分类号：TP391　　学校代码：10363
密　级：公开　　学　　号：2220920107

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　基于上下文增强的中文缩略词预测研究

论 文 作 者　王潇冉
指 导 教 师　陶皖（教授）
学科（专业）　软件工程
研 究 方 向　领域软件工程

论文提交日期：2025 年 5 月 9 日

分类号： TP391 单位代码： 10363
密　级： 公开 学　　号： 2220920107

题　目 基于上下文增强的中文缩略词预测研究

题　目 Research on Chinese Abbreviation Prediction Based on Context Enhancement

学生姓名： 王潇冉
指导教师： 陶皖（教授）
专　　业： 软件工程
研究方向： 领域软件工程

论文答辩日期： 2025年05月10日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：王慧慧

日期：2015年5月9日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 □，在_____年解密后适用本版权书。

本学位论文属于

不保密 ☑

学位论文作者签名：王萧由　　指导教师签名：陶晓

日期：2015年5月9日　　日期：2015年5月9日

基于上下文增强的中文缩略词预测研究

摘 要

作为语言表达中的重要形式，缩略词能够有效压缩文本并传递关键信息，从而使信息传递更加快捷。在语言交流中，缩略词通常承载着比原词更为精炼和精准的语义。随着信息技术和互联网的迅猛发展，信息量激增且内容复杂。在大规模文本数据处理中，缩略词的使用能够有效减轻语言表达的负担，避免冗长描述，便于在有效的时间和空间内获取核心信息，是信息检索、实体关联等诸多下游任务的基石，在自然语言处理的各类任务中发挥着关键作用。然而，缩略词常常具有多义性和上下文依赖性，因此，准确预测缩略词对于提升信息检索系统和搜索引擎等工具的准确性和效率至关重要。为此，缩略词预测任务应运而生，旨在预测实体完整形式的可能缩写形式，这些缩写通常由实体完整形式中的字符子序列组成。有效的缩略词预测不仅能为各种自然语言处理任务提供更为精确的输入，还能加深系统对语义理解的深度，从而提升其相关下游任务的准确率与可靠性。

目前的缩略词预测任务大体上可归结于基于特征的方法，其将缩略词预测问题建模为一个序列标记任务，即对输入序列数据中的每一个元素分配一个标签或类别，以指示该字符是否应保留。然而，这种方法往往难以充分利用整体上下文信息，尤其在中文缩略词的多义性较高的情况下。此外，缩略词的生成不仅依赖于局部上下文，还与其

他成分的语义关系密切相关。传统的序列标记方法通常将任务分割为独立的标记任务，忽视了与实体上下文文本中的丰富信息。本文以基于上下文增强的生成方法及融合主题一致性的评估方法为主要研究方向，以改进语义理解模型，提升缩略词预测任务的性能。主要研究内容如下：

（1）针对中文缩略词生成问题，本文提出一种基于上下文增强的中文缩略词生成方法，将中文缩略词预测任务视为序列生成任务。通过增加上下文信息来增强对缩略词的语义理解，借助预训练语言模型的强大表示能力，准确获取文本数据并有效捕捉上下文信息，实现对上下文的深层理解，从而提高生成缩略词的流畅性和可用性。同时，引入基于规则的候选缩略词筛选机制，剔除不合理或不符合规则的候选项，确保生成的候选缩略词更符合实际，并提高生成结果的准确性。

（2）针对中文缩略词的多义性和语境依赖性，本文提出一种基于上下文主题一致性的中文缩略词预测框架。通过构建“生成-评估”二阶段协同机制，实现语义准确性和主题一致性的双重提升。首先，在候选生成阶段，采用预训练语言模型对输入文本进行多粒度语义编码，通过注意力机制捕捉完整形式与上下文的关键关联，生成兼顾局部语境与全局语义的候选缩略词集合。随后，建立主题一致性评判体系，将候选缩略词动态嵌入原始语境，通过主题建模提取每个候选项的核心主题向量，以捕捉与原始文本主题相关的语义信息。通过对这些主题词的相似性进行评判，按照相似性评分对候选缩略词进行排序，确保生成的缩略词在语义上与原实体保持一致，从而进一步提升中文

缩略词预测的准确性，尤其在处理复杂上下文和多义性缩略词时，能够提供更为精准的预测结果。

关键词：自然语言处理，中文缩略词预测，序列生成模型，上下文增强，主题一致性

Research on Chinese Abbreviation Prediction Based on Context Enhancement

ABSTRACT

As a crucial form of linguistic expression, abbreviations effectively compress text and convey key information, thereby facilitating more efficient information transmission. In linguistic communication, abbreviations typically encapsulate meanings that are more concise and precise than their original terms. With the rapid development of information technology and the internet, the volume of information has surged, and content has become increasingly complex. In large-scale text data processing, the use of abbreviations helps alleviate the burden of linguistic expression, avoiding lengthy descriptions and enabling efficient retrieval of core information within limited time and space. Abbreviations serve as a cornerstone for various downstream tasks, such as information retrieval and entity linking, playing a vital role in numerous natural language processing (NLP) applications. However, abbreviations often exhibit polysemy and strong contextual dependence, making their accurate prediction crucial for enhancing the precision and efficiency of

information retrieval systems, search engines, and similar tools. Consequently, the task of abbreviation prediction has emerged, aiming to generate possible abbreviated forms of full entities, which are typically composed of character subsequences from the full entity name. Effective abbreviation prediction not only provides more precise input for various NLP tasks but also enhances the system's semantic understanding, thereby improving the accuracy and reliability of related downstream tasks.

Currently, abbreviation prediction tasks are predominantly based on feature-based approaches, which model abbreviation prediction as a sequence labeling task. In this approach, each element in an input sequence is assigned a label or category to indicate whether the character should be retained. However, such methods often fail to fully utilize global contextual information, especially given the high degree of polysemy in Chinese abbreviations. Furthermore, abbreviation generation is not solely dependent on local context but is also closely related to the semantic relationships among other components. Traditional sequence labeling methods typically divide the task into independent labeling steps, thereby overlooking the rich information present in the surrounding text related to the entity. This study focuses on a context-enhanced generation approach and an evaluation method incorporating thematic consistency to improve semantic understanding models and enhance the performance of

abbreviation prediction tasks. The main research contributions are as follows:

(1) To address the problem of Chinese abbreviation generation, this study proposes a context-enhanced approach that formulates abbreviation prediction as a sequence generation task. By incorporating additional contextual information, the proposed method enhances the semantic understanding of abbreviations. Leveraging the powerful representation capabilities of pretrained language models, it effectively captures textual and contextual information, facilitating a deeper comprehension of surrounding text and improving the fluency and usability of generated abbreviations. Additionally, a rule-based candidate filtering mechanism is introduced to eliminate unreasonable or nonconforming candidates, ensuring that the generated abbreviations better align with real-world usage and improving the accuracy of prediction results.

(2) To address the challenges posed by polysemy and contextual dependence in Chinese abbreviations, this study proposes a thematic consistency-based framework for abbreviation prediction. A two-stage "generation-evaluation" mechanism is designed to enhance both semantic accuracy and thematic consistency. First, in the candidate generation stage, a pretrained language model is employed to encode input text at multiple semantic levels. An attention mechanism is utilized to capture critical associations between the full form and its surrounding context,

generating a set of candidate abbreviations that consider both local context and global semantics. Subsequently, a thematic consistency evaluation system is established, dynamically embedding candidate abbreviations back into the original context. By employing topic modeling, core topic vectors are extracted for each candidate to capture semantic information related to the original text's theme. The similarity of these topic words is then assessed to rank candidate abbreviations based on their thematic relevance, ensuring that the generated abbreviations remain semantically consistent with the original entity. This approach significantly improves abbreviation prediction accuracy, particularly in handling complex contexts and ambiguous abbreviations, providing more precise predictions.

Wang Xiaoran(Software Engineering)
Supervised by Prof. Tao Wan, Assoc. Prof. Qiu Huili

KEY WORDS: Natural language processing, Chinese abbreviation prediction, sequence generation model, context enhancement, thematic consistency

目 录

摘 要 I
ABSTRACT IV
第 1 章 绪论 1
1.1 研究背景及意义 1
1.2 国内外研究现状及发展趋势 4
1.2.1 中文缩略词 4
1.2.2 序列生成模型 6
1.2.3 对比学习 7
1.2.4 主题模型 7
1.3 研究基本框架及主要贡献 8
1.4 研究内容及组织结构 12
第 2 章 关键技术和相关研究 14
2.1 文本表示方法 14
2.1.1 独热编码 14
2.1.2 词嵌入 15
2.1.3 ELMo 17
2.2 Transformer 18
2.2.1 注意力机制 20
2.3 对比学习 23
2.4 本章小结 24
第 3 章 中文缩略词上下文数据集构建 25
3.1 研究背景 25
3.2 基础数据架构 26
3.3 数据集上下文扩充与处理 28
3.4 本章小结 29
第 4 章 基于上下文增强的中文缩略词生成 31
4.1 研究背景 31
4.2 模型目标 32
4.3 模型结构 33
4.3.1 上下文增强编码 34
4.3.2 缩略解码 36
4.4 实验与分析 37
4.4.1 实验设置与评价指标 37

4.4.2 性能比较 37
4.4.3 基于上下文信息长度的参数实验 39
4.4.4 人工评估 39
4.5 本章小结 40
第 5 章 基于上下文主题一致性的中文缩略词预测 41
5.1 研究背景 41
5.2 模型目标 42
5.3 模型结构 43
5.3.1 候选生成阶段 44
5.3.2 上下文主题一致性评估阶段 45
5.4 实验与分析 47
5.4.1 实验设置与评价指标 48
5.4.2 候选缩略词数量确定 48
5.4.3 性能比较 49
5.4.4 基于上下文长度的参数实验 51
5.4.5 基于启发式规则的消融实验 51
5.4.6 人工评估 52
5.5 本章小结 53
第 6 章 总结与展望 54
6.1 工作总结 54
6.2 未来展望 55
参考文献 56
攻读学位期间发表的学术成果 63
致 谢 64

第 1 章 绪论

1.1 研究背景及意义

随着信息技术的飞速发展和互联网的普及，现代社会产生了海量的数据和信息。面对如此庞大的信息量，如何高效地从中提取核心内容、进行有效的交流，成为了信息处理领域的一个重要课题[1]。在这个过程中，结构精炼、传达精准且不失原意的词汇逐渐获得认可，这类词汇即所谓的缩略词（Abbreviations）。缩略词是指通过简化途径将长词或短语转化为一种文本缩写形态，它是一种蕴含意义的紧凑表达方式，通常由原始完整词汇中的部分字符构建而成。为便于书写与表达，在文本中指代某一特定实体时，人们倾向于采用其缩写形式而非全称[2]。鉴于其简洁性与能充分传达语义的特点，缩略词在各类自然语言处理任务中得到广泛应用。其不仅能有效压缩文本长度，减少冗余信息，还能在传递关键信息的同时提高表达的简洁性与精确度，尤其在信息检索、文本摘要、机器翻译等自然语言处理任务中，缩略词的正确处理对于提高系统效率和准确性具有重要意义。

缩略词预测（Abbreviation Prediction）[3]，旨在预测实体全称所对应的可能缩写形态，例如将“中国中央电视台”预测为“央视”或“中央台”。当前对缩略词的研究大多集中在英文首字母缩略词，如将“National Aeronautics and Space Administration”缩写为“NASA”。相较于英文缩略词，中文缩略词形态更为丰富多元，表 1-1 列举了中文缩略词的一些实例，一般认为其可归纳为四种类型：语素构成、中心词构成、混合构成与合并构成[4-6]。

（1） 语素构成：此类中文缩略词由源短语中各词的语素组合生成，所选语素均表达原词核心信息，但在原词内部位置并不固定，如“奥林匹克运动会”缩略为“奥运”。

（2） 中心词构成：该类中文缩略词以词为单位，抽取原词组中的关键成分构建而成，此类缩略词的原词组通常结构复杂，包含三个及以上词，如“上海迪士尼”缩略为“迪士尼”。

（3） 混合法构成：此类缩略词所抽取的成分包括语素、词以及音节的混合

体，形成一种较为特殊的缩略形式，如“中央电视台”缩略为“中央台”。

（4） 合并法构成：针对原词组中存在多项并列语言成分的情况，通过合并手段将其简化为中文缩略词，如“有理想、有道德、有文化、有纪律”缩略为“四有”。

表 1-1 中文缩略词构成方式示例

Table1-1 Example of Chinese Abbreviation Composition Method

构成方式	全称	缩略词
语素	奥林匹克运动	奥运
	安全保卫	安保
中心词	上海迪士尼	迪士尼
	人造地球卫星	人造卫星
混合法	中央电视台	中央台
	广播体操	广播操
合并法	包退、包换、包修	三包
	有理想、有道德、有文化、有纪律	四有

另外由于中文所含信息量大，中文缩略词含义往往依赖于上下文，同一个缩写在不同语境下可能表达不同意思，并且一个实体可能存在多个缩写形式。因此，中文缩略词预测成为自然语言处理领域的一个巨大挑战。

当前对于中文缩略词预测的研究主要集中于两类方法：基于规则的启发式算法[7]和基于特征的方法。基于规则的启发式算法是早期中文缩略词预测的主流方法。其核心假设在于缩略词的生成遵循特定的语言学规律与统计特征。该方法通过人工定义的语言学规则集，结合词频等统计指标，实现对目标缩略形式的推导。然而，人工制定的规则往往难以覆盖所有可能的缩略词，规则的制定往往依赖于专家的经验，可能存在主观性和不一致性。此外，随着缩略词数量的增加，维护

和扩展规则库的成本较高，容易出现遗漏。基于特征的方法则将缩略词预测问题模型化为一个序列标记任务。其核心思想是通过对原词中的每个字符进行二元或多元标签分类，确定哪些字符应被保留在缩略词中。尽管这种方法在中文缩略词预测方面取得了一些成就，但存在着以下问题：

(1) 它们仅仅依赖于转移矩阵来寻找最高概率的标签，未充分考虑标签之间的依赖关系。换句话说，它们仅预测当前字符的标签，而未考虑输入序列之前字符的标签。

(2) 它们几乎没有利用整个输入序列的整体语义信息，因为它们仅基于到目前为止的输入序列来预测当前字符的标签。

(3) 它们忽略了与实体相关文本中的丰富信息，因为它们仅利用实体本身的语义。

实际上，实体周围的文本可能包含关于实体的重要上下文信息，而这些信息在传统方法中被忽略了。最近的研究已着手处理这些问题，不过仍将缩略词预测视为序列标记任务，主要集中于提取更有效的文本特征。包括通过采用深度学习模型来自动挖掘连续文本特征，并基于这些特征对每个标记分配二进制标签等。这些深度学习方法的优势在于它们可以更好地捕捉上下文中的语义信息，从而提高了中文缩写的预测性能。

此外，随着大语言模型（Large Language Model，LLM）技术的突破，中文缩略词预测任务逐渐被形式化为一个序列生成任务（Sequence-to-Sequence，Seq2Seq）。在此框架下，模型需将完整命名实体（源语言）转化为其缩略形式（目标语言），这一过程可类比为机器翻译中的跨语言映射，但专注于同一语言内的信息压缩与重构。其利用序列生成模型建模输入和输出之间的全局依赖关系，并充分利用整个输入序列的语义。与序列标记方法相比，序列生成任务是被广泛认可的、针对中文缩略词进行合理处理的方法。其通过自注意力机制捕捉输入序列的跨位置依赖关系，避免了局部标注导致的语义割裂问题。此外，针对中文缩略词生成结果研究可知，候选缩略词的生成数量与最终命中率呈显著正相关。这代表着对预测的缩略词进行一个额外的评估筛选可以大大提升中文缩略词预测的准确率。

综上所述，基于上下文增强的中文缩略词预测研究能够提升预测任务性能，

优化数据信息的学习方式，强化模型对非连续语义压缩与动态场景适配的生成能力，对文本理解、信息检索、实体关联等下游任务的开发具有显著的意义与价值。

1.2 国内外研究现状及发展趋势

面对中文缩略词预测技术研究逐渐兴起，相关的理论知识和方法层出不穷。本课题的研究内容主要包含基于上下文信息的中文缩略词生成模型和基于上下文主题一致性的中文缩略词评估模型，其涉及到序列生成以及对比学习两方面内容。以下从中文缩略词预测、序列生成模型、对比学习和主题模型四个方面介绍其研究现状和发展趋势。

1.2.1 中文缩略词

中文缩略词，在构造形态上是由一个较长的中文短语经过压缩、概括等操作，形成的长度缩短、意义不变的特殊短语。由于其简洁并能够充分表达语义的效果，在各种自然语言处理任务[8-9]中被广泛应用。例如，如果能正确识别文本中的缩写，那么信息检索（Information Retrieval，IR）系统[10]的查询召回率和语音搜索[11]的准确率将会显著提高。此外，如果可以预测一个实体的缩写，那么文本中出现的缩写可以链接到知识库中正确的实体，从而进一步提高实体链接任务[8]的准确性。

在中文缩略词的早期探索中，研究者们主要依托于启发式规则与经典的机器学习算法。基于启发式规则的方法主要是利用语言学规则、词典匹配和启发式策略来生成缩略词。这类方法通常借助汉字拼音、词频信息及首字提取等规则进行处理。例如，“国际航空公司”，从每个词语的第一个字取出，组合成缩略词为“国航”。随着机器学习技术的发展，研究者们开始采用统计学习方法进行缩略词的预测。通过构建特征向量，利用分类算法进行预测。例如，Sun[12]使用基于上下文信息的支持向量机（Support Vector Machine，SVM）方法，有效地在上下文中识别出原实体完整形式和其缩略词。在此基础上，Sun[13]为了将非本地信息纳入缩略词生成任务中，进一步提出了隐式和显示两种解决方案：潜变量模型或全球信息的标签编码方法。Chang[14]则提出了一种基于隐马尔可夫（Hidden Markov Model，HMM）的单字符恢复模型，用于从文本语料库中提取大量的原

子缩略词及其全称。此外，Chang[15]还提出了一种基于 HMM 的生成模型，将该中文缩略词问题视为一种错误恢复问题，其中可疑的词根是需要从候选集合中恢复的“错误”。

