分　类　号：TP391　　单位代码：10183
研究生学号：2019561006　　密　　级：公 开

吉 林 大 学

博士学位论文

知识驱动的古文字字形识别与语义跟踪研究

Research on Knowledge Driven Ancient Chinese Glyph Identification and Semantic Tracking

作者姓名：迟杨

专　　业：计算机科学与技术

研究方向：自然语言处理，知识图谱

指导教师：福斯托·准奇利亚 教授

培养单位：人工智能学院

2025 年 3 月

知识驱动的古文字字形识别与语义跟踪研究

Research on Knowledge Driven Ancient Chinese Glyph Identification and Semantic Tracking

作者姓名：迟杨

专业名称：计算机科学与技术

指导教师：福斯托·准奇利亚 教授

学位类别：工学博士

论文答辩日期：2025 年 2 月 26 日

授予学位日期：　　年　　月　　日

答辩委员会组成：

	姓名	职称	工作单位
主席	付治国	教授	东北师范大学
委员	宫洵	研究员	吉林大学
	陈贺昌	研究员	吉林大学
	王鑫	研究员	吉林大学
	逄淑超	教授	南京理工大学

摘 要

知识驱动的古文字字形识别与语义跟踪研究

甲骨文、金文等中国古文字来自遥远的先秦时代，与现代汉字存在明显差异，书写也尚未统一，通常具有多种异体字形。至今仍有大量的古文字尚未被破译，在古文字考释的过程中，专家需要对历史上所有相近的字形进行比较研究，归纳文字演变的规律，从而对未知字形对应的汉字进行推理。但是历史上存在过数以万计的字形，它们的字形、字音、字义知识由非结构化的图像、文本数据所描述，这些数据分散在不同的历史典籍和平台中，难以被关联和检索，古文字研究高度依赖于专家的经验、记忆与推理能力。针对以上问题，本文的目标是利用人工智能（Artificial Intelligence， AI）方法，构建链接汉字的知识图谱，为古文字专家提供知识关联与检索服务；基于知识图谱，量化古文字字形相似度，为专家推荐字形相似的古文字；更进一步，预测未知字形所属的文字标签，启发专家进行考释研究；最后生成汉字字义的历时变化可视化分析。本研究对于古文字这一“冷门绝学”的发展与传承具有重要意义，研究内容与贡献具体包括：

第一，本文针对古文字字形、字义与字音领域知识的整合与关联问题，研究历时汉字知识图谱（ZiNet）的本体模式及构建方法。现有工作对古文字领域有形的实体进行链接，例如文物、拓本图像等，但是抽象层面的字形、字音与字义知识如何表示和互联的问题尚未得到解决，此外现有工作仅针对特定类型的古文字，如甲骨文，无法在历史长河中描述汉字的演变。针对这些问题，本文在 ZiNet 中为字形、字音与字义知识定义了专业的属性与关系，整合并关联甲骨文、金文、楚简、隶书等各历史时期的汉字。此外，提出与专家协作的半自动化拓本图像去噪、部件检测、释文通假字标注、字体库构建等数据处理与标注流程，提升该领域知识图谱的构建效率，并将 ZiNet 应用于智慧古文字检索平台。

第二，本文针对古文字字形的部件语义表示与相似度量化问题，研究基于知识图谱的古文字字形相似度计算方法。现有工作主要针对汉字的轮廓结构和视觉形态进行相似度计算，没有对部件之间的语义关系进行建模，而后者对于古文字研究非常重要，因为部件可以揭示汉字的字音、字义、字源与造字方法。针对该问题，本文基于随机游走和深度残差网络（ResNet）分别对知识图谱中的字形实

体以及古文字图像进行表示，并提出象形相似性（PicSim）、部件描述相似性（RLCSSim）、部件语义图谱相似性（GraphSim）三个古文字字形相似度的量化方法。实验与案例分析证明，该方法能够捕捉汉字的视觉与部件语义相似性，所推荐的相似古文字更符合领域专家的研究需求。

第三，本文针对古文字多模态特征融合与未知字形考释问题，研究部件语义知识驱动的古文字字形识别方法。AI 辅助的古文字考释处于起步阶段，现有工作基于计算机视觉模型，生成古文字图像对应的现代汉字图像，但是这类方法仅基于视觉特征，生成的汉字图像缺乏部件构成逻辑，难以模拟演变过程中部件构成发生较大变化的汉字，而人类专家可以根据未知字形的部件特征联想到与之具有字源关系的汉字。面对该问题，本文从另一视角，提出古文字字形识别的创新任务，旨在融合图像、部件描述文本、字形知识图谱的特征，将未知字形分类到所属的文字标签，并提出相应的由多模态编码器和解码器构成的汉字字形识别模型（Multimodal Character Glyph Identification，MCGI）。实验证明了方法的有效性，以及部件与视觉图像特征对于该任务的有效性。

第四，本文针对汉字语素含义的历时变化发现与分析问题，研究汉字语素含义挖掘与历时语义跟踪方法。现有工作对历史文献进行词义消歧，生成对目标词汇的义项随时间变化的频率分布可视化分析，但他们未对语素进行表示，汉字可以作为语素构成复合词汇来表达其语义。针对该问题，本文在语素层面对现有方法进行扩展，提出一个汉字语素含义挖掘与历时语义跟踪框架（Chinese Morpheme Sense Mining and Tracking，CMSMT），并提出汉字语素含义挖掘的方法，训练一个双编码器的汉字语境表征模型，并构建相应的数据集。实验表明该模型可以提升词义消歧与语素含义挖掘的效果，可视化案例分析表明该方法可以从词汇与语素两方面对汉字的语义变化进行表示。

关键词：

自然语言处理，知识图谱，多模态特征融合，古文字字形相似度计算，古文字字形识别，汉字历时语义跟踪

Abstract

Research on Knowledge Driven Ancient Chinese Glyph Identification and Semantic Tracking

The ancient Chinese characters, represented by the oracle bone inscriptions (Oracle) and bronze inscriptions (Bronze), originated from the pre-Qin dynasty. Their writing font had not yet been unified. The same ancient character can possess more than one different glyph, and there are obvious differences between them and the modern characters. Up to now, there are still a large number of ancient Chinese characters that have not been deciphered. In the ancient character deciphering research, experts should compare the related glyphs in the whole history, explore the laws of character evolution, and ultimately infer the possible corresponding Chinese character for it. However, there are tens of thousands of glyphs in Chinese history, and their glyph, pronunciation, and semantic knowledge are described by unstructured image and text data, which are scattered across different historical records and platforms and difficult to be associated and retrieved. Ancient characters deciphering highly depends on the human experts' experience, memory and reasoning ability. Facing the above problems, the goal of this paper is to use artificial intelligence (AI) to construct a knowledge graph linking Chinese characters and glyphs; based on the knowledge graph, to quantify the glyph similarity between ancient characters to recommend similar characters for experts; furthermore, to predict the Chinese character label for the unknown glyphs to inspire experts for deciphering research; finally, the visual analysis of the diachronic changes of the meanings of the target Chinese character is generated. This paper has significance for the inheritance of ancient Chinese characters. The details of the contents and contributions include:

Firstly, focusing on the problem of glyph, semantic and pronunciation knowledge integration and linkage in the field of ancient Chinese characters, this paper put forward on the ontology model and construction method of the diachronic Chinese character knowledge graph (ZiNet). The related works focused on linking tangible entities, such as cultural relics, rubbings, etc. However, the problem of how to represent and

interconnect abstract knowledge of glyphs, pronunciations, and meanings has not been solved. In addition, the existing work only focuses on one specific type of ancient characters, such as oracle bone inscriptions, and cannot describe the evolution of Chinese characters throughout the entire history. Focusing on these problems, this paper defines professional properties and relationships around the glyph, pronunciation, and meaning in ZiNet, integrating and linking Chinese characters from various historical periods such as oracle bone script, bronze inscriptions, Chu bamboo slips, and clerical script. In addition, this paper proposes a semi-automatic data processing and annotation process in collaboration with experts, including rubbing image denoising, radical detection, interchangeable character annotation, and font database construction. At present, ZiNet has been applied to an intelligent retrieval platform.

Secondly, focusing on the problem of ancient Chinese glyph representation and the ancient glyph similarity quantification, this paper put forward on the ancient character glyph similarity measurement method based on knowledge graph. Related works mainly calculate the structure and visual features of Chinese characters, without modeling the semantic relationships between radicals. Radical is significant for the ancient character deciphering, because it can reveal the pronunciation, meaning, origin, and the method of character creation. To solve these problems, this paper uses random walk and deep residual network (ResNet) algorithms to respectively represent the glyph entities in the knowledge graph and the character images, and proposes three similarity calculation methods, including pictographic similarity (PicSim), radical description similarity (RLCSSim), and radical semantic knowledge graph similarity (GraphSim). Experimental results and the qualitative analysis showed that this method can capture the visual and radical semantic similarities and meet the demands of experts.

Thirdly, focusing on the problem of multimodal feature fusion and unknown ancient Chinese character glyph deciphering, this paper put forward on the ancient Chinese glyph identification method powered by radical semantics. The existing works generated the corresponding modern character images for the Oracle glyph based on the image generation algorithms. However, these methods only use the visual features, and the generated images lack the logic of radical composition. Their effects will be

limited if the radical composition of the target glyph has large changes through the evolution. While human experts can associate the unknown glyphs with related characters based on their radical features. Facing these problems, this paper proposes an innovative task of ancient glyph identification, which fuses the features of image, radical description text, and character knowledge graph to classify the target glyph to the character label. And proposes a multimodal Chinese glyph identification framework (MCGI) composed of multimodal encoders and a decoder for this task. The experiment proved the effectiveness of this method, as well as the effectiveness of the radical and visual features for the task.

Finally, focusing on the problem of the discovery and visual analysis of diachronic semantic changes of Chinese characters and morphemes, this paper put forward on the method of character morpheme sense mining and diachronic semantic tracking. The existing works carried out word sense disambiguation on the contexts of historical documents to generate a visual analysis of the frequency distribution of the meanings of the target word over time. However, these works did not analyze the morpheme, the Chinese characters can be used as morphemes to form compound words to express their semantics. Faced with this problem, this paper extends the existing methods at the morpheme level, proposing a Chinese morpheme sense mining and tracking (CMSMT) framework, training a dual encoder Chinese character context representation model and constructing the corresponding dataset. The experiment showed that this model can enhance the effects of word sense disambiguation and morpheme sense mining. And the case study shows that this method can express the semantic changes of Chinese characters from two aspects of word and morpheme.

Keywords:

Natural language processing, Knowledge graph, Multimodal feature fusion, Glyph similarity measurement for ancient character, Ancient Chinese glyph identification, Diachronic semantic tracking for Chinese character

目 录

第 1 章 绪 论

1.1 本文的研究背景及意义

1.1.1 研究背景与挑战

中国汉字的发展历经 3000 余年，可以划分为古文字和今文字两个阶段，古文字特指秦以前留传下来的篆体汉字，主要包括商代出现的契刻在龟甲或兽骨上的甲骨文、铸刻在殷商与周朝青铜器上的金文、以及战国时期的文字。古文字记录了先秦时期祭祀、天文、历法、政治、社会和生产等各方面的事件，为研究我国历史、文化和语言提供了珍贵的资料。图 1.1 展示了汉字“春”在甲骨文、金文、楚简等古文字阶段，以及秦统一文字后的小篆、汉代隶书、楷书时期的字形演变过程。可以看出，古文字字形与今天的汉字相比存在很大的差异，且由于地理和政治隔绝等影响，古文字的书写不固定，字形尚未统一，图 1.1 展示了汉字“春”字的 9 个甲骨文字形，3 个金文字形和 3 个楚简字形，本文称为异体字形，异体字形之间的形态和部件结构都不相同，图中为每个字形标注了部件组成。

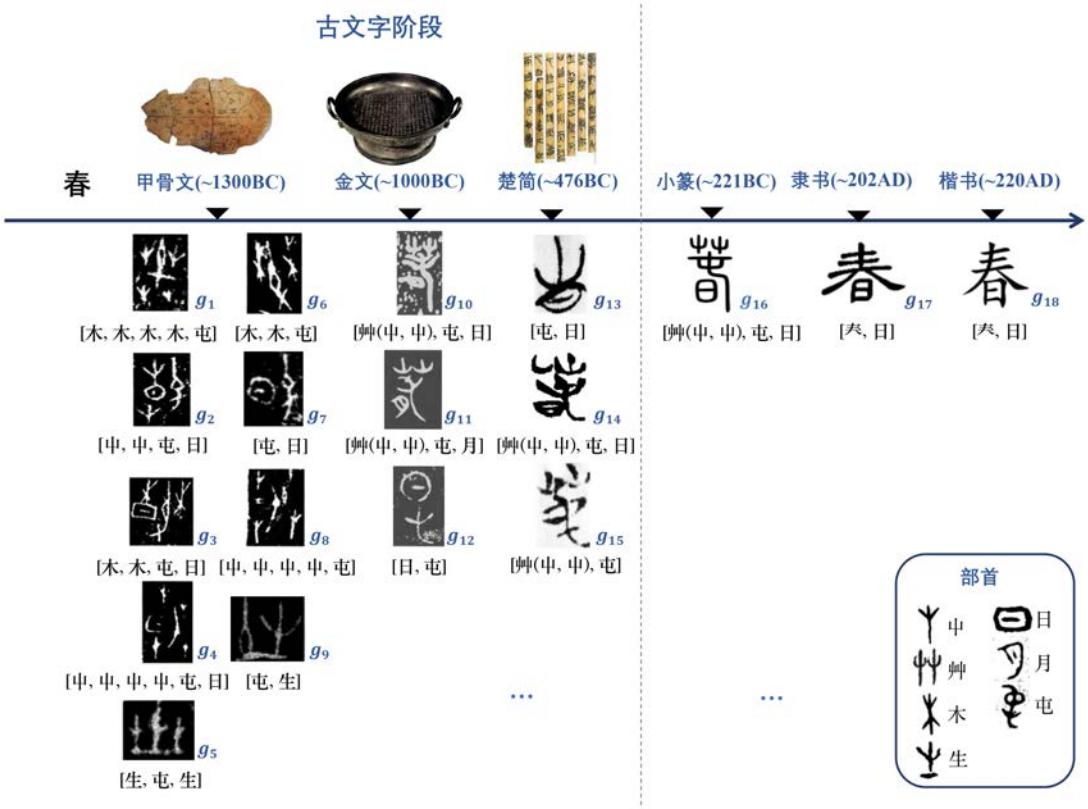


图 1.1 汉字“春”的历时演变及多种异体字形图例

古文字考释是一项非常具有挑战性的研究，在 4000 余个已发现的甲骨文字中，至今已经被完全释读的仅 1500 个左右。对于未释字来说，目前尚不清楚它们对应的现代汉字，也无法为其字音、字形、字义提供完善的解释。在字形层面的古文字考释中，专家需要对待考字形的部件进行拆分，通过形符部首与声符部首初步分析其字音与字义，并与历史上所有书写形态或部件构成相似的其他汉字进行比较和关联，基于从已知文字字形演变归纳出的潜在规律来分析它们之间是否具有字源关系，最终推理出该字形可能对应的现代汉字。这一过程考验着人类专家的记忆、经验与推理能力，长期以来，古文字研究面临以下困难：

第一，古文字领域知识的整合与检索面临挑战。古文字研究需要对各历史时期的文字的字形、字音、字义等属性信息进行查询、比较与分析，这些专业知识通常由专家从拓本图像、出土文献释文、字典、前人著作等多模态数据中学习和提取，这些数据大多分散在各类平台和专业书籍中，缺乏结构化的知识表示与整合，形义相近的古文字难以被高效地关联和查询。

图 1.2 展示了编号为“15625 正”的甲骨文物的主要信息，以及拓本图像与释文文本数据的案例，拓本图像是从文物上拓印获得的图像，释文是由古文字学家将用古文字书写的出土文献改写为对应的现代汉字，并对标点符号、通假字、缺失文字等进行标记所产生的文本数据（释文含义注释：“貞”：卜问；“翌”：未来；“癸酉”：纪年；“燎”：一种祭祀；A（B）：表示通假字）。

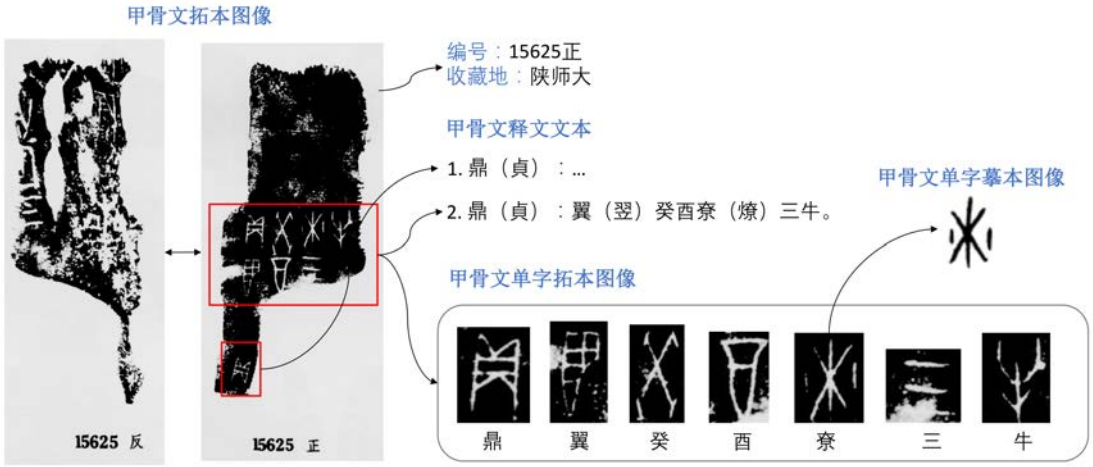


图 1.2 甲骨文拓本图像、文物信息与出土文献释文图例

第二，古文字相似字形发现面临挑战。中国历史上存在上万个字形，发现可以为古文字考释研究提供启发、与目标字形相似的其他汉字依赖于专家的记忆与经验，需要耗费大量的时间与精力。

第三，未知古文字字形的考释研究面临挑战。从上万个已知文字字形的历时演变中归纳出潜在规律，并对未知字形进行推理，将其关联到已知汉字的古文字破译研究挑战着人类专家的脑力极限。

第四，汉字的历时语义变化挖掘与分析面临挑战。从历史文献中分析汉字的历时语义变化对于古文字、语言学、词典编撰等研究具有重要意义，但是人工分析非常耗时且无法穷尽历史文献。

这些问题显示出基于人类经验和推理能力进行古文字研究与考释的局限性，近年来，人工智能（Artificial Intelligence， AI）的发展为古文字研究带来新机遇，知识图谱（Knowledge Graph）善于以人类的逻辑和认知表示和关联领域知识，以神经网络（Neural Network）为基础的深度学习算法善于对已有的数据特征进行归纳，从而对未知数据的性质进行预测。因此将 AI 应用于古文字研究是解决上述问题的有效途径。

1.1.2 研究意义

与上述古文字领域的四个实际问题相对应，本文的研究旨在将 AI 应用于古文字研究领域，构建一个关联所有古文字与字形的历时汉字知识图谱，为领域专家提供智能化的知识检索服务；研究基于知识图谱的古文字字形相似度计算方法，为专家自动化地推荐历史上的所有相似字形；研究面向古文字考释的字形识别方法，为专家提供对未知古文字字形所属的汉字的预测结果，辅助古文字考释工作；研究汉字的历时语义跟踪方法，为专家生成目标汉字字义的历时变化可视化分析。本文工作对于古文字领域专家的知识获取、思维启发，工作效率的提升都将具有重要帮助，其理论及应用意义具体如下：

在古文字领域知识建模以及多模态数据的关联与整合方面，已有的领域知识图谱[1]，链接甲骨文的图像、释文、文物、著录等多模态数据，但是工作局限于甲骨文，且未对古文字的字形、字音与字义知识进行进一步的建模。此外当前古文字多模态数据的处理、知识抽取和标注的过程严重依赖于专家劳动，制约了古

文字知识图谱的构建和更新效率。本文提出了一个历时汉字知识图谱，称为 ZiNet，致力于关联各历史时期字形、字义相似或相关的汉字。并且针对古文字领域的数据处理与标注的特殊问题，研究与古文字领域专家协作的半自动化知识图谱构建方法，提升专家的数据处理和标注效率，为后续的古文字知识建模与资源构建工作提供理论支撑。在应用层面，该知识图谱已经服务于智能化的古文字知识检索平台，并为该领域的 AI 算法提供支持。

在古文字字形相似度计算研究方面，尽管已有很多汉字字形相似度的量化方法[2-7]，但是由于古文字的很多特征与现代汉字不同，多数方法难以应用于古文字，例如古文字的象形程度高，无法拆分为笔画和固定结构，也没有计算机编码、拼音等现有方法中利用的属性。更关键的是，部首组成可以揭示文字的字义、字音、字源与造字逻辑，是古文字字形考释的重要线索，已有方法主要针对汉字的视觉形态、整体结构的相似性进行计算，未对汉字的部件语义相似性进行建模。例如 “頁”与“首”、“鳥”与“隹”、“攵”与“止”等古文字具有字源或近义关系，因此这些文字、以及包含两者作为部件的文字应该具有更大的相似性。本文提出了基于视觉特征与知识图谱的古文字字形相似度计算方法，该方法可以捕捉部件语义相关的文字，为专家提供更符合其研究需求的字形推荐服务。

在 AI 驱动的古文字考释研究方面，已有的相关研究利用图像生成算法模拟汉字图形的历时演变，输入古文字图像，生成与之对应的现代汉字的图像，从而启发专家进行考释研究[8-9]，或者识别古文字图像中的部件，然后将其重构为现代汉字对应的部件[10]。但是一个重要问题是，这些方法仅基于图像特征，未对汉字的部件特征进行表示，生成的现代汉字图像的笔画仍然较为随意，缺少部件层面的造字逻辑，也难以准确推理出对应的现代汉字的部件构成，而那些难以被破译的古文字的部件组成在演变中很可能已经发生了较大的变化。例如图 1.1 中“春”字的异体字形  （由部件“屯”、“生”组成）的部件组成与现代“春”字存在显著的区别，已有方法倾向于生成类似于“屯生”的现代汉字图像，很难准确推理出变化后的部件构成，因此难以将其映射到正确的现代汉字，但是人类专家可以根据形符部件“生”表示的含义，以及声符部件“屯”的发音将其与“春”字联系在一起。因此本文认为古文字考释系统应该具有对异体字形的部件语义特

征进行表示与归纳的能力，从而对部件发生变化的字形进行预测。基于该思想，本文提出利用图像和部件特征将未知字形分类到所属文字标签的创新性任务，以及对应的知识驱动的多模态模型，该方法在理论层面创新性地引入部件语义知识图谱，显式地学习所有已知字形的部件语义特征。该成果可以为领域专家推荐未知字形可能属于的文字排序，以及对应的相似字形，从而启发古文字考释工作。

在汉字历时语义跟踪研究方面，本文遵循 Hu 等人提出的词汇语义跟踪框架[11-12]，他们基于预训练语言模型对历史文献上下文进行词义消歧，即将上下文中的目标词汇分类到对应的词典义项中，然后对每个义项在每一个时间阶段的频率进行统计，通过多项式拟合后绘制目标词汇所有义项的历时频率分布函数的可视化分析。随着汉语的发展，汉字越来越多地作为语素组成复合词汇。因此，本文对已有的历时语义跟踪框架在语素层面进行扩展，研究汉字的语素含义挖掘的方法，从词汇和语素两方面表示汉字的语义变化，并针对汉字词汇和语素含义的历时变化相关性进行了有趣的统计分析。该成果可以提供汉字语义变化的可视化分析，辅助进行语言文字的相关研究以及词典编撰等工作。

综上所述，本研究在人工智能与古文字交叉研究领域不仅具有理论方法创新，也具有广泛的应用价值，其成果可以辅助古文字领域专家进行考释研究，加速我国古文字研究的数字化、智能化进程，对于这一“冷门绝学”的发展与传承具有重要价值，对于弘扬中华文化具有重要意义。

1.2 国内外研究现状与局限性分析

本章首先对人工智能应用于古文字研究的发展情况进行概述（1.2.1 节），然后分别对与本文四个研究内容对应的相关工作以及当前的局限性进行介绍和分析，包括古文字领域知识图谱构建相关研究（1.2.2 节）、古文字字形相似度计算相关研究（1.2.3 节），AI 驱动的古文字考释相关研究（1.2.4 节），以及词汇历时语义跟踪相关研究（1.2.5 节）。

1.2.1 人工智能驱动的古文字研究概述

古文字的载体为甲骨、竹简、帛书、器具等文物，通常以拓本图像的形式进行记录，古文字上下文也会由领域专家整理成便于阅读的文本形式进行传播。因此古文字图像处理和历史文献文本理解是 AI 驱动的古文字研究的两大方向，前

者基于计算机视觉（Computer Vision，CV）算法，后者与自然语言处理（Natural Language Processing，NLP）相关。这些研究成果已经被应用于古文字图像检测、识别、修复、历史文献文本分析、知识检索等多项服务中，如图 1.3 所示。

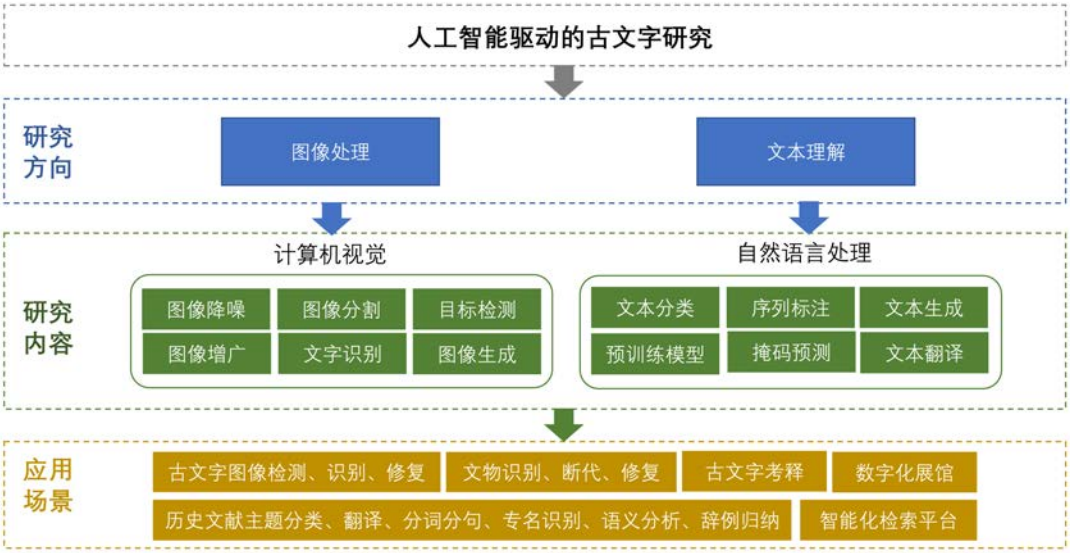


图 1.3 人工智能驱动古文字研究现状

在古文字图像处理方向，古文字图像降噪、古文字检测与识别是最常见的研究。出土文物长年受到风化与腐蚀，导致拓本图像具有很多噪声，因此图像降噪通常是古文字图像处理的第一步，Shi 等人基于自注意力机制（Self-Attention）重塑噪声样本图像[13]；Li 等人提出了一种中文汉字自动修复的扩散模型 [14]。在古文字检测与识别任务中，古文字数据集的小样本和长尾分布，即类别间样本数量不平衡，是该任务需要解决的重要问题。Li 等人构建了 HWOBC 甲骨文手写数据集，服务于甲骨文识别任务[15]。Wang 等人针对小样本问题，研发了无监督域自适应方法[16-17]。Zhang 等人提出了基于深度度量学习的分类方法应对长尾分布[18]。此外，对尾部类的样本进行扩充也是一种常见的策略，Li 等人提出了一个生成对抗框架来增强尾部类中的甲骨文样本[19]；Wang 等人提出了基于 CycleGAN 的数据扩充方法[20]；Li 等人提出了一种扩充少数类的样本的混合策略[21]；Yue 等人提出了一种改进的生成对抗网络来扩充数据集[22]。Xu 等人针对新的古文字不断被发现，训练阶段难以覆盖所有的汉字类别的问题，提出了一个大规模连续学习框架[23]。此外，一些工作致力于零样本学习，识别无训练数据的古文字，以解

决小样本问题，这些工作通常利用图像中的部首特征[24-28]。Diao 等人基于目标检测模型提取古文字图像中的部首信息，从而推理目标文字标签[28]。

计算机视觉方法也在文物修复、古文字考释、器物断代等方面取得广泛应用，Zhang 等人提出了一种用于甲骨文残片缀合的深度模板匹配算法[29]。Zhang 等人对未知图像返回一组相似的图像和相关的学术信息，指导古文字考释[30]。Chang 等人提出了一个图像到图像的翻译网络 Sundial-GAN，以生成古文字相应的现代汉字图像，指导古文字考释[8]。Xu 等人提出了 Conf-UNet 模型，将未识别的甲骨文与大篆联系起来，为专家推荐相关的篆书[32]。Zhou 等人引入了一个多头分类器和关系图实现了青铜鼎的考古测年[33]。

在历史文献文本理解方向，AI 已经被用于预训练语言模型训练、历史文献文本修复、文本归因、语言分析、诗歌等文本生成等任务。预训练语言模型（Pre-trained Language Models）在大规模历史文献语料上进行训练，并在特定下游任务上微调。目前的古汉语预训练模型包括 SikuBERT[34]、GuwenBERT[35]、CANCN-BERT[36]等。由于出土文献书写载体的破损，历史学家必须重建文本中丢失或难以辨认的部分，这一过程被称为文本修复。Ithaca 古希腊铭文修复工具通过字符级的缺失补全方法修复被破坏的铭文[37]。Luo 等人提出了一种序列到序列的神经网络模型来捕捉同源词之间的字符级对应关系，破译失传的语言[38]。Mo 等人基于在《四库全书》语料上训练的预训练语言模型 BERT（Bidirectional Encoder Representations from Transformers）[39]，模拟辞例归纳能力，预测历史文献中被遮蔽起来的文字[40]。文本归因旨在推断文本的时间归属、地理归属、作者归属、文体类别等[43-45]，比如属于散文、诗歌还是戏剧。历史文献的语言分析包含多种通用 NLP 任务，例如文本分类、分词、词性标注、命名实体识别、词义消歧、词义变化检测等，预训练语言模型以及早期的 Word2Vec[46]等方法在词汇表示学习的过程中被广泛使用。古文生成包括相似风格内容文本的生成与翻译[47]，相关研究致力于诗歌生成、诗歌意象分析[48]等与古文相关的任务。

上述技术的发展催生了一系列平台与应用，包括安阳师范学院的殷契文渊甲骨文大数据平台，是一个集甲骨文字库、著录库、文献库为一体的甲骨文知识共享平台；华东师范大学的商周金文智能镜古文字识别平台[49]，基于 AI 技术实现

古文字及其各种属性的系统识别、成篇文字材料的整体性识别、文字的文物载体的特征性影像识别等。河南大学的“缀多多”工具[29]，是一款利用 AI 技术进行甲骨残片缀合的软件；以及微软亚洲研究院与首都师范大学联合开发的甲骨文整理校重助手 Diviner，第一次将自监督 AI 模型引入到甲骨文的“校重”工作中，大幅提升了工作效率。

综上所述，人工智能驱动古文字研究近年来取得了突破性的进展，并在多个场景中应用，帮助相关学者高效处理、识别、修复和理解出土材料、检索信息。但是，现有的工作仍然存在局限性：（1）大多数工作致力于理解单一模态的数据，但是古文字的属性包括字形、字音、字义等多个维度，研究材料也包含图像、文本、视频等多模态数据，融合多模态信息才可以更完整地表示古文字的特征。（2）该领域的 AI 模型还不具备认知能力，现有的模型大多针对特定的任务，例如文物残片缀合、古文字图像识别、上下文缺失文字预测、人名与地名识别等，但是如何利用 AI 考释未知的古文字仍然是一个难题。类比人类学者根据古文字的字形结构、字典释义、历史文献上下文等领域知识，从字形、字音、字义、语用多角度进行关联和推理，目前的 AI 模型还不具备这样的推理能力。（3）古文字领域的各类数据集、知识图谱等数字化资源仍然稀缺，历史上字形、字义相关的文字与多模态数据缺乏整合与互联，不利于数字化检索和模型训练。

