的会员（会员）拥有俱乐部，并管理着俱乐部至今。ARI分数：15.53

可以看出，对从维基百科提取的皇家马德里相关片段进行处理后，遗传算法会选择更长、更难读的同义词，以提高 ARI分数。虽然很难想象有什么用例是为了降低文本对读者的可读性，但这种做法或许能在学习、教育等某些特定领域找到应用。

3.4. 实验研究

本节将详细介绍我们开展的一项实证研究，以了解新方法的可行性。为此，我们首先通过搭建实验设置，确定了实验条件。其次，我们进行了实验并获得了原始结果。随后，我们对所获得的数据进行了分析。

3.4.1. 实验设置

调整遗传算法的参数并非本研究的重点。因此，我们选择了经典的参数设置，这种设置已被证明效果良好。标准的参数调优（例如通过网格搜索）是未来可探索的方向之一。目前，我们在实验中使用的参数如下：

- 种群规模 $\{10, 15, 20\}$: 20
- 交配亲本数量 $\{10, 15, 20\}$: 10
- 基因数量：每个要替换为同义词的候选词对应一个基因
- 适应度函数：用户可从公式 1、2、3、4 中选择
- 停止条件： $\{100, 200, 300\}$：300 代

为了检验我们的方法能否优化文本可读性，我们从维基百科中半随机地选取了十篇文本。这些文本分为工程学、地理、历史、科学和体育等类别。有些度量指标（例如 SMOG）需要至少 30个句子才能应用其公式。对于句子数量不足 30个的情况，我们会重复文本，直到达到该句数。还应注意，每个用例所代表的可读性类别各不相同。例如，洋红色代表难度最高（FKGL 介于 15 和 18 之间），相当于一篇学术论文。红色代表中高难度（FKGL 介于 12 和 15 之间）。黑色代表中等难度（FKGL 介于 9 和 12 之间）。此外，蓝色代表中低难度。据估计，FKGL 介于 6 和 9 之间的文本可被多达 80% 的以英语为母语的人口理解，这对应于蓝色所代表的数值。

此外，每项实验均在一台标准计算机上运行，该计算机配备 32 GB 主存和一颗运行频率为 3.20 GHz 的 Intel Core i7-8700 处理器，操作系统为 Microsoft Windows 10 64 位。大部分功能均使用库 PyGAD[1], 实现；PyGAD 是一个用于构建遗传算法的开源 Python 库。

3.4.2. 实验

第一个实验是在实践中最有用的实验。它旨在尽可能降低所提出的十篇文本的可读性，使处理后的文本能够覆盖尽可能广泛的受众。为此，我们想使用可读性指标，根据音节数和词数等简单方面来估算文本的可读性。我们采用 FKGL 分数，因为几十年来，这一指标或其某些变体一直用于传统文本，而且它仍是最常见、应用最广泛的传统可读性度量之一。

[1]https://pypi.org/project/pygad/

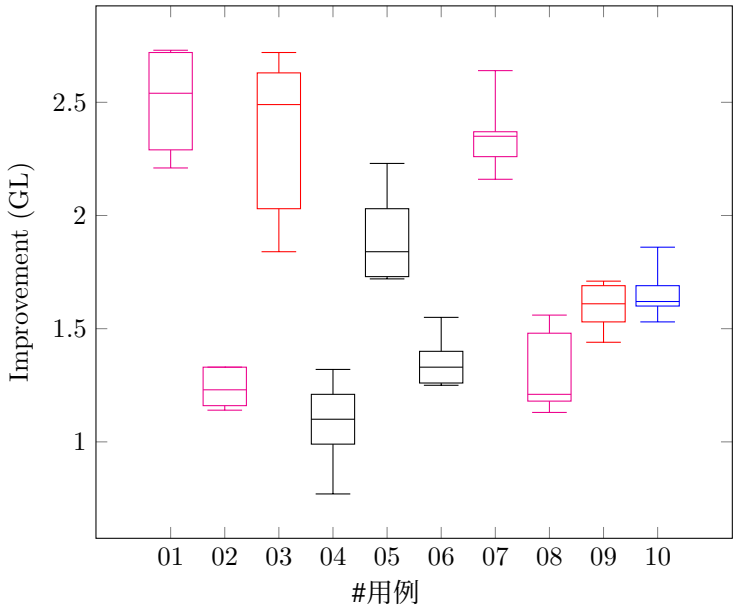


图 1：使用 WordNet 最小化 FKGL 分数的结果

图 1 展示了使用 WordNet 库降低相关文本 FKGL 分数后得到的结果。所示结果汇总了每个用例的十次独立运行。由于我们处理的是采用随机值的冷启动方法，而且变异过程中也包含随机成分，因此几乎每次都会得到不同的结果，所以必须以这种方式展示结果。

在进行的全部 100 次实验中（针对 10 个用例各运行 10 次）均取得了积极成果。此外，最小改进、平均改进甚至最大改进均处于 FKGL 量表的 0.77 至 2.73 分范围内。

图 2 展示了使用基于 word2vec 的同义词计算方法所获得的结果。可以看出，结果的方差非常大，这意味着使用该库会使结果难以预测。与此同时，与前一种情况一样，全部 100 次实验均实现了可读性提升，幅度介于 0.42 和 3.90 个年级水平之间。

图 3 展示了我们通过网页抓取技术搜索同义词所获得的结果。这次的结果方差小于前一种情况。此外，在进行的全部 100 次实验中，我们再次取得了有利结果。所实现的优化幅度介于 1.02 和 4.20 个年级水平之间。

表 1 展示了所进行的 300 次实验（每个同义词库各 100 次）的汇总结果。对于 WordNet，预期中位数优化幅度为 1.63 个年级水平，与.....非常相似

