arXiv:2302.00937v1 [cs.CL] 2 Feb 2023

分句越少越好：解构句子拆分中的可读性

Tadashi Nomoto日本文学研究
资料馆，立川市，东京 190-0014，
日本 nomoto@acm.org

摘要

本文聚焦于文本简化的一个分支——句子拆分。此研究的主要动机，是一个尚未得到证实的观点：如果将一个句子拆分成若干部分，它就会变得更容易理解。本文的首要目标是探究这一观点是否成立。具体而言，我们探讨：将一个句子拆分成两部分还是三部分，是否会产生影响？我们报告了基于亚马逊土耳其机器人所得出的研究结果。

更具体地说，我们引入了一个贝叶斯建模框架，以进一步探究拆分复杂句的特定方式在多大程度上会影响可读性，并考察了一系列从不同视角采用的其他参数，包括临床语言学和认知语言学。贝叶斯建模实验清楚地表明，将句子一分为二比将其分成三部分能更大程度地提升可读性。

1 引言

在文本简化研究中，人们常常忽略一个问题：他们所采用的技术是否真的有助于人们更好地理解文本。这种令人好奇的漠不关心，或许反映了人们默认承认研究所涵盖的数据集存在局限性（Xu 等人，2015），也可能是因为文本简化的目标本身有些模糊。

为应对这一情况，我们考察简化在文本可读性中的作用，尤其关注句子拆分。句子拆分的目标是将一个句子拆分成若干小片段，使它们合在一起能够保留原始含义。本文提出的一个主要问题是：文本拆分是否会影响可读性？鉴于过去在句子拆分方面投入了大量努力，令人意外的是，没有任何研究直接向人们提出这一问题；在大多数情况下，研究人员最终询问的是，生成文本是否“看起来比原文本更简单”，而不是原始未修改的版本（Zhang 和 Lapata，2017），这当然无法说明太多关于其可读性的情况。我们甚至不确定人们对于何谓简化是否存在任何共识。

另一个相关问题是，我们应该把一个句子拆分成多少部分？两部分、三部分，还是更多？本文聚焦于一种特定情形，探讨两句拆分和三句拆分在可读性上是否存在差异。我们还报告了微调语言模型生成的句子拆分与人工标注者的结果相比，效果有多好或多差。

本文采用的一般策略是征询人们的判断，了解简化是否使文本变得易读（第4节），并对他们的回答进行贝叶斯分析，以识别可能影响其决策的因素（第5节）。[1]

2 相关工作

长期以来，英语作为第二语言（ESL）领域一直广泛探索将简化作为提升第二语言（L2）学生阅读表现的一种方法。Crossley 等人(2014) 提出了一系列证据，表明简化文本确实能提升二语学习者的文本理解，这一结论是根据他们的阅读时间以及对相关问题作答的准确率得出的。他们还注意到，与较复杂的文本相比，简单文本的词汇多样性较低、词汇重叠更多，句子之间的语义相似性也更高。Crossley 等人(2011) 强调了连贯性作为影响可读性的因素的重要性。同时，研究发现，对文本进行扩展性修改有助于提高可读性，即通过添加信息使语言

[1]论文发表后不久，我们将在 GitHub 上公开为本研究创建的数据（应该可以在以下地址找到：https://github.com/tnomoto）。

表达不再含糊，修辞也更加明确。Ross 等人 (1991) 报告称，尽管文本长度有所增加，但对文本进行扩展性处理取得了积极效果：二语学生阅读修改后的文本时，在理解题上的得分高于阅读未经修改的原文时的得分。

过去，自然语言处理社区曾协同开展诸多工作，致力于开发据称可用于简化研究的度量指标和语料库（Zhang 和 Lapata，2017；Sulem 等人，2018a；Narayan 等人，2017；Botha 等人，2018；Niklaus 等人，2019；Kim 等人，2021；Xu 等人，2015），但这些工作在解决其研究如何帮助提升读者对文本的可理解性方面远远不够。我们的目标之一，是摆脱将可读性 relegated to a sideline 的主流观点。

3 方法

数据来自两个来源：拆分与改写基准（v1.0）（以下简称 SRB）（Narayan 等人，2017）和 WikiSplit（Botha 等人，2018）。SRB 包含与一组多句简化文本对齐的复杂句，这些简化文本的句子数量从两句到四句不等。WikiSplit 采用了类似格式，但每个复杂句仅对应一个两句简化文本。[2] 我们请亚马逊土耳其机器人工作者（以下简称 Turkers）根据语言特性为简化文本评分，并指明在可读性方面，他们是否更偏好两句版本还是三句版本。

我们从 SRB 中随机抽取了一部分，创建了测试数据（称为 $\mathcal{H}$），其中包含如下形式的三元组：$\langle S_0, A_0, B_0\rangle$、……，$\langle S_i, A_i, B_i\rangle$、……，$\langle S_m, A_m, B_m\rangle$，其中 S_i 是一个复杂句，A_i 是对应的两句简化文本，而 B_i 是其三句版本。尽管 A 在 BART 生成的版本和人工生成的版本之间交替，B 仅涉及人工简化文本。[3]更多说明请参见表 1。

