分类号：TP391　　学校代码：10363
密　级：公开　　学　　号：2220920101

安徽工程大学

Anhui Polytechnic University

硕士学位论文

题目　基于多特征融合的实体扩展技术研究

论 文 作 者　李道玉
指 导 教 师　皇苏斌 副教授
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：2220920101

题　　目 基于多特征融合的实体扩展技术研究

英文并列

题　　目 **Research on Entity Expansion Technology Based on Multi-Feature Fusion**

学生姓名：李道玉

指导教师：皇苏斌 副教授

专　　业：软件工程

研究方向：领域软件工程

论文答辩日期：2025 年 5 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：李益玉

日期：2025 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：李道玉　指导教师签名：

日期：2025 年 5 月 9 日　日期：2025 年 5 月 9 日

基于多特征融合的实体扩展技术研究

摘 要

在信息技术飞速发展的时代，海量数据和文本信息的增长使实体扩展成为自然语言处理和计算机科学领域的重要研究方向。实体扩展通过利用上下文信息、语义关联和多模态特征，拓展和丰富实体的语义表达，以满足用户对信息获取和语义理解的需求。该技术在搜索引擎、问答系统、分类体系构建等应用中对系统性能和用户体验具有显著影响，特别是在科研和实际应用领域，其重要性日益凸显。当前，实体扩展研究已超越传统的人名、地名、机构名领域，逐步关注网络数据。然而，网络数据的冗余性和非结构化特点为其带来了巨大挑战，例如，如何通过少量的社交媒体帖子识别出公众情绪的变化，甚至预测潜在的社会动荡，成为亟待解决的问题。此外，实体扩展技术正向产业情报分析、学术文献挖掘和知识管理等更广泛的领域拓展，帮助研究人员高效提取海量数据中的有价值信息，并构建语义丰富的知识图谱，从而为多领域研究和应用提供有力支持。因此，随着数据规模和复杂度的持续增长，实体扩展技术不仅具有重要的学术价值，也在工业界展现出广泛的应用前景。

本文的研究工作围绕实体扩展展开，主要从以下三个方面进行深入探讨：实体同义词扩展、实体缩略词扩展以及实体集合扩展。具体研究内容如下：

（1）提出了一种基于灵活感知域和多层上下文的实体同义词生

成方法，解决忽略实体全局信息和错误累计传播问题。该方法充分利用了灵活的感知域和多层次的上下文信息。首先，为了判断是否将新的候选实体纳入同义词集，本文设计了一种结合神经网络分类器与灵活感知域的策略。在分类器中，构建了一个三层特征交互网络，将实体层特征、集合层特征和句子层特征映射到同一嵌入空间，以提取更全面的同义词特征。其次，本文提出了一种基于动态权重的实体同义词扩展算法，利用训练好的神经网络分类器，从候选实体词汇中扩展最终的同义词。在三个公共数据集上的实验结果表明，所提出的方法在实体同义词生成的效果上优于现有方法。

（2）设计了一种基于多特征融合和全局上下文的中文缩略词扩展方法，解决忽略语义关联和标签之间的依赖关系问题。该方法包括候选缩略词生成、多特征优化和全局上下文评估 3 个阶段。在候选缩略词生成阶段，我们利用了一种新的中文预训练不平衡 Transformer（CPT）模型来生成多个候选缩略词，提高了预测模型的准确性和全面性。在多特征优化阶段，我们通过综合词性、笔画和词典特征对每个候选特征进行评估，然后分配一个质量分数以筛选出最有价值的候选特征。在全局语境评估阶段，我们采用对比损失的方法训练中文缩略词预测模型，并结合全局语境信息，以确保更有效的候选词获得更高的分数。选择具有最高质量分数的缩写作为最终输出。实验结果表明，该方法在缩略词扩展方面优于其他同类方法。

（3）构建了一种基于双向上下文的实体集合扩展方法，解决依赖高质量语料实体注释和稀疏监督问题。该方法包括监督信号增强、

双向上下文模块生成和迭代引导扩展三个阶段，通过监督信号增强模块缓解稀疏监督问题，利用上下文模式生成模块结合 XLNet 自动生成实体上下文模式，并通过 BERT 编码提取高维语义表示，最终采用生成模式引导的迭代扩展策略，根据上下文相似性评分逐步扩展实体集合，实现了无需人工标注语料的高效、语义一致的实体扩展。实验结果表明，该方法在实体集合扩展方面优于其他同类方法。

综上所述，本文围绕实体扩展技术展开研究，针对实体同义词、缩略词和实体集合扩展中的关键问题提出创新性解决方案。通过结合多层次上下文信息、多特征融合及生成式语言模型等技术，本文显著提升了实体扩展的精度和效率。研究成果不仅丰富了实体扩展的理论方法体系，也为知识图谱构建、信息提取等多领域应用提供了重要支持，展现了实体扩展技术的广阔发展前景。

关键词：实体扩展，同义词扩展，缩略词扩展，实体集合扩展，多层上下文，多特征融合

Research on Entity Expansion Technology Based on Multi-Feature Fusion

ABSTRACT

In the era of rapid advancements in information technology, the exponential growth of massive data and textual information has made entity expansion a critical research area in natural language processing and computer science. Entity expansion leverages contextual information, semantic associations, and multi-modal features to enrich and extend the semantic representation of entities, catering to user demands for information retrieval and semantic understanding. This technology significantly impacts system performance and user experience in applications such as search engines, question-answering systems, and taxonomy construction. Its importance is increasingly prominent in both scientific research and practical applications. Currently, entity expansion research has progressed beyond traditional domains such as names of people, locations, and organizations, gradually focusing on web-based data. However, the redundancy and unstructured nature of web data pose significant challenges. For example, how to identify changes in public sentiment from a small number of social media posts, and even predict potential social unrest, has become a problem that urgently needs to be addressed. Moreover, entity expansion techniques are expanding into

broader fields, including industrial intelligence analysis, academic literature mining, and knowledge management. These advancements help researchers efficiently extract valuable information from massive datasets and construct semantically rich knowledge graphs, thereby providing robust support for research and applications across multiple domains. Consequently, with the continuous growth in data volume and complexity, entity expansion technology not only holds substantial academic value but also demonstrates wide-ranging application prospects in industry.

This work focuses on entity expansion and conducts an in-depth exploration in three main areas: entity synonym expansion, entity abbreviation expansion, and entity set expansion. The specific research contents are as follows:

(1) A method for entity synonym generation based on a flexible perception domain and multi-layer context is proposed. This method addresses the issues of ignoring global entity information and error accumulation propagation. It fully leverages a flexible perception domain and multi-layer contextual information. First, to determine whether a new candidate entity should be included in the synonym set, this paper designs a strategy that combines a neural network classifier with a flexible perception domain. In the classifier, a three-layer feature interaction network is constructed, mapping entity-level features, set-level features, and sentence-level features into the same embedding space to extract

more comprehensive synonym features. Second, a dynamic weight-based entity synonym expansion algorithm is proposed, which utilizes the trained neural network classifier to expand the final set of synonyms from candidate entity words. Experimental results on three public datasets demonstrate that the proposed method outperforms existing approaches in generating entity synonym sets.

(2) A Chinese abbreviation expansion method based on multi-feature fusion and global context is proposed to address the issues of ignoring semantic associations and dependencies between labels. This method consists of three stages: candidate abbreviation generation, multi-feature optimization, and global context evaluation. In the candidate abbreviation generation stage, we employ a novel Chinese Pretrained Imbalanced Transform (CPT) model to generate multiple candidate abbreviations, enhancing the accuracy and comprehensiveness of the prediction model. In the multi-feature optimization stage, each candidate feature is evaluated by integrating part-of-speech, stroke, and dictionary features. A quality score is then assigned to filter out the most valuable candidate features. In the global context evaluation stage, a contrastive loss approach is used to train the Chinese abbreviation prediction model, incorporating global context information to ensure that more effective candidate words receive higher scores. The abbreviation with the highest quality score is selected as the final output. Experimental results

demonstrate that this method outperforms other similar approaches in abbreviation expansion.

(3) A bidirectional context-based entity set expansion method is proposed to address the challenges of relying on high-quality annotated corpora and sparse supervision. This method consists of three stages: supervised signal enhancement, bidirectional context module generation, and iterative guided expansion. First, the supervised signal enhancement module mitigates the issue of sparse supervision. Then, the context pattern generation module integrates XLNet to automatically generate entity context patterns, while BERT encoding is used to extract high-dimensional semantic representations.

Finally, an iterative expansion strategy guided by generated patterns is employed. Entities are progressively expanded based on contextual similarity scores, achieving efficient and semantically consistent entity expansion without requiring manually annotated corpora. Experimental results demonstrate that this method outperforms other similar approaches in entity set expansion.

In conclusion, this paper focuses on the research of entity expansion technology and proposes innovative solutions to key challenges in entity synonym expansion, abbreviation expansion, and entity set expansion. By integrating multi-level contextual information, multi-feature fusion, and generative language models, this study significantly enhances the

accuracy and efficiency of entity expansion. The research findings not only enrich the theoretical framework of entity expansion methods but also provide essential support for various applications, such as knowledge graph construction and information extraction, highlighting the broad development prospects of entity expansion technology.

KEY WORDS: Entity Expansion, Synonym Expansion, Abbreviation Expansion, Entity Set Expansion, Multi-level Context, Multi-feature Fusion

目 录

第 1 章　绪论

在信息爆炸与大数据快速发展的时代，知识的获取与组织变得愈发重要。知识图谱与自然语言处理等技术已在多个领域广泛应用，为信息的高效管理与智能分析提供了有力支撑。然而，面对多样化、动态化的文本数据，如何准确且高效地扩展和丰富实体的语义信息，仍然是一项具有挑战性的任务。

实体扩展作为构建知识图谱和语义理解中的关键任务，通过挖掘大规模文本或知识库中的信息，自动为目标实体寻找其相关的同义词、缩略词或属于同一类别的集合元素，从而提升系统对实体表示的完整性与准确性。该任务旨在解决实体表示中存在的不完整性和语义模糊问题，更好地支持下游任务，例如，文本分析[1]、信息检索[2]和智能问答[3]。

尽管近年来实体扩展领域取得了诸多研究进展，现有方法在泛化能力、语义上下文的理解以及应对数据稀缺性等方面仍面临显著挑战。为解决这些问题，本文聚焦于实体扩展任务，系统地探讨同义词扩展、缩略词扩展以及集合自动扩展的方法与应用。通过深入研究，本文希望为该领域提供新的思路，并为实际场景中的知识管理与智能化应用提供技术支持。

1.1 研究背景与意义

新兴信息技术的快速发展与广泛应用催生海量数据和文本信息，这使得实体扩展成为计算机科学和自然语言处理领域的一个重要研究方向。随着互联网的快速发展，用户对于更智能、准确的信息检索和语义理解的需求不断增加。实体扩展的研究旨在利用上下文信息、语义关联以及多模态特征，拓展和丰富实体的语义表达，从而更好地满足用户对信息获取和语义理解的迫切需求。在搜索引擎[4-7]、问答系统[8-10]、分类法构建[11-13]等众多应用中，实体扩展的研究成果直接影响系统的性能和用户体验。因此，在信息技术快速发展的当下，深入研究实体扩展技术具有重要的学术意义和广泛的应用价值。

实体（Entity）指现实世界中具体的物质或抽象的概念，例如“肯德基”、“中国”、“机器学习”、“环境保护”等都可以被视为实体。在实体扩展任务中，通过给定少量的种子实体名词，从大规模文本中自动发现与种子实体同类的新实体。

例如，给定“国家”类种子实体“中国”、“日本”、“英国”等，可以扩展出“俄罗斯”、“巴西”、“法国”等其他国家实体[14]；或给定“技术领域”种子实体“人工智能”、“大数据”，可以发现更多相关实体如“机器学习”、“深度学习”[15-17]等。

与传统的命名实体识别（NER）任务不同，实体扩展关注的是通过少量种子实体自动发现同类实体，而非提前定义固定类别[18,19]。尽管现有的命名实体识别技术已在许多应用场景中取得了良好的效果，但其局限性也不容忽视。例如，命名实体识别需要大量人工标注的训练语料，且类别定义通常较为通用或依赖于领域预先设定，难以适应多样化和动态化的实际需求。

在学术研究与实际应用中，实体扩展技术对搜索引擎、问答系统、推荐系统、语义分析等任务具有重要意义。例如，实体扩展可以提高搜索引擎对用户查询的理解深度，增强问答系统的准确性，以及帮助推荐系统更精确地捕捉用户需求。通过深入研究实体扩展技术，可以有效提升知识挖掘、信息检索和自然语言理解系统的性能，从而为各类信息系统和服务的智能化发展提供技术支撑。

1.2 国内外研究现状

随着自然语言处理技术的迅速发展，实体扩展已成为知识图谱构建和语义理解领域的重要研究方向。围绕实体扩展任务，国内外学者针对不同类型的实体扩展需求进行了广泛而深入的研究。本节将从实体同义词扩展、实体缩略词扩展以及实体集合扩展三个方面，对当前的研究进展与发展方向进行系统阐述。

1.2.1 实体同义词扩展

在自然语言处理中，实体同义词扩展的任务旨在发现与给定实体（如人、地点、组织等）具有相同或相似意义的其他实体。这在许多应用中至关重要，因为有效处理和理解实体之间的同义关系是构建更智能、更精确的文本理解系统的基础。随着信息检索[7]、问答系统[10]、知识图谱[20]等领域的发展，实体同义词扩展不仅能够提高搜索引擎的准确性，还能增强自然语言处理系统对用户查询的理解能力。实体同义词扩展的研究不仅关注提高算法的准确性和效果，还在不断拓展适用范围，以适应各种自然语言处理应用的需求。正因为如此，此研究领域得到

快速发展，涌现出丰富的方法。

早期的实体同义词扩展研究主要集中在结构化数据上。在这一阶段，研究者主要依赖于规则驱动和统计模型来挖掘实体的同义词关系。例如，Li 等人[21]通过关键词共现的统计信息，利用互信息（PMI）或潜在语义分析（LSA）等方法识别同义词。另一方面，StrucSyn[22]则探索了子查询的表达方式，并提出了一种异构的基于图的数据模型，通过捕捉同义词、网页和关键字之间的交互，解决了模糊实体名称识别的问题。

进入更大数据规模和计算能力快速提升的阶段后，研究者开始探索更加自动化的基于排名和分类的方法。通过构建候选实体的关联性得分排序，这些方法使用机器学习模型学习词汇之间的相似性，并根据得分筛选候选同义词。例如，Qu 等人[23]提出了集成分布特征和文本模式来自动发现具有知识库的实体同义词。SurfCon[24]结合了表面形式信息和分布特征，用于隐私感知临床数据的同义词发现。Chao 等人[25]利用语义信息，筛除了低相关性或不一致的候选项，从而确保同义词合的质量。而 Shen 等人[26]提出的 SetExpan 框架则将同义词扩展分为两个阶段：首先通过上下文特征选择检测同义词对，然后通过排名方法将这些对组合成完整的同义词集合。这些方法通过将上下文信息引入实体同义词扩展，提高了模型的表现和应用范围。

随着深度学习技术的飞速发展，研究者开始利用神经网络模型对实体同义词扩展问题进行建模，其核心在于结合多层次的上下文信息，进一步提升生成结果的质量和鲁棒性。例如，Shen 等人[27]引入了多头注意力机制，开发了一个实例分类器来区分实体同义词集合与查询实体，从而更加精准地捕获候选实体之间的语义联系。与此同时，Pei 等人[28]进一步利用实体同义词集合的信息，提出了通过两级网络排列实体与实体同义词集合的上下文信息的模型。该模型能够将实体和实体同义词集合映射到同一嵌入空间，通过灵活感受野对高阶上下文信息进行编码，从而有效促进实体同义词扩展的质量与效率。

1.2.2 实体缩略词扩展

实体缩略词扩展是自然语言处理领域中的一个任务，其主要目标是将文本中出现的缩略词（Abbreviation）映射到其完整的形式，也就是缩略词的全称或原

始表达。缩略词广泛用于口头和书面语言，消除了现代意识与任何语言有限的词汇资源之间的矛盾[29]。作为一种流行的词汇化形式，缩写在各种自然语言处理（NLP）应用中起着重要作用[30,31]。例如，在中文语音搜索中，将缩写词添加到词汇表中可以显著提高语音搜索的准确性，因为语音中出现的许多缩写词可以链接到数据库中的正确条目[32]。在信息检索（IR）中，使用缩写扩展查询将提高网络搜索的质量，因为许多网页只包含缩写而不是完整形式[33]。

中文缩略词预测的早期研究是基于启发式规则和传统的机器学习算法。Sun等人[34]结合启发式规则和支持向量机（SVM），以找到完整的形式和他们的缩写在一个上下文中。Sun[35]提出了一种判别潜变量模型和标签编码方法，以纳入全球信息。Chang 等人[36,37]使用隐马尔可夫模型并将任务建模为错误恢复问题。之后，中文缩略词预测主要被表述为一个序列标记问题。Dong Yang 等人[38]提出设计一系列特征函数，并使用 CRF 作为标注工具。Chen 等人[39]进一步结合了一阶逻辑，以模拟 CRF 无法捕获的长距离和全局约束。然而，所有这些作品都需要精心设计的功能。最近，神经网络被广泛用于自动提取特征，并用于中文缩略词预测，以取代手工特征[40,41]。然而，Dehong Ma 等人[42]提出序列标记方法并不善于把握完整形式的整体含义，基于神经网络的序列标记模型往往忽略标签依赖性，因为他们只使用转移矩阵来鼓励最可能的标签序列。为了解决这些问题，Chao Wang[43]首先将中文缩写预测公式化为序列生成问题，并提出了一个多级预训练模型来合并字符，单词和概念级嵌入，将搜索引擎返回的搜索结果与手工制作的功能一起重新排列 CRF 的预测。实体缩略词扩展的研究仍存在一些缺点。中文语言的多样性使得缩略词在不同地区行业和领域有着巨大的变体，而目前的研究可能未充分考虑这一点，导致扩展模型的泛化能力不足。此外，专业领域的差异也是一个问题，因为缩略词在不同领域中可能具有截然不同的含义，当前研究往往未能充分考虑到这些差异。

1.2.3 实体集合自动扩展

实体集合自动扩展是一种自然语言处理（NLP）任务，其目标是从已有的种子实体（Seed Entities）集合出发，利用算法和数据资源自动发现和生成更多符合相同类别或语义相关性的实体，构建一个更大的实体集合。这项任务广泛应用

于知识图谱构建、信息抽取、推荐系统和搜索引擎优化等领域。目前针对实体集合扩展问题的研究主要集中在三大类方法：基于分布的方法、基于模板的方法以及基于融合的方法。这些方法从不同的技术视角切入，结合统计、语义分析和多源数据融合等手段来提升实体集合扩展的准确性和全面性。

基于模板的方法的核心思想是通过设计或自学习生成一系列模板，从而利用这些模板从数据中提取候选实体，并通过打分排序选出最终的扩展结果。这些模板可以是人工定义的语义模板，也可以是从语料中自动学习到的模式。

如 Deepika 等人[44]在研究中提出的 Pattern-Based Bootstrapping 方法，展示了通过预定义模式从结构化语料中提取实体的高效性。此外，依靠自学习的方式生成模板也得到了深入研究，例如 Rong 等人[45]提出的方法，通过从种子实体上下文学习抽取模板。

这种方法的优势在于针对性较强，能够高效捕捉特定模式下的实体扩展需求，但其局限性在于模板的质量和覆盖范围对结果有较大影响。若模板不够完善或语料复杂多样，可能导致实体无法被识别或误识别。

基于分布的方法关注语料中上下文分布的统计特性，其核心在于构建词项分布矩阵并计算种子实体与候选实体的相关度。Ustalov[46]在其语义理论研究中首次提出了上下文分布的思想，奠定了基于分布方法的基础。基于这一思想，Bellavista 等人[47]进一步提出了 Distributional Similarity 的计算方法，通过上下文分布信息生成种子实体与候选实体之间的语义相关度排序。这种方法的优点在于无需人工干预，适合处理大规模数据，其扩展结果具有较高的通用性。然而，方法也存在不足：上下文分布对稀疏数据敏感，同时噪声数据可能导致结果偏差。例如，词频较低的实体可能难以生成可靠的上下文分布，从而影响相关度计算的准确性。

基于融合的方法通过结合多种数据类型和多种技术手段，试图达到更高的扩展准确性和稳健性。这种方法综合了基于模板和基于分布两种方法的优势，特别是在多源数据环境下，如网页文本、表格数据等。Markowitz[48]提出的方法通过融合网页数据和预定义模式扩展实体集合，实现了从多种信号中综合提取实体的目标。类似地，Pantel 等人[49]的研究则通过结合分布特性和结构化数据的方法，展示了融合策略的强大能力。融合方法的一个显著优势是可以降低单一方法在处

理复杂场景时的局限性。例如，当基于模板的方法因模式设计不完善出现错误时，基于分布的方法可以提供有效补充。这种方法对计算资源要求较高，且在设计时需综合考虑数据选择、方法组合及结果融合策略。

1.3 论文主要研究内容

综上分析可知，实体扩展研究仍然存在如下亟需解决的问题：

（1）忽略实体全局信息和错误累计传播问题：排序和修剪策略，首先根据候选项引用同一实体的概率来排列候选项。随后，修剪排序列表以创建实体同义词集。然而，这些方法单独处理每个候选项，并独立计算其引用查询实体的概率，从而忽略了候选实体之间共享的全局信息。两阶段任务方法包括实体同义词对的初始提取，然后组合成实体同义词集。这些两阶段方法在第一阶段中专门使用训练数据，而在第二阶段中不能充分利用训练信号。此外，在实体同义词集合的组织期间，所识别的实体同义词对通常是固定的，缺乏第一阶段和第二阶段之间的反馈机制。这种反馈的缺乏会导致错误的积累。

（2）忽略语义关联和标签之间的依赖关系问题：传统的汉语缩略词预测方法通常只关注实体的表面形式和频率，基于实体频率和长度等基本指标进行预测，而忽略了实体之间的语义关系及其周围的大量上下文信息。序列标记在预测当前字符的标签方面具有固有的缺点。这些方法仅依赖于转移矩阵来识别最可能的标签，而没有充分利用标签依赖性。它们预测字符的标签，而不考虑输入序列中前面字符的标签。

