

英语科技论文摘要语步结构语料库构建研究

李洪政* 王若锦* 冯冲[✉] 刘芳*

*北京理工大学外国语学院, 语言工程与认知计算工信部重点实验室/ 北京市房山区良乡 东路9号

†北京理工大学计算机学院/ 北京市海淀区中关村南大街5号

{lihongzheng,wangruojin,fengchong,liufang}@bit.edu.cn

摘要

语步结构是学术论文中的文本语篇单位, 在语步分析、论文写作等方面具有重要价值。尽管关于学术论文的语步研究非常丰富, 但语步标注数据资源仍然相对较少。本研究开发构建了一个英语科技论文摘要语步结构标注语料库, 目前已标注近3.4万个语步结构, 涵盖了自然语言处理、计算机视觉、通信工程、机械工程等学科领域, 同时进行了标注数据统计和分析。语料库构建的第一阶段依靠人工标注形成高质量语料, 在第二阶段也是主要阶段, 采用了基于BERT的自动识别与标注模型, 在保证标注质量的同时能够提升标注速度, 扩大标注规模。本研究基于构建的语料库开展了不同学科领域摘要语步结构识别实验, 对比了我们的模型与ChatGPT和Claude3等大语言模型的识别效果。结果显示我们的模型在各类语步识别上的F1指标均优于大语言模型, 表明了模型的有效性。该语料库目前可公开获取使用, 能够为科技论文信息抽取、英语写作智能批改等自然语言处理相关任务和学术用途英语等外语教学与研究等提供必要的数据库资源, 同时也能有效推动外语教育数字化转型。

关键词: 语步结构; 语料库; 研究论文; 摘要

Research on Construction of Corpus for Move Structures in Abstracts of English Scientific Research Articles

Hongzheng Li* Ruojin Wang* Chong Feng[✉] Fang Liu*

*School of Foreign Languages, Key Laboratory of Language, Cognition and Computation
Ministry of Industry and Information Technology, Beijing Institute of Technology /

No.9, Liangxiang East Road, Fangshan District, Beijing

†School of Computer Science, Beijing Institute of Technology /

No.5, Zhongguancun South St., Haidian District, Beijing

{lihongzheng,wangruojin,fengchong,liufang}@bit.edu.cn

Abstract

Move structures are discourse units in research articles (RA) and is of great value in move analysis, essay writing, etc. Although there is abundant research on move structures in academic articles, there are still relatively few move annotation data resources. This research developed and constructed a specific corpus for annotating move structures in English RA abstracts. Currently, nearly 34,000 Move structures have been annotated, covering the fields of Natural Language Processing (NLP), Computer Vision (CV), Communication Engineering and Mechanical Engineering. We also presented annotation statistics and analysis. The first stage of corpus construction relies on manual annotation to form high-quality corpus data. In the second and main stage, an automatic recognition and annotation model based on BERT is adopted, which can improve the annotation speed and expand the annotation scale while ensuring the annotation quality. We conducted move structure recognition experiments

based on the constructed corpus, and compared the performance of our model with large language models(LLM) including ChatGPT and Claude3. The experimental results show that the F1 scores of move structure recognition achieved by our model outperformed those of LLM, indicating the effectiveness of the proposed model. This corpus is currently publicly available and can provide necessary data resources for NLP related tasks such as scientific paper information extraction and English writing intelligent assistance, it is also beneficial to foreign language teaching and research such as English for Academic Purposes, it can effectively promote the digital transformation of foreign language education.

Keywords: Move structure , Corpus , Research article , Abstract

1 引言

语步(Move)是指具有具体交际功能和目的的文本切分单位和语义片段, 用来表示特定意义, 代表着学术论文各章节的表述结构与序列(Swales, 1990; Swales, 2004)。语步能够有效表征论文语篇的宏观结构, 并为实现整个语篇体裁的总体交际目的服务, 揭示论文语篇的目的性、步骤性和规约性(杨延宁 and 邹航, 2023)。

科技学术论文特别是研究论文 (Research Article, RA) 是科技和学术交流的重要载体。在研究论文中, 摘要是必不可少的组成部分, 通过简洁、清晰而有说服力的段落传达整篇文章的核心思想和亮点。好的摘要应该包括组织良好和逻辑合理的语步结构, 分别论述论文的研究背景、研究目的、研究方法、研究结论等基本信息, 以显示文章的重要性和价值。

下面是一个NLP领域的英文研究论文摘要示例, 其中清楚包含了几种常见的语步结构。

例句1: **〈研究背景〉**Previous works on cross-lingual NER are mostly based on label projection with pairwise texts or direct model transfer. **〈研究差距〉**However, such methods either are not applicable if the labeled data in the source languages is unavailable, or do not leverage information contained in unlabeled data in the target language. **〈研究目的〉**In this paper, we propose a teacher-student learning method to address such limitations. **〈研究方法〉**NER models in the source languages are used as teachers to train a student model on unlabeled data in the target language. **〈研究方法〉**The proposed method works for both single-source and multi-source cross-lingual NER. **〈研究结论〉**Extensive experiments for 3 target languages on benchmark datasets well demonstrate that our method outperforms existing state-of-the-art methods for both single-source and multi-source cross-lingual NER.