1.2.2 古文字知识图谱构建相关研究

知识图谱已被广泛应用于世界文化遗产知识的组织与关联，Europeana 项目将音乐、书籍和艺术品等数字遗产中的主题、时代和机构等元素联系起来，实现了对欧洲文化遗产资源的检索[50]。Ichpedia 项目构建了一个百科全书系统，使用户能够通过语义和地图检索来关联非物质文化财产元素[51]。Carriero 等人构建了意大利文化遗产的知识图谱[52]。在国内，文化遗产知识图谱覆盖传统书画、陶瓷文物、云锦工艺、石窟艺术、古都文化等各个领域[53-58]，这些项目通过将文化遗产知识在语义上联系起来，促进了对文化遗产数据管理和共享传播，提升了对文化现象本质的解释。在古文字知识图谱研究领域，Han 等人提出了一个甲骨文信息系统 IsOBS [59]，该系统记录甲骨文拓片、文档、甲骨文字及其所有的异体字形。Xiong 等人提出了一个多模态甲骨文知识图谱[1]，他们总结了甲骨文领域的

基础数据，包括但不限于：（1）甲骨文彩色照片（2）甲骨文拓本图像（3） 甲骨文摹本图像（4） 甲骨三维建模数据（5） 甲骨文 DNA 鉴定数据（6）甲骨文考古数据（7） 甲骨文书法（8） 甲骨碎片数据（9） 甲骨文数据库（10） 甲骨文释文文本（11）裁剪出的甲骨文单字拓本图像（12）甲骨文单字复制图像（13）甲骨文视频。甲骨文知识图谱致力于关联这些数据，以及相对应的字形、文物馆藏、研究学者、文献等信息，形成一张多模态知识网络。

在国外的词汇语义知识库的研究领域，最为著名的词汇知识库是英语 Princeton WordNet [60]，在它的体系结构中，同义词集（Synset）是整合单词的基本语义单元，每一个同义词集包含表示相同概念的词汇，对应一个唯一的标识符（ID）、简短的定义（Gloss）和一些使用样例。同义词集之间通过关系相互连接，形成概念语义网络，其中最常见的是上位关系（Hypernym）和下位关系（Hyponym），上位关系表示词义泛化，下位关系表示词义更为具体，例如“动物”是“哺乳动物”的上位概念，反之“哺乳动物”是“动物”的下位概念。以 WordNet 结构为基础产生了很多其他语言和跨语言的词汇知识库，Zhang 等人建立了一个先秦词汇 WordNet，称为 PQAC-WN[61]。多语言词汇知识库 Open Multilingual Wordnet（OMW）[62]、BabelNet[63-64]和 The Universal Knowledge Core（UKC）[65] 致力于以同义词集为纽带连接世界各地的词汇和概念。这些知识库已经被用于词义消歧、同源词发现等很多自然语言处理任务[66-69]。

根据调研，知识图谱已经被广泛应用于文化遗产以及词汇知识的组织和检索，但是古文字的领域知识图谱相关工作还很少。与本文最相关的工作是 Han 等人提出的甲骨文知识图谱，其局限性包括：（1）现有的古文字知识图谱致力于整合与链接古文字领域的多模态实体数据，例如文物、拓本图像，摹本图像、出土文献、著录等，未对更为抽象的古文字字形、字义知识进行表示与关联，且现有工作仅关注特定历史时期的古文字类型，未能在完整的历史中描述汉字的互联与演变。（2）古文字知识图谱的构建效率需要进一步提升。由于古文字领域知识的专业性，领域专家需要对出土的原始数据进行修复、切割、翻译、标注等大量的预处理、知识抽取和链接工作。如何利用自动化的方法减轻领域专家的工作负担，提升知识图谱构建和更新的效率有待进一步研究。

1.2.3 古文字字形相似度计算相关研究

汉字的字形相似度计算相比于印欧语言来说复杂很多，目前的相关工作针对现代汉字。Qi 等人将汉字的图像像素以及五笔编码的字符串分别转化为特征向量，利用二值化差值算法和改进后的 Jaro-Winkler Distance 算法分别计算相似度，融合后得到最终的相似度结果[3]。Hu 等人提出一种基于笔画的字向量表示方法，结合汉字的构词和拼音属性，将汉字映射为一个仅 32 维的空间向量[4]。Wang 等人将汉字表示为字形结构、字首部件和字尾部件的三元组，以部件为运算对象，字形结构为运算符，将汉字描述为前缀表达式，建立汉字字形相似度计算递归模型，使计算过程被逐层分解为原子部件间的相似性比较[5]。Su 等人提出了汉字字形增强词汇表示的模型，他们利用卷积自编码器（Convolutional Auto-encoder，convAE）从汉字的图像中学习字形特征[6]。Liu 等人构建了偏旁部首知识图谱，并提出相应的汉字部件拆分方法，基于该图谱提出 2CTransE 模型，学习汉字实体的向量表示，用于汉字字形的相似度计算[7]。Yang 等人利用视觉变换器（Vision Transformer，ViT）模型，在汉字的多种字体的数据上基于对比学习进行自监督训练，将汉字图像映射为高维向量来计算距离 [2]。

可以看出，相关工作对汉字的图像、笔画、拼音、部首、字形结构这些基本属性中的一种或多种进行特征表示，从而计算字形相似度，这些方法适用于现代汉字，但是对古文字进行相似计算时会面临如下问题：（1）多数方法无法迁移到古文字，因为古文字的很多特征与现代汉字不同，例如，古文字的象形程度高，难以拆分为规范的笔画；古文字部件位置的书写不固定，因此结构不同的字形也可能具有相似性，此外，古文字也不具有拼音、编码等现代汉字可以利用的属性信息。（2）相关方法未针对古文字的部件语义相似度进行计算，例如一些古文字虽然视觉形态差异较大，但是包含语义相关的部件，也应该具有字形相似性。已有工作的相似度计算多数是为了防止人类或机器的误认、误写而设计，因此主要针对文字的轮廓形态等视觉方面的相似性进行度量，而本文的目的不仅是为了检索形态相似的文字，更重要的是推荐字形所反映出的字义、字音、字源、造字方法上具有关联的文字，从而启发领域专家进行研究，因此需要对部件语义知识进行进一步的相似度表示。

1.2.4 人工智能驱动的古文字考释相关研究

古文字考释是一项非常具有挑战性的研究，需要从字形、字音、字义、语用多个角度进行推理，并对考释结果从各方面进行充分解释。现有的面向该任务的 AI 驱动的方法还非常少，这些工作多从字形角度入手，模拟和生成古文字演变后的现代汉字图像，为古文字专家提供参考。Zhang 等人对未知图像返回一组相似的图像和相关的学术信息，指导古文字考释[30]。Chang 等人提出了一个用于破译甲骨文的级联生成对抗网络（Generative Adversarial Networks，GAN）框架，称为 Sundial-GAN，该网络输入甲骨文图像，生成对应的现代汉字图像，网络由四个 GAN 结构按顺序级联构成，每一个结构对应汉字演变的一个代表性阶段[8]。Guan 等人提出了一种基于条件扩散的图像生成策略，生成甲骨文图像对应的现代汉字图像，特别地，为了使模型学习甲骨文的局部部件结构并将其映射到对应的现代汉字，该模型引入局部结构采样技术，增强了模型识别和解释复杂汉字模式的能力[9]。Wang 等人通过部件重构来破译甲骨文，他们将甲骨文解构为基本的部件，然后利用变换器模型将其重构为现代汉字[10]。

综上所述，AI 辅助的古文字考释研究尚处于起步阶段，现有方法的局限性在于：（1）相关工作大多仅基于从图像中提取的视觉特征来对字形进行表示，基于图像生成模型的方法致力于生成对应古文字的现代汉字形态，但是目前的算法难以理解汉字的笔画和部件，因此生成的现代汉字图像笔画混乱，缺乏逻辑。图 1.4（a）展示了 Guan 等人[9]提供的五个未知古文字考释结果案例，可以看出，尽管他们的方法通过局部结构采样将一部分甲骨文部件正确地映射为现代汉字的部件，例如将“”映射到“宀”，但是在很多情况下，生成的“汉字”的笔画和部件构成没有象形、会意、指事、形声的内部逻辑，因此难以与已知的汉字联系起来。（2）基于部件重构策略的方法将从图像中识别到的甲骨文部件映射为对应的现代汉字部件并重构为现代汉字，图 1.4（b）展示了 Wang 等人[10]的方法对两个未知甲骨文的考释结果。但是方法的局限性在于，对于难以被破译的字形来说，其部件组成与后代文字很可能具有较大的差异，无法完成直接的部件转换，考释的目的之一就是需要推理出变化后的部件组成。例如图 1.1 中“春”字的很多异体字形的部件组成与现代汉字“春”存在显著的区别，无法直接进行部

件映射，但是其形符部件（“木”、“艸”、“生”）表示的含义，以及声符部件（“屯”）的发音仍然与“春”字存在联系，因此本文认为古文字字形考释系统必须利用文字的部件语义信息，且具有对历史上所有异体字形的部件语义特征进行表示与归纳的能力，从而对部件组成发生变化的未知字形进行预测。

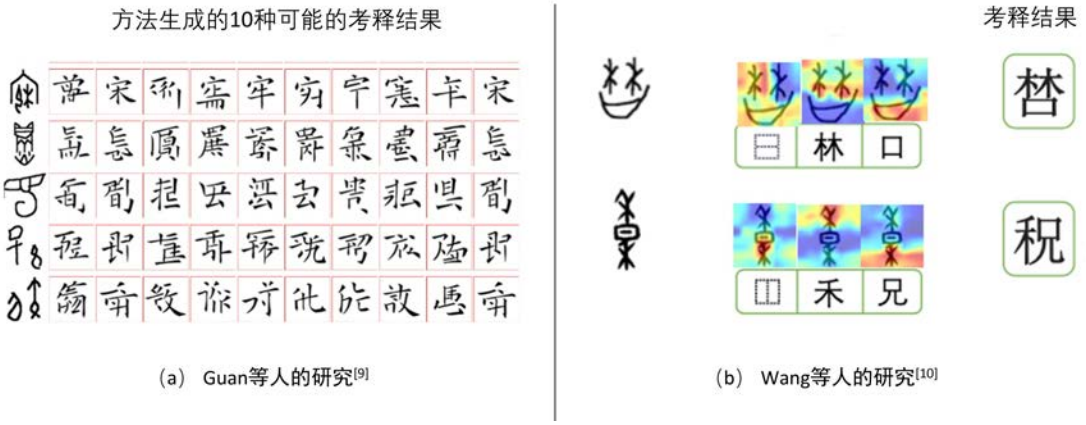


图 1.4 相关工作的古文字考释结果图例

1.2.5 词汇历时语义跟踪相关研究

NLP 领域中的词汇语义变化检测（LSCD）任务与本文工作较为相关。目前大多数学者使用上下文嵌入算法来表示词汇的语义特征。通过计算不同时间阶段的词嵌入向量的距离来计算词汇的语义是否发生变化。词汇向量可分为静态嵌入和语境化嵌入两种。静态模型，如跳字模型（Skip-Gram，SGNS）和连续词袋模型（Continuous Bag of Words，CBOW）[70]，将每个词汇表示为单一的嵌入向量，而语境化模型，如近年来兴起的预训练语言模型，可以为在不同语境中的同一个词汇生成不同的向量，以表示词汇的多义性。相对早期的一些研究使用静态模型来挖掘词义变化[71-77]，为了在细粒度的级别上表示词汇的每一个含义，近年的一些工作引入了语境化模型，Giulianelli 等人提出了第一个利用 BERT 词汇表证的无监督词汇语义变化检测方法[78]。Kurtyigit 等人证明，静态模型和语境化模型都可以成功地发现词义变化[79]。Rosin 等人提出了一种称为 TempoBERT 的时间上下文语言模型，该模型将时间信息作为文本的附加上下文来增强语义变化检测的性能[80]。Teodorescu 等人将词汇语义变化检测作为词义消歧问题[81]，并利用 XLM-RoBERTa 模型实现[82]。Hu 等人提出了一种词义的变化检测框架，该框架基于 BERT 和词典数据对目标词汇在历史语料库中的上下文进行词义消歧，从而

为词汇的每个含义提供随时间变化的平滑的频率分布表示[11]，Shu 等人的工作遵循该框架，对汉字从先秦至今的语义变化进行了表示和可视化分析，还为该任务构建了一个常用汉字的词义标注语料库[12]。

Teodorescu 等人[81]、Hu 等人[11]与 Shu 等人[12]的工作与本文的目标一致，他们将词义变化检测建模为词义消歧义问题，可以对词汇的每一个义项的历时变化生成细粒度的可视化，而不是仅提供对词汇语义是否变化的预测结果和对词汇含义的粗粒度表示。因此本文遵循他们的框架，但是当前还存在一些局限性：（1）现有的研究没有从语素的视角对汉字的语义变化进行表示。汉字在被创造之初被作为单音词使用，随着语言的发展越来越多地被用作构成复合词的语素，因此挖掘语素含义对于探索汉字语义的发展具有重要意义。（2）预训练的语言模型在汉字的词义消歧、语素含义挖掘任务中的效果有限，需要进一步对汉字语境表示模型进行训练，用于汉字表征训练的数据集也需要进一步扩展。

1.3 本文的研究内容与创新点

本文的研究内容与相应的研究问题如图 1.5 所示：

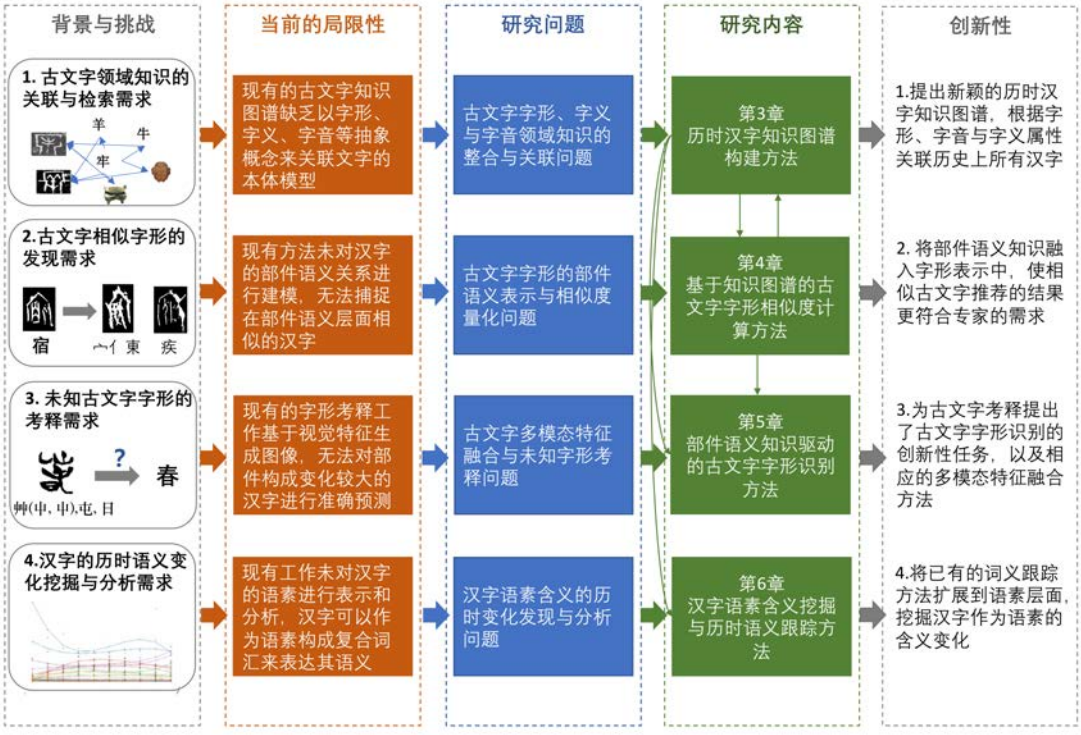


图 1.5 本文的背景、现有工作的局限性、研究问题、研究内容与创新点

本文面向 1.1 章描述的领域专家在古文字考释研究中的挑战与实际需求，基于 1.2 章介绍的当前人工智能驱动的古文字研究的局限性，分别针对古文字字形、字义与字音领域知识的整合与关联问题、古文字字形的部件语义表示与相似度量化问题、古文字多模态特征融合与未知字形考释问题、以及汉字语素含义的历时变化发现与分析问题开展创新研究，分别对应本文的四个章节。

四项研究之间具有内在联系：第一项研究构建的汉字知识图谱融合了汉字的字形与字义信息，并以此关联汉字，因此可以为第二、三、四项研究提供知识和数据支撑；第二项研究基于知识图谱的古文字相似字形计算方法所捕获的形近字信息被用来补全第一章汉字知识图谱中的语义关系，从而进一步链接相近的字形和文字，同时第二章的方法也被用于第三章的古文字字形识别研究，并与第三章提出的方法进行对比。

四项研究内容具体如下：

第一，针对古文字字形、字义与字音领域知识的整合与关联问题，本文研究历时汉字知识图谱及其构建方法。本文对古文字领域的实体和关系进行定义，提出了一个历时汉字知识图谱，称为 ZiNet，ZiNet 的本体层定义了文物、拓本图像、部件、字形、古文字、文字、词汇、义项、同义词九种概念，它对古文字的字形、字音和字义知识进行抽取和表示，并以此关联各历史时期的文字。此外本文提出了与专家协作的数据处理与整合方法，实现半自动化的古文字拓本图像去噪、部件检测、释文通假字标注、字体库构建等步骤。目前为止，该知识图谱整合包含甲骨文、金文、楚简在内的约 16000 余个古文字字形，并被应用于智慧古文字检索平台。

第二，针对古文字字形的部件语义表示与相似度量化问题，本文研究基于知识图谱的古文字字形相似度计算方法。本文基于随机游走和深度残差网络（ResNet）分别对知识图谱中的字形实体以及拓本图像进行特征表示，并提出了象形相似性（PicSim）、部件描述相似性（RLCSSim），以及部件语义图谱相似性（GraphSim）三个古文字字形相似度的量化方法，PicSim 适合独体的象形字，RLCSSim 和 GraphSim 适合由部件构成的会意字和形声字。定量评估证明，与基线方法相比，该方法与人类专家提供的样本相似度分数具有更高的正相关性，斯

皮尔曼相关系数为 0.8422，$p<0.05$。甲骨文相似字形推荐案例的定性分析表明，该方法能够捕捉部件之间的语义关系，符合古文字学者的专业需求。

第三，针对古文字多模态特征融合与未知字形考释问题，本文研究部件语义知识驱动的古文字字形识别方法。本文提出古文字字形识别的创新性任务，旨在利用字形的图像与部件信息，将其分类到所属的文字标签，启发领域专家进行研究。为此本文提出一个由基于 BERT、图注意力网络（Graph Attention Network，GAT）和 ResNet 组成的多模态编码器和一个无监督解码器构成的汉字字形识别模型（Multimodal Character Glyph Identification，MCGI），模型表示和融合字形的图像、部件描述文本以及汉字知识图谱的多模态特征，在历史上已知字形的数据集中进行训练，学习异体字形的视觉和部件语义特征，从而对部件发生较大变化的未知字形进行预测。实验表明，与基线方法相比，本文的方法取得了更好的效果，共时测试字形分类的准确率为 73.7%，历时组的准确率为 56.5%，并证明部件与视觉特征各自对于未知字形识别任务的有效性。

第四，针对汉字语素含义的历时变化发现与分析问题，本文研究汉字语素含义挖掘与历时语义跟踪方法。本文对已有方法在语素层面进行了扩展，提出了一个汉字语素含义挖掘与历时语义跟踪框架（Chinese Morpheme Sense Mining and Tracking，CMSMT），该框架包括汉字语境表示模型训练、词义消歧、语素含义挖掘、语义变化表示与可视化四个步骤。并且构建一个包含各个历史时期语境案例的义项-上下文数据集，用来训练基于 BERT 的双编码器汉字表征模型。实验表明该模型对汉字词义消歧与语素含义挖掘任务的效果具有显著提升，汉字语义跟踪的案例分析表明该方法可以从词汇与语素两方面对汉字的语义变化进行表示。最后本文进行了一个有趣的统计，发现汉字作为词汇与作为语素时的语义历时变化趋势具有正相关性。

本文的贡献与创新点如下：

第一，本文构建了一个新颖的历时汉字知识图谱（ZiNet），提出能够表示古文字字形、字音与字义知识的本体模式，关联各历史时期字形相似、字音相近、字义与字源相关的汉字及其多模态数据，并提出了与领域专家协作的数据处理与整合方法。

第二，本文提出了一种古文字字形相似度计算方法（GlyphSim）。该方法基于汉字知识图谱，创新性地将部件语义知识融入古文字的字形表示中，不仅可以计算古文字的视觉形态相似性，还可以捕捉字源和部件语义的相似性，使推荐的形近字更符合古文字领域专家的研究需求。

第三，本文面向古文字字形破译的难题，提出古文字字形识别的创新任务，并设计知识驱动的多模态汉字字形识别模型（MCGI）。该方法创新性地融合古文字的图像、部件描述文本与知识图谱的特征，从已知字形与对应汉字标签的映射中进行学习，对视觉形态和部件构成均发生较大变化的未知字形进行预测，为AI 辅助的古文字考释研究提供了一个新的策略。

第四，本文提出了一个汉字语素含义挖掘与历时语义跟踪框架（CMSMT），一个汉字语境表征模型，并构建了用于模型训练的历史文献义项-上下文数据集。该框架将已有的词义跟踪方法扩展到语素层面，挖掘汉字作为语素的含义变化，并就汉字的词汇含义和语素含义变化的相关性进行了有趣的统计分析。

1.4 本文的组织结构

本文总共分为七个章节，下面介绍各章的主要内容：

第 1 章为绪论。首先介绍研究的背景与意义，包括古文字研究面临的困难与挑战、领域专家的实际需求、研究的理论与应用意义，然后介绍国内外研究现状，包括人工智能驱动古文字研究的整体发展情况，以及本文四个研究内容对应的相关工作，并对已有方法的局限性进行分析和总结，最后介绍本文的四个研究内容与对应的研究问题、关键技术和创新点。

第 2 章为相关技术概述。将介绍本文研究使用的相关 AI 模型与算法，包括知识图谱、图注意力网络模型、预训练语言模型、计算机视觉模型。

第 3 章为历时汉字知识图谱构建方法。将介绍古文字领域数据与重点问题、汉字知识图谱的架构与本体模式、与领域专家协作的古文字领域知识图谱构建流程，包括拓本图像去噪、部件标注、字体库构建、释文通假字标注所使用的自动化方法，最后介绍目前为止知识图谱各个实体的统计信息，并展示基于该知识图谱建设的古文字智慧检索平台。

第 4 章为基于知识图谱的古文字字形相似度计算方法。首先将介绍该任务的定义，然后介绍本文的方法，包括古文字字形相似度量化的三个指标，最后介绍实验与评估，包括数据集与评价指标、实验设置、定量实验结果、定性分析与讨论。

第 5 章为面向考释的古文字字形识别方法。首先介绍异体字形的相关知识以及该任务的定义，然后介绍本文的方法，包括古文字图像、部件描述文本以及知识网络模态数据的表征与融合模型，接下来介绍实验与评估，包括实验所用的古文字数据集，基线方法、评价指标与实验设置，最后展示实验结果，进行讨论并分析各个模态的信息对该任务的贡献。

第 6 章为汉字的历时语义跟踪方法。首先将介绍汉字语素及历时语义变化的相关知识以及该任务的目标，然后介绍本文的方法，包括数据集、汉字语境表示模型训练、语素义项挖掘、以及历时语义变化表示与可视化，接下来介绍定量实验评估与汉字历时语义变化的可视化分析案例，最后介绍汉字词汇与语素义项变化的相关性分析。

第 7 章为总结与展望。总结了本文的研究内容、研究结果和主要贡献，并阐述了不足之处和未来工作的方向。

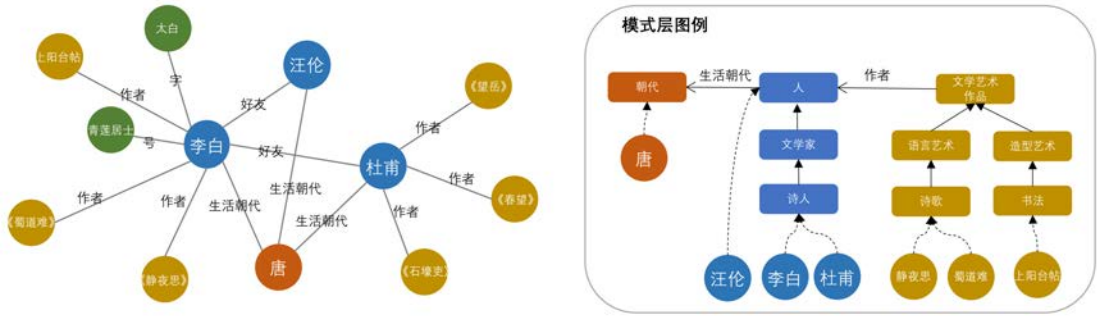
第 2 章　相关知识概述

本章将介绍与本文研究相关的知识图谱技术以及 NLP 和 CV 领域的算法。2.1 节介绍知识图谱的结构和构建技术、图注意力网络模型（Graph Attention Network，GAT），本文使用 GAT 对知识图谱中的实体节点进行表示。2.2 节介绍预训练语言模型，包括 Transformer 的基本架构、BERT 模型的预训练和使用范式，本文使用 BERT 来对古文字的文本数据进行表征。2.3 节介绍基于卷积神经网络（Convolutional Neural Network, CNN）和视觉变换器（Vision Transformer，ViT）的计算机视觉模型，本文使用 CV 模型对古文字的图像数据进行表征，此外，还介绍了融合文本和视觉特征的视觉语言模型 ViLT（Vision-and-Language Transformer），ViLT 对于本文的多模态融合方法具有指导作用。

2.1 知识图谱与图神经网络模型

2.1.1 知识图谱

知识图谱是结构化的语义知识库，用于描述物理世界中的概念及其相互关系。这一概念是 2012 年由谷歌首次提出的，目的是将互联网上的数据、信息及其链接关系聚集为知识网络，使这些资源更易于计算和处理，实现更智能的搜索引擎。知识图谱的基本单元是由“实体-关系-实体”或者“实体-属性-属性值”构成的三元组，实体指的是具有可区别性且独立存在的某种事物，例如 “唐朝”、“李白”、“《静夜思》”等。图 2.1 是一个知识图谱的示例。



（a）知识图谱数据层　　（b）知识图谱模式层

图 2.1 知识图谱案例

知识图谱在逻辑上可分为模式层与数据层两个层次，模式层也称为本体层。模式层构建在数据层之上，是知识图谱的概念模板，通常由本体库来进行管理。

例如“人-生活朝代-朝代”、“诗-作者-诗人”、“诗人-属于-人”，具体定义了知识图谱中的实体所属的概念，概念之间的关系和属性等。数据层是模式层的实例，主要是由物理世界的一系列的事实组成，例如三元组“《静夜思》-作者-李白”、“李白-生活朝代-唐朝”分别是模式层“文艺作品-作者-人”、“人-生活朝代-朝代”的实例。

知识图谱将现实世界中的知识表示为更接近人类认知的模式，提供了一种组织、管理和利用海量信息的方式。知识图谱的优势是使机器具备人类认知智能方面的推理能力和可解释性，结构化的知识也可以表示为高维嵌入向量的形式作为输入与深度学习的各类神经网络模型结合，目前已经被广泛应用于智能搜索、问答系统、个性化推荐、情报分析、反欺诈等多个领域。例如，在智能搜索中，系统会借助知识图谱对用户查询的关键词进行解析和推理，进而将其映射到知识图谱中的一个或一组概念之上，然后根据知识图谱的概念层次结构，向用户返回相关的信息，甚至直接推理出问题的答案。

知识图谱按照存储知识的类别可以分为通用知识图谱和垂直领域知识图谱两种，通用知识图谱存储通用的知识，通常来源于互联网和百科数据，知识覆盖范围广，实体种类丰富，目前比较著名的通用知识图谱包括 Yago[83]、Freebase[84]、Probase[85]、Dbpedia[86]、CN-DBpedia[87]等。领域知识图谱描述特定领域的知识，在金融、医学、电商、教育、法律等垂直领域都具有广泛的应用，它的实体类型相对较少，但与某个领域密切相关，且往往涉及到非常专业的领域知识。与通用知识图谱相比，领域知识图谱通常具有更深入和专业的模式层和更高质量的数据层，本文所构建的历时汉字知识图谱就属于领域知识图谱。

按照构建方法分类，知识图谱可被分为人工构建和自动构建两类。人工构建主要依赖于专家合作或者众包机制来建立，如非常知名的普林斯顿大学的英文语义词典 WordNet[60]，人工构建的知识图谱多为本体驱动，实体和关系具有较高的质量，缺点为由于认知局限和人工成本，图谱的规模较小。自动构建的知识图谱多从互联网上已经存在的数据资源进行整合，通常规模较大，但质量相对较低。知识图谱的构建和更新技术如图 2.2 所示，是一个迭代更新的过程：

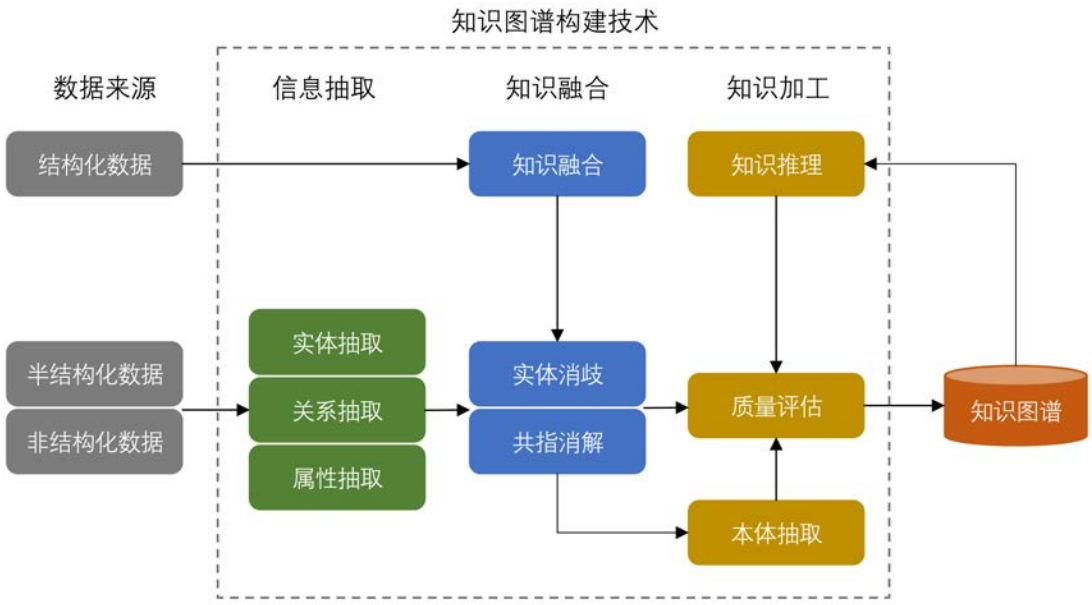


图 2.2 知识图谱的构建流程与构建技术

如图 2.2 所示，知识图谱的数据来源包括结构化数据、半结构化数据、非结构化数据三类，知识图谱的构建过程主要包含三个模块：信息抽取、知识融合、知识加工。信息抽取旨在从各种类型的数据源中提取出实体、属性以及实体间的关系，相关的 NLP 领域任务包括命名实体识别（Named Entity Recognition，NER）[88-91]、关系和属性抽取（Relation Extraction）[92-95]，主要是从自然语言文本数据中抽取中实体和关系属性。

知识融合旨在将从数据源中抽取到的实体对象链接到知识库中对应的正确实体对象的知识整合过程，对应的 NLP 任务为实体链接（Entity Linking）[97-100]，其中重要的步骤是实体消歧（Entity Disambiguation）和共指消解（Coreference Resolution），实体消歧针对某些多义词汇的称谓可能对应于多个不同的实体的情况，例如“苹果”这一词汇既可以指代一种水果，也可以指代苹果公司，实体消歧算法需要根据实体所处的上下文判断该词汇正确的含义，将该实体链接到正确的知识图谱中的实体上。共指消解用于解决多个指称对应同一实体对象的问题，共指消解技术将这些指称项合并到正确的实体对象。

知识加工包括本体构建（Ontology Construction）、知识推理和质量评估，是对融合的新知识进行质量评估的过程，本体构建是基于新知识对知识图谱模式层的进一步完善，可以采用人工构建或者自动构建技术实现，关键是对实体并列关

系和上下位关系抽取。知识推理，也可称为知识图谱补全（Knowledge Graph Completion）[101-105]，旨在通过已知的实体和关系，推理出知识图谱中缺失的关系，例如通过“《静夜思》-作者-李白”、“李白-生活朝代-唐朝”可以推理出“《静夜思》-写作朝代-唐朝”这一三元组知识。这种推理技术可以通过基于逻辑、基于图和基于深度学习的几种方法实现。质量评估是对知识的可信度进行量化，通过舍弃置信度较低的知识来保障知识库的质量。

