

AI翻译中的信息结构处理：问题与展望

邱林枫

四川大学外国语学院, 四川 成都

收稿日期: 2026年9月2日; 录用日期: 2026年10月2日; 发布日期: 2026年10月13日

摘要

信息结构是语言组织信息、突出交际重点并引导读者理解的重要机制, 主要涉及话题-述题、已知信息-新信息、焦点-背景等关系。在人工智能翻译中, 译文质量不仅取决于词汇和句法层面的对应, 还取决于原文信息结构能否得到适当重构。已有篇章级神经机器翻译通过分层注意力和上下文建模改善语篇处理, 大语言模型翻译研究也进一步探索了长上下文条件下的篇章翻译。因此, 当前问题不宜概括为模型普遍缺少跨句处理能力, 而应聚焦于上下文内容能否稳定转化为话题延续、信息状态识别和焦点功能保持。本文在梳理相关理论与技术研究的基础上, 分析语序调整、焦点识别和跨语言标记转换中的潜在偏差, 提出面向信息结构的改进框架: 将信息状态、焦点类型和共指关系形式化为分层标注与可计算特征, 通过上下文编码、多任务学习和对比损失增强功能保持, 并结合语篇专项测试与人工后编辑进行评价。上述方案旨在提供可检验的研究路径, 其有效性仍需通过跨语言、跨体裁实验验证。

关键词

AI翻译, 信息结构, 神经机器翻译, 大语言模型, 语篇连贯

Information Structure Processing in AI Translation: Problems and Prospects

Linfeng Qiu

College of Foreign Languages and Cultures, Sichuan University, Chengdu Sichuan

Received: September 2, 2026; accepted: October 2, 2026; published: October 13, 2026

Abstract

Information structure organizes discourse, highlights communicative emphasis, and guides interpretation through relations such as topic-comment, given-new, and focus-background. In AI translation, translation quality depends not only on lexical and syntactic correspondence but also on the appropriate reconstruction of these relations. Existing document-level neural machine translation research

has improved discourse processing through hierarchical attention and contextual modeling, while research on large language models has further explored document translation with extended context. The central issue is therefore not a general absence of cross-sentence processing, but whether contextual information can be reliably used to preserve topic continuity, information status, and focus function. Drawing on linguistic theories and relevant technical studies, this paper examines potential shifts arising from word-order adjustment, focus identification, and cross-linguistic differences in information-structural marking. It proposes an improved framework oriented towards information structure that formalizes information status, focus types, and coreference relations as layered annotations and computational features, integrates them through contextual encoding and multi-task learning, and introduces a contrastive objective for functional preservation. Discourse-specific testing and human post-editing are also incorporated into the evaluation framework. These proposals constitute testable research directions rather than established empirical findings, and their effectiveness requires validation across languages and genres.

Keywords

AI Translation, Information Structure, Neural Machine Translation, Large Language Models, Discourse Coherence

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

1.1. 研究背景

神经机器翻译(Neural Machine Translation, NMT)依靠编码器-解码器结构和注意力机制完成源语与目标语之间的转换。Transformer 模型出现后, 机器翻译在语法连贯性、词汇选择和长距离依赖处理方面取得了明显进步。与传统的统计机器翻译相比, NMT 不再主要依靠人工设计的短语规则, 而是通过大规模平行语料学习语言之间的对应关系。这类模型能够生成较为流畅的译文, 也能处理复杂的句子结构。

流畅并不等于准确。NMT 在处理信息结构时仍然存在明显不足。信息结构主要关注话语中的已知信息(Given)与新信息(New)如何安排。说话者通常会把听者能够识别的内容作为交际起点, 再把需要强调或补充的内容放在更突出的部分。不同语言对信息结构的表达方式并不相同。英语常通过句末位置、重音和分裂句突出焦点。