语步的概念最初来自修辞学, 最早由美国语言学家J.Swales提出。语步研究在专门用途英语 (English for Specific Purposes, ESP) 和学术用途英语(English for Academic Purposes, EAP)等领域具有比较悠久的历史, 在外语科技论文写作、语步分析等也具有重要的应用价值(Hyland, 2008; Moreno and Swales, 2018; Swales, 2019)。自从语步被提出以来, 很多研究基于不同学科领域学术论文的不同部分如摘要、引言、方法、结论等内容针对语步、修辞结构和语言学特征等开展了比较详细的分析和讨论(Cotos et al., 2017; Lu et al., 2021; Alsharif, 2023)。近年来也有一些研究工作关注语步结构的检测与自动识别等(Alliheedi et al., 2019; 张鑫 et al., 2024; 王末 et al., 2020; 王东波 et al., 2018; 杜新玉 and 李宁, 2024; 丁良萍 et al., 2019)。语步结构识别可以使文献中蕴含的知识显式地体现出来, 提高知识利用效率, 是内容抽取、文本摘要等任务应用的重要基础性工作(黄红 et al., 2022)。但是之前的研究工作更多地只关注某一特定学科领域论文中特定部分的语步分析, 或某个特定语步的识别, 几乎没有跨学科对比研究和全面的语步结构研究; 另一方面, 目前面向学术写作的语步结构自动标注的研究相对较少, 同时也缺乏针对科技研究论文摘要部分中的语步标注资源建设。这种不足为跨语言语步分析与识别、科技论文信息抽取等自然语言处理任务和计算机辅助语言学习 (Computer-Assisted Language Learning, CALL) 等应用场景带来了许多挑战和障碍。

为了解决这些问题, 本研究构建了一个大规模、多领域的英语科技论文摘要语步标注语料库, 当前已累计标注近3.4万个语步实例, 涵盖了人工智能、通信工程和机械工程等学科领域。

©2024 中国计算语言学大会

根据《Creative Commons Attribution 4.0 International License》许可出版

该语料库已在Github上开源发布⁰。标注主要分为两个阶段：在前期初始阶段，通过人工标注形成一部分高质量的标注数据；在后期阶段也是目前标注的主要阶段，在高质量标注数据的基础上，我们训练了一个BERT自动标注模型，实现语步自动标注，然后人工进行干预，修正系统标注不准确或者错误的语步，在快速扩大标注数量和规模的同时也能够保证标注质量。随后开展了语步结构自动识别实验，对比了我们的系统模型与GPT-4和Claude 3等大语言模型在不同学科领域的语步识别效果，同时针对具体实例进行了比较详细的分析。实验结果显示，本研究的模型在不同类型语步结构识别的F1指标分数均明显高于大语言模型的分数，表明了模型的有效性以及语料库建设的价值和必要性。

2 相关研究

2.1 学术领域语料数据资源

本研究构建的语料库属于学术领域语料库。学术文本语料库包含学术期刊、会议论文集等发表的研究论文、摘要等学术写作材料。学术语料库的建设有利于推动科学进步，在很多应用场景中具有实用价值，受到国内外学术界和业界的广泛关注。近年来，学术领域数据集的数量显著增加，同时也支持开源获取。比较具有代表性的语料库包括：ACL OCL语料库(Rohatgi et al., 2023)、S2ORC(Lo et al., 2020)、ARC(Bird et al., 2008)和AAN(Radev et al., 2009)等基于ACL Anthology构建的语料数据资源。这些数据集通常包含全面的元数据和科学论文全文，能够为研究人员提供可访问且重要的数据资源。

2.2 语步分类及语步标注

语步研究历史久远。不同时期的很多学者都针对语步进行了分类研究。例如：Swales开创性地提出了CARS语步模型(Swales, 1990; Swales, 2004)和语步分类体系(Swales and Feak, 2009)，为后来的语步研究带来了深远影响，很多研究者在此基础上不断改进语步的理论研究，包括Teufel等人提出的Argumentative Zoning (AZ) scheme(Teufel et al., 1999)及其修正后的理论等(Teufel et al., 2009; Teufel, 2010)。

随着语步理论研究的发展，也出现了一些在这些理论指导下建设的语步结构数据资源。Alliheedi等人重点关注生物化学领域论文方法部分中的语义角色和修辞语步(Alliheedi et al., 2019)。有研究者构建了选自《应用语言学》期刊(Applied Linguistics)摘要的数据集(刘霞, 2016)；Viera等人(2020)通过构建摘要语料库，调查了在英语母语国家和非英语国家发表的研究论文摘要中的修辞。已有的语步结构语料库通常都是面向特定单一领域的，而且规模相对较小，目前仍然缺乏多领域大规模的摘要语步语料库。因此，有必要构建涵盖多个学科领域的论文摘要标注语料库。

3 语料库构建过程

3.1 理论基础

本研究构建语料库的理论基础主要借鉴了Hyland关于论文摘要语步的五分类理论(Hyland, 2000)。他认为，学术论文全文中的语步分类可能并不完全适合分析摘要部分的语步结构。经过对纯科学、应用科学、人文科学和社会科学等学科门类的论文摘要进行了集中细致的考察，他提出学术论文摘要语步五分类法：

研究介绍 (Introduction)：论述研究说明，包括研究重要性、关键术语概念、研究差距等

研究目的 (Purpose)：指出该研究的一般或者特别目的

研究方法 (Method)：指明该研究的具体方法，包括数据、处理过程等

研究产出 (Product)：讨论该研究的主要发现和结果

研究结论 (Conclusion)：引出该研究的结论，包括研究重要性、研究不足、对未来研究的启示等

Hyland的分类法对本研究具有重要的借鉴意义。但根据仔细的前期调研和小规模的前导标注试验(Pilot Study)，我们发现Hyland的分类无法完全适合特定领域的摘要语步，因此需要进一步的详细分类体系。例如下面例句的类似表达在很多包括NLP在内的人工智能领域论文中

⁰<https://github.com/ljk1228/RAAMove>

都非常常见，一般是开源论文研究的相关信息，通常可以认为是对整个研究社区的贡献和价值，因此无法标注出Hyland分类中的任何一种。

例句2: We release source code for our models and experiments at Github.