随着研究的深入，一些研究人员开始将中文缩略词预测视为一项序列标记任务，并使用序列标记模型进行处理。例如，Sun 和 Yang[10][11][16]等人提出设计了一系列的特征功能，并使用条件随机场（Conditional Random Field，CRF）作为标记工具。Chen[17] 进一步结合了一阶逻辑，对 CRF 无法覆盖的长距离和全局依赖关系进行了建模。此外，Zhang[18]探索了结合搜索引擎返回结果的方法，通过手动调整对 CRF 预测结果的排序。Jiao[19]则利用 CRF 模型对完整形式进行预测，生成一定数量的缩略词候选，并通过搜索引擎获取的信息对这些候选进行评估和打分，最终通过重排序选择出最合适的缩略结果。近年来，随着神经网络技术的兴起，其在特征自动提取方面的优势逐渐被应用于中文缩略词预测任务中，从而取代了传统的手动特征设计方法。如 Zhang[20]为了解决序列标记任务生成的一些缩略词不符合中文的传统习惯，提出了一种新颖的神经网络架构。该结构结合了循环神经网络（Recurrent Neural Network，RNN）和一个确定给定字符序列是否可以构成词的机制来完成中文缩略词生成任务。Zhang[3]则利用双向长短时记忆网络（Bi-directional Long Short-Term Memory，BiLSTM）自动挖掘文本中的连续特征，以预测每个实体向量对应的标签。

针对前述研究，Ma[21]提出序列标记方法无法充分利用整个句子地整体意义，并且在处理标签之间地依赖关系时存在局限性。基于神经网络的序列标记模型往往忽略标签之间的依赖性，因为它们主要依赖转移矩阵来鼓励最有可能的标签序列生成。为了解决这些问题，Wang[22]首先将中文缩写预测作为一个序列生成问题，并提出了一个多层次的预训练模型来包含字符、单词和概念级的嵌入。Cao[23]进一步提出了一种具有缩写恢复策略的上下文增强 Transformer，主要是通过迭代译码过程来预测缩写序列，每一轮迭代都由一个缩写和恢复操作组成。Tong[24]从全新视角出发，聚焦于评估缩略词在其文本上下文中的质量，这一方法为缩略词预测的准确性和实用性提供了新的评价准则。

以上方法虽不同程度的改进了中文缩略词预测性能，但中文本身的特性使得中文缩略词形式复杂多样，而目前的研究未充分考虑到中文缩略词本身的多样性，

导致模型的泛化能力不足。此外由于一个缩略词可能在不同信息中含义存在差异，而现有的研究往往未能充分考虑到这些差异，这可能导致中文缩略词预测的准确性受到影响。

1.2.2 序列生成模型

Seq2Seq 模型[25-28]作为一种生成模型，最初应用于机器翻译[29]。所谓 Seq2Seq，即序列到序列模型，就是一种能够根据给定的序列，通过特定的生成方法生成另一个序列的方法，同时这两个序列可以不等长。这种结构又叫 Encoder-Decoder 模型，即编码-解码模型。在早期阶段，大多数基于神经网络的机器翻译模型都利用了递归神经网络，其中顺序和循环结构增加了训练难度。随后，Seq2Seq 被广泛应用于构建生成模型。然而，普通的 Seq2Seq 模型解析长文本的能力有限，因为其固定维的编码向量无法保留长文本的所有语义信息。为了解决这一问题，随着注意力机制在图像分类、语音识别、动作识别和文本摘要等许多领域的应用，许多自然语言处理（Natural Language Processing，NLP）模型也采用注意力机制来提高学习性能。Pham[30]在机器翻译模型中引入了两类有效的注意机制，包括全局方法和局部方法。Vaswani[31]提出了一种新的完全基于自注意网络和标准前馈网络相结合的深度 Transformer。然而，自回归框架并不能简单地适应并行令牌的生成。Gu[32]提出了一种非自回归变换器（NAT），该解码器可以并行地产生整个目标序列。Lee[33]提出在解码器中添加一个迭代细化阶段，以显著加快解码速度。Stern[34]提出了一种基于插入操作的序列生成迭代部分自回归模型。为了更有效地迭代译码，条件掩蔽语言模型（CMLM）[35]利用掩蔽语言建模目标进行模型训练，迭代预测标签。Levenshetin Transformer[36]也是一种新的部分自回归模型，由插入和删除操作组成，以实现更灵活的序列生成。G-Transformer[37]在 Transformer 中引入了局部性假设作为归纳偏差，以降低多句和文档级翻译的目标到源注意的复杂性。由于其有效性，许多基于注意力机制的 Seq2Seq 模型的变体被提出，并成功地应用于许多不同的领域，如语音识别[38]、视频总结[39]、对话系统[40]、医学总结[41]和交通预测[42]。

目前对于中文缩略词预测的研究热衷于将其看作为序列生成任务，但目前的大多数序列生成模型在语言理解和文本生成任务中存在的差距较大，这使得对中

文文本的处理面临着巨大的挑战。

1.2.3 对比学习

对比学习（Contrastive Learning），是一种无监督的学习方法，旨在拉近相似样本的距离，同时区分不相似样本。目前在计算机视觉（Computer Vision，CV）和 NLP 领域都取得了很大的进展。早期的对比学习工作[43-46]侧重于 CV，并将其应用于将图像的转换版本作为积极的例子，将两个不同的图像作为消极的例子，以自我监督的方式学习良好的视觉表征。近年来，对比学习在自然语言处理领域得到了广泛的应用，例如文本相似度计算、文本分类、词义表示学习和句子表示学习等。为了学习良好的文本表示，研究者提出了用不同的方法来构建积极的和消极的例子。SimCSE[47]使用 Dropout[48]噪声作为隐式数据增强，以一种无监督的方式构造正的例子。CLINE[49]使用 WordNet[50]将句子中的单词替换为同义词、反义词或随机单词来构造正和消极的例子。Taeuk[51]使用两种不同的 BERT[52]，而不是显式的数据增强来构建积极的和消极的例子。ConSERT[53]在标记嵌入层引入了四种数据增强策略来构造正实例。DeCLUTR[54]设计了几种策略来从不同的文档中取样阳性和阴性。

除了表征学习外，对比学习最近也被应用于不同的自然语言处理任务中。具体来说，在自然语言生成（Natural Language Generation，NLG）任务中，多应用于多语言机器翻译中的对比学习[55]，通过采样不同语言的相似句子作为积极的例子，将不同语言的句子表示到一个共享的空间中。SimCLS[56]将对比学习应用于一个额外的无参考评估阶段进行抽象总结，旨在通过相应的评估指标直接优化生成模型，以缩小生成模型的训练目标和评估指标之间的差距。

目前对比学习在 NLP 领域中受到了广泛应用，但很少有研究将对比学习与中文缩略词预测结合起来。本课题受到了 SimCLS[56]的启发，利用对比学习基于上下文信息使用预先训练好的大语言模型来评估由中文缩略词生成模型生成的候选缩略词质量。

1.2.4 主题模型

为了揭示文本中的共性主题及其内在叙事，主题模型已被证明是一种强大的

无监督文本挖掘工具。传统的主题模型以 Blei 提出的隐含狄利克雷分布（LatentDirichlet Allocation，LDA）为代表[57]，它假设每个文档是由多个主题生成的，通过贝叶斯推断方法来估计主题分布，从而实现主题的自动发现。在 LDA 的基础上，研究者们提出了多种扩展模型，如层次 LDA（Hierarchical LDA）[58]和相关主题模型，这些模型考虑了主题之间的相关性和层次结构，进一步提高了主题建模的灵活性和准确性。此外，深度学习的兴起使得研究者们开始探索基于深度学习的主题模型，如深度信念网络（Deep Belief Networks，DBN）[59]和变分自编码器（Variational Autoencoders，VAE）[60]，这些模型能够更好地捕捉文本的复杂特征。

近年来，文本嵌入技术在自然语言处理领域迅速流行，尤其是双向编码器表示（BERT）及其变体（如 RoBERTa[61]等）在生成上下文相关的词和句子向量表示方面表现出色。这些向量表示的语义特性使得文本的含义能够以一种方式编码，使得相似的文本在向量空间中彼此接近。研究者们开始将这些强大的上下文表示用于主题建模，Sia 等[62]展示了使用基于质心的技术对嵌入进行聚类的可行性，相较于传统方法如 LDA，作为表示主题的一种方式。

此外，随着神经主题模型越来越成功地利用神经网络来改进现有的主题建模技术[63]。将词嵌入纳入经典模型（如 LDA）展示了使用这些强大表示的可行性。在 LDA 类模型中，最近出现了一些主要围绕嵌入模型构建的主题建模技术，展示了基于嵌入的主题建模技术的潜力[64]。在此基础，Grootendorst 提出 BERTopic[65]，它将主题建模和文本嵌入技术相结合，创新型地采用了聚类方法，并结合基于类别地 TF-IDF 变体，生成连贯且准确地主题表示。

综上所述，随着预训练语言模型和嵌入技术地不断发展，主题建模在自然语言处理中的应用正不断扩展，为中文缩略词预测任务提供了一种新颖且高效的解决方案。尤其在处理复杂语境和多义性问题时，能够展现出更强的优势。

1.3 研究基本框架及主要贡献

在当代信息社会中，中文缩略词的使用愈发普遍，相关的预测任务也引起了广泛关注。然而，实际应用中仍然面临着诸多挑战。首先，中文缩略词具有较强的多义性和上下文依赖性。一个缩略词可能在不同的上下文中有不同的含义，这

使得准确理解和预测缩略词的具体含义变得复杂。例如，“华师”可以指代“华东师范大学”，也可以表示“华中师范大学”。其次，中文缩略词的生成质量通常受到上下文信息利用不足的限制。传统的缩略词生成方法往往忽视了文本中潜在的语义关联性，导致生成的缩略词可能与原始实体的语义不完全一致。尤其是生成的缩略词多为原始形式的简化组合，无法很好地表达其在上下文中的真实含义。这不仅影响了生成结果的流畅性，还导致生成的缩略词缺乏语义一致性。因此，如何改进生成模型，以更好地利用上下文信息，确保生成的缩略词既符合语法规范，又保持语义的一致性和流畅性，是当前的一个重要研究问题。

此外，候选缩略词的准确性和排序问题也是中文缩略词预测任务中的一个重要挑战。在预测过程中，生成的候选缩略词可能存在准确性不高或排序不合理的情况，影响了最终预测结果的质量。如果生成的缩略词未能准确反映原始实体的语义，或者候选缩略词的排序未遵循实际应用中的逻辑，那么预测结果将无法有效满足真实语境的需要。因此，设计有效的评估模型来优化候选缩略词的准确性和排序，将有助于提升中文缩略词预测任务的整体性能。

目前中文缩略词预测任务主要存在着以下挑战：

（1）如何提升生成结果的流畅性和语义关联问题：先前对于中文缩略词生成的研究虽不同程度的提升了缩略词预测性能，但是它们未充分利用上下文信息，导致生成的缩略词结果通常是原始形式的不完全组合，缺乏连贯性和流畅性。因此，关键问题在于如何改进现有生成模型，使其更好地利用上下文信息，以解决中文缩略词生成的流畅性和语义关联性问题。

（2）如何优化候选缩略词的准确性和排序问题：在中文缩略词预测任务中，生成的候选缩略词可能存在着准确性不足和排序不合理的问题，影响预测结果的质量。首先，候选缩略词的准确性涉及生成的候选缩略词是否能准确还原原始形式的语义。如果生成的缩略词与原始含义不匹配，会导致信息的误解和语义的混淆。其次，候选缩略词的排序问题指的是生成的缩略词在预测结果中的顺序是否合理和有序，这直接影响中文缩略词预测任务的准确性。因此，关键问题在于如何设计评估模型，以保持合理的排序顺序提升中文缩略词预测任务的性能。

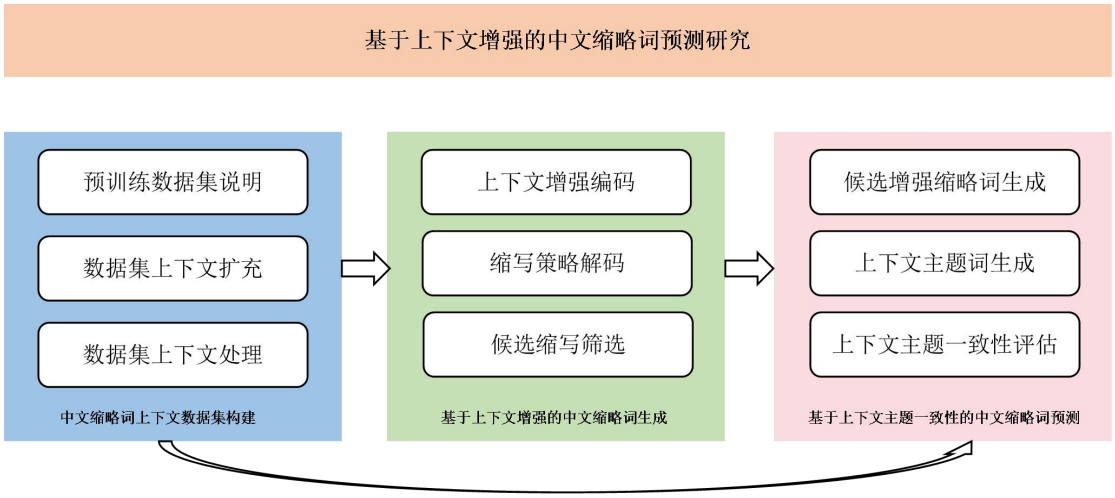

图 1-1 论文组织结构图

Figure1-1 Organizational Chart of the Paper

本文面向中文缩略词预测任务，针对中文缩略词存在的多义性和上下文依赖性和现有的序列生成模型存在的问题设计了两种基于上下文的中文缩略词预测方法。图 1-1 展示了本文的研究框架，具体来说，本文对基于上下文增强的中文缩略词预测方法展开了三个方向的研究，它们分别是中文缩略词上下文数据集构建、基于上下文增强的中文缩略词生成和基于上下文主题一致性的中文缩略词预测。主要研究内容如下：

(1) 中文缩略词上下文数据集构建

现存的中文缩略词数据集存在一些限制，即不包含条目的上下文信息或多数上下文信息仅由一句话组成。然而，仅靠一个单词一句话往往无法全面揭示中文缩略词的完整语义，无法满足一个合格的缩略词应该遵守的形式简洁、语义明确、约定俗成三个原则，因为这些词汇通常在特定的语境下才能正确理解。没有充分上下文信息的支持可能导致生成的中文缩略词无法正确还原原始形式的语义。因此，为了更全面的捕捉词语的语义范围，本课题将针对现有的中文缩略词数据集，对其的上下文信息进行扩充，这样有助于更深入理解中文缩略词，提高后续对中文缩略词的预测性能。

(2) 基于上下文增强的中文缩略词生成

近年来，针对中文缩略词的研究主要聚焦于实体本身，忽略了中文缩略词本身极大的语义依赖性。文本语境中的多样性和歧义性使得缩略词预测变得复杂。同一个缩略词在不同上下文中可能具有不同的含义，这种多义性对模型提出了更

高的理解和预测要求。现有方法主要依赖于概率模型或神经网络，通常涉及提取离散特征，例如分词信息、字符特征等，然后使用机器学习模型，对每个字符进行二元分类。但它们忽略了与实体相关的文本中的丰富信息，因为它们仅利用实体本身的语义。实际上，实体周围的文本可能包含关于实体的重要上下文信息，而这些信息在传统方法中被忽略了。因此，本课题提出一种基于上下文增强的中文缩略词生成。

1）针对完整形式的丰富上下文信息，通过注意力机制捕捉输入序列和目标序列之间的关联关系，获取上下文语义特征，采用预训练模型获取词与词之间的关联关系，实现对上下文的深层理解，最终提升生成缩写的可用性和流畅性。

2）采取规则处理的方法对生成的候选缩略词进行预先筛选，以提高序列生成任务性能。

(3) 基于上下文主题一致性的中文缩略词预测

1）候选缩略词数量确定

目前中文缩略词生成模型存在准确率不佳，未能充分满足自然语言处理下游任务期望的问题。对于生成模型生成的候选缩略词数量对于命中正确缩略词结果具有一定影响。理论上，生成的候选缩略词越多，候选缩略词中就越容易得到正确缩略词结果。然而，随着候选缩略词的数量增加，训练评估模型所需的时间和空间开销会显著增加。因此为了在命中率和时空开销之间取得平衡，本课题将通过小样本数据测试来确定候选缩略词的数量。这样可以在一定程度上平衡命中率和时空开销，确保后续评估任务的正确进行。

2）候选缩略词的重排序

在中文缩略词预测任务中，对于中文缩略词生成模型得到的候选缩略词难免会出现顺序混乱，语义关联不明的情况。因此对候选缩略词进行重排序是提高中文缩略词预测任务准确性和效率的关键。本课题采用上下文主题一致性评估的方法结合对比学习算法，对生成的候选缩略词集合进行处理。通过合理的特征选择、优化算法以及排序模型的训练，可以获得更准确、更有语义相关性的缩略词结果。

总的来说，上述三项研究内容构成了一个有机的整体，层层递进。第一项研究内容被视为本课题的前提，为后续的研究内容提供丰富的数据基础。第二项研究内容是本课题的基础，将中文缩略词预测任务定义为序列生成问题，融合先进

的预训练语言模型与详尽的上下文信息生成缩略词，旨在确保缩略词质量与可靠性。第三项研究内容是本课题的核心，其在第二项研究内容的基础上，将中文缩略词预测任务分为先生成后评估的二阶段。在第一阶段，生成缩略词候选集，并辅以规则处理进行初步筛选，确保候选集的质量。在第二阶段，通过引入上下文主题一致性的新视角对候选缩略词进行质量评估，从而完成缩略词预测任务。

1.4 研究内容及组织结构

本文的研究内容和组织结构如下:

第 1 章为绪论。简要介绍了基于上下文增强的中文缩略词预测的研究背景及意义，阐明了当前中文缩略词预测所面临的挑战，并对相关技术的国内外研究现状进行了简要的介绍。最后描述了本文研究的基本框架及主要贡献，并总结了本文的组织结构。

第 2 章为中文缩略词预测方法的相关理论研究技术介绍。简要讲述了中文缩略词预测的系统概述，同时介绍了一些相关理论技术的背景知识，包括注意力机制、对比学习、序列生成模型和主题模型等，并比较了它们的优势与局限性。

第 3 章为中文缩略词上下文数据集的构建。在本章中将重点介绍如何基于上下文信息构建中文缩略词预测任务的数据集。数据集的质量对于机器学习模型的训练和评估至关重要，尤其是在中文缩略词预测任务中，由于缩略词的多义性和上下文依赖性，构建一个有效的数据集需要特别关注上下文信息的提取和使用。本章分为两个部分：第一部分介绍上下文对中文缩略词预测任务的重要性，第二部分则详细描述数据获取的来源、数据处理方法及处理流程。

第 4 章为基于上下文增强的中文缩略词生成。首先介绍了该算法的整体框架，然后分别对各个模块进行分析，最后通过在第 3 章构建的中文缩略词数据集上的对比实验证明了该算法的有效性。