翻译不能只转换词汇和句法。译者还需要重建原文的信息层级。如果这种层级在译文中发生变化, 读者接收到的重点就可能不同。英语句子 “It was John who solved the problem” 把 “John” 设为排他性焦点。若译成 “约翰解决了问题”, 译文虽然保留了基本事实, 却没有完整传达原句的强调意义。“解决问题的人正是约翰” 更接近原文的交际功能。

针对单句翻译难以充分利用语境的问题, 篇章级 NMT 已经形成了较为系统的研究路径。Miculicich 等(2018) [1]通过分层注意力网络整合词级和句级上下文; Voita 等(2019) [2]则在英语-俄语字幕翻译中发现, 上下文建模能够改善指示表达、省略和词汇衔接等现象。这些研究说明, 跨句信息并非现有模型普遍缺失的输入, 关键在于模型是否能够提取与当前翻译决策相关的语境线索。不过, 共指或词汇一致性的改善并不自动证明焦点、对比关系和信息层级得到了完整保留。本文因此将问题进一步限定为: 已

有上下文建模机制如何支持信息结构的识别与跨语言重构,以及这种支持在不同语言和体裁中是否稳定。对于特殊构式的数据覆盖情况,也应通过具体语料统计判断,而不宜预先认定所有平行语料均存在相同的稀缺问题。

1.2. 研究意义

信息结构是翻译实现功能对等的重要部分。译文如果只保留命题内容,却改变了信息重点,读者对原文的理解仍然可能出现偏差。在广告翻译中,产品名称、功能特点和品牌信息往往具有不同的显著程度。焦点处理不当会削弱宣传效果。在法律翻译中,责任主体和责任范围必须保持清楚。焦点位置的变化可能造成条款歧义。在文学翻译中,倒装、停顿和省略有时承担着制造悬念的作用。机器译文的过度规范化会破坏原文的叙事节奏。

从理论角度看,信息结构研究能够连接功能语言学、翻译学和计算语言学。信息状态与交际参与者的认知状态有关。主题和焦点并不是固定的词汇属性,它们需要结合具体语境判断。相关理论为 NMT 处理语用信息提供了分析基础。

从实践角度看,研究信息结构可以帮助翻译系统减少机械直译。系统需要识别源语中的主题、焦点、对比关系和话题延续,再根据目标语的表达习惯重新安排句子。这个过程不是简单复制源语形式。它要求系统判断哪些信息应该前置,哪些信息应该后置,哪些内容需要显化,哪些内容则应该保留省略。

以条件似然最大化为主要目标的 NMT 能够从数据中隐式学习部分信息结构规律,但词元预测目标并不直接区分“命题内容基本正确”与“信息组织功能保持充分”这两种情况。即使模型已经接入篇章上下文,译文也可能在基本事实不变的情况下削弱原文的强调、对比或悬置效果。本文的研究意义在于,将这一差异转化为可操作的建模与评价问题:在输入端提供信息状态和焦点范围等结构化线索,在训练端增加信息结构辅助任务与功能保持约束,在评价端分别检验话题延续、焦点对应和指称衔接。需要强调的是,跨语言保持的是交际功能,而非源语词序或形式标签的机械复制。

2. 理论基础

2.1. 信息结构的基本概念

信息结构是说话者根据听者的知识状态组织话语的一种方式。Given 信息表示听者已经知道,或者可以从上下文推断出的内容。New 信息表示说话者希望听者注意的新内容。Given 信息经常使用代词、省略形式或定指名词短语表达。New 信息则常常位于句末,或者通过重音、特殊句式和焦点副词得到突出。

Given 与 New 不是完全固定的类别。同一个成分在不同语境中可能具有不同的信息状态。某个名词第一次出现时通常属于 New 信息。前文已经提到该名词后,它就可能转化为 Given 信息。信息结构因此具有动态性。机器翻译如果只根据单句判断信息状态,就很难处理跨句回指和话题延续。

