AMR-SELOR: 一种用于生成语义逻辑解释的神经符号框架

王耀彬[1)] 孙家兴[2)]

[1)](北方工业大学人工智能与计算机学院，北京市石景山区晋元庄路5号，100144)
[2)](北方工业大学人工智能与计算机学院，北京市石景山区晋元庄路5号，100144)

摘 要 随着深度学习在关键决策领域的广泛应用，如何为其决策过程提供忠实且可理解的解释成为构建可信人工智能的核心挑战。现有自解释模型(如基于逻辑规则推理的SELOR)在保证解释忠实性方面取得进展，但其依赖词汇或统计特征的原子，导致逻辑规则语义浅显，难以刻画复杂因果关系，且易受句法变化干扰。为弥合这一“语义鸿沟”，本文提出AMR-SELOR框架，将抽象语义表示(AMR)的深层语义结构与SELOR的逻辑推理能力结合，实现从“词法匹配”到“语义推理”的转变。具体而言，框架以AMR图中提取的语义三元组替代原子库，从而提升解释与人类认知模型的对齐度，并增强对文本复述的鲁棒性。本文进一步阐述AMR-SELOR的理论架构与关键实现技术(包括SPRING等先进AMR解析器)，并设计包含“复述鲁棒性”测试的综合验证方案，以探索更深刻、鲁棒且符合人类直觉的AI可解释性路径。

关键词 可解释人工智能(XAI)，自解释模型，神经符号推理，抽象语义表示(AMR)，逻辑规则推理

中图法分类号 TP393 **DOI号**: 10.1234/cjc.2025.000001

AMR-SELOR: A Neural Symbolic Framework for Generating Semantic Logic Interpretations

YaoBin Wang[1)] JiaXing Sun[2)]

[1)](School of Artificial Intelligence and Computer Science, North China University of Technology (NCUT), Beijing 100144, China)
[2)](School of Artificial Intelligence and Computer Science, North China University of Technology (NCUT), Beijing 100144, China)

Abstract

With the increasing deployment of deep learning in critical decision-making domains, providing faithful and human-understandable explanations has become a central challenge for trustworthy artificial intelligence (AI). Existing self-explanatory models, such as SELOR, achieve explanation fidelity by embedding logical reasoning into model structures. However, their reliance on lexical or statistical atoms yields semantically shallow rules, limiting the capture of complex causal relations and making them vulnerable to syntactic variations. To bridge this “semantic gap,” we propose AMR-SELOR, a neurosymbolic framework that integrates the deep semantic structures of Abstract Meaning Representation (AMR) with SELOR’s logical reasoning. Specifically, semantic triples extracted from AMR graphs replace the original atom library, enabling closer alignment with human cognitive models and improved robustness to paraphrasing. We present the theoretical architecture and key implementation techniques of AMR-SELOR, including the use of advanced AMR parsers such as SPRING, and design a comprehensive evaluation scheme incorporating a novel “paraphrasing robustness” test. This work aims to advance AI interpretability toward deeper, more robust, and cognitively aligned explanations.

Keywords Explainable artificial intelligence, self-explanatory models, neural symbolic reasoning, abstract semantic representation, logical rule. reasoning

---

收稿日期：2025.11.01；最终修改稿收到日期：2025.11.30. 王耀彬(第一作者)，男，2002年生，研究生，主要研究领域为自然语言处理、多模态大模型. E-mail: 2025354100103@mail.ncut.edu.cn

1 引言

1.1 研究背景与研究动机

在人工智能技术深度融入社会生产与生活的今天，深度学习模型已在自然语言处理、医疗诊断、金融风控等高风险领域展现出卓越的性能。然而，这些模型复杂的内部结构和非线性的计算过程使其决策逻辑往往不透明，形成了所谓的“黑箱”问题。这种不透明性不仅阻碍了用户对模型决策的信任，也为模型的调试、纠错和确保其公平性、合乎伦理带来了巨大挑战。因此，发展可解释人工智能(Explainable AI,XAI)技术，使模型的决策过程透明化、可理解，已成为AI领域一个紧迫且至关重要的研究方向。

1.2 现有方法与不足

XAI 领域的研究大致可分为两大范式：事后解释(post-hoc explanation)与自解释(self-explaining)[1]模型。事后解释方法，如LIME[2]和SHAP[3]，在模型训练完成后，通过分析其输入输出关系来推断决策依据。尽管这类方法具有模型无关的灵活性，但其解释与模型的实际推理过程相分离，可能存在“不忠实”(unfaithful)的风险，即生成的解释无法真实反映模型的内在逻辑。这种潜在的不一致性导致了实践者中普遍存在的“不安感”和信任问题[2]。