（3）依赖高质量语料实体注释和稀疏监督问题：传统的基于语料库的实体集合扩展方法需要一个明确标注了实体位置的高质量语料库。这种依赖在实际应用中往往难以满足，因为真实场景中的语料可能缺乏实体注释，或者注释质量较低。这种情况直接导致上下文模式提取的质量下降，从而限制了扩展性能。实体集合扩展任务通常以极少的种子实体（通常只有 3-5 个）作为输入，在面对包含数百甚至上千候选实体的情况下，监督信号极其稀疏，难以为模型提供足够的信息指导实体扩展。

针对上述问题，本文研究工作针对实体同义词扩展、实体缩略词扩展和实体集合扩展研究目标展开，主要结构如图 1-1 所示。图中本文的贡献和创新点主要

可总结为以下内容：

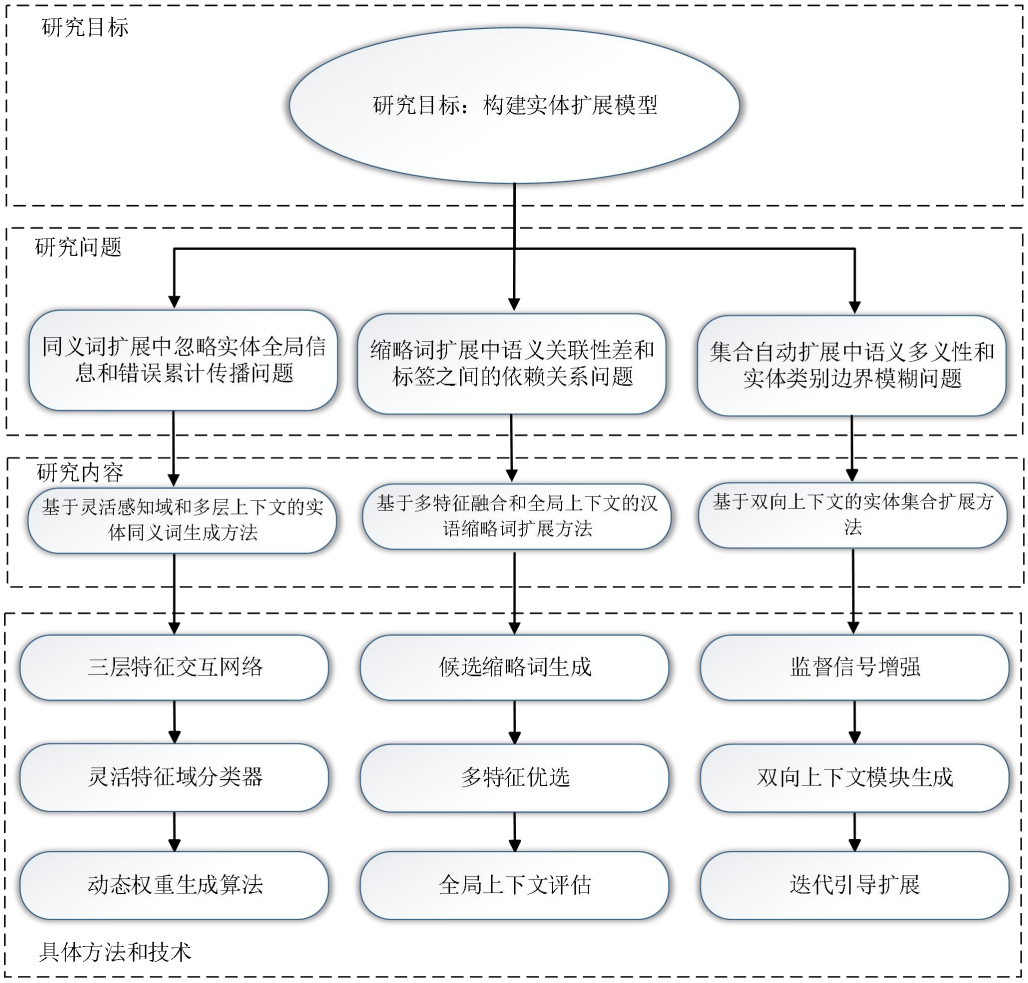

图 1-1 论文组织结构图

Fig. 1-1 Organization of the paper

（1）提出了一种结合灵活感知域与多层上下文信息框架。首先，为了判断是否将候选实体纳入同义词集，该方法结合神经网络分类器与灵活感知域，并构建三层交互式领域网络，将实体层、集合层和句子层映射至同一嵌入空间，以提取同义词特征。其次，本文设计了一种基于动态权重的实体同义词生成算法，利用训练好的神经网络分类器，从候选实体词汇表中筛选并构建同义词集。在三个公共数据集上的实验表明，该方法在实体同义词生成任务上优于现有方法。

（2）提出了一种融合多特征与全局上下文的汉语缩略词扩展方法。该方法包含候选缩略词生成、多特征优化和全局上下文评估三个阶段。首先，在候选缩略词生成阶段，采用改进的汉语预训练不平衡变换器（CPT）模型生成多个候选缩略词，以提升预测的准确性与覆盖度。随后，在多特征优化阶段，结合词性、笔画和词典特征对候选项进行评估，并分配质量分数，以筛选最优候选。最后，

在全局上下文评估阶段，引入对比损失训练缩略词预测模型，并结合全局语境信息，确保高质量候选词获得更高评分，最终选出最优缩写。实验结果表明，该方法在缩略词扩展任务中优于现有方法。

（3）提出了一种基于双向上下文的实体集合扩展方法。该方法包含监督信号增强、双向上下文模式生成和迭代引导扩展三个阶段。首先，通过监督信号增强模块缓解稀疏监督问题。随后，利用上下文模式生成模块结合 XLNet 自动生成实体上下文模式，并通过 BERT 编码提取高维语义表示。最后，采用基于生成模式引导的迭代扩展策略，通过上下文相似性评分逐步扩展实体集合，实现无人工标注情况下的高效、语义一致的实体扩展。实验结果表明，该方法在实体集合扩展任务中优于现有方法。

1.4 论文结构安排

全文共分为六章，每章的内容安排如下。

第一章绪论，阐述实体扩展背景以及研究意义，接着介绍与当前研究工作相关现状与亟需解决的问题，最后介绍本文工作内容。

第二章相关研究工作。介绍实体扩展基础理论与相关技术介绍，包括基本概念与知识。自然语言预处理、实体扩展方法和预训练模型。

第三、四、五章分别对 1.3 节中本文研究工作和贡献展开描述。每章主要内容包括相关问题描述、解决思路、研究方法描述以及仿真实验和结果分析。

第六章总结和展望。主要对本文工作进行总结，对工作的优势和不足进行分析，给出未来工作的研究方向。

第 2 章　基础理论与相关技术介绍

2.1 基本概念与知识

（1）实体识别：实体识别（Named Entity Recognition, NER）是自然语言处理（NLP）中的一项核心任务，旨在从非结构化文本中自动识别出具有特定语义意义的实体，并将其分类为预定义的类别，例如人名、地名、组织名称、时间、日期、货币单位等。NER 的任务包括识别实体在文本中的边界以及确定其所属的类别，例如在“Google was founded in California”中，识别“Google”为组织名称（ORG），而“California”为地名（LOC）。实体识别在许多应用场景中扮演重要角色，例如文本分类[1]、信息抽取[2]、搜索引擎优化[5]以及问答系统[8]等。它通过将非结构化文本转化为结构化信息，为后续的数据处理和分析奠定了基础。常见的实现方法包括基于规则的模式匹配（如正则表达式）、统计机器学习方法（如条件随机场 CRF）以及近年来广泛采用的深度学习技术（如 BiLSTM-CRF、基于预训练语言模型的 BERT 等）。深度学习方法能够利用上下文信息捕捉复杂的语义关系，从而提高 NER 的准确性与适应性。

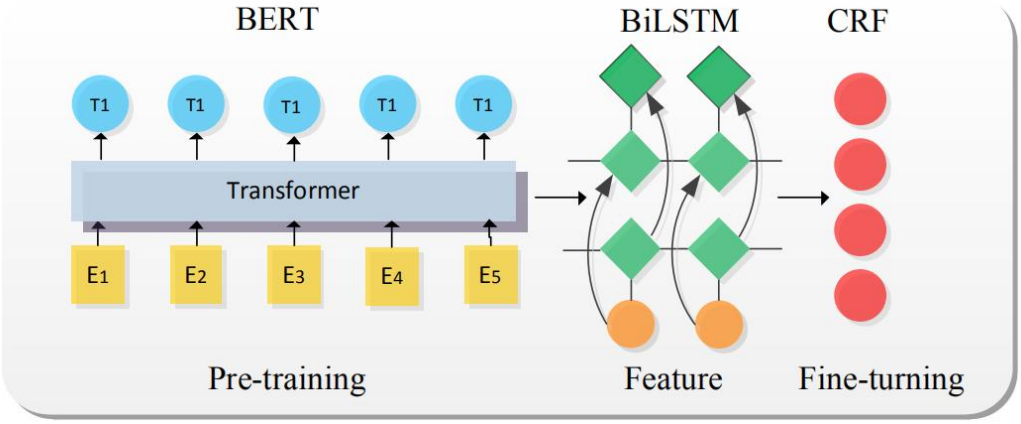

图 2-1 实体识别流程图

Fig. 2-1 Named Entity Recognition Process

如图 2-1，数据输入是实体识别任务的第一步，旨在为模型提供带有标注的高质量文本数据[50]。常用的标注数据集包括通用领域的 CoNLL-2003 和多领域的 OntoNotes，以及行业特定的自定义数据集（如医学、金融等）。这些数据集通常以 BIO 格式标注实体类别，表示实体的开始(B)、中间(I)和非实体(O)。例如，句

子"John is from New York"的标注为[John B-PER] [is O] [from O] [New B-LOC] [York I-LOC]，其中"John"，是一个人名实体(PER)，"New York" 是一个地点实体(LOC)。输入数据的质量和多样性直接决定了模型的泛化能力，因此在数据输入阶段，需要确保标注的一致性，并合理划分数据集（通常 80%用于训练，10%用于验证，10%用于测试）。此外，为低资源场景可以使用远程监督或数据增强技术来补充数据量。

数据预处理是实体识别模型训练的关键环节，用于将原始文本数据转换为模型可以理解的格式。首先，进行文本清洗，如移除多余的符号、无关字符或 HTML 标签。接下来是分词，将句子分解为单词或子词（根据具体任务需求）。例如，对于句子$S=\{w_1,w_2,\cdots,w_n\}$分词后可能得到子词序列$S^{'}=\{t_1,t_2,\cdots,t_n\}$。随后，需要对标注与分词后的子词序列进行对齐，确保实体标签正确映射到每个子词。例如，若“Apple”被分为“Ap”和“##ple”，则两个子词共享一个标注。然后，提取特征，包括静态词向量（如 Word2Vec、GloVe）或上下文相关嵌入（如 BERT、ELMo），将每个词或子词映射到高维向量空间。此外，还需对数据进行划分，将数据分为训练集、验证集和测试集，确保训练和评估的公平性和有效性。

模型选择是实体识别任务中的核心步骤，取决于任务需求、数据量以及计算资源的限制。对于小规模数据集，传统的条件随机场(CRF)是一种有效的选择，其通过定义特征函数$\varphi(y_{i-1},y_i,X)$，计算状态和转移的得分，目标函数为：

$$P=\frac{exp(\sum_i \varphi(y_{i-1},y_i,X))}{\sum_{y'} exp(\sum_i \varphi(y'_{i-1},y'_i,X))} \tag{2-1}$$

但对于大规模数据和复杂语境，现代深度学习模型表现更优。比如，BERT 使用多头注意力机制捕获全局上下文信息，其核心公式为：

$$Attention(Q,K,V)=softmax(\frac{QK^T}{\sqrt{d_k}}\cdot v) \tag{2-2}$$

其中，Q、K、V分别是查询、键和值矩阵，d_k是向量的维度。

将其与 BiLSTM-CRF 结合，可以充分利用上下文信息并对标签序列施加全局约束。最终的模型选择需根据场景权衡实时性、准确性和资源消耗。例如，轻量化模型（如 DistilBERT）适合部署到实时系统中。

模型训练是通过优化目标函数，让模型从输入数据中学习实体识别规则的过程。训练过程中，输入为分词后的序列$X=\{x_1,x_2,\cdots,x_n\}$和对应的标签$Y=$

$\{y_1, y_2, \cdots, y_n\}$模型的目标是最小化损失函数，如交叉熵损失：

$$\tau = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} log(P_{ij}) \quad (2\text{-}3)$$

其中，N是序列长度，C是类别数，P_{ij}是模型对标签j的预测概率。在使用CRF层时，损失函数会转化为：

$$\tau = -log(P(Y / X)) \quad (2\text{-}4)$$

训练过程包括前向传播（计算预测输出）、反向传播（计算梯度）和参数优化（如使用Adam优化器）。为了增强模型的泛化能力，可以通过数据增强技术（如同义词替换、随机插入）生成更多样本，或引入噪声提升模型鲁棒性。在多任务场景下，可以通过联合训练实体识别与其他任务（如关系抽取），进一步提升模型性能。

目前常用NER评价指标有准确率（Accuracy）、精确率（Precise）、召回率（Recall），其计算公式分别如下：

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2\text{-}5)$$

$$Precise = \frac{TP}{TP+FP} \quad (2\text{-}6)$$

$$Recall = \frac{TP}{TP+FN} \quad (2\text{-}7)$$

其中TP为真正类，FN为假负类，FP为假正类，TN为真负类。

（2）实体链接：实体链接（Entity Linking, EL）是一项关键的自然语言处理技术，其目的是将文本中的实体提及（mention）与知识库中的具体实体进行匹配和关联。实体链接通常包含三个核心环节：首先是从文本中识别出潜在的实体提及，这一步被称为实体识别；接下来，为每个提及生成候选实体集合，这些候选实体是可能对应的知识库实体；最后，通过语义分析和上下文匹配，从候选实体中选择与提及最相关的实体，即实体消歧。这项技术广泛应用于知识图谱构建、语义搜索和智能问答等领域。

如图2-1，在实体链接的过程中，实体提及识别作为第一个步骤，其任务是从自然语言文本中检测可能作为实体的短语或单词。这一过程可以通过基于规则的方法实现，例如正则表达式匹配和模板识别，也可以利用机器学习技术完成，如使用条件随机场（CRF）或双向LSTM（BiLSTM）等模型进行序列标注。识别出的实体提及可能会包含歧义，需要结合上下文和知识库信息加以进一步处理。

第二步候选实体生成的目标是通过字符串匹配、倒排索引或基于先验概率的方法，从知识库中快速检索出可能的候选实体集合。例如，通过分析 Wikipedia 的链接频率，可以建立提及与实体的关联概率，提升检索效率。

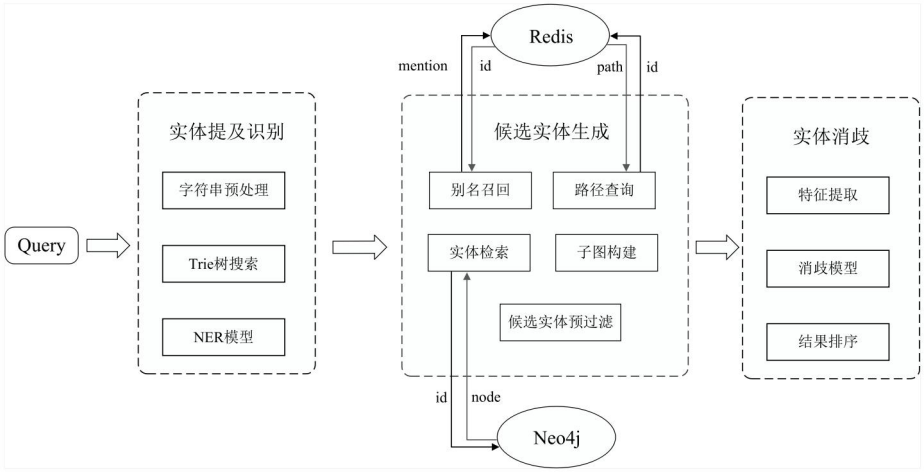

图 2-1 实体链接流程图

Fig. 2-1 Entity Linking Process

实体消歧是实体链接的关键环节，也是最具挑战性的部分。为了实现精准的消歧，需要对实体提及和候选实体的上下文进行语义匹配。基于词向量的表示方法，例如 Word2Vec 或 BERT，能够有效捕捉语义信息并应用于实体消歧。此外，知识图谱中的实体关系也可用于消歧，通过构建实体间的关系图并应用 PageRank 或随机游走等算法，可以增强上下文语义的利用效率。近年来，基于深度学习的端到端方法已成为实体消歧研究的热点，这类方法通过预训练模型直接学习上下文和实体之间的复杂语义关系，取得了显著的性能提升。

实体链接技术的应用场景十分广泛。在知识图谱构建中，实体链接被用于将结构化和非结构化数据中的实体统一到知识库中，形成跨领域的关联网络。在语义搜索和智能问答系统中，通过实体链接技术将查询中的关键术语与知识库中的实体关联，可以大幅提升搜索结果的准确性和问答系统的智能化水平。此外，在文本挖掘和情报分析中，实体链接能够帮助识别和组织海量文本中的核心实体信息，为决策提供支持。

尽管实体链接技术在实践中取得了广泛的应用，但其仍面临多重挑战。例如，实体名称的多义性问题使得消歧变得复杂，如“Apple”既可以指苹果公司，也可以指水果；同时，长尾实体由于知识库中信息不足，常常难以被准确识别和匹配。上下文信息不足的情况也可能导致系统难以确定提及的具体语义，尤其是在

短文本中。此外，随着新实体的不断出现，知识库需要持续更新以保持实体链接的时效性。

总的来说，实体链接技术是一项核心的自然语言处理任务，能够有效地连接文本和知识库，从而实现信息的深层次整合与利用。通过结合上下文语义表示、知识图谱关系和深度学习方法，实体链接技术在性能和应用范围上不断拓展，但也需要针对多义性、稀疏性等挑战进行持续研究和优化。

（3）知识库：实体扩展任务需要借助高质量的知识库作为核心支撑。知识库是实体扩展技术的重要基础，其主要作用是为实体扩展提供丰富的语义信息和结构化数据。知识库通常由实体、属性和实体之间的关系组成，通过这些信息，知识库能够为扩展提供上下文语义、候选实体集合以及验证扩展结果的依据。实体扩展相关的知识库可分为通用知识库和领域知识库两类，二者在应用场景和内容覆盖范围上各有侧重。

表2-1 知识库实例

Tab.2-1 knowledge base examples

知识库名称	覆盖范围	主要特点	适用场景
Wikipedia	通用	开放性高，内容丰富	通用实体扩展、信息检索
DBpedia	通用	基于 Wikipedia，结构化数据	实体消歧、上下位关系挖掘
Wikidata	通用	灵活性强，支持多语言	知识融合、多语言扩展
Freebase	通用	整合多个数据源	历史研究、通用数据挖掘
UMLS	医学领域	整合医学术语系统，专业性强	医学同义词扩展
OpenCorporates	金融领域	专注企业信息，高度准确	金融实体扩展、信息挖掘

如表 2-1 通用知识库是实体扩展任务中使用最广泛的资源之一，主要包括 Wikipedia、DBpedia、Wikidata 和 Freebase 等。这些知识库包含了来自各个领域的大量实体及其关系。例如，Wikipedia 作为全球最大的开放式百科全书，其内容结构化程度高，可以通过链接信息生成实体间的语义关系。DBpedia 和 Wikidata 则是在 Wikipedia 基础上提取和组织的知识库，它们通过语义网技术提供了结构化的数据形式，便于计算机处理和检索。此外，Freebase 作为早期的知识库项目，整合了多个数据源中的实体信息，尽管已停止更新，但仍是许多研究的参考数据集。这些通用知识库通常具有高覆盖率和多样性，适合用于广义的实体扩展任务。

领域知识库则专注于特定领域中的实体和关系，能够为实体扩展提供更加精确的上下文支持。在医学领域，UMLS（Unified Medical Language System）整合了多个医学术语系统，为医学术语的同义词扩展和缩略词解析提供了丰富的资源。

在金融领域，OpenCorporates 等知识库则专注于企业信息，能够支持与公司相关的实体扩展任务。相比通用知识库，领域知识库的特点是数据的专业性和高准确性，适合用于特定领域的实体扩展需求，但其覆盖范围往往有限。

知识库的结构化和多样性直接影响实体扩展的效果。在实体扩展过程中，知识库通常被用来生成候选实体、验证扩展结果以及挖掘实体之间的潜在关系。例如，在同义词扩展任务中，可以从知识库中提取实体的别名列表，以此生成同义词候选。在实体集合扩展任务中，知识库中的分类体系或实体关系能够为新实体的挖掘提供逻辑支持。此外，知识库中的上下位关系和语义网络还可以帮助扩展过程更好地理解实体之间的关联。

然而，知识库在实体扩展任务中也面临一些挑战。首先，知识库的时效性问题可能导致扩展结果无法及时反映最新的实体和关系。例如，新兴实体和动态变化的领域实体可能在旧版本知识库中缺失。其次，知识库的不完整性和长尾效应使得一些低频实体难以被识别和扩展。此外，不同知识库之间的格式和标准不统一，可能在数据整合时增加复杂性。因此，研究者在实体扩展任务中，通常需要结合多种知识库，并通过知识融合技术提升数据的覆盖率和质量。

总体来说，知识库在实体扩展任务中扮演了关键角色，为语义计算和候选生成提供了强有力的支持。随着知识库技术的发展，利用知识库构建跨领域、动态更新的扩展系统已经成为研究的重点方向。通过充分挖掘知识库的潜力，实体扩展技术能够在更多实际场景中发挥重要作用。

2.2 实体扩展方法

2.2.1 基于规则的实体扩展

基于规则的实体扩展是一种传统但高效的方法，广泛用于结构化和半结构化数据场景中。这种方法通过人为定义的规则从文本中提取目标实体并进行扩展，规则可以是基于正则表达式、固定模式或词典比对的。例如，在中文地名提取任务中，可以设计一个规则匹配“两个汉字加‘省’或‘市’”的模式，快速筛选出符合条件的候选实体。这种方法的实施成本较低，不需要复杂的模型训练，因此在小规模数据场景下表现优异。