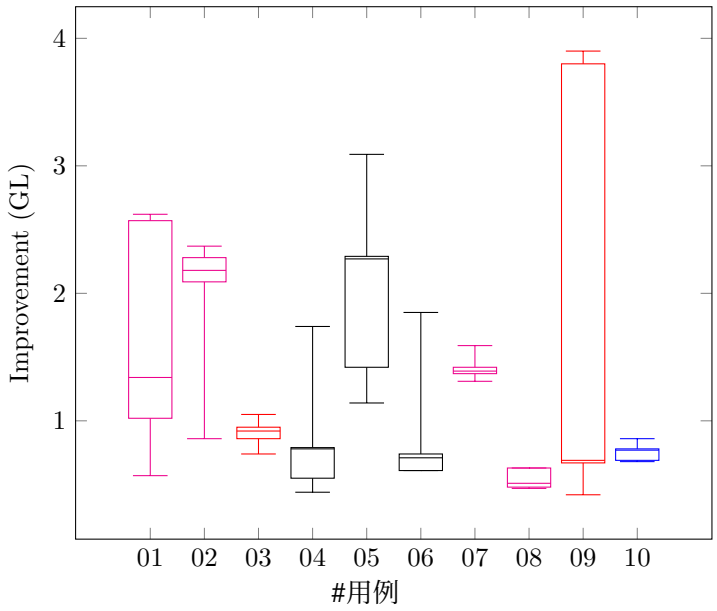


图2：使用 word2vec 最小化 FKGL 分数的结果

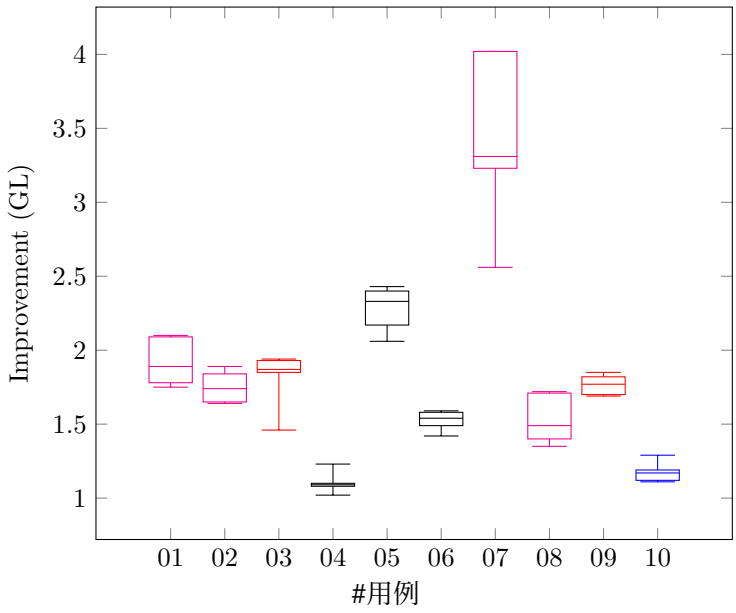


图3：使用网络爬取最小化 FKGL 分数的结果

库	最小值	中位数	最大值
WordNet	0.77	1.63	2.73
word2vec	0.42	0.96	3.90
网络爬取	1.02	1.61	4.20

表 1：使用不同同义词库所获得的结果汇总

用例	DCRF	SMOG	ARI	FKGL
01/科学	1.11 ± 0.17	2.03 ± 0.13	2.90 ± 0.27	2.53 ± 0.15
02/历史	0.29 ± 0.12	0.71 ± 0.08	0.84 ± 0.05	0.90 ± 0.04
03/地理	0.65 ± 0.07	0.91 ± 0.04	1.17 ± 0.12	1.33 ± 0.04
04/体育	0.78 ± 0.14	0.42 ± 0.12	0.83 ± 0.05	0.93 ± 0.10
05/工程学	0.59 ± 0.06	0.60 ± 0.10	0.66 ± 0.10	0.75 ± 0.09
06/地理	0.52 ± 0.15	0.70 ± 0.10	0.77 ± 0.19	0.77 ± 0.18
07/工程学	1.03 ± 0.15	0.51 ± 0.03	1.12 ± 0.21	1.32 ± 0.01
08/历史	0.77 ± 0.09	0.27 ± 0.06	0.44 ± 0.00	0.73 ± 0.17
09/体育	0.46 ± 0.06	1.06 ± 0.28	0.99 ± 0.12	0.81 ± 0.10
10/历史	0.74 ± 0.04	0.65 ± 0.03	0.77 ± 0.13	0.93 ± 0.04

表2：最小化实验结果汇总。数值表示使用 ORUGA 后取得的绝对改进。

与通过网络爬取获得的结果相比，使用 word2vec 进行优化的中位数在所考察的三个库中最差。请注意，我们在这里评判的不是替换内容的质量，而是优化输入文本时得到的数值，且不考虑最终结果的质量。

表2总结了我们获得的所有结果。这些文本是从维基百科中随机获取的。所示数值范围表示可达到的最小值和最大值（分别对应最小化和最大化问题）。

我们分析了认为最具代表性的四个公式，但分析其他公式也并非难事。使用 SMOG 时必须谨慎，因为待分析的文本至少需要包含30个句子。在这些实验中，我们将文本复制了一份，以便使用这30个句子。

表3总结了我们在最大化可读性分数时获得的所有结果。与上表对比可以看出，提高可读性水平要容易得多；换言之，让文本变得难以阅读，比反过来让文本变得易读更容易。

3.4.3. 讨论

我们已经了解了如何构建一种解，以优化不同性质文本的可读性。此外，这种优化是在尊重此类文本内容和形式的前提下进行的，并试图通过替换为更符合我们所研究的不同公式之可读性标准的同义词，将词语替换的影响降至最低。我们已经了解，优化可以通过两种方式实现。首先，可以通过降低可读性评分来实现，这

用例	DCRF	SMOG	ARI	FKGL
01/科学	1.98 ± 0.16	1.30 ± 0.06	1.80 ± 0.08	2.09 ± 0.14
02/历史	1.85 ± 0.25	1.99 ± 0.09	3.88 ± 0.28	3.02 ± 0.08
03/地理	2.38 ± 0.12	1.23 ± 0.08	4.44 ± 0.13	4.00 ± 0.44
04/体育	2.09 ± 0.40	2.79 ± 0.07	4.63 ± 0.40	4.26 ± 0.04
05/工程学	2.10 ± 0.43	2.16 ± 0.11	4.55 ± 0.10	4.23 ± 0.35
06/地理	1.53 ± 0.21	1.81 ± 0.22	2.96 ± 0.04	2.90 ± 0.41
07/工程学	1.49 ± 0.38	1.71 ± 0.04	3.87 ± 0.46	3.34 ± 0.16
08/历史	1.43 ± 0.18	3.09 ± 0.13	4.11 ± 0.63	3.70 ± 0.21
09/体育	1.60 ± 0.21	2.29 ± 0.16	3.54 ± 0.28	2.56 ± 0.04
10/历史	2.06 ± 0.08	2.64 ± 0.04	3.99 ± 0.25	3.27 ± 0.52

表3：最大化实验结果汇总。数值表示使用 ORUGA 后的绝对改进幅度

将使更多人能够完全理解文本。这无疑是实践中最实用的选项。此外，其次，也可以通过提高可读性评分来实现，使较少的人能够毫无歧义地理解处理后的文本。这个选项的实际效用并不像乍看之下那样明显。