2.1.2 图注意力网络模型

图神经网络（Graph Neural Networks, GNN）[106]是一类用于处理图结构数据的神经网络模型，如知识图谱，社交网络等，图是一种用顶点（Vertex）和边（Edge）表示实体及其关系的数学结构。一个图可以表示为 $G = (V, E)$，其中 V 是顶点集合，E 是边集合。与传统的神经网络，如卷积神经网络（Convolutional Neural Network，CNN）[107]、循环神经网络（Recurrent Neural Network，RNN)[108]等处理图像、文本等结构化的网格或序列数据不同，图神经网络需要考虑节点之间的关系，捕捉节点之间的关联性。

近年来出现了许多图神经网络的变种，例如图卷积网络（Graph Convolutional Network, GCN）、图注意力网络（Graph Attention Network, GAT）、图自编码器（Graph Autoencoders）、图生成网络（ Graph Generative Networks） 和图时空网络（Graph Spatial-temporal Networks）等以适应不同的任务和数据类型。GCN[109]是图神经网络模型的基础，它的思想是每个节点接受来自邻居节点的信号，将自身特征与邻居特征进行聚合，生成一个新的节点特征表示，这一过程可以迭代多次，最终形成更全面的图结构表示。GAT[110]基于注意力机制聚合邻居节点的特征信息，将注意机制用于确定节点邻域的权重，解决了 GCN 无法为邻居节点分配权重的问题，并且 GAT 是一种局部网络，无需了解整个图结构，只需知道每个节点的邻节点即可。

GAT 基于注意力机制更新节点的嵌入表示，注意力层的输入为:

$$h = \{h_1, \dots h_n\}, h_i \in \mathbb{R}^D \tag{2.1}$$

其中n表示图中的节点个数，h_i 表示节点当前的特征向量，D表示特征向量的维度，经过注意力层的处理后，输出为：

$$h' = \{h_1{}' \dots h_n{}'\}, h_i{}' \in \mathbb{R}^{D'} \tag{2.2}$$

其中$h_i{}'$为节点经过注意力层计算后的特征向量，D'为输出的节点特征向量维度。

$h_i{}'$的计算公式如下，其中N_i是节点i的所有邻居节点的集合，α_{ij}是注意力系数，表示邻居节点j对于节点i的重要程度，$\sigma(\cdot)$代表任意激活函数，W是网络参数矩阵，由模型经过训练获得，$W \in \mathbb{R}^{D \times D'}$。

$$h_i{}' = \sigma(\sum_{j \in N_i} \alpha_{ij} W h_j) \tag{2.3}$$

注意力系数的计算公式如下：

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a^T[Wh_i || Wh_j]))}{\sum_{k \in N^i} \exp(\text{LeakyReLU}(a^T[Wh_i || Wh_k]))} \tag{2.4}$$

这里 $a \in \mathbb{R}^{2D'}$，是网络要训练的参数，||表示矩阵的拼接操作，LeakyReLU是一种激活函数，公式如下，其中δ是负值斜率，使负值不等于 0。

$$y = \begin{cases} x, & x \geq 0 \\ \delta x, & x < 0 \end{cases} \tag{2.5}$$

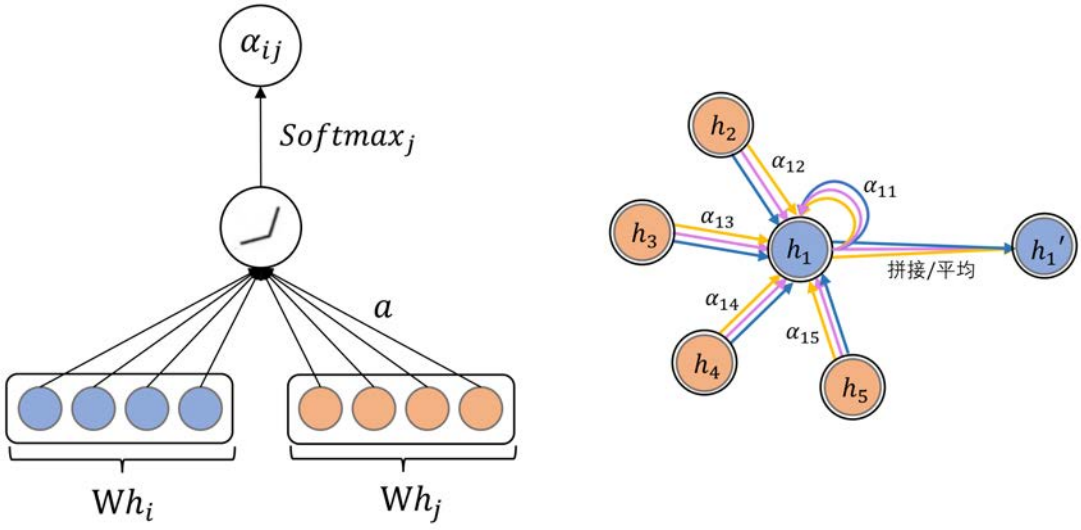
从以上公式可以看出，GAT 利用利用节点i和邻居节点j的特征作为输入，计算出j对于i的重要性权重α_{ij}，然后计算所有临居节点特征的加权平均，从而对节点i的表示进行更新。为了使这一过程的表示更加全面，GAT 采用了多头注意力机制，即独立的计算K个α，然后将其获得的特征拼接起来，公式如下：

$$h_i{}' = ||_{k=1}^{K} \sigma(\sum_{j \in N_i} \alpha_{ij}^k W^k h_j) \tag{2.6}$$

在网络的最后一层，继续采用拼接的方式会扩大特征维度，因此替换为求平均的操作：

$$h_i{}' = \sigma(\frac{1}{K} \sum_{k=1}^{K} \sum_{j \in N_i} \alpha_{ij}^k W^k h_j) \tag{2.7}$$

图 2.3 展示了 GAT 网络注意力系数的计算和节点特征向量的更新过程，其中三种颜色的线代表三个注意力头：



(a) 注意力系数α的计算过程　　(b) 节点向量的更新过程

图 2.3 图注意力网络模型注意力层计算过程

2.2 预训练语言模型

2.2.1 Transformer 模型

神经网络技术是 NLP 任务中普遍使用的 AI 算法，这些网络接收高维的文本特征向量作为输入，因此，文本表征经常作为诸多 NLP 任务中的首要工作，将词汇表示为稠密的词嵌入（Word Embedding）。词嵌入可以分为静态与动态两种，在静态词嵌入中，每个单词仅由一个静态向量表示，这种表示是上下文无关的，代表性的模型包括 Word2Vec[61] 和 GloVe[111]。而动态词嵌入模型可以随着上下文的不同为词汇生成不同的嵌入表示，能够应对一词多义的语境变化，捕获上下文中的句法结构等，因此得到了更广泛的应用。

动态词嵌入主要由预训练语言模型生成，比如 BERT[39]、T5[112] 以及 GPT[113]。这些预训练语言模型都是以 Transformer 的结构为基础设计的，如图 2.4 所示，Transformers 可以分为编码器（Encoder）和解码器（Decoder）两部分，编码器对输入的文本序列中的每一个词汇生成嵌入表示，然后这些词嵌入被送入多个自注意力层（Self-Attention），利用自注意力的原理融合上下文其他词嵌入的特征，对词嵌入进行更新，使预训练模型可以生成上下文相关的词嵌入，最后连接一层前馈网络，输出最终词汇的嵌入表示。解码器负责对编码器输出的嵌入进行解码，生成模型最终的预测结果。

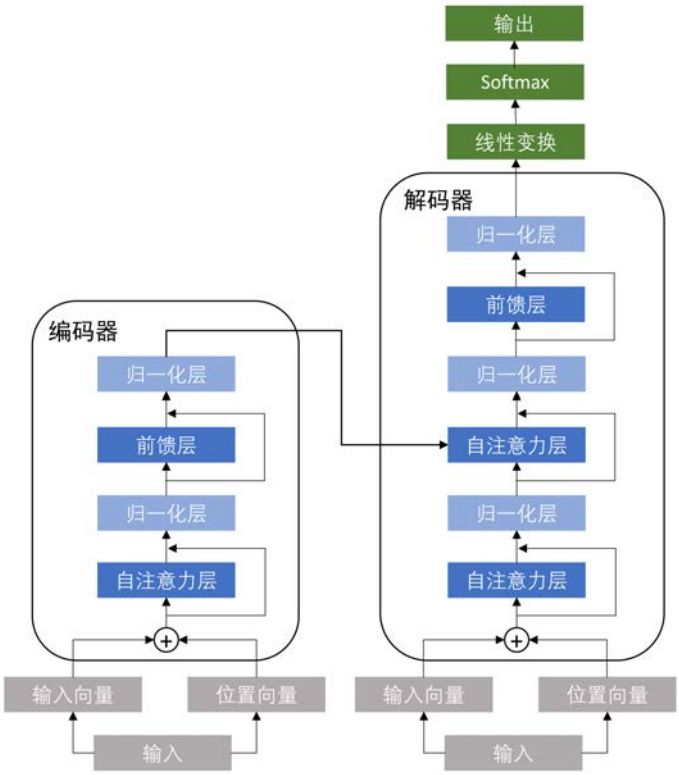


图 2.4 Transformer 结构图

预训练语言模型按照使用编码器和解码器的不同方式，可以分为自编码式（Auto-Encoder）、自回归式（Auto-Regressive）和混合模型。自编码式模型，例如 BERT[39]、RoBERTa[114]、ELECTRA[115]、DeBERTa[116] 等使用 Transformer 的编码器部分，通常以预测双向上下文输入序列中的掩盖字段为目标，因此善于解决 NLP 中的分类任务，比如文本分类、词义消歧、情感分析等。自回归模型使用解码器，通常从单向的序列中学习，将上一个时间步的值作为输入，以预测下一个时间步的值，因此被广泛用于自然语言生成任务，比如摘要、翻译、聊天机器人等，代表模型有 GPT[113]。混合模型使用编码器和解码器，通常被用于需要内容理解和生成的任务，比如机器翻译，代表模型有 T5[112]、BART[117]、BigBird[118]等。本文的任务使用 BERT 实现，下文将具体介绍 Transformer 使用的自注意力机制（Self-Attention）以及 BERT 模型的训练和应用模式。

2.2.2 自注意力机制

自注意力机制（Self-Attention）是 Transformer 最重要的部分，它的输入包括三个向量序列：查询（Query）、键（Key）和值（Value），它们是通过对输入

序列进行不同的线性变换获得的。自注意力机制通过将查询与键进行交互来计算注意力权重，然后将注意力权重应用于值向量，最后对加权后的值向量进行求和。这种交互能够捕捉到输入序列中每个位置与其他位置的关系，并为每个位置生成一个更新的表示。自注意力机制的公式如下，其中X是输入的词嵌入序列的矩阵，X'是经过自注意力机制更新后的词嵌入矩阵，W是网络通过训练所得到的参数矩阵，Q，K，V分别代表查询矩阵、键矩阵和值矩阵，d_k是键矩阵每一列特征的维度，最后利用$softmax$函数对结果进行归一化。

$$Q = W^Q X, \qquad K = W^K X, \qquad V = W^V X \tag{2.8}$$

$$X' = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.9}$$

公式对应的注意力机制的计算过程如图 2.5 所示：

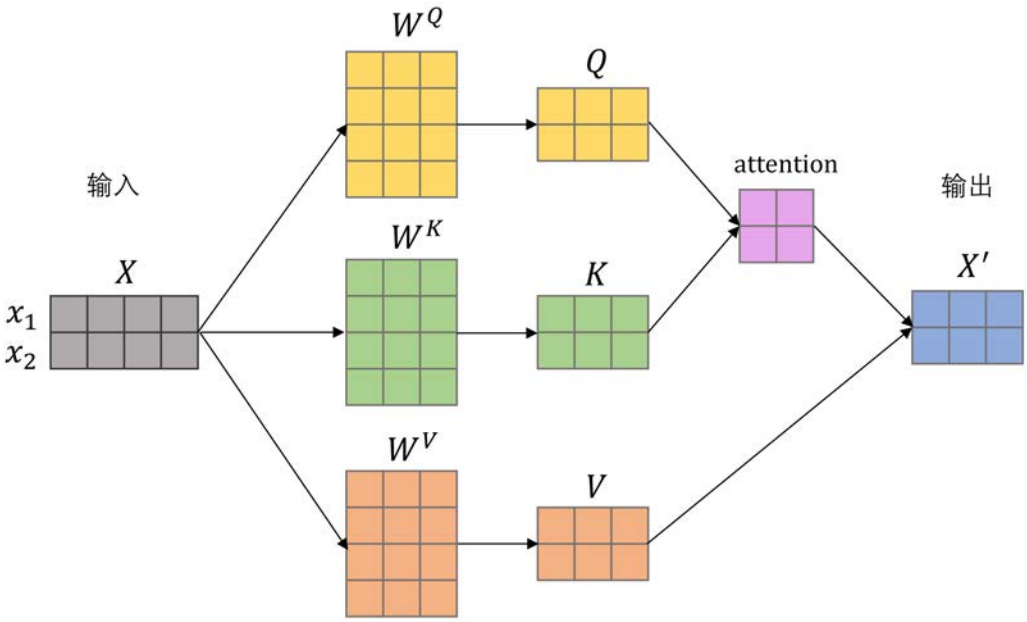


图 2.5 自注意力机制的计算过程

为实现更丰富的特征表达能力，当前的各种网络模型一般采用多头注意力机制（Multi Head-Attention），多头注意力机制将输入序列进行多次线性变换，得到多组查询、键和值向量，将多个头的输出进行拼接，并通过线性变换获得最终的注意力表示。

2.2.3 BERT 模型的训练与应用模式

BERT 模型的使用方法分为预训练-微调两个步骤，在预训练步骤中，BERT 在海量的文本上下文中进行训练，这一过程不针对特定的下游任务，目的是使 BERT 学习到自然语言的通用“语感”，BERT 最经典的两种预训练学习任务为掩码语言模型（Mask Language Model，MLM），和下一句子预测（Next Sentence Prediction，NSP），在 MLM 任务中，BERT 会对输入的文本进行随机的掩码处理，一般约 15%的词会被掩码成特定的标记“[MASK]”，另外一小部分词会被替换成随机的词，而其余的词保持不变，任务的目标是通过上下文中的其他词来预测掩码词。NSP 任务接收一对句子作为输入，任务是判断两个句子是否连续，旨在训练模型对句子级别的理解能力。模型的预训练过程如图 2.6 所示：

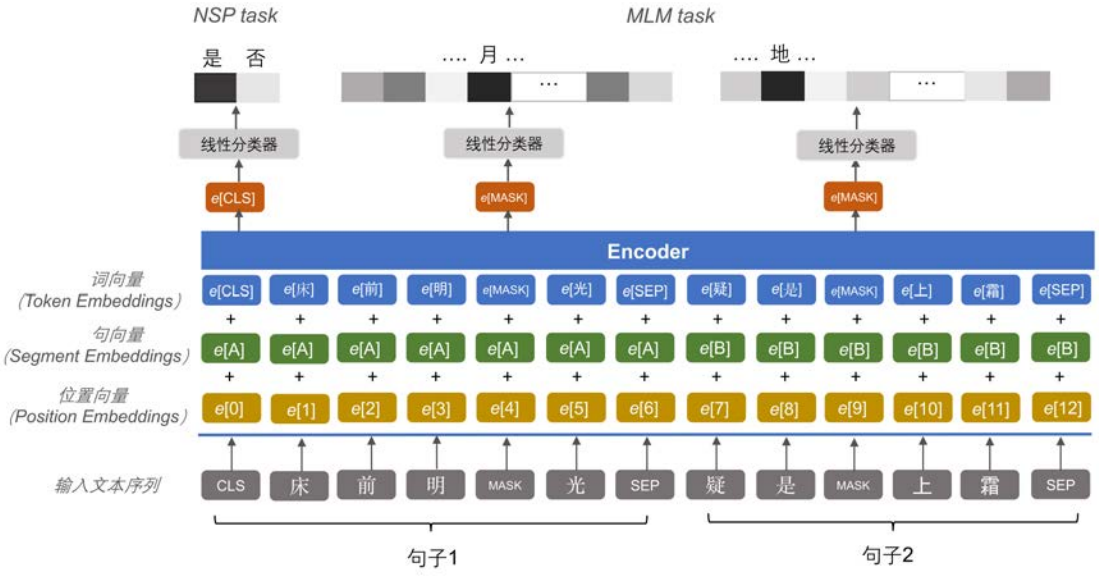


图 2.6 BERT 模型的预训练过程图例

如图 2.6 所示，在将文本输入到模型前，文本的字符（Token）序列需要被处理成包含首尾特殊符号和掩码的形式，通常句子的起始符号为“[CLS]”，结束符号为“[SEP]”，掩码符号为“[MASK]”，例如“床前明月光”将被处理为“[CLS]床前明[MASK]光[SEP]”，其中“月”字被随机替换为掩码标记，需要模型进行预测。在 BERT 的输入层，字符序列被转化为词嵌入的序列，其中每一个字符被表示为一个词嵌入向量。BERT 的词嵌入由 3 个子向量求和得到，分别是词向量（Token Embeddings），句子向量（Segment Embeddings），和位置向量（Position Embeddings）。词嵌入被初始化后，被输入 Encoder 层中进行进一步编码， BERT

的 Encoder 层是由上一节所述的多个相同结构的 Transformer 编码器组成，最终输出每个字符的嵌入表示。在 MLM 任务中，掩码字符所对应的输出经过线性变换和归一化函数，最终得到词汇表中所有词汇的概率分数，而 NSP 任务使用[CLS]起始符号所对应的输出向量，同样将其映射为二分类的概率分数，最终得到句子对是否连续的分类结果。

在预训练步骤结束后，为了将 BERT 模型应用于 NLP 的各类下游任务中，通常使用监督学习的方式，在下游任务人工标注的相对较小规模的数据集中对模型进行微调，可以调整模型的参数或者改变模型的架构等。通过使用不同部分的输出信息，BERT 可以适用于多种 NLP 任务，总结为 4 个经典的任务模式：（1）句子对匹配任务，主要应用于判断两个文本是否连续、相似，或者是否属于同一个主题，通常利用[CLS]位置对应的输出向量；（2）文本分类任务，应用于单个文本的主题分类或者情感分类，通常利用[CLS]位置对应的输出向量；（3）问答系统，每个字符对应一个输出，表示该字是答案的起始或结束位置的概率；（4）分词、命名实体识别等序列标注任务，每个字符对应一个输出，表示该位置词汇对应的序列标签概率。BERT 输出的向量经过线性变换和归一化函数即可映射为标签的概率，也可以作为其他神经网络模型的输入进行进一步的处理。

2.3 计算机视觉模型

2.3.1 卷积神经网络模型

卷积神经网络（Convolutional Neural Network，CNN）是计算机视觉领域的基础模型。CNN 分为输入层、卷积层（Convolution）、池化层（Pool）、全连接层和输出层。输入层将图像转换为像素值对应的矩阵，具有宽（Width）、高（High）和通道数（Channel）三个维度。卷积层通过多个卷积核（Convolution Kernel）提取图像的特征，输出与卷积核数量相同的特征图矩阵（Feature Map），每个特征图对应一个卷积核提取的特征。池化层又称为下采样层，旨在提取最具有代表性的特征，通常采用最大化或平均池化，降低特征图的维度，防止过拟合，消除平移的影响。在经过多次卷积和池化操作后，通过全连接层展开为一维的向量，汇总得到的图像的特征图信息，最终在输出层根据不同的任务计算概率。图 2.7 为一个单通道输入图像矩阵、单卷积核的计算过程图例。

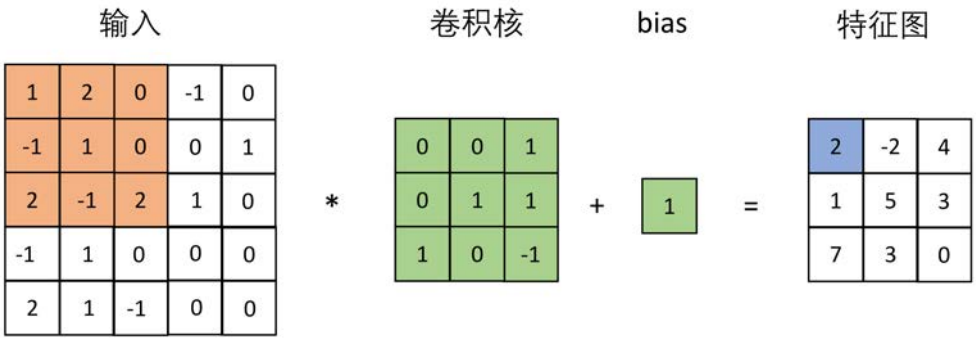


图 2.7 卷积计算过程图例

1998 年，CNN 之父 Lecun 提出了第一个 CNN 网络 LeNet-5 [119]，用于解决手写数字识别任务，包含两个卷积层，两个池化层和全连接层。2012 年，Alex Krizhevsky 等人提出的 AlexNet[120]网络包含 5 个卷积层，2 个全连接层，以及 1 个全连接输出层，与 LeNet-5 相比，大幅提升了网络参数的规模，在 ImageNet 2012 图像识别挑战赛上取得了突破性的成功，从此 CNN 成为了主流的计算机视觉模型。VGG-Nets[121]是 2014 年由牛津大学 VGG（Visual Geometry Group）提出的深度 CNN 模型，可以看作是加深版本的 AlexNet，包含十几个网络层。同年 GoogLeNet[122]模型，创新点是提出了 Inception 模块，在控制了计算量的参数量的同时，获得了非常好的性能。2015 年，何恺明提出的 ResNet[123]在网络通过引入残差模块来解决深度神经网络隐藏层过深时引起的退化问题，残差模块的基础结构如图 2.8 所示，这是 CNN 网络发展上的巨大创新，通过这一创新，ResNet 可以实现 152 层的深层网络，获得比 VGG 和 GoogLeNet 更高的准确率。

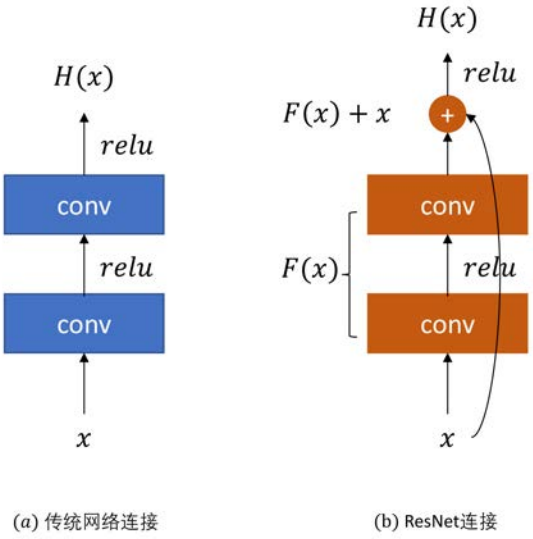


图 2.8 ResNet 网络残差模块基本结构

2017 年，康奈尔大学、清华大学、Facebook FAIR 实验室联合提出 DenseNet[124]，借鉴了 ResNet 的残差结构，相比 ResNet，DenseNet 网络中的每一层都直接与前面所有层相连，实现特征重用，强化特征传播，大幅度减少参数数量。

2.3.2 Vision Transformer 模型

Vision Transformer（ViT）[125]是 2020 年谷歌提出的将 NLP 领域的 Transformer（2.2.1 节）应用于图像分类的模型，它证明了 Transformer 模型处理图像任务的能力，启发了后续的相关研究。相对于 CNN 模型，ViT 更适合拥有大规模图像数据集的预训练和微调场景。ViT 模型的基本结构如图 2.9 所示，可以看出 ViT 采用了与 Transformer 处理文本序列类似的思想，每一个图像块相当于文本序列中的一个字符。

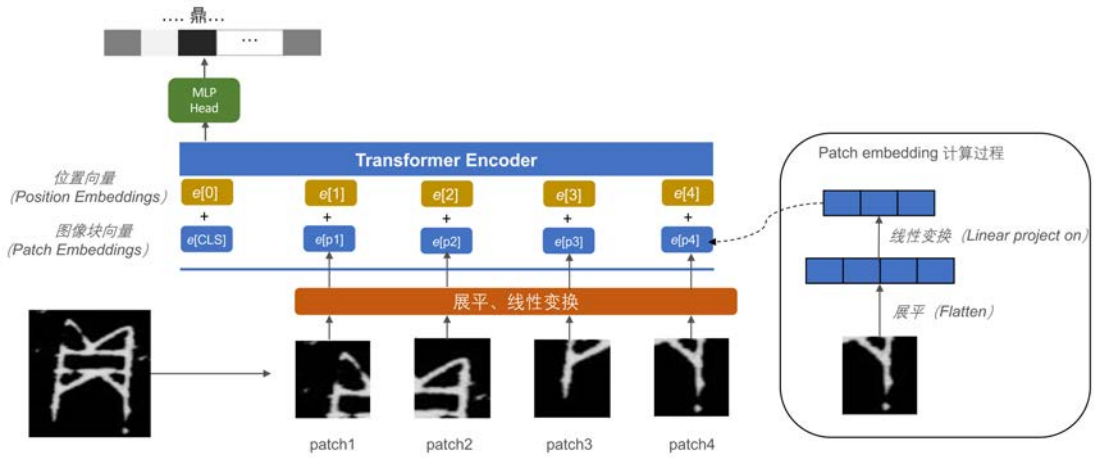


图 2.9 ViT 模型结构

ViT 模型将完整的图片切分成多个图像块（Patch），经过展平和线性变换后将图像块的三维矩阵拉伸为一个嵌入（Patch Embedding），在图像块序列的头部添加一个起始标记[CLS]，对应一个嵌入。然后将所有的 Patch Embedding 与位置嵌入（Position Embedding）相加，得到最终的输入向量，传入 Transformer Encoder 模块，最后将[CLS]对应的输出向量传入多层感知器（Multi-Layer Perceptron，MLP Head）模块，使用非线性激活函数最终得到分类标签的概率。

2.3.3 多模态视觉语言模型

在过去十年，单模态学习在 CV、NLP 领域的一系列任务上取得了先进的成果，但现实世界的问题往往是多模态的，近几年诞生了视觉-语言预训练 (Vision-and-Language Pre-training, VLP) 方法，通常基于图像文本匹配 (Image Text Matching，ITM) 和掩码语言建模 (Masked Language Modeling，MLM) 任务进行训练，并在各自的下游任务上面进行微调。VLP 模型中主要有三大部分，即视觉编码器（VE）、文本编码器（TE）和模态交互（MI）模块。VE 和 TE 模块分别对图像和文本的特征进行编码，而 MI 则将 VE 和 TE 提取的特征进行融合。

Wonjae Kim 等人[126]将视觉语言模型按照图像和文本两个模态的编码器是否具有相似的参数规模，以及两个模态的编码结果是否通过深度模型进行交互分为四类。第 1 类：视觉编码器占大部分算力，用简单的点积或浅注意层来对两种模态进行特征交互，代表模型是 SCAN[127]。第 2 类：视觉与文本模型都使用较大算力，用简单的点积或浅注意层来进行特征交互，代表模型是 CLIP[128]。第 3 类：视觉模型占大部分算力，使用 Transformer 对图像和文本特征的交互进行深层建模，代表模型是 FiLM[96]。第 4 类：视觉和文本都采用简单的编码方式，大部分计算力集中在模态的交互上面，使用 Transformer 对图像和文本特征的交互进行建模，代表模型是 ViLT[126]，模型结构如图 2.10 所示：

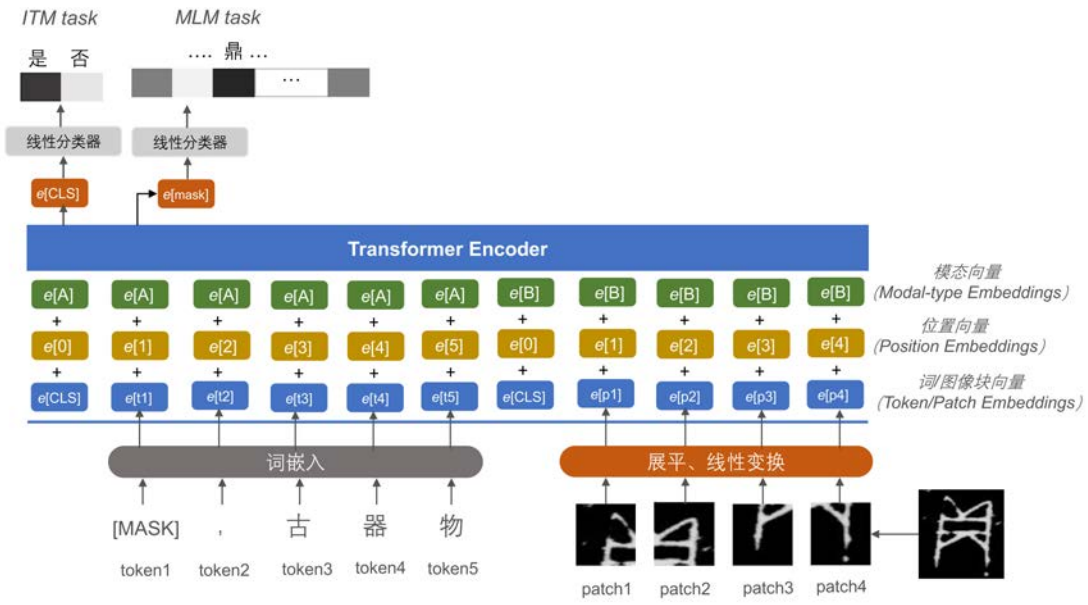


图 2.10 ViLT 视觉语言模型基本结构

ViLT 的 Transformer Encoder 以 ViT-B/32 的预训练权重做初始化，采用单流方法，即先对图像和文本的表征拼接，再输入 Transformer 的方式，基于 ITM 和 MLM 任务做预训练。不同于以往的 VLP 模型使用 ResNet 等复杂的视觉编码器，ViLT 受到 ViT 使用 Patch Embedding 编码图像特征的启发，使用极简的线性变换来抽取视觉特征，主要的计算量为基于 Transformer 的特征融合，极大地减小了模型的参数规模和运行时间，并可以取得相似或更好的性能。

2.4 本章小结

本章主要介绍了与本工作相关的算法和技术。首先对知识图谱的组成和构建技术进行概述，并介绍了可以对知识图结构进行表征的 GAT 模型。然后介绍 NLP 领域的 Transformer 架构的组成、自注意力机制，以及 BERT 模型的“预训练-微调”使用范式和训练方法。最后介绍了 CV 领域的 CNN 模型、ViT 模型，以及多模态的视觉语言模型，下一章将介绍 ZiNet 汉字知识图谱的构建方法。

第 3 章　历时汉字知识图谱构建方法

古文字的多模态数据，例如拓本与摹本图像、出土文献文本、字典词典、研究著录等分散在各类平台和专业著录中，这些碎片化的知识缺乏结构化、数字化的整合，历史上字形、字义与字音相似的古文字以及相关考古数据难以被互联查询和分析。针对古文字知识整合与关联问题，本文设计并构建了一个历时汉字知识图谱，称为 ZiNet，旨在对汉字的字形、字义与字音知识进行描述、抽取和链接，形成一张由各历史时期的字形、文字、部件、词汇、义项等实体节点和语义关系构成的汉字知识网络。此外，由于古文字领域多模态数据的处理和标注过程严重依赖于专家经验，需要耗费大量的时间和精力，本文研究半自动化的知识图谱构建方法，基于 NLP 与 CV 的相关技术，实现拓本图像去噪、部件识别、字体库构建、通假字标注等数据处理和知识提取的自动化方法，并与领域专家协作进行标注，在确保专业性的同时提升知识图谱的构建效率。本章 3.1 节为引言，3.2 节将介绍 ZiNet 知识图谱的架构及本体模式，3.3 节介绍 ZiNet 的构建过程，包括与古文字领域专家的协作方式以及采用的自动化的算法。3.4 节介绍 ZiNet 目前的统计信息，3.5 节介绍基于 ZiNet 知识图谱的潜在应用和落地平台。最后 3.6 节总结该章。

3.1 引言

目前古文字的字形、字义等领域知识以及描述它们的多模态数据，包括古文字拓本及摹本图像、出土文献释文、字典著作等的数字化整合尚不完善，相关知识难以被领域专家高效地获取和关联查询，专家往往需要依赖于经验和记忆力多方进行查阅，并将相关的文字联系在一起。尽管近年来古文字的数字化工作已经取得很大进展，将古籍史料转换为可供计算机检索和操作的数字化图像与文本资源，但是有关古文字知识库、知识图谱等结构化资源的构建研究仍然很少。与本文最相关的工作是甲骨文知识图谱[1]，他们总结了甲骨文的基础数据，包括图像、三维建模、DNA 鉴定、甲骨碎片、甲骨文释文等，并整合这些数据，以及相对应的字形、文物馆藏、研究学者、文献等信息，形成一张多模态知识网络。

本文认为古文字的知识表示与整合还存在以下挑战：（1）从古文字形义互联的视角来看，当前的工作关注于文物、图像、文本数据等有形实体之间的互联，而对于如何描述古文字的字义、字形、字音等抽象化的知识概念，并以此来链接