则匹配的核心可以形式化为如下描述：给定一个文本T和规则集合$R = \{r_1, r_2, \cdots, r_n\}$，实体扩展的目标是找到所有满足某个规则$r_i \epsilon R$的子字符串集合$E \subseteq T$，具体实现中，这可以通过正则表达式来描述，如从文本中提取符合邮箱格式的字符串：

$$T_{email} = [a - zA - ZO - 9._] + @[a - zA - Z] + [a - zA - Z] + \quad (2\text{-}8)$$

此外，规则方法可以通过语法规则增强其能力。例如，利用短语结构树或依存句法分析，捕捉目标实体的上下文特征，从而提高复杂语句中的准确率。

尽管规则方法在处理特定领域任务中表现优异，但其局限性也很明显。一方面，规则的设计高度依赖领域知识，无法轻松迁移到新任务；另一方面，当数据中包含大量非结构化内容或模式复杂时，规则匹配的效果可能大幅下降。然而，通过结合统计方法动态调整规则权重或设计更广泛的模式匹配策略，规则方法的适应性可以得到提升。

2.2.2 基于上下文的实体扩展

基于上下文的实体扩展通过分析目标实体的上下文关系，从语料库中挖掘出与之语义相关的其他实体。其理论基础是分布假设，即语义相似的词或实体通常出现在相似的上下文中。例如，“苹果公司”在描述科技公司的上下文中频繁出现，其他如“谷歌”“微软”等公司很可能共享类似的语义背景。这种方法适合处理开放域场景，尤其在大规模语料下表现出色。

上下文方法的核心是统计上下文中的共现信息。若目标实体e_t和候选实体e_c在语料中共现，其关联强度可以通过点互信息计算：

$$PMI(e_t, e_c) = log\frac{P(e_t, e_c)}{P(e_t)P(e_c)} \quad (2\text{-}9)$$

其中$P(e_t, e_c)$表示e_t和e_c同时出现在上下文窗口中的概率，而$P(e_t)$和$P(e_c)$分别是它们的单独出现概率。PMI越高，两个实体的语义关联性越强。

分布式表示进一步增强了上下文方法的表达能力。通过将实体和上下文表示为高维向量，可以通过余弦相似性评估实体之间的关系：

$$Similarity(e_t, e_c) = \frac{\overrightarrow{v_{e_t}} \cdot \overrightarrow{v_{e_c}}}{\|\overrightarrow{v_{e_t}}\| \|\overrightarrow{v_{e_c}}\|} \quad (2\text{-}10)$$

基于语言模型如 Word2Vec 或 BERT 的方法，能够在捕捉上下文信息的同时生成语义丰富的向量表示。例如，通过 BERT 的深层嵌入，我们可以在更加复杂

的语境中分析目标实体的语义特性。

上下文方法在开放域和无规则限制的任务中具有显著优势，但对数据质量和规模依赖较大。在语料稀疏或偏向性强时，这种方法可能引入噪声，导致扩展实体的质量下降。

2.2.3 基于深度学习的实体扩展

深度学习方法通过端到端的学习方式，捕捉实体与上下文之间的深层语义关系，为实体扩展任务提供了强大的工具。与规则方法和上下文方法不同，深度学习依赖神经网络的强大表达能力，能够在复杂的语言环境下提取出目标实体的特征，并识别与之相关的候选实体。这种方法广泛使用卷积神经网络、循环神经网络、长短期记忆网络及 Transformer 等架构。

深度学习的实体扩展流程可以概括为以下几个步骤。首先，文本中的每个词通过嵌入层映射到高维向量空间，形成特征表示：

$$Embedding(w) = W_e \cdot w \tag{2-11}$$

其中W_e是嵌入矩阵，w是输入的词索引向量。然后，使用 LSTM 或 Transformer 模型对输入序列进行上下文建模。LSTM 单元的状态更新公式如下：

$$h_t = \sigma(W_h x_t + U_h h_{t-1} + b_h) \tag{2-12}$$

其中h_t是当前时刻的隐状态，x_t是输入词向量，W_h、U_h和b_h是模型参数。在 Transformer 中，通过多头注意力机制，可以更灵活地捕捉目标实体与上下文的长程依赖关系：

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}} \cdot v) \tag{2-13}$$

其中Q、K、V分别是查询、键和值矩阵，d_k是向量的维度。

最后，模型的输出可以用于分类或生成。分类模型预测输入序列中的实体类别，而生成模型（如 GPT）则直接生成候选实体。通过迁移学习，预训练语言模型（如 BERT）可以快速适应特定任务，进一步提升扩展的精确性和鲁棒性。

深度学习方法在复杂任务和开放领域场景中表现卓越，但其高计算资源需求和对标注数据的依赖是需要解决的关键问题。在未来的研究中，通过结合知识图谱和少样本学习方法，深度学习模型有望在更广泛的任务中实现突破。

2.3 预训练模型

NLP 领域中的预训练模型（Pre-trained Model）是指事先已在大规模语料集上通过预训练任务进行训练的模型，同时保存模型预训练时的最佳参数，主要包含网络层数、神经元个数等。预训练模型能够部署在下游细粒度的 NLP 任务中，例如文本情感分析、文本分类、文本生成等，部署的过程也被称为微调（Fine Tune）。相比于传统机器学习和深度学习方法，预训练模型不仅部署简便，更是在 NLP 领域展现出了优越的模型性能，例如由谷歌团队在 2018 年发布的预训练模型 Bert[51]则是打破了十一项 NLP 任务记录，被誉为“NLP 领域的里程碑”。

在本文中，采用预训练模型用于后续候选生成和文本分类。因此，本节对所涉及的预训练模型给予理论介绍。

2.3.1 Bert

得益于注意力机制，Bert 模型能够动态学习到文本的上下文语义信息。图为 Bert 模型的结构图，其中 *Trm* 表示特征提取器 Transformer 中的编码器，*E* 表示输入序列，*T* 表示输出结果。有关 Transformer 的理论细节可参考文献[52]。在预训练阶段。Bert 模型的预训练任务如下：

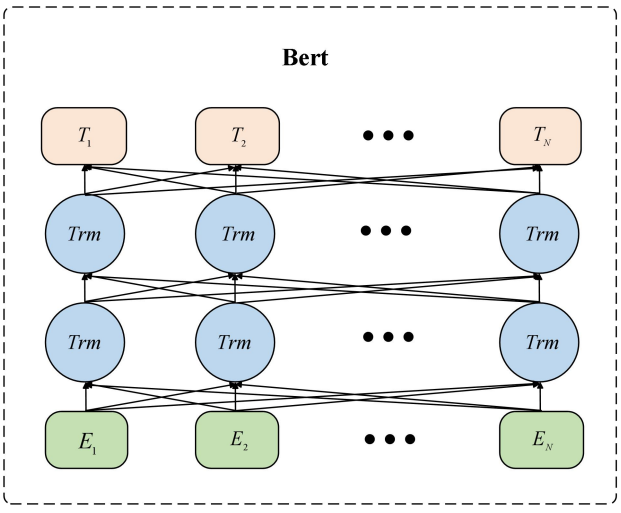

图 2-1 Bert 结构图

Fig. 2-1 The structure diagram of Bert

(1) MLM（Masked Language Model）：采用掩码“Mask”随机替换输入序列中的词汇，再根据已有的序列采用 Transformer 的编码器对已有的序列进行训练，

预测屏蔽词。

(2) NSP（Next Sentence Prediction），采用 Transformer 的编码器预测语句 B 与语句 A 的从属关系，即判断 A 和 B 在语料中是否为上下文关系。

在微调阶段。对于指定的 NLP 下游任务，首先需要按照 Bert 要求的格式预处理语料，然后在 Bert 模型的输出层添加指定的权重输出即可。例如针对句子对分类任务，Bert 模型最终输出的条件概率可表示为$P = softmax(CW^T)$其中C为输入语料的语义输出，W是 Bert 的参数矩阵。有关其他 NLP 下游任务的微调方式可在 Bert 原文中查询。

2.3.2 CPT

CPT（Chinese Pre-trained Transformer）专注于中文语言环境，是一种为中文自然语言处理任务量身定制的深度学习模型。受 BERT 语言模型启发，CPT 结合了 Transformer 架构与预训练技术，旨在为中文 NLP 任务提供高效的解决方案。通过在大规模中文语料上进行无监督预训练，CPT 能够深入捕捉中文语言的丰富语义与上下文信息。

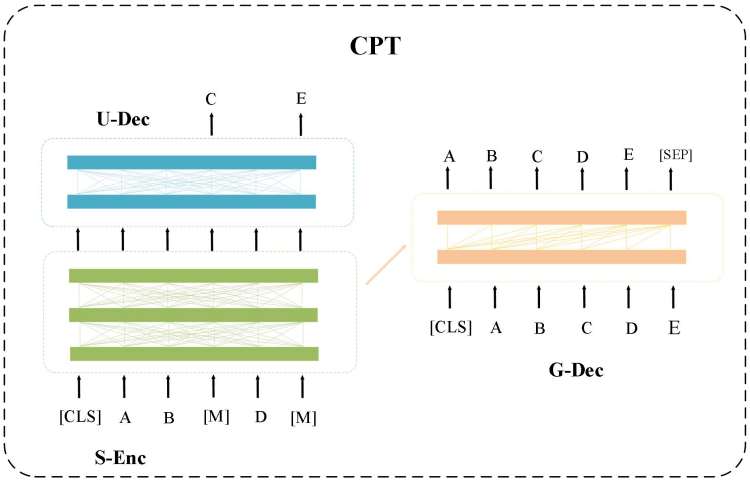

图 2-2 CPT 结构图

Fig. 2-2 The structure diagram of CPT

如图 2-2 所示，CPT 的架构是完整的 Transformer 的变体，由三个部分组成：

(1) 共享编码器 Shared Encoder（S-Enc）：具有全连接自注意力的 Transformer 编码器，旨在捕获语言理解和生成的公共语义表示。

(2) 理解解码器 Understanding Decoder（U-Dec）：一个具有全连接自注意力的浅 Transformer 编码器，专为 NLU 任务设计。U-Dec 的输入是 S-Enc 的输出。

(3) 生成解码器 Generation Decoder（G-Dec）：一个具有掩蔽自注意的 Transformer 解码器，它是为具有自回归方式的生成任务而设计的。G-Dec 利用交叉注意的 S-Enc 的输出。有了这两个特定的解码器，CPT 可以灵活地使用。例如，对于 NLU 任务，仅使用 S-Enc 和 U-Dec 就可以轻松微调 CPT，并且可以将其视为标准的 Transformer 编码器;而对于 NLG 任务，CPT 采用 S-Enc 和 G-dec，并形成 Transformer 编解码器。通过不同的组合，CPT 能够有效地应用于各种下游任务，充分利用预先训练参数，并获得有竞争力的性能。与大多数带有编解码器的 PTM 不同，我们为共享的编码器和解码器开发了一个深-浅框架。更具体地说，我们使用一个更深的编码器和两个浅的 CPT 解码器。我们假设浅层解码器保留了文本生成的性能并减少了解码时间，这已被证明对神经机器翻译[27]和拼写检查[28]有效。深-浅设置使得 CPT 对于具有较小参数开销的理解和生成任务更加通用。它还加速了用于文本生成的 CPT 的推理，因为 G-Dec 是一个轻型解码器。

2.4 本章小结

本章系统介绍了实体扩展技术的相关基础理论与技术，为后续研究提供了坚实的理论与技术支持。首先，阐述了实体识别与实体链接的基本概念及其在自然语言处理中的核心地位，并分析了知识库在实体扩展中的重要作用，包括通用知识库（如 Wikipedia、DBpedia）与领域知识库（如 UMLS）的应用场景和局限性。其次，详细介绍了几种主流的实体扩展方法，包括基于规则的实体扩展、基于上下文的实体扩展，以及近年来广泛应用的基于深度学习的实体扩展方法，深入探讨了这些方法的特点、适用场景和技术难点。最后，重点解读了预训练模型（如 BERT、GPT）的理论基础及其在实体扩展任务中的应用优势，通过预训练与微调的结合，有效提升了模型在语义理解和生成上的表现。本章内容为后续对实体扩展关键问题的深入研究和创新方法的提出奠定了全面的理论基础和技术框架。

第 3 章　基于灵活感知域和多层上下文的实体同义词生成

3.1 引言

实体同义词生成是自然语言处理领域中的重要研究课题，其核心目标是识别和归纳具有相同语义或指代同一事物的不同表达形式，例如“USA”、“U.S.”和“United States”均表示同一国家。实体同义词的生成在搜索引擎、问答系统、分类构建等下游任务中具有广泛应用价值。然而，现有方法在充分利用候选实体的全局信息以及应对误差传播等方面仍存在不足。本研究提出了一种名为 EnSynFields 的框架，通过三层特征交互域网络、灵活感知网络分类器及动态权重集合生成算法，旨在改进实体同义词的生成质量。

3.1.1 背景描述

随着智能系统对多样化语言表达需求的断增加，实体同义词生成逐渐成为解决自然语言理解问题的核心环节。例如，在分析用户查询“美国的首都是华盛顿”时，系统需要正确识别“美国”与“United States”的同义关系，以及“华盛顿”与“Washington”的指代对应性。这种理解能力对于满足用户的语义需求至关重要。

目前，关于实体同义词生成的研究主要可以分为两类方法：排序与剪枝方法以及两阶段任务方法。排序与剪枝方法：通过对候选词条按其可能性进行排序，并对排序结果进行剪枝，最终输出实体同义词。然而，这类方法往往独立地处理每个候选实体，忽视了全局上下文信息对候选实体关联性的潜在提升作用。两阶段任务方法：将实体同义词的生成过程分为两个步骤：首先进行同义词对的检测，然后将这些候选词对组合生成同义词。然而，这种方法缺乏跨阶段的反馈机制，容易受到第一阶段错误的累积影响，从而降低生成结果的准确性。

上述方法在处理复杂场景时面临两大主要局限：一是忽视候选实体间的全局信息交互，二是误差传播问题。因此，亟需一种能够综合全局信息并减少误差积累的新方法。

3.1.2 问题分析及解决思路

排序与剪枝方法在计算每个候选实体的关联性时，采用独立的概率计算方式，无法捕捉候选实体之间的全局关系。这一局限性可能导致全局信息缺失，从而使得一些高质量的候选实体由于缺乏上下文支持而被遗漏。

在两阶段方法中，候选词对的生成和最终同义词的生成是两个独立的过程，缺乏阶段间的动态反馈机制。第一阶段的错误会直接影响第二阶段的生成结果，导致误差传播并逐渐积累，从而影响整体性能。

为克服上述问题，本研究提出了 EnSynFields 框架，包含以下关键设计：

（1）三层特征交互域网络:通过构建实体层特征、集合层特征和句子层特征三层特征交互域网络，捕获候选实体之间的全局信息交互。通过双向传播机制，使不同层次的信息相互补充，显著增强实体特征表示。

（2）灵活感知域神经网络分类器:设计一种灵活感知域的神经网络分类器，全面建模实体、集合和句子间的关系。分类器在综合多层信息的基础上，动态决定是否将候选实体加入同义词。

（3）基于动态权重的集合生成算法:提出一种基于动态权重的集合生成算法，通过加权交叉熵损失函数平衡样本分布，提高生成同义词的质量，减小分类阶段误差传播的影响。

通过三层特征交互域网络联合学习实体和同义词的表示，本框架综合了多层次的上下文信息。灵活感知域分类器能够有效处理实体、集合和句子之间的复杂交互，全面建模这些关系。而基于动态权重的集合生成算法则增强了生成效果，并有效缓解了误差传播问题。EnSynFields 框架能够有效应对现有方法在实体同义词生成中面临的全局信息缺失与误差传播问题，提升生成结果的准确性和覆盖性。

3.2 EnSynFields 框架

通过上文分析，本节为解决好实体语义缺失和错误传播积累问题，设计了一个高效的实体同义词生成框架 EnSynFields，其名称来源于英文单词的组合：Entity Synonym Fields，意为“实体同义词域”。该名称反映了本方法的核心思想

—通过构建多个特征域（包括实体层、集合层和句子层），综合利用多层上下文信息来进行实体同义词扩展。其中，“En”取自“Entity”，“Syn”代表“Synonym”，“Fields”则强调该方法中多特征交互域的关键作用。其模型框架如图 3-1 所示：EnSynFields 框架包括三个关键部分：三层特征交互域网络、灵活感知域神经网络分类器和基于动态权重的集合生成算法。

（1）三层特征交互域网络：构造交互网络，用于捕获候选实体之间的全局信息。交互场网络包含实体层、集合层和句子层的上下文信息，可以减少单个场引起的潜在噪声。

（2）灵活感知域神经网络分类器：该分类器的开发目的是评估是否应将新的候选实体纳入当前实体集中。它具有一个具有灵活感知域的网络，能够学习额外的上下文特征。

（3）基于动态权值的同义词生成算法：设计了一种基于动态权值的实体同义词扩展算法。它重新采样生成的数据，并采用加权交叉熵损失函数，旨在平衡分布在不同的集合，从而提高实体同义词扩展任务的效率。

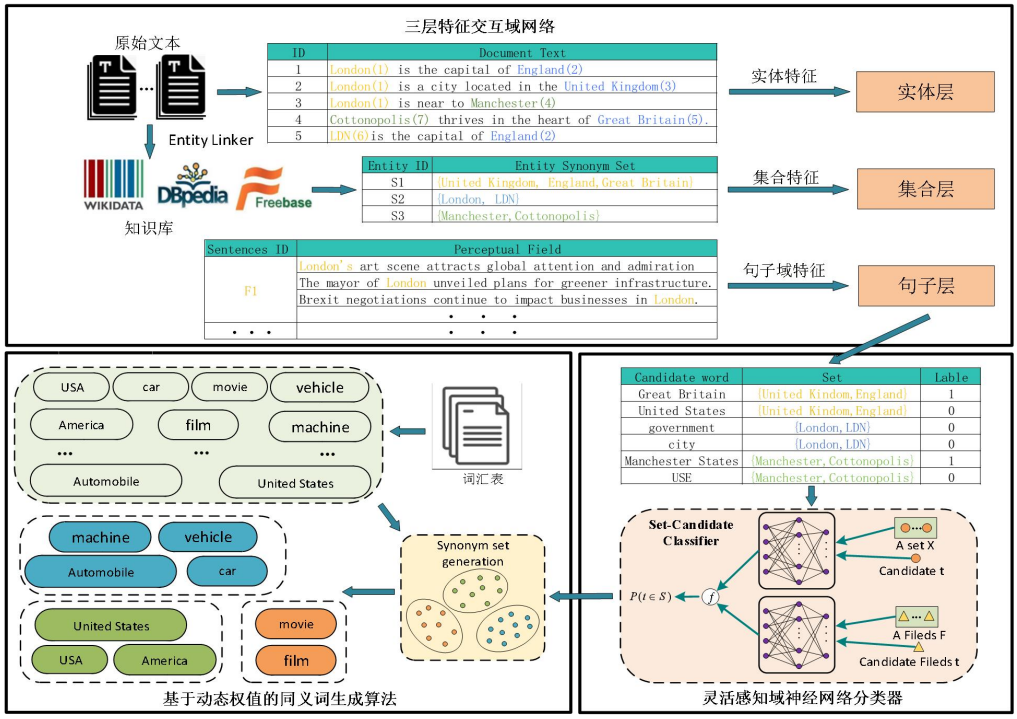

图 3-1 EnSynFields 框架概述

Fig.3-1 EnSynFields Framework Overview

3.2.1 三层特征交互域网络

如图 3-2 所示，三层特征交互域网络$N\{f(M-M),f(S-S),f(T-T)\}$由实体层场$f(M-M)$、集合层场$f(S-S)$和句子场$f(T-T)$组成。三层特征交互域网络保留了实体层、集合层和句子层的上下文信息，为实体、集合和句子的额外编码提供了灵活的接受场。

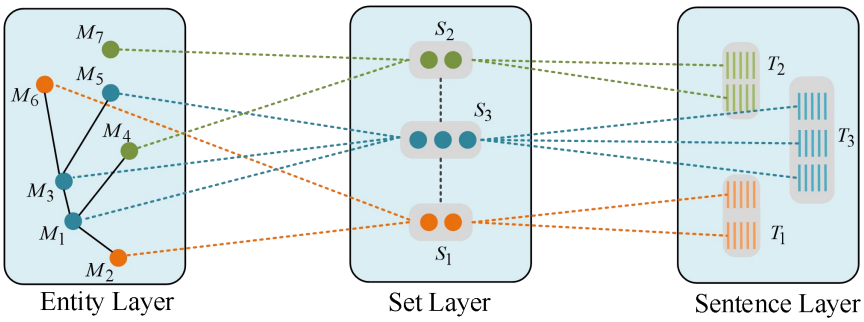

图 3-1 三层特征交互域网络框架

Fig.3-2 Three-layer feature interactive field network architecture

（1）数据预处理

为了构建同义词生成模型，本章节从数据预处理入手，重点介绍训练和验证数据的收集过程，并采用以下方法进行数据获取。

实体数据采集：现有的命名实体识别工具[50]用于标识文本语料库中提到的实体，即实体引用。

实体集数据获取：已建立的实体链接器，例如 DBPEDIA Spotlight[53]被用来将语料库中的实体引用与知识库中的对应实体相关联。来自链接语料库的同义词种子集可以提供远程监控信号，以帮助我们从原始文本语料库中发现不在知识库中的其他实体同义词。

句子集数据获取：BM25 信息检索算法[54]被用来从原始文本语料库中检索五个最相关的句子。之所以选择这个数字，是因为它平衡了提供足够的上下文与保持计算效率。这些选择的句子基于诸如词频、句子长度和句子频率的因素来评分，并且被用作用于实体的句子集。

（2）构建实体层特征网络

构建实体层特征网络的目的是建立实体到实体的语义关系。如图 3-3 所示，实体提及M是网络的节点。为了捕获实体到实体的语义关系，我们使用以下规则构建实体到实体的语义场网络$f(M-M)$：

规则 1：如果M_i是文本语料库中最接近M_j的实体，那么我们使用实体M_i和M_j的$Top-k$上下文构建网络$f(M_i-M_j)$。例如，M_1和M_2的网络是$f(M_1-M_2)=\{C_1,C_2,C_3,C_4,C_5,C_6\}$。