不同语言使用不同手段表达信息结构。英语具有较强的主语和句法结构限制。英语常把自然焦点放在句末,也会利用重音和分裂句凸显某个成分。例如,“It was the algorithm that improved the results”通过“It was...that”结构强调“the algorithm”。汉语的语序更加灵活。汉语可以通过话题前置、句末位置、语气副词和“是……的”等结构表达焦点。

相较于直接提高所有新信息的注意力权重,更可检验的做法是将信息结构识别和功能保持纳入联合训练目标。新信息不必然等于焦点,已知成分也可以承担对比焦点,因此不宜设置统一的显著性增益。本文提出如下待验证目标:

$$L = L_M T + \lambda_I S L_I S + \lambda_F$$

其中, $L_M T$ 为译文生成的负对数似然损失; $L_I S$ 为信息状态分类、话题与焦点范围识别等辅助任务损失的加权和; L_F 为焦点功能保持的对比损失。对给定源句 x 和语境 c , 构造功能保持的译文 y^+ 与命题内容尽量相近、但焦点范围或强调关系发生偏移的译文 y^- , 可采用:

$$L_F = \max 0, m - s_\theta(y^-)$$

其中, $m > 0$ 为间隔参数, s_θ 可取长度归一化后的条件对数概率。正负译文需经人工核验, 尽量控制流畅度、长度和词汇差异, 避免模型仅凭表层标记区分样本; 损失权重及间隔参数则在开发集上确定。该目标旨在提高模型对功能差异的敏感度, 保证焦点关系判定得到正确识别和生成。

2.2. 信息结构的语用作用

信息结构能够帮助读者建立理解路径。Given 信息为话语提供认知支点。New 信息在这个支点上继续推进内容。若译文把 New 信息提前, 或者把原本需要强调的成分处理成普通信息, 读者的注意方向就会发生变化。

焦点还可以表达对比、排他和说话者态度。“才”表示某一事件发生得晚于预期。“就”可以表示事件发生得早或条件较弱。“正是”能够加强确认语气。英语中的“only”“even”和“not until”也具有类似的焦点作用。这些词语不能只按照字面对应处理。译者需要判断它们在句中的信息功能, 再寻找目标语中的合适表达。

信息结构也会影响篇章衔接。汉语常通过话题链保持语篇连续。一个话题出现后, 后面的分句可以省略主语。英语通常需要显性主语和完整谓语。翻译系统如果没有识别前后句之间的回指关系, 就可能漏译主语, 或者加入错误主语。两种语言的差异要求 NMT 进行跨句分析。

2.3. NMT 的工作方式及其限制

NMT 首先将源语转换为连续向量表示。解码器再根据这些表示逐步生成目标语。Transformer 利用自注意力机制计算不同词语之间的联系。位置编码则帮助模型识别词语在句中的顺序。这种结构提高了模型处理长距离依赖的能力。

在分析模型限制时, 需要区分信息结构的隐式学习、显式监督和生成控制。Transformer 的上下文化表示可以编码部分话题、共指和信息状态线索, 缺少专门标注并不意味着模型完全无法识别这些关系; 同样, 注意力权重也不能直接视为语言学意义上的焦点标记。篇章级 NMT 通过上下文拼接、多编码器或分层注意力扩展信息来源, 大语言模型则可以利用更长的上下文和任务指令。Wang 等(2023) [3] 对大语言模型的篇章级翻译与语篇建模能力进行了评估, Choudhary 等(2025) [4] 进一步比较了不同上下文策略对语篇感知翻译的影响。这些进展使研究重点由“模型是否接入上下文”转向“模型是否利用了决定信息结构的上下文”, 以及“这种利用是否体现为目标语中的功能保持”。本文关注的是后两个层面, 而非将句级模型的限制推广至全部 AI 翻译系统。