相比之下，自解释模型将解释机制作为其架构不可或缺的一部分，强制模型在预测的同时生成一个人类可理解的解释。SELOR(Self-explaining deep models with logic rule reasoning)[4]是这一范式的杰出代表。

SELOR 框架通过一个严谨的概率公式：

$$p(y|x,b) \propto \sum_{\alpha} p(b|\alpha)p(y|\alpha)p(\alpha|x) \qquad (1\text{-}1)$$

将预测过程重构为寻找并评估逻辑规则的过程，从根本上保证了解释的忠实性。更重要的是，SELOR引入了“人类精度”(Human Precision)这一核心概念，将其定义为“人类对模型为其预测所提供理由的认同程度”，并将解释质量的评估标准从机器中心转向了人类中心。下图为SELOR的工作流程。

然而，尽管SELOR在架构上取得了突破，其推理的基础——逻辑规则的“原子”(atom)——仍然存在固有的局限性。SELOR的原子通常是基于词汇或统计特征的布尔表达式，例如“评论中‘awesome’一词出现次数≥ 1”或“用户年龄≥ 40”。这类原子本质上是浅层的、基于表面形式的，它们能够捕捉特征与标签之间的统计相关性，但往往无法触及更深层次的语义因果关系。SELOR论文中一个极具启发性的案例研究表明，模型可能仅仅因为“vegas”一词在数据集中与负面评论存在虚假相关，就生成了包含该词的解释规则。这对人类来说显然是不合理、不可信的。这种“语义鸿沟”的存在，即模型推理单元与人类认知单元之间的不匹配，是限制当前自解释模型达到更高层次可解释性的关键瓶颈。

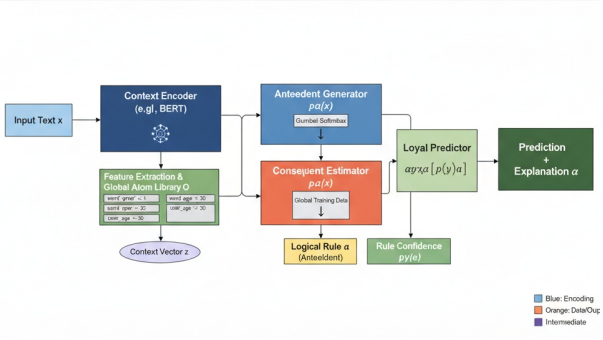

Fig. 1 图1-1 SELOR工作流程

1.3 本文贡献

针对上述问题，本论文提出了一项创新性的解决方案：将抽象语义表示(Abstract Meaning Representation, AMR)融入SELOR框架，构建一个名为AMR-SELOR的神经符号系统。AMR是一种现代语义形式化理论，它将句子的核心语义表示为一个有向无环图，其中的节点代表概念(如事件、实体)，边代表它们之间的语义关系(如施事、受事、地点等)。AMR的核心优势在于其能够“抽象掉”句法上的差异，对具有相同核心语义的不同句子给出相同的表示。本文的核心思想是用从AMR图中提取的、蕴含丰富语义信息的“语义三元组”(例如，‘(praise-01, :ARG0, staff)‘，意为“员工是‘表扬’这一动作的施事者”)来彻底取代SELOR原有的词法原子库。

通过这一根本性的变革，我们旨在将模型的推理基础从“句子中出现了什么词”提升到“句子表达了什么含义”。我们预期AMR-SELOR将带来以下主要贡献：

(1) 提出AMR-SELOR 框架：一个新颖的、深度集成的神经符号架构，首次将AMR的深层语义分析能力与SELOR的忠实逻辑推理机制相结合。
(2) 重构逻辑解释单元：将解释的“原子”从词法特征重新定义为语义三元组，使生成的逻辑规则在本质上更贴近人类的因果认知模型。
(3) 规划技术实现路径：提供一套完整的、具有

可操作性的技术实现方案，涵盖了从利用先进AMR解析器(如SPRING)进行语义解析，到改造SELOR核心模块(后件估计器与前件生成器)的全过程。

(4) 设计全面的验证策略：提出一套严谨的实验验证方案，不仅包括与原SELOR在各项指标上的直接对比，还引入了一项新颖的“复述鲁棒性”测试，以量化评估模型在语义层面上的稳定性。