相关文字的问题并未得到完善解决，而音形义相似文字的发现与分析对于古文字研究非常重要。（2）从古文字的古今互联的视角来看，已有的古文字知识图谱仅整合特定历史时期的古文字，如甲骨文，需要提供一个链接中国历史上所有的文字的模型框架，才能够描述汉字的演变过程。（3）从知识图谱构建的视角来看，为保证古文字数字化的专业性和正确性，领域专家需要对出土的原始数据进行修复、识别、标注等大量工作，很多相关工作基于 CV 算法对古文字图像数据进行降噪[13-14]与识别[15-23]，在 NLP 领域，SikuBERT[34]、GuwenBERT[35]、CANCN-BERT[36]等古汉语预训练模型已经被应用于辞例归纳，预测历史文献中被遮蔽起来的文字[40]，如果在知识图谱构建的过程中广泛应用这些技术，可以极大程度提升古文字知识图谱的构建效率。

针对上述问题，本文研究历时汉字知识图谱及其构建方法，本章的贡献是提出了 ZiNet，这是第一个通过字形和语义知识关联中国各个历史时期汉字及其多模态数据的知识图谱。ZiNet 定义了出土文物、文字拓本、字形、部件、古文字、文字、词汇、义项、同义词集九种概念，并为每个概念定义属性以及概念之间的语义关系，从而形成一张汉字知识网络。与已有工作不同，ZiNet 专注于对字形、字音、字义的抽象知识进行定义和描述，并根据这些抽象概念来关联古文字及相关数据，并且 ZiNet 是历时性的，它通过关联不同历史时期的汉字字形来描述汉字的演变，每种古文字类型的实体都相对独立地进行存储，因此可以方便地进行古文字类别的扩展和更新。此外，ZiNet 作为为古文字研究服务的领域知识图谱，对专业性和正确性要求很高，因此需要领域专家参与数据标注与知识图谱构建。为提升知识图谱的构建与更新效率，本文基于 NLP 与 CV 算法，实现了对拓本图像去噪、古文字部件识别、字体库构建、通假字标注的自动化方法，并探索了与领域专家协作的模式。

目前，ZiNet 已经包含来自甲骨文、金文、楚简的 12620 个古文字字形，6792 个已释字，2787 个未释读文字，52498 张拓本图像，以及 68901 条出土文献。ZiNet 在古文字研究领域具有很大的应用价值，不仅可以支撑智能化的检索平台，帮助领域专家高效地检索历史上相关联的文字和多模态信息，也可以与当前的 AI 模型相结合，改进当前的 AI 驱动的古文字算法的效果，例如古文字图像识别等，还可以捕捉在字形、字义、文字起源等方面可能相关的文字，对未知古文字

进行推理和预测，辅助领域专家进行未知古文字的考释工作。未来，随着大模型的兴起，ZiNet 也可以服务于大模型的训练与应用，为古文字研究作出更多贡献。

3.2 历时汉字知识图谱架构与本体模式

3.2.1 ZiNet 的概念和属性定义

本体论起源于哲学，后来在计算机科学中用于描述概念实体及其关系，古文字本体可以为知识图谱的构建提供清晰的结构和标准化的语义。本文与古文字专家合作，在 ZiNet 中定义了 9 种概念，包括出土文物、文字拓本、字形、部件、古文字、文字、词汇、义项、同义词集，每种概念具体的定义、实体案例和主要的数据属性见表 3.1。

表 3.1 ZiNet 的 9 种概念及其定义、实体案例和数据属性

概念	定义	实体案例	数据属性
出土文物 Unearthed relic	古文字的文物载体，如龟甲、兽骨、青铜器、简帛等。	伯武父鼎 甲骨合集 00006 号	文物名称、文物照片、拓本图像、文物类别、收录文献、出土地、收藏地、历史时期、出土文献释文等
文字拓本 Character rubbing	单个汉字的拓本图像		文字名称、拓本图像、摹本图像、历史时期
字形 Glyph	汉字的书写形态，本文将同一个文字部件组成相同，视觉形态相似的字体归纳为同一字形	“宿”的甲骨文具有两个字形: 1. 字形“侴”  2. 由部件“宀，侴”组成的字形 	文字名称、隶定字体、代表图像、部件描述

续表

概念	定义	实例	数据属性
部件 Radical	构成文字的最小结构单位	宀 佰 人 丙	部件名称、隶定字体、代表图像
古文字 Ancient character	特定类型的文字，由其对应的汉字和文字类型共同标识	宿_甲骨 宿_金文 宿_战国楚简 宿_小篆 宿_隶书	文字名称、文字类型、考释状态
文字 Character	文字实体具有两种类型：已释字和未释字，已释字对应现代汉字或者曾经使用的汉字，未释字是一个抽象的文字符号，没有对应的现代汉字	现代汉字“宿”	对应汉字、考释状态
词汇 Word	词汇是语言中最小的可以独立运用的单位，单音节词汇由单个文字组成，多音节词汇由多个文字组成	宿 归宿 宿舍	词汇名称
义项 Sense	词汇的含义	过夜，住宿 住宿站	定义、例句、词性、收录文献、历史时期
同义词集 Synset	表示同一概念的义项集合	Synset{过夜，住宿:宿、歇、顿、舍、住宿...}	定义

实体是概念的实例，例如“出土文物”概念的一个实例为“伯武父鼎”，ZiNet中的每一个实体都具有唯一的编码 ID 进行索引。数据属性连接实体和数据，数

据可以为文本（String）、图像（Image）、整数（Integer）等多种类型，用以描述实体的属性。例如，出土文物概念具有“文物名称”、“文物类别”、“ 历史时期”、“拓本图像”等属性，分别关联到描述这些属性的相应的文本和图像数据：“伯武父鼎”、“青铜器”、“西周中期”、“  ”。

3.2.2 ZiNet 的关系定义

本文为概念定义了 15 种主要关系，如表 3.2 所示：

表 3.2 ZiNet 中的 15 种主要关系

关系	键	值	三元组案例
包含 Contain	字形	部件	宿(宀佰) – 包含 – 宀
属于 Belong to	字形	古文字	宿(宀佰) – 属于 – 宿_甲骨
相似 Similar to	字形	字形	宿(宀佰) – 相似 – 寐（宀人爿口）
含义相近 Synonymy	部件	部件	金 – 含义相近 – 貝
读音相近 Homophone	部件	部件	父 – 读音相近 – 甫
包含部件 Contain radical	部件	部件	佰 – 包含部件 – 亻
对应文字 Correspond to	部件	文字	亻 – 对应文字 – 人
描述 Describe	文字拓本	字形	 – 描述 – 宿(宀佰)
来源于 Derived from	文字拓本	出土文物	 – 来源 – 伯武父鼎
演化为 Evolve into	古文字	文字	宿_甲骨 – 演化为 – 宿
通假 Tongjia	古文字	文字	宿_金文– 通假 – 肃
构成 Compose	文字	词汇	宿 – 构成 – 住宿
拥有含义 Process	词汇	义项	宿 – 拥有含义 –古代的住宿站
收录于 Recorded in	义项	同义词集	古代的住宿站– 收录于 – Synset{住宿站}
上位概念 Hypernymy	义项	义项	Synset{住宿站} – 上位概念 – Synset{建筑物}

关系通常是有方向的，连接两个概念，描述概念之间的语义关系。例如（字形-包含-部件）描述字形与部件之间存在包含关系。概念可以被实例化为实体，

这样就构成了知识图谱的最小语义单元，即“实体 a-关系-实体 b”三元组，例如“宿-包含部件-宀”，古文字知识图谱由三元组构成。通过这些关系，ZiNet 可以对字形、字音、字义相似或相关的文字进行互联和计算，并关联与之对应的图像、历史文献等多模态数据。

3.2.3 ZiNet 的架构

根据上文定义的概念、属性和关系，本文以“宿”字为例，对 ZiNet 的架构和优势进行介绍，图 3.1 展示“宿”字在知识图谱中的上下文。

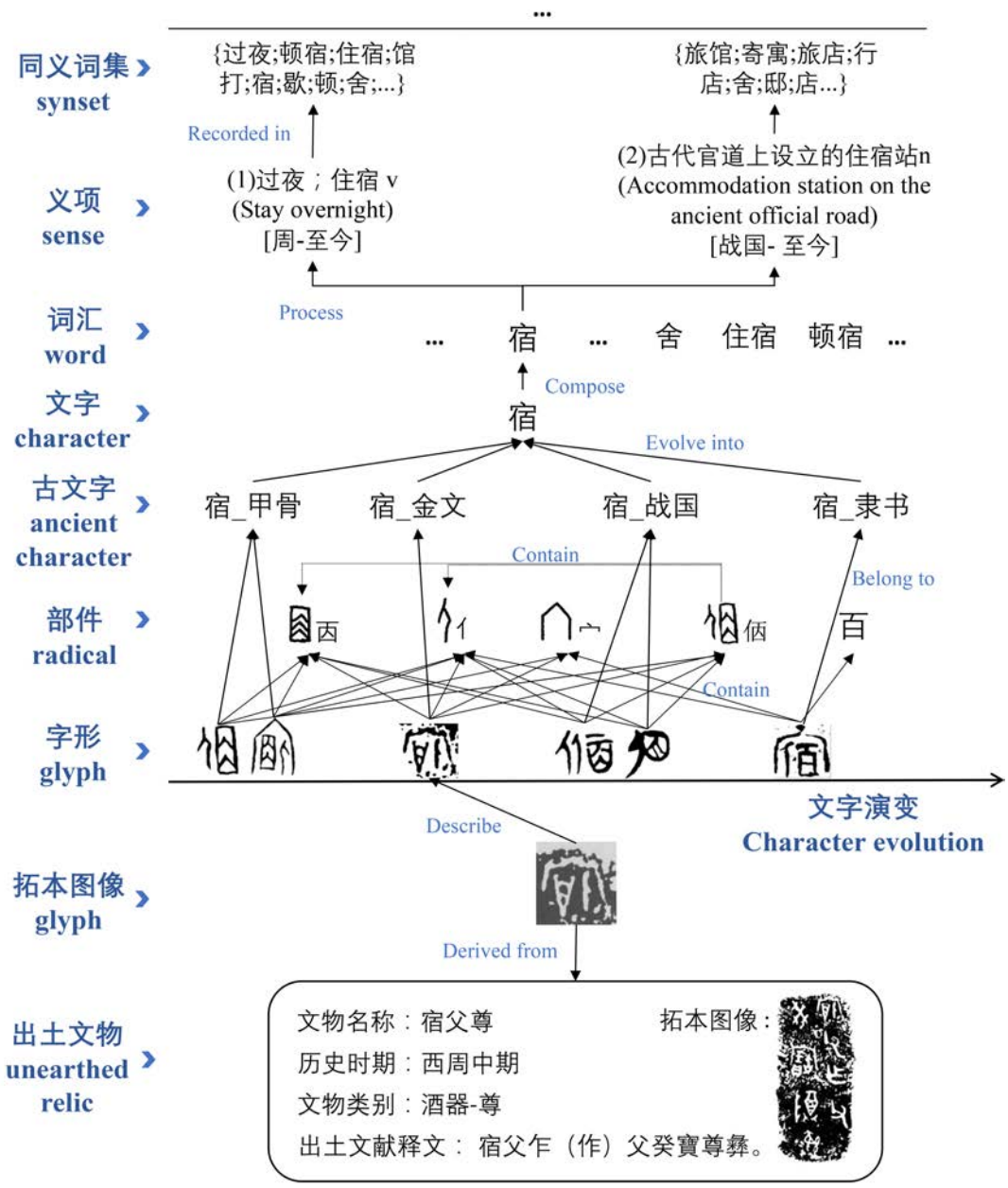


图 3.1 历时汉字知识图谱（ZiNet）架构图

可以看出，本文定义的关系将来自不同历史时期的文物、拓本、字形、部件、古文字、文字、词汇、含义、同义词集等实体根据字形、字音与字义彼此互联。“宿”字在甲骨文、战国楚简中分别有两个字形，分别由“宀佰”和“佰”部件构成，其部件组成不同，在金文和隶书时期有一个字形实体，ZiNet 将这些字形关联到对应的部件和古文字实体中，用以描述该字形的部件组成和文字类别，而后，四个古文字实体“宿_甲骨”、“宿_金文”、“宿_战国”和“宿_隶书”被关联到同一个文字实体“宿”中，表示它们属于同一汉字的不同演化阶段。同时部件之间定义了语义关系，如图中“佰”通过包含关系与“亻”、“ 百”互联。为描述字义知识，本文参考 WordNet 词汇知识库的经典结构，将文字实体进一步链接到构成的词汇中，词汇与义项相关联，义项最终被链接到同义词集中，以同义词集为枢纽将含义相同或相近的文字和词汇互联起来。ZiNet 还整合了出土文物和文字拓本的数据，每一条拓本图像都被链接到对应的字形实体和出土文物来源中，便于领域专家对字形的拓本图像、出土文物、以及文献上下文进行查询。

ZiNet 知识图谱的本体架构具有层次性，有以下特点和优势：

一．ZiNet 不仅专注于出土文物、文字拓本等有形数据之间的互联，还专注于对字形、字音、字义等抽象概念进行定义和描述，并根据这些抽象概念来关联古文字，具体表现在：

（1） 在字形层面，本文定义“字形”和“部件”概念，将部件组成相同，视觉形态相似的文字书写形态归纳为同一个字形，并以部件为纽带将字形互联，使包含同一个部件或者部件之间具有语义关系的字形具有更近距离。

（2） 在语义层面，本文设置了“义项”和“同义词集”概念，以同义词集为纽带，能够将含义相同或相似的词汇义项互联起来，此外部件也能够反应文字的语义，本文在部件层面设置“含义相近”的语义关系，通过该关系能够链接到近义部件及其构成的字形。

（3） 在语音层面，因为汉字的读音与声符部件相同或相近，本文在部件之间定义了“读音相近”的关系，通过该关系能够链接到同音部件及其构成的字形。同时本文在古文字与文字实体之间设置了“通假”关系，根

据这些关系可以检索到读音相同或相近的文字。

二．ZiNet 可以整合来自不同历史时期不同类型的古文字，因此可以描述汉字的历时演变过程。具体表现在：

（1） ZiNet 在对古文字概念的定义中区分了特定的文字种类，例如在图 3.1 中，“宿_甲骨”、“宿_金文” 、“宿_战国” 、“宿_隶书”分别表示 4 个不同历史时期的文字类型，关联到对应类型的字形、拓本图像和出土文物等实体，并通过将不同类型的古文字实体关联到同一个文字实体实现汉字的跨时代互联。

（2） ZiNet 对每种类型的古文字实体以及与它关联的字形、拓本图像、出土文物实体都相对独立地进行存储，因此可以方便地进行不同阶段古文字类别的扩展和更新。

（3） ZiNet 对出土文物、拓本图像、义项实体定义与时间相关的属性，可以检索该实体存在的历史时期。

3.3 历时汉字知识图谱的构建方法

3.3.1 历时汉字知识图谱构建流程

ZiNet 的构建流程框架如图 3.2 所示，它是与古文字领域专家协作构建的，图中带有“古文字专家”标记的步骤需要领域专家的参与。与通用知识图谱相比，ZiNet 的数据规模相对较小，但是对古文字领域知识的专业性和准确性有非常高的要求。在知识图谱构建的过程中，首先由领域专家参与进行架构与数据模式设计，即定义文章上一节中介绍的实体与关系类别，生成本体文档；然后从拓本图像、古文字著录、字典著录的多模态数据源提取古文字领域的结构化知识，主要的步骤包括古文字拓本图像提取、古文字字体库构建、部件标注、释文标注、词汇与义项提取等主要步骤；在提取知识后通过关系将其进行关联，并将从不同数据源中获取的知识进行融合，最后通过质量检测和知识补全步骤对知识图谱中的关系进行补全和完善，最终构建 ZiNet 汉字知识图谱，并生成一个本体文档，一个字体文档，一个字形数据库，一个出土文献数据库和一个词汇义项数据库，这些数据库和文档也是古文字领域重要的研究资源。参与 ZiNet 构建的领域专家共有 14 人，均为掌握古文字领域知识的专家、博士和硕士研究生。他们按照自身

的研究方向被分文甲骨文、金文、楚简和三个小组，每一组负责对应领域的古文字数据标注。

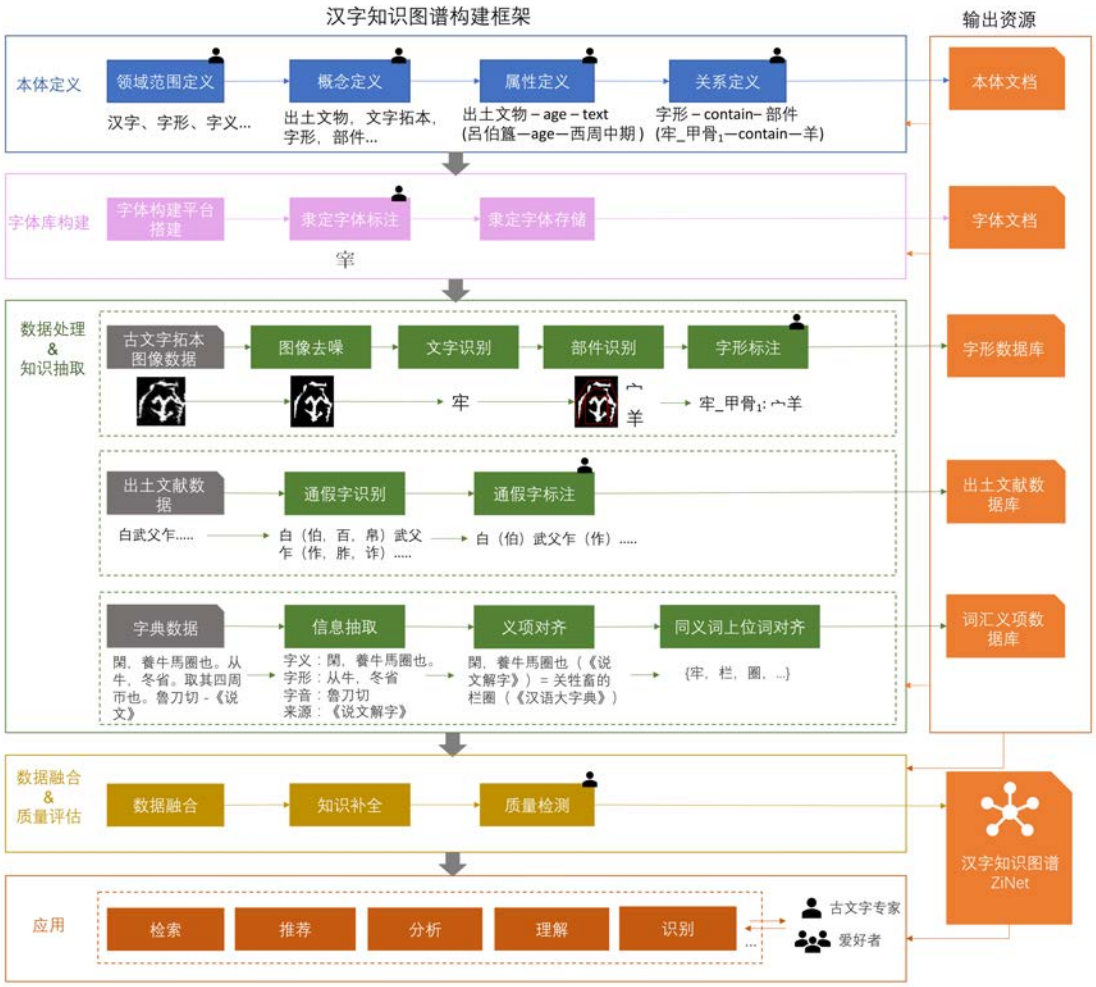


图 3.2 历时汉字知识图谱（ZiNet）的构建框架

ZiNet 构建的主要步骤如下：

（1） 知识图谱架构设计：与古文字领域专家合作定义古文字知识图谱的架构和数据模式，确定知识图谱包含的实体和关系，最后输出知识图谱的本体文档。

（2） 古文字拓本图像降噪与文字识别：基于 CV 领域的算法对拓本图像进行降噪处理，去除图像中由于出土文物表面腐蚀所产生的噪声，并对拓本中的文字进行目标检测，识别拓本图像中的古文字位置。

（3） 古文字提取：将文字识别的结果交由古文字领域专家进行裁剪，形成单字拓本图像，并由专家为每一个图像标注文字标签、文物来源、著录和历史时期等信息，融合其他来源的拓本图像，最终形成古文字单字的

拓本图像数据库。

（4） 古文字字体库构建：搭建古文字隶定字体库构建工具，该工具提供古文字领域专家操作的图形界面，可以通过打字由专家输入每一个部件，由平台拼接形成完整的隶定字并生成图像，也可以由专家上传图像，平台使用 Potrace 工具1批量将文字图像转化为矢量图，并使用 svg2ttf 工具2自动生成字体，最终输出一个古文字字体文件。

（5） 部件标注：基于目标检测算法识别古文字拓本图像中的部件，将提取的部件信息交由古文字领域专家进行进一步校对和标注，古文字专家根据需求模版，对不同级别的部件进行标记。

（6） 释文标注：古文字专家对拓本图像中的上下文进行标记，形成出土文献的释文文本，这一过程需要参考大量的已有的著录，并对其中的通假字进行标注和完善，本文通过自动化的方式初步标记释文中可能存在的通假字，将机器标记的结果交由领域专家进行进一步的标注，此外基于文本相似度计算方法关联上下文相似的历史文献，最终输出一个古文字文献数据库。

（7） 自动化校对：该步骤对由多位专家共同标注的内容进行自动化的一致性检查，不一致的结果由第三位专家重新标注。

（8） 词汇、义项提取：从字典数据中自动化抽取包括词汇、词性、义项定义和例句的信息。字典数据通常是半结构化的，利用正则表达式可以提取目标信息。

（9） 义项融合：对来源于不同的字典或词典的词汇的相同的义项实体进行对齐与合并。该步骤首先对义项的词性、定义和例句属性信息进行字符串匹配，对匹配程度高的义项进行初步合并，再计算义项定义的余弦相似度，对相似度超过一定阈值的义项进行进一步的合并。

（10） 同义与上位关系提取：利用基于规则的方法抽取义项定义文本中的同义关键词和上位关键词，首先根据字典定义的语言规律总结停留词表并删除定义中的停留词，例如“比喻”、“泛指”、“特指”等，接下来设

[1] https://gitcode.com/skyrpex/potrace

[2] https://gitcode.com/jaywcjlove/svgtofont

计规则：如果去除停留词后该定义为一个词汇并存在于词汇表中，则抽取该词汇为同义关键词，否则如果该定义中 “的”与“。”之间为一个词汇并存在于词汇表中，则抽取该词汇为上位关键词，例如“带汁水的菜肴。”可以提取上位关键词“菜肴”，最后将定义和词汇标签中包含相同同义关键词或上位关键词的义项分别标注为同义词和上位词。

（11） 数据融合：利用已经构建的古文字字体进行检索，即可获取同一文字在不同数据源、不同历史时期的关联的数据，例如字形、部件、释文、拓本图像、义项等，基于前文定义的数据模式的属性和关系将其关联起来形成知识图谱。对于不同数据来源、不同领域专家标记的图像和释文文本数据可能出现重复的情况，需要进行对齐与合并，对于拓本图像进行字形相似度计算，相似度超过一定的阈值则将两个字形合并为同一个字形，对释文文本及著录来源信息进行字符串匹配和余弦相似度计算，对相似度超过一定阈值的释文进行合并。

（12） 知识补全：本文主要对字形之间的相似关系进行补全，这是基于第 4 章中介绍的方法，对字形之间的相似性进行计算实现的，在相似度超过一定阈值的字形对之间建立相似关系。

（13） 质量检测：由古文字领域专家对知识图谱的各类型古文字进行抽样质量检测和评估，对于评估不合格的部分进行反工。

下文将对 ZiNet 构建中数据处理与标注的特殊工作进行介绍，包括古文字数据处理问题（3.3.2 节）、古文字拓本图像去噪（3.3.3 节）、古文字部件标注（3.3.4 节）、古文字字体库构建（3.3.5 节），以及古文字通假字标注（3.3.6 节）。

3.3.2 古文字领域数据处理问题

本节介绍古文字领域主要的多模态数据，以及古文字数据处理与标注中的特有问题。数据类型主要包括：（1）古文字的拓本图像与摹本图像数据：拓本图像是指对出土文物进行拓印得到的图像，摹本图像是对拓本进行数字化处理或临摹后形成的更为清晰的图像。古文字的出土文物主要包括龟甲或兽骨残片、青铜器、简帛、玺印、石碑等，出土的文字信息通常被制作成图像的形式进行保存，上下文中的每个文字会被裁剪为单字图像，便于记录古文字的书写形态。（2）出土文献的释文文本数据：由古文字专家对文物的拓本图像进行“翻译”后得到

的文本，“翻译”是指将古文字转换为对应的现代汉字字体进行书写，并且为其添加标点符号、通假字、残缺字等专业标注，以便相关学者对古代文字、语言、文学以及历史、文化等进行研究。（3）古文字著录的文本数据：包括历朝历代对汉字研究形成的字典、词典、字形集成、论文等著录，这些著录对于汉字的字形、字义、字音、语用、考释分析都进行了归纳和整理。图 3.3 提供了“伯武父鼎”文物的金文案例：

图 3.3 “伯武父鼎”金文多模态数据案例

通过对古文字专家进行访谈，本文关注到古文字领域数据处理和知识标注中的特殊问题，并在后文提供自动化的解决方案：

（1） 拓本图像的噪声问题：出土的文物由于常年遭受侵蚀，拓本图像包含很多噪声，这些噪声与文字的笔画混在一起难以辨识，去噪处理既有益于领域专家的阅读，也可为图像识别等算法提供更高质量的数据。

（2） 古文字部件标注问题：汉字的部首包括形符和声符，和语义和读音高度相关，明确每一个文字的部件组成对于古文字研究来说是必要的。因此本文在知识图谱中为每个字形关联部件信息，如何快速为古文字字形标注部件是一个问题。

（3） 古文字字体编码问题：很多未释读的古文字缺乏字体编码，古文字专

家经常使用图像对它们进行操作，这不利于后续处理。因此，对特殊的古文字进行隶定、编码、并生成统一的字体库是必要的步骤。隶定字是指使用通行的隶书或楷书笔法将古文字重新写定的方法，例如图 3.3 中的“  ”字，隶定字与古文字之间只有书体风格上的差异，没有部件结构上的差异。

（4） 出土文献的通假字标注问题：在古文字时期，由于汉字系统还未发展完善，以及古汉语中的用字不规范现象，经常存在一些用发音相同或相近的文字，即通假字来替代本字进行书写的情况，本字可能在当时已经存在，也可能在后世才被造出。例如“聞”是“婚”字的通假字。为了便于理解出土文献的含义，古文字专家需要对上下文中的通假情况进行判断，并加以标注，例如聞（婚），乍（作），白（伯）等，图 3.3 中展示了通假字标注的案例。

3.3.3 古文字拓本图像去噪方法

古文字拓本图像往往具有大量的复杂噪声。这些不同程度的复杂噪声严重干扰古文字的可视性和可读性，为后续的古文字检测与识别等视觉任务带来了极大的阻碍。与高斯白等简单的合成噪声不同，古文字拓本的噪声更为复杂，常见的图像降噪方法主要集中在去除简单噪声上，并且用户需要提前知道噪声的类别和等级，这在实践中很难满足，所以导致其在拓本图像中取得的效果不佳。

本文采用 Shi 等人[13]提出的基于生成对抗网络的图像降噪算法对古文字拓本图像进行降噪处理，模型结构如图 3.4 所示，该模型首先基于 U-Net 网络，从输入的噪声图像中提取出文字骨骼图像。U-Net 是 CNN 的一种，主要包括 10 个卷积层，其中有 5 个下采样层和 5 个上采样层。接下来基于生成对抗网络实现拓本图像的降噪，生成对抗网络包括生成器和判别器两部分，如图 3.4 所示，生成器将文字骨骼图像和与其对应的噪声图像作为输入，输出无噪声的文字图像，生成器以 U-Net 模型作为基础结构，包括 4 个下采样层和 4 个上采样层，每一层由多尺度特征融合残差块（MFR）组成。每个 MFR 包括四个子层，分别为多尺度集成层、连接层、卷积层和通道注意力机制层。判别器网络用于区分生成器产生的无噪声文字图像和原始无噪声文字图像，将生成器产生的无噪声图像作为输入，依次经过 5 个 MFR 的处理后，最终由全连接层对图像做二分类，使辨别网络能

够区分生成图像与真实图像。图 3.5 为甲骨文拓本图像去噪后的效果，可以看出，该方法去除了一部分古文字拓本中的腐蚀痕迹，使文字更易辨认，不仅有益于人类专家的阅读，也可以为后续的图像识别等操作提供更高质量的数据。

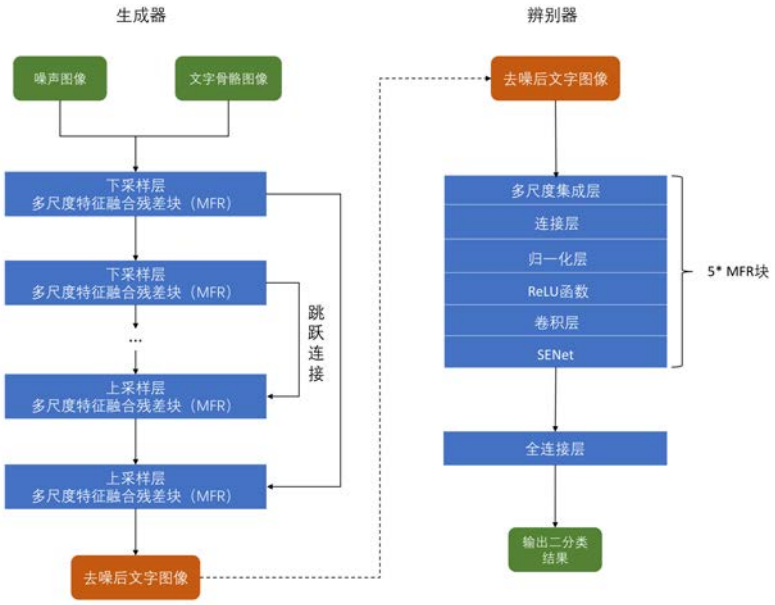


图 3.4 拓本图像去噪的生成对抗网络模型结构

图 3.5 古文字拓本图像去噪效果图

3.3.4 古文字部件标注方法

古文字的部件标注主要分为三个阶段，首先构建用于模型训练的古文字图像部件数据集，由于目前缺乏标记部件及其位置信息的古文字拓本图像数据集，因此数据集构建是首要的工作。然后，本文使用 Diao 等人提出的一种基于改进的 YOLOv4 （You Only Look Once Version 4）的神经网络目标检测模型对古文字拓本图像中的部件进行识别[28]，最后将识别的结果交由领域专家进行进一步的校对与标注。在古文字部件图像数据集的构建中，本文使用 labelimg 图像标注软件对古文字拓本图像中的部件进行标注，19 位接受过专业培训的标注者对古文字样本图像中的部件进行框选，并标注对应的部件名称。接下来对标注好部件的古文

字样本图像进行数据处理，以扩充古文字部件图像数据集，预处理的方法包括图片大小调整、色域变换和图片翻转和图片拼接。

Diao 等人的古文字部件识别模型主要分为三部分：第一部分通过主干特征提取网络进行初步的特征提取，包括纹理，颜色和形状等，获得三个有效的特征层。第二部分为基于 SPP（Spatial Pyramid Pooling）和 PANet（Path Aggregation Network）的加强特征提取网络，SPP 网络结构利用 4 个不同大小的池化核进行池化处理，再将不同的特征图进行拼接，可以分离出最显著的上下文特征。PANet 加强特征提取网络是一种实例分割算法，可以对于特征进行反复提取，该部分对三个特征层进行特征融合，获得三个更有效的特征层。第三部分的预测网络利用捕获的特征识别出古文字的部件。图 3.6 为古文字部件检测结果的案例，可以看出该算法输出古文字拓本图像中的部件位置和部件类别标签。本文在包含 2005 个甲骨文图像，202 个部件类别的测试集上对该算法进行评估，准确率为 58.1%。

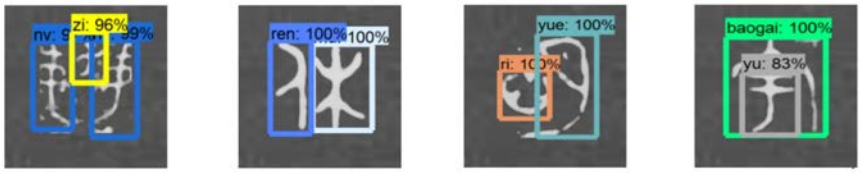


图 3.6 古文字部件目标检测效果图

最后由古文字专家进行进一步的标注，包括对机器标注的不正确的部件标签进行修改，以及对古文字部件之间的层级关系进行完善，最后按照规定的模版为每一个字形输出结构化的部件描述文本，模版以“，”作为标识符分隔部件，使用“（）”符号来描述部件之间的层级关系，例如甲骨文“宿”字第二个字形（图 3.1）的部件描述文本为：“宀，佰（亻，百）”。为回答使用模型是否能够提升专家的标注效率以及提升效率的原因，与 3 位参与标注的古文字领域专家进行一对一访谈，3 位受访者均表示在模型返回的结果中进行修改可以提升工作效率，首先因为直接使用模型回答正确的部件提升了标注速度，其次因为模型的返回结果对难以辨认的模糊的部件具有提示作用。