第 5 章为基于上下文主题一致性的中文缩略词预测。根据第 4 章对中文缩略词生成的研究内容，提出基于主题一致性的中文缩略词评估模型，旨在提升中文缩略词预测任务的准确性和实用性。针对第 4 章中生成的中文缩略词候选，评估模型通过计算候选缩略词与上下文主题的匹配度，对生成的缩略词进行主题一致性评分，从而优化缩略词的排序与选择。最后通过在中文缩略词数据集上的对比

实验证明了评估模型的有效性。

第 6 章为总结和展望。本章系统地回顾和整理了本文的研究工作，明确了提出算法的优点和局限性，并对下一步的优化方案进行了初步探究，最后对本研究的未来发展方向进行了展望。

第 2 章 关键技术和相关研究

本章主要介绍了与中文缩略词预测任务中相关的理论知识和技术背景。本文关于中文缩略词预测任务中所需要的模型方法主要包括：文本表示、Transformer、注意力机制及对比学习。

2.1 文本表示方法

文本表示（Text Representation）[66]，是自然语言处理中的核心基础任务，旨在解决非结构化语语言数据与计算机可处理形式之间的语义鸿沟问题。由于语言文字本身缺乏机器可理解的显示结构，且不同语言体系在语法规则、语义表达和信息区分度上存在显著差异，如何将文本转化为高效、可泛化的数字化表示形式，成为实现机器自主认知语义（如理解、推理、生成）的关键前提。对下游任务（如文本分类、信息检索和机器翻译）提供了可扩展的语义计算基础。

2.1.1 独热编码

独热编码（One-Hot Encoding）[67]，又称为一位有效编码，是自然语言处理中基础的离散文本表示方法。其核心思想是通过定长二进制向量对离散类别型数据进行唯一性编码，使得每个数据实例在正交向量空间中具有独立且互斥的表示。该方法通过构建语料库词典并建立词汇到索引的映射关系，将文本中的每个词映射为与词典长度相同的稀疏向量。如图 2-1，该向量仅在其对应索引位置为 1，其余位置均为 0，从而实现符号化语言到数值化形式的转换。

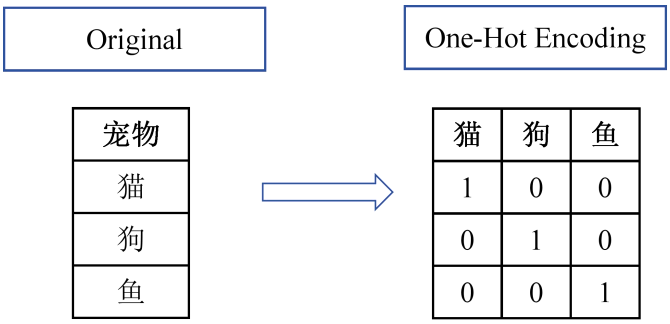

图 2-1 独热编码示例图

Figure2-1 One-hot encoding example diagram

独热编码的主要优势在于其能够清晰地区分不同类别的数据，增强模型对离散文本特征的处理能力。此外，由于其值仅为 0 和 1，能够在一定程度上优化数据的表示方式，使得类别之间的距离计算更加直观。然而，独热编码的缺点也很明显，主要体现在唯独爆炸问题：随着词汇量的增加，编码向量的维度会迅速膨胀，占用大量存储空间，并影响计算效率。此外，该方法未能捕捉单词之间的语义关联，即不同单词之间的相似性无法体系，使得模型难以理解词语之间的深层联系。

2.1.2 词嵌入

词嵌入（Word Embedding），是一种用于自然语言处理的表示技术，它将单词或短语从离散的词汇表映射到一个连续的向量空间[68]，使得单词的语义信息能够以数值形式表达。该技术的基本思想是：语义相似的词通常会出现在相似的上下文中。这种观点来源于分布式语义学神经网络模型的发展以及分布式表示（Distributed Representation）的应用。与独热编码相比，词嵌入能够揭示词汇之间的相似性和关系，因为相似的词汇在向量空间中会更接近。词嵌入技术不仅能够减少模型的复杂度，还能提升模型的表达能力和泛化性能。

从数学角度来看，Embedding 可以视作为一种映射关系，即映射 $maping(f:a\to b)$，其具有单射性和结构保留性[69]。在词嵌入的情况下，每个单词都会被唯一映射到一个向量，而映射后的空间仍然保留了原始数据的一定结构，例如，如果单词 a_1 和 a_2 在原始空间中的某种关系（如相似性或语义关联性）较强，那么它们在向量空间中的对应表示 b_1 和 b_2 也会保持相似的相对位置。如图 2-2 所示，具有相似属性的单词在嵌入空间中更为接近，这种特性使得词嵌入方法能够有效学习词语之间的语义关系，从而提升自然语言处理任务的性能。

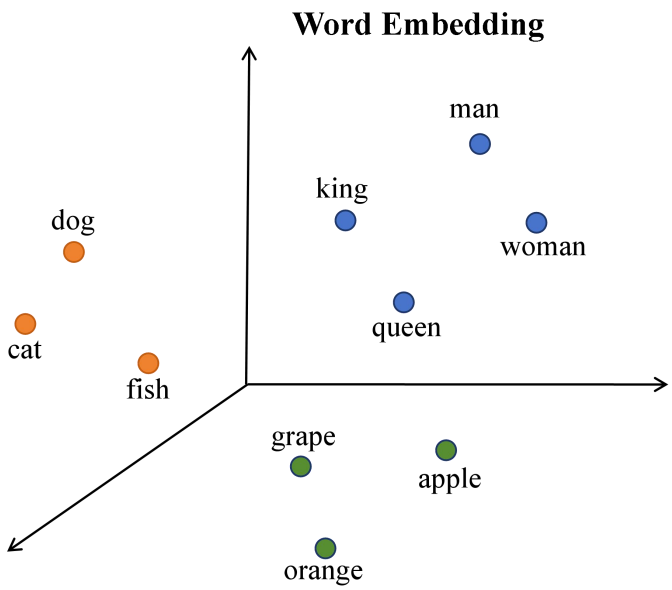

图 2-2 词嵌入空间结构示例图

Figure2-2 Word embedding space structure example diagram

在众多词嵌入方法中，Word2Vec 是最具代表性的一种，由 Mikolov 等人于 2013 年提出，包含两种基于神经网络的训练范式：连续词袋模型（Continuous Bag-of-Words, CBOW）和跳字模型（Skip-gram），并将训练得到的模型参数作为文本的向量表示[70]。这两种方法的训练过程如图 2-3 所示。

CBOW 通过上下文预测中心词。具体而言，给定一个目标词 $w(t)$，设定一个固定大小的窗口（如窗口大小为 2），并使用窗口内除 $w(t)$ 以外的词作为输入，即 $w(t-2)$ 到 $w(t+2)$，输入层神经网络将这些词的信息聚合后传递给投影层，最终用于预测 $w(t)$。训练完成后，输入层的网络参数即为学习到的词向量表示。而 Skip-gram 则是 CBOW 的逆过程，它使用目标词 $w(t)$ 作为输入，预测窗口内的其他词汇。相较于 CBOW 仅需预测一个目标词，Skip-gram 需要预测多个上下文词，因此其任务难度更大，但训练得到的词向量通常更具表达能力，尤其在语料库规模足够大的情况下，能够捕捉到更丰富的词语语义信息。

与传统的词袋模型相比，Word2Vec 在构建词向量时充分利用了上下文信息，提升了文本理解的能力，在多种自然语言处理任务中取得了显著成效。然而，对于缩略词的表示，Word2Vec 仍存在一定局限性。由于其窗口机制在上下文建模时存在固定长度的约束，难以有效捕捉长距离依赖关系，从而在处理包含复杂上下文的缩略词时表现较弱。这一问题限制了 Word2Vec 在缩略词预测任务中的应用，因此后续研究者探索了更高级的文本表示方法以弥补这一不足。

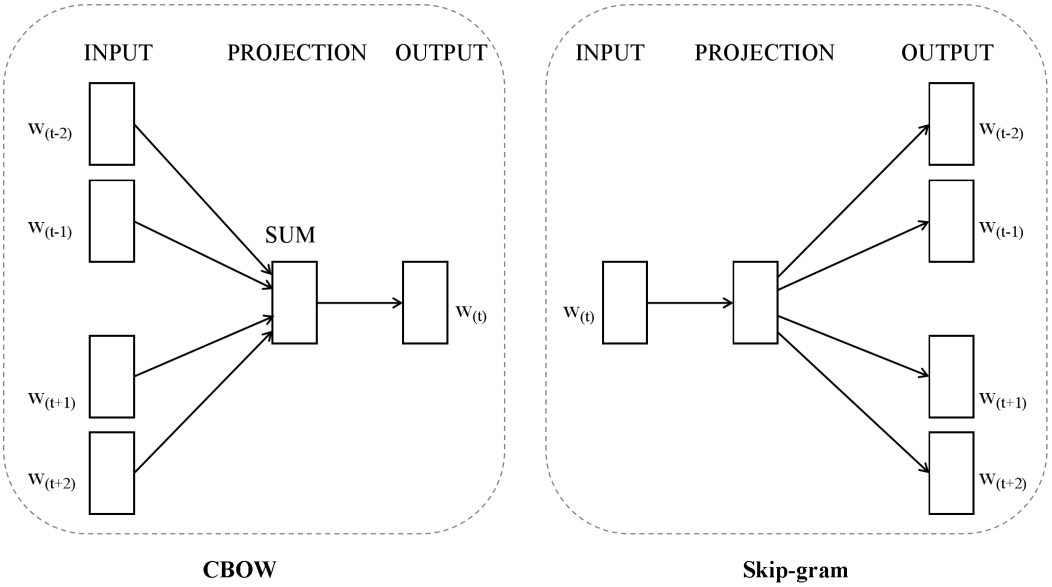

图 2-3 CBOW 和 Skip-gram 的训练过程

Figure2-3 The training processes of CBOW and Skip-gram

2.1.3 ELMo

传统的词向量表示方法由于采用固定的词向量映射方式，难以有效处理多义词及动态语境带来的挑战。为了解决这一问题，ELMo（Embeddings from Language Model）提出了一种全新的文本表示方式，不仅在多个自然语言处理任务中取得优异表现，还对双向语言模型和上下文感知词表示的发展产生了深远影响[71]。

ELMo 通过双向长短期记忆网络 BiLSTM 架构来建模上下文信息。ELMo 的训练过程可分为两个阶段。第一阶段为预训练阶段，在大规模语料库上训练语言模型。如图 2-4 所示，该网格结构采用了双层双向的长短期记忆网络（Long Short-Term Memory，LSTM），左侧的前向双层 LSTM 代表正方向编码器，输入的是从左向右顺序除的上文（不包括预测词），右侧的反向双层 LSTM 代表反方向编码器，输入的是从右向左的逆序的下文。第二个阶段则是下游任务的调整，在进行下游任务时，从预训练网络中提取对应单词网络各层的 Word Embedding 作为新特征补充到下游任务中。

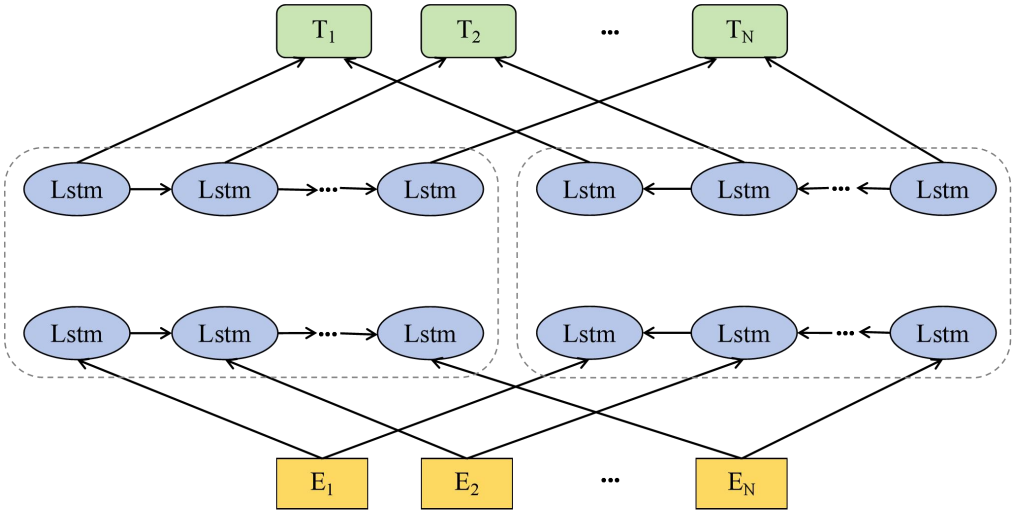

图 2-4 ELMo 架构图

Figure2-4 ELMo architecture diagram

相比于传统的静态词嵌入方法（如 Word2Vec、GloVe[72]），其核心创新点在于同一个词在不同的上下文中可拥有不同的词向量表示，从而有效解决多义词的问题，同时增强了对上下文信息的建模能力。但 LSTM 架构本身的串行计算模式限制了长距离依赖捕获能力，这一局限也成为了 ELMo 方法的局限之一，但也为后续基于 Transformer 架构之上的模型奠定了理论基础。

2.2 Transformer

Transformer 是一种基于自注意力机制的深度神经网络模型，用于解决序列到序列（seq2seq）任务，最初由谷歌提出，广泛应用于自然语言处理和语音识别等任务。传统的循环神经网络（RNN）和卷积神经网络（Convolutional Neural Network，CNN）通常通过递归或卷积操作来处理序列数据。尽管这些方法简单易用，但在处理长文本时常常面临着长期依赖关系难以捕捉和计算效率低下等问题。相比之下，Transformer 模型采用注意力机制来处理序列数据，克服了递归神经网络和卷积神经网络的局限性，从而在长文本处理上展现出更优的性能。作为当前机器翻译、文本生成等领域的先进模型，Transformer 以其卓越的性能和高效的并行计算能力而受到广泛关注。后续的理解模型和生成模型基本都是基于 Transformer 结构进行改进[73]。

如图 2-5 所示，Transformer 模型由两部分组成，分别是左边的编码器（Encoder）和右边的解码器（Decoder）。每个部分由多个相同的层（Layer）堆叠而成。其

中编码部分是由多层的 Encoder 组成，编码器负责对输入序列进行特征提取和表示学习。同理，解码部分也是由多层的 Decoder 组成。每一个编码器在结构上都是一样的，但是它们的权重参数是不同的。每一个编码器由多个堆叠的相同层组成，每一层包含两个主要部分：自注意力层（Self-Attention Layer）和前馈神经网络（Feed Forward Neural Network，FFNN）。自注意力层用于捕捉输入序列中各个位置之间的关系。具体来说，输入的每个词都会和其他层进行交互，从而生成表示这个词的上下文信息。前馈神经网络通常由两个全连接层组成，带有激活函数，用于进一步处理自注意力机制输出的表示。每层的输出经过层归一化和残差连接处理，以保持信息流的稳定。

解码器的作用是接受编码器的输出并生成最终的预测输出。解码器与编码器相似，也由多个堆叠的相同层组成，包含自注意力层、编码器-解码器注意力层和前馈神经网络。解码器的自注意力机制与编码器稍有不同，解码器的自注意力层会通过掩蔽（Masking）机制避免看到未来的信息，保证了自回归生成的顺序性。编码器-解码器注意力层的作用是通过注意力机制将解码器的每个词与编码器输出的表示进行交互，从而生成对目标任务有用的上下文表示。前馈神经网络则类似于编码器中的前馈神经网络，用于进一步处理输入信息。

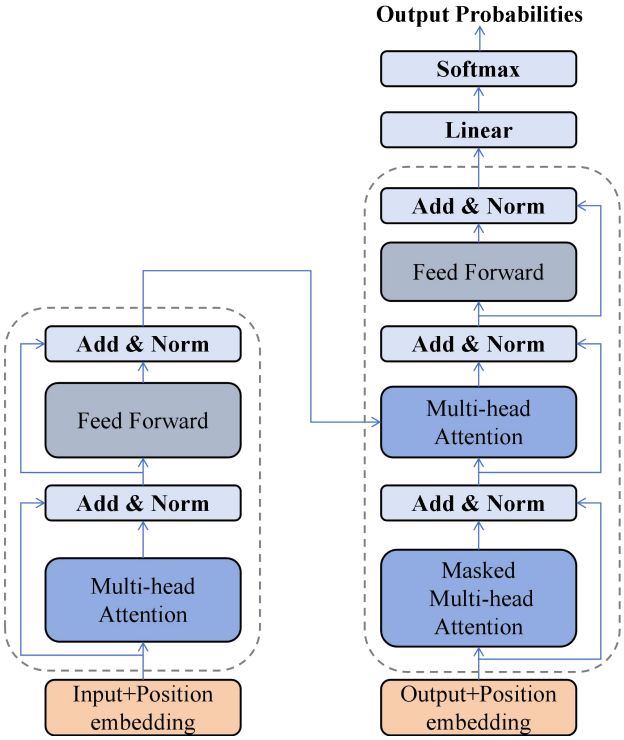

图 2-5 Transformer 模型架构

Figure2-5 Transformer model architecture

2.2.1 注意力机制

注意力机制[74-75]（Attention Mechanism）是一种模仿人类视觉注意力和认知系统的计算方法，最初起源于神经科学领域，用于描述一种选择性聚焦重要信息而忽略次要信息地生物本能。近年来，注意力机制被广泛应用于深度学习，通过引入注意力机制，神经网络能够动态调整对输入数据的关注程度，模仿人类聚焦重点的思维方式。其核心思想是根据输入数据不同部分的重要性动态调整模型的关注点。在处理复杂任务时，模型不必均匀地关注所有输入，而是可以根据上下文信息，选择性地聚焦于与当前任务最相关的部分。这种选择性关注的能力使得模型能够更有效地提取信息，从而提高整体性能。

注意力机制通常由查询（Query）、键（Key）和值（Value）三个组成部分构成。查询指的是查询范畴，自主提示，本质上是体现主观意识的特征向量。键是作为被拿来比对的对象，非自主提示，即反映物体显著特征的信息向量。值则是代表物体本身的特征向量，通常和键成对出现。注意力机制的关键流程是借助 Query 与 Key 之间的注意力汇聚，针对 Value 进行注意力权重的分配操作，进而生成最终的输出结果。具体计算过程如图 2-6 所示。