规则 2：如果M_i和M_j位于文本语料库中的同一个句子中，那么我们使用M_i和M_j的实体的$Top-k$上下文来构建网络$f(M_i-M_j)$。例如，M_2和M_3的网络是$f(M_2-M_3)=\{C_4,C_5,C_6,C_7,C_8,C_9,C_{10}\}$。

实体层特征网络隐式地对实体之间的语义信息进行编码，从而促进基于该特定层的语义关系的全面捕获。

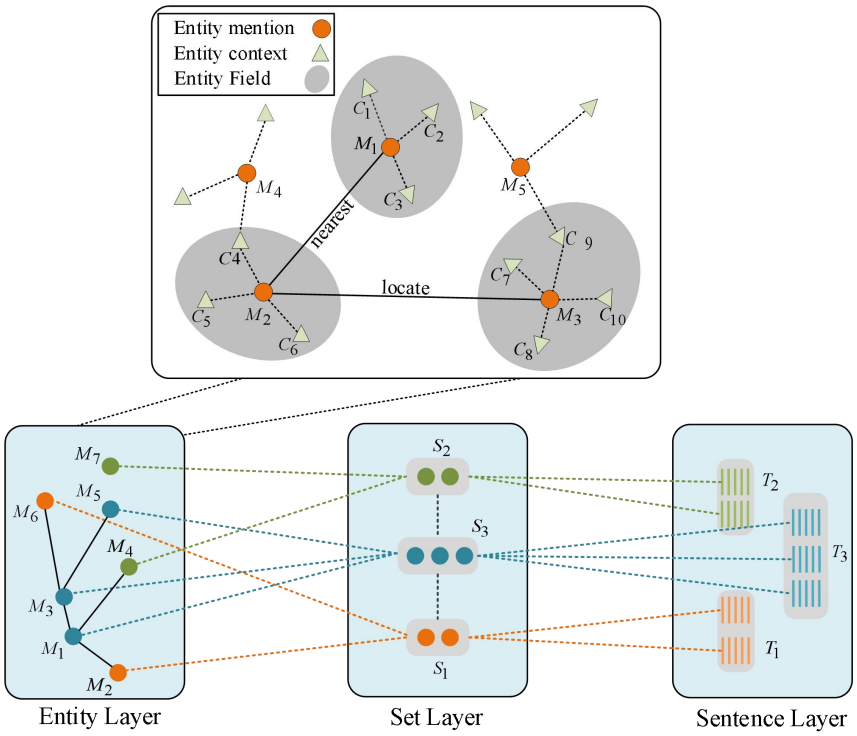

图 3-3 实体层特征网络

Fig.3-3 Entity-layer field network

（3）构建集合层特征网络

本节构建了集层特征网络来获取集层上下文信息。如图 3-4 所示，集合层特征网络由许多候选实体（candidate Entity）组成。为了捕捉集合到集合的语义关系，我们使用以下步骤构建集合到集合的语义场网络$f(S-S)$，如果实体提及M_i和M_j是同义词，则我们构造集合$S=\{M_i,M_j\}$以形成集合层特征。例如，$S_1=\{M_1,M_2\}$和$S_2=\{M_3,M_4,M_5\}$是两个集合特征。如果实体提及$M_i\in S_1$和$M_s\in S_2$是

最近的或位于实体层特征网络中，则我们构造一个集合层特征网络 $f(S_1 - S_2)$，并认为S_1和S_2是关联的。例如，S_1和S_2的网络是 $f(S_1 - S_2) = \{M_1, M_2, M_3, M_4, M_5\}$。

集合层特征网络隐式地在集合层提供上下文信息。这确保了所生成的同义词与它们的前后文本错综复杂地连接，从而增强了同义词生成的一致性和相关性。

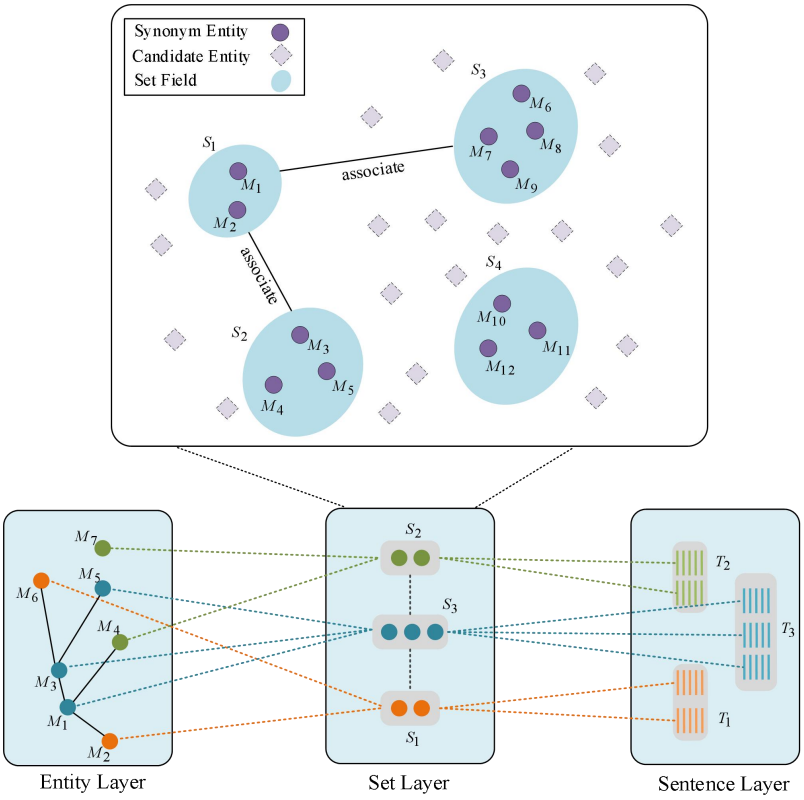

图 3-4 集合层特征网络

Fig.3-4 Set-layer field network

（4）构建句子层特征网络

本章节采用 BM25 信息检索算法来评估句子和实体提及（Entity Mention）之间的匹配度。此外，我们还综合考虑实体提及频率、句子长度和句子频率等因素，以确定中间层特征网络$f(T - T)$的上下文信息。

具体地，首先，计算句子的实体提及频率（术语频率TF）。TF表示在句子的上下文中提到的实体的重要性。其次，计算实体提及的逆句频。ICF表示实体提及项的稀有度，即，它在整个句子中的重要性。第三，基于TF、ICF和实体提及的其他参数计算 BM25 分数。BM25 的计算公式如下：

$$BM25 = ICF \cdot \frac{TF \cdot (K_1 + 1)}{TF + K_1(1 - b + b \cdot sen\_length / avg\_sen\_length))} \tag{3-1}$$

其中CF指示实体提及项的逆句子频率。TF表示实体提及词在句子中的词频。

K_1和B是缓和参数。sec_length表示句子的长度。avg_sec_length表示平均句子长度。最后，通过基于 BM25 分数对文本句子进行排序来执行排名。得分较高的句子在结果中的位置较高。基于语义层特征网络过滤出得分最高的前K个句子。

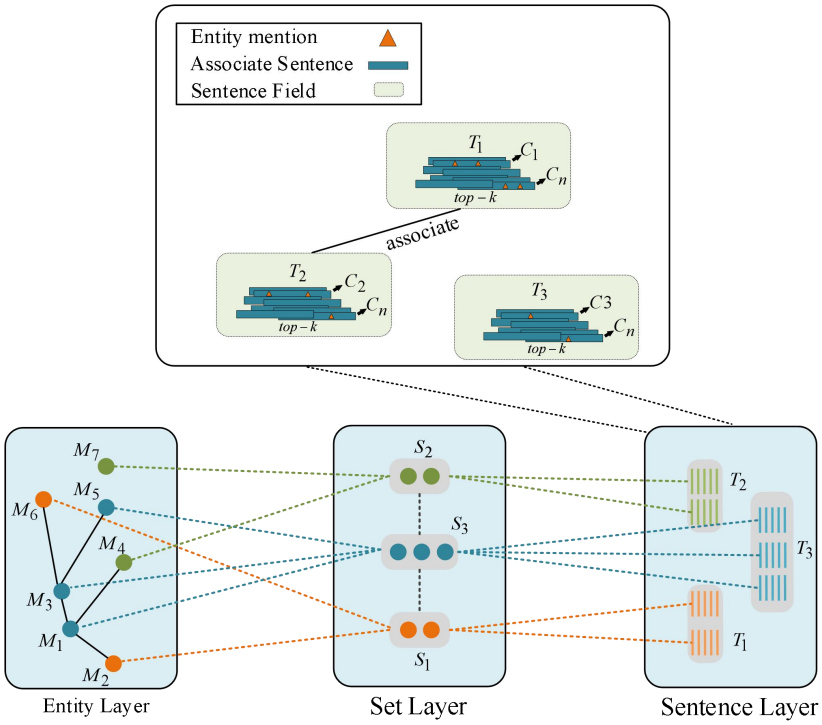

图 3-5 句子层特征网络

Fig.3-5 Sentence-layer field network

本章节构建句子层特征网络来捕获句子层上下文信息。如图 3-5 所示，句子层特征网络由许多候选句子组成。为捕捉句子到句子的语义关系，我们使用以下过程构建句子到句子的关联语义场网络$f(T-T)$：

首先，使用 BM25 信息检索算法，以获得最高的相关性的前K个句子作为实体的句子字段。其次，如果实体提及$M_i \in C_1$ $(C_1 \in T_1)$和$M_s \in C_s(C_2 \in T_2)$在集合层网络中相关联，则我们构造一个集合层网络$f(T_1-T_2)$。例如，C_1和C_2的网络是$f(T_1-T_2)=\{C_1, C_2, \cdots C_n\}$。

采用 BM25 算法的句子层特征网络能够有效过滤和排序候选句子，确保所选上下文信息与实体的语境高度相关，从而更精准地反映实体在不同上下文中的用法和含义。这种方法使生成的同义词具有更丰富的语义细节，提高了其实用性和准确性。

3.2.2 灵活感知域神经网络分类器

如图 3-6 所示，在数据预处理和三层交互式网络的构建之后，我们构建了一个灵活的感知域神经分类器，表示为$f(S,m)$，确定同义词S是否应该包含候选实体m。

灵活感知域神经网络分类器的基本目标是评估候选实体是否适合包含在同义词合中。先前的研究证明，同义词学习的关键在于当引入新术语时，术语分类器对同义词的替换不变性。早期的方法首先利用集合标记器来确定原始同义词合的分数，然后继续计算新形成同义词合的分数（即，新项被添加到集合中）。为了将这两个分数之间的差异转化为概率，他们采用了一个包含单位的目标函数。然而，该方法仅利用实体嵌入特征来捕获同义词信号，忽视了实体集合层和句子层之间关于实体同义关系的上下文信息。实际应用中，实体同义词关系的判定不仅依赖于实体本身的特征，还需要结合其在集合和句子中的上下文信息，以更全面地理解实体间的关联性。

基于三层特征交互域网络，本文构建了一个灵活的感知域神经分类器，对实体、集合和句子进行整体建模。如图 3-6 所示，图的下层是分类器$f(S,m)$的整体框架，主要包括多个平行评分器$Score(x)$。上图显示了评分器$Score(x)$的具体架构，它将集合 x 作为输入，通过评分系统，最终产生质量分数 Q(x)。这个分数 Q(x)反映了集合 x 的完整性和一致性。

如图 3-6 所示，给定同义词集$S(s_1,s_2,\cdots,s_n)$，候选实体m和三层特征交互域网络$N\{field(M-M),field(S-S),field(T-T)\}$，我们按照以下步骤构建灵活感知域神经网络分类器：首先，将集合$S(s_1,s_2,\cdots,s_n)$和$S\cup m(s_1,s_2,\cdots,s_n,m)$输入到集合分数，以分别获得两个分数$Q(S)$和$Q(S\cup m)$。计算$Q(S\cup m)$和$Q(S)$之间的差值，表示如下：

$$D(S,m)=Q(S\cup m)-Q(S) \tag{3-2}$$

其中$Q(S)$是输入同义词集合S的质量分数，$Q(S\cup m)$是将候选实体m添加到同义词集合S中的质量分数。

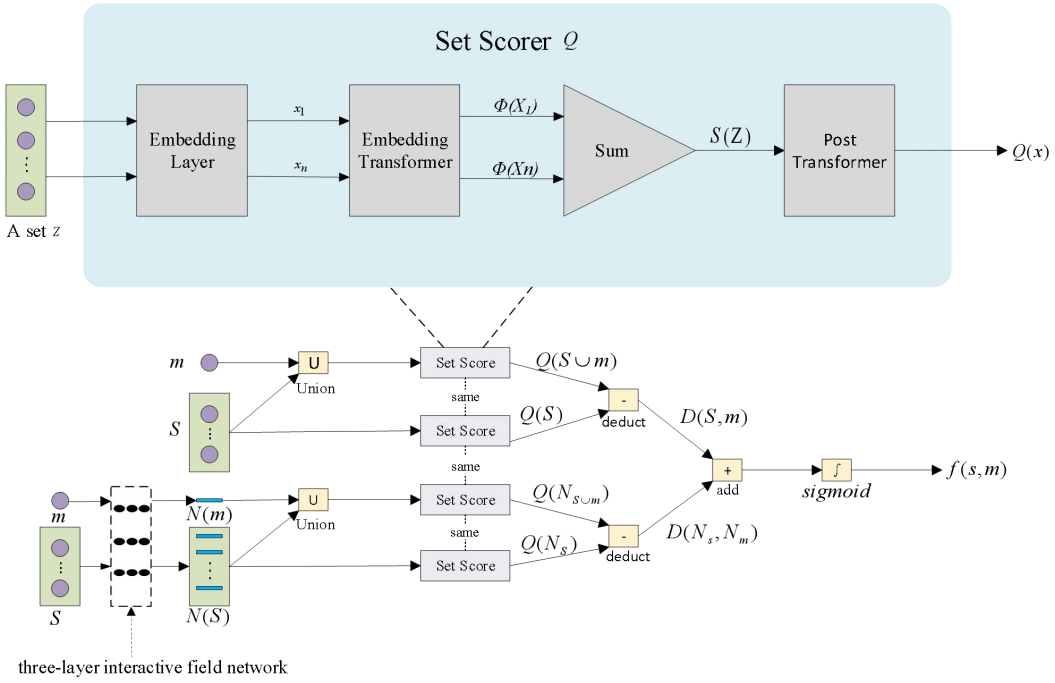

图 3-6 灵活感知域神经网络分类器架构

Fig.3-6 The architecture of flexible perceptual field neural network classifier

其次，在三层特征交互域网络的基础上，以相互增强的方式学习实体表示，得到一种灵活的感知域实体表示。将集合$N\{S(s_1,s_2,\cdots,s_n)\}$和$N\{S\cup m(s_1,s_2,\cdots,s_n,m)\}$输入到集合得分，以分别获得两个得分$Q(N_s)$和$Q(N_{S\cup m})$。计算$Q(N_s)$和$Q(N_{S\cup m})$之间的差，表示如下：

$$D(N_s,N_m)=Q(N_{S\cup m})-Q(N_s) \tag{3-3}$$

其中$Q(N_s)$是输入的灵活感知领域同义词集合N_s的质量分数，$Q(N_{S\cup m})$是将灵活感知领域候选实体N_m添加到同义词集合N_s的质量分数。

计算这两个差异质量分数的总和，并使用$sigmoid$函数将它们的质量分数之和转换为概率。为了确定候选实体是否应该被添加到同义词集合，概率$P(m\epsilon s)$表示如下：

$$P(m\epsilon s)=f(s,m)=sigmoid(D(S,m)+D(N_s,N_m)) \tag{3-4}$$

其中$sigmoid=\frac{1}{1+e^{-x}}$是非线性函数，$D(S,m)$是添加候选实体之后的差异质量分数。$D(N_s,N_m)$是在将候选实体添加到灵活感知域之后的差异质量分数。

我们使用 log-loss 作为损失函数来训练我们的分类器$f(S,m)$。$\vartheta(f)$表示如下：

$$\vartheta(f)=-log(f(S,m))-log(1-f(S,m))\cdot(1-y) \tag{3-5}$$

其中如果$m\epsilon S$，$y=1$，否则$y=0$。接下来，我们将描述质量评分体系结构$Q(\cdot)$。

3.2.3 基于动态权重的集合生成算法

在先前的研究中，引入了集合生成算法，利用集合评分器从候选实体$v_i \in V$生成同义词集合。该算法采用集合评分器来计算概率P以确定是否应该将新的候选实体v_i添加到同义词集合S_i中。如果P大于阈值α，则将实体v_i添加到集合S_i，否则创建新集合并将其添加到集合池$Pool$。

然而，该研究依靠实体嵌入特征来捕获同义词信号，忽略了集合层和句子层之间实体同义词的上下文信息。此外，随着集合数量的增加，集合生成算法可能面临不平衡的集合生成的问题，其中模型倾向于表现出对较大集合的偏向。此问题会导致集合生成算法的准确性降低。

为了解决上述问题，本研究提出基于动态权值的集合生成方法。该算法解决了集合生成中的不平衡问题，提高了集合生成的准确性。对于同义词较少的集合，动态权重可以增加其在训练中的影响，使模型更关注可能被较低频率集合模糊的重要特征。

为了解决不平衡问题的挑战，基于动态权重的集合生成算法采用了数据采样技术，并引入了加权交叉熵函数。该算法旨在通过调整集合权重来重新平衡样本在不同集合中的分布。通过在不同的集合中分配不同的权重，该函数被定制为引导模型有效地识别具有较少实体的集合。该算法根据同义词集合中实体的数量计算每个集合的权重。具体地，它利用逆类频率作为权重，即，计算实体总数除以每个集合中的实体数量。该算法的核心思想是：首先统计训练集中各类别的样本频率，并据此计算每类样本的逆类频率值ICF，随后，在训练过程中，模型对每个样本的损失值进行加权，其中权重即为其所属类别对应的ICF值。通过该机制，稀有类别样本在训练中将获得更高的关注，有效缓解了传统分类器易偏向高频类别的倾向，从而提高了同义词集合构建的鲁棒性与覆盖度。该算法通过向具有较少实例的集合分配较高权重来微调模型，这减轻了集合不平衡的影响。详细信息如下：

$$ICF = log(H/S\_c+1) \tag{3-6}$$

其中H是实体的总数，S_c是属于类别集合S的同义实体的数量。加 1 是为了避免

分母为 0 的情况。

然后使用加权交叉熵函数来考虑所设置的权重，该加权交叉熵函数将每个集合的交叉熵乘以其对应的权重并对所有集合样本进行平均。详情如下：

$$L = -\frac{1}{H}\sum_{i=1}^{H}(ICF_i \cdot Y_i \cdot log(P_i)) \quad (3\text{-}7)$$

其中，ICF_i表示第i个实体所属的集合的权重，Y_i表示真实标签，P_i表示预测概率。这种动态权重分配旨在通过在每次训练迭代期间自动调整各个集合的权重来改善不平衡的问题。因此，每次迭代中的候选实体主要与具有较大权重的同义词集进行比较。在本研究中，我们直接将概率阈值α设定为 0.5，并在后续分析中探讨其对聚类性能的影响。

算法 1：基于动态权重的同义词生成算法

输入:（1）灵活感知域神经网络分类器$f(S,m)$；（2）候选实体词汇表$V = (V_1, V_2, \cdots, V_n)$；（3）概率阈值$\alpha(0, 1)$。

输出： 实体同义词集$S = (S_1, S_2, \cdots, S_n)$，$S_i \subseteq V, \cup_{i=1}^{m} S_i = V, S_i \cap S_j = \emptyset$。

1 : 创建并添加第一个集群S_1；
2 : for 每个候选实体v_i in V do
3 : $best\_score \rightarrow 0$;
4 : $best\_j \rightarrow 1$;
5 : for each set S_j in Entity Synonym Set Pool do
6 : total_entities→ sum of entities in Synonym Set Pool;
7 : weight(S_i)→total_entities / (number_of_entities_in_set(S_i) + 1);
8 : S_{min_wight} →min(sets, key=get_set_weight);
9 : end for
10 : if $f(S_{min\_wight}, V_i) > best\_score$ then
11 : $f(S_{min\_wight}, V_i) \rightarrow best\_score$;
12 : $best\_j \rightarrow j$;
13 : end if
14 : if $best\_score > \alpha$ then
15 : S_{best_j}.add(V_i);
16 : else
17 : $S.append(\{S_i\})$;
18 : end if
19 : end for
20 : return S

如算法 1 所示，算法的输入包括三个部分：（1）灵活感知域神经网络分类器$f(S,m)$，（2）候选实体V_i的词汇表，（3）概率阈值α。该算法的输出是从词汇表$V =$

$(V_1, V_2, \cdots, V_n)$聚类的实体同义词集池$S = (S_1, S_2, \cdots, S_n)$。具体来说，该算法将$v_i \in V$枚举一次，并保留包含所有已识别集合$S_i$的池。对于词池中的集合，结合联合收割机动态权重分配，自动调整每个集合的权重，每次迭代输入的候选词优先与权重低的集合进行比较。对于候选实体v_i，我们应用灵活感知域神经网络分类器来计算将该实体添加到S中的每个识别集合中的概率。如果该概率超过阈值α，则将v_i添加到集合S_i。否则，如果概率阈值低于α，则使用该候选实体创建新集合$\{S_i\}$并将其添加到集合池。这个迭代过程一直持续到整个词汇表被遍历一次为止。该算法返回所有检测到的同义词集

3.3 实验与结果分析

在本节中，为了评估 EnSynFields 方法在生成同义词方面的有效性，我们在三个公共数据集上进行了实验。首先，我们概述了实验装置，然后介绍实验结果。最后，我们深入研究了 EnSynFields 的每个组件的综合分析，展示了几个具体的案例研究。

表3-1 使用数据集和知识库

Tab.3-1 Used datasets and knowledge bases

Dataset	Wiki	NYT	PubMed
#Documents	100,000	118,664	1,554,433
#Sentences	6,839,331	3,002,123	15,051,203
#Entity Mentions	98,664	31,702	96,588
#Entity Synonym Sets	4,920	1,494	17,972
#Entity.Fields	98,664	31,702	96,588
#Training Entities	8,731	2,600	72,672
#Training Synonym Sets	4,359	1,273	28,600
#Test Entities	891	389	1,743
#Test Synonym Sets	256	117	250
Knowledge base	Freebase	Freebase	NMLS