根据我们开展的研究，还可以进行一些改进。例如，多目标优化算法可以在尽可能少地影响单词的同时，对可读性评分进行优化。这样，我们的方法对原文的影响就会更加微不足道。解前沿可以让人工操作员决定评分调整与替换单词之间的权衡。

最后，作为我们方法的一项局限，可以发现处理后的文本有时会包含轻微的语法错误。这是因为我们尚未采用一些技术，例如根据上下文选择恰当的动词时态。对此有两种选择。一方面，我们可以实施更好的语法错误修正方法，以获得修正后的可读性分数；另一方面，我们可以让用户选择编辑单词，或选择最适合当下语境的单词。

4. 第二部分：尽量减少对原文形式的影响

虽然在第3节中我们已经看到，可以设计出用于优化文本可读性的有效解，但我们也看到，这种方法有时会造成较大干扰。也就是说，将许多单词替换为同义词可能会导致原文信息发生偏差。因此，在本节中，我们将通过尽可能少地替换单词，重点关注如何最大限度地降低 ORUGA 对原文的影响。为此，我们将构建一个基于多目标优化（MOO）的解，使我们能够优化可读性，同时最大限度地减少替换次数。本节包括技术预备知识、实现、一些说明性示例，以及一项用于测试多种不同性质文本的实证研究。

4.1. 技术预备知识

我们已经看到，在文本可读性领域，人们投入了大量精力来构建简单易懂的公式，使人们能够理解这些公式，并且这些公式与从人类角度看文本的易读程度具有良好的相关性。现在，我们更进一步，以获得更高质量的结果。多目标优化是一种同时优化两个或更多目标的策略。鉴于我们希望实现以下目标，这正是我们所面临的情况：一方面，提高文本的可读性；另一方面，尽可能减少为改善可读性而需要替换的单词数，从而最大限度地降低我们的方法对原文形式造成的影响。

此外，当决策必须在多个相互独立的目标之间权衡时，多目标优化很有用。我们的情况也是如此，因为我们既要最大限度地提高可读性，又要尽可能少地替换单词，这两个目标截然不同。在这种情况下，不同的解可以同时满足所有目标。因此，所有最优解都应被视为价值相等，无需人工操作员进行任何外部评估（Martinez-Gil & Chaves-Gonzalez，2021）。更正式地说，我们可以按照公式 5 所示对多目标优化问题进行建模。

$$\begin{gathered}\min\left(s_1(\vec{x}), s_2(\vec{x}), \ldots, s_n(\vec{x})\right) \\ subject\ to\ \vec{x} \in X\end{gathered} \tag{5}$$

在多目标优化中，没有任何一个解能够同时处理所有目标函数。因此，应优先寻找这样的解：若要改善其中任何一个目标，至少就会使其他目标之一变差。因此，如果满足公式 6 中所述的条件，则称一个解 $\vec{x}_1 \in X$ 支配另一个解 $\vec{x}_2 \in X$。

$$\begin{gathered}s_i(\vec{x}_1) \leq s_i(\vec{x}_2)\ \forall i \in \{1, 2, 3, \ldots, n\} \\ s_j(\vec{x}_1) < s_j(\vec{x}_2)\ \exists j \in \{1, 2, 3, \ldots, n\}\end{gathered} \tag{6}$$

因此，如果不存在任何解能够支配某个解 $\vec{x} \in X$，则该解是最优的。如果元素 x 在所有目标上都不劣于元素 y，并且至少在一个目标上严格优于元素 y，则称元素 x 支配另一个元素 y。搜索空间中未被支配的元素构成帕累托前沿，它代表了正交目标的最佳可能解。

4.2. 实现

针对多目标优化策略，有多种实现方式，例如 Kukkonen & Lampinen (2005年)；Zhang & Li (2007)。本文不打算逐一讨论。不过，我们将考察其中表现最佳的算法之一：NSGA-II（Deb 等人，2002年）。该策略总结于算法 2 中。其运行方式基于前沿和拥挤距离的概念。

NSGA-II 遵循遗传算法的基本结构，但采用了不同的交配和生存选择方法。在 NSGA-II 中，第一步是按前沿排序来选择个体。这样一来，就可能需要拆分某个前沿，因为无法让其中所有个体都存活下来。

根据该分割前沿的拥挤距离选择解。在目标空间的参数范围内，曼哈顿距离对应于拥挤距离。另一方面，为了确保每一代都保留极端点，因此将无限拥挤距离赋予这些点。算法 2 展示了这种方法的一种可能实现。

算法 2 用于优化文本可读性的多目标优化技术

```
过程 ORUGA2-MOO
    种群 ←初始化种群 ()
    种群 ←生成随机个体 (种群)
    计算可读性评分 (种群)
    根据帕累托为种群分配等级
    辅助种群 ←生成子代种群（种群）
    当尚未达到停止条件时，执行
        对于种群和辅助种群中的每个个体，执行
            解 ←计算种群的可读性评分
            解 ←根据帕累托为种群分配等级
            解 ←generateNonDominateSolutions（种群）
            解 ←determiningCrowdingDistance（种群）
            对于（每个解）执行
                种群 ←添加解至下一代（种群）
            结束循环
        结束循环
        种群 ←选择低前沿高拥挤距离点（种群）
        种群 ←生成下一代种群（种群）
    结束循环
    优化后的文本 ←最优个体（解）
    优化后的文本 ←必要时纠正错误（优化后的文本）
    返回优化后的文本
```

NSGA-II 是一种遗传算法，具有以下三个特点：它采用精英原则，即只有种群中最优的成员才能传递到后续世代。此外，它还采用了专门用于保持多样性的机制（拥挤距离）。因此，它能够识别非支配解。

4.3. 示例说明

由于我们试图同时优化两个正交目标，因此不可能提供单一的解。尽管如此，我们可以向人工操作员提供一系列解，涵盖从修改最重要的词数以实现全面优化，到只改动最少数量的单词以实现小幅优化。

从通过修改最多词数实现全面优化的解，到通过改动最小值词数实现轻微优化的解。由人工操作员决定选择哪种解。

示例 4 展示了一个真实示例，该示例使用称为 NSGA-II 的多目标优化技术控制待替换词语的词数。其处理的文本内容与地理有关，取自维基百科。我们必须再次强调，尽管理论上可以手动编辑单词以满足给定指标的标准，但这种方法具有透明性，并且能够针对任何所需指标自动运行。请注意，蓝色单词是可用同义词替换的候选词。