此外，我们从 WikiSplit 中提取了

[2]我们仅使用 WikiSplit 和部分 SRB 对 BART 进行微调，使其成为生成单次拆分（双句）简化文本的模型；同时使用 SRB 开发测试数据，供人工标注者进行语言评估。SRB 源自 WebNLG（Gardent 等人，2017），通过利用与文本片段相关联的资源描述框架来编制简化文本。[3]HSplit（Sulem 等人，2018a）是另一个能够生成多次拆分简化文本的数据集（基于 Zhang 和 Lapata(2017)）。我们在此未采用该数据集，因为其中只有 359 个句子，拆分方式的变化也有限。

	BART	HUM
A（双句拆分）	113	108
B（三句拆分）	–	221

表1：$\mathcal{H}$ 的细分情况。其中，113 条为由 BART-large 生成的 A 类（二分拆分），108 条为人工标注者创建的 A 类。另有 221 条 B 类（三分拆分），均由人工标注者生成。

训练集	DEV
1,135,009 (989,944)	13,797(5,000)

表2：BART 的训练设置。数据来自 SRB（Narayan 等人，2017）和 WikiSplit（Botha 等人，2018）。括号中的数字表示源自 WikiSplit（Botha 等人，2018）。

SRB，另一个数据集 $\mathcal{B}$，包含作为原文的复杂句子，以及作为目标的双句简化文本（表2），即 $\mathcal{B} = \{\langle S'_0, A'_0\rangle, \ldots, \langle S'_n, A'_n\rangle\}$，用于微调语言模型（BART-large）。[4] 微调使用 GitHub 上提供的代码完成。[5]

我们让 Turker 完成的一项任务（或用亚马逊的说法，称为一项人类智能任务）是参加一个包含三个部分的语言测验。第一部分的问题介绍让参与者阅读三段对应于一个三元组 $\langle S_i, A_i, B_i\rangle$ 的短文；第二部分询问他们从三个维度评价 A_i 和 B_i 的语言特性：含义、语法和流畅度；第三部分则提出两个比较问题：（1）A_i 和 B_i 是否比 S_i更易读；（2）A_i 和 B_i 中哪一个更容易理解。

图1 展示了该任务初始部分的屏幕截图。在 **原文** 是一个复杂句或 S_i 对于某些 i。**文本A** 和 **文本B**分别对应于 A_i 和 B_i，其显示顺序是随机的。

总共有 221 个人工智能任务（表1），每个任务均由七人完成。所有参与者均自报为英语母语者，且拥有大学或更高学历。参与资格仅限于居住在美国、加拿大、英国、澳大利亚和新西兰的人员。

[4]https://huggingface.co/facebook/bart-large[5]https://github.com/huggingface/transformers/blob/master/examples/pytorch/translation/run_translation.py