3.3.1 实验数据集

我们在三个公共数据集上评估了 EnSynFields。Wiki 数据集包含大约 100,000 篇从维基百科中提取的文章，总计约 6,839,331 个句子，并使用了 Freebase 知识库。NYT 数据集包括 118,664 篇来自《纽约时报》的新闻文章，共 3,002,123 个句子，同样采用了 Freebase 知识库。PubMed 数据集由来自公共医学研究所的约

150 万篇论文摘要组成，总计 15,051,203 个句子，并应用了 UMLS 知识库。DBpedia[55]是 Wiki 和 NYT 数据集的实体链接器，而 PubMed 则使用了 PubTator[56]。此外，Freebase[57]和 UMLS[58]等知识库也被应用于这些数据集。在测试集创建过程中，我们随机选择了一部分链接实体，剩余的则用于训练。数据集的详细信息见表 3-1。

3.3.2 对比方法

本节使用六种对比方法。

•Kmeans[59]：Kmeans 是一种无监督的聚类中，该方法将候选实体嵌入作为算法的输入，并且输出是检测到的同义词的集合。它用于对实体词汇表中的同义词集进行聚类。

•Louvain[60]：Louvain 是一种基于模块化优化的网络社区结构发现的无监督算法，该方法构造了候选实体的图，其中每个候选实体代表一个节点，然后使用余弦相似度计算每个候选实体的嵌入，并且如果相似度超过设定的阈值，则向图添加边。

•SVM+Louvain[61]：SVM+Louvain 是一种两阶段的监督方法，第一阶段利用 SVM 进行同义词对预测，第二阶段利用第一阶段预测的同义词对构造一个图，并将其与 Louvain 算法相结合，得到一个同义词集。

•SetExpan+Louvain[61]： SetExpan+Louvain 是两阶段监督方法，初始阶段使用 SetExpan（弱监督集合扩展算法），以在词汇表中找到每个候选实体的 K 个最近邻居，然后构造 k-NN 图，并且在第二阶段，将上述 Louvain 算法应用于该图，以获得同义词集。

•COP-Kmeans[62]：COP-Kmeans 是该算法的半监督变体，其组合了用于聚类的 COP 和 Kmeans 算法，该算法利用约束信息指导聚类过程，提高了聚类的准确性和稳定性。该方法为每个数据集设置聚类的预言数 K，并将训练同义词转化为成对约束。

•SynSetMine[27]：SynSetMine 是一种两阶段的监督方法，用初始阶段训练分类器。在第二阶段中，采用集合生成算法来生成同义词集合。

•EnSynFields：EnSynFields 是我们提出的基于灵活感知域的实体同义词生成

方法。该算法结合多层上下文信息和基于动态权重的同义词生成算法，从候选实体词典中生成同义词。

3.3.3 超参数设置

对于实体嵌入，我们参考由 Shen 等人出版的 50 维实体嵌入方法[27]。为了对每个数据集的语料库进行公平比较，我们将关联项嵌入的维数固定为 50。为了调整模型的超参数，我们采用了五重交叉验证方法。对于 EnSynFields，我们使用了两个尺寸为 50 和 250 的隐藏层嵌入式转换器用于嵌入式特征层，并使用了三个尺寸为 250、500 和 250 的后隐藏层转换器。我们采用初始学习率为 0.001 的 Adam 优化算法来优化 EnSynFields 模型，表 3-2 中提供了其他模型超参数。

表3-2 超参数设置

Tab.3-2 Hyper-Parameter settings

Hyper-parameter name	NYT	WIKI	PubMed
Learning rate	0.001	0.001	0.003
Dropout rate	0.3	0.4	0.3
FieldSize	5	5	5
α	0.5	0.5	0.5

3.3.4 实验结果分析

实验的目的是评估实体同义词集生成的效率和有效性。为了充分理解实验结果，本节将分为两个主要部分：聚类性能和柔性感知域神经网络分类器预测性能。

表 3-3 是聚类性能的比较（最好的结果用下划线标记）。我们从结果中看到，EnSynFields 在三个公开数据集上的表现优于比较方法。

如表 3-3 所示，无监督方法 Kmeans 和 Louvain 的性能低于 EnSynFields，主要是因为它们不能使用标记信号，这限制了标签可用时的性能。与 EnSynFields 相比，SetExpan+Louvain 表现出较低的性能，但明显上级 Kmeans 和 Louvain，这是因为 SetExpan+Louvain 是一种两阶段监督算法，利用 K 最近邻信息沿着 Louvain 算法。与 Kmeans 算法相比，COP-Kmeans 算法集成了来自训练集的额外监督数据，从而通过明智的聚类决策来提高其性能。我们观察到 SVM+Louvain 的性能显着低于其他比较的方法，调味使用 SVM 捕获监督信息作为生成同义词的性能差。SVM+Louvain 的主要缺点是由于它们的学习模型没有整体集合视图，

并且基于成对相似性。

表3-3 实体同义词生成性能比较

Tab.3-3 Performance comparison of entity synonym set generation

Approach	NYT			Wiki			PubMed		
	ARI(%)	FMI(%)	NMI(%)	ARI(%)	FMI(%)	NMI(%)	ARI(%)	FMI(%)	NMI(%)
Kmeans	28.9	30.9	83.7	34.4	35.5	87.0	48.7	49.9	88.0
Louvain	21.8	30.6	90.1	42.3	46.5	92.6	46.6	52.8	90.5
SVM+Louvain	3.6	5.1	21.0	6.0	7.6	25.4	7.8	8.8	31.1
SetExpan+Louvain	43.9	44.3	90.3	44.8	45.0	92.1	58.9	61.9	92.2
COP-Kmeans	33.8	34.6	87.9	38.8	40.0	90.3	49.1	51.9	89.9
SynSetMine	44.9	46.4	90.6	56.4	57.1	93.0	74.3	74.5	95.0
EnSynFields	**48.6**	**50.0**	**91.5**	**58.7**	**60.5**	**93.8**	**77.6**	**78.2**	**96.3**

SynSetMine 在性能方面优于其他方法，但不及 EnSynFields。这是因为 SynSetMine 只利用了实体层的语义关系，忽略了实体的全局信息，这可以大大提高同义词发现的质量。上述分析表明，灵活的感知域神经网络分类器和基于动态权重的集合生成算法可以提高实体同义词扩展任务的效率。

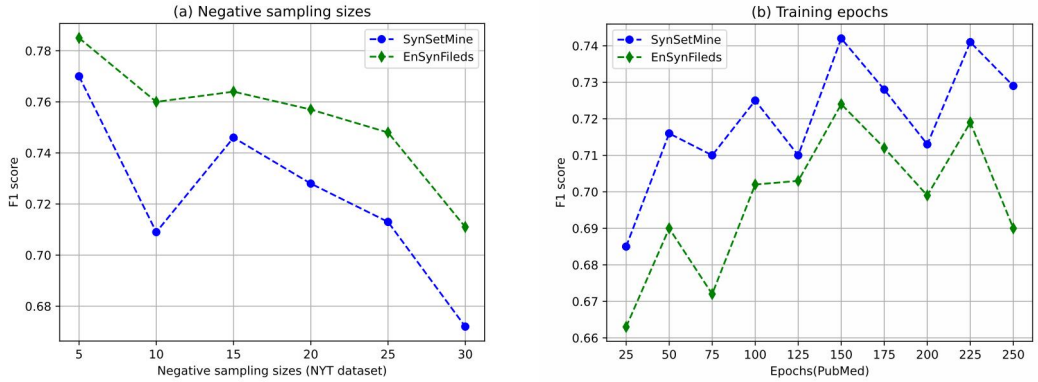

图 3-7 比较不同负样本大小和时期的 F1 分数

Fig.3-7 Comparing F1 scores across different negative sampling sizes and epochs

为了评估具有灵活感知域的神经网络分类器，使用来自 NYT 和 PubMed 的数据集比较了 EnSynFields 和 SynSetMine 的 F1 分数。图 3-7（a）显示了基于 NYT 数据集的不同阴性样本量的这些评分。观察结果显示，与 SynSetMine 相比，EnSynFields 始终获得更高的 F1 评分。如图 3-7（b）所示，这种趋势在 PubMed 数据集的各个训练阶段都很明显，EnSynFields 始终保持上级 F1 评分。以上结果表明，灵活感知域神经网络分类器可以将实体、集合和句子作为一个整体捕捉，并将其整合到动态权重整合生成算法中。这种集成有效地增强了模型对同义词的预测能力。

3.3.5 模型分析

本节将从多个角度分析隐藏层大小、嵌入维数和阈值α等各种参数如何影响性能。

(1)不同隐藏层大小的分析

为了更全面地评估不同隐藏层对评分器架构的影响，我们使用不同的隐藏层大小进行分析。对于评分器架构的嵌入式 Transformer，不同的隐藏层大小表示为 200、250、300、350、400。对于评分器架构的后隐藏层 Transformer，不同的隐藏层大小被表示为 200*2、250*2、300*2、350*2、400*2。图 3-8 显示了三个数据集上不同评分器架构的 ARI 和 FMI。可以观察到，当隐藏层大小从 200 转变为 250 时，每个架构的性能改进是显著的。然而，每种架构的性能通常随着隐藏层的增加而下降。因此，我们认为，嵌入 Transformer 隐藏层大小范围从 200 到 250，和 posthidden 层 Transformer 大小范围从 400 到 500，是足够的捕获所需的基本同义信号识别实体同义关系。

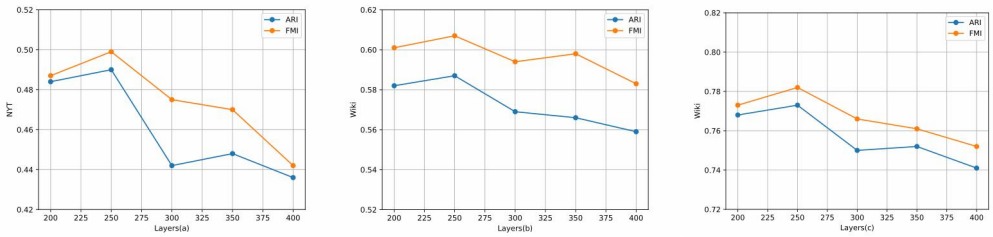

图 3-8 记分器架构的不同层大小的结果

Fig.3-8 Results of different layer sizes for scorer architecture

(2)不同 Embeddings 维度的分析

在表示学习中，嵌入的大小会显著影响机器学习模型的有效性。我们通过固定所有其他超参数并在 25，50，75，100，125，150 之间改变嵌入维数来探索这一点。图 3-9 展示了嵌入大小对使用 NYT 和 Wiki 数据集的 SynSetMine 和 EnSynFields 性能的影响。我们看到，EnSynFields 始终优于 SynSetMine，如果维度不够大，模型会出现欠拟合合并和收敛困难。例如，该模型在 25 维中不起作用。当维数大于 50 时，它缓慢增加。此外，实验结果表明，我们的模型在不同的尺寸下的稳定性。

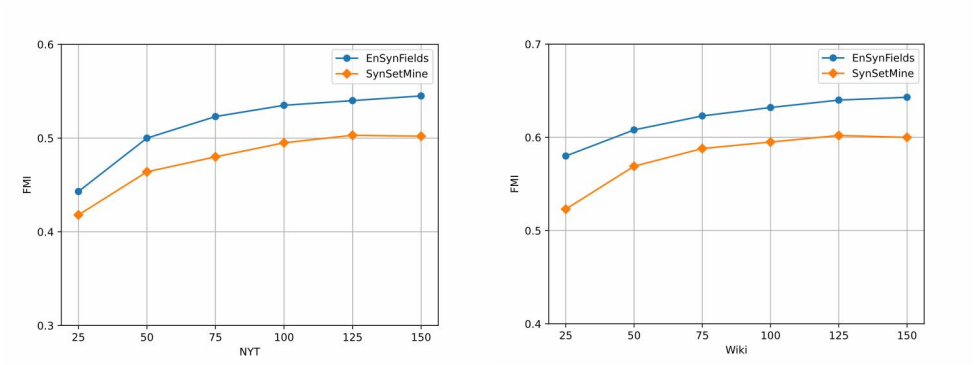

图 3-9 不同 embeddings 维度，不同方法在数据集 NYT 和 Wiki 上的结果

Fig.3-9 Results of different approach on datasets NYT and Wiki when varying the dimension of embeddings

（3）不同概率阈值α的影响

算法 1 概述了我们的集合生成方法，该方法使用概率阈值来决定在现有集合中包含候选实体。较高的阈值意味着算法采用更保守的方法，导致创建额外的集合。这个实验的重点是检查这个超参数对实验结果的影响。我们使用一个灵活的感知域神经网络分类器进行集生成算法，并改变阈值来观察其效果。图 3-10 显示了结果。最初观察到的是在不同阈值下聚类有效性的稳定性，0.4 至 0.6 的范围通常更合适。此外，0.5 的阈值是一个稳健的选择，在不同的数据集上始终产生有利的结果。

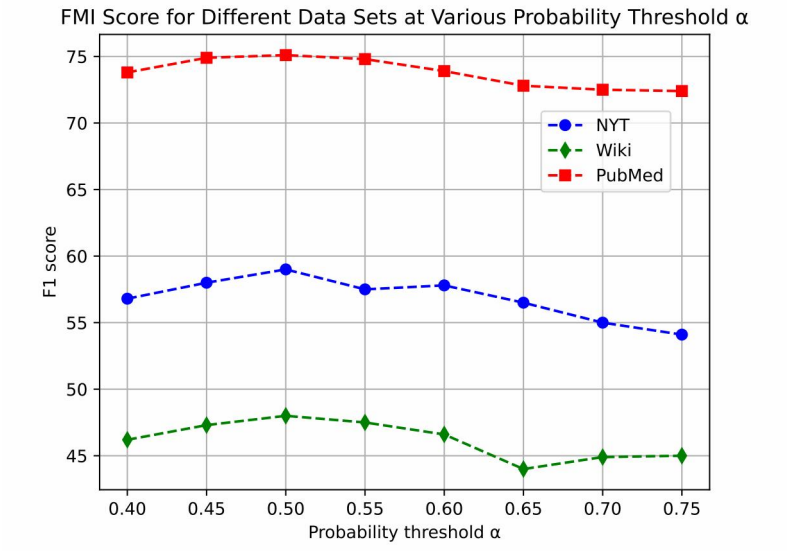

图 3-10 不同概率阈值的结果α

Fig.3-10 Results of different probability thresholds α

3.4 本章小结

本节探讨了一种从文本语料库中提取实体同义词的方法。该方法构建了一个三层特征交互网络，旨在捕捉实体层、集合层和句子层之间的上下文信息。为了判断候选实体是否应被纳入同义词集合，提出了一种灵活的感知域神经网络分类器，并结合了基于动态权重的实体同义词生成算法。在评估该方法时，实验分别在三个公共数据集上进行，并与多种现有的最先进方法进行了对比。实验结果表明，该方法在实体同义词生成方面的性能优于当前的最先进技术。

在后续的研究中，我们将计划重点探讨如何利用实体的属性信息来提升实体同义词生成方法的效率。此外，还将设计一种多模态策略，以有效区分近义实体与同义实体，并处理词汇的多义性问题，从而提高实体同义词扩展的准确性。

第 4 章 基于多特征融合和全局上下文的中文缩略词预测

4.1 引言

现有方法通常依赖序列标注进行缩略词预测，并且仅关注与实体本身相关的局部上下文信息，导致无法保证所生成的缩略词保持原有含义。为此，本节提出了一种基于多特征融合与全局上下文信息的创新型中文缩略词预测方法。具体而言，该方法首先利用预训练的汉语非平衡 Transformer 生成模型生成多个候选缩略词；随后，通过多特征优化阶段筛选出高质量候选项；最后，在全局上下文中对候选项的质量进行综合评估。实验结果表明，该方法在缩略词预测任务中优于现有同类方法，有效提升了预测准确性与语义一致性。

4.1.1 背景描述

缩略词是通过压缩或省略部分单词或短语而形成的一种简化表达形式，在自然语言处理中具有重要意义。缩略词的广泛使用不仅可以提高语言表达的简洁性，还在信息检索、实体链接及问答系统等领域中发挥了关键作用。例如，“ABC”代表“农业银行（Agricultural Bank of China）”，而“清华”是“清华大学（Tsinghua University）”的缩写。在信息检索（IR）中，通过扩展查询中包含的缩略词，可以有效提升搜索引擎的查询处理能力及结果的相关性，从而更好地满足用户需求。

然而，与英语缩略词通常基于全称首字母规则构成不同，中文缩略词的形式更加复杂和多样化。中文缩略词可能由连续字符、间隔字符甚至具有语义关联的词语组合而成。例如，“非典型肺炎（Severe Acute Respiratory Syndrome）”的缩写“非典”由词语的前两个字符构成，而“人民代表大会（National People’s Congress）”的缩写“人大”则通过选择不连续的关键字符实现。此外，中文缩略词还可能表现出高度语义化特征，例如“数学期望（mathematical expectation）”的缩写“期望”。这种复杂性和多样性使得中文缩略词预测成为一项具有挑战性的研究任务。

目前，大多数中文缩略词预测方法将该问题建模为序列标注任务，即预测全称中每个字符是否被保留。例如，条件随机场（CRF）模型通过对字符序列进行

标注，捕获字符之间的相互作用和依赖关系。随着深度学习技术的发展，双向长短时记忆网络（Bi-LSTM）等模型被广泛应用于缩略词预测任务，通过自动提取连续文本特征对字符进行二分类。然而，现有方法在捕获语义关联和上下文信息，以及处理标签依赖性等方面存在明显的局限性。

4.1.2 问题分析及解决思路

传统中文缩略词预测方法忽视语义关联和上下文信息。通常仅关注实体的表面特征（如字符频率和长度），而忽略了实体间的语义关联性和上下文信息。例如，在预测“国际足球联合会（FIFA）”的缩写“国际足联”时，仅依赖字符表面的特征可能无法充分理解其语义关联。

现有序列标注方法通常依赖简单的转移矩阵来预测当前字符标签，而未充分利用标签之间的复杂依赖关系。这种局限性可能导致模型无法准确捕捉字符之间的交互，降低预测的准确性和鲁棒性。

中文缩略词的生成过程往往涉及复杂的字符选择规则和语义信息，而现有方法在处理不规则字符选择及语义化缩略词时表现出局限性，难以生成语义一致且具有代表性的高质量缩略词。

为了解决上述问题，本章节提出一种基于多特征融合和全局上下文的中文缩略词预测方法，通过以下三个阶段改进预测过程：

（1）候选生成模块

利用最新的中文预训练不平衡变换器（CPT）模型生成多个候选缩略词，提升候选集合的多样性和覆盖率，为后续优化提供充分的选择空间。

（2）多特征优选模块

结合词性、笔画和词典特征对每个候选缩略词进行评估，赋予候选词质量得分，从中筛选出最具语义关联性和上下文适应性的候选词。通过多特征融合，显著提升缩略词的预测准确性和语义一致性。

（3）全局上下文评估模块

在全局上下文评估阶段，利用对比损失函数训练模型，整合全局上下文信息以确保候选缩略词的语义一致性。通过结合上下文信息为缩略词分配权重，进一步优化预测结果，最终选择质量得分最高的缩略词作为输出。

本研究在预测过程中引入多维特征，包括语义、字符结构和词典信息，全面提升模型的语义理解能力，借助对比损失函数和全局上下文信息，改善缩略词预测的语义一致性，使得预测结果更符合实际语义表达，并通过候选生成、多特征优化和上下文评估三个阶段的联动，确保候选缩略词生成的多样性及最终结果的高质量。综上所述，本研究提出的方法通过整合多层次的特征与全局信息，克服了现有方法的局限性，为中文缩略词预测提供了新的思路，并在准确性和鲁棒性方面实现了显著提升。

4.2 中文缩略词预测框架

如图 4-1 所示，中文缩略词预测框架包括三个主要组件：候选缩略词生成、多特征优化和全局上下文评估。

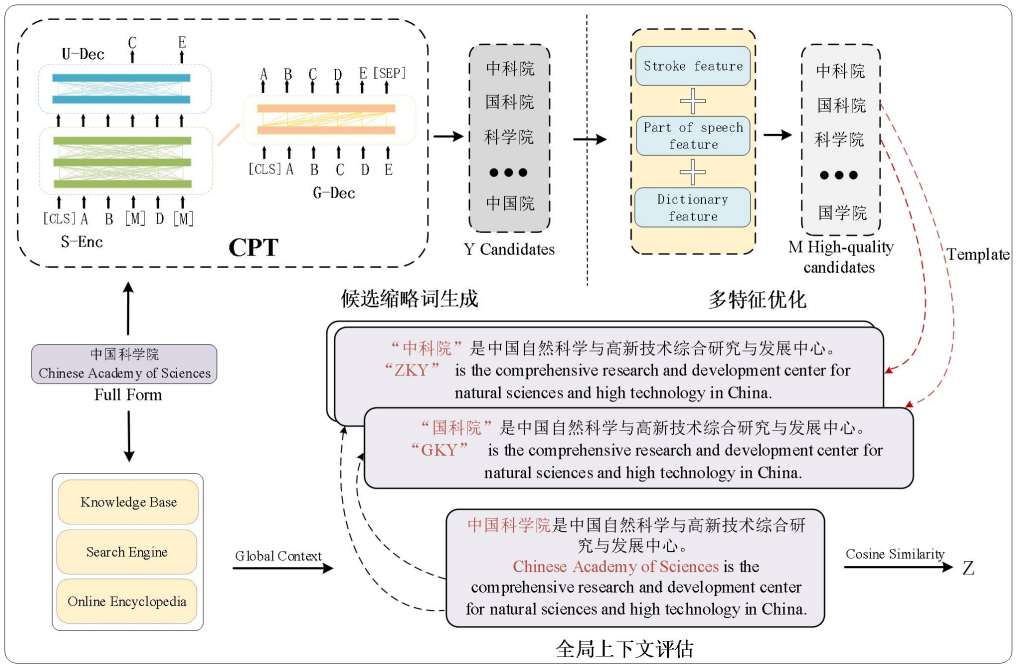

图 4-1 中文缩略词预测框架的总体架构

Fig.4-1 The overall architecture of the Chinese abbreviation prediction framework