3.3.5 古文字字体库构建方法

古文字具有图画性强，字无定型等特点，很多古文字，尤其是未释字没有计算机字体，这对古文字学者的计算机阅读和处理、以及后续的智能化算法运行带

来极大的困难。因此本文构建了一个古文字隶定字体库构建工具，该工具首先根据隶定字的结构和部件标签生成文字表达式，然后基于文字表达式生成字体的矢量图，最后将其转换为字体保存到 ttf 格式的字体文件中。隶定是指使用通行的隶书或楷书笔法将古文字重新写定的方法，隶定字与古文字之间只有书体风格上的差异，没有部件结构上的差异。

文字表达式由文字结构编号和部件名称组成，有两种输入方法，一是由古文字专家直接在平台界面对应的文字结构预设框中输入组成文字的各个部件，二是上传一个隶定字体的图像，通过基于 YOLOv4 的目标检测模型与汉字部件位置关系推理模型自动得到图像中的各个部件和文字结构。本文的文字结构的包括 14 种，分别为“01 独体字”，“02 上下结构”，“03 左右结构”，“04 包围结构”，“05 半包围-左上右下”，“06 半包围-左下右上”，“07 左中右结构”，“08 上中下结构”，“09 左右结构-上下结构”，“10 上下结构-左右结构”，“11 包围结构-上下结构”，“12 品字结构”，“13 上下结构-品字结构”和“14 镶嵌结构”。例如“”这一隶定字属于“09 左右结构-上下结构”，生成的文字表达式为：“09 女頁刀”，表达式中部件的排列顺序是按照文字结构中部件序号的先后顺序排列的，且文字结构中部件的序号是按照从左到右、从上到下、从内到外顺序进行编号。

在生成文字表达式后，根据表达式从现代汉字与部首的矢量图库中提取对应的矢量图，再把部件矢量图根据表达式中的结构放入到结构框预设的相应位置中，按照汉字的比例对各个部件进行缩放、拉伸、平移、旋转等变换。在文字生成后，将图片自动转为一个整体的矢量图保存，最终转化成 ttf 文件，完成字体构建。图 3.7 是古文字隶定字体库的图例，除字体外还存储了对应的文字表达式，即文字结构和部件标签。

图 3. 7 古文字隶定字体库存储案例

3.3.6 古文字通假字标注方法

本文基于 BERT 实现古文字释文文本的通假字标注任务。本文统称释文文本（例如图 3.3）括号中的文字为本字，括号外的文字为通假字，例如“聞（婚）”，“乍（作）”，“白（伯）”等。在上古时期，由于本字还未被造出，或者书写者的用字不规范，出土文献中存在很多借用发音相同或相近的通假字来替代本字进行书写的情况，在通假字标注中，模型需要识别文本序列中的通假字，并将其映射到对应的本字。

本文将通假字标注建模为 NLP 中的序列标注任务，模型如图 3.7 所示，该任务对输入文本序列中的每一个古文字字符进行分类，对于每一个输入的目标字符，如果模型将其分类到类别编号“0”中，则表示该字符为非通假字，如果分类到其他类别编号，则说明该字符为通假字，对应的本字为类别编号对应的文字标签。需要注意的是，由于通假字是发音相同或相近的文字，因此，本文参考上古读音与拼音，仅将与目标字符音近或音同的文字设置为有效的类别标签，如果概率得分最高的文字不是有效标签，则对得分最高的前 N 个文字标签按分数从高到低排序，依次进行检验，将第一个有效标签设置为对应的本字，如果前 N 个文字中无有效标签，则将该字符分类为非通假字，本文将 N 设置为 50。

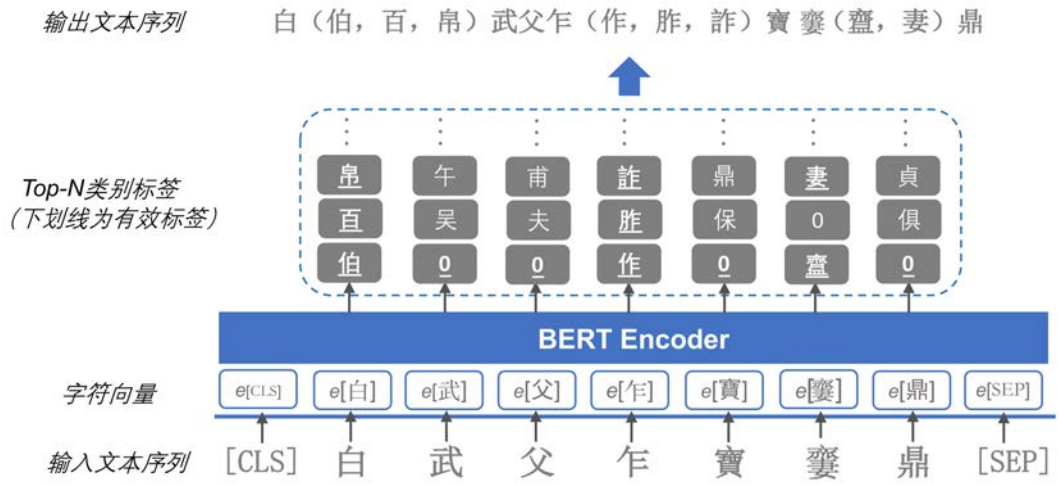


图 3.8 基于 BERT 的古文字通假字标注模型结构图

本文的 BERT 模型是以 bert-ancient-chinese[3]为基础，在甲骨文、金文和楚简的出土文献数据集上继续训练获得的，bert-ancient-chinese 模型基于谷歌 bert-

[3] https://huggingface.co/Jihuai/bert-ancient-chinese

base-chinese[4]在古汉语大规模传世文献数据上继续训练获得。本文将 BERT 模型的输出直接接入线性分类器进行分类得到文本序列中每一个字符的类别标签。本文在 57766 条已经标注通假字的出土文献数据集中微调 BERT 模型，对 2548 条来自甲骨文、金文、楚简的上下文进行测试，每一个测试文本均包含至少一个通假字，结果显示模型的精度为 90.89%，F1 分数为 35.99%。对结果进行分析发现对于经常使用的通假字，例如“白（伯）”、“乍（作）”等，模型的效果很好，而对于很少使用，甚至未出现在训练集中的通假字的效果较差，但是经常使用的通假字在文本中的比例很大，因此自动化标注仍可以提升专家的标注效率。

本文对模型的通假字标注结果按照释文文本形式整理，即图 3.8 中的输出文本序列，在括号中提供前 N 个结果中所有有效的文字标签排序作为候选，为领域专家提供参考，例如“.....白（伯，百，帛）.....”，该文本序列将交由古文字专家进行进一步修改和标注，最终获得标记后的释文，即“..... 白（伯）.....”。

3.4 历时汉字知识图谱实体统计

本节对 ZiNet 的实体规模进行统计，并讨论该知识图谱的潜在应用与未来的挑战。ZiNet 当前主要包含甲骨文、金文、楚简三种古文字类型，相关的图像、释文、字典等数据源是由古文字领域专家提供的专业数据。截止到目前，ZiNet 主要实体的数量如表 3.3 所示，其中“=”表示该字形的数量，“≤”表示在该时间节点之前共有的实体数量。

ZiNet 的一个潜在应用是为古文字专家和爱好者提供知识检索服务，与传统的数字化平台相比，可以为用户智能化地推荐具有关联关系的文字、文献上下文、拓本图像等多模态信息。ZiNet 作为古文字领域的知识资源，可以为计算机视觉、自然语言处理领域的模型提供数据和知识支撑，服务于智能化的古文字出土材料的处理、以及未知古文字的考释研究。例如利用古文字部件结构的信息改进古文字图像切割、降噪、修复、识别等模型的效果；计算来自各个历史时期的字形之间的相似度，为古文字领域专家推荐相关字形（第 4 章）；预测未知字形所属的文字标签（第 5 章）；对出土文献与传世文献进行对读，为专家推荐相关的历史文献上下文；对词汇的语义变化进行可视化分析（第 6 章）等。当前，大模型已

[4] https://huggingface.co/bert-base-chinese/tree/main

经在多个领域的人工智能应用中取得突破性进展，在未来，可以利用 ZiNet 中的数据训练一个集古文字识别与理解为一体的多模态大语言模型，以支持文字缀合、考释、器物断代等一系列复杂、高难度、具有实际应用价值的任务。

表 3.3 历时汉字知识图谱（ZiNet）的实体统计

实体	统计
拓本图像	15123=甲骨； 14482=金文； 22893=楚简
字形	2892=甲骨； 3289=金文； 6439=楚简
部件	甲骨=584；金文=853；楚简=868
古文字	2419=甲骨； 2373=金文； 4787=楚简
文字	已释字：1255=甲骨；3079≤金文；6792≤楚简；18966≤现汉；未释字：1164=甲骨；1713≤金文；2787≤楚简
词汇	423997
义项	69825≤206BC；177570≤618AD；315181≤1368AD；386949≤1840AD；570764≤现代
同义词集	366544
释文条目	41689=甲骨；21610=金文；5602=楚简

但是，ZiNet 仍有一些局限性，目前的知识图谱的构建仍然主要依赖于古文字领域专家大量的人工标注，会受到人类体力、脑力、记忆力的局限，无法覆盖大规模的数据。在未来如果要将知识图谱在短时间内扩展到更多种类的古文字，并建立可持续的高效率的更新机制，必须进一步研发和改进各类自动化出土材料处理与知识抽取、知识融合的方法，也需要研发高效的出土材料处理和标注工具，建立更高效的人机交互和众包机制，进一步提升领域专家处理和标注古文字材料的效率。

3.5 基于历时汉字知识图谱的智慧平台

基于 ZiNet 古文字知识图谱以及所实现的自然语言处理、计算机视觉技术，本文设计并开发了“古文字智慧平台”，已经投入使用。截止到目前，平台已实现了“古文字检索”、“电子字编检索”、 “古文字图像降噪”、 “古文字图

像识别”、“释文检索”、“古文造字”、“辞例分析”等功能，为古文字相关研究人员提供了实用性强、准确率高的智能服务。下面是各个功能的简单介绍：

（1） 古文字检索：古文字检索是对汉字基本信息的检索功能，用户输入汉字，输出包括该字从甲骨文、金文到楚简时期的字形演变信息、该文字的可视化知识图谱、该文字的所有异体字形与部件信息、该文字的字典解释，以及字形相似、字义相似的其他文字、词汇的推荐链接等。图 3.9 展示了“宿”字的检索案例。

(a) 字形演变与知识图谱　(b) 字形图例

(c) 字典解释　(d) 相似字形推荐

图 3.9 古文字检索功能案例

（2） 电子字编检索：电子字编检索是对所有古文字单字拓本图像的检索功能，用户输入单个汉字，输出包括该字从甲骨文、金文到楚简时期的所有拓本图像，以及该图像的文物、时期与著录来源信息。

（3） 古文字图像降噪与图像识别：该功能利用 3.4.2 节与 3.4.3 节所述的拓本图像降噪与基于部件目标检测的文字识别算法实现，用户上传古文字拓本图像，降噪功能返回降噪后的图像，并支持下载，古文字识别功能返回得分最高的几个文字标签与对应的概率分数，图 3.10 展示了“宿”

字的图像降噪与识别案例。

(a) 古文字图像降噪功能

(b) 古文字识别功能

图 3.10 古文字拓本图像降噪与识别功能案例

（4） 释文检索：该功能为甲骨文、金文、战国文字的出土文献文本信息的检索功能。用户输入在释文中出现的关键词，也可以通过高级检索，选择该释文的著录来源、时期、文字类别、文物类别等信息，系统返回符合检索要求的所有释文条目以及相关信息，用户可以指定古文字的类别和关键属性对结果进行筛选过滤，关键字“宿”的释文检索案例见图 3.11。

甲骨文　金文　楚简

为您找到相关结果6个，当前为第1页　　按器物主分类-子分类筛选：酒器　全部

释文编号	时代	器类	器名	释文	著录信息
金文09137	西周中期	酒器-爵	史宿爵（史宬爵）	史宬（宿）乍（作）父庚寶尊（尊）彝。	銘圖08564
金文12176	西周中期	酒器-尊	宿父尊	宿父乍（作）父癸寶尊（尊）彝。	銘圖11689
金文12807A	春秋早期	酒器-壺、鐘	鄀季宽車壺（冥季宿車壺）	鄀（冥）季車宾（宽）自乍（作）行鼎。子孫永寶用之。	集成09658.2，銘圖12326A
金文12807B	春秋早期	酒器-壺、鐘	鄀季宽車壺蓋（冥季宿車壺蓋）	鄀（冥）季宾（宽）車自乍（作）行△。子孫永用之。	集成09658.1，銘圖12326B
金文14412	春秋晚期	酒器-缶	宽兒缶甲（宿兒缶）	隹（唯）正八月初吉壬申。蘇公之孫宾（宿）兒嬭（擇）其吉金。自乍（作）行缶。朁（沫-眉）壽無疆（期）。永保用之。	銘圖14091
金文14413	春秋晚期	酒器-缶	宽兒缶乙（宿兒缶）	隹（唯）正八月初吉壬申。蘇公之孫宾（宿）兒乍（作）嬭（擇）其吉金。自乍（作）行缶。朁（沫-眉）壽無疆（期）。永保用之。	銘圖14092

图 3.11 “宿”字的释文检索案例

（5） 古文造字：该功能基于 3.4.4 节所述方法实现字体库构建，用户输入 svg 格式的字体图像，系统将该文字提交到字体库并生成可复制的字体。

（6） 辞例分析：该功能利用 3.4.5 节训练的 BERT 模型，对出土文献缺失文

字进行预测，用户输入句子并将需要预测的一个或几个字用 [MASK] 代替，系统返回[MASK]位置的文字预测结果，图 3.12 展示了一个甲骨文辞例的分析案例。

图 3.12 辞例分析功能案例

3.6 本章小结

本章介绍了 ZiNet 历时汉字知识图谱， ZiNet 对部件、字形、文字、词汇、义项等专业概念的系统性关系进行了描述，它可以从字形和字义层面关联不同历史时期的文字，描述汉字的演变过程，这是本文在汉字知识图谱结构设计上的创新之处。接下来介绍了拓本图像去噪、部件标注、字体库构建以及释文通假字标注几个重点任务的自动化方法，以及与领域专家的协作方式，一定程度地提升了知识图谱的构建效率。ZiNet 已经覆盖甲骨文、金文、楚简文字中的 12620 个古文字字形，在人工智能辅助的古文字考释方面都具有潜在的应用价值，目前已经基于 ZiNet 建设了智慧古文字平台，为领域专家提供丰富的检索与数据处理功能。

第 4 章　基于知识图谱的古文字字形相似度计算方法

中国历史上存在过数以万计的字形，古文字专家在研究的过程中经常需要将相似的字形进行对比，从而归纳和推理出文字演变的规律，对未知文字进行考释。但是发现字形相似的文字很大程度上依赖于专家的经验和记忆力，无法穷尽所有的字形。目前自动化的字形相似度计算方法主要针对现代汉字，不适用于古文字，一是由于古文字不具有现代汉字的笔画等一些结构化特征，二是这些方法没有针对古文字的部件语义逻辑进行设计，无法捕获在字源或者造字逻辑上相关的文字。本章基于 ZiNet 汉字知识图谱，提出一个古文字字形相似度的计算方法，该方法综合考虑汉字在视觉上的象形相似度和部件语义的相似度，对古文字之间的字形相似度进行自动化评分，在实际应用中为领域专家推荐字形相似的古文字。本文通过定量实验与定性分析证明了该方法的有效性。本章 4.1 节为引言，4.2 节介绍字形相似度计算任务的定义，4.3 节将介绍提出的几种相似度计算方法，4.4 节介绍实验设计，数据集，定量评估和定性分析的结果与讨论。4.5 节对本章的工作进行总结。

4.1 引言

在古文字研究中，专家经常需要将待分析的古文字字形与历史上的字形进行比较分析，从而得出字形演变的规律和字形之间是否具有联系的结论。然而，中国历史上存在过数以万计的汉字字形，发现相似的汉字很大程度上依赖于专家经验，必然受到人类记忆的限制，研究自动化的古文字字形相似度计算方法可以为领域专家推荐相关的字形，服务于古文字研究，具有实际应用意义。

汉字字形相似度的测量相对于印欧语言来说更为复杂，因为汉字由图形和部首组成，无法直接使用编辑距离进行度量。相关工作通常对现代汉字基本属性中的一种或多种进行表示，将汉字义映射为高维向量，包括汉字的五笔编码表示方法[3]、基于笔画、构词和拼音的汉字向量表示方法[4]、基于偏旁部首知识图谱的汉字字形表示方法[7]、基于 ViT 的汉字图像表示方法[2]等，然后利用如 Jaro-Winkler Distance 等距离度量算法计算向量之间的距离，获得汉字的相似性分数。多数方法难以应用于古文字，其中的问题包括：（1）古文字与现代汉字的特征不同导致方法无法通用。古文字的书写不固定，部件位置灵活，象形程度高，难以总结出固定的结构和笔画。（2）现有方法未对汉字部件语义的相似度进行表

示与计算。现代汉字的相似度计算方法主要为了防止人类或机器的误认、误写而设计，因此针对文字结构、整体轮廓与形态、笔画数量等方面的相似性进行计算，但是对于古文字专家的研究来说，除视觉上的相似外，部件构成所反映出的语音、语义、字源上的潜在联系非常重要，如果能够捕获这些具有潜在关联的相似文字，对于古文字研究将更有价值。

基于以上的背景和问题，本章的贡献包括：（1）提出了一个古文字字形相似度的计算方法，称为 GlyphSim（Chinese Character Glyph Similarity），对古汉字之间的字形相似度进行自动化评分，服务于相似字形推荐的实际应用。该方法基于计算机视觉技术和第 3 章中介绍的 ZiNet 汉字知识图谱，不仅针对汉字在视觉上的象形相似度，也针对汉字在部件结构逻辑上的相似度。具体来说，本文设计了象形相似性（Pictographic Similarity, PicSim）、部件描述相似性（Radical Longest Common Subsequence Similarity, RLCSSim）、以及部件语义图谱相似性（Radical Semantic Graph Similarity，GraphSim）三种相似度计算方法，并将其进行融合。部件语义图谱基于 ZiNet 构建，对古文字的字形以及部件之间的近义、分化等专业关系进行表示，使本文的方法能够表示文字的部件语义特征。（2）本文在构建的甲骨文数据集中通过定量实验和定性分析来对方法进行评估，定量实验的结果显示，本章方法对数据集中甲骨文字对的相似度打分与专家打分之间呈显著的正相关关系，Spearman's 相关性分数为 0.8422，而仅基于视觉图像相似的方法的相关性分数仅为 0.3241，此外，在相似古文字推荐测试中，57.13%被标注的相似字对可以被该方法的前 5 位（Top-5）推荐结果覆盖，证明了该方法的有效性，此外，对 Top-5 推荐结果的定性分析可以表明该方法能够捕捉到可能在词源或语义上相关的字形。

4.2 任务定义

给定（c_1，c_2），c_1与c_2是待计算的两个古文字，c_1与c_2各拥有一个或多个字形：G_{c_1}=$\{g_1^{c_1},..g_n^{c_1}\}$，$G_{c_2}$=$\{g_1^{c_2},..g_m^{c_2}\}$，其中$G$是字形集合，$g$表示字形，$n$和$m$分别是$c_1$和$c_2$的字形数量。每一个字形$g_i$拥有一组描述它的图像：$P_{g_i}=\{p_1,\dots,p_{|P_{g_i}|}\}$，$P_{g_i}$表示$g_i$的图像集合，同时$g_i$拥有一个组成它的部件序列$R_{g_i}$：$R_{g_i}$=$\{r_1,\dots,r_{|R_{g_i}|}\}$，$r$表示部件，$R_{g_i}$序列中的部件按照位置分布以汉字书写的规则

排列。有一个知识图谱$KG(V，Rel)$，其中V表示知识图谱中的所有节点，Rel表示连接节点的边，知识图谱KG中共有文字、字形与部件三种实体作为节点，文字集合C、字形集合G、部件集合R中的元素，以及由它们之间的关系构成的边均包含在KG中。

本文希望构建一个汉字字形相似度计算模型，称为 GlyphSim，输入文字c_1与c_2的字形集合G_{c_1}，G_{c_2}，每个字形对应的图像集合P_{g_i}与部件序列R_{g_i}，以及字形知识图谱KG，模型输出c_1与c_2的相似度得分，分数在 0-1 之间，越接近 1 表示c_1与c_2的字形越相似。

4.3 古文字字形相似度计算方法

4.3.1 方法原理概述

本文的字形相似度计算方法不仅针对汉字视觉上的相似性，还包括它们的部件系统语义之间的相似性。具体的评价标准和原理如下：

一．视觉相似性：如果两个字形的整体轮廓与书写形态相似，则它们具有相似性。视觉相似性主要是为象形字和指事字这一类独体字设计，在最早的甲骨文中，象形字的占比相对较高，大约占 40%，例如 “鸟”、“牛”、“刀”，它们是通过直接用笔画描绘物体的图案被创造出来的文字，指事字是通过在象形文字的图案上增加抽象化的指事符号被创造出来的，例如“刃”通过在“刀”上增加符号来表示“刀刃”的含义，它们在视觉形态上也是相似的。如果两个独体字的视觉形态相似，那么它们的语义或者文字演变很可能相关，例如“隹”与“鸟”，“羊”与“牛”等。

二．部件系统的语义相似性：如果两个文字共享部件，或者部件之间具有潜在关系，则它们具有部件语义相似性。汉字除了象形字和指事字外，还包括会意字和形声字，它们是通过使用部件组成合体字的两种造字方式，其中形声是由表意符号和表音符号两部分组合成字的造字法，例如“材”字，形符为“木”，声符为“才”，会意是用两个或两个以上的独体字会合起来表示一个新的意义的造字法，例如“鸣”由“口”和“鸟”组成，用于表示鸟的叫声。形声字的数量随着汉字系统发展成熟而迅速增加，在西周金文中，形声字占比已经接近 60%，会意字大约占 30%，而象形字和指事字则缩减到 10%左右。因此，仅考虑汉字视觉

上的相似性是不够的，部件与古文字的本义和读音高度相关，文字的部件系统之间的相似性可以反映出古文字的字义、字音、演变发展之间的关联，本文中部件语义相似性的具体指标如下：

（1） 部件共享：两个字形包含相同的部件，则它们具有部件语义之间的相似性。

（2） 部件具有异体、近义、同音、派生或指事方面的语义关系：如果两个字形包含语义相关的部件，则它们具有部件语义相似性。例如：“虎”与 “虍”属于异体字； “木”与“屮”表示部件时的含义相关；“匕”与“比”的读音相同，在古文中经常作为通假字；“東” 与 “束”是从同一个字分化出来的，具有派生关系；“刀”与“刃”具有指事关系。

（3） 部件在造字时具有通用关系：如果两个字形包含通用部件，则它们具有部件语义相似性。通用部件是通过探索同一文字的异体字形发现的，如果同一文字两个异体字形除某一对部件外，其他部分组成完全相同，则这一对部件在创造该字时是通用的。例如，“牢”字在甲骨文中有两个异体字形，第一个包含部件“宀”和“牛”，第二个包含“宀”和“羊”，在这个案例中，“牛”和“羊”是一对通用部件，在造字中表示“家畜”的含义。如果一对部件在多个古文字的造字案例中发现通用，则本文将它们设置为通用部件。

图 4.1 展示视觉相似性与部件语义相似性的案例：

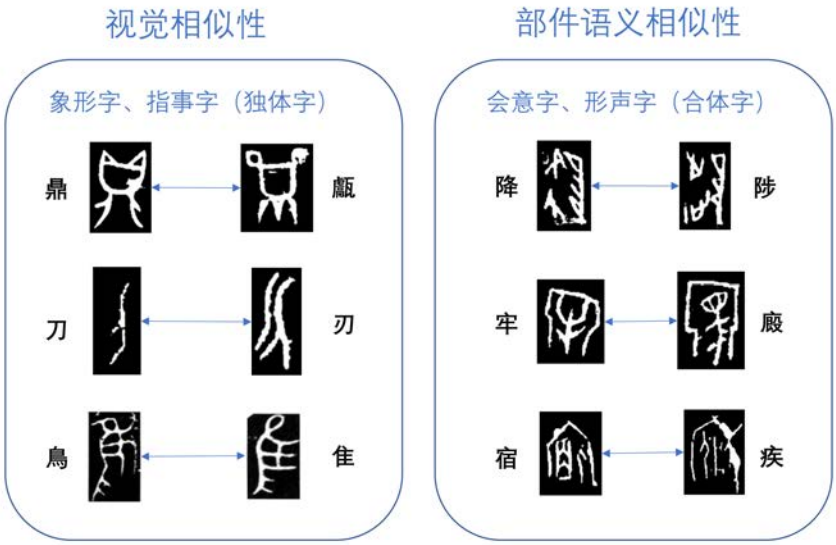


图 4.1 视觉相似性与部件语义相似性图例

根据以上字形相似度的评价标准，本文设计了象形相似度（PicSim）、部件描述相似度（RLCSSim）、部件语义图谱相似度（GraphSim）三种字形相似度的计算方法，分别将在 4.3.2、4.3.3 和 4.3.4 节介绍。其中 PicSim 基于计算机视觉技术从图像中提取字形的形态特征，而 RLCSSim 和 GraphSim 则从部件文本和 ZiNet 提供的部件语义图谱中提取部件系统的特征来计算相似度。

4.3.2 象形相似度计算方法

本文使用 ResNet 网络提取古文字图像中的视觉特征，方法如图 4.2 所示。假设文字集合为$C = \{c_1，\ c_2，\ \ldots, c_n\}$，所有的图像集合为$P = \{p_1，\ p_2，\ \ldots, p_m\}$，网络任务是将图像$p$分类到相应的文字标签 c 中：$f(p|\varphi) = c$，其中$\varphi$是需要通过训练获取的参数。网络输入图像$p$，输出一个$|C|$维向量，表示每一个文字标签$c$的概率分数。在训练步骤中，提供图像及其对应的汉字标签，通过最小化交叉熵损失函数得到最优参数φ。

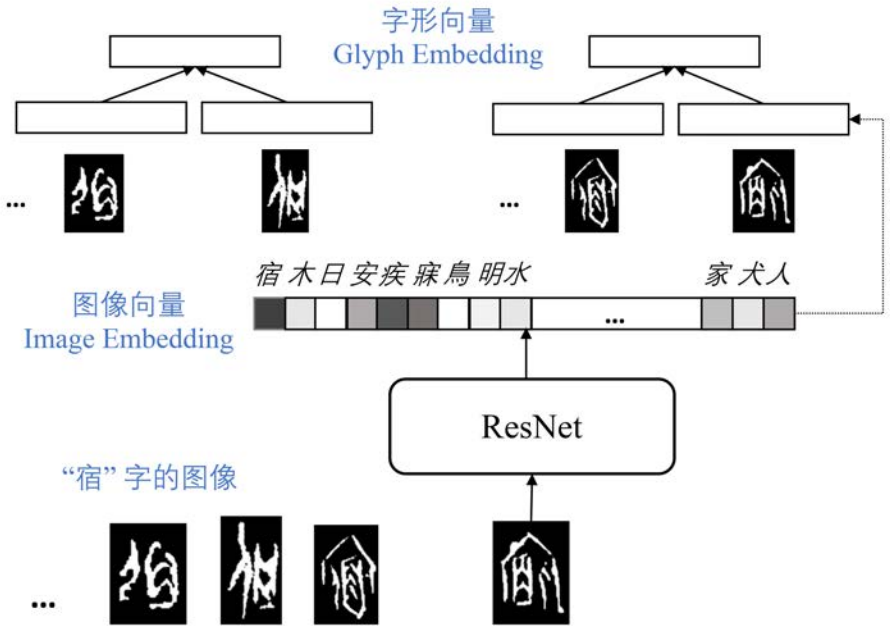


图 4.2 象形相似度计算方法

本文使用网络输出的$|C|$维向量作为输入图像的特征表示$\boldsymbol{e^v}$。对字形g的表示$\boldsymbol{e^g}$被设置为该字形的图像集合P_g中所有图像向量的平均值：

$$\boldsymbol{e^g} = \frac{1}{|P_g|} \sum_{p_j \in P_g} \boldsymbol{e_j^v} \tag{4.1}$$

得到所有字形的嵌入$\boldsymbol{e^g}$后，本文计算字形向量之间的余弦相似度得到字形对(g_i, g_j)之间的相似度分数$Picsim(g_i, g_j)$：

$$PicSim(g_i, g_j) = \alpha Cosine(\boldsymbol{e_i^g}, \boldsymbol{e_j^g})$$
$$\begin{cases} \alpha = 1, & R'_{g_i} \cap R'_{g_j} \neq \emptyset \\ 0 < \alpha < 1, & Otherwise \end{cases} \tag{4.2}$$

$Cosine$表示余弦相似度计算，这里需要乘以一个超参数α，仅当两个字形共享相同或有语义关系的部件时，$\alpha = 1$，否则，α将被设置为一个大于 0 且小于 1 的值。R'_g为字形g包含的所有部件以及与包含的部件具有语义关系的部件集合。部件的语义关系包括 4.3.1 节中所述的异体、近义、同音、派生、指事，以及通用关系。

最后，给定待测古文字对（c_1，c_2）的字形集合G_{c_1}，G_{c_2}，c_1与c_2之间的象形相似性分数$PicSim(c_1, c_2)$被设置为其所有字形对组合的相似度最大值。

$$PicSim(c_1, c_2) = Max\{PicSim(g_i, g_j)\}, \quad (g_i \in G_{c_1}, g_j \in G_{c_2}) \tag{4.3}$$

4.3.3 部件描述相似度计算方法

RLCSSim 旨在度量部件描述序列之间的相似性。本文将一对待测试的字形(g_i, g_j)表示为组成它们的部件文本序列R_{g_i}={$r_1, \ldots, r_{|R_{g_i}|}$}，$R_{g_j}$={$r_1, \ldots, r_{|R_{g_j}|}$}，部件在序列中的排序是按照汉字的笔画顺序决定的，一般遵从先左后右，先上后下，先内后外的规则。部件序列中只包含最小级别的部件，而不包含中间的次级部件，例如甲骨文中“宿”字第二个字形的部件序列为{宀，亻，丙}。如果一个字形是独体字，那么它的部件序列中只包含它本身。

公式 4.4 为字形g_i与字形g_j的部件描述相似度的计算方法。RLCS(R_{g_i}, R_{g_j})表示部件序列R_{g_i}和R_{g_j}之间的最长公共子序列。值得注意的是，$\left|RLCS(R_{g_i}, R_{g_j})\right|$，即最长公共序列长度的计算方法不仅针对两个部件序列中相同的部件，还包括 4.3.1 节中所述的语义相关的部件，如果两个序列中对应的两个部件相同，则最长公共子序列的长度数值将加 1，如果对应的两个部件不同，但是语义相关，则最长公共子序列的长度将加一个超参数θ，0<θ<1。在获得所有字形之间的相似

度分数$RLCSSim$后，可根据公式 4.5 计算两个古文字之间的相似性，古文字之间的相似度与公式 4.3 一样被设置为它们所有字形组合之间的相似度的最大值。

$$RLCSSim(g_i, g_j) = \frac{2 \times \left|RLCS(R_{g_i}, R_{g_j})\right|}{\left|R_{g_i}\right| + \left|R_{g_j}\right|} \tag{4.4}$$

$$RLCSSim(c_1, c_2) = Max\{RLCSSim(g_i, g_j)\}, \quad (g_i \in G_{c_1}, g_j \in G_{c_2}) \tag{4.5}$$

4.3.4 部件语义图谱相似度计算方法

RLCSSim 的相似性分数是离散的，为了在高维向量中表示字形，并获得连续的相似度分数，本文引入了部件语义图谱相似性（GraphSim），构造一个包含部件r、字形g与文字c三种实体节点的部件语义知识图谱，通过部件将所有汉字字形关联起来。如图 4.3 所示，图中包含三种关系：$R1(c，g)$，$R2(g，r)$，以及$R3(r，r)$：$R1$描述文字和字形之间的从属关系，如果某一字形节点属于某文字节点，则在它们之间建立一条边；$R2$描述字形和部件之间的包含关系，如果某一字形节点包含某部件节点，则在它们之间建立一条边，包括任一级别的部件；$R3$描述部件之间的语义关系，如果两个部件之间存在第 4.3.1 节中所述的部件共享、异体、近义、同音、派生、指事、或者通用关系，那么在这两个部件之间建立一条边。该部件语义知识图的节点均来自 ZiNet 汉字知识图谱，$R1$与$R2$关系可以从 ZiNet 中直接获取，$R3$关系由古文字专家进一步标注获得。