在Hyland分类的基础上，我们根据具体标注需求，对该分类方法进行了调整和补充，提出了表 1所示的八种基本的语步类型，并为每种语步设置了标注标签。

语步	含义	标签
背景 (Background)	陈述研究领域，并提供关于研究的相关信息。	BAC
差距 (Gap)	表明当前的研究差距，提出开展研究的积极理由。	GAP
目的 (Purpose)	表明研究目的、假设等，概述论文背后的意图。	PUR
方法 (Method)	提供研究的设计、程序、假设、方法、数据等信息。	MTD
结果 (Result)	说明研究的主要发现或结果或已完成的工作。	RST
结论 (Conclusion)	总结研究结果或将结果扩展到论文范围之外。	CLN
启示 (Implication)	得出尚未明确陈述的推论以及对后续研究的启发等。	IMP
贡献 (Contribution)	指出研究带来的理论和实践价值。	CTN

Table 1: 本研究设计的语步类型、含义及其标签

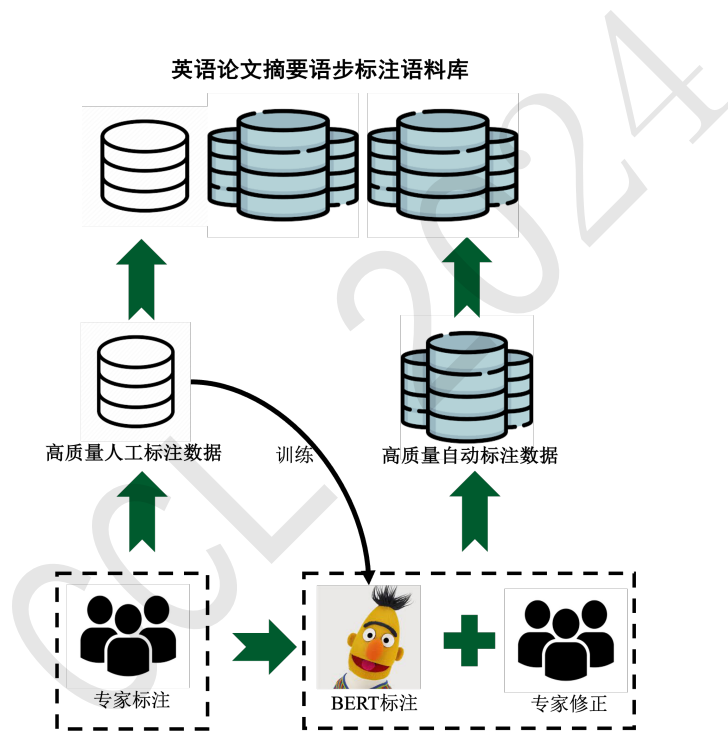


Figure 1: 语料库构建框架图

3.2 语料来源及处理

根据各专业领域的教授专家推荐，选择本领域的英文顶级期刊（JCR Q1）和国际知名顶级会议论文中的摘要作为标注语料，涵盖人工智能和工程学科大类，其中人工智能学科包括NLP和计算机视觉（CV）领域；工程大类包括通信工程（Communication Engineering, CE）和机械工程（Mechanical Engineering, ME）领域。

对于人工智能领域，我们选择了ACL大会主会论文和AAAI大会CV Track的会议论文。从ACL Anthology官网¹可以直接下载收录了所有论文的BibTex文件，从中筛选出ACL2020-ACL2022三年的长论文摘要作为标注数据；另外补充了AAAI 2022大会CV Track的会议论文。

¹<https://aclanthology.org/>

对于工程大类领域，选择了三本顶级期刊²。在Web of Science (WOS)检索平台中检索每本期刊的信息，然后将检索得到的论文题目、摘要等基本信息进行处理。为了方便标注语步结构，我们把两个学科领域的所有摘要文本段落根据主要的句末标点符号预处理为一句话一行的格式。同时在保留原文基本意义的前提下，对摘要文本进行必要的清洗，忽略可能会影响标注的特殊符号、引用格式等信息，以保证标注文本的质量。

3.3 语料库构建

如图1所示，语料库构建主要分为两个大的阶段：第一阶段是人工标注，第二阶段是系统自动标注+人工修正。我们采用开源标注工具Doccano³作为在线标注平台（图2）。

3.3.1 人工标注

为了保证语步标注的质量和准确性，我们的语料标注团队由四名外国语学院的专业教师和一名外语语言学专业的博士生组成，同时每周举行例会讨论标注过程中存在的各种疑难问题。在标注过程中，原则上以完整的句子为单位，标注者需要为每个句子选择最合适的语步标签。