2 相关工作

本研究位于可解释人工智能(XAI)、计算语言学和神经符号AI三个领域的交叉点。本章将分别对这些领域中的相关工作进行综述，以明确AMR-SELOR框架的学术定位和创新性。

2.1 可解释性AI范式

可解释AI的研究旨在打开深度学习的“黑箱”，其发展历程中形成了两大主流技术范式。

事后解释(Post-hoc Explanation)：这是早期XAI研究的焦点。这类方法作用于一个已训练好的模型，试图通过外部扰动或内部探查来解释其行为。代表性工作包括LIME[2](Local Interpretable Model-agnostic Explanations)和Anchor[5]。LIME通过在预测实例的局部邻域内用一个简单的、可解释的模型(如线性模型)来近似复杂模型的行为。Anchor人则旨在寻找一个“锚点”，即一个能充分固定预测结果的输入特征子集。尽管这些方法因其模型无关性而应用广泛，但它们的核心缺陷在于解释与模型决策过程的分离，导致了解释的“忠实性”问题备受质疑。研究者指出，事后解释可能被恶意攻击，产生误导性的结果或无法捕捉到特征间的复杂交互，从而不能真实反映模型的内在逻辑。

自解释模型(Self-explaining Models)：为了从根本上解决忠实性问题，研究重心逐渐转向自解释模型。这类模型将解释的生成过程作为其体系结构的一个内在组成部分。SENN[1](Self-Explaining Neural Networks)是一个早期的代表，它将模型分解为对可解释基概念的线性组合，从而生成基于特征重要性的解释。然而，SENN的解释形式较为简单，表达能力有限。SELOR框架则在SENN的基础上迈出了重要一步，它将解释形式从线性权重提升为更具表达力的逻辑规则1。通过强制模型经由一条全局一致且局部连贯的逻辑规则进行预测，SELOR不仅保证了忠实性，还显著提升了解释的“人类精度”。本研究正是建立在SELOR的坚实基础之上，旨在通过引入深层语义信息，进一步提升其逻辑规则的质量和可理解性。

2.2 抽象语义表示AMR

可解释AI的研究旨在打开深度学习的“黑箱”，其发展历程中形成了两大主流技术范式。

基础理论与形式化：抽象语义表示(AMR)[6]由Banarescu等人于2013年正式提出，旨在为自然语言句子提供一种规范化的、捕捉核心语义的图表示。AMR的核心设计原则是抽象掉表层句法结构，使得意义相同但表述方式不同的句子能够映射到同一个AMR图上。一个AMR图是一个有根、有向的无环图，其中节点代表概念(通常是词语的词义或PropBank中的谓词-论元框架)，边则代表它们之间的语义关系(如':ARG0'代表施事，':ARG1'代表受事，':location'代表地点等)。AMR通常使用PENMAN[7]格式进行文本序列化，例如，对于句子“The boy wants to go”，其PENMAN表示为'(w/want-01 :arg0 (b/boy) :arg1 (g/go-01 :arg0 b))'。这种结构化的表示方法为进行深度的语义分析和推理提供了可能。

先进的解析技术：将自然语言文本自动转换成AMR图的过程称为AMR解析(parsing)。近年来，随着预训练语言模型的发展，AMR解析的性能取得了长足的进步。其中，由罗马大学Sapienza NLP实验室开发的SPRING[8]框架是当前最先进的AMR解析器之一。SPRING将AMR解析和生成任务统一视为序列到序列(seq2seq)的转换问题，通过巧妙的图线性化技术，利用强大的Transformer架构实现了端到端的解析，取得了顶尖的性能，且无需复杂的预处理流水线。SPRING的开源和高性能为本研究将AMR集成到SELOR框架中提供了坚实的技术基础。

应用领域：AMR作为一种强大的语义中间表示，已被成功应用于多种下游NLP任务，如机器翻译、文本摘要、问答系统和信息抽取，证明了其在捕捉和利用句子核心语义方面的有效性。

2.3 神经符号AI前沿

AMR-SELOR的构想本质上是一种神经符号(Neuro-Symbolic, NeSy[9])AI方法。NeSy旨在融合神经网络强大的模式识别、泛化能力与符号系统明确的知识表示、逻辑推理能力，以期构建出既能学习又能推理的更强大、更鲁棒的AI系统。这一研究方向与认知科学中的双过程理论相呼应，将神经网络类比为快速、直觉的“系统1”，而将符号推理类比为缓慢、审慎的“系统2”。