(1)候选缩略词生成：给出一个缩略词$E = [E_1, E_2, \cdots, E_x]$的完整形式，使用中文预训练模型 CPT 生成$Y$个候选缩略词$N = [N_1, N_2, \cdots, N_y]$。

(2)多特征优化:给定Y个候选缩略词$N = [N_1, N_2, \cdots, N_y]$，综合候选缩略词的词性特征、笔画特征和词典特征，剔除低质量的候选缩略词。最终，我们获得 M 个高质量候选缩略词$N = [N_1, N_2, \cdots, N_m]$。

(3)全局上下文评估:我们采用 BM 25 算法通过将上下文嵌入模板来收集广泛的上下文信息。然后将这些填充的模板放置在上下文之后，以丰富每个候选者的上下文。我们使用对比学习来训练评估模型，使用它们的余弦相似度作为评估分数。评估模型给有效度高的候选词分配更高的分数。

4.2.1 候选缩略词生成

在本节中，我们的目标是生成各种潜在的缩写，以实现高召回率。

我们利用基于 transformer 的人工智能模型 CPT，在中文文本数据上进行预训练，作为我们的生成模型。如图 4-1 所示，CPT 的架构是完整 Transformer 的变体，由三个部分组成：共享编码器（S-Enc）、理解解码器（U-Dec）和生成解码器（G-Dec）。

具体来说，我们将 CPT 的嵌入层表示为$Embedding(\cdot)$，将输入序列转换为 Transformer 块的向量。Transformer 编码块由$TransEnc(\cdot)$表示，而 Transformer 解码块由$TransDecE(\cdot,\cdot)$表示。给定一个缩写 E 的完整形式，CPT 的具体过程如下：

$$T_0^e = Embedding(M) \tag{4-1}$$

$$T_i^e = TransEnc(T_{i-1}^e), i\epsilon[1, X] \tag{4-2}$$

其中 X 表示 Transformer 编码块的数量，T_i^e表示 Transformer 编码器第 i 层的隐藏状态。第 t 个时间步的解码过程概述如下：

$$T_0^d = Embedding(N < t), t\epsilon[1, y] \tag{4-3}$$

$$T_i^d = TransDec(T_{i-1}^e, T_X^e), i\epsilon[1, Y], Y < X \tag{4-4}$$

$$N = g(T_Y^d) \tag{4-5}$$

其中Y表示 Transformer 解码块的数量，T_i^d表示解码器的i层处的隐藏状态，并且$g(\cdot)$表示用于从词汇表中选择令牌的 softmax 分类器。T_i^d用作输入，N表示输出序列的第T个令牌，其中令牌在输出序列中的第 t 个令牌之前被考虑。因此，CPT 能够生成Y个候选缩写$N = [N_1, N_2, \cdots, N_y]$。

4.2.2 多特征优化

给定全称缩写 E 及其Y个候选缩写 N，多特征优化阶段的目标是识别高质量的候选。随着 CPT 模型生成的候选缩写Y的数量增加，一些标记可能不对应于缩写的完整形式。因此，我们使用多特征优化阶段来细化和选择有价值的候选项。

在本节中，我们根据特定的优先级顺序对高质量的候选缩写进行优先级排序和选择，直到识别出总共M个候选缩写。

•词性特征：通常，实体名称的名词部分是最独特和最具指示性的，使其成为缩写生成期间保留的主要候选者。动词和形容词，虽然提供不同的性质或行为，通常被排除在缩略词之外，因为它们可能缺乏代表性。本文使用中文分词工具 jieba 1 对候选缩略词进行分词和词性标注。每个词性都被赋予一个权重W，名词的权重更高。我们计算每个候选缩写中词性的累积权重。如果权重超过某个阈值，则将候选项标记为高质量；否则，将其分类为低质量。

•笔画特征：汉字的笔画数影响缩略词的生成，偏向于笔画较少的汉字。例如，在《北京大学(北京大学)》中，笔划是：北(5 划)、京(8 划)、大(3 划)和学(7 划)。选择笔画较少的字符将得到缩写“北大(北大)”。对于候选缩写，我们计算每个字符的平均笔画数。已经为行程计数设置了一个阈值，表示为α。平均笔划数低于此阈值的候选被标记为高质量，而超过阈值的候选被标记为低质量。

•词典特征：字典功能可以指示实体名称是否在专用字典或特定于域的词典中列出。特定领域的标准专业术语或专有名词可能具有遵循特定惯例或规定的缩写。对于每一个候选缩写，我们检查它的存在，在一个中文缩写字典。如果候选缩写在字典中不存在，则将其标记为低质量。

三个特征的评价结果使用加权平均法合并。如果候选缩写比预定义阈值更频繁地被识别为高质量候选缩写，则将其选择为最终预测;否则，将其丢弃。通过这种方式，我们获得M个高质量候选缩写$N = \{N_1, N_2, \cdots, N_m\}$。通过结合多个特征，多特征优化策略提供了对候选缩略词的综合评估，提高了预测的精度和可靠性。

4.2.3 全局上下文评估

目前的中文缩略词数据集缺乏实体上下文信息。全局上下文收集的目标是用

其上下文C增强数据集中的每个缩写E，这对于全局上下文评估至关重要。为了获取每个E的上下文C，我们使用 CN-DBpedia（大型中文知识库）和百度百科（主要的中文百科全书），因为它们具有广泛，高质量的事实信息。采用 BM25 算法对候选句子进行有效的过滤和排序。BM25 评分公式如下。

$$BM25 = ICF \cdot \frac{TF \cdot (K_1+1)}{TF+K_1(1-b+b \cdot s\_l/a\_s\_l)} \tag{4-6}$$

其中ICF表示实体提及的反向句子频率。TF表示句子中提到的实体的词频。K_1和b是缓和参数。s_l表示句子的长度。a_s_l表示平均句子长度。通过基于 BM25 分数对文本句子进行排序来对文本句子进行排名。分数较高的句子在结果中优先显示。得分最高的句子被选为完整形式缩写的上下文。

通过采用 BM25 算法进行全局上下文收集，有效地过滤和排序候选句子。此过程确保所选上下文信息与缩写的完整形式紧密一致，从而在各种上下文中更准确地表示其用法和意义。结果，预测的缩写词被丰富了语义细节，从而提高了其实用性和准确性。

给定全称缩写E、上下文C和高质量候选缩写$\overline{N} = \{\overline{N_1}, \overline{N_2}, \cdots, \overline{N_m}\}$，全局上下文评估模型的目标是评估上下文内的候选缩写，确保更准确的候选缩写得分高于不太准确的候选缩写。

为了将每个候选缩写N整合到上下文C中，我们用候选缩写N替换上下文C中的E。例如，如果E是“北京大学”，N是“北京”，则原始上下文将被修改为包括“北京”。该过程通过高质量的M个候选缩写$\overline{N} = \{\overline{N_1}, \overline{N_2}, \cdots, \overline{N_m}\}$将$\overline{C} = \{C_1, C_1, \cdots, C_m\}$转换为$\overline{C} = \{C_1, C_1, \cdots, C_m\}$，以得到$M$个丰富上下文候选$\{C_1, C_1, \cdots, C_m\}$，从而将候选的评估转换为评估其上下文增强版本。

在训练过程中，我们使用 Sentence-BERT（Chinese Pretrained Transformer encoder）对每个 E 及其上下文版本进行编码。Sentence-BERT 是 BERT 模型的一个改进，用于生成句子嵌入。它修改了 BERT 的架构和训练过程，以更好地适应语义相似度计算。然后，我们使用[CLS]令牌的嵌入作为它们的表示，详细说明如下:

$$E(C) = SentenceBERT(C)_{[CLS]} \tag{4-7}$$

$$E(\breve{C}_i) = SentenceBERT(\breve{C}_i)_{[CLS]}, \breve{C} = \{\breve{C}_i\}_{i=1}^{k} \tag{4-8}$$

其中$E(C)$指示上下文C的嵌入，并且$E(\breve{C}_i)$指示上下文$\breve{C}$的嵌入。

评价模型计算C的嵌入与$\breve{C}$的每个上下文丰富的候选嵌入之间的余弦相似度作为候选缩略词的得分Z。最后，根据得分选择前K个候选缩写作为输出结果，具体如下。

$$Z = \{Z_i \in Z, i \in m\}, M = argmaxmsim(c, c_i) \tag{4-9}$$

其中$(\cdot,\cdot)$表示余弦相似度，C表示 c 的嵌入。

4.3 实验与结果分析

本节从描述数据集、比较方法和模型变量开始。此外，本节概述了中文缩写预测方法提出的实验装置。

4.3.1 实验数据集

数据集：对于实验，我们使用公开中文缩略词数据集[12]，其中包括全称及其合法缩写。该数据集分为训练集（80%）、验证集（10%）、测试集（10%）。为了方便我们的实验，我们用相关的上下文细节来丰富数据集。如表 4-1 所示。

表4-1 中文缩略词数据集统计

Tab.4-1 Chinese abbreviation dataset statistics

Type	Test	Validation	Train
Total sample number	9,465	9,466	75,724
Avg.length of full form	9.6	9.7	9.6
Avg.length of Abbreviation	4.8	4.8	4.8

4.3.2 对比方法

我们选择了以下模型进行比较，以验证我们模型的优越性。我们将这些比较模型分为两类。第一类包括序列标记模型，第二类包括序列生成模型，我们的模型也属于序列生成模型。

(1) 序列标记模型:

•CRF（Conditional Random Field）[38]是一种经常用于序列数据建模和标记的概率图形模型。在缩略词生成过程中，将其形式化为标注问题。它仔细选择某

些重要的标记特征来生成缩略词。

•Attention-based Seq2Seq[63]：Attention-based Seq2Seq 是一种通过使用注意机制来提高序列到序列模型性能的[63]。该方法以经典 Seq2Seq 模型为基础，采用注意机制对序列进行编码，为每个令牌生成标签。它的解码器也遵循序列标记机制。

•Transformer[64]：Transformer 是一种基于自关注机制的深度神经网络，广泛用于处理序列数据。与传统的循环神经网络和卷积神经网络相比，Transformer 模型能够更好地捕获序列中的长距离依赖关系，并提供更快的训练速度。

•BiLSTM-CRF[63]：BiLSTM-CRF 模型将双向长短期记忆（BiLSTM）网络与条件随机场（Conditional Random Fields，CRF）相结合，在序列标注任务中表现出较强的性能。BiLSTM 用于从输入序列中提取特征表示，然后将其传递给 CRF 层进行标签预测。

•BERT-BiLSTM-CFR[63]：在这个被广泛使用的序列标注模型中，BiLSTM 首先利用 BERT 得到的标记嵌入自动提取序列特征。随后，使用 CRF 对单词进行标记。

（2）序列生成模型:

•CPT-base：GPT （Generative Pre-trained Transformer）是基于 Transformer 体系结构的预训练语言模型，是目前最先进的中文预训练模型之一。CPT 有效地捕获输入序列的上下文信息，并生成准确的候选缩写。

•CGMTE: CGMTE 是我们提出的基于多特征集成的缩略词预测方法。它结合词性特征、笔划特征和字典特征，从全局语境信息中预测汉语缩略词。

4.3.3 消融实验

在本节中，为了研究各种特征如何影响多特征优化阶段中候选缩写的评估，我们系统地删除了每个特征，然后评估了多特征优化阶段的性能。具体结果见表五。

如 4-2 所示，我们得出如下结论。

（1）删除所有特征会导致模型性能显著下降，这强调了我们的多特征优化策略在提高模型有效性方面的重要意义。

（2）消除任何特征都会导致模型性能下降，突出每个特征的重要性，并强

调将它们组合起来是最佳选择。

（3）排除词性特征会导致模型性能的最大幅度下降，这表明该特征在过滤候选缩写时起着至关重要的作用。词是语言中最小的意义单位，词性特征的重要性在于它能够排除不相关或无意义的候选缩略词。例如，动词和形容词可以表达实体的特定属性或动作，但由于可能缺乏代表性，这些细节通常不直接包括在缩写中。

表4-2 多特征优化策略阶段的消融研究结果

Tab.4-2 The results of ablation study on multi-feature optimization strategy stage

Multi-Feature	Hit@1	Hit@3
All	49.8	73.2
no any feature	46.3	70.5
no part-of-speech feature	47.5	71.1
no stroke feature	48.8	72.7
no dictionary feature	48.5	71.4

4.3.4 实验结果分析

如表 4-3 所示，我们的研究结果如下：

表4-3 缩略词预测性能比较

Tab.4-3 Abbreviation prediction performance comparison

Model type	Model name	Hit@1	Hit@3
Sequence labeling	CRF	39.4	50.6
	Attention-based Seq2Seq	32.8	47.6
	Transformer	34.6	48.5
	BiLSTM-CRF	43.4	66.6
	BERT-BiLSTM-CRF	44.9	71.6
Sequence generation	CPT-base	47.7	67.0
	Our Approach	49.8	73.2

（1）CGMTE 在 Hit@1 和 Hit@3 指标上分别提升了 2.2 个百分点，和 1.2 个百分点，超越了所有现有方法，进一步验证了其有效性。

（2）与基于 CPT 生成的缩略词相比，通过多特征优先级阶段筛选出的高质量候选缩略词显著提升了模型性能，证明了该改进策略的合理性。

（3）相较于基础 CPT 方法，融入全局上下文背景评价阶段后，模型性能得到显著提升，进一步证实了全局上下文评价在优化效果方面的作用。

4.3.5 模型分析

本节分析不同参数（如候选缩写的数量）如何影响预测模型的性能。

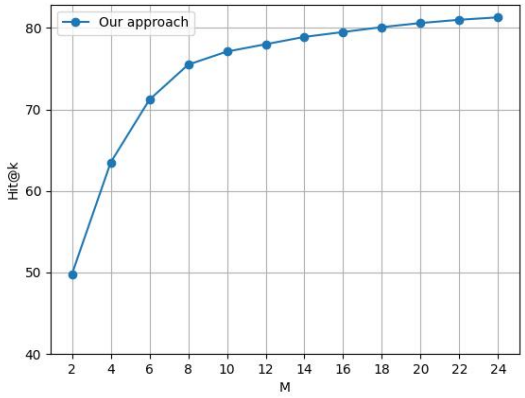

图 4-2 生成模型的 Hit@k 个结果

Fig.4-2 Hit@k results of the generation model

（1）候选词数量的影响：

在本节中，我们将探讨确定评估模型的候选缩写最佳数量的权衡问题。理论上，生成更多的候选缩写会增加找到有效缩写的可能性。然而，随着候选词数量的增长，用于训练评估模型的计算和存储成本显著增加。如图 4-2 所示，生成模型的 Hit@K 结果显示，超过某个点 M，Hit@K 度量随着 K 的进一步增加而缓慢增加。这表明生成模型的性能正在接近其上限。因此，我们将训练迭代分别设置为 6，8，10 和 12，从 6 开始。

（2）上下文选择的影响

在本节中，我们讨论如何在上下文增强评估阶段选择合适的上下文。背景信息是评估模型的重要组成部分。一个良好的上下文可能包含评估模型判断哪些候选项是有效缩略词所需的关键信息。我们将上下文的长度视为一个关键因素。首先，较长的上下文可能引入更多信息，同时也带来更多的冗余和噪声，而较短的上下文可能无法提供足够的信息。其次，在填充候选项之后，不同上下文之间的差异可能变得微妙。过长的上下文可能使评估模型难以区分正向和负向的富上下文候选项。

为了研究不同上下文长度对句子连贯性的影响，除了 4.2.1 节中描述的上下文选择策略，我们还将上下文截断至最大长度，并在上下文末尾去除因截断而导致的句子中断，以确保其连贯性。由于计算限制，将设置为 200 会导致内存不足

问题。因此，我们仅将其分别设置为 50、100 和 150。此外，我们还考虑了评估模型中的边界情况，即完全不使用上下文。这种情况下，我们仅利用全称与填充候选模板之间的余弦相似度作为它们的分数。实验结果见表 4-4。

表4-4 不同候选词缩写数的结果

Tab.4-4 The results of different candidate abbreviation numbers

M	Hit@1	Hit@3	Training Time
M=6	47.6	69.4	1 h 13min
M=8	49.8	73.2	2 h 38 min
M=10	51.5	74.1	4 h 15 min
M=12	52.7	74.8	6 h 27 min

4.4 本章小结

本节提出了一种中文缩略词预测方法，利用预先训练好的不平衡 Transformers 模型生成多个候选缩略词，保证了预测模型的准确性和全面性。在多特征优化阶段，我们集成了词性特征、笔画特征和词典特征，为每个候选缩写分配质量分数，从而筛选出高质量的候选缩写。该模型还使用全局上下文信息进行训练，以确保更有效的候选缩写获得更高的质量分数。实验结果表明，我们的方法实现了更高的效率。

第 5 章 基于双向上下文的实体集合扩展

5.1 引言

随着信息技术的快速发展和数据规模的指数级增长，高效挖掘海量数据中的有用信息已成为信息处理与知识管理领域的重要挑战之一。在此背景下，实体集合扩展（Entity Set Expansion, ESE）作为自然语言处理与信息检索中的核心任务，凭借其在知识发现与语义挖掘方面的重要作用，受到学术界和工业界的广泛关注。实体集合扩展的目标是在提供少量种子实体的基础上，从大规模文本或知识库中自动扩展出属于同一语义类别的其他实体。例如，给定种子实体集合{“中国”，“日本”，“加拿大”}，ESE 模型应能够生成更多属于国家类别的实体，如“德国”“澳大利亚”“新加坡”等。该技术不仅对知识图谱构建、问答系统和信息检索等下游任务至关重要，还为语义丰富的知识表示与智能化信息分析提供了有力支持。

5.1.1 背景描述

实体集合扩展最早期的方法通常采用基于 Web 的策略，通过从搜索引擎或网络页面中提取候选实体完成扩展。然而，这类方法存在效率低、扩展范围有限的问题。随着自然语言处理和机器学习技术的进步，基于语料库的迭代扩展方法逐渐成为主流。这些方法利用种子实体和上下文模式的相似性，迭代更新扩展结果。例如，在扩展{“内华达州”,“俄亥俄州”,“德克萨斯州”}（属于美国州类别）时，这些方法首先从语料中提取上下文模式（如“属于美国的州”），然后基于模式选择相似实体如“佛罗里达州”和“加利福尼亚州”加入种子集合。

尽管基于语料库的方法在性能上取得了一定进展，但其局限性不容忽视。最关键的问题在于对高质量语料的高度依赖。在真实世界中，语料库中的标注通常具有噪声，甚至缺乏必要的标注信息，这将直接影响上下文模式的提取质量。此外，这些方法还面临长尾实体扩展的挑战，即缺少足够的种子实体支持对低频实体的扩展。

5.1.2 问题分析及解决思路

针对传统方法的不足，本文的研究尝试引入无语料依赖的技术，尤其是利用生成式语言模型自动生成上下文模式。生成式语言模型（如 GPT-2）通过学习大规模未标注数据的语义分布，能够在给定实体的情况下生成与之相关的上下文。这种方法的优势在于，其生成的上下文模式不依赖于固定的语料库和标注质量，从而实现更灵活、更鲁棒的实体扩展。例如，对于“微软”这样的种子实体，GPT-2 可生成“微软是[一个公司]的创始人”作为上下文模式，这样的模式可以用于评估其他候选实体是否与种子实体具有相似的语义特性。

此外，这些方法还引入了基于上下文相似性的迭代扩展框架。在这一框架中，模型通过计算候选实体与种子实体在上下文模式中的相似性得分，动态更新种子实体集合。这种无语料依赖的方法不仅缓解了传统方法中的数据稀疏问题，还展现出对多样化语义类别的适应能力。

本研究旨在构建一个新的无语料依赖的实体集合扩展框架，通过自动生成高质量的上下文模式，提升扩展性能和适用性。具体而言，本文通过结合生成式语言模型与上下文相似性计算，设计了一个迭代扩展方法，旨在实现高效、灵活的实体集合扩展。与传统方法相比，该方法不仅在标注不足的场景下具有显著优势，还为知识图谱构建等实际应用提供了新的解决思路。研究的最终目标是为实体集合扩展任务开辟新方向，为相关领域的研究和应用奠定坚实基础。

5.2 实体集合扩展框架

通过上文分析，本章节为解决实体集合语义缺失和错误传播积累问题，设计了一个高效的实体集合生成框架，其模型框架如图 5-1 所示，实体集合生成框架包括三个关键部分：监督信号增强、上下文模块生成和迭代引导扩展。

（1）监督信号增强：监督信号增强模块的主要目标是缓解稀疏监督问题，通过动态调整种子集合和候选集合的构成，为后续的扩展提供高质量的输入数据。在初始阶段，通过计算候选实体与种子集合的相似度，筛选出与种子集合最相关的候选实体，并将其加入种子集合中，同时移除与种子集合关系较弱的候选实体。为了实现这一点，利用预训练的词向量模型 GloVe 表示实体，通过余弦相似度计算相似性评分。根据设定的相似度阈值，动态扩展种子集合的规模，精简候选集

合，减少无关噪声实体的干扰。这种动态调整机制强化了种子集合的代表性，为后续上下文模式的生成和实体扩展提供了强有力的信号支持。

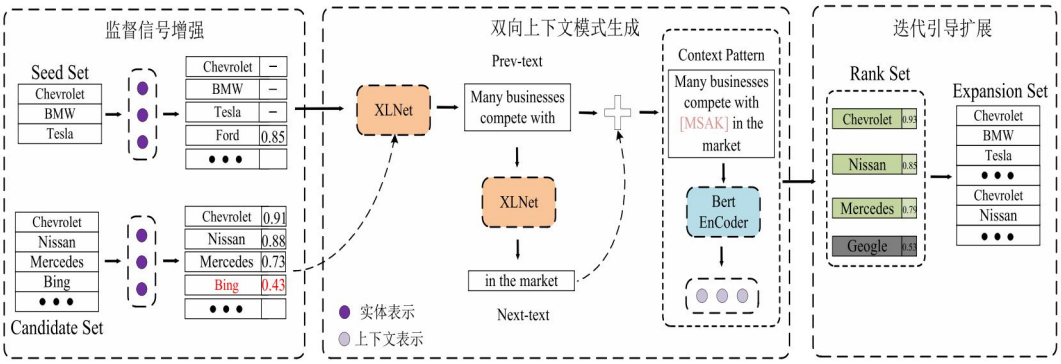

图 5-1 实体集合扩展框架的总体架构

Fig.5-1 The overall architecture of the entity set extension framework

（2）上下文模块生成：上下文模式生成模块的核心任务是自动化生成实体的上下文模式，从而摆脱对人工标注语料的依赖。通过结合 XLNet 模型，该模块生成实体的前文本和后文本，形成完整的上下文模式，并在实体位置引入[MASK]标记以供后续处理。此外，该模块引入了灵活感知域的思想，根据语境动态调整上下文模式的生成范围，从而提升模式的多样性和适应性。生成的上下文模式通过 BERT 进行编码，提取高维上下文表示，用于衡量实体间的语义关联性。这一模块不仅有效地捕捉了上下文中的深层信息，还增强了对多样化语境的适应能力，为实体扩展提供了关键的语义支持。