在标注平台中，选中一个句子以后即可弹出设计好的标注标签，标注者可以选择一个标签完成标注。如果需要修改，可以点击标签，选择其他合适的标签。

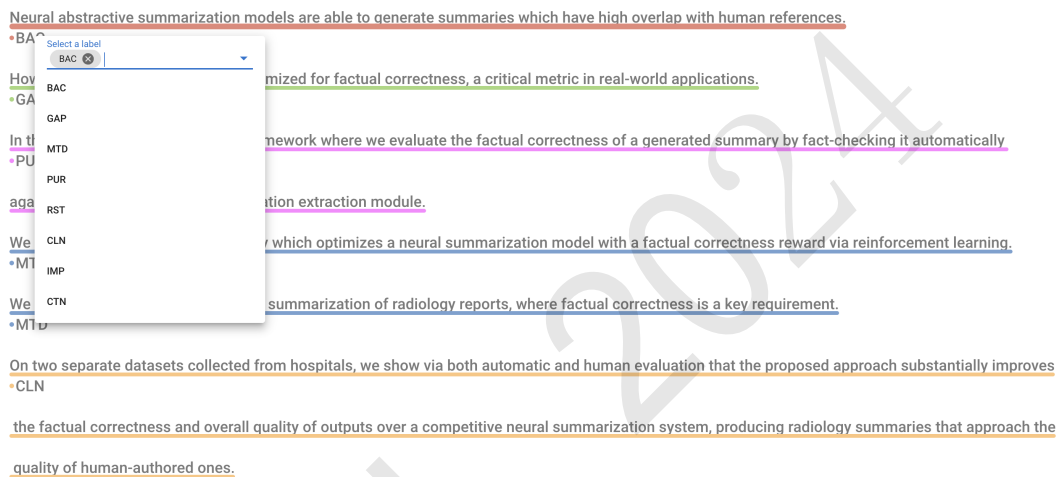


Figure 2: 标注平台界面

3.3.2 自动标注

人工标注虽然能够保证标注质量，但无法提升标注速度和语料库标注规模，因此需要自动标注。本研究采用了Lin等人(2023)提出的基于BERT(Devlin et al., 2019)的语步自动标注模型(图3)，将语步识别标注视为多标签识别和分类问题。

需要说明的是，在标注语步的时候，句中的不同单词对于预测句子语步类型具有不同的影响。像图3中的例句，句中有“results”“performance”等非常明显的关键词，它们对于快速识别为“RST”语步类型具有更高的贡献程度。考虑到这种普遍情况，该模型引入显著性注意力(saliency attention)，句子中的每个单词都被视为一个特征，并计算其对特定语步类型的贡献(显著性值)。模型从人工标注的高质量数据中学习句子的语义特征，每个句子首先被分配一个表达整体语义的语步标签，然后设计了词语显著性向量(word saliency embeddings)与BERT的其他三种向量(即token、segment与position embeddings)一起作为输入表示来捕捉词语对于语步的贡献情况，提升模型识别语步的能力，最终完成语步识别和标注。

自动标注结果上传到标注平台以后，标注团队开始检查标注结果，修正标注不准确或者错误的语步结构。标注模型根据人工反馈的数据，可以不断进行迭代优化，进一步提升标注效

²三本期刊分别是机械工程领域的Journal of Mechanical Design、International Journal of Heat and Mass Transfer、通信工程领域的IEEE Journal on Selected Areas in Communications

³<https://github.com/doccano/doccano>

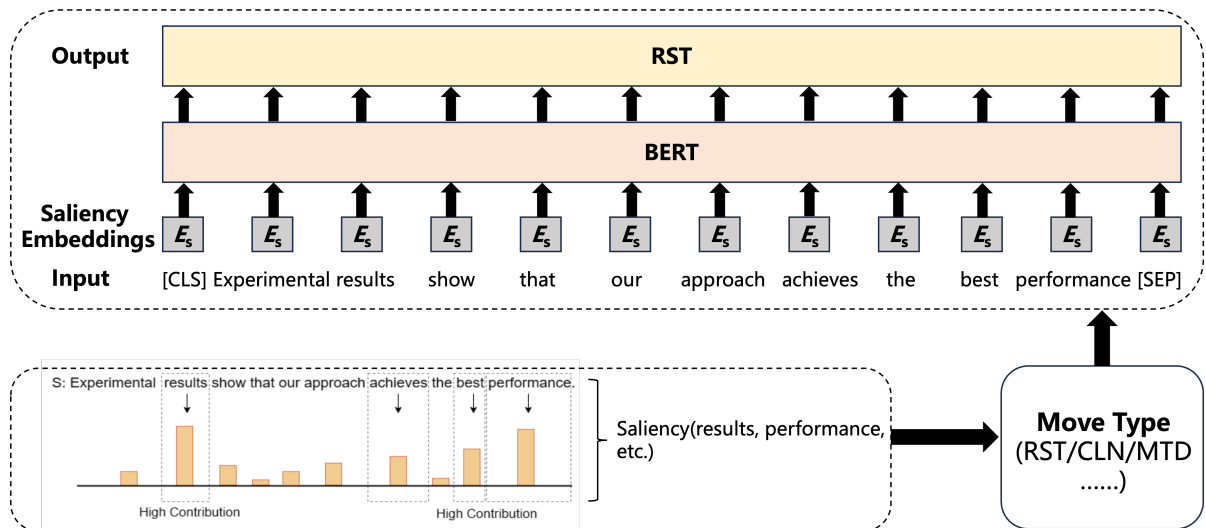


Figure 3: 语步自动识别模型架构

果。通过这种方式，可以快速提高标注速度和效率，在保证标注质量的同时扩大标注规模。语料标注完成以后，可以直接从标注平台中导出如下面实例所示的JSONL格式标注文件。

```
{ "id": 20, "data": "Words can have multiple senses. Compositional
distributional models of meaning have been argued to deal well with finer
shades of meaning variation known as polysemy, but are not so well equipped
to handle word senses that are etymologically unrelated, or homonymy.",
"label": [[0, 31, "BAC"], [32, 265, "GAP"]] }
```

4 语料库数据统计

本研究目前已标注2,670篇人工智能领域（NLP方向1,340篇，CV方向1,330篇）和2000篇工程大类（通信工程和信息工程各1,000篇）的论文摘要，共得到33,988个语步结构。两个学科领域论文摘要文本中的数据统计信息如表 2所示。其中，平均句子数、平均词语数和平均语步数分别指每篇摘要中平均包含的句子数量、单词数量和语步数量。