美国国防部高级研究计划局(DARPA)的ANSR[10](Assured Neuro Symbolic Learning and Reasoning)等项目的大力投入，也反映了学术界和工业界对NeSy作为构建可信AI关键路径的共识。近年来，已有研究开始探索将AMR作

为符号知识整合到神经模型中。例如，有工作将AMR解析与指代消解等符号模块结合[11]，以提升语言理解的准确性和鲁棒性。这些工作为本研究提供了重要的思想借鉴，即AMR可以作为连接神经网络表示与符号逻辑推理的有效桥梁。

2.4 可解释性中的语义表示

将语义信息用于提升AI可解释性的探索尚处于起步阶段，但已展现出巨大潜力。近期的一项代表性工作是AMREx[12]，一个用于可解释事实核查的系统。AMREx利用AMR之间的相似度(通过Smatch度量)来判断一个声明与证据之间是否存在语义蕴含或矛盾关系，并能通过对齐的AMR节点来提供部分解释。AMREx的成功表明，AMR能够为解释任务提供有价值的语义grounding。

然而，AMREx等现有方法通常以流水线或事后分析的方式使用AMR，AMR分析的结果作为外部信息提供给下游的分类或解释模块。这与AMR-SELOR的核心理念存在本质区别。AMR-SELOR旨在将语义推理内化为模型决策过程的核心环节，模型的预测直接依赖于其能否找到一条基于AMR三元组的、具有高置信度的逻辑规则。这种深度、原生的集成方式，是对现有工作的进一步发展，有望实现一种更彻底、更忠实的语义可解释性。

3 主要成果论述

本章将详细阐述AMR-SELOR框架的核心设计理念、系统架构，并论证其相较于原版SELOR在可解释性质量上的潜在优势。本框架的核心贡献在于，它通过引入抽象语义表示(AMR)，从根本上重塑了自解释模型的推理基石，推动解释生成从浅层词法匹配向深层语义推理的深刻转变。

3.1 概念架构

AMR-SELOR的整体架构继承了SELOR的概率推理框架，但对其核心组件进行了根本性的改造，以无缝集成AMR的语义信息。整个信息处理流程可分解为五个关键步骤。

(1) 并行语义与上下文编码：当一个输入文本x进入系统时，它被并行送入两个模块：

- 上下文编码器：与SELOR一样，一个预训练的深度模型(如BERT 或RoBERTa)将文本x编码为一个高维的上下文向量z。这个向量捕捉了文本的深层句法和语境信息。
- 语义解析器：一个先进的AMR解析器(如SPRING)将文本x解析为其对应的AMR图G。这个图显式地表示了文本的核心语义结构。

(2) 实例级原子库提取：一个新增的“原子提取器”(Atom Extractor)模块负责遍历AMR图G，并提取其中所有的语义三元组。这些三元组构成了该输入实例x专属的、动态的候选原子库$A_x = \{t_1, t_2, ..., t_n\}$。例如，对于句子“The staff was amazing”，解析后可能提取出三元组’(staff, :domain-of, amazing-01)’。

(3) 语义前件生成：改造后的“深度前件生成器”(Deep Antecedent Generator)接收上下文向量z和实例级原子库A_x作为输入。它的任务不再是从一个全局固定的、庞大的词汇库中进行搜索，而是学习如何从这个小而精的、与当前输入高度相关的语义三元组集合A_x中，选择出最具解释力的三元组，并组合成一条逻辑规则(前件)α。

(4) 全局后件评估：与SELOR类似，“后件估计器”(Consequent Estimator)模块负责评估生成的规则α的全局置信度$p(y|\alpha)$。这个模块预先在整个训练数据集上进行了训练，学习了不同语义规则组合与最终预测标签y之间的全局统计关联。

(5) 忠实预测生成：模型的最终预测输出是基于规则置信度的函数，完全遵循SELOR的“无作弊”原则。模型必须找到一条能够被全局验证的、高质量的语义规则，才能对预测结果产生高置信度。