EXAMPLE 4. GEOGRAPHY 来源： WIKIPEDIA

目标：使用 WordNet 同义词和 NSGA-II，最小化 FKGL 分数，并尽量减少待替换的单词。

“尼亚加拉瀑布是一组由三座瀑布组成的瀑布群，位于尼亚加拉峡谷南端，横跨边界，分隔加拿大的安大略省和美国的纽约州。三座瀑布中最大的是马蹄瀑布，它横跨两国之间的国际边界，分隔这两个国家。它也被称为加拿大瀑布。较小的美国瀑布和新娘面纱瀑布位于美国境内。新娘面纱瀑布由山羊岛与马蹄瀑布分隔，并由月亮岛与美国瀑布分隔；两座岛屿均位于纽约州。尼亚加拉河由伊利湖流入安大略湖，并由此形成了这些瀑布群；其流量在北美洲所有垂直落差超过 50 米（160 英尺）的瀑布中最高。在每日白天游客参观时间内，每分钟都有超过 168,000 立方米（590 万立方英尺）的水流越过瀑布顶部。”
FKGL 分数：10.72。

待替换词语	预期 FKGL
5	10.43（▽ 2.71%）
6	10.28（▽ 4.10%）
7	10.13（▽ 5.50%）
8	9.98（▽ 6.90%）
9	9.91（▽ 7.56%）
10	9.84（▽ 8.21%）
11	9.76（▽ 8.96%）
12	9.68（▽ 9.70%）
13	9.61（▽ 10.35%）

可以看出，只需对原文稍作修改，就能优化可读性。形式上的微小改动在意义上是否也微不足道，这仍是一个悬而未决的问题。不过，本文稍后会探讨这个问题。

4.4. 实验研究

本节将详细说明我们开展的一项实证研究，以确定这种新方法是否可行。为此，我们首先进行了实验设置，以确定实验的参数。其次，我们完成了测试并收集了未处理数据。接下来，我们分析了收集到的数据，这便是我们的第三点，也是最后一点。

4.4.1. 实验设置

与前一部分一样，本文的重点并非对多目标优化策略的参数进行细致微调。因此，在基于传统参数设置方案进行简短的初步研究后，我们发现有一种配置效果相当不错。因此，我们在实验中使用的参数如下：

- 种群规模 $\{10, 15, 20\}$: 20
- 交配亲本数量 $\{10, 15, 20\}$: 20
- 基因数量：每个要用同义词替换的候选词对应一个基因
- 适应度函数：用户可以在公式 1、2、3、4 和待替换词语之间进行选择
- 停止条件： $\{300, 600, 900\}$：900 代

此外，每项实验都在一台标准计算机上运行，该计算机配备 32 GB 主存和一颗运行频率为 3.20 GHz 的 Intel Core i7-8700 处理器，操作系统为 64 位 Microsoft Windows 10。大部分功能均使用 jMetalPy[2]，实现；这是一款用于设计和实现多目标优化策略的开源 Python库（Benítez-Hidalgo 等，2019）。

4.4.2. 实验

现在，我们将开展有关尽量减少对原始消息形式影响的实验。此前我们关注的是纯粹的优化，而在这里，我们着重于尽量减少被替换的词数，以使消息的外观尽可能少受影响。

图 4 总结了我们实验后得到的结果。在本研究采用的十个用例中，我们尝试同时降低可读性评分（FKGL）和待替换词语的数量。可以看出，我们总是得到一个解的帕累托前沿，这表明替换的词语越少，所进行的优化也越少。不过还应注意，替换的词语越少，对原始消息形式的影响也越微不足道。与此前的实验一样，我们使用不同颜色表示帕累托前沿。每种颜色代表相应案例研究的难度。