	摘要总数	句子总数	词语总数	平均句子数	平均词语数	平均语步数
人工智能领域	2,670	17,391	381,734	7	143	4
工程领域	2,000	16,597	406,244	8	202	4

Table 2: 两个学科领域的摘要文本数据统计

语步	人工智能	工程领域	小计	%
BAC	3,301	3,165	6,466	19.02
GAP	2,031	1,241	3,272	9.63
PUR	2,694	2,180	4,874	14.34
MTD	4,972	6,554	11,526	33.91
RST	2,181	1,551	3,732	10.98
CLN	1,494	1,512	3,006	8.84
CTN	601	229	830	2.44
IMP	117	165	282	0.83
总计	17,391	16,597	33,988	100

Table 3: 语料库中各类语步结构的标注和分布情况

从表 3可以看到，两个领域中标注数量前三位的语步依次为MTD、BAC、PUR。不同

类型的语步分布存在很大差异：“MTD”语步在两个学科中的标注次数都是最多的，两个领域的标注总量占比高达约34%，甚至超过了第二位“BAC”（6,466）和第三位“PUR”（4,874）的总和；而“IMP”的次数最少，只占到低于1%的比例。这表明不同领域的论文摘要均更关注MTD这一语步。这也跟直觉判断是一致的，因为方法确实是摘要中最有吸引力和最重要的部分之一，在摘要中通常会重点说明论文研究使用的方法，体现出该论文的价值和亮点工作。在标注过程中，我们也发现，很多论文摘要中的方法语步包括不止一个句子。

NLP	[BAC]Knowledge graph (KG) entity typing aims at inferring possible missing entity type instances in KG. [GAP]It is a very significant but still under-explored subtask of knowledge graph completion. [PUR]In this paper, we propose a novel approach for KG entity typing which is trained by jointly utilizing local typing knowledge from existing entity type assertions and global triple knowledge in KGs. [MTD]Specifically, we present two distinct knowledge-driven effective mechanisms of entity type inference. ... [CLN]Experimental results on two real-world datasets (Freebase and YAGO) demonstrate the effectiveness of our proposed mechanisms and models for improving KG entity typing. [CTN]The source code and data of this paper can be obtained from Github...
ME	[PUR]This paper proposes a novel density-based method for structural design considering restrictions of multi-axis machining processes. [MTD]A new mathematical formulation based on Heaviside function is presented to transform the design field into a geometry which can be manufactured by multi-axis machining process. [MTD]The formulation is developed for 5-axis machining, which can be also applied to 2.5D milling restriction. The filter techniques are incorporated to effectively control the minimum size of void region. [MTD]The filter techniques are incorporated to effectively control the minimum size of void region. [CLN]The proposed method is demonstrated by solving the compliance minimization problem for different machinable freeform designs.

Table 4: NLP领域和ME领域摘要语步标注实例对比

语步类型	人工智能领域		工程领域	
	#摘要	占比 (%)	#摘要	占比 (%)
BAC	2,003	75.02	1,528	76.29
GAP	1,518	56.85	891	44.48
PUR	2,333	87.38	1,901	94.91
MTD	2,245	84.08	1,873	93.51
RST	1,540	57.68	953	47.58
CLN	1,192	44.64	1,079	53.87
IMP	112	4.19	159	7.94
CTN	544	20.37	215	10.73

Table 5: 两个学科领域中包括各类语步的摘要数量及占比情况

对比两个学科领域，可以看到不同学科的摘要文本特点和语步分布也存在较大差异。结合表2和表3，工程领域的摘要文本数量比人工智能少了近700篇，摘要文本句子总量整体也比人工智能领域少了近800句，但平均词语数量远高于人工智能，表明工程领域的摘要句子长度更长。语步分布上，工程领域的“MTD”语步数量远高于人工智能领域，其他语步如“BAC”“GAP”“PUR”等数量均少于人工智能。这表明通信工程和机械工程这两个工程学科的论文摘要更倾向用较多的语句来详细论述研究方法。标注团队在实践中也确实发现，很多论文摘要开头甚至很少交代研究背景和研究差距，而是直接用一句话交代研究目的，随即用多个连续的句子讨论研究采用了哪些方法，同时也不太注重提及研究本身的价值以及对于本领域的启发和贡献等。由于人工智能领域整体上具有明显的开源特点，很多论文摘要中经常提到研究

相关的代码、数据等面向公众开源，体现了对于研究社区的贡献，我们会把这种表述统一标注为“CTN”，因此这一语步在人工智能领域的数量更多。表 4通过两个标注实例对比了NLP和机械工程（ME）两个不同领域的摘要语步特点。

在语步分布的基础上，如果不考虑每种语步在摘要中的标注次数（一次或者多次），只要标注了该语步，则认为该摘要文本中包括该语步类型。我们也对这种情况进行了统计。表 5 显示了在2,670篇人工智能论文摘要和2,000篇工程领域论文摘要中，标注每种语步类型的摘要数量以及占全部摘要的比重。标注“PUR”的摘要数量最多，超过2,300篇的人工智能论文摘要中都有该语步，工程领域论文也类似。出现数量第二多的是“MTD”。对比表3和表4中的“PUR”和“MTD”语步，在全部摘要中标注数量最多的“MTD”高于“PUR”，是因为在一篇摘要中，通常只有一个句子被标注为“PUR”，但可能有多个句子被标注为“MTD”。

5 实验及分析

本部分基于已构建的语料库开展了语步结构的自动识别实验，分别从人工智能学科领域和工程学科领域中各抽样选择50篇研究论文摘要作为开放测试集，对比我们的标注模型与官方的大语言模型GPT-4和Claude 3(Opus)的识别效果。实验采用F1值作为评价指标。