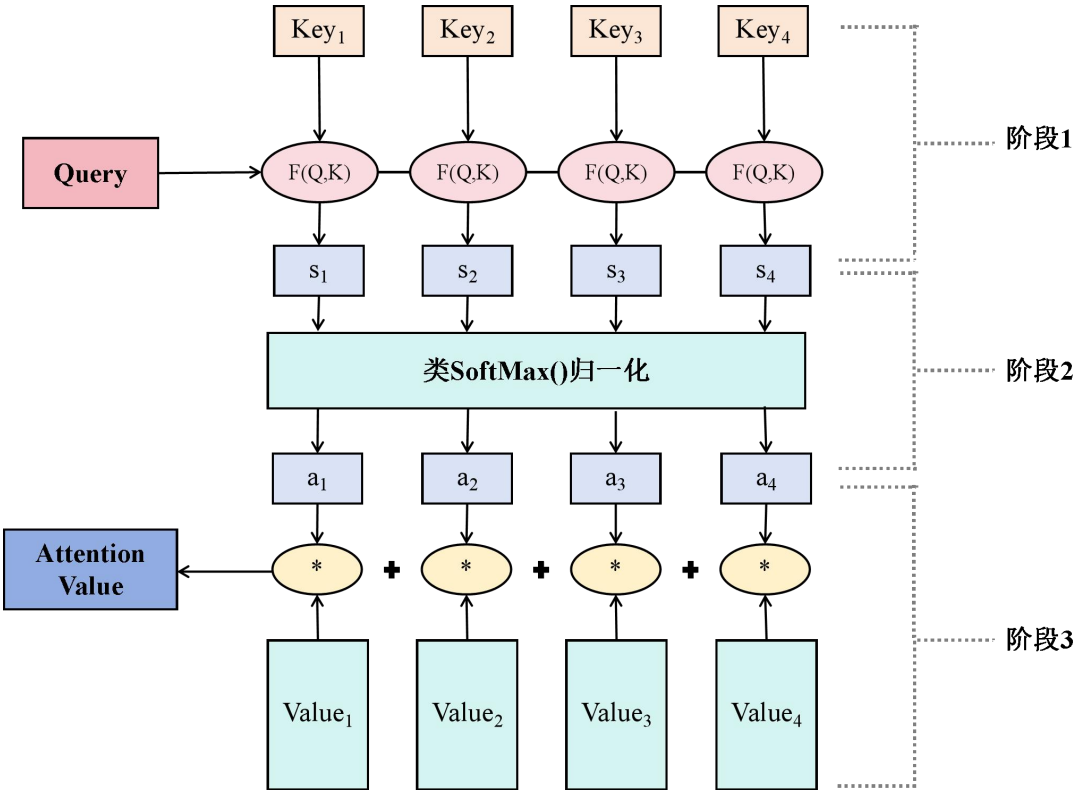

图 2-6 注意力机制计算过程

Figure2-6 The calculation process of the attention mechanism

目前常见的注意力机制包括自注意力机制和多头自注意力机制等，这些注意力机制在各种自然语言处理任务中发挥了重要作用。

（1） 自注意力机制

自注意力机制（Self-Attention Mechanism）[76]旨在通过建立序列内部元素间的全连接权重矩阵，实现全局依赖关系的动态捕获。与 RNN 和 LSTM 等时序模型不同，自注意力机制摆脱位置递推的计算约束，直接计算任意两元素间的关联强度。它想要解决的问题是神经网络接收的输入包含众多尺寸各异的向量，并且这些不同向量之间存在特定关联。然而，在实际训练时却难以将这些输入之间的关系充分利用起来，这就导致模型训练结果不理想。针对全连接神经网络对于多个相关的输入无法建立起相关性的这个问题，通过自注意力机制来解决，自注意力机制实际上是想让机器能够留意到整个输入里不同部分之间的相关性。

自注意力机制是注意力机制的变体，其减少了对外部信息的依赖，在捕捉数据或特征的内部相关性方面更具优势。在注意力机制中，其 Query 和 Key 是不同来源的。而与注意力机制不同，自注意力机制中的 Query、Key 和 Value 是同一个东西，或者三者来源于同一个 X 。通过 X 找到 X 里面的关键点，从而更关注 X 的关键信息，忽略 X 的不重要信息。不是输入语句和输出语句之间的注意力机制，而是输入语句内部元素之间或者输出语句内部元素之间发生的注意力机制。

由于自注意力机制能够直接建模全局依赖关系，它在处理长序列数据时比 RNN 更高效且能并行计算，大幅提升了训练速度。这使得自注意力机制成为 Transformer 结构的核心组件，并广泛应用于机器翻译、文本生成和问答系统等自然语言处理任务。

（2） 多头注意力机制

多头注意力机制（Multi-Head Attention Mechanism）[77]是对传统注意力机制的改进，旨在通过分割输入特征为多个注意力头（Attention Heads）并独立处理每个头部来提高模型的表达能力和学习能力。如图 2-7 所示，在多头注意力机制中，将输入的特征通过多个独立的、并行运行的注意力头进行处理。每个注意力头都会独立的计算注意力得分，并生成一个注意力加权后的输出。这些输出随后通过拼接或平均被合并以形成一个最终的更复杂的表示。注意力机制通过聚焦于

关键信息，提高了模型对输入数据的理解和处理能力。而多头注意力机制通过并行处理和集成多个注意力头的结果，从不同角度捕捉数据的多样性，进一步增强了模型的学习能力和表达能力。

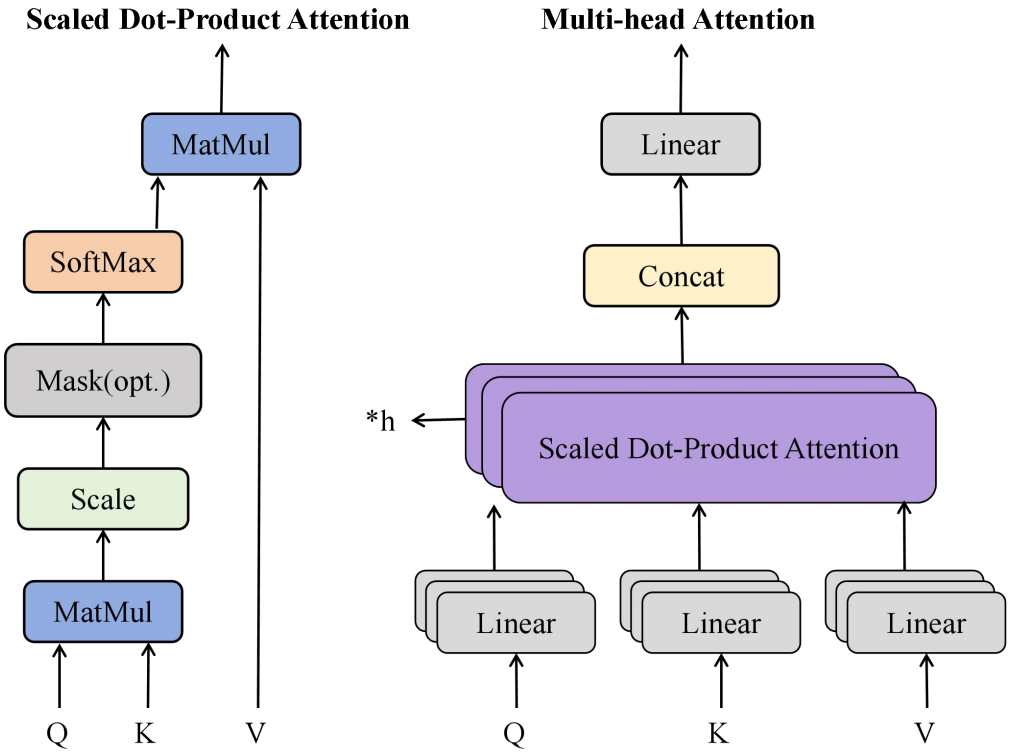

图 2-7 多头注意力机制结构示意图

Figure2-7 Schematic diagram of the multi-head attention mechanism structure

通过引入多个注意力头，模型能够在多个子空间中独立学习信息。每个注意力头可以关注输入数据的不同方面或特征，从而使模型能够捕捉到更丰富和多样化的表示。这种能力对于处理复杂的输入数据尤为重要。在序列数据中，尤其是自然语言处理任务中，词与词之间的关系可能相隔很远。多头注意力机制通过并行计算多个注意力头，能够有效捕捉长距离依赖关系，克服传统递归神经网络在处理长序列时的局限性。这使得模型在理解上下文时更加灵活和准确。此外，多头注意力机制的并行计算使得模型在处理大规模数据时更加高效，能够加速训练和推理过程。这在实际应用中，尤其是在需要实时处理的任务中，具有显著的优势。

2.3 对比学习

对比学习（Contrastive Learning）[78]是一种无监督学习方法，旨在通过比较不同样本之间的相似性和差异性，学习到有效的特征表示。不同于传统的监督学习方法，对比学习无需依赖大量标注数据，而是通过构造正负样本对来优化模型，使相似样本的表示更加接近，而不相似的样本推进。其优点在于能够利用未标记数据进行训练，同时提高特征表示的泛化能力，使其适用于不同任务和领域。此外，对比学习还能通过合理的设计和调整，缓解类别不平衡和噪声数据等监督学习方法常见的问题。

在对比学习中，模型需要学习将同类别样本映射到相邻的表示空间区域，并将不同类别的样本区分开。为实现这一目标，通常采用特定的损失函数，如对比损失（Contrastive Loss）和信息量损失（InfoNCE Loss）。对比损失通过最小化同类别样本间的距离并最大化不同类别样本间的距离来优化模型，而 InfoNCE Loss 利用 Softmax 计算正样本的概率，从而增强模型对区分样本的能力。为了进一步提升性能，对比学习还结合了数据增强策略（如旋转、裁剪、颜色变换等），以增加样本多样性，提高模型的鲁棒性和泛化能力。此外，不同的对比策略（如采样方法和相似度计算方式）也被广泛用于探索更优的训练方案。

对比学习已在多个领域取得广泛应用。例如，在计算机视觉任务（如图像分类、目标检测、人脸识别等）中，代表性方法 SimCLR 通过数据增强和对比策略优化图像特征表示，显著提升了分类效果。在自然语言处理任务（如文本分类、问答、语言生成等）中，对比学习通过将文本映射到低维向量空间，并结合注意力机制（如 Transformer）来建模文本之间的相似性。中文缩略词预测的核心任务是根据给定的实体名称预测可能的缩略词形式。然而，由于缩略词的构造方式多样，可能存在多个可行的缩略词候选，使得传统基于规则或简单统计的方法难以精确匹配正确的缩略词。对比学习通过优化损失函数，使得模型能够更清晰地区分不同类别的缩略词表示，减少因语料噪声或上下文歧义导致的错误匹配。此外，由于缩略词的实际含义往往依赖于上下文信息，对比学习可以通过构造不同的上下文变换，使模型学习缩略词在不同语境下的表现，并确保在相似上下文中，缩略词的语义一致性。这样模型不仅能够预测常见的缩略词，还能结合上下文推测更符合特定场景的缩略词，提高预测的可靠性。

2.4 本章小结

根据本文的研究目标和研究内容，本章对本研究将要利用的相关技术做了介绍，这些技术理论为本文的研究工作提供了重要的启发，并为后续的研究奠定了坚实的基础。首先，本章对自然语言处理领域中基础的文本表示方法进行概述，然后介绍了本文模型中所涉及到的关键技术：Transformer、注意力机制方法框架和对比学习等相关内容，并对每个内容进行了深入分析。

第 3 章 中文缩略词上下文数据集构建

中文缩略词的动态性与多义性赋予其在自然语言处理领域中极高的复杂性与挑战性。当前的研究大多受限于传统的“实体-缩略词”二元数据范式，未能充分考虑到缩略词在不同语境下的多重含义及其语境依赖性，这导致现有的预测模型在实际应用中常常受到语义歧义的干扰，表现出较为显著的局限性。因此，构建一个能够融入丰富上下文信息的中文缩略词数据集，显得尤为关键。这不仅是提升缩略词预测准确度的基础，也为模型在复杂语境下的鲁棒性提供了强有力的支持。本章将深入探讨中文缩略词上下文数据集的构建过程，具体而言，3.1 节首先引入任务的设计背景，分析并阐述上下文信息在缩略词消歧、提升模型泛化能力及增强语义理解等方面的深远意义，进而指出现有数据集所面临的若干不足与限制。紧接着，3.2 节将详细描述数据集的基础架构，包括数据来源、结构设计及其整体框架。在此基础上，3.3 节针对现有数据集中上下文信息不足的显著问题，提出了一种创新的上下文扩充和处理方案，该方案结合了百度百科和生成式语言模型的优势，旨在弥补数据集在实际应用中的不足，进一步提升其实用性和完整性。

3.1 研究背景

中文缩略词作为语言动态演变中的重要表现形式，广泛出现在新闻报道、社交媒体、学术文献以及众多专业领域的文本中。其生成规则既具有显著的规律性，又充满了灵活性，既遵循语言学规律，如“核心语素提取”和“语义组合压缩”等，又因语境的动态变化而常常产生歧义性与领域依赖性。例如，缩略词“高工”在不同的语境中可能分别指代“高级工程师”或“高精度工业设备”。同样，“北航”既可以指代“北京航空航天大学”，也可在特定语境下表示“北方航空枢纽”。这种多义性和上下文敏感性使得缩略词的预测与消歧任务高度依赖上下文信息的支持。

然而，当前中文缩略词研究面临的一个显著问题是数据的局限性。主流的公开数据集大多采用“实体-缩略词”二元组形式来存储，缺乏对原始上下文语境的记录。尽管这些数据集能支持基础的缩略词提取与映射任务，但却无法提供足

够的语境信息，这会直接影响到缩略词预测的精度，进而降低模型在实际应用中的泛化能力。因为中文中许多缩略词具有歧义，且其含义往往依赖于具体语境。在传统的缩略词预测任务中，模型通常仅依赖实体与缩略词之间的直接匹配进行学习和推理。然而，缩略词的真正意义往往依赖于其所在句子或段落的语境。这种简单的建模方法难以捕捉到缩略词与其真实意义之间的深层联系。因此，如何有效地构建并补充包含上下文信息的缩略词数据集，成为提升中文缩略词预测性能的关键所在。

从多个维度来看，上下文信息在中文缩略词预测中的重要性尤为突出：

（1）消除歧义性：许多缩略词在不同语境中可能存在多个含义，通过引入上下文信息，模型能够更准确地判断特定语境下缩略词的具体含义，从而有效消除歧义。

（2）提高模型的泛化能力：上下文信息帮助模型识别语境中的潜在规律，增强其对未见数据的预测能力。当模型能够结合上下文进行推理时，缩略词的预测能力得到了显著提高，进而增强了模型在未知数据上的鲁棒性与适应性。

（3）增强模型的语义理解：上下文的引入不仅有助于提升模型对缩略词语义的把握，还能够加深模型对语言的全面理解。传统的词处理方法虽然能够捕捉到单一词汇的语义，但往往无法充分考虑词汇在具体语境中的实际效果，这使得模型在处理多义词和歧义时显得力不从心。

综合来看，构建一个充实上下文信息的中文缩略词预测数据集无疑是提升缩略词预测任务性能的关键。本文旨在通过引入多源上下文信息（如百科知识库与生成式语言模型的补充），有效弥补数据集中的上下文缺失，增强其语义表达能力，为后续的缩略词预测模型提供更加全面的训练数据。这一数据集的构建，不仅能优化后续模型的训练过程，还能为模型的实际应用提供强有力的支持和保障。

3.2 基础数据架构

本文以 Wang[22]发布的基准数据集作为基础架构，其采用了两种策略，分别从 CN-DBpedia[79]和百度百科两个数据源进行信息收集，确保数据的多样性和准确性。

（1） CN-DBpedia

CN-DBpedia 是中文领域最具代表性的开放域结构化知识库之一，基于 DBpedia 项目而构建。DBpedia 是一个将维基百科中的结构化信息提取并转换为可查询的知识库的项目，CN-DBpedia 则专注于中文内容的提取和整理。CN-DBpedia 中的数据以三元组的形式组织，采用（头实体，关系，尾实体）形式存储知识，例如（中国，首都，北京）、（莫言，获奖，诺贝尔文学奖）。为了从中提取有效的缩略词数据，其首先从 CN-DBpedia 中选择关系为缩略词、全名、其他名称等的三元组。通过分析三元组数据，其确定了以下规则用于选择有效的三元组：

1) 缩略词与全称字符一致性：选择的三元组中的缩略词必须由全称中的字符组成。这是为了保证缩略词和全称之间具有语义上的一致性，避免因字符不相关而造成的上下文不匹配。

2) 字符顺序一致性：缩略词中的字符顺序应与其对应的全称中的字符顺序保持一致。该条件的目的是确保从全称到缩略词的映射具有明确的顺序逻辑，避免由于顺序混乱导致的上下文偏差。

在满足上述两个条件后，就从 CN-DBpedia 中筛选出了符合要求的三元组

（2） 百度百科

为了获取更多的缩略词数据集，Wang[22]进一步从百度百科上进行收集。百度百科是中国最大的在线百科全书之一，涵盖了丰富的中文实体信息。百度百科中的每个条目都有一个标题，该标题通常直接表示了该条目的全称。例如，“中国国家铁路集团有限公司”即为该条目的标题，该标题下的描述部分通常会详细介绍该实体的相关背景信息，包括常用的缩略词或别名。因此，将百度百科条目的标题视为该实体的全称，并通过正则表达式从该条目的描述部分来提取与其对应的缩略词。具体而言，正则表达式根据缩略词的常见表现形式（如：全称后面跟随的括号中的缩略词，或者是缩写形式）进行匹配。例如，“中国国家铁路集团有限公司（简称：中铁集团）”中的“中铁集团”即为该全称的缩略词。通过这种方式，可以有效地从百度百科的描述中提取出各类缩略词。

通过上述方法，构建了一个大型的中文缩略词数据集，数据集统计情况详见表 3-1，具体而言，该数据集中的每个条目都由实体全称和其对应的缩略词组成。数据集以 8：1：1 的比例分为训练集、验证集和测试集。

表 3-1 数据集统计

Table3-1 Dataset statistics

类别	训练集	验证集	测试集
实体总数	75,724	9,466	9,465
实体全称平均长度	9.6	9.7	9.6
缩略词平均长度	4.8	4.8	4.8

3.3 数据集上下文扩充与处理

中文缩略词的预测任务对上下文信息的依赖程度极高，因此，获取具有代表性且高质量的上下文信息对数据集的构建至关重要。原始数据集仅包含实体全称与对应的缩略词，缺乏足够的上下文支持，难以充分体现缩略词在不同语境中的多义性和依赖性。为此，本文对数据集进行了进一步的上下文补充，以提高数据集的语义丰富度和预测任务的准确性。具体补充过程如下：

（1） 从中文百度百科知识库中获取上下文信息

原始数据集中的部分数据是通过百度百科获取的，百度百科的条目标题被视为实体的全称，描述部分则用于提取对应的缩略词。由于百度百科条目的描述部分是对实体全称最为权威且丰富的上下文资源，本文利用百度百科的 API 接口自动化地检索每个实体的信息，获取相关上下文。具体过程如下：