[2]https://jmetal.github.io/jMetalPy/

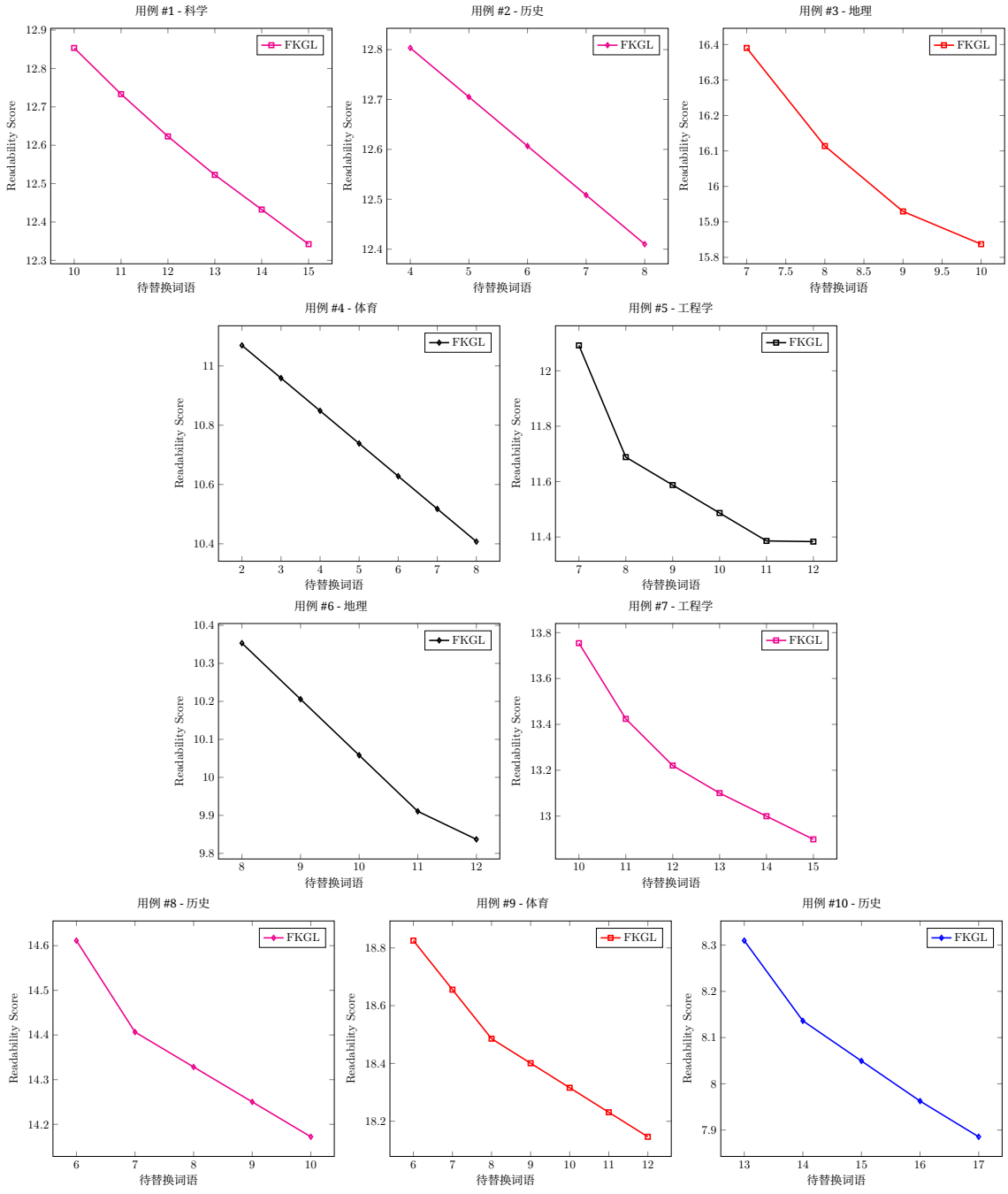


图 4：使用 NSGA-II 获得的十个用例的非支配解

4.4.3. 讨论

我们展示了如何设计一种能够提升文本可读性的解，并通过多目标优化技术实现了该解。此外，这种优化会兼顾相关文本的形式，尽量减少将单词替换为更符合我们所研究的各类公式之可读性标准的同义词所造成的影响。

正如我们在论文前文所述，这类优化也可以通过两种截然不同的方式进行：一种是降低可读性评分，从而让更多人能够理解文本；另一种则是提高可读性评分，这会使能够理解处理后文本的个体数量减少。

尽管解始终会以帕累托前沿的形式呈现，供人工操作员选择，但也可以定义保守型、中间型和激进型配置，分别直接采用保守、中等或更大胆的版本；同时，也应允许手动编辑不符合上下文的单词。

5. 第三部分：保留原文的精髓

在第3节中，我们已经看到，如何设计并实现一种实用的解决方案，以优化任意文本的可读性。尽管这种解决方案似乎运行良好，但仍不能称为完全自动化，因为人工用户仍需判断，用来替换原有词语的词是否与原文语境相符。在第4节中，我们采用多目标优化策略，确保只替换原文中的最少数量的单词，以保留原文形式；我们甚至更进一步，允许人工操作员决定 ORUGA 对原文形式的影响程度。在第5节中，我们将探讨一个尚未解决的问题：如何控制原文与生成文本之间的语义距离。距离较短，意味着原文没有被扭曲；距离较长，则意味着我们大幅提升了可读性，但代价是改变了原文的含义。这第三部分，也是最后一部分，之所以重要，是因为只有这一策略取得成功，才能确保该解决方案完全无需人工监督。

5.1. 技术预备知识

这一部分将加入的额外因素，是测量原文与待发送文本之间的语义距离。这样，我们就能控制消息含义发生的任何显著变化。该语义距离将成为多目标优化策略中的第三个目标。

文献中有一些方法，能够以有意义的方式确定两段文本之间的语义距离，即使它们没有任何共同的单词。为此，这种方法会将单词嵌入为向量。研究表明，在K近邻分类中，该方法的表现优于许多被视为研究现状的方法。

借助词移距离（WMD）（Kusner 等人，2015年），并利用预训练词嵌入，我们可以通过计算一段文本中的嵌入单词到达另一段文本中相应单词所需移动的最小距离，自动评估两段文本之间的语义距离。举例来说，我们可以使用 WordNet 库计算同义词，从而优化文本可读性。同时，我们可以使用 word2vec 来监督生成文本与原文之间的语义差异，使其最小化。我们可以用对抗过程来比喻这一做法。

这种方法的妙处在于，不同的同义词库不再彼此竞争，而是协同合作，尝试衡量（并因此便于控制）原文与最终文本之间的语义距离。接近零的数值表示所考察文本的含义实际上等同，而趋近于无穷大的距离则表示这些文本截然不同。

此外，我们不必担心遗传算法将一个单词替换为一个或多个单词（例如，将“考虑”替换为“纳入考量”），因为词移距离（WMD）的设计已考虑到这种情况，它假设文本长度不必相同。使用 WMD 时，每个单词都会与其他所有单词进行匹配，但匹配权重由流矩阵 T 决定，从而确保即使需要匹配的单词数不相等，语义距离仍具有对称性。

5.2 实现

在本研究中，我们决定使用 WMD 来测量初始文本与流程结束时生成的文本之间的语义距离。这样可以确保用于替换候选词的同义词恰当连贯。之所以这样选择，是因为 WMD 可以通过比较两段文本中彼此相关的重要单词，测量两段文本之间的语义距离，即使它们没有任何共同单词，也能做到这一点。

除此之外，该方法还使用了一种称为词袋表示的表示方式。该方法背后的理念是，通过找出移动源文本与目标文本的分布的最佳方式，应该能够确定两段不同文本之间的距离。

我们可以对策略进行形式化定义：令 d 和 d' 为两个文本的嵌入表示，$T \in \mathcal{R}^{n \times n}$，其中 $T_{ij} \geq 0$ 表示词语 i 在 d 中流向词语 j 在 d'中的量。此外，i与 j之间的距离可以表示为 $c(i, j) = \|x_i - x_j\|$。用 $c(i, j)$表示

从一个词语移动到另一个词语的成本。为了将 d 转换为 d'，必须确保来自 i 的流量等于 d_i，从而使 $\sum_j T_{ij} = d_i$。这样，在满足所有这些约束的前提下，将 d 移动到 d'的最小累计成本由公式 7 所示的解给出。

$$
\begin{aligned}
&arg\ min \sum_{i,j=1}^{n} T_{ij} c(i,j) \\
&subject\ to \sum_{j=1}^{n} T_{ij} = d_i\ \forall i \in \{1,2,3 \cdots n\}\ \wedge \\
&\sum_{i=1}^{n} T_{ij} = d'_j\ \forall j \in \{1,2,3 \cdots n\}
\end{aligned}
\tag{7}
$$

因此，我们使用一个函数来计算两段文本之间的距离：将一段文本中的每个单词移动到另一段文本中最近单词所需的最小距离累加起来。本文将使用由 word2vec 得到的词嵌入，因为它们能够保留单词使用语境中的关键特征。

如今，词移距离常用于计算语义距离，这也是原因之一。唯一的问题是，在最坏情况下，计算受约束的最小累计成本的复杂度为 $O(p^3 \log u)$，其中 u 是文本中不同词的数量（Skianis等，2020年）。因此，对于包含大量不同词的文本，词移距离的表现可能不佳。不过，有一些技术可以提升其性能。

