基于规则的原子句提取方法

Lineesha Kamana宾夕法尼亚州立大学 lpk5305@psu.edu

Akshita Ananda Subramanian宾夕法尼亚州立大学 apa6321@psu.edu

Mehuli Ghosh
宾夕法尼亚州立大学
mvg6189@psu.edu

Suman Saha
宾夕法尼亚州立大学
szs339@psu.edu

arXiv:2601.00506v1 [cs.CL] 1 Jan 2026

摘要

自然语言常将多个想法组合成复杂句。原子句子抽取旨在将复杂句拆解为更简单的句子，使每个句子都表达一个单独的想法；这有助于提升信息检索、问答和自动推理系统的性能。以往研究对“拆分与改写”任务进行了形式化定义，并建立了评估指标；采用大型语言模型的机器学习方法也提高了抽取准确率。然而，这些方法缺乏可解释性，也无法深入揭示哪些语言结构会导致抽取失败。尽管一些研究探索了基于依存关系抽取主谓宾三元组和从句，但尚无系统性分析考察哪些特定从句结构和依存关系会导致抽取困难。

本研究通过分析复杂句结构（包括关系从句、状语从句、并列结构和被动结构）如何影响基于规则的原子句子抽取性能，弥补了这一研究空白。我们使用 WikiSplit 数据集，在 spaCy 中实现了基于依存关系的抽取规则，生成了 100 组金标准原子句集合，并使用 ROUGE 和 BERTScore 评估性能。该系统取得了 **ROUGE-1 F1 = 0.6714**、**ROUGE-2 F1 = 0.478**、**ROUGE-L F1 = 0.650** 和**BERTScore F1 = 0.5898**，表明其在词汇、结构和语义层面具有中等至较高的对齐度。具有挑战性的结构包括关系从句、同位语、并列谓词、状语从句和被动结构。总体而言，基于规则的抽取具有尚可的准确率，但容易受到句法复杂性的影响。

1 引言

许多自然语言句子包含多个命题，这些命题通过嵌入从句和其他复杂的语言结构表达出来。对于问答和自动推理等任务，这些句子可能包含过多交织在一起的想法，使计算机系统更难清晰地理解其含义，从而导致它们在此类任务中的表现变差。

原子句子抽取有助于绕过这一挑战。它将复杂句拆解为简单的单元，每个单元都表达一个单一的观点。这些更小的“原子”单元更易于系统理解，因为它们能更清楚地呈现主语、动词和谓语之间的关系。

以往的研究通过拆分与改写问题探索了这一领域，并且严重依赖在数据集上训练的大型神经网络模型。尽管这些模型表现出色，但它们并未说明为何选择特定的拆分方式，而且很难理解哪些语言结构会导致它们出错。另一方面，基于规则的方法和基于依存关系的方法透明度高得多，因为它们依赖句法本身。

为了探究哪些特定语言结构会给基于规则的原子句子抽取系统带来最大问题，本文将回答以下两个研究问题：

哪些类型的句子结构（如关系从句、条件句或否定句）最难正确拆分？

句子抽取的准确度如何？

为回答这些问题，我们使用 spaCy 的依存句法分析器构建了一个基于规则的提取系统，并在 WikiSplit 数据集的一个子集上对其进行了评估。其目标是明确哪些特征会导致错误，并确定未来的系统如何更好地处理更加复杂的句子。

2 背景与文献综述

将复杂句拆分为更简单的原子单元的研究，源于拆分

并改写任务。Narayan等人（2017年）对这一方法进行了形式化。他们的研究引入了 BLEU 和 SARI 等评估指标，用于衡量拆分后句子的准确率和实用性。这为后续研究奠定了坚实的基础。

基于这一框架，Botha等人（2018年）推出了 WikiSplit。这是一个大型数据集，由维基百科编辑历史中自然产生的句子拆分构建而成。这项工作有助于证明，大规模数据集能够显著提升人工智能模型在拆分与改写方法上的表现。与以往工作使用的数据集相比，WikiSplit 数据集还提供了更多样的句子结构，因此是评估现实场景中句子拆分效果的宝贵基准。

与神经网络方法相比，基于依存关系的抽取方法侧重于句子结构句法的可解释性。Nikalus等人（2018年）对开放信息抽取（OpenIE）系统进行了综述，并着重介绍了如何利用依存句法分析抽取主语-动词-宾语模式和类从句单元。对 nsubj、dobj、advcl 和 elcl 等静态关系的讨论，直接影响了我们制定识别原子句子的规则。该综述还强调了基于规则的系统的优势与局限，进一步表明有必要详细分析哪些语言结构会给基于依存关系的抽取带来问题。