训练数据的影响不仅体现为句式出现频率, 还体现为语境是否完整、功能对立是否得到覆盖, 以及参考译文是否保留了相应的语用差异。即使某一焦点副词出现频繁, 缺少其不同作用域和使用语境的对照样本, 模型仍可能学习到过于单一的译法。相反, 零代词等现象在部分语言和体裁中并不罕见, 真正的困难可能是回指关系缺少标注, 或者句对抽取过程切断了先行词所在的上下文。因此, 应分别统计句式频率、完整语境覆盖率和功能对立样本数量, 再判断译文规范化倾向与哪些数据因素有关, 避免将信息结构偏差统一归因于低频表达。NMT 还受到文本模态的限制。语音中的重音、音高和停顿可以提示焦点位置。纯文本模型无法直接获得这些信息。影视翻译和口译场景中的信息结构因此更加复杂。模型需要结合音频、画面和上下文, 才能较准确地判断说话者的真正重点。

3. NMT 处理信息结构面临的主要问题

3.1. 语序僵化

语序既表达句法关系，也组织信息流。英语倒装句 “In the valley lived an old hermit” 把地点放在句首，把新出现的人物放在句末。地点为读者提供场景，隐士则成为叙事焦点。若系统译成 “一个老隐士住在山谷里”，译文虽然基本正确，却改变了原句的展开方式。“山谷里住着一位老隐士” 更能保留存现效果。

汉语话题链在英译中也会出现结构断裂。“这批设备调试完毕，随即投入试运行” 没有重复主语，因为前一句已经提供了话题。英语通常需要明确的主语和时态。若系统只进行句内转换，就可能生成缺少主语的译文，或者把陈述误处理为命令。

语序僵化的根源包括数据偏差和模型偏好。训练语料中的常规主谓结构会形成强烈的输出惯性。位置编码能够记录词序，却不能独立说明词序所承担的语用作用。模型可以知道某个成分位于句首，却未必知道该位置是在标记主题、场景还是对比焦点。

3.2. 焦点识别偏差

句法焦点是 NMT 容易忽略的部分。“It was the newly introduced policy that stabilized the market” 并不是普通的 “新政策稳定了市场”。原句强调，稳定市场的正是这项政策，而不是其他因素。若译文只使用普通主谓句，原文的排他性就会减弱。

汉语焦点副词也常被机器忽略。“他昨晚十一点才交卷” 中的 “才” 表示时间晚于预期。“He submitted the paper at eleven last night” 只传达了时间和动作，没有表达迟延意义。“He did not submit the paper until eleven last night” 才保留了原句的焦点关系。

隐性焦点的识别不仅需要上下文，还需要判断当前话语正在回答什么问题，即话语中的当前问题 (Question under Discussion, QUD)。例如，同一句 “John solved the problem” 在回答 “Who solved the problem?” 时以 “John” 为焦点，在回答 “What did John do?” 时则可能以谓语部分为焦点。即使篇章级模型能够读取前文，也仍需检验其是否根据问题变化调整译文的信息突出方式。对于文学省略，还应区分为满足目标语句法要求而进行的必要显化，以及改变叙事留白的过度补全。前者可能有助于理解，后者则可能损害原文的交际或审美功能，二者不宜一概视为错误。

3.3. 跨语言标记差异

汉语和英语之间也存在明显差异。汉语可以通过零代词延续话题。英语通常需要显性主语。英语还更加依赖冠词来区分定指和非定指。汉语则没有完全对应的冠词系统。NMT 如果只按照词汇和句法寻找对应关系，就很难在目标语中补充必要的信息状态标记。

这些问题会产生不同的交际后果。广告译文可能失去重点。文学译文可能失去节奏。法律译文可能产生责任主体不清的问题。技术文本也可能因主语和话题关系不明而降低可读性。错误不一定表现为明显的语法错误。很多时候，译文表面流畅，信息结构却已经发生改变。