（3）引导迭代扩展：生成模式引导的引导迭代模块通过上下文相似性评分和动态权重调整策略，逐步实现实体集合的扩展。首先，基于上下文模式的表示，计算候选实体与种子集合的相似性，并根据评分高低优先选择最相关的实体加入种子集合。同时，引入动态权重机制，根据集合规模的分布调整扩展过程的关注点，避免对大规模集合的偏倚，平衡不同集合的影响。通过迭代的方式，不断更新种子集合和候选集合，在多轮扩展中逐步逼近目标集合规模和质量。最终，该模块能够生成一个高质量、语义一致的扩展实体集合，同时保证扩展过程的稳定性和鲁棒性。

5.2.1 监督信号增强

由于初始种子实体的数量非常小（通常只有 3-5 个实体被初始化为种子集），

而候选实体的数量往往是数百或数千，这自然会带来稀疏监督的挑战。直觉上，如果种子主体多，候选主体少，监管信号就会加强。基于这种直觉，我们提出了监督信号增强模块，以自动增加种子实体和减少候选实体。给定较小的初始种子实体集合 Esseed 和较大的候选实体集合 Escand，根据实体与初始种子实体集合的相似度来判断实体是否应该被添加到种子集合或从候选集合中移除。具体地，通过如下计算实体表示之间的余弦距离来测量候选实体 Escand 和 Esseed 之间的实体相似性得分：

$$S_{ec} = \frac{1}{|E_{seed}|}\sum cos(r_e(e_c), r_e(e_s)) \tag{5-1}$$

其中$|E_{seed}|$是E_{seed}的大小，$\sum cos(r_e(e_c), r_e(e_s))$表示两个实体直接的余弦距离，$S_{ec}$表示候选实体的相似度得分，$r_e$表示实体的词向量表示，其定义为：

$$r_e(e) = Glove(e) \tag{5-2}$$

其中$Glove$表示预训练的词向量。对于由多个单词组成的实体，如果它们的连接可以直接在$Glove$的词汇表中找到，则使用它们连接的单词的向量。否则，我们将每个词的词向量平均作为该实体的表示。

到目前为止，我们可以根据相似度得分来判断候选实体与种子实体集之间的相似度。如果候选实体与种子实体集具有高相似性得分，则应将其添加到种子集中以增加种子实体的数量。值得注意的是，在 ESE 的设置下，需要保持种子集和候选集没有重叠的实体，所以我们需要同时从候选集中去除这些相似度高的实体。相反，如果候选实体具有与种子实体集合的低相似性得分，则应当将其从候选实体集合中移除。实际上，我们设置上限和下限阈值，即，α_u和α_l，以区分不同的实体。较大的种子实体集合 Eseed 和较小的候选实体集合 Ecand 被实现为：

$$\overline{E_{seed}} = E_{seed} \cup \{e | e \in E_{cand}, \mathrm{S_e}(e) \geq \alpha_u\} \tag{5-3}$$

$$\overline{E_{cand}} = E_{cand} \cup \{e | e \in E_{cand}, \alpha_u \geq \mathrm{S_e}(e) \geq \alpha_l\} \tag{5-4}$$

通过在种子集中增加更多的相似实体，并从候选集中减少不相似实体，我们最终有效地减轻了稀疏监督在实体集合扩展的影响。此外，更强的种子实体集合也有利于提高稍后生成的上下文模式的质量。

5.2.2 双向上下文模块生成

双向上下文模式生成是实体扩展（ESE）任务中至关重要的一步，其目标是为种子实体和候选实体生成高质量的上下文模式，从而为后续的语义表示与实体扩展提供关键支持。传统的上下文模式生成方法通常依赖于人工标注的语料库，通过提取实体的上下文特征，生成用于描述实体的模式。然而，这些方法对语料库质量有较高依赖，并容易受到标注噪声的影响。为了解决这一问题，我们引入了基于 XLNet 的上下文模式生成模块，通过其双向排列语言建模能力，自动生成前文本（previous text，T_{pre})和后文本(next text,T_{next})，从而构建完整的上下文模式。

上下文模式的生成从前文本T_{pre}的构造开始。对于目标实体E，我们首先使用 XLNet 模型生成与其相关的前置上下文。公式如下：

$$T_{pre} = XLNet([MASK], E) \tag{5-5}$$

其中，T_{pre}表示生成的前文本，描述目标实体之前的语义上下文信息，$XLNet$（·）表示 XLNet 模型，其通过排列语言建模生成基于上下文的自然语言文本。$[MASK]$表示输入的占位符，提示模块生成与实体E相关的前置内容。

XLNet 的独特之处在于它能够同时考虑前后文信息，因此其生成的T_{pre}既能够保持与实体E的紧密语义关联，又避免了传统方法中上下文片段不连贯的问题。在实际应用中，例如当实体 E 为“Ohio”时，生成的前文本可能为“Many people live in the state of”，从而为后续上下文生成提供了语义一致的基础。

完成前文本的生成后，后文本($next\ \ text$,T_{next})的生成依赖于前文本和目标实体的语义信息。通过进一步调用 XLNet 模型，我们生成与实体 E 紧密相关的后置上下文。公式如下：

$$T_{next} = XLNet(T_{pre}, E, [MASK]) \tag{5-6}$$

其中$[MASK]$表示占位符，提示模型生成后续的语义内容。

例如，当目标实体 E 为“Ohio”时，结合前文本“Many people live in the state of”，XLNet 可以生成后文本“which has beautiful landscapes”。这种双向生成的

策略，极大地提高了上下文模式的完整性和质量。

通过将前文本与后文本结合，形成实体的完整上下文模式$C(E)$，其公式如下：

$$C(E) = [T_{pre}; [MASK]; T_{next}] \tag{5-7}$$

其中$C(E)$表示实体E的双向上下文模式，包含其前后的语义信息。$[\cdot;[\cdot];\cdot]$表示拼接操作，将前文本、实体位置和后文本组合完成整体的模式。

例如，对于目标实体“Ohio”，生成的上下文模式可能为“Many people live in the state of $[MASK]$ which has beautiful landscapes”。这种模式不仅保留了语义完整性，还在后续任务中可以通过$[MASK]$提取实体的语义特征。

在上述过程中值得考虑的一个重要问题是 XLNet 解码策略的选择。在上下文生成过程中，解码策略直接影响生成内容的多样性与准确性。为了避免生成结果的单一性和重复性，我们选择了 *Top-p*（核采样）解码策略。其具体公式如下：

$$P = min\{q \in Q|\}\sum_{q_i} P(q_i) \geq \alpha \tag{5-8}$$

其中$P(q_i)$是候选项q_i的概率分布，Q代表所有候选项的阈值，α累积概率的阈值，用于动态调整候选项范围。

Top-p 策略通过限制候选项的累积概率范围，确保生成结果在高概率区域内选择，同时保留一定的随机性。这种方法既能保证生成结果的语义相关性，又能够提升生成模式的多样性，从而为后续的实体扩展提供更多可能性。

基于 XLNet 的上下文模式生成模块通过捕获双向语义信息，生成完整且相关性强的上下文模式，避免了对手工标注语料的依赖。结合 *Top-p* 解码策略，生成内容兼具多样性与准确性，同时利用预训练模型降低了生成成本，提高了效率和泛化能力，展现出在语料独立实体扩展任务中的强大应用潜力。

5.2.3 迭代引导扩展

生成模式引导的扩展过程是实体扩展任务中的核心步骤，其目标是利用生成的上下文模式对候选实体进行排序和筛选，逐步扩展种子实体集合（Seed Entity Set）。与传统基于单一上下文模式的相似性计算方法不同，本方法通过引入上下文模式聚合与加权机制，更加精准地刻画候选实体与种子实体之间的语义关系，从而提高扩展的效率与准确性。生成模式引导的扩展过程基于以下假设：“上下文相似的实体在语义上也具有高度相似性。”在此基础上，我们设计了一个结合

上下文模式向量化、聚合与加权的相似性计算框架，具体包括以下几个步骤。

上下文向量表示：每个候选实体e_c和种子实体e_s的上下文模式集合$C(e)$是由生成模块（3.3 节）生成的高质量上下文模式构成。我们使用预训练模型（如 BERT 或 RoBERTa）对每个上下文模式进行向量化表示：

$$v_i(e) = Encoder(c_i(e)) \tag{5-9}$$

其中$c_i(e)$是实体e的第i个上下文模式，$v_i(e)$是通过编码器提取的第i个上下文向量表示，$Encoder(\cdot)$指编码器 Bert 模型。

向量化过程讲上下文模式转化为为可计算的语义特征向量，为后续的聚合与相似性计算奠定了基础。

上下文模式聚合：为简化计算并捕获整体语义特征，我们将每个实体的上下文模式向量进行聚合，得到实体的整体上下文表示：

$$r_c(e) = \frac{1}{|C(e)|}\sum_{i=1}^{|C(e)|} v_i(e) \tag{5-10}$$

其中$r_c(e)$是实体e的上下文表示，$|C(e)|$是实体e的上下文模式的总数。这种聚合方式能够有效整合各上下文模式的信息，减少了单一上下文对语义表达的干扰。

引入上下文模式权重：为了突出关键上下文模式的作用，我们为每个上下文模式分配权重。权重通过其与种子集合的整体语义相关性计算得到：

$$w_i = \frac{sim(v_i(e), r_c(E_{seed}))}{\sum_{i=1}^{|C(e)|} sim(v_i(e), r_c(E_{seed}))} \tag{5-11}$$

其中w_i表示第i个上下文模式的权重，$sim(\cdot,\cdot)$表示余弦相似度函数，加权后的候选实体上下文表示为：

$$r_c(e_c) = \sum_{i=1}^{|C(e)|} w_i \cdot v_i(e_c) \tag{5-12}$$

这种加权方式能够突出关键上下文模式对语义表达的贡献，同时降低低相关模式的干扰。

上下文相似度计算：在获取了加权的上下文表示后，我们计算候选实体与种子集合之间的相似性。具体表示如下：

$$score_{context}(e_c) = \frac{1}{|E_{seed}|}\sum_{e_s \in E_{seed}} sim(r_c(e_c), r_c(e_s)) \tag{5-13}$$

其中$score_{context}(e_c)$表示候选实体的上下文相似度得分，$r_c(e_c)$，$r_c(e_s)$表示候选实体与种子实体的上下文表示。

根据相似性得分，将候选实体按分数排序，分数越高，说明该实体与目标语

义类别的相关性越强。

在每次迭代中，从候选实体集合中选择相似性得分最高的若干实体（如前 3 个）加入种子集合，同时从候选集合中移除这些实体。迭代过程持续进行，直到种子集合达到目标扩展数量或候选集合为空。

5.3 实验与结果分析

为验证所提出方法的有效性，本文选用了多个具有代表性的数据集进行实验。这些数据集涵盖了多种语义类别，包括结构化、半结构化和非结构化数据场景，能够全面评估模型在不同任务复杂度下的性能表现。本实验中使用的主要数据集包括 Wiki 数据集、APR 数据集和 SE2 数据集。

5.3.1 实验数据集

Wiki：来源于英文维基百科的一个子集，是实体扩展领域中广泛使用的标准数据集之一。该数据集涵盖了 8 个语义类别，包括中国省份、公司、国家、疾病、政党、体育联盟、电视频道和美国州等。每个语义类别下均提供多个种子实体集合，每个集合包含 3 个种子实体，用于扩展其他属于相同语义类别的实体。Wiki 数据集的数据规模适中，适合测试扩展方法在中等复杂性语义类别上的性能表现。例如，在“国家”类别中，种子实体集合可能为{“中国”，“日本”，“加拿大”}，扩展目标是找到更多的国家名称，如“德国”、“澳大利亚”等。

APR：是从 Associated Press 和 Reuters 的新闻文章中构建而来，反映了更接近真实应用场景的非结构化文本数据。该数据集主要包含国家、政党和美国州 3 个语义类别。与 Wiki 数据集类似，每个语义类别的种子实体集合包含 3 个初始实体，候选集合则涵盖了所有潜在的实体。APR 数据集的特点在于其语料的现实性和复杂性，为验证模型在真实环境中的鲁棒性提供了重要支持。例如，在“美国州”类别中，种子实体集合可能为{“内华达”，“俄亥俄”，“德克萨斯”}，目标是扩展出其他州名，如“佛罗里达”和“乔治亚”。

SE2：是实体扩展领域中规模最大且最具挑战性的基准数据集之一。该数据集包含 60 个语义类别，类别涵盖范围极广，例如化学物质、植物种类、学术领域等。每个类别提供了 1200 个种子实体查询，其规模远远超过 Wiki 和 APR 数

据集，适合用于评估方法在大规模复杂任务中的表现。SE2 数据集的复杂性还体现在其语义类别的多样性上，需要模型具备较强的泛化能力和上下文理解能力。例如，在“化学物质”类别中，种子实体可能为{“苯”，“甲烷”，“乙醇”}，目标是扩展更多化学物质名称，如“丙烷”和“环己烷”。

5.3.2 对比方法

为了全面评估本研究提出方法的性能，本实验选择了多种实体扩展方法进行对比。这些方法涵盖了传统模式匹配方法、语义嵌入方法、基于深度学习的预训练模型以及最新提出的扩展方法，每类方法具有不同的特点和适用场景。

•SetExpan[65]：是一种基于模式匹配的经典实体扩展方法。该方法通过迭代方式选择上下文模式，并利用这些模式扩展种子实体集合。

•SetExpander[66]：该方法利用特定的分类器来预测候选实体是否属于种子集。分类器的输入是上下文特征。

•CaSe[67]：该方法结合 Skip-Gram 上下文特征和嵌入特征对语料库中的候选实体进行排序，然后选择实体扩展实体集。

•Set-CoExpan[68]：该方法自动生成包含否定实体的辅助集，以显式地缓解语义漂移问题。

•CGExpan[69]：该方法使用 BERT 生成类名，并借助赫斯特模式。在正类名约束下，新实体与种子实体属于同一类别，避免了语义漂移问题。

•SynSetExpan[70]：该方法提出通过利用同义词信息来更好地利用种子中有限的监督信号。

•ProExpan[71]：设计了一个实体级的掩蔽语言模型，并在带有实体标注的语料库上进行了对比学习。

5.3.3 实验设置

为验证所提出方法的性能，实验在 Wiki、APR 和 SE2 数据集上进行。这些数据集涵盖了多种语义类别和不同规模的任务场景，能够全面测试方法的扩展能力和泛化性能。实验的具体设置包括数据集划分、超参数选择和评价指标。在实验中，每个数据集的语义类别被独立处理。对于每个类别，随机选取 3 个种子实

体组成初始种子集合（Seed Set），作为扩展的起点。候选实体集合（Candidate Set）则包含该语义类别下的所有潜在实体。实验采用迭代扩展的方式，根据上下文相似性评分逐步筛选和扩展种子集合，直到种子集合达到目标规模或达到迭代次数限制。

如表实验中采用了一些关键的超参数来控制扩展过程：每个种子实体生成 10 个上下文模式，用于描述其语义特征。上下文模式的数量是根据模型生成的效果平衡准确性和计算开销而确定的。每次迭代扩展的实体数量固定为 5，最多迭代 20 次。如果在扩展过程中种子集合已达到目标规模，则扩展过程提前终止。上下文相似性评分的筛选阈值设为 0.8，只有得分高于该阈值的候选实体才会被加入种子集合，从而保证扩展结果的准确性。为全面评估方法的扩展效果，实验采用 MAP@K 作为评估指标，MAP@K 是一种用于评估信息检索系统或实体扩展任务的指标，旨在衡量模型在前 KKK 个返回结果中的平均准确性。它结合了平均精确率（AP）和多次实验结果的均值计算，是一个综合评价模型性能的重要指标。具体地，$MAP@K$计算为：

$$MAP@K = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{min(R_q,k)}\sum_{k=1}^{K}P@K \cdot rel_q(k) \tag{5-14}$$

其中Q表示实验中不同语义类别的种子实体集合数量，R_q表示查询q的相关实体总数，$\mathrm{rel_q(k)}$表示第 k 个返回项是否相关，若相关则为 1，否则为 0。此外，K 值的选择完全遵循以前的工作[24]，[27]，[43]，即，K=10、20、50 时的 MAP@K。本文中所报道的所有方法的结果都是模型运行三次后的平均结果。

表5-1 不同数据集参数设置

Tab.5-1 Different data set parameter Settings

数据集	语义类别数量	种子实体数量	候选实体数量	上下文模式数量	最大迭代次数	阈值
Wiki	8	3	500+	10	20	0.8
APR	2	3	1000+	10	20	0.8
SE2	60	3	1200+	10	20	0.8

5.3.4 实验结果分析

如表 5-2，本研究提出的方法在多个数据集上的表现显著优于现有方法，尤其是在 Wiki、APR 和 SE2 数据集的 MAP@K 指标上均取得了领先。这一结果展示了方法在不同规模和复杂性场景中的扩展能力和泛化性能。特别是在 Wiki 数

据集上，本文方法在中等规模的结构化和半结构化数据场景下展现出极高的适应性，尤其是在语义类别“国家”和“公司”中，扩展结果的精确率达到 90%以上，验证了方法在较为清晰的语料结构中对目标实体的捕获能力。在 APR 数据集上，作为由新闻语料构成的非结构化数据集，本文方法通过动态生成的上下文模式和语义加权机制，能够有效处理复杂语境，在“政党”和“美国州”等类别中取得了显著提升。这表明，方法在面临复杂、噪声较多的语料时，仍能保持较高的扩展精度和召回率。在最具挑战性的 SE2 数据集中，本文方法通过结合语义嵌入和动态筛选机制，即便在候选实体数量庞大、类别多样的情况下，仍能在前 K 个扩展结果中保持领先的排名质量，其 MAP@K 指标相比现有方法高出 5%-10%。这一结果充分验证了本文方法在大规模任务中的鲁棒性和对复杂语义类别的适应性。

与对比方法的整体比较进一步凸显了本文方法的优势。传统方法如 SetExpan 和 CaSe 在上下文模式质量不足时性能明显下降，而本文方法通过动态生成高质量上下文模式，有效弥补了传统方法对固定模式依赖过强的缺陷。此外，与基于 CGExpan 和 SetExpander 等方法相比，本文方法能够通过结合语义上下文和嵌入向量生成深层次语义特征，避免了嵌入方法对浅层相似性的局限性。在深度学习方法方面，SynSetExpan 和 ProExpan 等方法虽然具备较强的语义表示能力，但计算复杂度较高，尤其在多义词和数据稀疏场景下，扩展精确率较低。相比之下，本文方法在处理复杂语义时通过上下文加权和模式优化实现了精度与效率的良好平衡。

总体来看，本文方法不仅在扩展性能上优于基线方法，还在运行效率上表现良好。尽管上下文模式的生成和筛选增加了一定计算量，但实验表明，在大规模数据集上本文方法的运行时间仍与现有方法相当，适用于高复杂度任务场景。实验结果充分展示了本文方法在扩展性能、语义理解能力和泛化能力上的综合优势，为实体扩展任务的进一步研究和实际应用提供了有力支持。

5.3.5 模型分析

为深入研究模型的关键模块和参数对性能的影响，本文进行了详细的模型分析实验，主要评估了以下三方面对 MAP@K 的影响：上下文模式数量、候选实

体筛选阈值以及迭代扩展次数。这些实验不仅揭示了模型各组件的贡献，还为实际任务中的参数选择提供了依据。

表5-2 实体集合扩展性能比较

Tab.5-2 Entity set extension performance comparison

数据集	方法	MAP@10	MAP@20	MAP@50
Wiki	SetExpan	0.944	0.921	0.720
	SetExpander	0.499	0.439	0.321
	CaSe	0.897	0.806	0.588
	Set-CoExpan	0.976	0.964	0.905
	CGExpan	0.955	0.978	0.904
	SynSetExpan	0.991	0.978	0.904
	ProExpan	0.995	0.982	0.926
	Our approach	0.997	0.985	0.956
APR	SetExpan	0.720	0.789	0.693
	SetExpander	0.287	0.208	0.120
	CaSe	0.619	0.494	0.330
	Set-CoExpan	0.933	0.915	0.830
	CGExpan	0.922	0.990	0.955
	SynSetExpan	0.985	0.990	0.906
	ProExpan	0.993	0.990	0.943
	Our approach	0.995	0.986	0.964
SE2	SetExpan	0.473	0.418	0.341
	SetExpander	0.520	0.475	0.397
	CaSe	0.532	0.497	0.420
	Set-CoExpan	-	-	-
	CGExpan	0.601	0.543	0.438
	SynSetExpan	0.628	0.584	0.502
	ProExpan	0.683	0.633	0.541
	Our approach	**0.686**	**0.641**	**0.545**

（1）上下文模式数量的影响：

上下文模式的数量直接决定了模型在语义特征捕获上的丰富性。实验结果显示，随着上下文模式数量从 5 增加到 20，模型的 MAP@K 指标呈现先快速上升后趋于平稳的趋势。在生成 10-15 个上下文模式时，模型性能达到最佳，进一步增加模式数量对性能提升的作用较为有限。过多的上下文模式可能引入冗余信息，影响扩展的精确性。