通过引入部件语义知识图谱使语义相似的部件及相关字形在高维向量空间中具有更近的距离。接下来，如公式 4.6 所示，使用随机游走算法（node2vec）生成字形节点的高维向量$\boldsymbol{e^g}$，node2vec 是一种基于深度学习和随机游走策略的网络节点表示算法。在生成所有字形的高维向量表示后，根据公式 4.7，计算待测字形的向量之间的余弦相似度得到字形之间的相似度得分。最后，给定古文字拥有的所有字形，两个古文字的部件语义图谱相似度的计算方式与上文中公式 4.3、4.5 相同，即设置为它们所有字形组合之间的相似度的最大值（公式 4.8）。

$$\boldsymbol{e_i^g} = node2vec(KG)[i] \tag{4.6}$$

$$GraphSim(g_i, g_j) = Cosine(\boldsymbol{e_i^g}, \boldsymbol{e_j^g}) \tag{4.7}$$

$$GraphSim(c_1, c_2) = Max\{GraphSim(g_i, g_j)\},$$
$$(g_i \in G_{c_1}, g_j \in G_{c_2}) \tag{4.8}$$

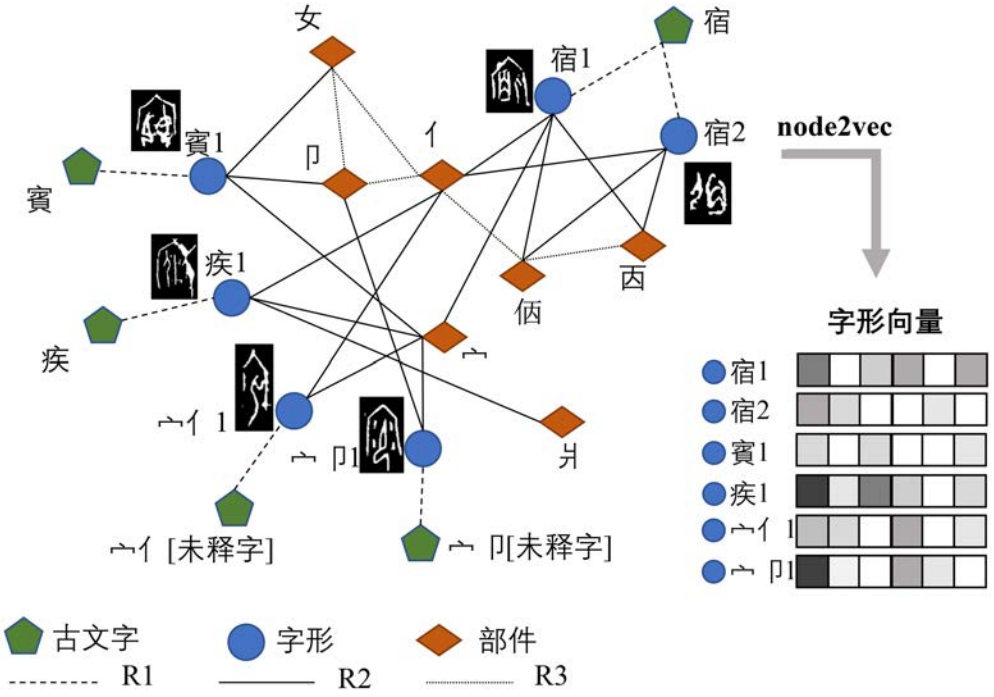


图 4.3 部件语义图谱相似度计算方法

最后，根据公式 4.9，本文直接对$PicSim$、$RLCSSim$、$GraphSim$三种相似度分数加权求和来获取两个文字（c_1，c_2）之间最终的字形相似得分$GlyphSim(c_1, c_2)$，其中γ_1、γ_2、γ_3是三种分数的权重，它们的取值大于 0 小于 1。

$$\begin{aligned} GlyphSim(c_1, c_2) \\ &= \gamma_1 RLCSSim(c_1, c_2) + \gamma_2 PicSim(c_1, c_2) \\ &+ \gamma_3 GraphSim(c_1, c_2) \end{aligned} \tag{4.9}$$

4.4 实验与评估

4.4.1 实验数据集

本实验的数据源为 ZiNet 中的一部分甲骨文数据，共包含 2543 个甲骨文文字，其中 1260 个文字为已释字，1283 个文字为未释字，这些文字共包含 2912 个字形，586 个部件种类和 15175 个拓本图像，平均每个字形拥有 5.2 个拓本图像。由古文字领域专家对这些甲骨文的相似信息进行标注，建立两个数据集：

（1） **甲骨文字对相似性得分数据集：**该数据集包含 5400 个甲骨文字对，以

及它们对应的人类专家标注的相似性分数序列：$HSimScoreSeq\{Score(c_{11},c_{12}),Score(c_{21},c_{22}),\ldots,Score(c_{i1},c_{i2})\},i=5400$。5400 个甲骨文字对是从 2543 个甲骨文中随机抽取的，专家们被要求按照他们的标准对每一对甲骨文字的字形相似性进行评分，打分范围是 0 到 10 的整数，两个文字字形的相似程度越高，则得分越接近 10，反之越接近 0。三位专家参与了这项标注工作，我们选择他们标注的中位数作为每对甲骨文字的最终得分。

（2） **甲骨文相似文字数据集：**该数据集包含人类专家标注的 6405 个相似甲骨文字对：$HSimSet\{(c_1,c_{11}),(c_1,c_{12}),\ldots,(c_i,c_{ik})\},i=2543,k\leq 5$。专家被要求为 2543 个甲骨文字中的每一个提供不超过 5 个他们认为最相似的其它文字，候选文字同样来自该甲骨文字集。在标注过程中，由一位领域专家首先注释相似的文字，然后，由另一位专家进行了验证，删除他认为不正确的文字，最终得到 6405 个相似甲骨文字对，每个甲骨文字平均被标注了 2.5 个相似文字。

4.4.2 评价指标设置

本文设置了三个评价指标：

一．人类与模型相似性打分的相关性系数（Correlation）：本文使用斯皮尔曼相关性系数（Spearman’s Correlation）来评估甲骨文字对相似性得分数据集中 5400 组专家标注的相似度分数与本文方法计算得出的相似度分数之间的相关性，两组数据的正相关性越强，则该值将更接近 1，负相关性越强，则更接近-1，设 $MSimScoreSeq$ 为与人类标注的 $HSimScoreSeq$ 对应的由本文方法给出的相似度分数序列，则相关性系数为：

$$Correlation=Spearman(HSimScoreSeq,MSimScoreSeq) \tag{4.10}$$

二．Top-n 相似文字对甲骨文相似文字数据集的覆盖率（Coverage）：本文的模型为数据集中所有 2543 个甲骨文字的组合计算相似度分数，并按照得分高低排序，得到为每个甲骨文推荐的前 n 名（top-n）相似文字，实验中 n 的取值为 5，10，20，50，100，200。本文计算模型为所有 2543 个甲骨文推荐的 top-n 相似文字与甲骨文相似文字数据集中 6405 个相似字对的重合比例，该指标旨在评估用户需要浏览多少信息量才能查询到相似的文字。设本文方法给出的 top-n 相

似文字对的集合为：$MSimSet\{(c_1,c_{11}),(c_1,c_{12}),\dots,(c_i,c_{in})\},i=2543,n=(5,10,20,50,100,200)$，则覆盖率的计算公式为：

$$Coverage=\frac{|HSimSet\cap MSimSet|}{|HSimSet|} \tag{4.11}$$

三．定性分析：本文通过相似字形推荐的案例和可视化分析，对方法进行评价：

（1） 甲骨文 top-5 相似字推荐案例分析。本文为 6 个包含独体字和合体字在内的甲骨文提供 top-5 相似甲骨文字的推荐结果，通过分析相似文字的整体形态与部件语义是否具有相似性，展示该方法捕获字形轮廓相似度与部件语义级别的相似度的能力。

（2） 甲骨文字形向量可视化分析，本文为甲骨文数据集中所有 2912 个字形生成高维向量并进行可视化，以此来分析方法对部件语义特征的表示能力。

4.4.3 模型与参数设置

实验对两种基线方法、本文提出的三种字形相似度计算方法以及它们的组合方法进行了测试，具体包括：

（1） ResNet

（2） RadicalGraph

（3） PicSim

（4） RLCSSim

（5） GraphSim

（6） γ_{11}RLCSSim+γ_{12}PicSim

（7） γ_{21}RLCSSim+ γ_{22}GraphSim

（8） γ_1RLCSSim+ γ_2PicSim+ γ_3GraphSim

其中，方法（1）与方法（2）是本文的基线方法。由于很多已有的基于笔画、结构、拼音等现代汉字属性特征的字形相似度计算工作无法应用于古文字数据，因此本文使用仅利用图像数据的视觉方法以及仅利用部件信息的知识图谱方法来进行对比。其中方法（1）ResNet 是仅基于图像相似度的基线方法， ResNet 的计算过程与 PicSim 相同，但是在最后的字形向量余弦相似度计算时不区分是否包含相同或语义相关部件的字形对，即不计算权重α（公式 4.2）。设置该基线的

目的是验证仅基于图像特征的字形相似性度量的有效性。方法（2）RadicalGraph 是仅基于知识图谱的基线方法，与本文提出的 GraphSim 方法不同，该知识图谱遵循 Liu 等人提出的偏旁部首知识图谱结构[7]，没有包含本文提出的古文字部件之间的语义关系。需要注意的是，Liu 等人的偏旁部首知识图谱在字形与部件之间建立“主要组成”与“次要组成”两种关系，以此来确定对视觉相似影响更大的部件，但是古文字数据无法提供该信息，因此本文将其简化为单一的组成关系，生成字形嵌入的方式与 GraphSim 方法相同。

方法（3）、方法（4）和方法（5）对应本文提出的三个字形相似度度量方法。方法（6）、方法（7）和方法（8）为本文提出的三种方法的组合方法，目的是验证三种方法是否具有互补性，以及怎样组合使用可以获得最好的效果。

γ是取值大于 0 小于 1 的超参数，本实验将γ_1、γ_2、γ_3分别设置为 0.4、0.3 和 0.4，将γ_{11}、γ_{12}、γ_{21}、γ_{22}均设置为 0.5。PicSim 方法的超参数α设置为 0.4，RLCSSim 方法的超参数θ设置为 0.7。为设置这些超参数，本文额外注释了 100 个相似古文字对作为校验集，对多个候选参数组合进行计算，得到使 top-10 覆盖率最大的参数组合。实验使用的 ResNet 网络层数为 18，批量大小（Batch Size）为 64，学习率（Learning Rate）为 0.001，网络经过 90 个 Epoch 的训练。本文利用 OpenNE[5]工具实现 node2vec 算法，采用工具的默认参数，输出的字形向量的维度为 50。

4.4.4 定量实验结果与讨论

七种方法的 Spearman 相关性得分如表 4.1 所示，从结果中可以分析出以下结论：（1）本文提出的三种相似性度量方法均显示出正相关性，其中 PicSim 为 0.3241，RLCSSim 为 0.8188，GraphSim 为 0.7763，证明了三种方法的有效性。（2）通过对三种方法进行组合可以提升相关性分数，其中 RLCSSim+ PicSim+ GraphSim 与人类注释的相关性最强，得分为 0.8422，这说明本文提出的几种古文字字形相似性度量方法具有互补性，融合图像、部件描述和部件语义图谱的特征可以得到更好的效果。（3）基于部件的 RLCSSim 与 GraphSim 方法的得分高于基于视觉图像的 PicSim 方法，这说明部件语义对于古文字专家来说是非常重

[5] https://github.com/hengdos/OpenNE

要的字形特征，对字形相似性度量起到重要作用。（4） ResNet 基线方法没有显示出正相关性，相关性得分为-0.0248，PicSim 的相关性得分相对于后两个基于部件的方法也显著降低，这说明仅基于从图像中提取的特征不足以表示古文字字形并进行相似性计算，因为虽然古文字的象形程度更高，但是符合象形造字法的汉字所占的比例很小，仅局限于一些独体字，大部分的汉字为由部件组成的会意字或者形声字，很多古文字虽然整体的形态轮廓相似，但其部件组成却毫无关系，因此不会被领域专家判断为相似字形。（5）RadicalGraph 基线方法的相关性得分为 0.7016，低于 GraphSim 方法，说明本文在知识图谱中加入部件之间的语义关系有益于字形相似度度量，与古文字专家的认知相同。

表 4.1 斯皮尔曼相关性系数实验结果

方法	Correlation	
	Score	**p-value**
ResNet	-0.0248	>.05
RadicalGraph	0.7016	<.005
PicSim	0.3241	<.005
RLCSSim	0.8188	<.005
GraphSim	0.7763	<.005
RLCSSim+ PicSim	0.7614	<.005
RLCSSim+ GraphSim	0.8391	<.005
RLCSSim+ PicSim+ GraphSim	**0.8422**	<.005

覆盖率指标的结果如表 4.2 所示：从结果的分析中可以得出以下重要结论：（1） 实验结果可以证明本文方法的有效性， top-5 的覆盖率最高可以达到 59.75%，top-100 最高可以达到 95.08%，top-200 为 97.86%。这说明在实际的相似古文字推荐应用中，top-100 的推荐即可覆盖绝大部分的相似古文字。（2）在小推荐量（top5-top10）的情况下，RLCSSim+GraphSim 的效果最好，而在推荐量较大（top20-top200）的情况下，CLCSSim+PicSim+GraphSim 的效果最好。这是因为 PicSim 有利于表示不包含部件的独体象形文字，在推荐量较大时提升了覆盖率。

综上所述，实验证明了本文提出的三种方法的有效性，也证明三种方法各有优势，可以通过组合来提升总体的字形推荐效果。部件系统是人类专家考虑的字形相似性的重要指标，有必要表示和计算古文字的部件语义的潜在关系，而不是仅计算文字的视觉相似性。在实际应用的大多数场景中，CLCSSim+GraphSim 是最佳选择。在较大规模的推荐场景中，可以选择 CLCSSim+PicSim+GraphSim，而在规模较小的推荐场景中，CLCSSim 是一种方便、简单的推荐方法。PicSim 更适合于独体的象形文字。

表 4.2 Top5-Top200 相似文字覆盖率实验结果（%）

方法	Top5	Top10	Top20	Top50	Top100	Top200
ResNet	11.55	13.74	16.71	22.26	28.79	43.45
RadicalGraph	52.07	61.28	69.93	79.14	85.42	88.93
PicSim	19.53	24.03	29.74	41.25	50.27	59.25
RLCSSim	52.63	65.21	74.91	86.15	91.83	95.93
GraphSim	53.90	64.84	74.96	85.92	91.69	96.03
RLCSSim+PicSim	42.39	52.51	64.59	78.61	87.63	94.53
RLCSSim+GraphSim	**59.75**	**70.37**	78.86	88.70	93.99	97.38
RLCSSim+PicSim+GraphSim	57.13	69.49	**79.75**	**89.41**	**95.08**	**97.86**

4.4.5 形近字推荐案例分析

图 4.4 展示了本文提出的方法 GlyphSim（RLCSSim+PicSim+GraphSim）、以及 ResNet 与 RadicalGraph 两个基线方法对“刀”、“鼎”、“月”、“降”、“宿”、“牢”六个甲骨文字的 top-5 相似字形推荐结果，前三例是独体的象形文字，后三例是由多个部件组成的合体文字。为便于分析，本文为每一个文字图像在下方标注了它的甲骨文部件构成，例如“宿[宀人丙]”表示该甲骨文对应的现代汉字是“宿”，甲骨文的部件组成为“宀”、“ 人”、“丙”，如果该甲骨文是一个未释字，则用“-”表示。

从案例中可以看出，对于独体的象形字来说，本文的 GlyphSim 方法可以根据 PicSim 方法捕捉到视觉形态相似性进行推荐。例如“刀”与“刃”、 “刀”与“亡”、 “鼎”与“具”、 “鼎”与“鬲”、“月” 与“夕”等甲骨文均具

有视觉上的相似性，它们的本义也基于象形文字所描绘的图案彼此相关，例如“刀”的相似文字多与刀口和切割行为有关，“刃”的本义为刀刃，“亡”的本义为锋芒，而“鼎”的相似文字大多与祭祀容器和食物有关，“具”的本义为准备（食物），“甗”是一种蒸食物的炊器。一些相似字形在文字演变方面具有潜在的关系，例如，“鼎”与“貞”、“月”与“夕”在甲骨文时期源于同一个字，“舟”与“月”在作为部件时随着文字演变发生替代关系，一些汉字，如“前”、“服”、“朕”等，在古文字时期的部件是“舟”，但在后来演变为“月”。

图 4.4 甲骨文 top-5 相似字形推荐案例

对于合体字来说，本文提出的 GlyphSim 方法可以捕捉部件语义的相似性，从而对其相似文字进行推荐。对于“宿”字，本文推荐的五个相似甲骨文字均包含部件“宀”，且内部也多包含与“人”相关的部件，表示人在家中进行某些活动的含义，例如“疾”字包含部件“宀、人、丬”表示人卧病在床，本义为伤病。“降”字的本义为从高处走下来，它的相似文字均包含部件“阜”，与山相关，本文捕捉到的第一个相似文字为“陟”，本义为“攀登”，其构型和含义与“降”相反。“牢”的本义为家畜的栏圈，其甲骨文由部首“宀”和“牛”组成，本文捕捉到的它的三个相似的文字也由“宀”和某种动物组合而成，例如，最相似的字符“廄”的甲骨文包含“宀”与“馬”，其本义为马舍，“家”字包含“宀”与“豕”，会意为在家中豢养牲畜。

与 ResNet 方法进行比较，可以看出，ResNet 方法仅基于视觉特征，所推荐的一些形近字在整体轮廓上具有视觉相似性，但是部件之间并无语义联系，因此推荐结果很多不符合古文字专家的需求，例如为“刀”的推荐的“允”（象人突显头部，表示点头允许）、“午”（杵的初文，木棒）等字、以及为“宿”推荐的未释字“[宜刀]”等。与 RadicalGraph 方法进行比较，可以看出，RadicalGraph 方法推荐了很多包含相同部件的形近字，但是由于缺乏对视觉特征的表示，推荐的字形在视觉相似方面不如本文方法，此外 RadicalGraph 未对部件语义进行建模，从结果可以看出，虽然其为 “牢”字推荐的形近字包含“宀”或“牛”相同的部件，但是没有捕捉到“宀”+某一动物的会意造字逻辑，而本文的方法为“牢”字推荐“廄”、“家”、“ 宀犬”等形近字，通过对比可以总结本文 GlyphSim 方法的优势：

（1） 会意字、形声字等合体古文字占比超过 70%，古文字的部件摆放位置相对灵活随意，因此部件构成相近的文字视觉形态上未必相似，反之亦然，这使得视觉单模态的字形相似度度量难以符合古文字专家的需求。GlyphSim 通过对部件特征进行表示与计算，可以解决这一问题。

（2） 通过在知识图谱中对部件的语义关系进行表示，GlyphSim 能够捕捉到更多具有语义关联的文字，这类相似字对于古文字领域专家的研究更有价值。在同样包含“宀”的情况下，“牛”与“馬”、“豕”这些表示动物的部件语义相似度要高于“人”等其他部件，因此本文为“牢”字推荐了“廄”、“家”等更为相似的文字，而非“宿”、“疾”等。

4.4.6 字形向量可视化分析

图 4.5 是对使用 GraphSim 方法生成的 2912 个甲骨文字形的高维向量的可视化。图中可以发现 30 个明显的字形簇，图中每个簇的标签是该簇中古文字的共同组成部件。可以看出，同一簇中的部件具有语义相关性，例如第 8 组的中心部件“止”、“ 彳”、“走”、“ 夂”均与“行走”相关；第 13 组的部件“戈”、“刀”、“斤”、“戊”、“戌”、“我”和“殳”都是古代武器；第 21 组“皿”、“鼎”、“食”、“酉”与器皿和食物有关；第 27 组的中心部件“木”、“屮”、“禾”、“葉”、“生”均与植物有关。此外，不同聚类簇之间也存在语义相似性。例如，第 1 组（“女”、“母”、“每”）、第 2 组（“人”）、第 3 组

（“子”）和第 4 组（“大”）都是表示人类的部件，在空间上非常接近。第 19 组（“魚”）、20 组（“牛”、“羊”、“犬”、“ 豕”、“鹿”）、22 组（“隹”、“鳥”）、23 组（“皃”）、24 组（“龜”）和 25 组（“虎”）的部件都与动物相关，它们在空间中非常接近，21 组聚类（“皿”、“鼎”、“食”、“酉”）表示食物和器皿，与这些动物簇也非常接近。以上可以证明本文提出的部件语义图谱可以表示古文字的部件语义特征，部件系统相似的文字在空间中更接近。

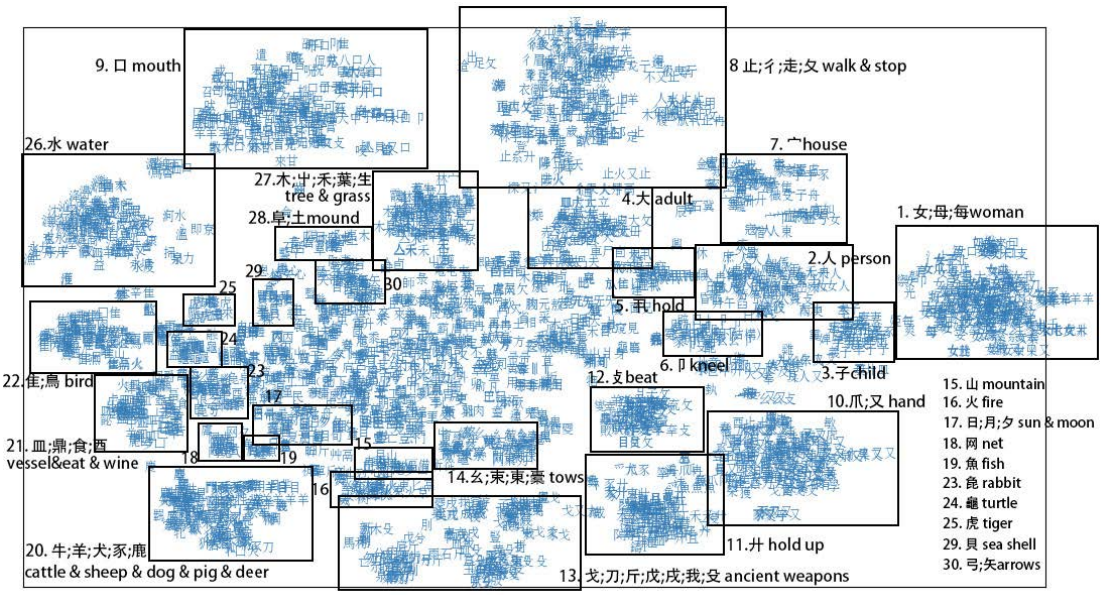


图 4.5 甲骨文字形向量可视化展示

4.5 本章小结

本章介绍了古文字字形相似度的三种计算的方法，包括象形相似度、部件描述相似度和部件语义图谱相似度。本文在甲骨文数据集上分别对三种方法以及它们的组合方法进行了测试，结果证明了本文方法的有效性，三种方法的相似度分数与专家标注的分数都具有正相关关系，并且三种方法具有互补性，通过组合可以获得最好的效果，证明仅使用古文字图像的视觉特征不足以完整描述古文字字形，部件特征对于古文字领域专家来说也非常重要。最后本文对甲骨文字的 Top-5 相似字形推荐结果进行了定性分析，证明本文的方法可以捕捉到部件语义方面相似的古文字。本章介绍的古文字相似度度量方法具有很强的应用创新性，可以实现古文字相似字形的推荐应用，服务于古文字领域的信息检索和研究。

第 5 章　部件语义知识驱动的古文字字形识别方法

在古文字考释研究中，领域专家需要对已知文字的演变进行归纳，并对历史上所有相似的字形进行比较分析，最终推理出未知字形对应的现代汉字。然而，历史上存在数以万计的字形，同一个文字在不同的历史节点也拥有很多异体字形，它们的部件和形态有着复杂的联系，这一研究挑战着专家的脑力极限。AI 辅助的古文字考释研究尚处于起步阶段，相关工作基于图像生成算法，生成甲骨文图像对应的现代汉字图像，启发专家进行考释。这些方法的局限性在于仅利用图像模态的特征，生成的汉字图像缺乏造字逻辑，并且很多文字发展至今，部件组成发生了很大改变，已有方法对于这些文字的效果有限。本章提出古文字字形识别的创新任务，旨在融合图像与部件信息，将未知字形分类到所属的文字标签。具体来说，本章构建部件语义知识图谱（CGKG），关联不同历史时期的汉字字形，并提出了一个融合图像、部件描述文本和部件语义图谱特征的多模态汉字字形识别模型（MCGI），模型显式地表示汉字部件构成特征，并对其演变进行归纳，在已知字形上进行训练，对未知字形的汉字标签进行预测。实验是在一个跨越 1000 多年的真实古文字字形数据集上设计的，证明方法的有效性，并报告每种模态信息对这项任务的贡献，本工作可以为古文字的自动破译提供初步参考。本章 5.1 节为引言，5.2 节介绍异体字形的分类以及字形识别任务定义，5.3 节将介古文字字形识别方法的细节，5.4 节介绍实验设计、数据集、实验结果与讨论等。5.5 节对本章的工作进行总结。

5.1 引言

古文字阶段，汉字的书写尚未统一，同一个文字经常拥有多种异体字形，并且随着发展，文字的形态也会发生较大的变化。图 5.1 展示了汉字“春”在甲骨文、金文、楚简、小篆、隶书时期的多种异体字形和它们的部件组成。尽管这些异体字形在视觉形态上有很大不同，但它们的部首组成之间存在着潜在的语义联系，它们的部件系统都通过描绘“植物在阳光下生长”的画面来表达“春天”的含义。因此，许多字形包含“屮”、“木”、“生”这些表示植物的部件，一些字形包含部件“日”来表示太阳，几乎所有字形都包含表示字音的部件“屯”，“屯”与“春”发音相同。

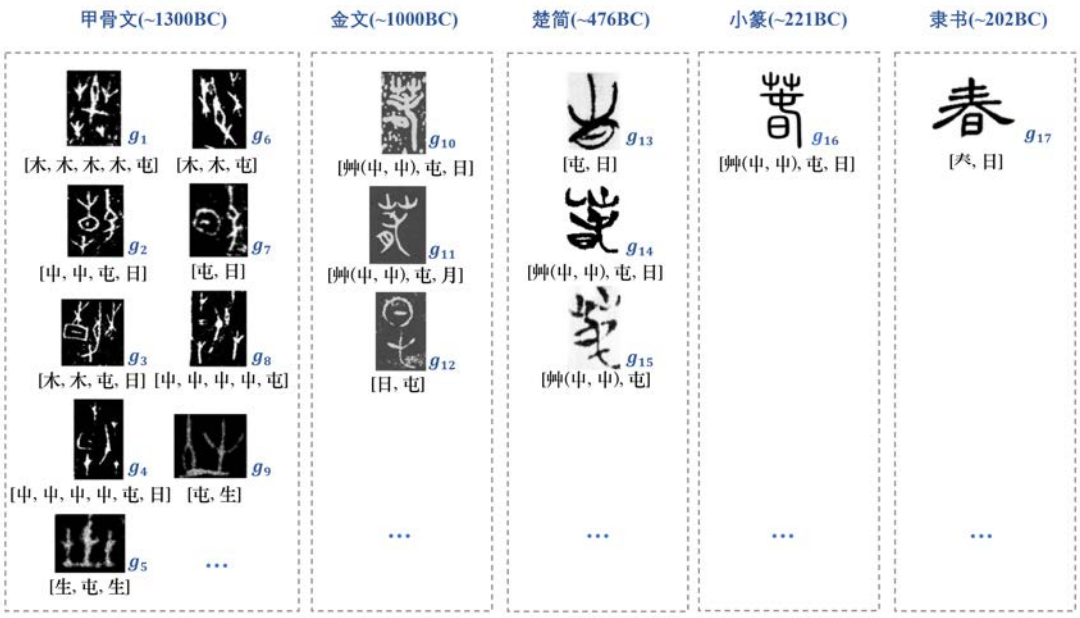


图 5.1 “春”字在甲骨文、金文、楚简、小篆、隶书时期的部分异体字形

目前，很多古文字字形未被破译，这意味着尚不确定它们所对应的现代汉字，专家们必须对历史上相关的字形进行比较分析，尝试根据已知文字的演变逻辑进行推理，判断它们是否属于同一个文字。然而，在不同的历史阶段，分布着数以万计的字形，它们的部首和形态之间有着复杂的关系，这使得古文字考释工作具有挑战性，严重依赖于专家的记忆和经验。目前的 AI 辅助的古文字考释研究尚处于起步阶段，相关研究利用图像生成算法模拟汉字图形的历时演变，输入古文字图像，生成与之对应的现代汉字的图像，从而启发专家进行考释研究[8-9]，或者对古文字的图像进行部件分解并重构为现代汉字对应的部件[10]。这些方法的局限性在于仅利用图像模态的特征，生成的汉字图像的随机性很大，大多不符合造字逻辑，并且很多文字发展至今，部件组成发生了很大改变，因此才难以被古文字专家破译，已有方法对于这些文字的识别效果有限。

基于以上问题，本文认为古文字破译需要同时考虑汉字的视觉形态和部件语义信息，因此本文提出一项创新性的古文字字形识别任务，旨在融合图像与部件语义特征，将未知字形分类到所属的文字标签，推荐概率分数最高的文字，从而启发古文字专家进行古文字考释。多模态融合的相关工作分布在视觉语言模型、多模态知识图谱领域，遵循通用的工作流程：通过预训练的单编码器，如 BERT[39]，ResNet[124]，GAT[110]等为每种模态的数据生成嵌入，随后这些嵌入被馈入到多模态编码器进行融合，多模态编码器可以使用简单的点积运算、多模态注意力机制

或更复杂的变换器。ViLT[126]采用浅嵌入层，并在文本和视觉嵌入交互中使用深层次的 Transformer。HRGAT[42]提出了一种基于文本和视觉共同关注的多模态融合方法。IMF[41]提出了一种用于链路预测的两阶段交互式多模态融合框架。本文针对古文字领域数据的特点，提出了一种由部件语义知识驱动的多模态方法，该方法利用汉字字形知识图谱（CGKG）来表示所有字形与部件之间的专业关系，并提出融合视觉图像、部件描述文本和图谱特征的多模态字形识别模型（MCGI），该模型在已知字形上进行训练，对未知字形的汉字标签进行预测。

具体来说，本文的贡献包括：（1）提出古文字字形识别任务。（2） 提出融合视觉图像、部件描述文本和图谱特征的多模态字形识别模型（MCGI）。（3）在一个跨越 1000 多年的字形数据集上从共时和历时两个角度对该方法进行了评估，测试样本大部分为部件构成发生变化的字形，实验结果显示，本文提出的方法对共时字形识别的准确率可以达到 73.7%，对历时字形识别的准确率为 56.5%，通过与基线方法进行比较，证明该方法的有效性，并给出各模态信息对该任务贡献的结论。方法最大的创新之处是引入了汉字的部件语义知识，这与汉字的形声、会意的造字方法以及人类专家的考释方法是一致的，即根据部件对语义和发音相关的字形进行比较与关联，并且由于模型学习到异体字形部件语义的共同特征，对于部件结构发生一定变化的未知字形也具有识别能力，可以为未来古文字的破译提供初步的参考。

5.2 任务定义

5.2.1 异体字形相关知识与分类

异体字是指一组含义、发音、用法均相同但书写形态不同的字形，在本文的任务中，它们属于同一个汉字类别标签。在现代汉字中，异体字的代表案例是繁体字（如“鳥”）和相应的简体字（如“鸟”）。在古文字阶段，汉字系统还未发展成熟，由于书写的灵活性以及地理和政治的隔绝，古文字的异形现象非常常见。如图 5.2 所示，本文根据字形识别任务的需求对异体字形进行更进一步的界定，对属于同一文字的任一一组字形（g_1，g_2），如果字形g_1与字形g_2满足以下任意一种情况，则本文认为它们互为异体字形：