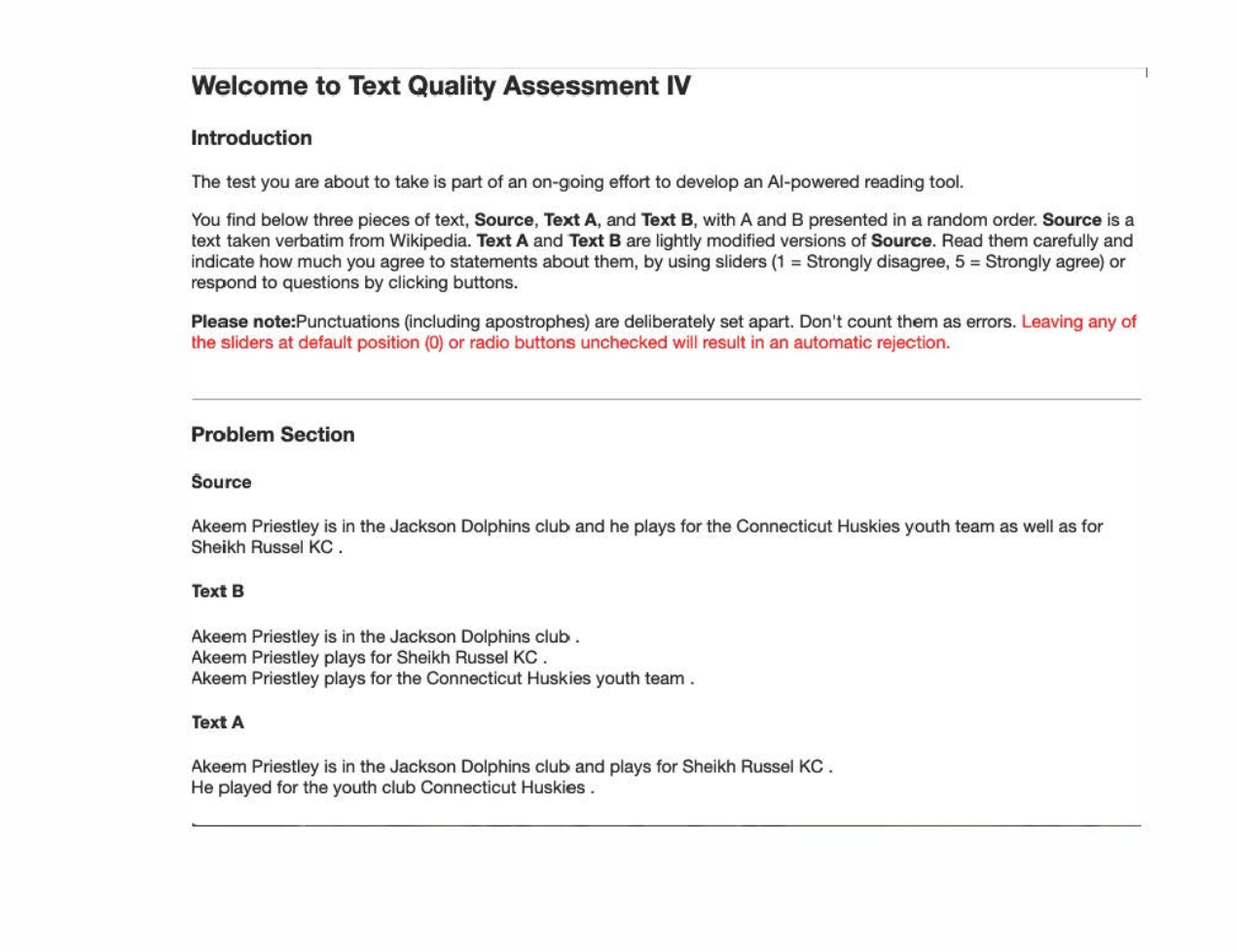

图1：人类智能任务的屏幕截图。这是 Turker 参加测试时看到的画面 t.

4 初步分析

表 3 总结了比较问题的结果。一个标记为《S, BART-A》\|q的问题询问 Turker，在原文和 BART-A 中，哪一个更容易理解；其中，BART-A 是由 BART 生成的两句简化文本。我们共收到 791 份（113×7）回答，其中 32% 表示更喜欢原文，67% 表示更喜欢 BART，另有 1% 的人回答不确定。另一个标记为《S, HUM-A》\|q的问题，将原文与 HUM-A 进行比较，后者是由人工完成的两句拆分文本。该问题收到了 756 份回答（）。结果总体上与《S, BART-A》\|q相似。大多数人更倾向于两句拆分文本，而不是复杂句。三句版本同样受到偏好这一事实表明，无论一个复杂句最终被拆分成多少部分，将其拆开都能提高可读性。

表 4统计了关于双句和三句拆分的比较问题的回答。选择二分式而非三分式简化文本的人更多。

表 5和 6 展示了简化文本在流畅度、语法和意义保留方面的得分，对比了 BART-A 与 HUM-B，[6] 以及 HUM-A 与 HUM-S 两组，评分范围为 1（较差）到 5（优秀）。无论哪种情况，我们都发现 A 和 B 在语法和含义方面的差异不大，但在流畅度方面的差异最大。t 检验发现这种差异具有统计显著性。双句简化文本的流畅度得分通常较高（超过 4.0），而对应的三句拆分得分较低（低于 4.0）。表7 展示了 BART-A 和 HUM-A/B 生成文本的示例。

5 贝叶斯视角

此时我们很好奇的一个问题是，哪些因素促使亚马逊土耳其机器人工作者做出这些决策

[6]如表 5 和 6 所示，BART-A 的表现总体上与 HUM-A 相当，这表明其生成内容大多与人工标注者生成的内容难以区分。

问题	可选项				
	S	BART-A	HUM-B	不确定总数	
《源文本，BART-A》|q	254 (0.32)	527 (0.67)	–	10 (0.01)	791
源文本, HUM-B|q	290 (0.37)	–	490 (0.62)	11 (0.01)	791
	S	HUM-A	HUM-B	不确定总数	
源文本, HUM-A|q	253 (0.33)	494 (0.65)	–	9 (0.01)	756
源文本, HUM-B|q	288 (0.38)	–	463 (0.61)	5 (0.01)	756

表3：比较部分的结果。这里展示了选择每个可选项的亚马逊土耳其机器人工作者人数。源文本：原文。BART-A：由 BART 生成的两句简化文本。HUM-A：人工两句简化文本。HUM-B：人工三句简化文本。S, BART-A\|q 询问亚马逊土耳其机器人工作者，原文和 BART-A 哪个更容易理解。67% 的人表示更倾向于 BART-A，32% 倾向于原文，另有 1% 不确定。S, HUM-B\|q 比较原文和 HUM-B 的可读性。S, HUM-A\|q 考察原文和 HUM-A。

问题	可选项			
	BART-A	HUM-B	不确定总数	
《BART-A, HUM-B》|q	460 (0.58)	316 (0.40)	15 (0.02)	791
	HUM-A	HUM-B	不确定总数	
《HUM-A，HUM-B》|q	439 (0.58)	301 (0.40)	16 (0.02)	756

表4：两句简化与三句简化的比较。无论采用何种生成方式，大多数人都选择了两句简化文本。

类别	HUM-A	HUM-B
**流畅度	4.04 (0.39)	3.75 (0.38)
语法	4.12 (0.32)	4.10 (0.32)
含义	4.31 (0.36)	4.33 (0.28)

表5：HUM-A 和 HUM-B 的平均分和标准差。HUM-A 比 HUM-B 更流畅。注：** = $p < 0.01$。

类别	BART-A	HUM-B
**流畅度	4.04 (0.37)	3.72 (0.36)
语法	4.07 (0.30)	4.05 (0.34)
含义	4.21 (0.38)	4.25 (0.35)

表6：BART-A 和对应 HUM-B 的平均分和标准差。BART-A 的流畅度显著高于 HUM-B。“**”表示这两组在 0.01 水平上存在显著差异。

他们所做的。我们通过构建一个贝叶斯模型来回答这个问题，该模型基于从过去关于可读性及相关领域的文献中汇集的预测因子。

5.1 模型

我们采用贝叶斯逻辑回归。[7]

$$
\begin{aligned}
Y_j &\backsim Ber(\lambda), \\
\text{logit}(\lambda) &= \beta_0 + \sum_i^m \beta_i X_i, \\
\beta_i &\backsim \mathcal{N}(0, \sigma_i) \;\; (0 \leq i \leq m)
\end{aligned}
\tag{1}
$$