4. 信息结构处理的改进路径

4.1. 规则与深度学习结合

在已有篇章级 NMT 的基础上，可以构建 “当前句编码 - 上下文编码 - 信息结构预测 - 条件生成” 的多任务框架，而非仅在单句模型外附加规则。具体而言，当前句编码器处理待译句，上下文编码器处

理相关前文，信息结构预测模块识别实体的信息状态、话题范围和焦点范围，再将预测结果以标签嵌入或软概率分布的形式与文本表示融合。融合过程可采用门控机制，由模型学习不同语境下结构化线索的使用程度，并结合预测置信度降低错误标签的影响。对于超出局部窗口的话题链，可维护记录实体身份、最近提及位置及跨句关联的记忆表，检索相关提及后再进行编码。该方案是在既有上下文建模基础上增加显式语用监督，其有效性需要与仅使用相同上下文的模型进行比较。

当系统识别到“*It is...that*”结构时，它可以把该结构标记为焦点构式。汉译时，系统可以优先考虑“……的是……”或“正是……”等表达。当系统识别到汉语“才”时，它需要判断“才”是否表示时间迟延、条件限制或数量不足，再选择“*not until*”“*only then*”等目标语结构。规则不需要替代神经模型。规则的作用是限制明显偏离语用。

4.2. 语用标注与数据增强

信息结构标注宜采用相互关联但不混同的分层体系。第一层在实体提及层面标记“篇章已提及”“可由语境推知”“篇章新引入”等信息状态，并另外记录听者可识别性，避免将篇章新信息简单等同于听者未知信息。第二层标记话题与述题、焦点与背景的跨度，这些标签不与信息状态互斥，例如已知实体仍可构成对比焦点。第三层标记焦点副词的作用域、共指链以及零代词与先行词的关系。在计算实现上，可将标签映射为可学习嵌入，将距上次提及的句数、窗口内提及次数、语法角色和焦点标记等作为辅助特征，再与上下文化表示融合。对于无法仅凭文本确定的焦点，应设置“不确定”标签或保留多个候选，避免把标注者推断强制转化为唯一训练答案。

焦点标注还应记录其判定依据，区分回答当前问题的信息焦点，以及涉及替代项比较的对比焦点；对于纠正或排他解释，则进一步注明是否存在明确的语境或形式证据，避免把所有对比焦点都解释为排他关系。在有显式问句的语料中，可以问答关系辅助确定焦点跨度；在没有问句的语料中，可由标注者依据上下文提出候选当前问题，并记录判定置信度。跨语言对齐应以实体、事件和焦点作用域为锚点，允许一个源语焦点在目标语中通过不同词序、词汇或句式实现，而不强求逐词标签相同。标注流程可采用自动预标注、双人独立复核和争议仲裁，并分别报告信息状态的一致性、话题、焦点跨度的一致性。

数据增强应区分“功能保持的形式改写”与“功能改变的对照构造”。前者是在固定语境下，以不同目标语表达保留同一焦点和对比关系，可作为多个有效参考；后者则有意改变焦点范围、焦点标记或必要的指称形式，用于训练和测试模型能否识别功能偏移。普通陈述句、倒装句与分裂句即使命题内容接近，也不能默认在所有语境下等价。因此，增强样本应同时保留源句、上下文、信息结构标注及相应译文，并由人工核验语境适切性。划分数据时，应先按文档划分训练集、开发集和测试集，再在各集合内部进行增强，防止同一原始样本的改写版本跨集合泄漏。

跨句语料的建设应与上下文选择策略同步设计。可先以当前句及前两句作为局部上下文基线，再比较更长窗口、整段输入及实体关联检索等方案；窗口大小仅作为实验变量，不应预先认定输入越长越好。Choudhary 等(2025) [4]对不同上下文策略的比较，为检验上下文组织方式对语篇翻译的影响提供了参照。针对话题链和零代词，可优先检索潜在先行词所在句，并记录相关信息与当前句之间的距离。评估时还应加入无关上下文和上下文打乱条件，观察模型是否真正利用了有效关系。若使用目标语历史，测试阶段应采用系统自身已经生成的译文，而非人工参考译文，并单独分析历史错误的累积影响。