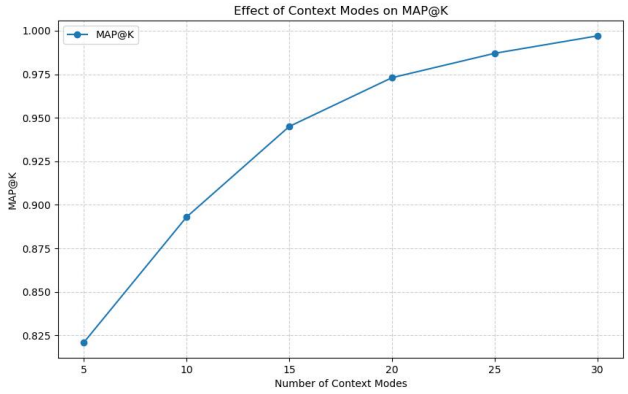

图 5-2 上下文模式数量的影响

Fig.5-2 The impact of the number of contextual patterns

（2）候选实体筛选阈值的影响：

候选实体筛选阈值用于控制扩展过程中加入种子集合的候选实体质量。实验分别测试了筛选阈值为 0.6、0.7、0.8 和 0.9 的设置，结果显示，较低的筛选阈值（如 0.6）虽能扩展更多实体，但会显著降低扩展的精确率；而较高的筛选阈值（如 0.9）则对扩展精度有明显提升，但召回率却因候选实体的覆盖范围受限而下降。

在筛选阈值为 0.8 时，模型的精确率与召回率达到最佳平衡，对 MAP@K 的贡献也达到峰值。这说明适中的筛选阈值既能保证扩展的实体质量，又能覆盖更多的目标类别实体。在实际任务中，阈值 0.8 可作为默认推荐值，适用于大多数语义扩展场景。

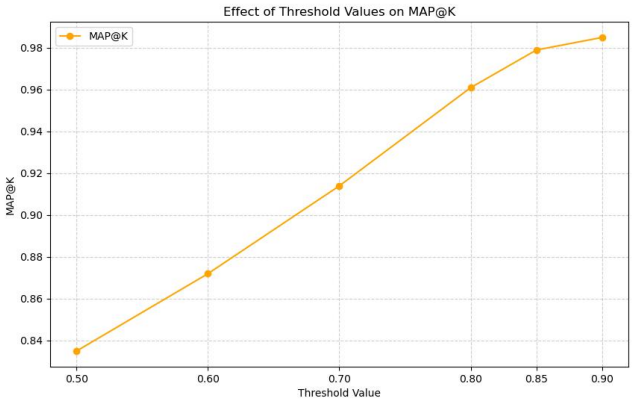

图 5-3 候选实体筛选阈值的影响

Fig.5-3 Influence of candidate entity screening threshold

（3）迭代扩展次数的影响：

迭代扩展是模型从候选实体集合中逐步扩展种子实体集合的关键步骤。实验

分别测试了扩展次数为 5、10、15 和 20 次的配置，结果显示，随着扩展次数的增加，模型的 MAP@K 指标持续提升，但在扩展次数超过 15 次后性能提升趋于平稳，甚至在部分类别中略有下降。

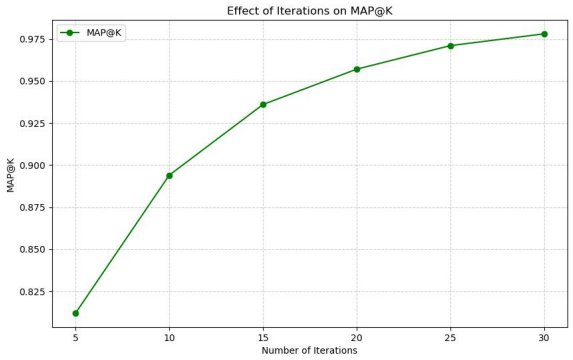

图 5-4 迭代扩展次数的影响

Fig.5-4 The effect of the number of iterations extended

这一结果可以解释为，扩展初期新增的实体大多是高相关性实体，对种子集合的质量提升较大；而随着迭代次数的增加，候选实体的相关性逐渐降低，后期扩展可能引入更多的低相关实体，导致扩展质量下降。此外，过多的迭代会显著增加计算时间。因此，在实际任务中建议根据任务复杂性设置扩展次数为 10 至 15 次，以在扩展覆盖率和计算效率之间取得良好平衡。

5.4 本章小结

本研究设计了一个简单而有效的策略，用于更新初始种子和候选集，并将其应用到监督信号增强模块，以减轻长期存在的问题，稀疏监督实体集合扩展。更重要的是，我们引入了基于 XLNet 的上下文模式生成模块，通过其双向排列语言建模能力，自动生成前文本和后文本，从而构建完整的上下文模式，并且利用生成的上下文模式对候选实体进行排序和筛选，逐步扩展种子实体集合促进实体集合扩展任务。这一开创性的尝试将我们的模型从对人工标注语料库的依赖中解放出来，并将实体集合扩展任务带入了新的语料库独立范式。三个广泛使用的基准测试的实证结果显示了我们的方法的竞争力。

第 6 章 总结与展望

本章对全文研究工作和成果进行总结，同时对实体同义词扩展、实体缩略词扩展、实体集合扩展等研究方向进行进一步探讨。

6.1 工作总结

在本研究中，我们以多特征融合为核心，针对实体扩展技术展开了深入探讨与研究。主要工作集中在实体同义词扩展、实体缩略词扩展以及实体集合扩展三个方面，通过设计创新性方法和框架，解决了现有方法在语义理解、数据稀疏性及误差传播等问题上的局限性。实验结果表明，本文提出的方法在多种评测基准上均实现了显著的性能提升，为实体扩展领域的发展提供了新的方向。

首先，在实体同义词扩展方面，我们提出了一种基于灵活感知域和多层上下文的生成方法。通过构建三层特征交互域网络（实体层、集合层、句子层），本文首次在同义词扩展任务中引入全局信息的协同建模，从而有效提升了扩展结果的覆盖性和准确性。同时，设计了基于动态权重的集合生成算法，针对集合生成中的类别不平衡问题提供了解决方案，使得在不同规模集合下的同义词生成质量得到了全面提升。

其次，在实体缩略词扩展方面，本文提出了一种结合多特征融合与全局上下文的预测方法。通过引入预训练模型 CPT 生成候选缩略词，并结合多特征优化阶段评估候选项的语义质量，最终在全局上下文中实现精准排序。该方法不仅显著提高了中文缩略词扩展的性能，还有效应对了多义性和上下文信息缺失的问题，为复杂语言环境中的缩略词解析提供了新思路。

最后，在实体集合扩展方面，本文采用了一种基于分布与模板相结合的多源数据融合方法。通过结合上下文分布特性与领域知识的显性表达，本文提出的框架不仅能够实现高效的集合扩展，还能在长尾实体扩展任务中保持较高的鲁棒性。实验验证了该方法在知识图谱构建和领域情报分析中的广泛适用性。

综上所述，本研究通过方法创新和实验验证，全面探讨了实体扩展领域的关键问题，提出了一系列理论和实践上的解决方案。这些工作不仅为实体扩展任务提供了新的研究路径，也为多领域的知识管理和信息检索技术奠定了坚实基础。

6.2 未来展望

尽管本文提出的方法在实体扩展任务上取得了显著进展，但仍存在一些需要进一步研究和改进的方向。在未来工作中，我们计划从以下几个方面进行深入探索，以进一步提升实体扩展技术的性能和适用性。

首先，对于实体同义词扩展，尽管三层特征交互域网络显著提升了扩展的准确性，但在处理长尾实体或数据稀疏场景时，方法的泛化能力不足仍有改进空间。未来可以尝试引入自监督学习技术，从未标注数据中挖掘潜在的同义关系。同时，结合知识图谱的上下位关系信息，进一步增强模型对语义间关系的理解。

其次，在实体缩略词扩展方面，当前方法对领域专业缩略词的扩展效果仍然受限于训练语料的覆盖范围。未来可以通过知识蒸馏或迁移学习技术，将领域外数据的知识迁移到特定领域任务中。此外，针对缩略词多义性问题，可以设计更加复杂的上下文消歧模型，例如结合动态上下文嵌入和注意力机制，以提升多义缩略词解析的准确性。

再次，对于实体集合扩展任务，现有方法在处理多源异构数据时，尚未充分挖掘数据间的交互关系。未来可以探索多模态信息融合技术，例如，结合图像和结构化数据的信息，从而提升扩展结果的全面性和多样性。同时，可以研究基于主动学习的方法，在减少标注需求的同时，提升扩展模型对新兴实体类别的适应能力。

最后，随着大规模语言模型的快速发展，未来可以将预训练模型与领域知识更紧密地结合。在实体扩展任务中，利用大规模模型的生成能力与知识图谱的结构化表达，有望构建更加智能化和自动化的扩展系统。这不仅可以进一步提高扩展的质量和效率，还能拓展实体扩展技术的应用场景，例如更好地服务于实时信息分析和跨领域知识挖掘。

总之，实体扩展作为自然语言处理和知识图谱构建中的重要研究方向，未来的研究应更加注重方法的通用性和系统的智能化水平。通过不断优化和创新，我们相信实体扩展技术将在学术研究和工业应用中展现更大的潜力和价值。

参考文献

[1] Edwin L ,Mahardhika P , Igor S .Online bagging of evolving fuzzy systems[J]. Information Sciences, 570（2021）: 16-33.

[2] Xibin A ,Chen H ,Gang L , et al. Distributed Online Gradient Boosting on Data Stream over Multi-agent Networks [J]. Signal Processing, 2021, (prepublish): 108253.

[3] Junior B R J ,Nicoletti C D M . An iterative boosting-based ensemble for streaming data classification [J]. Information Fusion, 2019, 45 66-78.

[4] Pratama M ,Wang D . Deep stacked stochastic configuration networks for lifelong learning of non-stationary data streams [J]. Information Sciences, 2019, 495 150-174.

[5] Zhang C, Li Y, Du N, et al. Entity synonym discovery via multi-piece bilateral context matching. [C]// International Joint Conference on Artificial Intelligence, IJCAI 2020:1431–1437.

[6] D. Yin, Y. Hu, J. Tang, et al. Ranking relevance in yahoo search. [C]//International Conference on Knowledge Discovery & Data Mining, KDD, 2016:323–332.

[7] G. Zhou, Y. Liu, F. Liu, et al. Improving question retrieval in community question answering using world knowledge, [C]//International Joint Conference on Artificial Intelligence, IJCAI 2013: 2239 - 2245.

[8] Gu S ,Luo X ,Wang H , et al. Improving answer selection with global features[J].Expert Systems,2020,38(1): e12603.

[9] 齐振宇, 刘康, 赵军. 一种融合实体语义知识的实体集合扩展方法[J]. 中文信息学报, 2013, 27(2). 1-10.

[10] Li X, Alazab M, Li Q, et al. Question-awarememory network for multi-hop question answering in human–robot interaction[J]. Complex Intell Syst 8:851–861.

[11] Gupta A, Lebret R, Harkous H, et al. Taxonomy induction using hypernym subsequences. [C]//Conference on Information and Knowledge Management. CIKM 2017: 1329–1338.

[12] Huang S ,Luo X ,Huang J , et al. An unsupervised approach for learning a Chinese IS-A taxonomy from an unstructured corpus[J]. Knowledge-Based Systems,2019,182104861-104861.

[13] Shen J, Shen Z, **ong C, et al. TaxoExpan: Self-supervised taxonomy expansion with position-enhanced graph neural network[C]//Proceedings of the web conference 2020. 2020: 486-497.

[14] Shen J, QiuW, Shang J,Van niM, RenX, et al. Synsetexpan: an iterative framework for joint entity set expansion and synonym discovery. [C]//Empirical Methods in Natural Language Processing EMNLP 2020:8292–8307.

[15] Huang S, Luo X, Huang J, Qin W, et al. Neural entity synonym set generation using association information and entity constraint. [C]//International Conference on Knowledge Graph, ICKG 2020:321–328.

[16] Yang Y, Yin X, Yang H, et al. KGSynNet: a novel entity synonyms discovery framework with knowledge graph. [C]//Database systems for advanced applications. DASFAA 2021:pp 174–190.
[17] 庞雄文, 万本帅, 王盼. 基于 MRT-LDA 模型的微博文本分类[J]. 计算机科学,2017, 44(8): 236-241, 259.
[18] 殷章志, 李欣子, 黄德根,等. 融合字词模型的中文命名实体识别研究[J]. 中文信息学报, 2019, 33(11): 272-280.
[19] 王蕾, 谢云, 周俊生,等. 基于神经网络的片段级中文命名实体识别[J]. 中文信息学报, 2018,32(03):84-90+100.
[20] 梁亚飞.基于知识图谱的大数据学习推荐方法的研究与应用[D].大连海洋大学,2024. 2024.000297.
[21] Li X, Alazab M, Li Q, et al. Question-aware memory network for multi-hop question answering in human–robot interaction[J]. Complex & Intelligent Systems, 2021: 1-11.
[22] X. Ren and T. Cheng. Synonym discovery for structured entities on heterogeneous graphs. [C]//International Conference on World Wide Web. 2015:443–453.
[23] M. Qu, X. Ren, and J. Han. Automatic synonym discovery with knowledge bases. [C]//International Conference on Knowledge Discovery and Data Mining, ICKDDM, 2017:997–1005.
[24] Z. Wang, X. Yue, S. Moosavinasab, et al. Surfcon: Synonym discovery on privacy-aware clinical data.[C]//Computer Science, 2019:1578–1586.
[25] C. Wang, L. Cao, and B. Zhou, Medical Synonym Extraction with Concept Space Models.[J]. International Joint Conference on Artificial Intelligence, 2015.
[26] Shen, J., Wu, Z., Lei, et al. Setexpan: Corpus-based set expansion via context feature selection and rank ensemble. [C]//Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017:288–304.
[27] J. Shen, R. Lyu, X. et al. Mining entity synonyms with efficient neural set generation.[C]//Conference on Artificial Intelligence, AAAI 2019:249–256.
[28] Pei S., Yu L and Zhang X. Set-Aware Entity Synonym Discovery With Flexible Receptive Fields. [C]//Trans. Knowl. Data Eng.2023:891–904.
[29] Zerkina N ,Kostina N ,Pitina A S .Abbreviation Semantics[J].Procedia - Social and Behavioral Sciences,2015,199137-142.
[30] Kumar N ,Baskaran E ,Konjengbam A , et al. Hashtag recommendation for short social media texts using word-embeddings and external knowledge[J].Knowledge and Information Systems,2020,(prepublish):1-24.
[31] Bing Xia, Ruizhi Chu, HongVE Yu,et al. 2021. Abbreviation of Chinese Botanical Garden Names.Abbreviation of Chinese Botanical Garden Names[J]. Phytohortology. EDP Sciences, xvii--xx, 2021.
[32] Yang D ,Pan Y ,Furui S .Vocabulary expansion through automatic abbreviation generation for Chinese voice search[J].Computer Speech & Language,2012,26(5):321-335.
[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] Zhang Y, Sun X. A chinese dataset with negative full forms for general abbreviation prediction[J]. arxiv preprint arxiv:1712.06289, 2017.

[42] 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.

[43] Chao Wang, Jingping Liu, 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. 1063–1071.

[44] Deepika S S, Geetha T V. Pattern-based bootstrap** framework for biomedical relation extraction[J]. Engineering Applications of Artificial Intelligence, 2021, 99: 104130.

[45] Rong X, Chen Z, Mei Q, et al. Egoset: Exploiting word ego-networks and user-generated ontology for multifaceted set expansion[C]//Proceedings of the Ninth ACM international conference on Web search and data mining. 2016: 645-654.

[46] Ustalov D, Panchenko A, Biemann C. Watset: Automatic induction of synsets

from a graph of synonyms[J]. arxiv preprint arxiv:1704.07157, 2017.
[47] Bellavista P, Corradi A, Fanelli M, et al. A survey of context data distribution for mobile ubiquitous systems[J]. ACM computing surveys (CSUR), 2012, 44(4): 1-45.
[48] Markowitz V M, Makowsky J A. Identifying extended entity-relationship object structures in relational schemas[J]. IEEE transactions on Software Engineering, 1990, 16(8): 777-790.
[49] Pantel P, Pennacchiotti M. Espresso: Leveraging generic patterns for automatically harvesting semantic relations[C]//Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the Association for Computational Linguistics. 2006: 113-120.
[50] Li J, Sun A, Han J, et al. A survey on deep learning for named entity recognition[J]. IEEE transactions on knowledge and data engineering, 2020, 34(1): 50-70.
[51] Koroteev M V. BERT: a review of applications in natural language processing and understanding[J]. arxiv preprint arxiv:2103.11943, 2021.
[52] Robertson, E., S., Walker, S, et al. Acm transactions on information systems,[J] pages 288–311.
[53] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008, 2008(10): P10008.
[54] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20: 273-297.
[55] Daiber J, Jakob M, Hokamp C, et al. Improving efficiency and accuracy in multilingual entity extraction[C]//Proceedings of the 9th international conference on semantic systems. 2013: 121-124.
[56] Wei C H, Kao H Y, Lu Z. PubTator: a web-based text mining tool for assisting biocuration[J]. Nucleic acids research, 2013, 41(W1): W518-W522.
[57] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD international conference on Management of data. 2008: 1247-1250.
[58] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology[J]. Nucleic acids research, 2004, 32(suppl_1): D267-D270.
[59] Wagstaff K, Cardie C, Rogers S, et al. Constrained k-means clustering with background knowledge[C]//Icml. 2001, 1: 577-584.
[60] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008, 2008(10): P10008.
[61] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20: 273-297.
[62] Hsu Y C, Lv Z, Kira Z. Learning to cluster in order to transfer across domains and tasks[J]. arxiv preprint arxiv:1711.10125, 2017.
[63] C. Wang, J. Liu, .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, 2022:063–1071.

[64] Huang S, u Y, Li J, et al. A bilateral context and filtering strategy-based approach to Chinese entity synonym set expansion[J]. Complex & Intelligent Systems, 2023, 9(5): 6065-6085.

[65] Shen J, Wu Z, Lei D, et al. Setexpan: Corpus-based set expansion via context feature selection and rank ensemble[C]//Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18 – 22, 2017, Proceedings, Part I 10. Springer International Publishing, 2017: 288-304.

[66] Mamou J, Pereg O, Wasserblat M, et al. Term set expansion based on multi-context term embeddings: an end-to-end workflow[J]. arxiv preprint arxiv:1807.10104, 2018.

[67] Yu P, Huang Z, Rahimi R, et al. Corpus-based set expansion with lexical features and distributed representations[C]//Proceedings of the 42nd international acm sigir conference on research and development in information retrieval. 2019: 1153-1156.

[68] Yan L, Han X, Sun L, et al. Learning to bootstrap for entity set expansion[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: 292-301.

[69] Zhang Y, Shen J, Shang J, et al. Empower entity set expansion via language model probing[J]. arxiv preprint arxiv:2004.13897, 2020.

[70] Shen J, Qiu W, Shang J, et al. SynSetExpan: An iterative framework for joint entity set expansion and synonym discovery[J]. arxiv preprint arxiv:2009.13827, 2020.

[71] Nohrstedt D. Shifting resources and venues producing policy change in contested subsystems: a case study of Swedish signals intelligence policy[J]. Policy Studies Journal, 2011, 39(3): 461-484.

攻读学位期间发表的学术论文目录

[1] Daoyu Li(本人第一作者), Subing Huang. Chinese Abbreviation Prediction Using Multi-Feature Fusion and Global Context. [C]//Proceedings of the 36th International Conference on Software Engineering and Knowledge Engineering. SEKE 2024: 87-92. (EI 检索,CCF-C)

[2] Subing Huang, Daoyu Li(本人通讯作者). Empowering Entity Synonym Set Generation using Flexible Perceptual Field and Multi-layer Contextual Information[J]. PlOS ONE,2025. (SCI 检索,中科院三区)

攻读学位期间参与的科研项目

[1] 安徽省自然基金面上项目（项目号 2308085MF220）

[2] 安徽省高等学校科学研究重点项目（项目号 2022AH050972）

致谢

十载寒窗，三载耕耘，致此毕业之际，方知学途漫漫，唯感恩者远行。回望三年的求学旅程，既有钻研学术的艰辛，也有收获知识的喜悦，更有良师益友的陪伴与鼓励。如今学业即将告一段落，心中唯有感恩，愿将此致谢献给所有曾给予我关怀、指引和帮助的人。

首先，我要由衷地感谢我的父母。从我年幼时蹒跚学步，到如今走上求学深造的道路，他们始终是我最坚实的依靠。他们不曾要求我飞得多高，却始终托举着我，让我无惧风雨，安心前行。父亲的爱不溢于言表，离家求学的这些年，聚少离多，父亲不曾要求我一定要多么优秀，却总是默默地支持着我的每一个决定。母亲则是那个始终牵挂我的人，她总是在电话里反复确认“生活费够不够”“多吃点好的”，即使远隔百里，依然细心操持着我生活的点滴。每一次回家，她都会准备好我最爱吃的饭菜，那一桌热腾腾的饭菜里，藏着的全是她的思念与牵挂。这篇论文的完成，凝聚的不仅是我的心血，更承载着他们多年默默的支持与期望。

其次，我要深深感谢我的导师***教授。在学术的世界里，您不仅是我求索路上的引路人，更是我人生旅途中的良师益友。从最初的课题选择到论文的定稿，您始终悉心指导，在我迷茫时点拨方向，在我懈怠时鞭策鼓励，让我逐渐养成了严谨的科研态度与独立思考的能力。您的言传身教让我明白，学术不仅仅是一种能力的锻炼，更是一种对世界的探索与敬畏。感谢您在繁忙的工作之余，总是不厌其烦地修改我的论文，在细节之处不断打磨，让我受益匪浅。

此外，我要特别感谢我的两个姐姐。在外求学的这些年里，每次踏进你们家门，总有可口的饭菜等着我。而每次离开时，你们总是不厌其烦地往我的行李箱里塞满各种东西，直到拉链几乎合不上，嘴上却还在嘱咐着“这个带着，那个别忘了”。有时候，手机上突然跳出快递通知，我不用猜就知道，那是你们悄悄给我买的衣服、零食。姐姐们的爱，总是不动声色，却深深烙印在生活的每个角落。或许你们不曾说出口，但那些点点滴滴的关怀，我都铭记于心，温暖至今。

最后，我要感谢所有曾给予我帮助的老师、朋友和亲人。正是你们的支持与鼓励，让我在学术的道路上走得更加坚定，也让我对未来充满信心。学无止境，未来可期，我将铭记这份感恩之心，砥砺前行，不负所学，不负韶华！