Ber（λ）服从参数为λ的伯努利分布。 β_i 表示与随机变量（预测变量） X_i相关的系数，其中 β_0 为截距。我们假设 β_i（包括截距）服从均值为 0、方差为 σ_i的正态分布。 Y_i 取值为 1 或 0。如果 Turker 认为两句简化文本更易读，则 $Y = 1$；如果更偏好三句版本，则 $Y = 0$。

[7]贝叶斯树方法在解释潜在原因与结果之间的关系方面同样有用（Chipman 等人，2010；Linero，2017；Nuti 等人，2019），但我们在此不作探讨。当预测变量与结果之间存在广泛的非线性关系时，后者可能成为一个可行的选择。[8]https://github.com/jasonyux/FastKASSIM[9]https://github.com/luozhouyang/python-string-similarity[10]https://github.com/shivam5992/textstat

文本类型：原文 奥尔德尼机场服务于奥尔德尼岛，其第一条跑道铺有禾本科植物，跑道长 497 米。BART-A 奥尔德尼机场服务于奥尔德尼岛。奥尔德尼机场的第一条跑道铺有禾本科植物，长 497 米。HUM-A 奥尔德尼机场的跑道长 497.0 米，第一条跑道表面铺有禾本科植物。奥尔德尼机场服务于奥尔德尼岛。HUM-B 奥尔德尼机场第一条跑道的表面铺有禾本科植物。奥尔德尼机场的跑道长 497.0 米。奥尔德尼机场服务于奥尔德尼岛。

表7：原始与修改后

类别	变量名	描述	值
合成	**bart**	如果简化由 BART 生成，则为 true；否则为 false。	分类的
衔接性	**ted1**	原文与其拟议简化版本之间的树编辑距离（TED）。[8] 其中 TED 表示数量所需的编辑操作（插入、删除、替换）将一棵解析树转换为另一棵解析树；数量越多，相似度就越低（Boghrati 等，2018；Zhang 和 Shasha，1989）。	连续型
	ted2	简化中所含句子之间的树编辑距离。	连续型
	子集	基于子集的树核（Collins 和 Duffy，2002；Moschitti，2006；Chen 等，2022）[8]	连续型
	子树	基于子树的树核 (Collins 和 Duffy，2002；Moschitti，2006；Chen 等，2022)[8]	连续型
	重叠	Szymkiewicz-Simpson 系数，一种标准化的基数——值，表示两个词集交集的基数（Vijaymeena 和 Kavitha，2016）。[9]	连续型
认知	**Frazier**	从终端节点到根节点或第一个祖先节点的距离最左侧出现的（Frazier，1985）。	连续型
	Yngve	每个单词右侧出现的非终结符数量推导树中单词的（Yngve，1960年）。	连续型
	依存长度	句法分析中每个词元的依存关系数量（Magerman，1995年；Roark等人，2007年）。	连续型
	树节点	解析树中的节点数量（Roark等人，2007年）	连续型
经典	**dale**	Dale-Chall 可读性评分（Chall 和 Dale，1995年）[10]	连续型
	ease	Flesch 阅读易读度（Flesch，1979）[10]	连续型
	FK等级	Flesch-Kincaid 年级水平 (Kincaid等人, 1975)[10]	连续型
感知	**语法**	语法完整性（人工编码）	连续型
	含义	语义忠实度（人工编码）	连续型
	流畅度	语言自然度（人工编码）	连续型
结构性	**拆分**	如果句子被二分，则为 true；否则为 false。	分类的
信息型	**s**	衡量目标文本中保留了多少原始内容（Sulem 等人，2018b）。	连续型
	AMSA		

表8：预测因子

5.2 预测因子

我们使用表8中所示的预测变量。它们分为六类：合成、衔接性、认知、经典、感知和结构性。合成特征表明该简化是否由 BART 创建：若是，则取值为真；否则为假。衔接性类别下的特征是我们对 SYNSTRUT 和 CRFCWO 的改编；这两者属于 McNamara 等人(2014) 为衡量句间衔接性而创建的多种特征。SYSTRUCT 通过考察句法相似性，或统计相邻句子共享的公共子图中的节点数，来衡量句子之间的一致性和连贯性。我们以**树编辑距离** (Boghrati 等, 2018)替代 SYSTRUCT，因为与只寻找单个公共子图的 SYSTRUCT 相比，它使我们能够处理多个子图。CR- FCWO 给出两个相邻句子中共同出现的词元的归一化计数。我们在此使用 Szymkiewicz-Simpson 系数对其进行了模拟，其表达式为 $O(X,Y) = \frac{|X \cap Y|}{\min(|X|,|Y|)}$。

认知类别中的预测因子取自临床语言学和认知语言学领域的研究（Roark等人，2007年；Boghrati 等，2018）。它们反映了衡量句子认知复杂度的各种方法。例如，**Yngve** 评分将一个单词的认知需求定义为：在推导规则中位于该单词右侧且尚未处理的非终结符数量。

