

Structure-aware translation of scientific pages

A. Rivera, M. Chen, S. Patel

摘要

科学论文依赖于页面结构。公式、图注和参考文献必须与其周围的文本保持关联 [1]。我们研究了一种面向翻译后 PDF 文档的版面感知方法。

1. 介绍

当阅读顺序丢失时，双栏文章很难翻译。在此示例中，翻译后的页面应保持标题、引用 [1] 和相邻插图的位置不变。

忽略页面几何结构的方法可能会将图注与其插图分离，或将公式移动到错误的章节中。

我们的方法将每个文本块视为一个单元，在替换原始语言的同时保留其位置。

参考文献

[1] Rivera 等人. 文档布局研究, 2025.

2. 方法

对于每个块 b_i ，翻译其文本，同时保持页面位置与阅读顺序之间的关系。

$$T(b_i) \rightarrow b'_i$$

转换后的块被放置在同一列中。图2比较了五种文档结构。

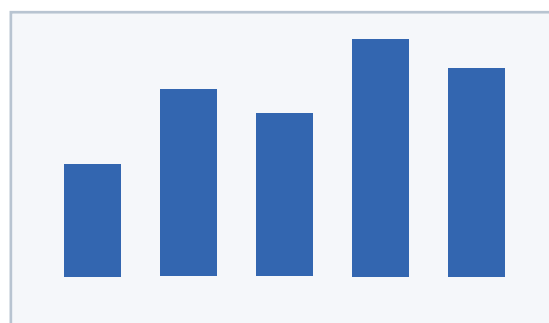


图2. 基于文档结构的识别。

翻译后，图题仍保持在插图下方。

3. 结论

可读的翻译取决于保留视觉语境。