句法简化研究表明，通过拆出关系从句、分离从属从句以及拆分并列动词短语等操作降低语法复杂度，可以显著提升读者和下游自然语言处理系统对文本的理解能力（Siddharthan，2006）。

Liu、Cohen 和 Lapata（2018）介绍了用于解析话语表征结构（DRS）的方法，强调了可解释的基于规则的推导。他们的研究展示了结构化语义句子表征的价值。这也支持了我们的决定：构建基于规则的抽取模型，而非神经模型，以确保能够追溯特定句法如何影响抽取效果。

这些研究为我们的工作提供了概念和方法论基础。先前的研究帮助确定了数据集和评估框架，并展示了依存信息的实用性，同时强调了可解释性的重要性。然而，这些研究都没有考察哪些类型的从句会给基于规则的原子句子抽取带来挑战。

3 方法论

3.1 数据集选择

我们从 WikiSplit 数据集中抽取了 100 个复杂句。WikiSplit 是一个为拆分与改写任务构建的大规模语料库。不过，WikiSplit 只提供二分拆分，不足以评估多从句原子分解的效果。我们的目标是将一个句子拆解为表达单一观点的所有命题，因此将复杂句作为基础，用于构建更细粒度的原子句。

3.2 金标准标注流程

将选出的 100 个句子逐一手动拆分为 2 至 5 个原子句，具体数量取决于整个复合句中彼此独立的意思数量。标注过程中严格遵循准则，以确保所有句子的拆分方式一致。

- 每个原子句必须只表达一种状态或事件。
- 关系从句，例如（“Diana，曾担任市长……”），被拆分为独立的原子句。
- 并列谓词，例如（“Alice 唱歌并跳舞”），被拆分为多个句子，例如“Alice 唱歌”和“Alice 跳舞”。
- 诸如“before”和“although”引导的状语从句，被拆分为各自独立的句子。

3.3 数据清洗与分词

为了准备数据集以供拆分并最终进行评估，首先使用 Pandas 删除重复句子。随后，将所有句子转换为小写以便匹配，这样就不会因缺少大写字母而影响原始句子拆分规则，也不会影响后续的 ROUGE 或 BERT 分数。接着，使用了 spaCy 的“en_core_web_sm”模型，因为它高效且轻量。该模型用于分词、标注词性以及进行依存句法分析。

3.4 基于规则的拆分

基于规则的抽取系统尝试通过检测并列连词来拆分每个句子，检测依据可以是单词本身（“and”和“or”），也可以是词性（“conj”）。系统还识别出被标记为从属结构的从句，例如状语从句和关系从句（“advcl”和“relcl”）。此外，系统还尝试根据依存句法分析，为短语分配正确的主语。

> “Anna 吃了一个苹果和一根香蕉” →“Anna 吃了一个苹果”，“Anna 吃了一根香蕉”

在此情况下，系统的并列连词检测会从两个句子中分别去除“and”，并为“a banana”单独创建一个“ate”；随后，主语重构会将“Anna”分配给“ate a banana”。

3.5 与黄金标准数据的评估

模型生成的原子句与人工创建的黄金标准集合中的对应句子进行了配对。为计算生成的原子句的准确率，计算了以下指标:

- ROUGE-1、ROUGE-2、ROUGE-L，用于量化模型生成的原子句与黄金标准数据之间的词汇和结构重叠程度。
- BERTScore（使用 RoBERTa-large），用于表示模型生成的原子句与黄金标准数据之间的语义相似度。

3.6 错误类别

此外，通过观察模型生成的原子句集合中的错误，我们确定了基于规则的系统失败的多个方面，并从数据中为每个类别提供了一个示例。

主语缺失

模型输出：“改名为魏忠贤。”标准答案：“这对夫妇将他改名为魏忠贤。”→ 模型以动词短语开头，丢失了原句的名词短语主语。

宾语缺失

模型输出：“纳达尔直落两盘落败。”标准答案：“纳达尔直落两盘输掉了比赛。”→ 模型保留了事件，却省略了宾语“比赛”，导致谓词未充分说明。

并列结构错误

模型输出：“她自学了如何画画。”标准答案：“她自学了画画，并在高中时开始出售漫画。”→ 并列谓词“开始出售漫画”完全丢失了。

关系从句错误

模型输出：“Marks & Co. 位于这座五层楼建筑内。”标准答案：“Marks & Co. 所在的那座五层楼建筑至今仍然存在。”→ 关系从句“Marks & Co. 所在的位置”被当作主句处理，导致意思完全颠倒。