5.2.1 Yngve

请参见图 2。 **Yngve** 为句法分析中的每条边赋予一个数字，以反映其认知负荷。名词短语 NP 得到“1”，因为其右侧有一个作为姐妹节点的动词短语 VP。一个单词的认知负荷定义为从根节点到该单词路径上各数字之和。在图 2中，“Vanya”将得到 $1+0+0=1$，而“home”为 0。对单词的负荷取平均值，即可得到 Yngve 复杂度。

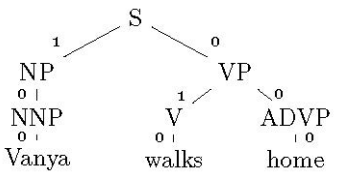

图 2：Yngve 评分

5.2.2 Frazier

Frazier 将单词的句法深度（从叶节点到在某条推导规则中首次位于最左侧的祖先节点的距离）视为决定句子复杂度的最重要因素。如果我们对 **Frazier** 应用于图2中的句子，就会得到类似图3所示的分数。‘Vanya’得 $1+1.5=2.5$，‘walks’得1，而‘home’得0（它没有最左侧祖先节点）。Roark等人(2007年) re-

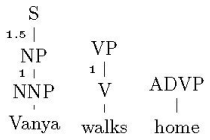

图 3：Frazier 评分

报告称，**Yngve** 和**Frazier** 在区分具有轻度记忆障碍的受试者方面都表现良好。

5.2.3 依存长度

依存长度（依存长度）和 **树节点**（树节点）也是Roark等人（2007年）发现有效的特征之一。前者衡量依存句法分析中的依存关系数量，后者则衡量短语结构树中的节点数量。

5.2.4 子集和 子树

子集 和 **子树** 都是基于 树核 (Collins 和 Duffy，2002；Moschitti，2006； Chen 等，2022）[11] 前者考察两个句法分析共享多少子图，后者则考察共享多少子树。请注意，子树是以终端节点结尾的结构。

5.2.5 经典可读性特征

我们还纳入了可读性研究中早已确立为标准的特征，即 Dale-Chall可读性指数、Flesch 阅读易读度和 Flesch-Kincaid 年级水平（Chall 和 Dale，1995 年；Flesch，1979； Kincaid等人，1975）。

5.2.6 感知特征

感知类别中的特征来自亚马逊土耳其机器人工作者对我们要求他们评估的简化文本质量所作的判断。我们没有

[11]树核是一个定义为 $K(T_1, T_2) = \sum_{n_1 \in N(T_1)} \sum_{n_2 \in N(T_2)} \Delta(n_1, n_2)$ 的函数，其中

$$\Delta(a,b) = \begin{cases} 0 & \text{if } a \neq b; \\ 1 & \text{if } a = b; \\ \prod_i^{C(a)} (\sigma + \Delta(c_a^{(i)}, c_b^{(i)})) & \text{otherwise.} \end{cases}$$

$C(a)$ = a 的子节点数量， $c_a^{(i)}$表示 a的第 i个子节点。令 $\sigma > 0$。

未对任务中的语法正确性、含义和流畅性作出任何具体定义或说明。因此，他们的回答很可能是自发且基于直觉的。

5.2.7 拆分和 SAMSA

最后，我们有 **拆分**，用于记录简化是否为二分式：如果是，则取值为 真，否则取值为假 。**SAMSA** 是近期加入简化指标体系的一项指标，用于考察句子拆分后，原文中的命题内容保留了多少（Sulem 等人，2018b）。（越高越好。）我们将所有特征都标准化为 **BART** 和 **拆分**以外的 zz 分数，其中 $z = \frac{x-\bar{x}}{\sigma}$ 。

5.3 评估

我们使用 BAMBI（Capretto 等人，2020）[12] 训练模型（公式 1）。预热期为 50,000 次，抽样 4,000 次，使用 4 条 MCMC 链（哈密顿）。为了单独考察每个预测变量的效应（或重要性），我们做了两件事：首先查看每个因素的后验分布，即与某个预测变量相关的系数 β ，该预测变量与预测变量

-

变量，并观察它与 0 相差多远；一个

-

另一种方法是进行消融研究，考察某个特征的缺失如何影响模型性能。我们使用一种称为“渡边－赤池信息准则”（WAIC）的指标来衡量模型性能（Watanabe，2010；Vehtari 等人，2016），它是 AIC 的一种贝叶斯版本（Burnham 和 Anderson，2003）。[13]

图4 展示了使用马尔可夫链蒙特卡洛进行 4,000 次抽样迭代后，与预测变量相关的参数后验分布。与这些参数相关的链均未出现ˆ发散。我们得到 R 介于 1.0 和 1.02 之间，所有 β_i的稳定性都相当可靠（Gelman 和 Rubin，1992），这表明所有相关参数都已成功收敛。[14]