下图为本文所设计的AMR-SELOR框架工作流程图。

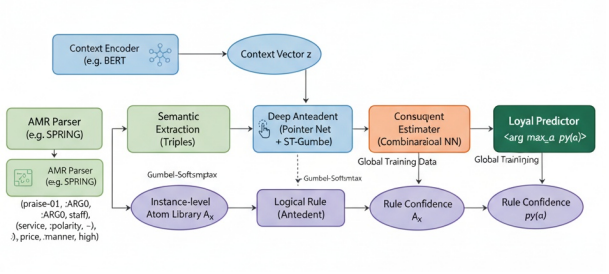

Fig. 2 图3-1 AMR-SELOR工作流程

这一架构的精妙之处在于，它将复杂的语义理解任务(由AMR解析器完成)与忠实的逻辑推理任务(由SELOR的核心框架完成)进行了解耦和串联。AMR解析器扮演了一个强大的“语义预处理器”或“结构化注意力机制”的角色，它极大地缩小了前件生成器的搜索空间，使其能够专注于在高度相关的语义单元上进行推理，而不是在海量的词汇海洋中进行盲目探索。

3.2　语义原子：从词素到三元组

AMR-SELOR最核心的创新在于对“原子”这一基本解释单元的重新定义。下表与下图清晰地对比了标准SELOR原子与我们提出的AMR-SELOR语义原子的本质区别：

表3-1　语义原子对比

特征	标准SELOR原子	AMR-SELOR
基本单元	词汇/特征的存在性	语义关系(三元组)
示例	`"amazing"≥1`	`(staff, :ARG0-of, praise-01)`
来源	全局词汇表	输入AMR图
含义	存在”amazing”	员工是“表扬”的施事者

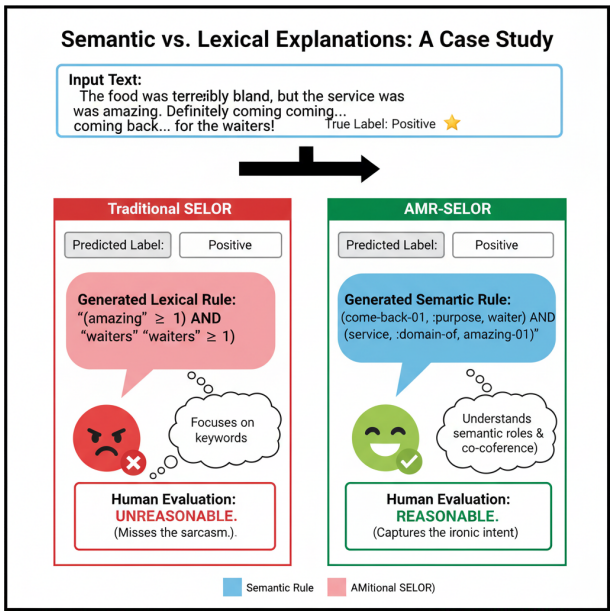

Fig. 3　图3-2　原子对比图

在SELOR中，一个原子是一个作用于输入特征的布尔函数，例如$o_i(x) = (\text{词频}(\text{”}tasty\text{”}) \geq 1)$。它回答的问题是“某个表层特征是否存在？”。而在AMR-SELOR中，一个语义原子(即语义三元组)是一个作用于输入文本语义图的布尔函数，例如$t_j(x) = ((staff, :ARG0\text{-of}, praise\text{-}01) \in \text{Triples}(\text{Parse}(x)))$。它回答的问题是“某个深层语义关系是否存在？”。

这种转变意义重大。它使得模型的解释语言从一种机器易于处理但人类难以直观理解的“相关性语言”，转变为一种更接近人类认知模型的“因果性语言”。

3.3　假设优势

本文预测，这种从词法到语义的根本性转变将为模型带来三个关键优势：

3.3.1　卓越的人类精度

SELOR框架的核心目标是提升“人类精度”。我们认为，AMR-SELOR将在此指标上取得质的飞跃。人类在理解和解释语言时，依赖的是对语义角色和事件结构的认知。例如，当被问及为何一条评论是负面时，一个令人信服的理由是“因为评论者抱怨价格过高”，而不是“因为评论中出现了‘price’和‘high’这两个词”。AMR-SELOR生成的解释，如’(price, :manner, rapacious-01) ⇒ 负面情绪’，直接对应了前一种人类认知模式。相比之下，标准SELOR可能会生成的解释，如’(”vegas” ≥ 1) ⇒ 负面情绪’，则仅仅反映了数据中的统计噪声。通过将解释的语言与人类的思维语言对齐，AMR-SELOR有望生成真正符合人类直觉、具有说服力的理由，从而大幅提升人类精度。

3.3.2　增强的复述鲁棒性

这是AMR-SELOR带来的一个独特且关键的优势。AMR的核心设计哲学之一就是对句法结构的不变性，即语义相同的句子，无论其句法结构如何(如主动语态vs. 被动语态、名词化结构vs. 动词结构)，都应被解析为同一个AMR图。这一特性天然地赋予了AMR-SELOR对文本复述(paraphrasing)的鲁棒性。

例如，考虑以下两个语义等价的句子：

- x_1: ”The staff was amazing and impressed us.”
- x_2: ”We were impressed by the amazing staff.”