在本文中，我们使用 gensim[3] 库实现的词移距离，并以 word2vec 的词嵌入作为输入（Mikolov等，2013年）。这样一来，原本相互竞争的同义词库便转变为一个对抗性库，有助于控制语义距离。

5.3. 示例说明

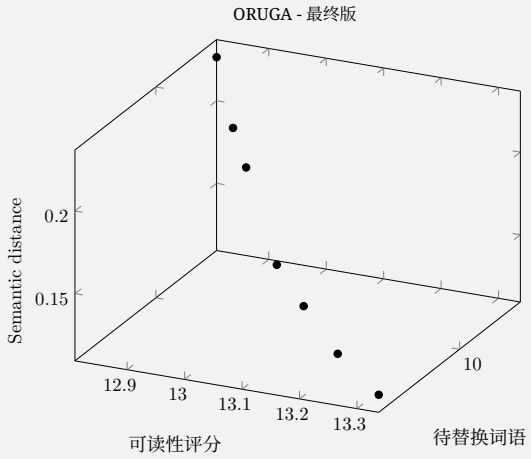
示例 5 向我们介绍了取自维基百科的工程学内容。我们希望观察 ORUGA 在尝试同时降低 FKGL可读性分数、需要替换的词数，以及原文与处理后文本之间的语义距离时如何运作。我们借助 WordNet 替换同义词，并借助 NSGA-II 进行多目标优化。最终目标是确保传达的原始信息遭到扭曲的风险极低。

[3]https://pypi.org/project/gensim/

Example 5. Engineering

原文来源：维基百科

“大数据是指规模过大或复杂度过高、无法使用传统数据处理应用软件处理的数据集。包含许多字段（行）的数据能提供更高的统计效力，而复杂度较高（属性或列更多）的数据则可能导致更高的错误发现率。大数据分析面临的挑战包括数据采集、数据存储、数据分析、搜索、共享、传输、可视化、查询、更新、信息隐私和数据源。大数据最初与三个关键概念相关：规模、多样性和速度。大数据分析在抽样方面面临挑战，因此此前只能进行观察和抽样。由此引出的第四个概念“真实性”，指的是数据的质量或洞察力。” ARI分数：14.43。



以尽可能低的失真风险降低 ARI分数

大数据指的是规模过于庞大或复杂，传统的数据处理应用软件无法处理的数据集。具有许多字段（行）的数据具有更高的统计效力，而复杂度更高（属性或列更多）的数据可能导致更高的虚假发现率。大数据分析面临的挑战包括数据采集、数据存储、数据分析、搜索、共享、传输、可视化、查询、更新、信息隐私以及数据源。大数据最初与三个关键概念相关：规模、多样性和速度。大数据分析在抽样方面带来了挑战，因此此前允许仅进行观察和抽样。因此，第四个概念“真实性”指的是数据的质量或准确性。ARI分数：13.33 - ▽ **7.62%**

该示例表明，我们不再像以前那样盲目操作。现在，我们还会尽量减少待替换词语，以及初始文本与生成文本之间的语义距离。因此，结果更加合理，也可以放心地应用于实际应用环境。可读性评分优化的效果不如以前显著，但无论在形式还是内容上，扭曲原始信息的风险都大幅降低了。

5.4. 实验研究

本节中，我们开展了实证研究，以测试此版本的 ORUGA。为此，我们通过实验设置设计了实验，随后进行了实验并收集了原始数据。最后，我们对收集到的数据进行了分析。

5.4.1. 实验设置

与本研究前面部分的情况一样，精准调整多目标优化策略的参数并非此处的重点。基于常规的参数设置策略进行简短的初步研究后，我们发现一种配置效果相当好。这是上述内容所带来的结果。以下是我们使用的参数列表：

- 种群规模 $\{10, 15, 20\}$: 20
- 交配亲本数量 $\{10, 15, 20\}$: 20
- 基因数量：每个待用同义词替换的候选词对应一个基因
- 适应度函数：公式 1、2、3、4、待替换词语和词移距离（Kusner 等人，2015年）。
- 停止条件： $\{300, 600, 900\}$：900 代

5.4.2. 实验

接下来，我们将开展实验，探究如何最大限度地减少对原始消息含义的影响。现在，我们的重点是生成能够投入实际应用（或接近投入实际应用）的结果。为此，我们将进行实验，看看能否控制原始消息内容与生成文本之间的含义差异。

图 5 汇总了本研究中一直在考察的十个用例所取得的全部结果。可以看出，每个用例无论主题多么具体或具有挑战性，都对应着数量可观的解，范围从最保守的方案（扭曲原始消息风险最低的方案）到最激进的方案（以扭曲原始消息为代价，更显著地降低可读性评分的方案）。根据此前的实验，每种颜色代表一个难度等级。