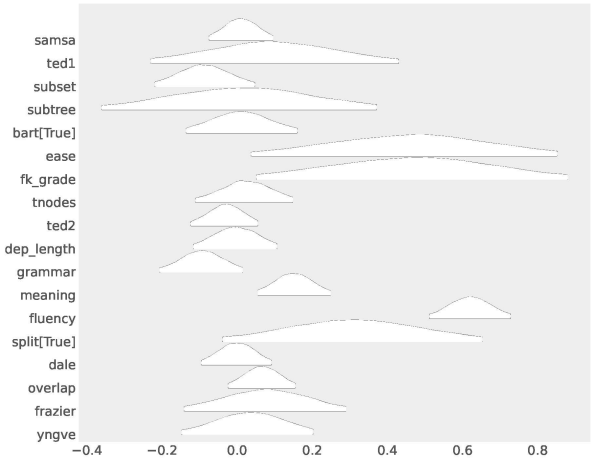

图4：完整模型中系数（β）的后验分布。分布离 0 越远，它对预测结果就越重要。

乍一看，要理解图4有些困难，但一个普遍接受的经验法则是：假设分布集中在 0 附近的因素，在解释观测结果方面不如那些分布偏离零值的因素重要。按照这一法则，最有可能影响可读性的因素是：**易读性**、**子集**、**FK等级**、**语法**、**含义**、**流畅度**、**拆分**和**重叠**。尚不清楚的是，这些预测因子在多大程度上影响了可读性。

一种有效的判断方法是进行消融研究，即通过考察从模型中移除某个单独因素会在多大程度上降低模型性能，来隔离该因素效应的方法。研究结果如表9所示。每一行都表示移除某个特定预测变量后模型在 WAIC 上的表现。因此，表9 中的“ted1”表示一个包含表8中所有预测因子、但不包含 **ted1**的模型。蓝色行表示没有停用任何特征的完整模型。某项结果高于基底模型，说明移除某个特征产生了正效应，也就是说该特征是冗余的。低于基底模型则意味着移除该特征产生了负效应，表明不应舍弃该特征。沿着表格向下，特征的重要性逐渐增加；向上则逐渐降低。因此，最重要的是**流畅度**，其次是**含义**；最不重要的是**子树**，其次是**dale**，以此类推。从表 9 可以看出，为解释可读性，我们需要保留哪些预测因子：它们是

[12] https://bambinos.github.io/bambi/main/index.html[13] WAIC 的定义如下。

$$\text{WAIC} = \sum_{i}^{n} \log \mathbb{E}[p(y_i|\theta)] - \sum_{i}^{n} \mathbb{V}[\log p(y_i|\theta)]. \quad (2)$$

$\mathbb{E}[p(y_i|\theta)]$ 表示在 θ 的后验分布下的平均似然， $\mathbb{V}[\alpha]$ 表示 α 的样本方差，即 $\mathbb{V}[\alpha] = \frac{1}{S-1}\sum_{1}^{S}(\alpha_s - \bar{\alpha})$，其中 α_s 是从 $p(\alpha)$ 中抽取的一个样本。WAIC 分数越高，模型越好。 n 是数据点的数量。ˆ

[14] R = 链内方差与链间方差之比，是检查收敛情况的标准工具（Lambert，2018）。

该比值越接近 1，MCMC 链越有可能已经收敛。

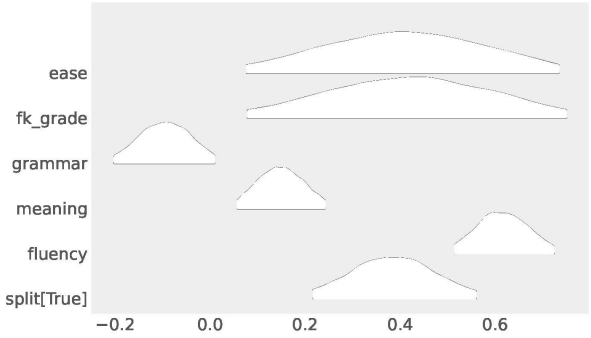

图5：简化模型中系数参数的后验分布。

语法、**拆分**、**FK等级**、**易读性**、**含义**和**流畅度**（称为“所选特征”）。注意，**BART**处于负值区域，这意味着从可读性的角度来看，人们并不在意简化是由人工还是机器完成的。**SAMSA** 也被发现处于负值区域，这意味着从信息的角度来看，将一个句子拆分为两个句子所包含的信息，与将其拆分为三个部分一样多。

为了进一步明确它们的重要程度，我们又进行了一次仅涉及所选特征的消融实验。结果见表10。最底部是**流畅度**，倒数第二的是**拆分**，接下来是**含义**，依此类推。在表格中越往上，特征的重要性就越低。这些特征的后验分布见图5。[15] 不出所料，它们的分布都远离零，其中**流畅度**离零最远。结果表明，与认为经典可读性指标（如**易读性**和**FK等级**）用处不大的普遍看法相反，当人们被问及可读性时，这些指标对他们所做的决策有很大影响。

6 结论

在这项研究中，我们提出了两个问题：把一个句子拆开，能否帮助读者更好地理解文本？如果能，拆成多少部分是否重要？我们发现，拆分确实能让读者更好地与文本互动（表 3），而且，双句简化文本明显优于三句简化（表 3、9、10）。为什么两句拆分能带来更好的简化，仍是个谜。一个可能的答案或许在于拆分可能造成的潜在对句子层面话语结构的破坏；Crossley 等人 (2011、2014)认为，话语结构的完整性是简化的关键组成部分。我们认为，这一问题值得未来进一步探讨。

7 局限性

• 我们没有考虑将一个句子拆分成超过三个句子的情况。这主要是因为我们未能找到包含大量长度超过三句的人工简化文本的数据集。虽然我们的研究结论不太可能不适用于超过三句的情况，但如果能在涉及超过三个句子的案例上检验这一假设，将会更为理想。

• 我们为当前研究招募的这批人总体上是受过良好教育、以英语为第一语言的成年人。因此，我们在本研究中得到的结果未必适用于二语学习者、未成年人或未接受大学程度教育的人群。