1） 对于数据集中的每个实体地全称，首先通过 API 接口请求百度百科的条目。如果条目存在，则直接返回该条目的摘要段落；

2） 在获取到相关的条目时，本文提取该条目的“简介”或“概述”部分作为上下文内容，为后续缩略词预测任务提供背景知识。

（2） 使用生成式语言模型补充上下文信息

尽管百度百科提供了大量的实体信息，但并非所有数据集中的实体都能够在其数据库中找到相关条目，尤其是一些领域特有的缩略词或较为冷门的实体。因此，为了进一步扩充数据集的上下文信息，本文引入了生成式语言模型（GPT-3.5)来补充缺失的上下文。

1） GPT3.5 生成上下文信息

当百度百科中无法找到某个实体的相关条目时，本文通过 GPT3.5 来进一步补充上下文数据。GPT3.5 在训练过程中接触了大量的文本数据，涵盖了广泛的主题和领域。这使得它能够生成关于各种实体的相关描述，提供准确和详细的信

息。因此，本文通过调用 OpenAI 提供的 GPT-3.5 接口，利用提问来补充上下文信息。

2） 控制生成内容的质量

为了避免生成的上下文信息过于单一或片面，本文设计了多轮对话式提问的机制。每个实体的上下文生成不仅依赖于一次提问，而是通过多个问题的轮次逐步引导模型，获取更为详细和多样化的信息。通过设计连续性强的问题序列，确保生成的上下文内容能够涵盖实体的多个方面，避免局限于单一的定义或描述。此外，为了确保生成内容的权威性和代表性，本文还通过反向提问来进一步验证生成内容的准确性。

（3） 上下文信息处理

为确保所收集的上下文信息在后续实验中能够有效利用，避免冗余信息的干扰，本文对收集到的上下文进行了系统的处理。

1） 为了防止数据泄露以造成对实验结果的干扰，针对每个三元组条目（实体全称，缩略词，上下文信息），对上下文中包含基本事实缩略词的句子进行删除。

2） 考虑到长文本输入带来的时空复杂度问题，本文采用了截断策略来缩短上下文并保留其基本位置。首先，对于每个上下文找到实体全称出现的句子。如果实体全称不在上下文中，则定位上下文的第一个句子。其次，以定位的句子为中心实施双向扩展，逐句扩展定位的句子直至满足预设的上下文窗口约束条件[min_length，max_length]。min_length 和 max_length 均为超参数。至此，所有定位的句子都被保留为上下文。

通过上述处理，本文成功获取并补充了完整的中文缩略词数据集 CED（Context Enhanced Dataset）。该数据集涵盖了丰富的上下文信息，以确保其在后续研究中的有效性和可靠性。

3.4 本章小结

本章首先明确了中文缩略词预测任务的设计背景，重点强调了上下文信息在该任务中的至关重要性。随后，基于已发布的基准数据集，详细阐述了数据集的构建过程，特别是通过 CN-DBpedia 和百度百科两个数据源进行信息收集，以确保数据集的多样性和准确性。在此基础上，针对数据集缺乏充分上下文信息的问

题，本文提出并实施了双重策略：一方面通过百度百科获取上下文信息，另一方面利用生成式语言模型（GPT-3.5）补充缺失的上下文内容。此外，考虑到长文本带来的时空复杂度问题，本文对上下文信息进一步处理。最终得到涵盖了丰富上下文信息的中文缩略词数据集 CED，为后续研究提供了坚持的数据支持。

第 4 章 基于上下文增强的中文缩略词生成

本章提出了一种基于上下文增强的中文缩略词生成方法，旨在提高中文缩略词预测的性能。4.1 节分析了当前中文缩略词预测方法存在的局限性，并针对性地提出了优化解决方案。4.2 节明确了中文缩略词预测的核心目标，并基于任务需求给出了形式化的定义，为后续的研究工作提供了理论基础。4.3 节详细介绍了模型的整体架构，模型由上下文增强编码模块和解码模块组成。通过对每个模块的功能与工作原理的深入分析，阐述了如何通过这些模块提升模型的预测精度。4.4 节对比了本章提出的模型与现有的主流方法，首先界定了研究问题，并详细说明了实验环境、对比方法及其参数设置；然后，展示了模型与对比方法在标准数据集上的实验结果，并对其进行了详细的对比分析。为了进一步验证模型的效果，还引入了人工评估，分析了评估结果，得出结论：基于上下文增强的中文缩略词生成方法，相较于现有的对比方法，显著提升了中文缩略词预测任务的性能。

4.1 研究背景

在自然语言处理领域，缩略词的生成一直是一个复杂的任务。尤其在中文环境中，由于语言自身的形态学特性，缩略词的构造往往并非简单的字符截取，而是涉及多层次的语义抽象、信息浓缩以及高度依赖上下文的推理过程。随着信息化时代的深入，中文文本中缩略词的使用频率激增，涵盖领域广泛。这不仅使中文缩略词生成任务变得愈加重要，也使得该任务面临更为严峻的挑战。

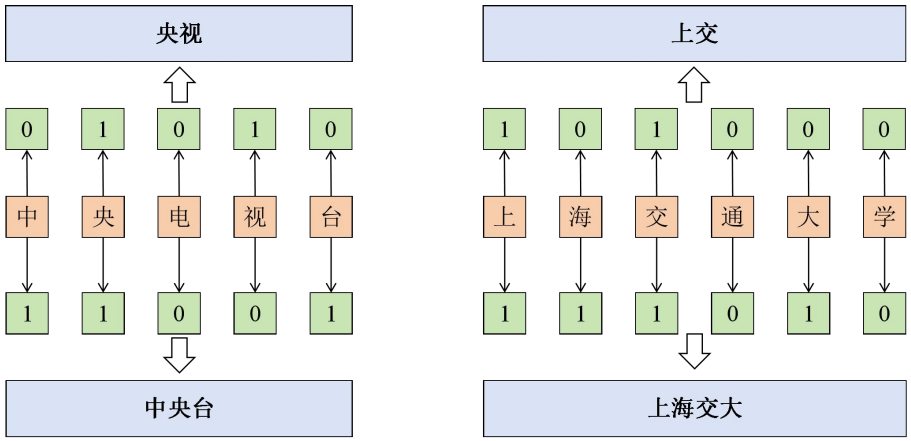

图 4-1 序列标记方法示例

Figure4-1 Example of sequence labeling methods

如图 4-1 所示，大多数中文缩略词预测任务通常被建模为序列标记任务，其核心思路是对给定实体中的每个字符进行标记，指示该字符是否应当被保留在最终的缩略词中，或者在缩写过程中予以剔除。尽管基于特征的序列标注方法在中文缩略词预测任务中取得了一定进展，但这些方法依然存在明显的局限性。它们过于依赖当前字符的局部信息来进行标签预测，缺乏对先前字符标签信息以及整个输入序列全局上下文的考虑。这一局限性意味着，在实际预测过程中，模型只依赖于已经扫描过的字符的局部上下文，而未能有效挖掘实体前后文中的潜在信息。然而，实体的前后文，包括相邻单词、短语乃至篇章层面的语义关系，可能直接影响实体的缩略形式。传统的序列标注模型未能充分利用这些潜在的关联信息，这就导致了在处理具有复杂语境依赖性的中文缩略词时，模型的预测精度受到显著限制。

为了克服这一问题，本文提出将中文缩略词预测任务转变为序列生成问题，并结合序列生成模型与丰富的上下文信息，期望能突破传统方法的瓶颈。与传统的基于局部信息的序列标记方法不同，序列生成方法在生成过程中能够动态地处理和利用全局上下文。这种方法不仅能够在生成时充分考虑到上下文的影响，还能通过引入多维度的上下文信息，为模型提供更多的语义线索，从而帮助其更准确地理解并运用全局语境。在此过程中，模型能更好地捕捉到与缩略词生成相关的语义信息，提高生成结果的准确性。通过这种创新，模型不仅能够提升对缩略词生成的精确度，还能够增强对复杂语境的认知能力和泛化能力，从而在实际应用中展现出更为优异的预测效果。

4.2 模型目标

中文缩略词预测任务的核心目标在于对源实体完整形式可能的缩写进行准确预测。具体而言，考虑一个包含 n 个字符的源实体的全称表示为 $X=[x_1,x_2,\ldots,x_n]$，其中字符 $x_i(1\le i\le n)$。同时考虑长度为 m 与该实体相关的文本上下文信息为 $D=[d_1,d_2,\ldots,d_m]$，其中 m 为上下文窗户大小，字符 $d_i(1\le i\le m)$。在中文缩略词预测任务中，目标是预测一个有效的缩写序列 $Y=[y_1,y_2,\ldots,y_k]$，其中 k 表示缩写长度，满足 $y_j\in X$，$1\le j\le k$。这一问题可形式化为：

$$\hat{Y} = \arg\max_{Y}[y_1, y_2, ..., y_k] \tag{4-1}$$

其中，$\hat{Y}$ 表示预测的最优缩略词序列。

本文的任务是通过序列生成模型，充分利用上下文信息，提高对缩略词的准确预测。通过优化模型，目标是确保生成的缩略词序列能够准确且精确地反映源实体的完整语义。在这个任务中，本文关注的是通过综合全称和相关上下文信息来提高预测准确性，使得生成的缩略词能够更贴切地表达源实体的完整含义。这种方法旨在克服传统方法在捕捉语境信息方面的局限性，从而提高中文缩略词预测任务的性能和实用性。

4.3 模型结构

模型的整体框架如图 4-2 所示。首先，本文引入了上下文增强编码器，该编码器接受源实体的完整形式 X 及其相关文本（上下文信息）D 的标记序列作为输入。为了确保模型能够高效地捕捉字符间地深层语义关系，本文利用与 BERT 一致的初始化操作获得输入标记的初始嵌入和位置嵌入，两者共同被输入到一个编码器块中，从而生成一组关键的特征表示。这些特征表示对输入标记包含的重要信息进行编码，其中一部分信息将在后续的解码过程中发挥关键作用。

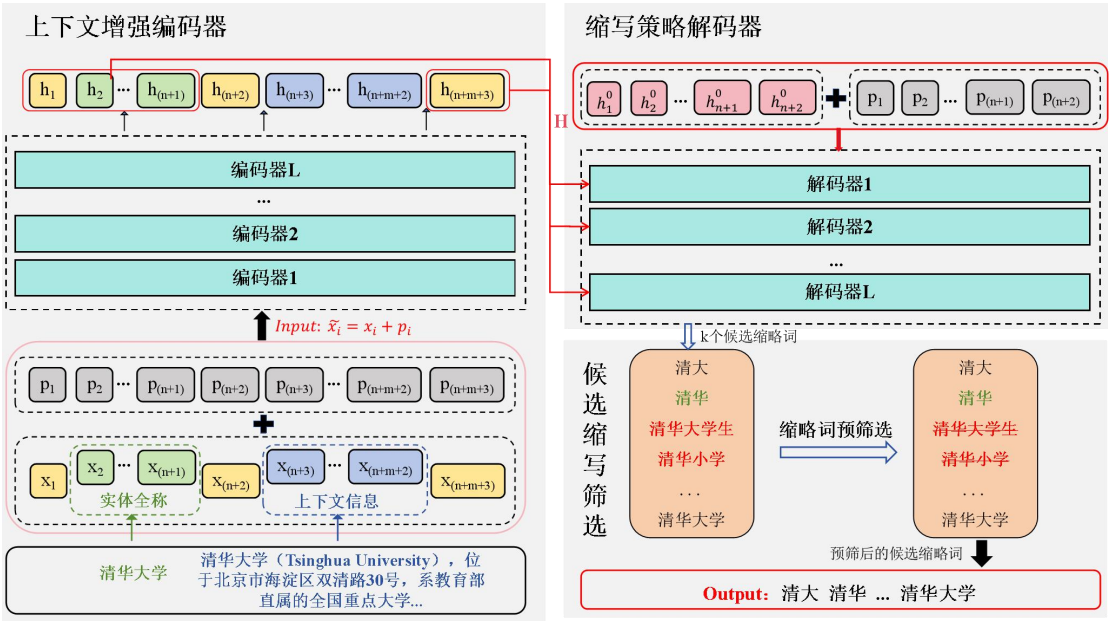

图 4-2 模型结构图

Figure4-2 Model structure diagram

其次，为了增强模型的生成能力，本文采用了一个预训练的中文BARTTransformer CPT[80]的解码部分。CPT 作为一种适用于中文自然语言理解（NLU）和生成（NLG）任务的预训练语言模型。它采用了不平衡的 Transformer 架构，包括一个用于建模通用语义特征被称为共享编码器（S-Enc）的 Transformer 编码器，一个用于自然语言理解任务的浅层理解解码器（U-Dec），以及一个用于文本生成任务的浅层生成解码器（G-Dec）。在预训练阶段，CPT 采用掩蔽语言建模（MLM）和去噪自动编码（DAE）任务进行预训练，以增强模型的上下文理解能力。并采用 CPT 的生成解码器 G-Dec，利用输入序列的高维特征，逐步生成目标缩略词序列 Y。

最后，由于缩略词必须严格由原实体的字符组成，为了确保生成结果符合约束，本文进一步引入规则处理模块，对生成的缩略词进行筛选。首先检查生成的缩略词是否完全来源于原实体字符，若包含非实体字符，则剔除该候选结果；此外，对于可能存在的冗余字符或无效输出，利用启发式规则进行调整，以确保最终缩略词符合中文缩写习惯。接下来将对三个模块进行详细描述。

4.3.1 上下文增强编码

该单元旨在获取输入标记序列的连续特征表示，以在后续解码过程中预测缩略词序列。除了源实体的全称和相关文本的标记外，本文还添加了特殊标记 $[BOS]$ 和 $[EOS]$，分别代表着输入序列的开始和结束位置，以及引入 $[SEP]$ 来分隔 X 和 D。输入序列可表示为：

$$S=[BOS]\oplus X\oplus[SEP]\oplus D\oplus[EOS] \tag{4-2}$$

如图 4-2 所示，本文为每个输入标记执行初始化嵌入发的操作，即将初始化偏置设为 0，并从 $N(0,0.02)$ 中采样权重，其中层归一化参数为 $\beta=0$，$\gamma=1$。同时，为了准确捕捉输入文本的时序模式，每个标记的初始嵌入都添加了一个位置嵌入，然后将它们输入到 Transformer 编码器块中。具体来说，对于第 i 个标记 $1\le i\le n+m+3$，$x_i\in R^{de}$ 表示其字符嵌入，它将每个输入标记映射到一个低维的向量空间中，使得具有相似语义的标记在向量空间中距离更近。位置嵌入为 $p_i\in R^{de}$。其目的是为了编码输入序列的顺序信息。因为模型本身是基于注意力

机制的，它对输入序列的顺序不敏感，而位置嵌入能够为模型提供关于标记在序列中位置的信息，从而让模型能够更好地捕捉序列的上下文信息。每个标记的初始嵌入由其字符嵌入与位置嵌入之和表示：

$$\tilde{x_i} = x_i + p_i \tag{4-3}$$

接下来将这些嵌入输入至编码器中，该编码器由 L 个 Transformer 块组成，每个块都包含 8 个自注意力机制头。正如前文所述，由于 Transformer 的特性，每个标记的表示不仅包含自身信息，还编码了除自身之外的所有其他输入标记的信息。因此，基于 Transformer 的编码器有助于使模型充分利用所有输入文本信息，以便进行准确的预测。具体而言[32]，假设由编码器输出的第 i 个标记的特征表示为 $h_i \in R^{de}$，则编码过程可以表示为：

$$h_1, h_2, \ldots, h_{n+m+3} = Transformer_{1\sim L}(x_1, x_2, \ldots, x_{n+m+3}) \tag{4-4}$$

其中，每个 h_i 表示对应输入标记的编码表示。由于 Transformer 的自注意力机制，每个标记的表示不仅包含其自身信息，还编码了整个输入序列的全局信息。这使得模型能够更好地理解输入文本的语义和上下文关系，从而为后续的缩略词预测提供更丰富的特征。

然而，由于每个 h_i 在某种程度上编码了整个输入文本信息，而过多的特征表示可能引入噪音，因此，为了降低噪声对后续解码的影响，本文仅将全程标记 $h_1, h_2, \ldots, h_{\mathrm{n+1}}$ 以及 $h_{\mathrm{n+m+3}}$ 传递给解码器模块，以供后续使用。换句话说，仅全称标记的特征表示以及[BOS]和[EOS]的特征表示，这些被表示为 H。

$$H = [h_1, h_2, \ldots, h_n, h_{n+m+3}] \tag{4-5}$$

其中，H 作为上下文增强特征，将被输入至解码器进行缩略词生成。通过这种特征筛选的方式，能够聚焦于与缩略词生成最相关的信息，提高模型的预测准确性和效率。

4.3.2 缩略解码

本文使用符号 *Embedding*() 表示 CPT 的嵌入层，该层将输入序列转换为 Transformer 块的输入向量。使用符号 *TransEnc*() 表示 Transformer 编码器块，符号 *TransDec*() 表示 Transformer 解码器块。在给定输入 X ， D 的情况下，CPT 的编码过程可被描述为：

$$H_0^e = Embedding(X, D) \tag{4-6}$$

$$H_i^e = TransEnc(H_{i\text{-}1}^e), i \in [1, M] \tag{4-7}$$

其中，M 表示 Transformer 编码器块的数量，H_i^e 表示 Transformer 编码器的第 i 层的隐藏状态。在获得 H_M^e 之后，解码过程的第 t 个时间步可以被描述为：

$$H_0^d = EmbStart(\hat{y} < t) + p_t, \quad t \in [1, n+2] \tag{4-8}$$

$$H_t^{d,1} = TransDec(H_t^0, H_M^e) \tag{4-9}$$

$$H_t^{d,l} = TransDec(H_t^{d,l-1}, H_M^e), \quad 2 \le l \le N, N \le M \tag{4-10}$$

其中，N 表示 Transformer 解码器块的数量，$H_t^{d,l}$ 表示 Transformer 解码器的第 l 层的隐藏状态，p_t 表示位置嵌入信息，H_t^0 是当前时间步输入的初始嵌入，包括前一轮输出的 token 及其位置嵌入。

最终，解码器的输出通过一个分类器进行标记预测：

$$\hat{y}_t = g(H_t^{d,N}) \tag{4-11}$$

其中，$g()$ 表示一个基于 softmax 的分类器，用于以 $H_t^{d,N}$ 作为输入的语义表示中选择一个标记。$\hat{y}_t$ 代表输出序列的第 t 个标记，$\hat{y} < t$ 表示输出序列第 t 个标记之前的标记。因此，模型可以生成输出序列 Y 。

4.4 实验与分析

为验证本章提出方法的性能，本节首先介绍了实验设置，接着在真实数据集上分别开展了实验，并对实验结果进行分析。

4.4.1 实验设置与评价指标

对于实验部分，本文使用 jieba 库作为分词算法。在预训练阶段，本文使用 Adam W 优化器在数据集上进行微调训练，训练参数如下：默认批处理大小设置为 128，学习率为 2e-5，训练轮次设置为 10 轮。为防止过拟合并加速模型收敛，本文利用 Dropout 操作进行优化，使用 AdamW 优化器最小化模型损失，在每批次操作后进行参数规范以提高模型性能。

为准确有效地评价本文模型的性能表现，本文在实验中使用了命中率（Hit）作为评价指标。如果测试样本的预测缩写正好是其实际缩写，则将该样本视为命中样本（Match）。命中率是命中样本在所有测试样本中的比例。较高的命中率表示模型可以更接近每个实体的真实缩写。

4.4.2 性能比较

为了评估模型的有效性，本文基于 CED 数据集，将方法与若干基线模型进行比较。基线模型有序列标记模型和序列生成模型。其中，序列标记模型包括：CRF，RNN-RADD，Bi-LSTM+CRF，BERT+Bi-LSTM+CRF。序列生成模型包括：Attention-based Seq2seq，Wang et.al[22]，Transformer，CPT。

（1） CRF[83]：CRF 是一种经典的机器学习模型，其是给定随机变量 X 的条件下，随机变量 Y 的马尔可夫随机场。广泛应用于序列标记、实体识别等任务。在缩略词生成任务中，它被用来为给定的全称生成候选缩写。该方法通过引入手写特征来辅助生成。

（2） RNN-RADD：RNN-RADD 方法将 RNN 与用于建模动态字典（RADD）的机制相结合。动态字典机制有助于弥补单个汉字表示的不足，提高对上下文信息的建模能力.