标准SELOR可能会为这两个句子生成不同的解释规则。对于x_1，规则可能是’(”staff” ≥ 1) AND (”amazing” ≥ 1)’；对于x_2，规则可能是’(”impressed” ≥ 1) AND (”staff” ≥ 1)’。尽管都是合理的，但解释的不一致性可能会让用户感到困惑。

而AMR-SELOR则有望为这两个句子生成完全相同的解释。因为它们的AMR图都会包含类似’(staff, :domain-of, amazing-01)’和’(staff, :ARG0-of, impress-01)’这样的核心语义三元组。因此，生成的规则可能是’(staff, :domain-of, amazing-01)⇒正面情绪’。这种在语义层面上的稳定性，是纯词法模型难以企及的，它代表了一种更高级、更本质的鲁棒性。

3.3.3 更强的表达能力

AMR的表示体系能够精细地刻画多种复杂的语言现象，这极大地扩展了AMR-SELOR规则的表达能力2。

- 否定(Negation)：AMR使用'polarity -' 来表示否定。这使得模型可以学习到如'(service, :polarity, -) ⇒ 负面情绪'这样的规则，直接捕捉到"服务不好"的核心语义，而不是依赖于"not"、"bad"等否定词的组合。
- 情态(Modality)：AMR可以表示情态动词，允许模型区分事实与可能性，例如，区分"The service is good"和"The service could be good"。
- 共指(Co-reference)：AMR通过变量复用显式地处理共指关系。这使得模型能够构建跨越多个子句的复杂推理链。例如，对于"The waiter was rude, because he ignored us"，模型可以生成规则'(waiter, :instance, he) AND (he, :ARG0-of, ignore-01) ⇒ 负面情绪'，准确地将"粗鲁"的行为归因于"服务员"。

综上所述，AMR-SELOR框架不仅是对SELOR的一个简单升级，更是一次深刻的范式革新。通过将坚实的语义理论基础注入到忠实的自解释架构中，它有望生成更精确、更鲁棒、更具洞察力的解释，从而在通往真正可信赖AI的道路上迈出坚实的一步。

4 关键实现技术

Working...

5 验证

Waiting

6 结论

本报告深入探讨并系统性地提出了一个名为AMR-SELOR的新型自解释AI框架。该框架的核心创新在于，通过将抽象语义表示(AMR)的深层语义分析能力与SELOR框架的忠实逻辑推理机制进行深度融合，旨在解决当前自解释模型普遍存在的"语义鸿沟"问题。我们详细阐述了AMR-SELOR的理论基础、系统架构、关键实现技术以及一套全面的验证方案，旨在推动AI可解释性从基于表层特征的统计归纳，向基于深层语义的认知推理迈进。

框架回顾与潜在影响：AMR-SELOR用从AMR图中提取的语义三元组替换了SELOR原有的词汇原子，将解释的基本单元从"词"提升到了"义"。这一根本性的转变有望带来多重收益：通过生成与人类认知模型更为一致的、基于语义角色的解释，显著提升"人类精度"；利用AMR对句法结构的抽象能力，获得对文本复述的强大鲁棒性；借助AMR丰富的表示体系，增强解释规则的表达能力，以捕捉否定、情态等复杂语言现象。如果得到成功验证，AMR-SELOR将为构建更值得信赖、更易于人类协作的AI系统提供一个强有力的范例，证明了将符号化的语义知识原生集成到深度学习模型决策回路中的巨大潜力。