状语从句错误

模型输出：“他在百老汇演出。”标准答案：“在美国居住的二十年间，他曾在百老汇演出。”→ 状语修饰语“在美国居住的二十年间”被省略了。

同位语错误

模型输出：“Maria Johanna Görtz 是一位瑞典艺术家。”参考答案：“Maria Johanna Görtz，也以 Jeanette Görtz 之名为人所知，是一位瑞典艺术家。”→ 同位语“也以 Jeanette Görtz 之名为人所知”被省略了。

截断

模型输出：“这张专辑已发行”参考答案：“这张专辑限量发行，仅推出 12 英寸黑胶唱片版本。”

→ 该模型只提取主要动词短语，并截去其余内容，即使这些内容可能还可用于另一个原子句中。

其他

模型输出：“一些冰山崩解后漂出了峡湾。”
黄金标准：“格陵兰约有 10% 的冰山每年崩解并流出峡湾。”

→ 存在多个问题，包括缺少数值细节、缺少修饰语“每年”，以及时态错误。

4 结果

4.1 总体表现

基于规则的原子句抽取系统以此前讨论过的、由人工标注的 100 个 WikiSplit 源句组成的黄金标准集进行评估，该数据集生成了 252 对对齐的原子句对。系统表现通过 ROUGE 和 BERTScore 进行度量，这些指标共同用于评估句对在词汇、结构和语义方面的相似度。

指标	精确率	召回率	F1
ROUGE-1	0.7645	0.6328	0.6714
ROUGE-2	0.5595	0.4473	0.4776
ROUGE-L	0.7427	0.6098	0.6495

表1：比较模型生成的原子句与黄金标准的 ROUGE 分数。

ROUGE-1 F1 分数表明，黄金标准与模型生成分数之间的词元级相似度处于中高水平；ROUGE-L F1 分数则表明，基于规则的抽取系统保留了黄金标准的句子结构。然而，ROUGE-2 F1 分数显示，基于规则的系统在保留局部短语相似度方面表现不佳。

指标	精确率	召回率	F1
BERTScore	0.5898	0.5414	0.5648

表2：模型生成的原子句与黄金标准原子句之间的 BERTScore 语义相似度。

BERTScore 处于中等水平，这表明基于规则的抽取在一定程度上保留了语义相似度，但在处理更深层的语义内容时可能会失败。这通常是由于论元缺失或分句重构不完整所致。

4.2 长度和动词特征

为了帮助量化基于规则的分解系统生成的原子句与黄金标准之间的结构差异，我们考察了平均词元长度和动词数量，以判断是否遗漏了重要的结构部分。

度量	模型生成	黄金标准
平均原子句子长度	11.3769	12.6666
每个原子句的平均动词数	1.2857	1.4285

表3：模型生成的原子句与标准原子句在长度和动词数量上的比较。

4.3 错误类别

利用依存特征对错误类型进行人工分类后，确定了 8 个主要类别，其中包括一个“其他”类别，该类别涵盖同一个原子句中的多种错误类型。

错误类型	错误比例
宾语缺失	0.444444
多重错误／其他	0.392857
并列结构错误	0.051587
主语缺失	0.043651
同位语错误	0.027778
状语从句错误	0.019841
关系从句错误	0.015873
截断	0.003968

表4：模型生成的原子句中错误类型的分布。

最常见的错误类别是宾语缺失，即系统正确识别了谓词，却未能包含相应的宾语。基于规则的系统生成的一个符合该错误类别的不完整句子示例是“柳花夫人得到了”，其中缺少宾语。第二常见的错误类别包括基于规则的抽取器处理较复杂的依存关系时出错所产生的杂项错误或错误输出组合。该类别所占比例表明此类错误普遍存在，许多原子句都包含多个错误。

并列结构也可能导致错误，例如系统未能拆分并列谓词，或错误地将并列名词短语视为独立事件。主语缺失错误可能发生在主语传播未能跨越

合并从句时；这种情况在被动结构中更为常见。此外，关系从句、状语从句和同位语从句错误出现的频率要低得多，但也反映出识别从句边界的难度。在 WikiSplit 数据集这类包含嵌套从句的复杂句中，用于判断从句边界的句法线索往往模棱两可。

4.4 结果总结

总体而言，这项工作的结果表明，词汇重叠度较高，但语法完整性仍不稳定。因此，基于规则的系统生成的原子句与金标准句子之间的语义相似度仅属中等，这说明许多预测出的原子句保留了正确含义，却缺少必要的论元。