（3） Bi-LSTM+CRF：该模型基于 Bi-LSTM 以及 CRF 联合。并在联合模

型中，使用了词分割和词性信息作为特征。

（4） BERT+Bi-LSTM+CRF：该模型首先使用 BERT 获取全称中每个字符的预训练向量，然后将其输入到双向 LSTM 中，最后使用 CRF 执行中文缩略语预测任务。

（5） Attention-based Seq2seq：基于注意力的 Seq2seq 是一种典型的生成模型，使用编码器-解码器结构。在中文缩略语生成中，它试图使用注意力机制来更好地捕捉输入序列和输出序列之间的关系。

（6） Transformer：Transformer 是一种经典的编码器-解码器模型，广泛应用于生成任务。它通过自注意力机制有效地捕捉输入序列中的长距离依赖关系，对于中文缩略语生成等任务具有很好的性能。

（7） Wang et.al[22]：首先尝试将中文缩略语预测任务形式化为序列生成问题，并提出了一种新的 seq2seq 模型和一种多级预训练模型。

（8） CPT-base：兼顾理解和生成的中文预训练模型，在许多 NLG 任务上显示出有效性。

表 4-1 实验结果

Table4-1 Experimental Result

模型	Hit
CRF	39.4
RNN-RADD	42.5
Bi-LSTM+CRF	43.4
BERT+Bi-LSTM+CRF	44.9
Attention-based Seq2seq	32.8
Transformer	34.6
Wang et.al[22]	47.6
CPT-base	46.5
Our Method	48.3

所有比较方法的总体结果如表 4-1 所示。可以看出本文方法显著优于基线模型。实验结果充分表明了基于上下文增强的中文缩略词预测方法在中文缩略词预测任务上的有效性。更具体地，在该公开数据集上，本文模型 Hit 值比基线模型的最佳结果提高了 0.7 分，优于传统方法。这说明本文使用预训练语言模型可以有效捕获实体和上下文信息的语义信息，提升中文缩略词预测的效果。当本文引入上下文信息时，与 CPT-base 相比，性能显著提高，证明了上下文信息对中文缩略词预测任务的有效性。

4.4.3 基于上下文信息长度的参数实验

在本文提出的中文缩略词生成方法中，上下文信息采集为关键环节。从机理层面分析，扩展上下文信息的覆盖范围有助于捕捉潜在的关联特征，但同时也面临着信息冗余和噪声的干扰。另一方面，过长的上下文可能使模型难以聚焦于关键信息。反之，过度压缩上下文窗口虽然能够规避噪声污染，却可能造成关键语义单元缺少，信息不饱和。

为了深入探讨不同上下文长度对模型性能的影响，本文进行了针对上下文长度的参数实验。由于计算能力的限制，本文将上下文长度限制到 200 以内。如图 4-3 所示，当上下文长度在 0-150 区间时，模型命中率呈现稳定上升趋势。而过长的上下文（150-200）则导致了命中率的下降。同时，当上下文长度为 0 时，即未使用任何上下文信息，模型的性能也显著下降。再次验证了上下文信息对本研究的重要性。

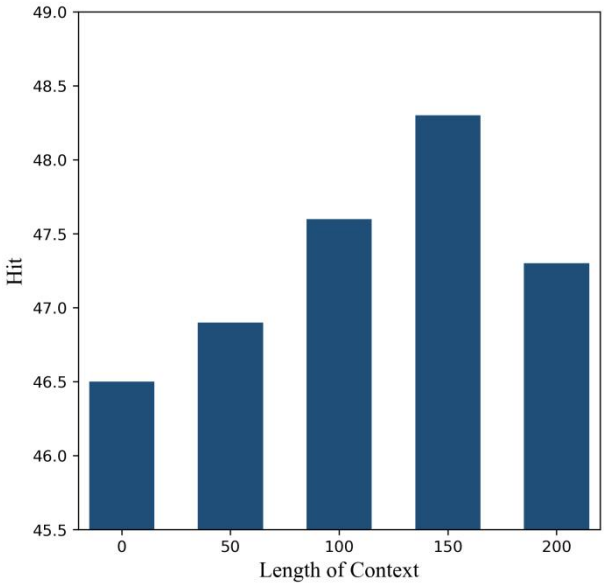

图 4-3 上下文长度对模型命中率的影响

Figure 4-3 The effect of context length on the model's hit rate

4.4.4 人工评估

为了更公平和准确的比较，进一步考虑了以下指标，这些指标基于测试集中随机抽取的 1000 个样本分别进行人工评估计算出来的。

（1） *CC*：具有缩略词特性，能够有效表达实体原始形式，但不是数据集的唯一注释缩略词，则表示为*CC* 。

（2） *CE*：具有正常的语言结构，但不能有效表达实体原始语义的缩略词，

则表示为*CE*。

（3） *EE*：不能有效表达实体原始语义且语言结构异常的缩略词，则记为*EE*。

具体而言，*CC*、*CE*、*EE*三种类型举例如表 4-2 所示。

表 4-2 缩略词预测的 4 种类型

Table4-2 Four types of abbreviation prediction

类型	完整形式	正确缩略词	预测缩略词
Match	北京大学	北大	北大
CC	中央电视台	央视	中央台
CE	奥林匹克运动	奥运	奥林匹克
EE	华中师范大学	华师/华中师大	华范

对于测试集随机抽取的 1000 个样本的人工评估结果见表 4-3。由表 4-3 可知，在 1000 个随机样本中，样本命中率相对较高，这代表模型在一半以上的样本中成功预测了唯一的注释缩略词。其次，R_{CC}值均高于R_{CE}、R_{EE}，说明模型在一定程度上倾向于生成新的缩略词，而不仅仅是在训练集中见过的缩略词，这可能有助于模型更好地捕捉数据集中的潜在规律。另外，结果中R_{CE}占有一定比例，说明模型在语义理解方面还存在着挑战。

表 4-3 人工抽样结果

Table4-3 Manual sampling results

Judges	Hit	R_{CC}	R_{CE}	R_{EE}
1	52.4%	23.7%	20.6%	3.3%
2	52.1%	24.3%	23.2%	4.0%
3	51.9%	25.1%	19.3%	3.7%
Average	52.1%	24.4%	21.0%	3.7%

4.5 本章小结

在本章，研究了一种基于上下文增强的中文缩略词预测方法，旨在解决现有研究在处理中文缩略词时的局限性。该方法将中文缩略词预测任务重新定义为序列生成问题，并结合预训练语言模型和复杂的上下文信息，有效增强了实体全称及其上下文的嵌入表示，从而捕获了实体的深层语义信息。在文章的最后，通过对比实验验证了该文所提方法的有效性，并通过基于上下文长度的参数实验消融实验说明了模型中上下文信息采集的重要性。

第 5 章 基于上下文主题一致性的中文缩略词预测

本章提出了基于上下文主题一致性的中文缩略词预测方法，以提升中文缩略词的预测性能。5.1 节讨论了传统单阶段生成模型在中文缩略词预测中的局限性，重点探讨了其在候选集质量和语义关联性方面的不足之处。5.2 节明确了中文缩略词预测的核心目标，即准确预测源实体的缩写，并将该任务建模为一个优化问题。5.3 节首先介绍了方法的整体框架，包括候选生成阶段和上下文主题一致性评估阶段，并详细阐述了各个模块的功能和工作原理。5.4 节则将本章提出的方法在公开数据集上进行了一系列实验，并对实验结果进行对比分析。

5.1 研究背景

随着自然语言处理任务对中文缩略词生成准确性的要求不断提高，传统的单阶段生成模型正逐步暴露出其固有的瓶颈，尤其是在候选集质量和语义关联性方面的缺陷愈发明显。单阶段方法往往依赖固定规则或简单的统计策略来直接生成缩略词，这种方式虽然计算高效，但在真实应用场景中却存在显著的局限性。首先，由于缺乏对语境的深层建模，所生成的候选词极易受到字符表面特征的误导，导致生成质量不稳定——既可能遗漏高度概括的优质缩略词，又可能生成冗余甚至不合理的词汇。此外，单阶段方法在处理复杂语言现象时显得力不从心，尤其是面对多义词、同义词及其在特定语境下的细微变体时，往往无法精准捕捉其与全称之间的微妙语义关系。例如，“人工智能”在不同场景下可能缩略为“AI”或“人智”，但传统方法缺乏对语境的动态理解，难以判断哪种缩略形式更为自然、规范。又如“国家体育总局”可能缩写为“国体”，也可能缩写为“体育总局”，而单阶段方法因缺乏充分的上下文对比信息，常常在多个可选方案中难以做出合理决策。由此可见，单阶段模型的缺陷主要体现在：

（1） 候选生成质量不稳定：容易遗漏合理候选，或生成不符合语言规范的缩略词；

（2） 语义匹配能力不足：无法精准捕捉缩略词与全称之间的深层语义关联；

（3） 对多义性与上下文变化的适应性差：无法灵活调整缩略策略，以适应不同语境需求。

因此，为了应对这些挑战，本文创新性的提出了一种两阶段优化的中文缩略词预测框架，从“候选生成”和“上下文一致性评估”两大核心环节入手，借助预训练语言模型的强大表征能力，结合对比学习的惊喜筛选策略，最终实现更高质量、更具语义一致性的缩略词预测。其核心借鉴了检索增强生成（Retrieval-Augmented Generation，RAG）的分阶段优化理念[82]。第一阶段候选生成采用基于 Transformer 的编码器-解码器架构，通过预训练语言模型的序列生成能力，在确保语义完整性的前提下，生成高覆盖率的缩略词候选集，尽可能多地捕捉潜在地优质缩略词选项。第二阶段上下文主题一致性评估引入对比学习框架，通过构建正负样本训练语义匹配模型，对候选缩略词进行筛选，确保最终输出的缩略词既符合语境需求，又具备高度概括性和可读性。

具体而言，本研究将中文缩略词预测任务视为序列生成任务，并巧妙借助预训练语言模型卓越的表征学习能力，深入挖掘上下文中的隐含信息，从而生成既符合语言习惯又高度概括的候选缩略词。在此基础上，通过引入精心设计的评估模型，对候选缩略词进行二次筛选，选取出可用性强、流程性高的缩略词，进一步提升了中文缩略词预测的准确性和鲁棒性。

5.2 模型目标

中文缩略词预测任务的核心目标在于对源实体完整形式可能的缩写进行准确预测。具体而言，考虑一个包含 n 个字符的源实体的全称表示为 $X=[x_1,x_2,...,x_n]$，其中 n 表示全称长度，x_i 为字符集合，同时考虑与该实体相关的 m 个字符的文本上下文信息为 $D=[d_1,d_2,...,d_m]$，其中 m 为上下文窗户大小。在中文缩略词预测任务中，目标是预测一个有效的缩写 Y。

本文将缩略词预测任务建模为一个优化问题，旨在最大化生成的缩略词与完整实体形式之间的语义一致性。具体而言，本文定义一个 C_{cand} 为候选缩略词集合，包含 k 个可能的缩略词，即 $C_{and}=\{C_1,C_2,...,C_k\}$。$Sim(c_i,X)$ 表示候选缩略词 c_i 与完整实体形式 X 之间的语义相似度，采用余弦相似度进行计算。θ 为主题一致性阈值，用于评估候选缩略词与上下文主题的一致性。通过以下优化目标来选择最佳的缩略词：

$$\hat{Y} = \arg \max_{c_i \in Ccand} Sim(c_i, X) , Sim(c_i, X) \geq \theta \tag{5-1}$$

其中，$\hat{Y}$ 表示最终预测的缩略词。

在基于上下文主题一致性的中文缩略词预测方法中，目标是根据候选缩略词在当前上下文信息中与原始实体的主题一致性。为此，本文为所有候选项生成一个一致性排名列表，并根据模型对每个候选缩略词的主题一致性预测，排序所有项目。在此排名列表中，排名第一的候选缩略词则会作为最终的缩略词预测结果。通过这种方式得到的结果更为准确并更贴近原实体语义。

5.3 模型结构

本文提出一种基于上下文主题一致性的中文缩略词预测方法，模型框架如图 5-1 所示，该模型主要包含两部分：候选生成阶段和上下文增强评估阶段。

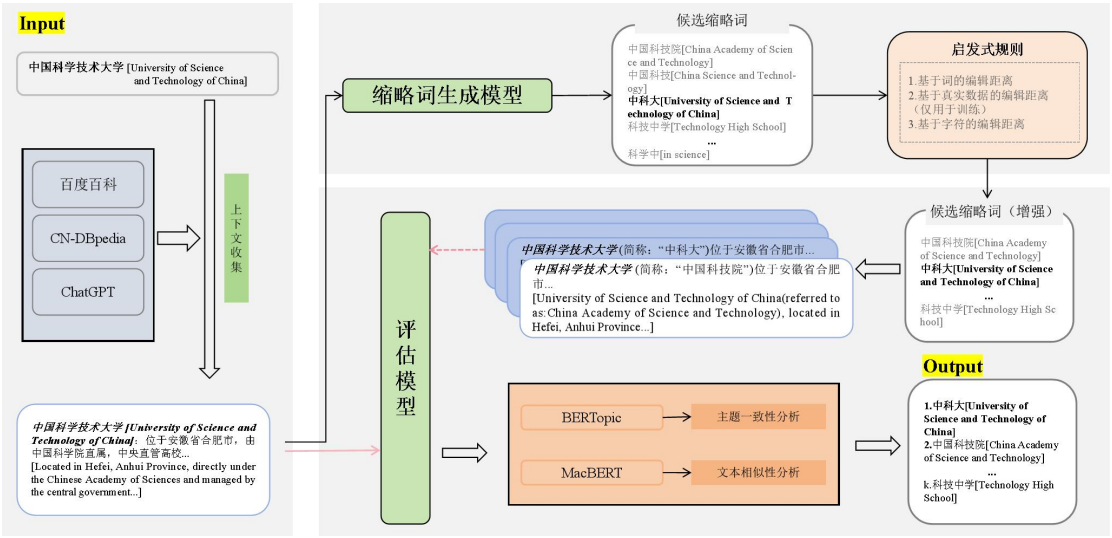

图 5-1 模型结构图

Figure5-1 Model structure diagram

（1）候选生成阶段：此阶段基于 BART Transformer CPT 预训练模型获取文本语义信息，从而生成候选缩略词。具体而言，给定一个全称 X 和其相关的上下文信息 D，首先将 (X, D) 输入到模型中，训练模型以生成 k 个候选缩略词 $Y = \{y_1, y_2, ..., y_k\}$。此外，本文利用一组启发式规则对于候选缩略词预先筛选以

生成一组增强候选缩略词$\hat{Y}=\{\hat{y}_1,\hat{y}_2,...,\hat{y}_k\}$。

（2）上下文主题一致性评估阶段：在此阶段中，本文从主题一致性的角度出发来评判每个候选缩略词的合理性。具体而言，给定(X,D)和$\hat{Y}=\{\hat{y}_1,\hat{y}_2,...,\hat{y}_k\}$，为了增强候选集$\hat{y}$的上下文信息，首先对于不包含实体全称$X$的上下文$D$，本文在上下文$D$段首增加“$X$”，为此确保所有上下文中都包含$X$。其次，用“（简称：“$\hat{y}_i$”）”进行填充，将$X$右侧填入“X（简称：“$\hat{y}_i$”）”。这样可以构造出上下文增强版本$\hat{D}=\{\hat{D}_1,\hat{D}_2,...,\hat{D}_k\}$，其中每个$\hat{D}_i$是在上下文 D 中用候选缩略词$\hat{y}_i$替换原实体$X$后的版本。在得到$\hat{D}$后，本文使用 BERTopic 主题模型分别对原始上下文D和增强后的上下文$\hat{D}$进行主题建模。分别得到f个主题$T=\{t_1,t_2,...,t_f\}$和$\hat{T}=\{\hat{t}_1,\hat{t}_2,...,\hat{t}_f\}$。接着，使用对比学习训练评估模型，分别对$T$和每个$\hat{t}_i$进行编码，并利用它们之间的余弦相似度作为$\hat{y}_i$的评分。这样，评估模型对于与原实体主题一致性更高的候选缩略词给予更高的评分。最终，根据对应的评分从$\hat{y}$中选择出最终输出的Y。

5.3.1 候选生成阶段

候选生成阶段，旨在生成多个高质量的候选缩略词，以确保命中率。为此，本文以一个预训练的中文 BARTTransformer CPT 模型为基础进行候选缩略词的生成。CPT 作为一种适用于中文自然语言理解（NLU）和生成（NLG）任务的预训练语言模型，它采用了不平衡的 Transformer 架构，包括一个被称为共享编码器（S-Enc）的 Transformer 编码器，一个用于理解的浅层 Transformer 编码器（U-Dec），以及一个用于生成的浅层 Transformer 解码器（G-Dec）。两个具有共享编码器的解码器分别通过掩蔽语言建模（MLM）和去噪自动编码（DAE）任务进行预训练。