5.1 实验数据及设置

表 6是实验数据的基本统计信息。平均句子数和平均句长分别是每篇摘要中平均包含的句子数量和句子中的单词数量。

	句子总数	平均句子数	平均句长	词语总数	词汇量	语步总数
人工智能领域	318	6.36	27.76	8,884	1,902	318
工程领域	397	7.94	32.12	12,752	2,415	397

Table 6: 数据集统计

为了尽可能地提高LLM识别语步类型的准确性，我们设计了下面的Prompt，并分别将其输入至官网的对话框中。大模型根据prompt给出反馈以后，正式开始语步识别任务。每个完整的摘要文本在两个模型中均只标注一次以确保标注公平。

“Move structures are important semantic and discourse units in research articles(RA). You are a senior expert in the field of EAP and are very good at analyzing the move structures in English RAs from different disciplines. You will analyze the move structures with the pre-defined move types and their labels as follows:

(1) Background(BAC): States the research area and provides any historical, theoretical, or empirical related information. (2) Gap(GAP): Establishes a niche: indicates a gap, adds to what is known, presents positive justification. (3) Purpose(PUR): Indicates purpose, thesis or hypothesis, outlines the intention behind the paper. (4) Method(MTD): Provides information on design, procedures, assumptions, approach, data, etc. (5) Result(RST): States main findings or results or what was accomplished. (6) Conclusion(CLN): Summarizes the results or extends results beyond scope of paper. (7) Implication(IMP): Draws inferences which has not been explicitly stated. (8) Contribution(CTN): Points out the theoretical and practical value of the methods used in the articles.

Please identify the most suitable move type and annotate it for [each sentence] in the abstract texts. That is, every complete sentence [must] have a move label. Here is an example:

[BAC] Recent neural models for relation extraction with distant supervision alleviate the impact of irrelevant sentences in a bag by learning importance weights for the sentences. [GAP]Efforts thus far have focused on improving extraction accuracy but little is known about their explainability.”

5.2 实验结果及分析

本研究的自动标注模型与两个大语言模型识别各类语步的对比结果如表 7所示。其中第二列是经人工核实后，数据集里每种语步类型的标准正确数量。图 4是八种语步识别F1指标的均

	正确语步总数	Ours			GPT-4			Claude3		
		#识别	#正确	F1(%)	#识别	#正确	F1(%)	#识别	#正确	F1(%)
AI领域	BAC(61)	64	60	96.00	50	42	75.68	65	52	82.54
	GAP(32)	31	30	95.24	42	29	78.38	32	22	68.75
	PUR(48)	49	45	92.78	46	40	85.11	53	43	85.15
	MTD(101)	94	92	94.36	94	84	86.15	86	78	83.42
	RST(25)	36	25	81.97	45	21	60.00	65	23	51.11
	CLN(45)	34	34	86.08	29	20	54.05	4	3	12.24
	IMP(0)	0	0	0	0	0	0	1	0	0
	CTN(6)	9	6	80.00	0	0	0	15	6	57.14
工程领域	BAC(52)	67	51	85.71	54	33	62.26	56	41	75.93
	GAP(32)	31	30	95.24	24	16	57.14	33	26	80.00
	PUR(53)	65	51	86.44	49	42	82.35	72	50	80.00
	MTD(182)	162	156	90.70	157	140	82.60	143	138	84.92
	RST(11)	28	11	56.41	62	10	27.40	82	11	23.66
	CLN(62)	39	38	75.25	5	4	11.94	2	2	6.25
	IMP(3)	3	3	100	4	1	28.57	0	0	0
	CTN(2)	2	2	100	0	0	0	9	1	18.18

Table 7: 两个学科领域语步自动识别标注实验结果

值。

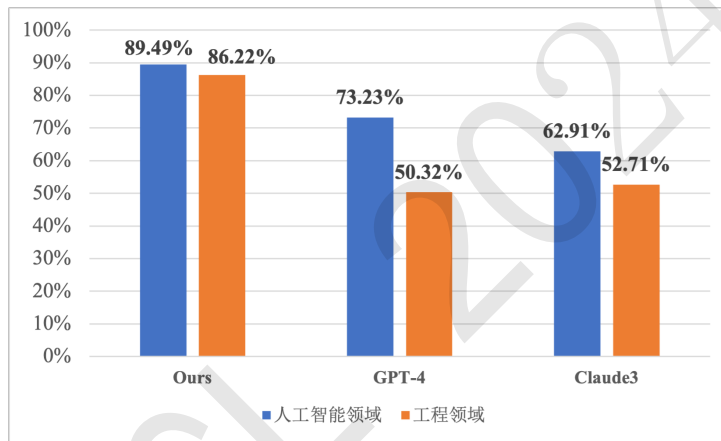


Figure 4: 三个模型整体识别所有语步的平均F1值

5.2.1 自动标注模型与大语言模型识别效果对比

表 7显示，在当前实验中，本研究采用的基于BERT的识别模型在两个领域的语步识别F1指标均明显地远高于两个大语言模型的结果，而且识别每种语步类型的F1也相对更加稳定。

在人工智能领域和工程领域，我们的模型识别效果最好的F1值均在95%以上，所有语步整体识别的平均F1值也分别达到了89%和86%（图 4）；而GPT-4和Claude3的最高F1值仅接近85%，整体平均F1则低于70%。这表明了本文模型的有效性。但值得说明的是，并不是所有语句的识别效果都优于大模型。通过具体分析三个模型的识别结果，也发现了在某些摘要文本的语步识别中存在我们的模型识别错误、而大模型识别正确的情况。例如下面的实例：

例句3: [PUR]In this paper, we propose a novel bipartite flat-graph network (BiFlaG) for nested named entity recognition (NER), which contains two subgraph modules: a flat NER module for outermost entities and a graph module for all the entities located in inner layers. [MTD]Bidirectional LSTM (BiLSTM) and graph convolutional network (GCN) are adopted to jointly learn flat entities and their inner dependencies.