8 致谢

我们感谢匿名审稿人向我们分享他们的意见和想法。对于他们付出的努力，我们深表感激和谢意。

参考文献

Reihane Boghrati、Joe Hoover、Kate M. Johnson、Justin Garten 和 Morteza Dehghani。2018。对话层面句法相似度指标。行为研究方法，50(3)：1055–1073。 [链接]。

Jan A. Botha、Manaal Faruqui、John Alex、Jason Baldridge 和 Dipanjan Das。2018。学习如何根据维基百科编辑历史拆分并改写。载于*2018*年自然语言处理实证方法会议论文集，第732–737页，比利时布鲁塞尔。计算语言学协会。 [链接]。

K.P. Burnham 和 D.R. Anderson。2003年。模型选择与多模型推断：一种实用的信息论方法。Springer New York。 [链接]。

Tomás Capretto、Camen Piho、Ravin Kumar、Jacob Westfall、Tal Yarkoni 和 Osvaldo A. Martin。2020。Bambi：一种用于拟合贝叶斯线性模型的简易 Python 接口。 [链接]。

J.S. Chall 和 E. Dale。1995年。可读性再探：新版戴尔-查尔可读性公式。Brook- lineBooks。 [链接]。

[15]我们发现其稳定性达到了 $1.0 \leq \hat{R} \leq 1.01$，近乎完美。马尔可夫链蒙特卡洛（MCMC）的设置，即预热迭代次数和抽样次数，均与之前相同。

效应	预测变量	排名↑	WAIC↑	p_WAIC↓	d_WAIC↓	se↓	dse↓
	子树	0	-1899.249	17.797	0.000	17.787	0.000
	dale	1	-1899.287	17.852	0.038	17.791	0.207
	依存_长度	2	-1899.362	17.916	0.113	17.777	0.211
	Yngve	3	-1899.406	17.904	0.157	17.777	0.464
	树节点	4	-1899.414	17.898	0.165	17.797	0.408
	bart	5	-1899.421	17.967	0.172	17.786	0.216
−	SAMSA	6	-1899.450	18.018	0.201	17.776	0.315
	ted1	7	-1899.557	17.996	0.308	17.771	0.575
	ted2	8	-1899.632	18.019	0.383	17.782	0.624
	Frazier	9	-1899.740	18.096	0.492	17.779	0.708
	子集	10	-1900.069	17.811	0.820	17.741	1.282
	重叠	11	-1900.431	17.966	1.182	17.750	1.511
ref.	base	12	-1900.532	19.089	1.283	17.787	0.208
	语法	13	-1900.780	17.979	1.531	17.698	1.657
	拆分	14	-1900.852	18.030	1.603	17.697	1.776
	ease	15	-1901.657	17.962	2.408	17.670	2.064
+	fk_grade	16	-1901.710	18.030	2.462	17.685	2.049
	含义	17	-1903.795	17.885	4.546	17.425	3.071
	流畅度	18	-1965.386	17.938	66.137	14.067	11.349

表9：WAIC比较。*p_waic* = 有效参数数量（Spiegelhalter 等人，2002），用于估计模型的复杂度：数值越大，模型越复杂。*d_waic* = 与最佳模型的 WAIC 差距。*se* = WAIC 估计值的标准误。*dse* = 最佳模型与其余各模型之间 WAIC 估计差值的标准误。↑ 表示数值越大越好。↓ 表示相反的情况。

预测变量	排名↑	waic↑	p_waic↓	d_waic↓	se↓	dse↓
base	0	-1891.901	7.181	0.000	17.485	0.000
语法	1	-1892.235	6.183	0.335	17.365	1.672
ease	2	-1893.515	6.137	1.614	17.350	2.324
fk_grade	3	-1893.626	6.161	1.726	17.366	2.358
含义	4	-1895.308	6.145	3.407	17.111	3.059
拆分	5	-1900.028	6.169	8.127	17.038	4.247
流畅度	6	-1956.041	5.935	64.140	13.784	11.289

简化模型的 WAIC 结果

Maximillian Chen、Caitlyn Chen、Xiao Yu 和 Zhou Yu。2022。FastKASSIM：一种基于快速树核的句法相似度指标。*arXiv* 预印本 *arXiv:2203.08299*。

Hugh A. Chipman、Edward I. George 和 Robert E. McCulloch。2010。BART：贝叶斯加性回归树。《应用统计年鉴》，4(1)。[link]。