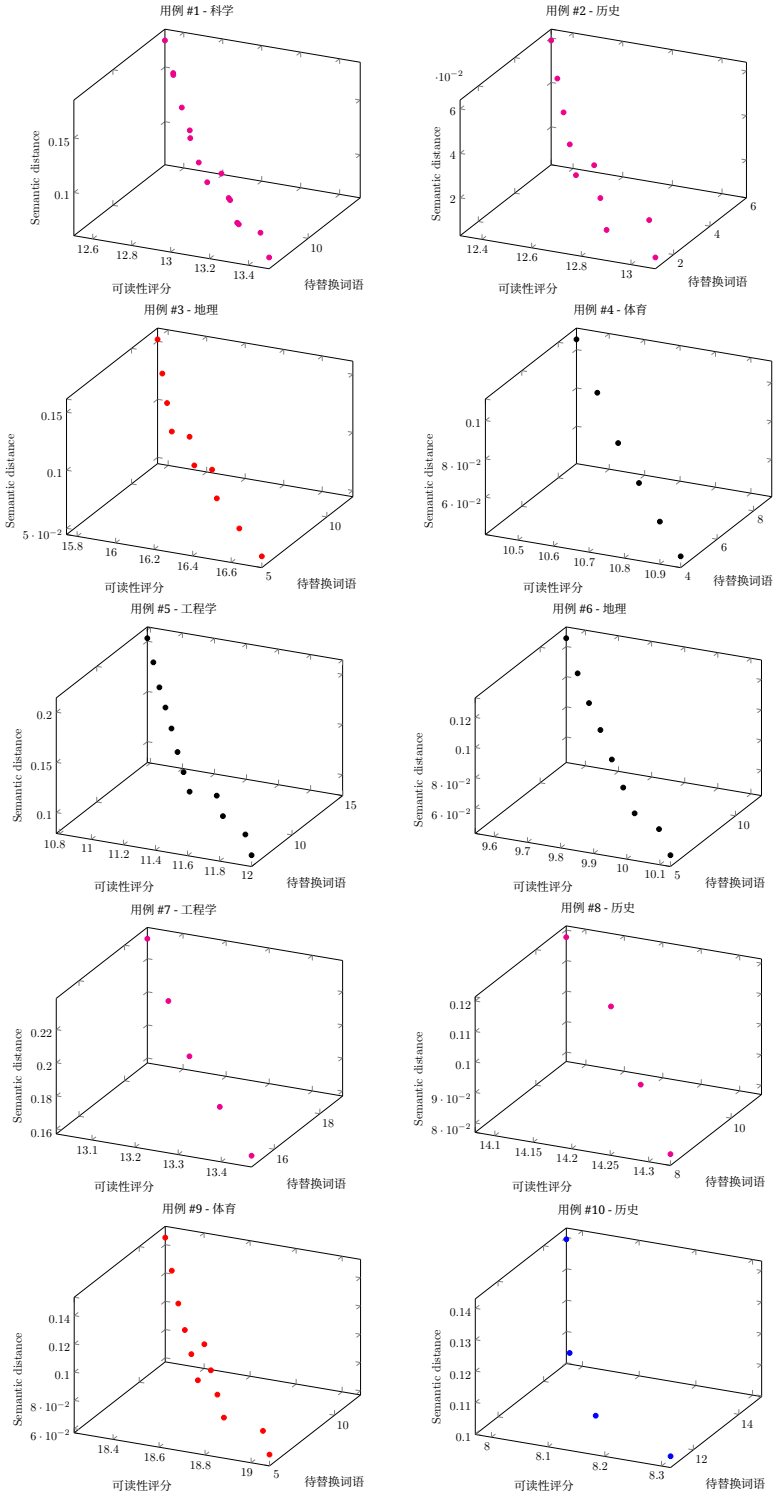


图 5：ORUGA 第三版（也是最终版）所得结果汇总27

5.4.3. 讨论

我们已经了解了如何构建一种解，以提升不同自然属性和可读性水平文本的可读性，并且已经实现了该解。此外，这类优化会兼顾相关文本的内容和结构，尽量减少用更符合我们所研究的各种公式之可读性标准的同义词替换单词所带来的影响。

在整个研究过程中，我们已经看到如何利用遗传算法提升任何英语文本的可读性。正如我们之前所解释的，优化是双向的，也就是说，它可以自动提高或降低文本的可读性。确定优化的基础后，我们便能够控制可替换的词数；还能够确保原文与最终文本之间的语义距离不会急剧增大。因此，我们最初设定的目标——构建一种能够提升任何文本可读性，同时又不显著改变其形式或内容的解——已经实现。

6. 结论

在这项研究中，我们展示了 ORUGA 如何利用遗传算法自动优化文本的可读性。我们证明，通过自动将待优化文本中的某些单词替换为其同义词，可以按照我们期望的方向（降低或提高）优化文本的可读性水平。由于通过多种额外的多目标优化技术，力求将对原文的改动降至最低，因此文本的内容和形式都不会改变。尽管理论上可以利用神经语言模型构建类似的解，但这种方法具有一项显著优势：它是无监督的，无需训练。

一项全面的实证研究表明，我们已成功完成所有实验。首先，这些实验使用三个同义词库和不同的可读性指标，对不同主题（历史、地理、体育、自然和科学）的文本进行了处理。其次，我们设计了一种多目标优化解，以控制待替换词语的词数。我们测试了不同文本，以评估并比较这种方法的可行性。此外，第三，我们测量并控制原文与最终输出文本之间的语义距离。为此，我们采用了一种新颖的技术：利用一个同义词库，通过对抗过程控制使用另一个不同同义词库所获得的结果。

结果表明，本文开篇提出的关于文本可读性优化的假设是成立的。尽管这项研究取得了成功，但仍需谨记，简单公式通常更容易付诸实践，这也是这项工作的局限之一。这些公式有一个根本性的缺陷：无法对单词在语境中的用法进行语义建模，而这对于捕捉更丰富的文本难度概念是必要的。

作为未来的研究方向，一种可行的方法（但计算成本较高）是利用大型数据集训练上下文嵌入模型，例如 BERT。接下来，需要存储单词及其对应的上下文表示，并采用近邻方法来识别包含语境信息的同义词。这样一来，文本处理的影响将进一步减小。

源代码

该方法的源代码以 MIT 许可证发布于以下 GitHub 仓库：https://github.com/jorge-martinez-gil/oruga。

致谢

本文所报告的研究由联邦气候行动、环境、能源、交通、创新与技术部（BMK）、联邦数字化与经济事务部（BMDW）以及上奥地利州资助，研究属于由奥地利研究促进机构 FFG 管理的 COMET 卓越技术能力中心计划框架下的 SCCH 中心项目。

参考文献

Ante，L.（2022）。可读性与科学影响力之间的关系：来自新兴技术话语的证据。《信息计量学杂志》，16，101252。URL：https://doi.org/10.1016/j.joi.2022.101252。doi：10.1016/j.joi.2022.101252。

Ben´ıtez-Hidalgo, A., Nebro, A. J., Garc´ıa-Nieto, J., Oregi, I., & Ser, J. D.（2019）。j metalpy：一种用于多目标优化的元启发式 Python 框架。Swarm and Evolutionary Computation，（第 100598 页）。网址：http://www.sciencedirect.com/science/article/pii/S2210650219301397。doi：https://doi.org/10.1016/j.swevo.2019.100598。Chall, J. S., & Dale, E.（1995）。重新审视可读性：新版 Dale-Chall可读性公式。Brookline Books。Chandrasekaran, D., & Mago, V.（2021）。语义相似度的发展——综述。ACM《计算调查》，54，41:1–41:37。doi：10.1145/3440755。Collins-Thompson, K.（2014）。文本可读性的计算评估：当前及未来研究综述。ITL-International Journal of Applied Linguistics，165，97–135。