这些错误大多源于谓词不完整，具体而言，是宾语和补语缺失。此外，我们可以看到，复杂的句法结构，包括并列结构、同位语以及关系从句和状语从句，是句子分解中大多数结构性错误的原因。黄金标准集中的每个原子句包含的动词和词元也更多，这表明基于规则的系统可能生成信息不充分的结果，或遗漏原句中的重要信息。

这些发现表明，基于规则的系统能够较好地生成原子句，但人类语言的多样性极大，且某些结构存在歧义，因此基于依存关系的方法难以区分不同的想法，尤其是当这些想法在句法上似乎属于同一单元时。

5 结论与未来工作

本研究旨在考察不同句法结构如何影响基于规则的原子句抽取系统的性能。通过使用 WikiSplit 数据集中的 100 个经人工标注的句子对系统进行评估，我们发现，基于规则的方法能够可靠地处理简单的主谓宾结构，但面对关系从句、同位语、并列谓词和状语从句等复杂形式时，往往更难应对。这些结构常常导致主语或宾语缺失，或从句不正确，凸显了句法复杂性如何直接造成谓词不完整和边界错误。尽管存在这些局限，该系统仍取得了中等水平的性能指标，与黄金标准相比表现尚可；这表明，基于规则的抽取能够提供一种更透明、可解释的解决方案，同时保留原文的大部分含义。

我们使用前文讨论过的、由人工标注的100个 WikiSplit 源句组成的黄金标准集，对基于规则的原子句子抽取系统进行了评估，并由此得到252对对齐的原子句对。我们使用 ROUGE 和 BERTScore 进行度量，这些指标共同评估句对的词汇、结构和语义相似度。结果表明，尽管词汇和结构层面的重合度较高，但语义完整性仍有提升空间，尤其是在涉及嵌入式从句或并列从句的情况下。

研究已表明，基于规则的从句切分有助于信息抽取任务。例如，在药理学文本挖掘中，一种语言学从句切分算法提升了从复杂生物医学句子中抽取药物相互作用的效果（Segura-Bedmar 等，2011年）。在此基础上，未来可以探索原子句子抽取在问答和事实核查等更具体应用中的表现。此外，扩展并改进规则集，使其更好地处理嵌套从句、并列结构和同位语，或者结合其他组件来恢复缺失的论元，都有望提高整个系统的准确率。

参考文献

Botha, J. A., Faruqui, M., Alex, B., & Ganchev, K.（2018年）。学习拆分与改写维基百科编辑历史中的文本。arXiv。https://arxiv.org/abs/1808.09468

Liu, J., Cohen, S. B., & Lapata, M.（2018年）。话语表征结构解析。《计算语言学协会汇刊》，6，147–160。https://doi.org/10.1162/tacl_a_00011

Narayan, S., Gardent, C., Shimorina, A., & Perez-Beltrachini, L.（2017年）。拆分与改写。2017年自然语言处理实证方法会议（EMNLP）论文集，606–616。

https://doi.org/10.18653/v1/D17-1063

Niklaus, C., Cetto, M., Freitas, A., & Hand- schuh, S.（2018年）。开放信息抽取综述。第27届国际计算语言学会议论文集，3866–3878。https://aclanthology.org/C18-1326

Zhu, Z., Bernhard, D., & Gurevych, I. (2010). 用于句子简化的单语树结构翻译模型。《第23届国际计算语言学会议论文集》，1353–1361。https://aclanthology.org/C10-1158

Siddharthan, A. (2006). 句法简化与文本衔接。《语言与计算研究》，4，77–109。https://doi.org/10.1007/s11168-006-9011-1

Zhang, X., & Lapata, M. (2017). 利用深度强化学习进行句子简化。《2017年自然语言处理经验方法会议论文集》，584–594。https://doi.org/10.18653/v1/D17-1062

Chatterjee, N., & Agarwal, R. (2021). DEP- SYM：一种利用依存树的轻量级句法文本简化方法。《计算文本简化研讨会（CTTS）论文集》。

Kim, J., Maddela, M., Kriz, R., Xu, W., & Callison-Burch, C. (2021). BiSECT：利用双语平行文本学习拆分与改写句子。《2021年自然语言处理实证方法会议论文集》，6193–6209。https://doi.org/10.18653/v1/2021.emnlp-main.500

Segura-Bedmar, I., Martínez, P., & Sánchez- Cisneros, D. (2011). 一种基于语言规则的方法，用于从药理学文档中提取药物相互作用。《BMC生物信息学》，12(Suppl 2)，S1。https://doi.org/10.1186/1471- 2105-12-S2-S1