局限性与挑战：与任何前沿研究一样，AMR-SELOR的实现也面临着诸多挑战与固有的局限性，这与SELOR原作者所秉持的严谨学术态度一致。

- 对解析器质量的依赖：整个框架的性能上限受制于上游AMR解析器的准确性。解析错误会直接引入错误的语义原子，对后续的规则生成和评估造成干扰。尽管SPRING等解析器已达到较高水平，但在特定领域或面对复杂、模糊的语言时，其错误仍不可避免。
- 计算复杂性：引入AMR解析步骤无疑会增加模型的推理时间。更重要的是，为后件估计器生成预训练语料库需要对整个数据集进行离线解析和大规模规则采样，这是一个计算密集型的过程，对计算资源提出了更高的要求。
- 抽象概念的处理：AMR图中包含一些非词汇化的抽象概念(如amr-unknown)，以及复杂的图结构。如何让前件生成器和后件估计器有效地学习和利用这些高度抽象的符号信息，是一个开放的研究问题。

未来研究方向：AMR-SELOR框架为未来的研究开辟了广阔的空间。

- 迈向一阶逻辑：SELOR的作者指出，其模型局限于命题逻辑，而向一阶逻辑的演进是未来的重要方向。AMR的结构化特性为实现这一目标提供了天然的跳板。未来可以研究如何从AMR三元组中学习带有变量和量词的规则(例如，FORALL(x) such that (x, :instance, positive-word),...)，从而实现更具泛化能力的抽象推理。
- 跨句子与文档级推理：当前的AMR解析和AMR-SELOR框架主要针对单个句子。将该框架扩展到处理文档级的AMR图，有望让模型能够理解和解释段落级的语义连贯、论点发展和篇章关系，从而为文本摘要、对话系统等更复杂的任务提供高质量的解释。

- 端到端联合学习：尽管本报告为保证可行性提出了分阶段的训练流程，但探索AMR解析器与SELOR模块的端到端联合训练是一个富有吸引力的长期目标。通过联合优化，AMR解析器或许能学会在保证语义准确性的同时，生成更“有利于解释”的图结构，从而实现语义理解与逻辑推理之间更深层次的协同。

总之，AMR-SELOR代表了在追求真正可理解AI道路上的一次有原则的、前瞻性的探索。它不仅是一个具体的模型提案，更是一种思想的倡导：即真正的可解释性必须植根于对“意义”的深刻理解。

致 谢 感谢idea的提出者孙家兴师兄，感谢辛苦上课的任课教师

参 考 文 献

[1] Melis D A, Jaakkola T. (2018). Towards Robust Interpretability with Self-Explaining Neural Networks[C]//Advances in Neural Information Processing Systems, 31. Curran Associates, 2018: 7775–7784.

[2] Ribeiro M T, Singh S, Guestrin C. (2016). ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1135–1144.

[3] Lundberg S M, Lee S I. (2017). A Unified Approach to Interpreting Model Predictions[C]//Advances in Neural Information Processing Systems, 30. Curran Associates, 2017: 4765–4774. SHAP

[4] Lee S, Yi X, Wang X, Xie X, Han S, Cha M. (2022). Self-explaining Deep Models with Logic Rule Reasoning[C]//Advances in Neural Information Processing Systems, 35. Curran Associates, 2022: 30161–30174.

[5] Ribeiro M T, Singh S, Guestrin C. (2018). Anchors: High-Precision Model-Agnostic Explanations[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). AAAI Press, 2018.

[6] Banarescu L, Bonial C, Cai S, Georgescu M, Griffitt K, Hermjakob U, Knight K, Koehn P, Palmer M, Schneider N. (2013). Abstract Meaning Representation for Sembanking[C]//Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. Sofia: ACL, 2013: 178–186.

[7] Goodman M W. (2020). Penman: An Open-Source Library and Tool for AMR Graphs[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Online: ACL, 2020: 312–319.

[8] Blloshmi R, Bevilacqua M, Fabiano E, Caruso V, Navigli R. (2021). SPRING Goes Online: End-to-End AMR Parsing and Generation[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Punta Cana: ACL, 2021: 134–142.

[9] Minsky M L. (2022). Neuro-Symbolic AI: The 3rd Wave[OL]. arXiv:2208.13678.

[10] The ANSR Program. (n.d.). DARPA[EB/OL]. Available: https://www.darpa.mil/research/programs/assured-neuro-symbolic-learning-and-reasoning.

[11] Li Z, Gildea D. (2024). A Hybrid Neuro-Symbolic Pipeline for Natural Language Understanding[J]. Applied Sciences, 16(7): 529.

[12] Jayaweera C, Youm S, Dorr B J. (2024). AMREx: AMR for Explainable Fact Verification[C]//Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER). ACL, 2024: 234–244.