Dale, E., & Chall, J. S.（1948年）。用于预测可读性的公式：说明。《教育研究通报》，（第 37–54 页）。Deb, K., Agrawal, S., Pratap, A., & Meyarivan, T.（2002年）。一种快速且具精英策略的多目标遗传算法：NSGA-II。《IEEE进化计算汇刊》，6，182–197。网址：https://doi.org/10.1109/4235.996017。doi：10.1109/4235.996017。Devlin, J., Chang, M., Lee, K., & Toutanova, K.（2019）。BERT：用于语言理解的深度双向变换器预训练。载于 J. Burstein, C. Doran, & T. Solorio（编），《2019年计算语言学协会北美分会大会：人类语言技术论文集》，NAACL-HLT 2019，美国明尼苏达州明尼阿波利斯，2019年6月2日至7日，第1卷（长篇与短篇论文）（第 4171–4186 页）。计算语言学协会。doi：10.18653/v1/n19-1423。Forrest, S.（1996）。遗传算法。ACM《计算调查》，28，77–80。Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S.（1975）。为海军士兵推导新的可读性公式（自动可读性指数、雾指数和弗莱士阅读易读性公式）。海军技术训练司令部技术报告，米灵顿，田纳西州，研究处。Kukkonen, S., & Lampinen, J.（2005）。GDE3：广义差分进化的第三个演进步骤。载于《2005年IEEE进化计算大会》（第 443–450 页）。IEEE，第1卷。Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K.（2015）。从词嵌入到文档距离。载于《国际机器学习大会》（第 957–966 页）。PMLR。Madrazo Azpiazu, I., & Pera, M. S.（2020）。跨语言可读性评估是否可行？《信息科学与技术协会期刊》，71，644–656。Maqsood, S., Shahid, A., Afzal, M. T., Roman, M., Khan, Z., Nawaz, Z., & Aziz, M. H.（2022年）。使用机器学习模型评估英语句子的可读性。《PeerJ计算机科学》，8，e818。网址：https://doi.org/10.7717/peerj-cs.818。doi：10.7717/peerj-cs.818。Martinc, M., Pollak, S., & Robnik-ˇSikonja, M.（2021）。文本可读性的有监督与无监督神经方法。《计算语言学》，47，141–179。Martinez-Gil, J.（2022年）。语义相似度测量堆叠方法的全面综述。《机器学习及其应用》，10，100423。Martinez-Gil, J., & Chaves-Gonzalez, J. M.（2019）。基于模糊逻辑的语义相似度控制器自动设计。《Expert Systems with Applications》，131，45–59。doi：10.1016/j.eswa.2019.04.046。

Martinez-Gil, J.，& Chaves-Gonzalez, J. M.（2020）。一种基于符号回归的可解释语义相似度测量新方法。《专家系统应用》，160，113663。doi:10.1016/j.eswa.2020.113663。
Martinez-Gil, J.，& Chaves-Gonzalez, J. M.（2021）。语义相似度控制器：准确性与可解释性之间的权衡。《知识型系统》，234，107609。doi:10.1016/j.knosys.2021.107609。
Martinez-Gil, J.，& Chaves-Gonzalez, J. M.（2022）。可持续语义相似度评估。《智能与模糊系统杂志》，43，6163–6174。doi:10.3233/JIFS-220137。Mc Laughlin, G. H.（1969年）。SMOG分级——一种新的可读性公式。《阅读杂志》，12，639–646。Meade, C. D.，& S
mith, C. F.（1991）。可读性公式：注意事项与标准。《患者教育与咨询》，17，153–158。
Mikolov, T.，Sutskever, I.，Chen, K.，Corrado, G. S.，& Dean, J.（2013）。词语和短语的分布式表示及其组合性。载于《神经信息处理系统进展》26：第27届神经信息处理系统年会2013。会议论文集，会议于2013年12月5日至8日在美国内华达州太浩湖举行。（第3111–3119页）。
Miller, G. A.（1995年）。WordNet：英语词汇数据库。《美国计算机协会通讯》，38，39–41。
Mitchell, R.（2018）。使用Python进行网络爬取：从现代网络收集更多数据。” O’Reilly Media, Inc.”。Navigli, R.，& Martelli, F.（2019）。词语与词义相似度概述。《自然语言工程》，25，693–714。doi:10.1017/S1351324919000305。Pantula, M.，& Kuppusamy, K. S.（2022年）。一种基于机器学习的模型，用于评估网页的可读性并确定年级水平。《计算机杂志》，65，831–842。URL: https://doi.org/10.1093/comjnl/bxaa113。
doi:10.1093/comjnl/bxaa113。Pedersen, T.，Patwardhan, S.，& Michelizzi, J.（2004）。WordNet：相似度——衡量概念的相关性。载于D. L. McGuinness与G. Ferguson（编），《第十九届全国人工智能大会、第十六届人工智能创新应用大会论文集》，2004年7月25日至29日，美国加利福尼亚州圣何塞（第1024–1025页）。AAAI出版社 / 麻省理工学院出版社。Qin, Y.（2021）。机器生成新闻与人工新闻的可读性比较研究。载于第6届国际控制、机器人与控制论会议，CRC 2021，中国上海，2021年10月9日至11日（

第339–343页）。IEEE。URL: https://doi.org/10.1109/CRC52766.2021.9620136。doi:10.1109/CRC52766.2021.9620136。Rus, V.，Lintean, M. C.，Banjade, R.，Niraula, N. B.，& Stefanescu, D.（2013）。SEMILAR：语义相似度工具包。载于第51届计算语言学协会年会，ACL 2013，《会议系统演示》论文集，2013年8月4日至9日，保加利亚索非亚（第163–168页）。Senter, R.，& Smith, E. A.（1967）。自动可读性指数。技术报告：俄亥俄州辛辛那提大学。Skianis, K.，Malliaros, F. D.，Tziortziotis, N.，& Vazirgiannis, M.（2020）。提升词移动距离的技巧。载于《国际人工神经网络会议》（第761–772页）。施普林格。Todirascu, A.，Fraņcois, T.，Bernhard, D.，Gala, N.，& Ligozat, A.-L.（2016）。衔接特征是否与文本可读性评估相关？载于第26届国际计算语言学会议（COLING 2016）（第987–997页）。Wilkins, D. A.（1972）。《语言教学中的语言学》第111卷。Edward Arnold，伦敦。Zhang, Q.，& Li, H.（2007）。MOEA/D：一种基于分解的多目标进化算法。《IEEE进化计算学报》，11，712–731。URL: https://doi.org/10.1109/TEVC.2007.892759。doi:10.1109/TEVC.2007.892759。Zhu, G.，& Iglesias, C. A.（2017）。计算知识图谱中概念的语义相似度。《IEEE知识与数据工程汇刊》，29，72–85。URL: https://doi.org/10.1109/TKDE.2016.2610428。doi:10.1109/TKDE.2016.2610428。