4.3. 人机协同与后编辑

人机协同可以围绕“模型预标注 - 译者确认 - 受控生成 - 差异检查”展开。系统首先显示候选话题、焦点范围及回指关系；对于存在多种合理分析或置信度较低的片段，由译者确认或修改，再据此生成候

选译文。界面中的“高亮”应区分语言学焦点与任务关注项：法律责任主体或广告产品名称即使需要重点检查，也未必就是当前语境中的焦点，不能一律要求模型增加强调表达。生成后，系统可将焦点范围变化、指称对象变化和显化处理等标为复核项。交互效果则可通过信息结构错误率、人工修改次数和后编辑耗时评估，而非仅以译者是否偏好某一版本判断。

质量评价应同时覆盖整体翻译质量、信息结构识别和实际生成功能。整体质量可采用 BLEU、COMET 等指标，并结合人工检查遗漏、增译和命题内容错误；识别层面分别报告信息状态分类的宏平均 F1、话题与焦点跨度 F1，以及零代词先行词识别准确率；生成层面则由双语评审者在完整语境下判断焦点所对应的实体或事件、对比关系和指称链是否得到保持。可借鉴 Voita 等(2019) [2] 的语篇现象对比测试思路，构建正确译文与信息结构偏移译文的最小对照组，报告模型偏好正确候选的比例，但不能以候选评分代替对自由生成译文的评价。实验至少应包括句级基线、仅加入上下文的篇章级基线，以及增加信息结构监督的模型，并进一步移除结构化特征、辅助任务或对比损失开展消融分析。大语言模型则应在相同上下文和提示预算下进行组内比较。最终按语言方向、体裁和焦点显隐类型报告结果，同时记录计算开销，以判断改进是否稳定且具有实际应用价值。

5. 结论

NMT 提高了机器翻译的流畅度，却没有彻底解决信息结构重构问题。语序僵化、焦点识别偏差和跨语言标记差异仍然会影响译文的交际效果。系统可能准确转换词汇，却改变 Given/New 关系。系统也可能生成语法正确的句子，却削弱原文的强调、对比和悬置意义。

信息结构处理需要语言学知识、语篇语境和模型计算能力共同参与。规则模块可以识别显性结构。语用标注可以帮助模型学习信息状态。多模态输入能够补充重音和视觉焦点。译者后编辑则可以根据文本目的完成最后判断。任何单一方法都难以覆盖所有场景。

未来研究需要扩大多语种语用标注语料，特别是增加低资源语言和跨句语料。研究者也需要改进信息结构的自动评估方法。BLEU 等指标不能充分反映焦点丢失和话题断裂。多模态翻译还需要进一步结合语音、图像和文本信息。NMT 只有从词语对应走向交际功能重建，才能更接近真实翻译活动的要求。

参考文献

- [1] Miculicich, L., Ram, D., Pappas, N. and Henderson, J. (2018) Document-Level Neural Machine Translation with Hierarchical Attention Networks. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, 31 October-4 November 2018, 2947-2954. <https://doi.org/10.18653/v1/d18-1325>
- [2] Voita, E., Sennrich, R. and Titov, I. (2019) When a Good Translation Is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, 28 July-2 August 2023, 1198-1212. <https://doi.org/10.18653/v1/p19-1116>
- [3] Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., et al. (2023) Document-Level Machine Translation with Large Language Models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 16646-16661. <https://doi.org/10.18653/v1/2023.emnlp-main.1036>
- [4] Choudhary, R., Hida, R., Hamada, M., Futami, H. and Sekiya, T. (2025) Exploring Context Strategies in LLMs for Discourse-Aware Machine Translation. *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, November 2025, 24382-24391. <https://doi.org/10.18653/v1/2025.findings-emnlp.1324>