当给定的候选缩略词数量k逐渐增大时，候选缩略词中可能会出现不属于原短语子序列的字符。这些候选缩略词很明显不能成为有效的缩写。因此，这些缩

写需要过滤掉。此时，为了补足 k 个候选缩略词，本文利用一系列启发式规则[24]来构造复杂负样本，以得到最终的 k 个增强候选缩略词。

（1） 基于词的编辑距离（word-based rule）：在中文中，词是最小的语义单位，从 X 中删除词后并不能影响原文本的流畅性。因此可以通过从完整形式 X 中删除少数词来生成负样本，这样既能保持负样本的流畅性，又能保证其意义与原实体不一致。

（2） 基于真实数据的编辑距离（truth-based rule）：由于在训练阶段可以获得真实数据，因此与真实数据相似的候选缩略词可以被认为是负样本（仅用于训练）。

（3） 基于字符的编辑距离（character-based rule）：与原始短语 X 在字面上相似的候选项也可构成负样本，因为它们保留了原始短语中的一部分字符。

最终，在过滤掉不合理的候选项并通过启发式规则构造复杂负样本后，就获取了最终用于上下文主题一致性评估阶段的 k 个增强候选缩略词。

5.3.2 上下文主题一致性评估阶段

第二阶段是上下文主题一致性评估阶段，其目标是为了评估候选缩略词在其上下文的准确性。在此阶段中又可分为主题词生成和主题一致性评判两个阶段。

（1） 主题词生成

为了揭示文本中的共同主题和潜在的信息，主题模型已被证实是一个强大的无监督工具。在此阶段中，本文使用 BERTopic 主题模型分别对 D 和每个 $\hat{D}_i$ 进行主题分析。BERTopic 能够更好地捕捉文本的深层语义信息，使得主题生成更具相关性和准确性，尤其在捕捉长文本或复杂句子之间的关系时表现优异。其主要通过预训练模型进行文本嵌入、对嵌入进行降维及聚类、主题表示三个步骤进行主题挖掘[83]。

1） 文本嵌入构建

在 BERTopic 中对文档进行嵌入，以在向量空间中创建表示，这些表示可以在语义上进行比较。BERTopic 采用 Sentence-BERT 框架，该框架允许用户使用预训练语言模型将句子和段落转换为密集的向量表示，并在各种句子嵌入任务中实现了先进的性能。给定一组文本数据 D，每个文本 d_i 被映射到一个嵌入空间中，生成嵌入矩阵 E：

$$E = \text{BERT}(D) \in \mathbb{R}^{m \times d} \tag{5-2}$$

其中 m 是文本的数量，d 是嵌入的维度。

2） 嵌入降维及聚类

在高维向量空间中，空间局部性的概念变得模糊，距离测量差异较小。BERTopic 中使用 UMAP[84]降维算法，降维后的嵌入表示为 E_{UMAP}，其在较低纬度上保留了更多的高维数据的局部和全局特征，同时经过降维处理后向量的聚类效果也将会得到提升。

$$E_{UMAP} = \text{UMAP}(E) \in \mathbb{R}^{n \times k} \tag{5-3}$$

其中，k 是降维后的维度。

接着，使用 HDBSCAN[85]算法对降维后的向量进行聚类，该算法对噪声点具有良好的鲁棒性，在聚类过程中可以识别异常值并进行排除，以此提高聚类性能。聚类的结果为每个文本分配了一个主题标签 T：

$$T = HDBSCAN(E_{UMAP}) \tag{5-4}$$

3） 研究主题挖掘与表征

基于聚类所得的句向量类团，使用 c-TF-IDF 算法对类团进行主题表示，提取主题标签词。c-TF-IDF 对 TF-IDF 算法进行了改进，它考虑了词在特定类别（即聚类）中的重要性，选择每个聚类中的代表性词形成主题，能够更加准确地表征聚类主题。

$$W_{t,c} = tf_{t,c} \cdot \log\left(1 + \frac{A}{tf_t}\right) \tag{5-5}$$

其中，词频表示词 t 在类 c 中的频率。在此实例中，类 c 是每个聚类中的文档连接而成的单个文档。然后，逆文档频率被替换为逆类频率，以衡量一个词对类提供了多少信息。它通过将每类的平均词数 A 除以所有类中词 t 的频率再取对数来计算。为了输出正值，本文在对数中的除法中加一。因此，这种基于类的 TF-IDF 过程表示的是词在聚类中的重要性，而不是单个文档。这使得能够为每个文档聚类

生成主题词分布。最后，通过迭代合并最不常见主题与其最相似的主题的c-TF-IDF表示，并且可以将主题数量减少到用户指定的值。

（2） 主题一致性评判

评估模型的目标是根据原上下文信息和候选上下文信息，使得有效的候选缩略词获取比无效候选缩略词选项更高的评分。为此，本文引入主题一致性评判。

给定T和$\hat{T}_i$，本文使用中文预训练Transformer编码器MacBERT[86]分别对T和$\hat{T}_i$进行编码，并使用$[CLS]$标记的嵌入作为它们的表示，如方程所示：

$$h_T = MacBERT(T)[CLS] \tag{5-6}$$

$$h_{\hat{T}_i} = MacBERT(\hat{T}_i)[CLS], \hat{T} = \{\hat{T}_i\}_{i=1}^{k} \tag{5-7}$$

其中，h_T表示T的嵌入，$h_{\hat{T}_i}$表示$\hat{T}_i$的嵌入。由于在训练阶段真实数据总是处于候选缩略词中，因此本文将原上下文信息生成的主题词T作为正样本，记作$\hat{T}^+$。而其它$k-1$个候选缩略词视为负样本。$\hat{T}^+$的嵌入表示为$h_{\hat{T}^+}$，其中$h_{\hat{T}^+} \in \{h_{\hat{T}_i}\}_{i=1}^{k}$。接着通过以下方程中描述的对比损失来训练评估模型，其中$sim()$表示余弦相似度，τ表示温度超参数：

$$L_{eval} = -\log \frac{e^{sim(h_T, h_{\hat{T}^+})/\tau}}{\sum_{f=1}^{e} e^{sim(h_T, h_{\hat{T}_i})/\tau}} \tag{5-8}$$

在推理时，评估模型将使用T的嵌入和每个候选$\hat{T}$的嵌入之间的余弦相似度作为$\hat{T}$的评分，这也是其对应候选项$\hat{Y}$的评分。然后，根据评分从$\hat{Y}$中选择出最终的结果Y：

$$\mathrm{Y} = \{\hat{Y}_i \mid \hat{Y}_i \in \hat{\mathrm{Y}}, i \in \kappa\}, \kappa = \underset{i \in \{1,2,\ldots,k\}}{\arg\max}\ _k sim(h_T, h_{\hat{T}_i}) \tag{5-9}$$

5.4 实验与分析

为验证本章提出方法的优越性，首先在CED数据集上将该方法与其他几种方法进行了对比实验。

5.4.1 实验设置与评价指标

在候选生成阶段，本文使用 Adam W[87]优化器在数据集上进行微调训练，训练参数如下：默认批处理大小设置为 128，学习率为 2e-5，训练轮次设置为 10 轮。在评估阶段，训练参数如下：批处理大小设置为 4，学习率为 4e-5，训练轮次设置为 10。为防止过拟合并加速模型收敛，所有模型进一步利用 Dropout 操作进行优化，使用 Adam W 优化器最小化模型损失，在每批次操作后进行参数规范以提高模型性能[88]。

为准确有效地评价本文模型的性能表现，本文在实验中使用了 Hit（命中率）作为评价指标。较高的命中率表示模型可以更接近每个实体的真实缩写。

$$Hit@n = \frac{|Numn|}{|NS|} \tag{5-10}$$

其中，$|Numn|$表示正确缩略词排前 n 名的次数。

5.4.2 候选缩略词数量确定

对于生成模型生成的候选缩略词数量对于命中缩略词结果具有一定影响。理论上，生成的候选缩略词越多，候选缩略词中就越容易得到正确缩略词结果，然而，随着候选缩略词数量增加,训练评估模型所需的时间和空间开销会显著增加。因此为了在命中率和时空开销之间取得平衡，本文通过实验来确定候选缩略词的数量。结果如图 5-2 所示：

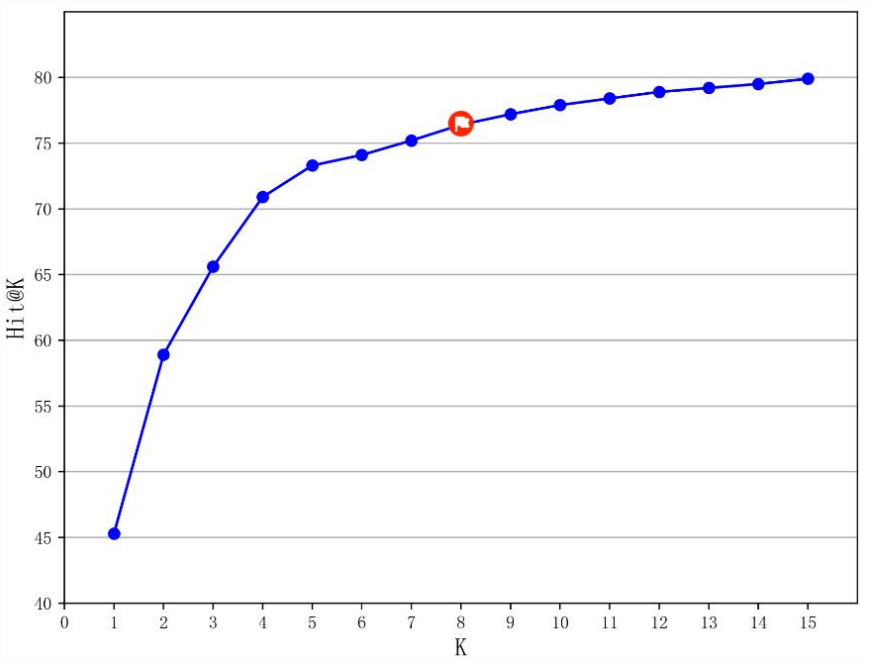

图 5-2 候选缩略词数量对命中率的影响

Figure5-2 The impact of the number of candidate abbreviations on the hit rate

由图 5-2 可知，当 K 值逐渐增大时，命中率也在随之提高。当 K 从 8 到 15 时，增加候选缩略词数量所带来的效益逐渐减小。此外，考虑到较高候选数量可能带来的计算复杂度和资源消耗，相比于更高的候选数量，K 为 8 时能够在较小的计算负担下实现较好的性能。因此，本实验将候选缩略词数量定为 8。

5.4.3 性能比较

为了评估模型的有效性，本文将与同一数据集上的几个基线模型进行比较。基线模型有序列标记模型和序列生成模型。其中，序列标记模型包括：CRF，RNN-RADD，Bi-LSTM+CRF，BERT+Bi-LSTM+CRF。序列生成模型包括：Attention-based Seq2seq，Transformer，CPT。

（1） CRF：CRF 是一种经典的机器学习模型，广泛应用于序列标记、实体识别等任务。在缩略词生成任务中，它被用来为给定的全称生成候选缩写。该方法通过引入手写特征来辅助生成。

（2） RNN-RADD：动态字典机制有助于弥补单个汉字表示的不足，RNN-RADD 方法将循环神经网络与用于建模动态字典（RADD）的机制相结合，以提高对上下文信息的建模能力。

（3） Bi-LSTM+CRF：该模型使用了词分割和词性信息作为特征来辅助生

成。

（4） BERT+Bi-LSTM+CRF：该模型首先使用 BERT 获取全称中每个字符的预训练向量，然后将其输入到双向 LSTM 中，最后使用 CRF 执行中文缩略语预测任务。

（5） Attention-based Seq2seq：基于注意力的 Seq2seq 是一种典型的生成模型，使用编码器-解码器结构。在中文缩略语生成中，它试图使用注意力机制来更好地捕捉输入序列和输出序列之间的关系。

（6） Transformer：Transformer 是一种经典的编码器-解码器模型，广泛应用于生成任务。它通过自注意力机制有效地捕捉输入序列中的长距离依赖关系，对于中文缩略语生成等任务具有很好的性能。

（7） CPT-base：兼顾理解和生成的中文预训练模型，在许多 NLG 任务上显示出有效性。

表 5-1 实验结果

Table5-1 Experimental Result

模型	Hit@1	Hit@3
CRF	39.4	50.6
RNN-RADD	42.5	65.4
Bi-LSTM+CRF	43.4	66.6
BERT+Bi-LSTM+CRF	44.9	**68.8**
Attention-based Seq2seq	32.8	47.6
Transformer	34.6	48.5
CPT-base	**45.3**	65.6
Our Method	**50.1(+4.8)**	**70.9(+2.1)**

表 5-1 总结了所有比较方法的总体结果。实验结果表明，本文提出的方法在中文缩略词预测任务中表现优异，超越了所有基线模型。具体而言，在中文缩略词公开数据集上，本文方法在 *Hit*@1 和 *Hit*@3 指标上分别取得了 50.1% 和 70.9%的成绩，较基线模型中的最佳结果分别提升了 4.8 和 2.1 个百分点，优于传统方法。

此外，与 CPT-base 相比，当本文引入上下文主题一致性评估阶段时，模型性能得到了提升，进一步证实了该评估阶段在整个中文缩略词预测流程中的关键作用和有效性。这是因为评估阶段能够对模型预测结果进行进一步的筛选和优化，确保预测结果与上下文主题保持高度一致，从而提高了预测的准确性和可靠性。

5.4.4 基于上下文长度的参数实验

上下文信息为本章研究的重要一环，为了深入探究不同上下文长度对模型性能的影响，本文进行了基于上下文信息长度的消融实验。同第四章，由于计算能力的限制，本章仍将上下文长度限制为 200。此外，上下文长度为 0，即仅计算了实体全称与候选缩略词之间的余弦相似度，而不把上下文纳入计算。实验结果如表 5-2 所示。

表 5-2 上下文长度对模型命中率的影响

Table5-2 The impact of context length on model hit rate

上下文长度 L	Hit@1	Hit@3
0	45.3	65.6
50	48.7	68.4
100	50.1	70.9
150	47.3	71.6

由表 5-2 可知，随着上下文长度的减小，其对应的 *Hit*@1 值也相应减少，但 *Hit*@3 相对稳定。其次，当上下文长度在 100-150 区间内时，其 *Hit*@1 值呈下降趋势，但其值仍呈上升趋势。这是因为一个实体往往可能对应着多个有效的缩写形式。但是需要注意的是 *Hit*@1 值是对缩略词预测的严格指标。此外，当实验删除上下文信息时， *Hit*@1 和 *Hit*@3 性能均下降，这再一次验证了上下文信息对本研究的重要所在。

5.4.5 基于启发式规则的消融实验

为了研究不同的启发式规则对实验结果的影响，本文分别删除了启发式规则在对方法性能进行评估。评估结果见表 5-3。

表 5-3 启发式规则对模型命中率的影响

Table5-3 The impact of heuristic rules on model hit rate

启发式规则	Hit@1	Hit@3
All	50.1	70.9
w/o any heuristic rule	47.3	68.7
w/o word-based rule	42.1	65.2
w/o truth-based rule	48.7	69.2
w/o character-based rule	47.9	68.3

根据表 5-3 的结果，可以得出以下结论：

（1）当移除任意一种启发式规则后均会导致模型性能下降，这充分证明了启发式规则在优化模型性能中的重要作用。

（2）当移除基于词的编辑距离规则时，模型性能出现了最为显著的下降，这意味着该规则对于模型性能的提升贡献最大。这一现象的原因在于，词作为语言中最小的语义单位，在给定实体完整形式的情况下，移除部分词汇可能会改变其语义，从而生成最具挑战性的负样本，这些样本对模型的识别能力提出了更高的要求，进而推动了模型性能的提升。

5.4.6 人工评估

为了更公平和准确的比较，本文基于 4.4.4 所述人工评估框架，对随机抽取的 1000 个样本进行人工标注，统计了 *CC*（有效缩略词但不唯一）、*CE*（语义异常）和 *EE*（语义与结构均异常）三类实体的分布比例。评估结果见表 5-4。

表 5-4 人工抽样结果

Table5-4 Manual sampling results

Judges	Hit	R_{CC}	R_{CE}	R_{EE}	Hit+R_{CC}
1	45.7%	29.3%	17.8%	7.2%	75%
2	46.3%	28.6%	18.5%	6.6%	74.9%
3	45.5%	29.7%	17.3%	7.5%	75.2%
Average	45.8%	29.2%	17.9%	7.1%	75.0%

由表 5-4 可知，在 1000 个随机样本中，样本 *Hit* 值和 R_{CC} 值相对较高，且

二者相加的平均值达到 75.0%，这表明模型在预测中文缩略词方面有一定的成效，能够成功处理一半以上的样本，在命中目标以及处理缩略词特性方面表现良好。不过，结果中 R_{CE} 占有一定比例，说明模型在语义理解方面还存在着挑战。

5.5 本章小结

在本研究中，主要聚焦于中文缩略词预测问题，并提出了一种创新性的评估缩略词质量的方法。传统的中文缩略词预测方法多采用序列标记技术，尽管这一方法在某些应用中取得了初步成效，但它固有的局限性，尤其是在处理复杂语境中的多义词和同义词时，往往导致生成的缩略词不准确或缺乏语义一致性。为了解决这一问题，本研究创新性地将中文缩略词预测问题定义为序列生成任务。这一转变有效地突破了传统序列标记方法的局限，允许模型从更高层次的上下文信息中生成候选缩略词，提升了预测的灵活性和准确度。

进一步地，本文提出了通过引入“上下文主题一致性”来评估缩略词质量。这一方法通过比较候选缩略词与其上下文之间的主题一致性，能够准确筛选出最符合原实体含义的缩略词，而不单纯依赖于字符层面的相似性。这种评估机制大大提高了缩略词预测的准确性，尤其在处理含有多义词、同义词或歧义性的情况下，表现出了强大的优势。

为了验证所提方法的有效性，本文与当前主流的几种中文缩略词预测方法进行了对比实验。实验结果表明，本文方法在多个评价指标上均优于传统方法，充分证明了上下文主题一致性评估对于提升缩略词预测精度的作用。

展望未来，尽管本研究取得了一定的成果，但在模型的上下文建模和语义推理方面仍然存在提升空间。因此，未来的研究可以着重探索更为复杂且精细的上下文建模技术，如深度语义分析、跨句子语境理解等，以进一步提升模型的性能和泛化能力。这将为中文缩略词预测技术的进一步发展奠定基础，并使其能够在更多自然语言处理应用中发挥作用。