（1） 部件种类不同：异体字形包含不同类别的部件，如图 5.2（a）。

（2） 部件数量、位置或部件字形不同：异体字形的部件种类相同，但是数

量、分布位置或者部件的形态不同，如图 5.2（b，c，d）。

（3） 古文字类型或历史时期不同：例如甲骨文、金文和楚简文字是三种不同的古文字类型，如图 5.2（e）。

此外，本文按照异体字形是否属于相同的古文字类型，对其分为共时异体字形和历时异体字形两类：

（1） 共时异体字形：如果一组异体字形属于相同的古文字类型，本文称它们为共时异体字形，如图 5.2（a，b，c）。

（2） 历时异体字形：如果一组异体字形不属于同一种古文字类型，本文称它们为历时异体字形，如图 5.2（d，e）。

相应的，在后文的实验中，本文也将测试集中的待预测字形分为共时测试字形与历时测试字形两类，分别进行评估，因为后者更难识别，并且在实际应用中很常见：

（1） **共时测试字形：**如果训练集中至少存在一个该字形的共时异体字形，则将该字形称为共时测试字形。

（2） **历时测试字形：**如果训练集中不存在该字形的共时异体字形，则称该字形为历时测试字形。

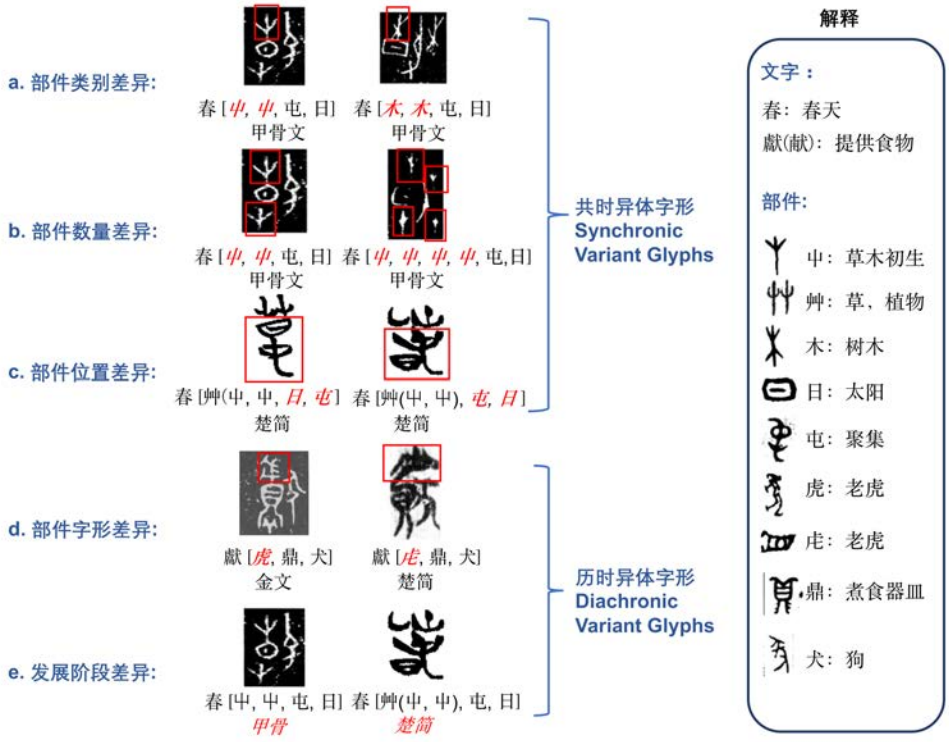


图 5.2 异体字形分类图例

5.2.2 古文字字形识别任务定义

古文字字形识别任务如图 5.3 所示：

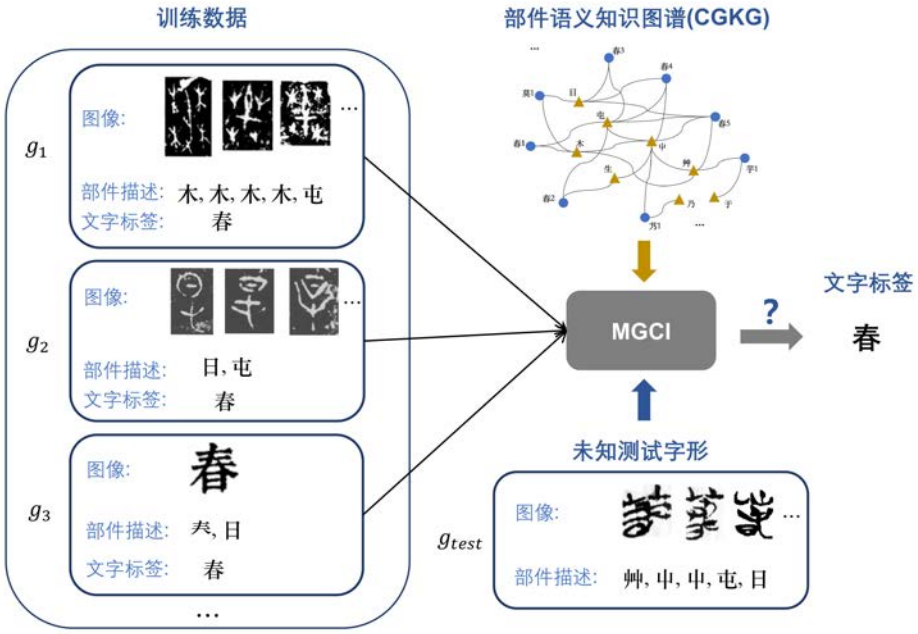


图 5.3 古文字字形识别任务

设有一个包含文字类别标签的集合C：$C = \{c_i \mid i = 1，2，\ \ldots，\mid C \mid\}$，$c$表示文字；一个字形训练集合$G_{train}$：$G_{train} = \{g_i \mid i = 1，2，\ \ldots，\mid G_{train} \mid\}$，包含所有已知的分布在任何历史阶段的汉字字形，$g$表示字形；一个字形测试集合$G_{test}$：$G_{test} = \{g_i \mid i = 1，2，\ \ldots，\mid G_{test} \mid\}$，包含古文字阶段的未知的待测试的字形样本，$G_{train} \cap G_{test} = \emptyset$。每个字形属于一个汉字类别标签。因此，该任务的训练数据集为：$\{(g, c) \mid g \in G_{train}，\ c \in C\}$，相应的，测试数据集为：$\{(g, c) \mid g \in G_{test}，\ c \in C\}$。

本任务中的字形由图像和部件文本两种数据描述。对每个字形g_i，有一个或多个图像，其图像集合定义为P_{g_i}，$P_{g_i} = \{p_k \mid k = 1，2，\ \ldots，\mid P_{g_i} \mid\}$，$p$代表图像。还有一个部件序列$R_{g_i}$，$R_{g_i} = \{r_k \mid k = 1，2，\ \ldots，\mid R_{g_i} \mid\}$，它包含该字形按书写顺序排列的部件，$r$代表部件。

设有一个字形知识图谱KG，$KG = (V，\ Rel)$，$V = V_g \cup V_r$，其中V是图的节点集合，图中包含字形与部件两种实体节点，V_g是字形实体的节点集合，V_r是部件

实体的节点集合。Rel是节点之间的边集。KG包含G_{train}中的所有字形、它们的部件以及它们之间的边。

本任务的目标是将未知字形g，$g \in G_{test}$，分类到对应的文字标签c中：$f(P_g, R_g, KG\ |\theta) = c$，$c \in C$，其中$\theta$是通过训练得到的模型参数，所有的监督学习都基于训练数据集和KG实现。

5.3 古文字字形识别方法

5.3.1 模型框架概述

古文字字形识别模型（MCGI）的体系结构如图 5.4 所示，给定一个目标字形 g_k，模型的输入包括 g_k的图像集合P_{g_k}、部件描述文本序列T_{g_k}，T_{g_k}是R_{g_k}对应的文本，以及字形知识图谱KG。

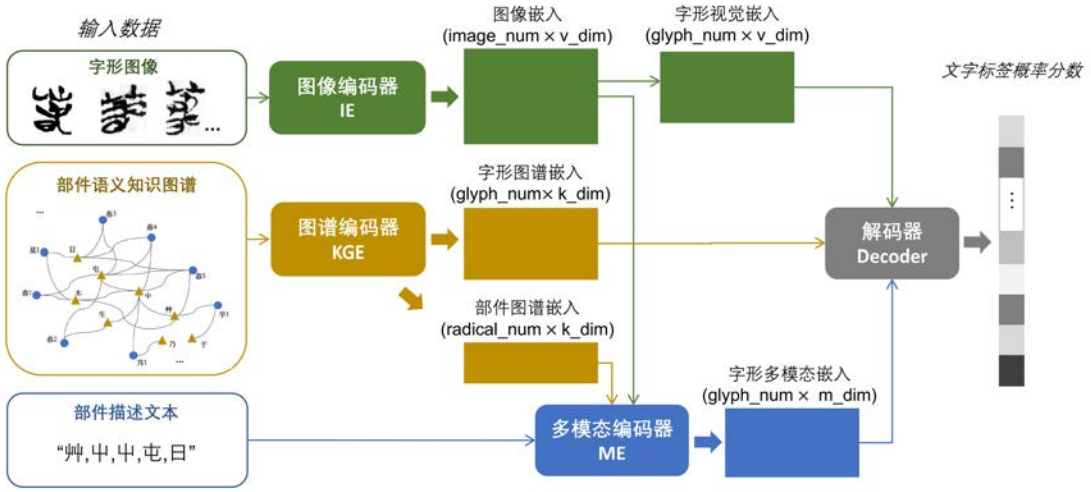


图 5.4 古文字字形识别模型框架

MCGI 由三个编码器模块和一个解码器模块组成，编码器用于提取字形 g_k 各个模态的特征，生成高维向量表示，包括：（1）图像编码器（IE）：从P_{g_k}的图像中提取视觉特征，最终输出字形g_k的高维视觉嵌入$\boldsymbol{e}_{g_k}^{\boldsymbol{v}}$。（2）图谱编码器（KGE）：从字形知识图谱$KG$中提取特征，输出字形$g_k$的高维图谱嵌入$\boldsymbol{e}_{g_k}^{\boldsymbol{kg}}$，以及所有部件实体的高维嵌入$\boldsymbol{e}^{\boldsymbol{r}}$。（3）多模态编码器（ME）：从部件描述文本序列$T_{g_k}$中提取特征，并融合 IE 和 KGE 编码器生成的视觉嵌入和部件嵌入，最终输出字形g_k的多模态嵌入$\boldsymbol{e}_{g_k}^{\boldsymbol{m}}$。解码器对三个编码器生成的字形表示向量进行解

码，最终输出目标字形 g_k 与文字类别标签之间的概率分数：$Decoder\left(\boldsymbol{e}_{g_k}^{\boldsymbol{v}}, \boldsymbol{e}_{g_k}^{\boldsymbol{kg}}, \boldsymbol{e}_{g_k}^{\boldsymbol{m}}\right) = \{s(g_k, c_i) | i = 1,2,\ldots,|C|\}$，$s(g,c)$是字形和文字标签之间最终的概率分数。

本章 5.3.2 节将介绍汉字字形知识图谱的结构，第 5.3.3 节将介绍图像编码器（IE），第 5.3.4 节将介绍图谱编码器（KGE），第 5.3.5 节将介绍多模态编码器（ME），第 5.3.6 节将介绍解码器，最后第 5.3.7 节介绍损失函数以及 MCGI 的工作步骤。

5.3.2 汉字字形知识图谱

汉字字形知识图谱（CGKG）是基于 ZiNet 构建的，结构如图 5.5 所示。

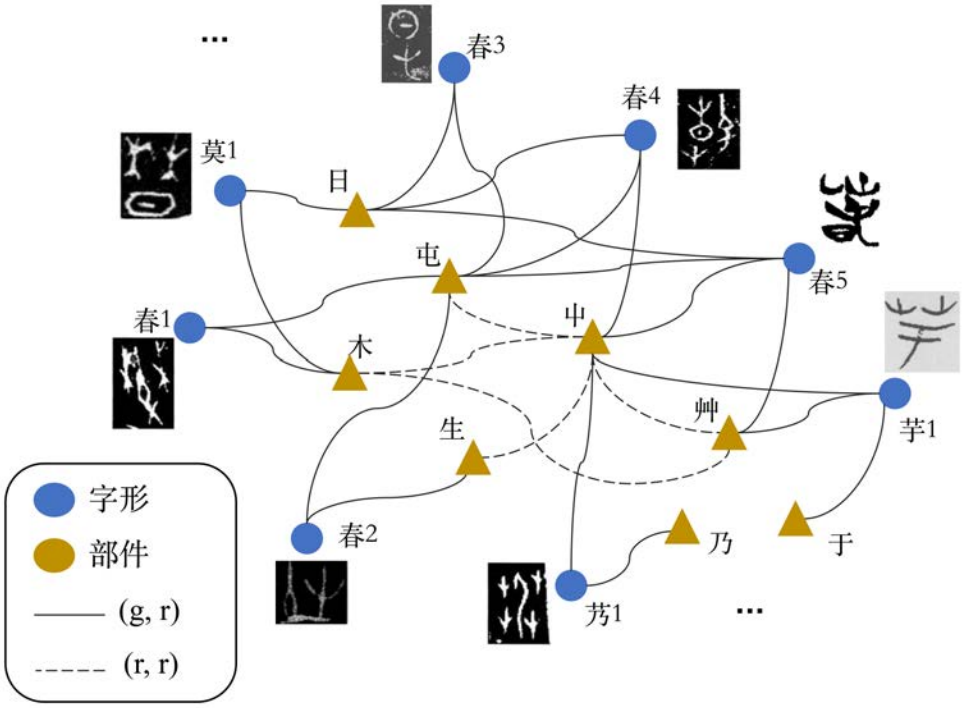


图 5.5 汉字字形知识图谱图例

CGKG 包含字形和部件两种类型的实体以及两种关系：$R1(g, r)$，$R2(r, r)$，$R1$描述字形和部件之间的包含关系，如果某一字形节点包含某部件节点，则在它们之间建立一条边，这里包括任一级别的部件；$R2$描述部件之间的语义关系，如果两个部件之间存在以下关系，那么在这两个部件之间建立一条边：（1）它们具有包含关系，例如“艸”与“中”；（2）它们属于同一个字符，例如 “虎”与“虍”；（3）它们具有派生关系，例如 “東”与“束”；（4）它

们具有指示关系，例如 “生”与“屮”；（5）它们表达相同或相似的含义，例如 “屮”与“木”；（6）它们的发音相同或相似，在古代汉语中通常可以相互借用（通假），例如 “匕”与“比”；（7）它们是通用的部件，如“屮”与“木”是在“春”字的异体字形对（“屮、屮、屯、日”， “木、木、屯、日”）中观察到的一对可互换的部件，因为除“屮”与“木”外，两个字形他部分的部件组成均相同。其中关系（1）、（2）可以从 ZiNet 中直接获取，关系（7）是从 ZiNet 提供的训练数据集自动挖掘得到的，本文遍历数据集中所有的异体字形对，提取互相替换的部件对，如果该异体字形对的部件组成之间只有该部件对发生替换，其他部分相同，则将其设置为候选部件对，最后出现 2 次以上的部件对将被认为是造字中可以通用的部件，添加到 CGKG 中。其他关系由专家注释得到。以上部件之间的这些关系不是独立的，而是高度相关的，因此本文在 CGKG 中不指定关系的类别，如果两个部件满足上述任一专业关系，本文将在它们之间添加一条边，最后构造一个无向图。最后，在 657 个部件节点之间标注了 1907 条边。

5.3.3 图像编码器模型

对于一个目标字形g_k，图像编码器（IE）为该字形的图像集合P_{g_k}中的每一个图像生成高维向量表示，图像向量的集合为$E(P_{g_k}) = \{\boldsymbol{e}_{\boldsymbol{g_k}}^{\boldsymbol{v_i}} | \boldsymbol{e}_{\boldsymbol{g_k}}^{\boldsymbol{v_i}} \in \mathbb{R}^{d_v}, i = 1,2,\dots,|P_{g_k}|\}$，$d_v$是图像特征向量的维度。最后字形$g_k$的视觉特征向量$\boldsymbol{e}_{\boldsymbol{g_k}}^{\boldsymbol{v}}$被设置为$E(P_{g_k})$中所有图像向量的平均值：

$$\boldsymbol{e}_{\boldsymbol{g_k}}^{\boldsymbol{v}} = \frac{1}{|P_{g_k}|} \sum_{p_i \in P_{g_k}} IE(p_i) \tag{5.1}$$

视觉编码器 IE 是计算机视觉模型，本文使用 ResNet 模型，IE 在图像分类任务上进行预训练，该任务将训练集G_{train}中字形拥有的全部图像样本分类到对应的文字标签中：$f_{p->c}$。

5.3.4 知识图谱编码器模型

知识图谱编码器（KGE）为汉字字形知识图谱 CGKG 中的每一个字形和部件实体生成高维向量表示，字形向量的集合为$E(KG_g) = \{\boldsymbol{e}_{\boldsymbol{g_i}}^{\boldsymbol{kg}} | \boldsymbol{e}_{\boldsymbol{g_i}}^{\boldsymbol{kg}} \in \mathbb{R}^{d_{kg}}, i =$

$1,2,\dots,|V_g|\}$，部件向量的集合为$E(KG_r)=\{\boldsymbol{e}_{r_i}|\boldsymbol{e}_{r_i}\in\mathbb{R}^{d_{kg}},i=1,2,\dots,|V_r|\}$，$d_{kg}$是知识图谱实体节点的特征向量维度。对于目标字形$g_k$，它的图谱特征向量$\boldsymbol{e}_{g_k}^{kg}$为：

$$\boldsymbol{e}_{g_k}^{kg}=KGE(KG)[k] \tag{5.2}$$

图谱编码器（KGE）是一个知识图谱表示模型，本文使用 node2vec 来初始化 CGKG 中的实体节点，然后进一步在字形实体分类任务上训练一个图注意力网络（GAT），对 KG 中的实体节点进行进一步的表示，GAT 利用 CGKG 中邻居节点的特征对实体节点的特征表示进行更新，详细介绍见 2.1.2 节。字形实体分类任务将训练集G_{train}中的字形分类到对应的文字标签中：$f_{g->c}$。

5.3.5 多模态编码器模型

本文的多模态编码器（ME）是基于 2.2 章介绍的 BERT 模型实现的，它的体系结构如图 5.6 所示。

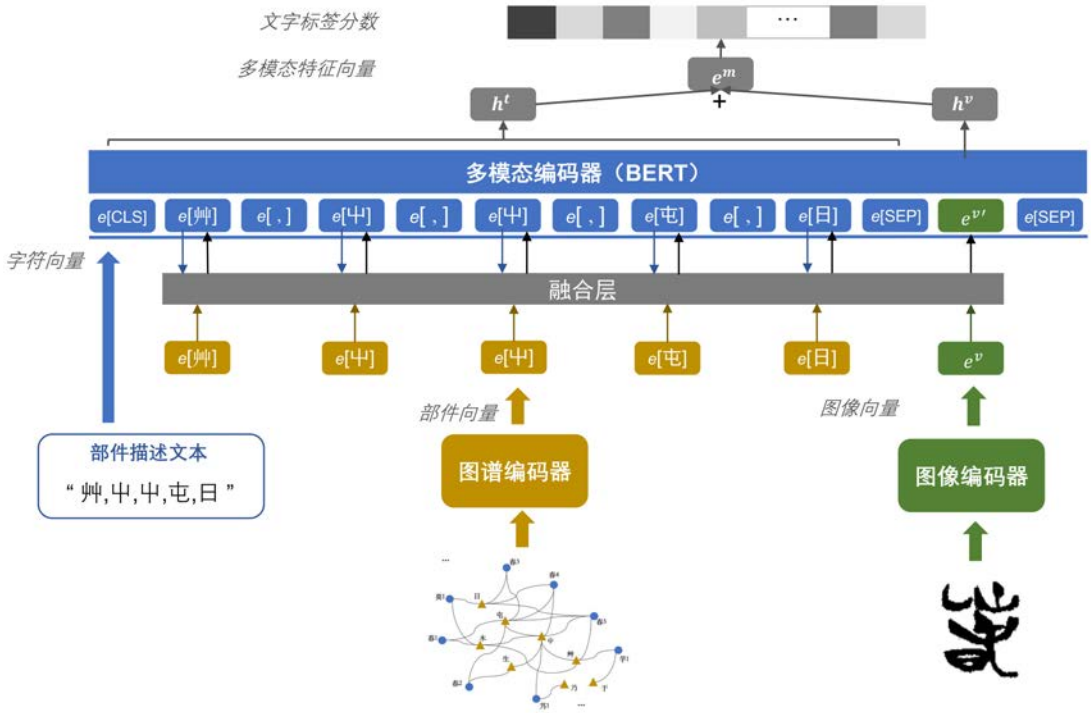


图 5.6 多模态编码器体系结构

对于一个目标字形g_k，ME 的输入包括：（1）g_k的一个图像向量$\boldsymbol{e}_{g_k}^{v_i}$，$\boldsymbol{e}_{g_k}^{v_i}\in E(P_{g_k})$；（2）$g_k$的部件描述文本向量序列$E(T_{g_k})=\{\boldsymbol{e}_{g_k}^{t_i}|\boldsymbol{e}_{g_k}^{t_i}\in\mathbb{R}^{d_t},i=1,2,\dots,|T_{g_k}|\}$，该向量是由 BERT 模型对部件描述文本字符序列T_{g_k}进行初始化得

到的，d_t是字符向量的维度；（3）g_k的部件实体向量$E\left(KG_{r_{(g_k)}}\right)$，是由 KGE 编码器生成的$g_k$包含的所有部件的特征向量。ME 编码器的输出为多模态特征向量$\boldsymbol{e}_{\boldsymbol{g_k}}^{\boldsymbol{m_i}}$：

$$\boldsymbol{e}_{\boldsymbol{g_k}}^{\boldsymbol{m_i}} = ME\left(\boldsymbol{e}_{\boldsymbol{g_k}}^{\boldsymbol{v_i}}, E\left(T_{g_k}\right), E\left(KG_{r_{(g_k)}}\right)\right) \tag{5.3}$$

由于字形g_k可能拥有不止一个图像，因此本文使用 ME 编码器得到每一个图像对应的多模态特征向量$\{\boldsymbol{e}_{\boldsymbol{g_k}}^{\boldsymbol{m_i}} | \boldsymbol{e}_{\boldsymbol{g_k}}^{\boldsymbol{m_i}} \in \mathbb{R}^{d_m}, i = 1,2,\dots,|P_{g_k}|\}$，并以该集合中所有向量的平均作为$g_k$最终的多模态向量表示$\boldsymbol{e}_{\boldsymbol{g_k}}^{\boldsymbol{m}}$，$d_m$是多模态特征向量的维度。

具体来说，对于部件描述文本序列T，本文按照从上到下、从左到右、从外到内的汉字书写顺序排列部件，以“，”为间隔符号。如果文字是独体字，则部件为其本身，如果文字的部件包含次级部件，则次级部件写在该部件之后，例如，图 5.1 中“春”的字形 g14 的部件描述文本序列T为{艸，屮，屮，屯，日}, 其中“屮，屮”为构成“艸”的二级部件，在 BERT 对T中的字符进行词嵌入初始化之前，需要在T中增加 BERT 规定的特殊符号：T={[CLS]，t_2，…，t_{n-3}、[SEP]，[IMG]，[SEP]}，其中n为T 增加符号后的文本序列长度，[CLS]、[SEP]为 BERT 规定的的起始符号和结束符号，[IMG]为图像嵌入$\boldsymbol{e}^{\boldsymbol{v}}$后续将要填充的位置，$t_2$至$t_{n-3}$为文本字符。

随后T被 BERT 初始化为向量序列$E(T) = \{\boldsymbol{e}^{\boldsymbol{t_i}} | \boldsymbol{e}^{\boldsymbol{t_i}} \in \mathbb{R}^{d_t}, i = 1,2,\dots,|T|\}$。在被送入 BERT 的编码器处理之前，在融合层需要与图像特征向量$\boldsymbol{e}^{\boldsymbol{v}}$，以及部件特征向量$E\left(KG_{r_{(g_k)}}\right)$进行融合，生成融合后的输入向量序列$E(O) = \{\boldsymbol{e}^{\boldsymbol{o_i}} | \boldsymbol{e}^{\boldsymbol{o_i}} \in \mathbb{R}^{d_t}, i = 1,2,\dots,|T|\}$，融合层的计算方法为：

$$\begin{cases} \boldsymbol{e}^{\boldsymbol{o_i}} = Linear(\boldsymbol{e}^{\boldsymbol{v}}), & i = |T| - 1 \\ \boldsymbol{e}^{\boldsymbol{o_i}} = Linear(\boldsymbol{e}^{\boldsymbol{t_i}} || \boldsymbol{e}_{\boldsymbol{r_{(t_i)}}}), & r_{(t_i)} \in V_r' \\ \boldsymbol{e}^{\boldsymbol{o_i}} = \boldsymbol{e}^{\boldsymbol{t_i}}, & Otherwise \end{cases} \tag{5.4}$$

其中，$Linear$表示神经网络线性层，||代表向量拼接操作，$r_{(t_i)}$是字符t_i对应的 CGKG 中的部件实体，$\boldsymbol{e}_{\boldsymbol{r_{(t_i)}}}$是 KGE 输出的该实体的特征向量，$V_r'$是 CGKG 中中具有部件之间语义关系$R2(r，r)$的部件实体的集合。接下来将融合后的向量序列$E(O)$馈送入 BERT 的编码器，在最后一层输出$o_{n-1}$位置对应的视觉隐藏向量

$\boldsymbol{h^v}$，以及文本隐藏向量$\boldsymbol{h^t}$，$\boldsymbol{h^t}$是o_1到o_{n-2}位置输出向量的平均。多模态字形嵌入$\boldsymbol{e^m}$被设置为二者的和：

$$\boldsymbol{e^m} = \boldsymbol{h^v} + \boldsymbol{h^t} \tag{5.5}$$

本文在字形分类任务上利用训练数据对 ME 进行预训练，$\boldsymbol{e^m}$随后将通过线性变换层和$softmax$函数最终得到文字标签的分数向量。

5.3.6 解码器模型

本文采用一个简单的无监督解码器，每一个文字标签c_i对应一个字形集合$G_{(c_i)}$，该集合中包含c_i在训练集中的全部字形：$G_{(c_i)} = \{g_j^{c_i} | g_j^{c_i} \in G_{train}, j = 1,2,\dots,m\}$，$m$是$c_i$在训练集中的字形数量。本文计算待测试字形$g_k$的字形向量与$G_{(c_i)}$中每一个字形的字形向量之间的余弦相似度，并将最大的余弦相似度值作为文字标签c_i对于字形g_k的概率分数$score(g_k, c_i)$。

由于本文的三种编码器 IE、KGE 与 ME 分别为字形g_k生成视觉特征表示向量$\boldsymbol{e_{g_k}^v}$，图谱特征向量$\boldsymbol{e_{g_k}^{kg}}$，和多模态特征向量$\boldsymbol{e_{g_k}^m}$，因此对每个向量按照上述方法计算余弦相似度，可以得到g_k与c_i之间的三个概率分数：$s^v(g_k, c_i)$，$s^{kg}(g_k, c_i)$，和$s^m(g_k, c_i)$。最后通过对三个分数计算加权平均得到最终的概率分数$s(g_k, c_i)$：

$$s = \alpha s^m + \beta s^v + \gamma s^{kg} \tag{5.6}$$

其中，α、β、γ是大于 0，小于 1 的超参数。

为了进行比较，本文也设置了一个基于监督学习的解码器，该解码器使用训练集中的数据在字形分类任务上进行训练，将字形分类至文字标签，$f_{g->c}$，最终字形g_k的隐藏向量$\boldsymbol{e_{g_k}}$被设置为：

$$\boldsymbol{e_{g_k}} = \alpha Linear(\boldsymbol{e_{g_k}^m}) + \beta Linear(\boldsymbol{e_{g_k}^v}) + \gamma Linear(\boldsymbol{e_{g_k}^{kg}}) \tag{5.7}$$

在该监督解码器中，α、β、γ是通过网络训练的到的参数，网络将$\boldsymbol{e_{g_k}^m}$、$\boldsymbol{e_{g_k}^v}$、$\boldsymbol{e_{g_k}^{kg}}$三个特征向量分别输入三个线性变换层，最终得到融合后的隐藏向量$\boldsymbol{e_{g_k}}$，最终$\boldsymbol{e_{g_k}}$经过$Softmax$函数得到最终的文字标签的概率分数。

5.3.7 算法步骤与损失函数

MCGI 遵循分阶段训练和预测步骤，训练阶段的步骤如下：

（1） 使用训练集中的图像数据对 IE 编码器进行训练，得到训练后的 IE 编码器模型、训练集中所有图像的向量表示，以及训练集中字形的视觉向量表示$\boldsymbol{e^v}$；

（2） 使用训练集数据在字形图谱 CGKG 上对 KGE 编码器进行训练，得到训练后的 KGE 编码器模型、图谱中所有字形实体节点的向量$\boldsymbol{e^{kg}}$、以及部件节点的向量表示；

（3） 使用训练集中数据对 ME 编码器进行训练，同时需要 IE 编码器和 KGE 编码器输出的图像向量和部件向量作为输入，得到训练后的 ME 编码器模型，以及训练集中字形的多模态向量表示$\boldsymbol{e^m}$。

在预测阶段，对于待分类的字形g_{test}，方法遵循如下步骤：

（1） 通过训练好的 IE 编码器获取g_{test}所有图像的向量，以及g_{test}的字形视觉向量$\boldsymbol{e^v_{g_{test}}}$；

（2） 将g_{test}添加到字形知识图谱 CGKG 中，即创建一个新的字形节点，并将其连接到相应的部件节点，训练 KGE 编码器以获取g_{test}的字形图谱嵌入$\boldsymbol{e^{kg}_{g_{test}}}$，以及所有字形实体节点的嵌入表示；

（3） 通过训练好的 ME 编码器获取g_{test}多模态字形向量$\boldsymbol{e^m_{g_{test}}}$，该步骤需要 IE 编码器和 KGE 编码器输出的g_{test}对应的图像向量和部件向量作为输入；

（4） 将g_{test}以及训练集所有字形样本的三种字形向量$\boldsymbol{e^v}$、$\boldsymbol{e^{kg}}$、$\boldsymbol{e^m}$输入到解码器中进行计算，最终输出g_{test}对所有文字标签的概率分数。

本文使用交叉熵损失函数（Cross-Entropy）对三个编码器进行训练，对于$(x，c_i)$，x是目标字形或图像，$s(x，c_i)$是x和文字c_i之间的分数，交叉熵损失函数写为：

$$loss = -s(x，c_i) + log\sum_{l=1}^{|C|}\exp\left(s(x，c_l)\right) \tag{5.8}$$

5.4 实验与评估

5.4.1 实验数据集

本文基于 ZiNet 构建了实验使用训练集和测试集，具体的统计信息见表 5.1。它共包含 6941 个文字标签和 1974 个部首。训练集中共有 14931 个字形，其中包括 2478 个甲骨文字形、2839 个金文字形、6042 个楚简字形，还有 3572 个小篆和隶书字形，这些字形共拥有 53452 个图像。测试集中共有 1279 个古文字字形，其中包括 414 个甲骨文字形、468 个金文字形、397 个楚简字形，这些测试字形共拥有 5319 个图像。

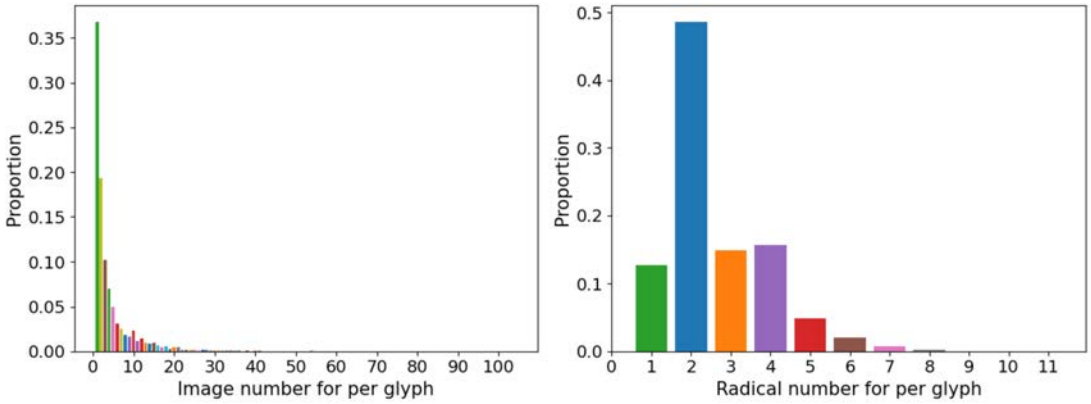
表 5.1 数据集统计信息

数据集	数据类型	统计
-	文字	6941
	部件	1974
训练集	字形	14931 [甲骨文 2478；金文 2839；楚简 6042；其他 3572]
	图像	53452
测试集	字形	1279 [甲骨文 414；金文 468；楚简 397]
	图像	5319

