大语言模型捕捉语境信息

汇报人：李子明

武汉大学金融系

2025.11.02

语言学体系中的信息层次

- 在语言学体系中，语言中的信息通常被划分为四个层次：
  - 词汇层：语言的表层形式，包括词汇使用、形态变化和句法结构
  - 语义层：语言单位本身具有的意义，词语和句子所表示的内容、主题和逻辑
  - 语境层：语言的上下文，围绕特定语言单位并为其解释提供信息的解读框架
  - 语用层：语言在语境中的使用意图、交际目的与受众效果

传统文本分析对语境层信息关注较少

- 传统文本分析主要聚焦于词汇层和语义层信息
  - 词汇层信息：
    - 词汇的使用（积极词、消极词、模糊词、专业术语），衡量语气倾向或专业性
    - 词频与分布（高频词、罕见词、停用词比例），反映语言风格或复杂度
    - 句法结构（平均句长、从句比例、被动语态使用率），衡量文本可读性与语言复杂度
    - 词汇多样性（不同词汇的丰富度），衡量语言表达的精细化程度
    - 可读性指标（Fog Index、Flesch-Kincaid、LIX 指数），衡量普通读者理解难度
    - 语言风格（第一人称/第三人称使用、模糊表达比例），分析叙述语气与主体性表达
  - 语义层信息：
    - 内容主题（战略转型、创新、资本开支、竞争格局等）
    - 情感态度（乐观悲观、风险提示、不确定性）
    - 概念关系（词语和主题之间的语义联系）
    - 相似度（语义上的距离）
    - 潜在意义（隐含的意义模式）
  - 过往文献采用的词典匹配、词频统计、生成主题模型以及词向量等词项级的文本分析方法主要针对词汇层和语义层信息，忽视了语境层信息。

大语言模型捕捉语境信息

- 词项级分析方法的局限性
  - 语义本质上是多维的，词项级方法将文本压缩到单一维度，难以涵盖全部信息。
  - 固定形式的指标构建表示整篇文本，无法挖掘文本信息的深层非线性关系和交互作用。
- 大语言模型的优势
  - 通过大规模预训练，能够很好地编码文本序列中的综合语义。
  - Transformer 架构能够识别上下文并保留大量语境信息。
- 文本语境：The News in Earnings Announcement Disclosures: Capturing Word Context Using LLM Methods (MS, 2025)
- 数字语境：Context-Based Interpretation of Financial Information (JAR, 2024)