第 6 章 总结与展望

本章首先总结本文对基于上下文增强的中文缩略词预测方法所做的研究工作，随后针对该领域未来的研究进行展望。

6.1 工作总结

中文缩略词预测作为自然语言处理领域中的一项关键任务，近年来得到了广泛的关注和研究。随着中文信息化进程的推进，缩略词在各类文本中逐渐增多，其准确预测对于文本理解、信息检索及问答系统等应用具有重要意义。传统的中文缩略词预测方法大多依赖于基于规则的手工构建或简单的机器学习模型，然而这些方法往往忽视了上下文信息的深层次语义联系，导致预测效果不尽如人意。因此，如何在复杂的语境下准确生成与上下文一致的缩略词，成为当前研究的一个重要挑战。针对上述问题，本研究基于序列生成模型和语义匹配双重视角出发，提出了一种基于上下文增强的中文缩略词预测方法，以提高模型在复杂语境下的适应能力。下面将对本文的研究内容进行总结：

（1）基于上下文增强的中文缩略词生成方法。传统缩略词生成方法往往仅依赖于源实体的字符特征，而忽略了缩略词的形成过程通常受到其所在文本环境的影响。为此，本文提出了一种上下文增强的序列生成框架，通过引入预训练语言模型，使得模型在学习缩略词生成模式时，能够充分利用上下文的深层语义信息。具体而言，该方法采用编码-解码架构，首先通过上下文增强编码器提取源实体及其周围文本的特征表示，然后利用预训练 Transformer 解码器生成符合语境的缩略词。实验表明，该方法相比传统基线在多种评测指标上均取得了显著提升，尤其在高语境依赖的任务场景下，生成结果更加精准且符合语言习惯。

（2）基于上下文主题一致性的中文缩略词预测方法。由于缩略词的形成方式通常具有多样性，不同的候选可能在字面形式上均符合缩略原则，但其语义适配性却存在显著差异。为解决这一问题，本文进一步提出了一种基于上下文主题一致性的缩略词评估策略。该方法借鉴对比学习思想，构造正负样本对，通过计算候选缩略词填充到原上下文中的主题一致性得分，来衡量候选缩略词与源实体的匹配程度。最终，模型依据该评分对生成的候选集进行排序，并选取最优解作

为最终输出。实验结果表明，该方法能够有效提升缩略词的可读性和语境适应性，避免了单纯基于字符匹配可能带来的歧义问题。

6.2 未来展望

针对现存的一些问题，本研究虽然对现有的中文缩略词预测方法进行了部分改进，但由于研究时间等的限制，算法性能还可以进一步提升，未来仍有值得改进的方向：

（1）扩展多模态信息融合：当前方法主要依赖文本信息进行缩略词预测，而在实际应用中，实体的缩略方式往往受到图像、音频、知识图谱等多模态信息的影响。例如，医疗领域中的术语缩略可能受到影像报告的辅助解释，而新闻报道中的缩略词则可能依赖于视频字幕的上下文提示。因此，未来的研究可探索如何融合视觉与文本特征，利用跨模态学习提升缩略词生成的精准度。

（2）构建更细致的语境建模机制：尽管本文的上下文增强方法已显著提升了模型对语境信息的利用能力，但中文语境的复杂性远超当前方法的建模能力。例如，某些实体在不同领域或特定场景下的缩略方式可能完全不同，而现有模型在跨领域预测时仍然存在一定的泛化问题。未来可以探索领域自适应建模，结合知识图谱、因果推理等技术，使模型能够动态调整缩略词预测策略，以适应不同场景的需求。

参考文献

[1] 李盖玛吉. 汉语动词性缩略语与原语词的句法语义对比研究[D]. 华中师范大学,2017.

[2] 刘友强. 基于双语平行语料的中文缩略语识别研究[D]. 南京大学,2015.

[3] Zhang Y, Sun X. A Chinese Dataset with Negative Full Forms for General Abbreviation Prediction[C]//Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.

[4] 牛晓雁.现代汉语缩略语研究及规范[D]. 河北师范大学硕士学位论文,2004.

[5] 武子英, 郑家恒. 现代汉语缩略语自动识别的方法研究[J]. 计算机工程与设计, 2007, 28(16): 4052-4054.

[6] 李盖玛吉.现代汉语动词性缩略语的多维考察[D].华中师范大学,2024.

[7] 丁浩,胡广伟,齐江蕾,等.基于随机森林和关键词查询扩展的医学文献推荐方法[J].数据分析与知识发现,2022,6(07):32-43.

[8] Sil A, Kundu G, Florian R, et al. Neural cross-lingual entity linking[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[9] Xia B, Chu R, Yu H V E, et al. Abbreviation of Chinese Botanical Garden Names[J]. Phytohortology. EDP Sciences, xvii--xx, 2021.

[10] Sun X, Li W, Meng F, et al. Generalized abbreviation prediction with negative full forms and its application on improving chinese web search[C]//Proceedings of the Sixth International Joint Conference on Natural Language Processing. 2013: 641-647.

[11] Yang D, Pan Y C, Furui S. Vocabulary expansion through automatic abbreviation generation for Chinese voice search[J]. Computer Speech & Language, 2012, 26(5): 321-335.

[12] Sun X, Wang H, Zhang Y. Chinese abbreviation-definition identification: A SVM approach using context information[C]//Pacific Rim International Conference on Artificial Intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 495-504.

[13] Sun X, Okazaki N, Tsujii J. Robust approach to abbreviating terms: A discriminative latent variable model with global information[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 2009: 905-913.

[14] Chang J S, Teng W L. Mining atomic chinese abbreviations with a probabilistic single character recovery model[J]. Language Resources and Evaluation, 2006, 40: 367-374.

[15] Chang J S, Lai Y T. A preliminary study on probabilistic models for Chinese abbreviations[C]//Proceedings of the Third SIGHAN Workshop on Chinese Language Processing. 2004: 9-16.

[16] Yang D, Pan Y, Furui S. Automatic chinese abbreviation generation using conditional random field[C]//Proceedings of Human Language Technologies:

The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. 2009: 273-276.

[17] Chen H, Zhang Q, Qian J, et al. Chinese named entity abbreviation generation using first-order logic[C]//Proceedings of the Sixth International Joint Conference on Natural Language Processing. 2013: 320-328.

[18] Zhang L, Li S, Wang H, et al. Constructing Chinese abbreviation dictionary: A stacked approach[C]//Proceedings of COLING 2012. 2012: 3055-3070.

[19] 焦妍,王厚峰. 基于机器学习方法与搜索引擎验证的缩略语预测[C]//中国中文信息学会.中国计算语言学研究前沿进展（2009-2011）.清华大学出版社,2011:6.

[20] Zhang Q, Qian J, Guo Y, et al. Generating abbreviations for chinese named entities using recurrent neural network with dynamic dictionary[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 721-730.

[21] Ma D, Li S, Wu F, et al. Exploring sequence-to-sequence learning in aspect term extraction[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. 2019: 3538-3547.

[22] Wang C, Liu J, Zhuang T, et al. A Sequence-to-Sequence Model for Large-scale Chinese Abbreviation Database Construction[C]//Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. 2022: 1063-1071.

[23] Cao K, Yang D, Liu J, et al. A Context-Enhanced Transformer with Abbr-Recover Policy for Chinese Abbreviation Prediction[C]//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022: 2944-2953.

[24] Tong H, Xie C, Liang J, et al. A Context-Enhanced Generate-then-Evaluate Framework for Chinese Abbreviation Prediction[C]//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022: 1945-1954.

[25] Feng W, Lan L, Zhang X, et al. Learning sequence-to-sequence affinity metric for near-online multi-object tracking[J]. Knowledge and Information Systems, 2020, 62: 3911-3930.

[26] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[27] Makris D, Agres K R, Herremans D. Generating lead sheets with affect: A novel conditional seq2seq framework[C]//2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-8.

[28] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2. 2014: 3104-3112.

[29] Cho K, van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[C]//Proceedings of the 2014 Conference on Empirical Methods in

Natural Language Processing (EMNLP). Association for Computational Linguistics, 2014: 1724.

[30] Pham H, Manning C D, Luong M T. Effective approaches to attention-based neural machine translation[J]. Computer Ence, 2015, 2015: 1-2.

[31] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need in Advances in Neural Information Processing Systems[J]. Search PubMed, 2017, 30: 5998-6008.

[32] Gu J, Bradbury J, Xiong C, et al. Non-autoregressive neural machine translation[C]//International Conference on Learning Representations (ICLR). 2018.

[33] Lee J, Mansimov E, Cho K. Deterministic Non-Autoregressive Neural Sequence Modeling by Iterative Refinement[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2018.

[34] Stern M, Chan W, Kiros J, et al. Insertion transformer: Flexible sequence generation via insertion operations[C]//International Conference on Machine Learning. PMLR, 2019: 5976-5985.

[35] Ghazvininejad M, Levy O, Liu Y, et al. Mask-Predict: Parallel Decoding of Conditional Masked Language Models[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 6112-6121.

[36] Gu J, Wang C, Zhao J. Levenshtein transformer[J]. Advances in Neural Information Processing Systems 32, 2019: 11179-11189..

[37] Bao G, Zhang Y, Teng Z, et al. G-Transformer for Document-Level Machine Translation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 3442-3455.

[38] Prabhavalkar R, Rao K, Sainath T N, et al. A Comparison of sequence-to-sequence models for speech recognition[C]//Interspeech. 2017: 939-943.

[39] Ji Z, Xiong K, Pang Y, et al. Video summarization with attention-based encoder-decoder networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(6): 1709-1717.

[40] Geng B, Yuan F, Xu Q, et al. Continual learning for task-oriented dialogue system with iterative network pruning, expanding and masking[J]. arXiv preprint arXiv:2107.08173, 2021.

[41] Xu L, Zhang Y, Hong L, et al. ChicHealth@ MEDIQA 2021: Exploring the limits of pre-trained seq2seq models for medical summarization[C]//Proceedings of the 20th Workshop on Biomedical Language Processing. 2021: 263-267.

[42] Ye J, Zheng F, Zhao J, et al. Incorporating Reachability Knowledge into a Multi-Spatial Graph Convolution Based Seq2Seq Model for Traffic Forecasting[J]. arXiv preprint arXiv:2107.01528, 2021.

[43] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. PMLR, 2020: 1597-1607.

[44] Chen T, Kornblith S, Swersky K, et al. Big self-supervised models are strong semi-supervised learners[J]. Advances in neural information processing systems, 2020, 33: 22243-22255.

[45] Chen X, Fan H, Girshick R, et al. Improved Baselines with Momentum Contrastive Learning[J]. arXiv e-prints, 2020: arXiv: 2003.04297.

[46] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9729-9738.

[47] Gao T, Yao X, Chen D. SimCSE: Simple Contrastive Learning of Sentence Embeddings[C]//2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021. Association for Computational Linguistics (ACL), 2021: 6894-6910.

[48] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.

[49] Wang D, Ding N, Li P, et al. CLINE: Contrastive Learning with Semantic Negative Examples for Natural Language Understanding[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 2332-2342.

[50] Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.

[51] Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-Guided Contrastive Learning for BERT Sentence Representations. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2528-2540.

[52] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 4171-4186.

[53] Yan Y, Li R, Wang S, et al. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 5065-5075.

[54] Giorgi J, Nitski O, Wang B, et al. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1:

Long Papers). 2021: 879-895.

[55] Pan X, Wang M, Wu L, et al. Contrastive Learning for Many-to-many Multilingual Neural Machine Translation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 244-258.

[56] Liu Y, Liu P. SimCLS: A Simple Framework for Contrastive Learning of Abstzuractive Summarization[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2021: 1065-1072.

[57] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[58] Griffiths T, Jordan M, Tenenbaum J, et al. Hierarchical topic models and the nested Chinese restaurant process[J]. Advances in neural information processing systems, 2004, 17-24.

[59] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. science, 2006, 313(5786): 504-507.

[60] Kingma D P, Welling M. Auto-encoding variational bayes[EB/OL].(2013-12-20)

[61] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[62] Sia S, Dalmia A, Mielke S J. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too![J]. arXiv preprint arXiv:2004.14914, 2020.

[63] Zhao H, Phung D, Huynh V, et al. Topic modelling meets deep neural networks: A survey[J]. arXiv preprint arXiv:2103.00498, 2021.

[64] Qiang J, Qian Z, Li Y, et al. Short text topic modeling techniques, applications, and performance: a survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(3): 1427-1445.

[65] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure[J]. arXiv preprint arXiv:2203.05794, 2022.

[66] 韩旭.基于神经网络的文本特征表示关键技术研究[D].北京：北京邮电大学,2019.

[67] 梁杰,陈嘉豪,张雪芹,等.基于独热编码和卷积神经网络的异常检测[J].清华大学学报(自然科学版),2019,59(07):523-529.

[68] 殷复莲,潘幸艺,柴剑平.基于词向量的电影评论情感分析方法[J].现代电影技术,2017,(08):4-9.

[69] Zhao Y, Huang S, Dai X, et al. Learning word embeddings from dependency relations[C]//2014 international conference on asian language processing (ialp). IEEE, 2014: 123-127.

[70] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2.

2013: 3111-3119.

[71] Peters M E, Neumann M, Iyyer M, et al. Deep Contextualized Word Representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018: 2227-2237.

[72] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[73] 朱张莉,饶元,吴渊,等.注意力机制在深度学习中的研究进展[J].中文信息学报,2019,33(06):1-11.

[74] 任欢,王旭光.注意力机制综述[J].计算机应用,2021,41(S1):1-6.

[75] Astudillo A, Barrera A, Guindel C, et al. DAttNet: monocular depth estimation network based on attention mechanisms[J]. Neural Computing and Applications, 2024, 36(7): 3347-3356.

[76] Jin Y, Xie J, Guo W, et al. LSTM-CRF neural network with gated self attention for Chinese NER[J]. IEEE Access, 2019, 7: 136694-136703.

[77] Guo Q, Huang J, Xiong N, et al. MS-pointer network: abstractive text summary based on multi-head self-attention[J]. IEEE Access, 2019, 7: 138603-138613.

[78] Wang X, Qi G J. Contrastive learning with stronger augmentations[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(5): 5549-5560.

[79] Xu B, Xu Y, Liang J, et al. CN-DBpedia: A never-ending Chinese knowledge extraction system[C]//International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Cham: Springer International Publishing, 2017: 428-438.

[80] Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. arXiv preprint arXiv:2109.05729 (2021).

[81] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Icml. 2001, 1(2): 3.

[82] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. Advances in neural information processing systems, 2020, 33: 9459-9474.

[83] 李豪,张柏苑,邵蝶语,等.融合BERTopic和Prompt的学者研究兴趣生成模型——以计算机科学领域为例[J/OL].情报科学,1-21[2024-08-25].http://kns.cnki.net/kcms/detail/22.1264.G2.20240726.1726.010.html.

[84] McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction[J]. arxiv preprint arxiv:1802.03426, 2018.

[85] MCINNES L, HEALY J, ASTELS S. hdbscan: Hierarchical density based clustering[J].The Journal of Open Source Software, 2017, 2(11).

[86] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping

Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In Findings of the Association for Computational Linguistics: EMNLP 2020. 657-668.

[87] Loshchilov I, Hutter F. Decoupled weight decay regularization[J]. arxiv preprint arxiv:1711.05101, 2017.

[88] 马志远,高颖,张强,等.融合语义与结构信息的知识图谱补全模型研究[J].数据分析与知识发现,2024,8(04):39-49.

攻读学位期间发表的学术成果

已录用的论文：

[1] 王潇冉，陶皖，皇苏斌. 基于上下文增强的中文缩略词预测方法研究[J]. 天津理工大学学报.（第一作者，已录用）

[2] Wan T, Wang X, Zhang Q. A Study on Chinese Acronym Prediction Based on Contextual Thematic Consistency[C], International Conference on Brain Inspired Cognitive Systems. Singapore: Springer Nature Singapore.（第二作者，国际学术会议，已录用）

[3] 张强,王潇冉,高颖,等.ChatGPT 生成与学者撰写文献摘要的对比研究——以信息资源管理领域为例[J].图书情报工作,2024,68(08):35-47.（第二作者，CSSCI）

致 谢

行文至此，我的硕士生涯也即将落下帷幕，这一次是真的要对学生时代的自己说一声：“再见了”。二十余载求学路，我深感这一路走来属实不易，纵然有万般不舍，但也要继续向前走。回想在安工程这三年的时光，如烟花一般，绚烂而短暂。

首先，我要由衷的感谢我的导师陶皖教授，感谢陶老师从本科以来对我的指导和帮助。初次见到陶老师是在大一，她的亲切与温柔如春风拂面，让我倍感温暖。陶老师一直是我所追随的女性榜样，她优雅地平衡着自我、工作和家庭。本科至今，老师对我产生的影响不仅仅是学术上要脚踏实地，同时还有为人处事的看法以及生活与规划上的宝贵建议，她以严谨的学术态度和乐观的人生态度，激励我不断进步和成长。其次，我要感谢我的大师兄张强博士，在读研期间，师兄以渊博的专业知识和平易近人的处事态度让我受益良多。在我做学术研究遇到困难时，他经常为我指引方向，在我对未来迷茫的时候，给予我很多的人生建议。最后，感谢给予我帮助的所有老师以及 721 的小伙伴们，特别感谢师妹金晶、师弟董纬、祝玉宣（排名不分前后）。山水一程，有幸相遇。在这即将分别的时刻，我衷心祝愿老师工作顺利、身体健康；祝愿师兄和小伙伴们前程似锦、梦想成真；祝愿课题组越来越好，蒸蒸日上，在学术的道路上取得更加辉煌的成就。

一路走来幸得朋友们一直陪伴在我身边，给予我无限包容、安慰和鼓励。感谢读研期间王青、孙先兰、唐朝军、朱毅、杨小涵（排名不分前后）的帮助与陪伴。希望我们能在各自的领域闪闪发光也好，普普通通也好，都能成为自己想要成为的人。愿大家前程似锦，万事胜意。

父母之恩，常记于心。漫漫求学路离不开父母的无条件支持，感谢父母二十余载养育恩情，感谢你们对我每一次选择的理解与支持，家永远是我最温暖的港湾和最坚硬的后盾。养育之恩，无以为报，希望未来我可以像你们于我一样，成为你们的依靠。愿父母身体健康、万事顺意。

始于初秋，终于盛夏。我想感谢这三年里迷茫、焦虑、内耗、遗憾过以及努力、幸福、感动、快乐过的自己，尽管一路跌跌撞撞的前进，但也从未想过放弃。无数个自我否定，自我治愈的日子，让我发现并学会了接受自己的平庸。特别感

谢这段人生经历，尤其是那些宝贵又温暖的帮助，让我不断地了解自己、认识自己，让我能够以更从容、更平和的心态迎接人生的新篇章。愿自己赤诚勇敢，热烈自由。

衷心感激曾经遇见的每一个人，所有的经历与相遇，于我是收获亦是礼物。

最后，感谢所有百忙之中抽出时间参加我论文评阅、评议和答辩的专家们，感谢你们宝贵的意见与建议。