Michael Collins 和 Nigel Duffy。2002。用于句法分析和词性标注的新型排序算法：离散结构上的核函数与投票感知机。载于计算语言学协会第*40*届年会论文集，第263–270页，费城，宾夕法尼亚州，美国。计算语言学协会。 [link]。Scott A. Crossley、David B. Allen 和 Danielle S. McNamara。2011。文本可读性与直觉式简化：可读性公式比较。《外语阅读》，23(1):84–101。Scott A. Crossley、Hae Sung Yang 和 Danielle S. McNamara。2014。简化文本究竟有何简单之处？一项关于文本理解与文本处理的计算和心理语言学研究。《外语阅读》，26(1):92–113。R. Flesch。1979。《如何写通俗英语：律师与消费者指南》。Harper & Row。 [link]。Lyn Frazier。1985。句法复杂度。载于 David R. Dowty、Lauri Karttunen 和 Arnold M. Zwicky 编，自然语言句法分析：心理学、计算与理论视角，《自然语言处理研究》，第129–189页。剑桥大学出版社。 [link]。Claire Gardent、Anastasia Shimorina、Shashi Narayan 和 Laura Perez-Beltrachini。2017。为自然语言生成微规划器创建训练语料库。载于计算语言学协会第*55*届年会论文集（第*1*卷：长篇论文），第179–188页，温哥华，加拿大。计算语言学协会。 [link]。Andrew Gelman 和 Donald B. Rubin。1992。使用多个序列进行迭代模拟的推断。《统计科学》，7(4):457–472。 [link]。Joongwon Kim、Mounica Maddela、Reno Kriz、Wei Xu 和 Chris Callison-Burch。2021。BiSECT：利用双语平行文本学习拆分和改写句子。载于*2021*年自然语言处理实证方法会议论文集，第6193–6209页，线上和蓬塔卡纳，多米尼加共和国。计算语言学协会。 [link]。J. Peter Kincaid、Robert P. Fishburne Jr.、Richard L. Rogers 和 Brad S. Chissom。1975。为海军士兵推导新的可读性公式（自动可读性指数、雾指数和弗莱施阅读易读度公式）。技术报告，海军技术训练司令部。Ben Lambert. 2018.《贝叶斯统计学学生指南》。SAGE。Antonio R. Linero. 2017. 树模型贝叶斯方法综述。《统计应用与方法通讯》，24:543–559。David M. Magerman. 1995. 用于句法分析的统计决策树模型。载于第*33*届计算语言学协会年会，第276–283页，马萨诸塞州剑桥，美国。计算语言学协会。 [link]。Danielle S. McNamara、Arthur C. Graesser、Philip M. McCarthy和Zhiqiang Cai. 2014.《运用 *Coh-Metrix* 自动评估文本与话语》。剑桥大学出版社。Alessandro Moschitti. 2006. 让树核适用于自然语言学习。在第*11*届欧洲分会计算语言学协会会议上，第113–120页，意大利特伦托。计算语言学协会。 [link]。Shashi Narayan、Claire Gardent、Shay B. Cohen和Anastasia Shimorina. 2017. 拆分并改述。载于*2017*年自然语言处理实证方法会议论文集，第606–616页，丹麦哥本哈根。计算语言学协会。 [link]。Christina Niklaus、André Freitas和Siegfried Handschuh. 2019. MinWikiSplit：一个包含最少命题的句子拆分语料库。载于第*12*届国际自然语言生成会议论文集，第118–123页，日本东京。计算语言学协会。 [link]。Giuseppe Nuti、Lluís Antoni Jiménez Rugama和Andreea-Ingrid Cross. 2019. 一种可解释的贝叶斯决策树算法。*arXiv:1901.03214v3 [*统计机器学习*]*。 [link]。Brian Roark、Margaret Mitchell和Kristy Hollingshead. 2007. 用于检测轻度认知障碍的句法复杂度测量方法。载于生物、转化与临床语言处理，第1–8页，捷克共和国布拉格。计算语言学协会。 [link]。Steven Ross、Michael H. Long和Yasukata Yano. 1991. 简化还是扩展？两种文本改写对外语阅读理解的影响。夏威夷大学英语作为第二语言工作论文，10(22):1–32。David J Spiegelhalter、Nicola G Best、Bradley P Carlin和Angelika Van Der Linde. 2002. 模型复杂度与拟合的贝叶斯度量。《英国皇家统计学会杂志：*B*辑（统计方法学）》，64(4):583–639。Elior Sulem、Omri Abend和Ari Rappoport. 2018a. BLEU 不适用于文本简化评估。载于2018年会议论文集

自然语言处理实证方法，第738–744页。计算语言学协会。 [link]。

Elior Sulem、Omri Abend 和 Ari Rappoport。2018b。用于文本简化的语义结构评估。载于 *2018* 年北美计算语言学协会分会会议论文集：人类语言技术，第 *1* 卷（长篇论文），第 685–696 页，美国路易斯安那州新奥尔良。计算语言学协会。 [link]。

Aki Vehtari、Andrew Gelman 和 Jonah Gabry。2016。使用留一交叉验证和 WAIC 进行实用的贝叶斯模型评估。统计与计算，27(5)：1413–1432。 [link]。

M. K. Vijaymeena 和 K. Kavitha。2016。文本挖掘中的相似度量综述。机器学习与应用：国际期刊，3(1)：19–28。 [link]。

Sumio Watanabe。2010。奇异学习理论中贝叶斯交叉验证与广泛适用信息准则的渐近等价性。*JournalofMachineLearning Research*，第 3571–3594 页。

Wei Xu、Chris Callison-Burch 和 Courtney Napoles。2015。当前文本简化研究中的问题：新数据大有帮助。计算语言学协会会刊，3：283–297。[链接]。

Victor H. Yngve。1960年。语言结构的一种模型和假说。美国哲学学会会刊，104(5)：444–466。 [链接]。

Kaizhong Zhang 和 Dennis Shasha。1989。树之间编辑距离及相关问题的简单快速算法。*SIAM* 计算杂志，18(6)：1245–1262。 [链接]。

Xingxing Zhang 和 Mirella Lapata。2017。利用深度强化学习进行句子简化。载于*2017*年自然语言处理实证方法会议论文集，第584–594页，丹麦哥本哈根。计算语言学协会。 [链接]。