在该例句中，第二个句子出现在PUR语步之后，而且句中存在单词“adopted”，是比较明显的方法语步的信号。这个语步被我们的模型错误识别为BAC，但被GPT-4正确地识别

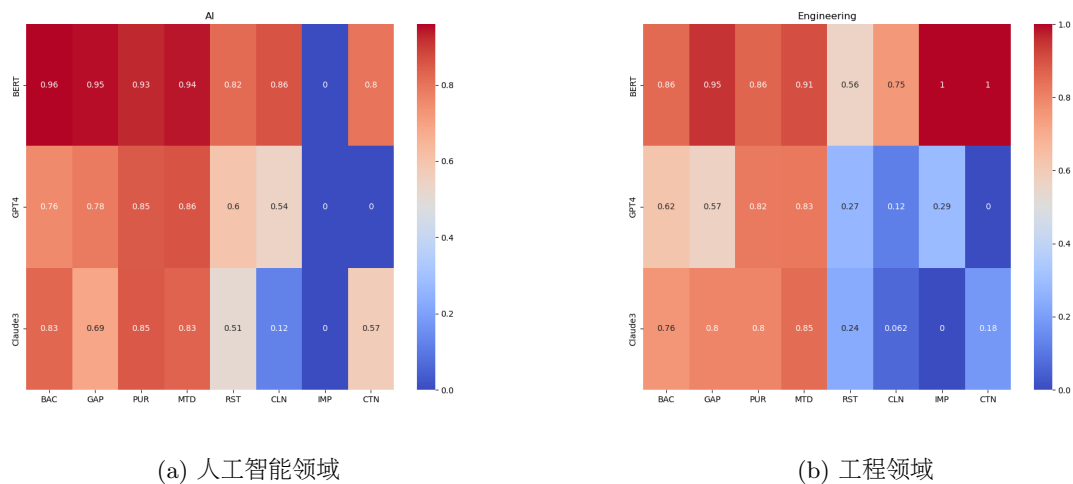


Figure 5: 两个学科领域的语步识别F1指标热力图

为MTD。从中可以看出GPT对于文本语义的理解能力。

5.2.2 两个领域的识别效果对比

图 5是根据表 7绘制的语步识别F1指标热力图。从表 7和图 5中可以观察到，三个模型在AI领域的语步识别效果均明显优于工程领域。特别是工程领域，RST、CLN和CTN语步的识别效果都比较差。具体从两个领域来看，领域内各类语步的识别情况存在较大差异。在AI领域中，三个模型识别效果最好的语步类型均不相同，分别是BAC（我们的模型）、MTD（GPT-4）和PUR（Claude3）；在工程领域，我们的模型除了在IMP和CTN两个语步识别完全正确以外，识别最好的是GAP语步，其次是MTD，GPT4和Claude3则均在MTD上识别最好。这也再次说明了MTD语步在不同领域论文摘要中的重要性以及在语句表达上有其特殊之处，比其他语步更容易被识别。

5.2.3 大语言模型之间的识别效果对比

图 5显示，GPT-4在AI领域的识别效果明显好于Claude3，平均F1值比Claude3高了10个百分点；而在工程领域，Claude3的效果略好于GPT-4，但二者都在50%附近，远低于AI领域的平均F1，主要是受到了RST、CLN和CTN语步的影响。由此可见，整体上AI领域的摘要语步类型特征更加明显，更容易被识别出来。可以认为AI领域论文摘要写作的规范性和可读性好于工程领域。

尽管两个大语言模型的识别效果在两个学科领域中存在差异，但二者的识别表现也具有一定的趋同性。例如：在两个领域中，结论语步（CLN）的识别数量及准确识别数量均远远低于标准的正确数量，甚至出现了个位数的识别量，导致F1值急剧下降。我们试图跟GPT-4了解CLN识别过少的原因，它给出的解释是“CLN通常出现在章节或文档的末尾，如果提供的文本大部分来自文档的引言或方法部分，那么自然会导致CLN的实例减少。”换句话说，由于我们在Prompt中明确提到识别的是摘要文本中的语步，因此GPT-4认为摘要中不应该出现过多的CLN语步。

与CLN相反，两个领域中结果语步（RST）的识别数量则均超过了标准正确数量，也就是把大量原本不属于RST的语步识别为RST。这种情况在我们的识别模型中也很普遍。例如下面的实例：

例句4: [MTD]We propose a simple, effective transition-based model with generic neural encoding for discontinuous NER. [CLN]Through extensive experiments on three biomedical data sets, we show that our model can effectively recognize discontinuous mentions without sacrificing the accuracy on continuous mentions.

在该例句中，第二个句子的语步都被两个大模型识别为RST，但实际上应该是CLN。因为如果句子中存在诸如“show”“demonstrate”之类的单词，一般就能比较明确地判断为CLN。

我们分析了GPT和Claude识别效果不稳定的原因,认为可能主要表现在两个方面:(1)我们只使用了官方网页版的大模型聊天模式,并未像专门训练我们的BERT模型那样有针对性地训练大模型,因此大模型无法充分学习到摘要文本中的语义信息。(2)我们在prompt中也没有为大模型提供太多的提示信息,包括能够快速且准确识别某种语步结构的关键信息等,例如“show”“demonstrate”等词语可以有助于快速判断为CLN语步。我们相信如果进一步优化prompt,将会提升大语言模型的语步识别效果。

6 结束语

本文面向英语科技论文写作,构建了一个涵盖多个学科领域、较大规模的论文摘要语步结构标注语料库。主要通过人工标注与自动标注+人工校对的方式构建,采用了基于BERT的语步自动识别与标注模型,能够保证语料库建设的规模和质量。在语步自动识别实验中,本文的模型在不同学科领域中的各类语步类型识别效果均优于大语言模型GPT4和Claude3的识别效果,体现了所构建的语料库和自动识别模型的有效性和价值。该语料库能够为科技论文信息抽取、科技论文智能辅助写作和批改等自然语言处理任务以及跨学科领域语步分析等外语教学和研究提供必要的数据库支持,帮助二语学习者更好地理解论文语步结构,提升写作能力等,同时通过语言智能技术赋能外语场景,有助于推动外语教育数字化转型。

在未来的研究中,我们将在现有工作成果的基础上继续标注更多学科领域的摘要文本语步结构,持续扩大语料库的规模。我们同时也考虑标注科技论文中的其他组成部分,例如引言和方法部分,目标是建设大规模、多领域、多元化的语步数据资源,为开展科技论文写作智能批改等后续研究提供坚实基础。

致谢

本研究受到国家社科基金项目(23BYY166)资助。感谢项目课题组全体成员为本研究做出的贡献。感谢匿名评审为本文提出的宝贵意见和建议。

参考文献

- Mohammed Alliheedi, Robert E. Mercer, and Robin Cohen. 2019. Annotation of rhetorical moves in biochemistry articles. In *Proceedings of the 6th Workshop on Argument Mining*, pages 113–123, Florence, Italy, August. Association for Computational Linguistics.
- Maryam Alsharif. 2023. Rhetorical move structure in business management research article introductions. *Journal of Language and Linguistic Studies*, 18(4).
- Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).
- Elena Cotos, Sarah Huffman, and Stephanie Link. 2017. A move/step model for methods sections: Demonstrating rigour and credibility. *English for Specific Purposes*, 46:90–106.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, volume 1, pages 4171–4186.
- Ken Hyland. 2000. *Disciplinary discourses: Social Interactions in academic writing*. Longman.
- Ken Hyland. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for specific purposes*, 27(1):4–21.
- Jinkun Lin, Hongzheng Li, Chong Feng, Fang Liu, Ge Shi, Lei Lei, Xing Lv, Ruojin Wang, Yangguang Mei, and Lingnan Xu. 2023. Move structure recognition in scientific papers with saliency attribution. In *China Conference on Knowledge Graph and Semantic Computing*, pages 246–258. Springer.

- Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, July. Association for Computational Linguistics (ACL).
- Xiaofei Lu, Jungwan Yoon, and Olesya Kisselev. 2021. Matching phrase-frames to rhetorical moves in social science research article introductions. *English for Specific Purposes*, 61:63–83.
- Ana I Moreno and John M Swales. 2018. Strengthening move analysis methodology towards bridging the function-form gap. *English for specific purposes*, 50:40–63.
- Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The ACL Anthology network corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLP4DL)*, pages 54–61, Suntec City, Singapore, August. Association for Computational Linguistics (ACL).
- Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: Advancing open science in computational linguistics. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361, Singapore. Association for Computational Linguistics.
- John Swales and Christine B. Feak. 2009. *Abstracts and the writing of abstracts*. University of Michigan Press.
- John Swales. 1990. *Genre analysis: English in academic and research settings*. Cambridge University Press.
- John Swales. 2004. *Research genres: Explorations and applications*. Cambridge University Press.
- John M Swales. 2019. The futures of eap genre studies: A personal viewpoint. *Journal of English for Academic Purposes*, 38:75–82.
- Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway, June. Association for Computational Linguistics (ACL).
- Simone Teufel, Advait Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1493–1502.
- Simone Teufel. 2010. The structure of scientific articles: Applications to citation indexing and summarization. *CA: CSLI Publications*.
- Rodrigo Tovar Viera. 2020. Rhetorical move structure in abstracts of research articles published in ecuadorian and american english-speaking contexts. *Arab World English Journal (AWEJ)*, 10.
- 丁良萍, 张智雄, and 刘欢. 2019. 影响支持向量机模型语步自动识别效果的因素研究. *数据分析与知识发现*, 3(11):16–23.
- 刘霞. 2016. 英语学术论文摘要语步结构自动识别模型的构建. Ph.D. thesis, 北京外国语大学.
- 张鑫, 许海云, 杨宁, 方肖, and 赵爽. 2024. 有限样本下的科技文献语步识别方法探讨. *图书情报工作*, 68:117–129.
- 杜新玉 and 李宁. 2024. 中文学术论文全文语步识别研究. *数据分析与知识发现*, 8:74–83.
- 杨延宁 and 邹航. 2023. 基于语步结构的学术论文语篇构式研究. *外语教学理论与实践*, 182(2):1–18.
- 王东波, 高瑞卿, 叶文豪, 周鑫, and 朱丹浩. 2018. 不同特征下的学术文本结构功能自动识别研究. *情报学报*, 37:997–1008.
- 王末, 崔运鹏, 陈丽, and 李欢. 2020. 基于深度学习的学术论文语步结构分类方法研究. *数据分析与知识发现*, 4:60–68.
- 黄红, 陈冲, and 张婧莹. 2022. 科技文献内容语义识别研究综述. *情报学报*, 41(09):991–1002.