根据进一步统计，训练集中每个文字标签平均只有 2.15（14931/6941）个字形样本，这是古文字领域当前字形数据集的不完整性导致的。因此，在目前的研究中，测试集的字形只覆盖 805 个较为常用的文字标签，且每个测试字形在训练集中至少有一个属于同一个文字标签的异体字形。本文进一步统计了训练集中字形样本包含的图像和部件数量，每个字形包含的图像数量和部件数量的比例分布如图 5.7 所示。训练集中每个字形平均拥有 3.73 个图像，包含平均 2.62 个部件，约 60%的字形仅包含 1-2 个部件，且由于文字使用频率的差异，图像数据集具有非常明显的长尾效应，约 55%的字形仅包含 1-2 个图像样本，它符合古文字的数量在真实世界中分布，古文字领域数据的这些特点使该任务具有一定挑战。

本文对测试集中的字形样本进行分类统计，表 5.2 显示，在 1279 个测试字形中，有 598 个属于共时测试字形（定义见第 5.2.1 节），681 个属于历时测试字形，本文在实验中将对这两组字形分别进行评估，因为后者更难识别。此外，还

统计了测试字形与其训练集中异体字形的部件种类一致性情况。“部件一致”代表测试字形在训练集中至少拥有一个部件种类与之相同的异体字形，即它们包含的部件种类相同，仅在部件位置、数量、字形或文字类别方面不同（见第 5.2.1 节），而“部件不一致”组的测试字形对于其训练集中的异体字形来说，部件的种类发生了变化，包括部件种类的增减或者部件替换。经过统计，共有 1024 个部件种类不一致的测试字形，只有 255 个部件一致的测试字形，这说明模型需要学习部件构成的潜在语义特征，而不能仅依靠部件匹配实现字形识别。



（a）字形样本的图像数量比例分布　　（b）字形样本的部件数量比例分布

图 5.7 训练字形样本的图像和部件数量比例分布

表 5.2 测试集字形的分类统计

	部件一致	部件不一致	总和
共时测试字形	145	453	598
历时测试字形	110	571	681
总和	255	1024	1279

5.4.2 基线方法设置

本文将提出的方法与多个单模态与多模态基线模型进行对比，将这些方法按照使用数据的模态来进行分类，[v]代表基于视觉图像模态数据的方法，[t]代表基于文本模态数据方法，[k]代表基于知识图谱的方法，[t+v+k]代表多模态方法。

1．**视觉单模态模型（[v]）：**本文将模型在图像分类任务上进行训练，将字形与文字标签之间的分数设置为字形所有图像与该文字标签的概率分数之和，具体模型包括：

（1） DesNet：2.3.1 节介绍的 DesNet121 模型。

（2） ResNet（ours）：本文使用 ResNet50 模型，基于无监督对比学习，在古汉字图像数据集上进行进一步的预训练得到该模型。

2. **知识图谱单模态模型（[k]）：**本文基于汉字字形知识图谱 CGKG，将模型在字形分类任务上进行训练：

（1） GAT：2.1.2 节中介绍的图注意力网络模型。

3. **文本单模态模型（[t]）：**本文基于字形的部件描述文本数据，将基于监督学习的模型在字形分类任务上进行训练，具体模型包括：

（1） 部件描述文本的最长公共子序列（LCS）。

（2） SikuBERT6：在古汉语文献的大规模语料库中训练的 BERT 模型。

（3） BERT（ours）：本文对 SikuBERT 在甲骨文、金文、战国文字的出土文献数据集上继续训练得到的 BERT 模型。

4. **多模态模型：**本文设置了四种多模态基线方法，由于古文字字形识别是一个创新性的任务，目前没有完全相同的工作，因此，本文选择其他多模态融合的模型，在本文的字形识别任务和数据集上进行复现，以显示潜在的可用方法在本任务上的效果。具体方法包括：

（1） [t+v] ViLT：2.3.3 节介绍的计算机视觉语言模型，ViLT 采用极简的线性变换来抽取视觉特征，基于 Transformer 实现图像和文本的特征融合，在减小模型的参数规模的同时取得更好的性能。

（2） [t+v] HRGAT[42]：一种基于文本和视觉协同注意力机制的多模态融合方法，分别在文本信息的引导下获取加权视觉特征，以及在视觉信息的引导下获取加权的文本特征，然后，使用门控融合来评估不同模态的特征的重要性，并将其整合到最终的多模态表示。

（3） [t+v+k] IMF[41]：一种用于知识图谱链接预测的两阶段交互式多模态融合框架，对不同模态进行独立表示，利用双线性池化进行融合。

[6] https://huggingface.co/SIKU- BERT/ sikubert

（4） [t+v+k] GlyphSim：本文第 4 章提出的字形相似度计算方法，该方法不是针对本章的任务设计的，但是可以通过计算训练集与测试集字形之间的相似度，将测试集字形分类到相似度最大的字形对应的文字标签中。

5.4.3 消融实验设置

本文的 IE、KGE 和 ME 编码器分别使用 ResNet50（ours）、GAT 和 BERT（ours）模型实现。在消融实验中，本文首先对多模态编码器（ME）进行消融，评估 ME 分别使用三种多模态信息，即部件描述文本（T）、字形知识图谱（KG）、视觉图像（P）的组合时的效果，以此来验证 ME 编码器设计的合理性。然后，本文对完整的 MCGI 模型进行消融实验，对比 MCGI 在三种编码器组合使用时的效果，以及解码器分别采用监督和无监督学习时的效果，以此验证本文每一种编码器、解码器设计的合理性。待评估的模型如下：

1. **多模态编码器（ME）的消融实验：**

（1） ME(t+v)：仅使用部件描述文本和图像特征的 ME 编码器，即在网络的融合层不融合部件特征。

（2） ME(t+k)：仅使用部件描述文本和部件特征的 ME 编码器，即在网络的融合层不融合字形的图像特征。

（3） ME(t+v+k)：完整的 ME 编码器，即 5.3.5 节介绍的方法。

2. **MCGI 完整方法的消融实验：**

（1） ME(t+v+k)+IE：MCGI 使用 ME 与 IE 编码器生成的$\boldsymbol{e^m}$和$\boldsymbol{e^v}$特征向量进行解码，解码器采用无监督模型。

（2） ME(t+v+k)+KGE：MCGI 使用 ME 与 KGE 编码器生成的$\boldsymbol{e^m}$和$\boldsymbol{e^{kg}}$特征向量进行解码，解码器采用无监督模型。

（3） ME(t)+KGE+IE：MCGI 使用 ME、IE、KGE 三种编码器生成的$\boldsymbol{e^m}$、$\boldsymbol{e^v}$和$\boldsymbol{e^{kg}}$特征向量进行解码，解码器采用无监督模型，但是 ME 编码器仅使用部件描述文本信息。

（4） ME(t+v)+KGE+IE：MCGI 使用 ME、IE、KGE 三种编码器生成的$\boldsymbol{e^m}$、$\boldsymbol{e^v}$和$\boldsymbol{e^{kg}}$特征向量进行解码，解码器采用无监督模型，但是 ME 编码器仅融合部件描述文本和图像信息。

（5） ME(t+v+k)+KGE+IE：MCGI 使用 ME、IE、KGE 三种编码器生成的$\boldsymbol{e^m}$、

$\boldsymbol{e^v}$和$\boldsymbol{e^{kg}}$特征向量进行解码，解码器采用无监督模型，即本文提出的 MCGI 模型的完整方法。

（6） ME(+v+k)+KGE+IE-s：MCGI 使用 ME、IE、KGE 三种编码器生成的$\boldsymbol{e^m}$、$\boldsymbol{e^v}$和$\boldsymbol{e^{kg}}$特征向量进行解码，解码器采用监督学习模型。

5.4.4 评价指标与参数设置

本实验的评价指标包括：

（1）正确标签分别在前 1、3、10 位排序中的平均比例（**R@1**、**R@3**、**R@10**）。

（2）正确标签的平均排名（**MR**）和平均倒数排名（**MRR**）。

本文选择基于排序的评价指标，因为这项任务的潜在应用场景是在古文字研究和信息检索应用中为领域专家推荐未知字形最可能属于的一组文字标签，以及对应的相关字形，以启发他们的想法并缩小候选文字的检索范围。正确标签在前 n 位的平均比例以及 MR、MRR 得分不仅可以评估算法的准确性，也可以反映实际应用中的文字推荐的数量和专家发现正确答案所需要的浏览量。

在 IE 编码器的训练中，本文设置批量大小（Batch Size）为 64，学习率（Learning Rate）为 0.001，epoch 为 150，图像特征的维度d_v为 768；在 ME 编码器的训练中，本文使用 768 维的 BERT 基础版本，批量大小为 64，学习率为 0.00002，多模态特征向量的维度d_m为 768，epoch 为 100。对于 KGE 编码器，本文使用 OpenNE2 工具提供的 node2vec 算法初始化图谱节点的嵌入，本文使用该工具的默认参数，输出向量的维度为 1000，接下来在 GAT 网络训练 5 个 epoch，学习率为 0.0005，批量大小为 256，图谱特征向量的维度d_k为 2000。解码器中的超参数α、β和γ分别设置为 0.4、0.5 和 0.7，这是通过在 0.1 和 1.0 之间，间隔为 0.1 的组合上对 50 个额外的字形样本进行校验设置的最优参数。对于所有的监督学习模型，本文共运行 3 次，选择最好的分数作为结果。

5.4.5 实验结果与讨论

实验结果如表 5.3 所示，本文将测试集按照 5.4.1 节中介绍的标准分为共时测试字形与历时测试字形分别进行评估，并将 5.4.2、5.4.3 节中介绍的待评价模型分为四个大组，第一组为单模态基线模型，第二组为多模态基线模型，第三组为本文的 ME 编码器的消融实验，第四组为本文整体方法的消融实验。表中粗体

标注的 **ME(t+v+k)+KGE+IE** 是本文提出的完整方法，最佳结果以红色斜体突出显示，四个子组中的最佳结果以蓝色突出显示。

表 5. 3 字形识别实验结果

方法	共时测试字形					历时测试字形				
	R@1	**R@3**	**R@10**	**MR**	**MRR**	**R@1**	**R@3**	**R@10**	**MR**	**MRR**
[v]DenseNet	29.9%	39.6%	50.5%	411	0.369	23.8%	35.1%	47.9%	591	0.316
[v]ResNet(ours)	35.6%	48.7%	61.4%	292	0.441	28.8%	40.2%	53.2%	369	0.369
[k]GAT	58.7%	77.8%	85.5%	34	0.691	36.9%	51.5%	68.0%	153	0.469
[t]LCS	50.5%	66.1%	76.9%	58	0.599	28.6%	43.3%	52.9%	480	0.378
[t]SikuBERT	67.4%	79.4%	85.5%	86	0.742	35.7%	47.6%	57.7%	482	0.435
[t] BERT(ours)	70.6%	82.8%	88.0%	82	0.777	40.4%	53.5%	62.4%	454	0.485
[t+v]ViLT	66.7%	81.1%	87.5%	70	0.743	51.7%	63.3%	70.8%	348	0.587
[t+v]HRGAT	65.2%	74.4%	81.6%	185	0.710	36.1%	46.0%	53.3%	722	0.424
[t+v+k]IMF	63.0%	76.1%	85.6%	68	0.710	38.6%	54.8%	68.4%	157	0.448
[t+v+k]GraphSim	69.6%	85.8%	91.6%	*9*	0.784	44.1%	62.0%	79.1%	*62*	0.557
ME(t+v)	68.6%	80.3%	85.5%	84	0.755	55.9%	64.3%	72.1%	260	0.618
ME(t+k)	71.9%	84.1%	89.0%	70	0.782	41.1%	53.7%	61.7%	363	0.487
ME(t+v+k)	69.2%	79.3%	85.8%	94	0.754	*56.5%*	65.9%	73.3%	271	0.625
ME(t+v+k)+IE	69.4%	83.6%	89.0%	32	0.772	50.4%	65.8%	75.3%	131	0.594
ME(t+v+k)+KGE	70.7%	85.3%	90.5%	19	0.785	52.7%	69.3%	79.1%	97	0.625
ME(t)+KGE+IE	72.1%	87.1%	92.8%	13	0.801	51.7%	68.1%	80.9%	73	0.619
ME(t+v)+KGE+IE	72.6%	88.3%	*93.0%*	12	0.808	54.8%	71.1%	*82.5%*	69	0.648
ME(t+v+k)+KGE+IE	73.5%	*88.5%*	*93.0%*	13	*0.812*	*56.5%*	*71.8%*	82.1%	70	*0.659*
ME(t+v+k)+KGE+IE-s	*73.7%*	86.5%	92.0%	30	0.810	55.7%	70.6%	80.8%	91	0.646

首先讨论方法的有效性，从表 5.3 中可以得到如下结论：（1）与基线相比，本文的模型在除 MR 以外的所有指标上都取得了最好的结果，能够证明该方法的有效性。（2）该方法在实际使用中可以用于未知字形的文字标签推荐，对共时测试字形来说最高的 R@1 准确率可以达到 73.7%，前 10 位的推荐即可覆盖 93%

的正确标签，但是对历时测试字形的效果较差，最高的 R@1 准确率为 56.5%，最高 R@10 准确率为 82.1%，在实际应用中，历时测试字形需要被推荐更多的前 n 位文字标签才能覆盖更多的正确文字。（3）对 ME 编码器的消融实验可以看出，完整的多模态信息组合方法（ME(t+v+k)）对历时组的实验结果显著优于 ME(t+v)和 ME(t+k)组合，但是共时组中 ME(t+k)组合具有最好的性能，综合来看，完整的 ME 模型仍然取得了最好的效果，可以证明 ME 编码器设计的合理性。（4）在对整体方法的消融实验中，通过比较 ME(t+v+k)+IE 和 ME(t+v+k)+KGE+IE、以及 ME(t+v+k)和 ME(t+v+k)+KGE 组合方法的实验结果，可以证明本文提出的 KGE 编码器在所有指标上的有效性。特别是 KGE 编码器被观察到对提升正确文字标签的平均排名（MR）指标具有非常显著的作用，ME(t+v+k)+KGE+IE 的 MR 较 ME(t+v+k)+IE 降低了 20.2。（5）在对整体方法的消融实验中，通过比较 ME(t)+KGE+IE 与 ME(t+v+k)+KGE+IE 的实验结果，可以证明本文的 ME 编码器中视觉和知识图谱模块的有效性，特别是对于历时组的结果具有明显改善。（6）在对整体方法的消融实验中，通过比较 ME(t+v+k)+KGE 与 ME(t+v+k)+KGE+IE 的结果，可以证明 IE 编码器的有效性，ME(t+v+k)+KGE+IE 对于 ME(t+v+k)+KGE 在所有指标上均有提升。（7）本文没有在实验结果中观察到监督解码器（ME(t+v+k)+KGE+IE-s）相对于使用无监督解码器（ME(t+v+k)+KGE+IE）在大多数指标上的改善。这是因为在字形识别任务中，一个文字标签只有有限个字形用于训练，属于小样本分类问题，因此无监督的解码器可以取得更好的效果。

接下来讨论图像、文本、和图谱模态的信息分别对于该任务的贡献。从单模态方法的结果可以看出，基于文本（[t]）模型和图谱（[k]）模型的结果均优于视觉（[v]）模型，[t]模型在 R@1、R@3、R@10 指标的得分较高，[k]模型可以显著提升 MR 指标，因为古文字异体字形之间的书写形态差异过大，这限制了[v]方法的性能。但是，本文也注意到基于部件方法的局限性，根据 5.4.1 节的统计，大多数字形的部首数量仅为 1-3 个，这不足以区分部件组成相似的字形特征，而部件组成相似是现实中常见的情况。本文随机选取了 150 个 ME(t+k)模型的分类错误的字形样本，发现其中 108 个被分类到部件组合非常相似的文字标签中。通过融合多模态信息可以对特征进行互补，通过对比 ME(t+k)与 ME(t+v+k)编码器

的结果，本文发现知识图谱中获取的部件语义特征对共时测试组的贡献更大（ME(t+k)），但对历时组的效果有限，通过进一步融合视觉图像特征则可以显著提升历时测试组的效果（ME(t+v+k)），原因是共时异体字形之间存在更大的部件语义关联，使它们从字形知识图谱中受益更多，而历时组则需要更多的图像特征作为补充。

最后讨论工作的局限性：（1）本文只测试了 805 个常用文字标签下的 1279 个已被考释的字形。对于未释和生僻的文字的字形来说，它们的形态和部件的变化很可能更大，因此也更难进行预测，方法的效果预计将会降低。（2）本文的方法只适用于在训练集中存在异体字形的测试字形，并且对于独体字形和部件未与训练字形共享的字形，该方法的效果略差，但这种情况在本任务中并不常见。（3）除部件和视觉特征外，历史文献的上下文、词典记录、字音、前人研究等都是古文字考释研究非常重要的信息，多模态考释方法有待进一步研究。

5.5 本章小结

本文根据古文字考释研究的实际需求，提出了一个字形识别的创新任务，设计了一种基于视觉和字根语义信息的知识驱动的多模态识别方法，基于 ZiNet 汉字知识图谱构造了汉字字形的语义知识图谱，提出了融合视觉、文本和图谱特征的字形识别框架，实验结果显示，本文提出的方法对共时字形识别的准确率可以达到 73.7%，对历时字形识别的准确率为 56.5%，与基线相比，本文的模型在大多数指标上都取得了最好的结果，证明了该方法的有效性。并且通过对单模态模型的评估，证明了视觉模态与部件模态的信息对字形识别任务的有效性，视觉信息对历时测试组的贡献更大，而后者对共时测试组的效果更好，两者互补可以取得最好的效果。

第 6 章　汉字语素含义挖掘与历时语义跟踪方法

词汇的语义变化可以揭示语言、社会和文化发展的复杂过程。近年来， NLP 方法被使用从大规模的历史语料中挖掘词义的历时变化，例如判断词义是否发生变化、新义项的出现或旧义项消失时间等。对于这项任务，目前仍然缺乏支持模型训练的相关古汉语语料库和数据集，并且现有工作对词汇语境进行词义消歧，生成对目标词汇的义项随时间变化的频率分布可视化分析，但未对语素进行表示，汉字可以作为语素构成复合词汇来表达其语义。针对这些问题，本文致力于从词汇和语素两个层面对汉字在较长历史阶段的语义变化进行挖掘和可视化分析，具体来说，本文基于 ZiNet 中的数据构建一个跨越 3000 年的词汇义项数据集，训练一个汉字语境表征模型，并将现有的词汇语义跟踪方法扩展到语素层面，提出一个汉字语素含义挖掘与历时语义跟踪框架（CMSMT），该框架包括汉字语境表示模型训练、义项识别、语素含义挖掘和语义变化表示四个主要步骤。定量实验结果和定性分析可以表明该方法的有效性。最后，在一个有趣的统计中，本文发现汉字的词汇与语素的语义变化趋势有很强的正相关性。本章 6.1 节为引言，6.2 节介绍汉字、词汇、义项与语素的基本定义，6.3 节将介绍历时语义跟踪方法的细节，6.4 节介绍对该方法的定量和定性评估，6.5 节介绍词汇与语素义项变化的相关性分析，6.6 节对本章的工作进行总结。

6.1 引言

词汇的含义随着时间的推移而不断变化，可以反映出语言和社会文化的复杂发展过程。例如，在商周时期，汉字“师”的本义是“军队”，后来又出现了“专责教育的官职”的新含义，并由此引申出“教师”的含义。对词汇语义变化进行挖掘，如分析哪些词义是稳定的，新的词义在什么时间出现、使用增加或减少等情况，有助于推动历史语言学、文字学、历史学、词典编纂等领域的研究，但是仅依靠专家的人力来阅读和分析大规模的历史语料是不现实且不全面的。

近年来，NLP 和历史语料库的发展推动了该领域自动化方法的应用。像 BERT 这样的预训练语言模型能够捕获复杂的上下文特征，并在不同的上下文中对目标词进行表示，基于这种语境化的表征，Hu 等人[11]首次引入了历时语义建模，以细粒度、平滑的方式表示目标词汇每一个义项的频率分布历时变化，他们对特定语境下的词汇进行词义消歧，然后对每个义项在每一个历史时期的使用频

率进行统计，从而生成该词汇的义项随时间变化的频率分布表示和可视化分析。Shu 等人的工作遵循该框架，对汉字从先秦至今的语义变化进行表示与可视化[12]。然而，与印欧语系的“词”不同，汉字可以作为单音词汇使用，也可以作为语素构成多音节复合词，因此，挖掘语素层面的字义变化对于探索语言和文字的发展同样具有意义，此外，该领域缺乏用于汉字语义消歧和语素挖掘的数据集与模型。

针对以上背景和问题，本章的内容与贡献包括：（1）构建用于汉字语义跟踪的数据集，包括一个用于汉字语境表示模型训练的义项-上下文数据集，该数据集包含汉字、词汇、义项和带有朝代注释例句，以及一个中国历史文献语料库。（2）本文将 Hu 等人[11]提出的语义跟踪方法扩展到语素层面，具体来说，本文使用义项-上下文数据集训练了一个词汇表示模型，然后利用该模型识别历史文献语料库上下文中的词汇的含义，并挖掘可作为语素使用的字义，最后从词汇和语素两方面为汉字各个义项的使用频率变化提供平滑的表示，其中模型训练与语素义项挖掘是本文的创新部分。（3）本文在 100 个常见汉字的样本中，对汉字的词汇含义与对应语素含义的频率分布变化进行统计分析，发现二者具有强相关性。定量实验结果表明，与原 BERT 模型相比，本文训练的词汇表示模型具有更好的义项识别和语素义项挖掘性能，义项识别的准确率由 55.08%提高到 74.19%，语素义项挖掘的 F1 分数由 59.21%提高到 75.14%。

6.2 汉字与语素相关知识

本章将对汉字、词汇、义项、语素的概念和它们的关系进行简要介绍：

词汇：词汇是由语素组成的最小的造句单位，本文将汉语词汇简单地分为两类：单音节词和多音节复合词，单音节词由一个汉字组成，复合词由两个或两个以上的汉字组成。根据对《汉语大词典》中 256293 个复合词的统计，85%（217967）的复合词由两个汉字组成，6.7%（17148）由三个汉字组成，7.8%（20016）由四个汉字组成。

义项：义项是词汇的含义，一个词可以有多个义项，相同或相近的义项也可以由多个词来表示。

汉字：汉字是汉语的记录符号。从表达意义的角度来说，汉字可以看作是单音节词，也可以看作是构成复合词的一个语素，如复合词“教师”由“教”和“师”两个语素组成。

语素：语素是一个词中最小的表达含义的语言单位，在汉语中，大多数语素是用一个汉字来表示的。一般来说，复合词的含义与其语素的含义相关。

汉字是汉语起源时自然产生的最小语义单位，随着社会的发展和认知的提升，由于汉字造字的不可持续性而产生了复合词。汉字语义的表达和使用模式可以自然而清晰地分为两类：作为单音节词汇使用或作为语素组成其他复合词。汉字每一个义项的使用情况随着时间的推移可能会发生变化，例如，在现代，许多汉字的某些含义已经不能直接作为词汇使用，但这并不意味着这些义项已经消失，它们可能仍然保留在语素中。因此，有必要将语素概念引入到汉字语义跟踪框架中。

6.3 汉字历时语义跟踪方法

6.3.1 方法框架概述

本章将介绍汉字历时语义跟踪方法，汉字语素含义挖掘与历时语义跟踪框架（CMSMT）如图 6.1 所示，它包括四个步骤：（1）汉字语境表示模型训练；（2）义项识别，将上下文中的目标词汇分类到相应的义项标签；（3）语素含义挖掘，将文字义项与可以组成的复合词的义项匹配；（4）语义变化表示与可视化，根据（2）和（3）的结果为每个义项提供平滑的频率分布表示。

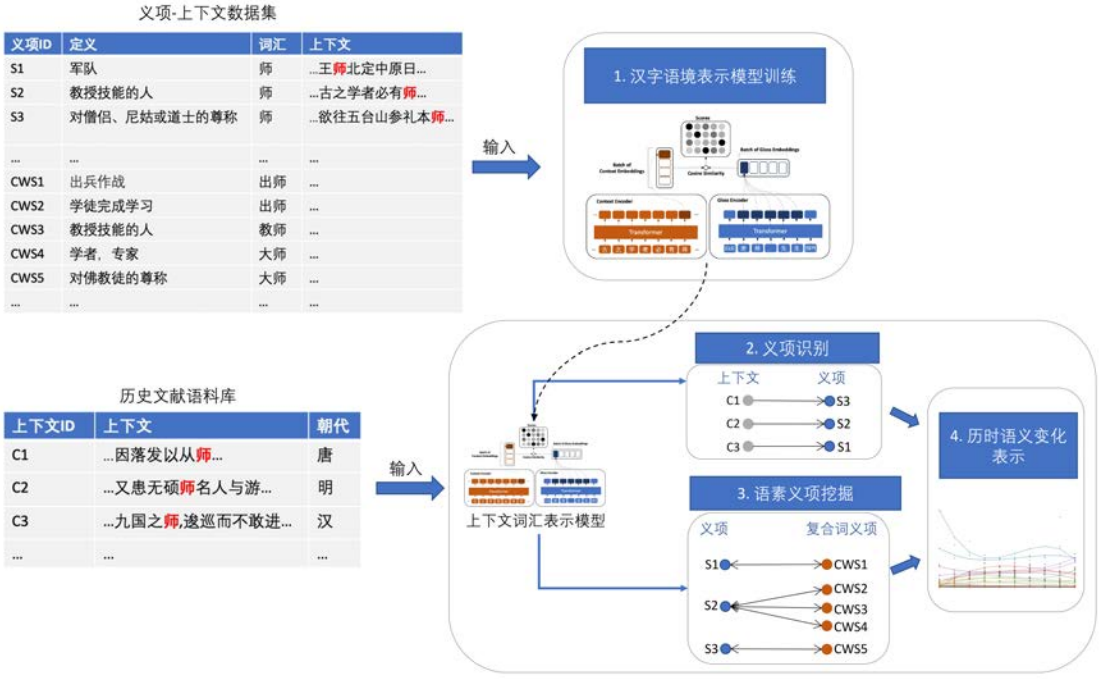


图 6.1 历时语义跟踪框架（CMSMT）

首先，本文基于 ZiNet 中的数据构建一个义项-上下文数据集，数据集包含汉语词汇、义项、每个义项的定义和分布在各个历史阶段的上下文案例。本文使用该数据集训练一个汉字与词汇的语境表示模型，该模型为上下文语境中的汉字或词汇生成高维向量表示，使具有相似含义的词汇嵌入在空间中更接近，并可以对词汇的不同义项进行消歧。

给定待分析的汉字，本文首先在构建的大规模历史文献语料库中检索包含目标汉字的所有上下文，使用上下文词汇表示模型生成每一个上下文中该汉字或词汇的高维向量表示。然后基于该模型实现义项识别和语素义项挖掘，义项识别任务的目的是通过词义消歧将上下文中的目标词汇分类到正确的候选义项中。例如，图 6.1 中，汉字“师”有三个候选义项：“军队”、“教授技能的人”和“对僧侣、尼姑或道士的尊称”，上下文“……九国之师, 逡巡而不敢进……”中的“师”字需要被分类到“军队”义项中。

语素义项挖掘任务的目的是回答目标汉字的哪些义项可以参与构成复合词、可以构成哪些复合词。本文将目标汉字参与构成复合词时表达的义项称为“语素义项”。因此，该任务的主要过程是将目标汉字的每个义项与其可以组成的复合词的义项进行匹配。例如，图 6.1 中“师”字的三个义项：“军队”、 “教授技能的人”、“ 对僧侣、尼姑或道士的尊称”可以分别与“出师”词汇的“出兵作战”义项、“教师”的“教授技能的人”义项，以及“大师”的“ 对佛教徒的尊称”义项相匹配。

最后，对于目标汉字的每个义项，在对每一历史时间阶段的语境使用数量以及组成的复合词数量进行统计后，本文可以从词汇和语素两个角度来表示该义项的历时频率分布，为相关领域专家生成汉字语义变化的可视化分析。本文第 6.3.2 节将介绍数据集构建，第 6.4.3 节介绍汉字语境表示模型，第 6.4.4 节介绍语素义项挖掘，第 4.5 节介绍历时语义变化表示。

6.3.2 数据集介绍

本文构建了一个用于该任务的义项-上下文数据集，用于语境表示模型的训练，以及一个包含中国各个历史时期文学作品的历史文献语料库，用于从中挖掘汉字的语义变化。其中义项-上下文数据集旨在整合汉字和词汇的所有义项、以及义项在历史文献中的语境上下文样本信息，包含 9 个条目：

（1） 上下文：目标词所处的上下文，通常是一个句子。

（2） 时间阶段：上下文所处的历史时间段。本文用相对宽泛的朝代来划分时间区间，如图图 6.2 所示，因为朝代是划分我国政治、生活、文化等历史阶段最直接、最重要的依据，虽然语言词汇的变化节点与朝代更替并不完全一致，但是更细粒度的时间划分很难清晰界定与标注。

（3） 作者：上下文的作者。

（4） 作品名称：上下文来源的文学作品名称。

（5） 词汇索引：目标汉字或词汇在上下文中的位置索引。

（6） 词汇 ID：目标词的唯一识别码。

（7） 词汇：上下文中的目标词名称，可以是单音节词（汉字）或复合词。

（8） 义项 ID：上下文中目标词所属义项的唯一识别码。

（9） 义项定义：对所属义项含义的描述，通常是一到两个句子。

图 6. 2 汉字历时语义跟踪的时间阶段划分

原始数据来自 ZiNet 收集的开放汉语词典资源。本文从词典中提取出词汇（包括汉字）、义项、例句、作者、文章标题等结构化信息，接下来对定义进行简化，仅保留前三个句子，因为长定义的后文对义项的描述性变弱，最后通过自动搜索百度百科[7]，根据作者和著录信息对每一个上下文进行朝代注释。最终，义项-上下文数据集中共包括 265289 个词汇、392213 个义项和 844236 条上下文，平均每个复合词汇包含 1.35 个义项，每个汉字包含 5.26 个义项，每个词汇义项有 1.91 个上下文样本，每个汉字的义项有 3.91 个上下文样本，本文分别统计了汉字和复合词的数量，以及它们的义项和上下文数量，见表 6.1，每个朝代阶段的上下文数量比例如图 6.3 所示。

[7] https://baike.baidu.com/

本文的历史文献语料库包括历史文献及其所属朝代的标注。语料库的数据源来自《殆知阁语料库》[8]，该语料库规模超过 20 亿字，包含中国历史上流传下来的各类文献著作，包括诗歌、小说、散文、戏剧等各类文学体裁，本文通过自动搜索百度百科为每篇历史文献标注朝代信息，删除重复的文献、以及朝代未知或无法标注的文献，最终共获得分布于各朝代的 4569 篇传世文献，共计 4.13GB。

表 6.1 义项-上下文数据集的统计信息

词汇类别	词汇数量	义项数量	上下文数量
单音节词汇 (汉字)	8996	47,341	185,071
复合词	256,293	344,872	659,165

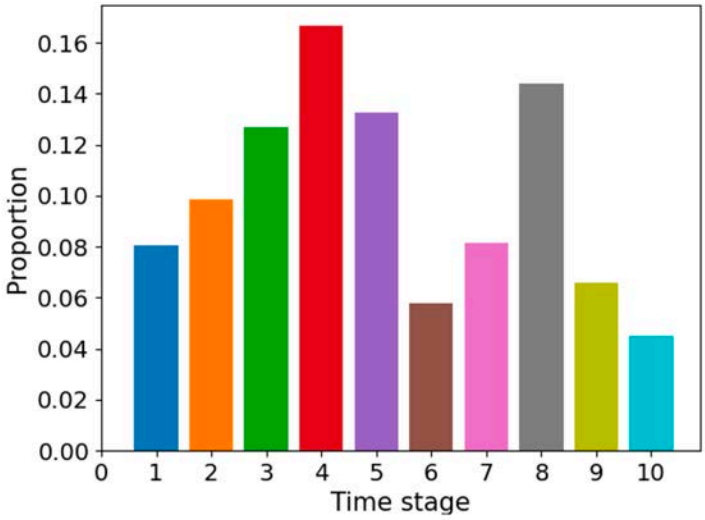


图 6.3 各历史时间阶段的上下文数量比例分布，1 – 10 为顺序的朝代编号

6.3.3 汉字语境表示模型

虽然 BERT 模型已经从预训练过程中学习到通用的上下文语义特征，但是还不足以识别近似的义项和区分不同的义项，因此本文使用义项-上下文数据集在词义消歧（WSD）任务上对 BERT 模型进行进一步微调。给定一个文字或复合词汇w和上下文c，WSD 任务的目的是找到一个函数f，使得$f(w,c)=s$，$s \in S_w$。其中S_w是w的所有候选义项的集合，s是上下文c中w表示的正确义项。

[8] https://github.com/garychowcmu/daizhigev20

本工作构建一个联合优化的双编码器模型，包括语境上下文编码器和义项编码器，模型的体系结构如图 6.4 所示。上下文编码器的输入是包含目标汉字w的上下文文本序列c，$c = \{[\text{CLS}], d_1, \ldots w, \ldots, d_n, [\text{SEP}]\}$，其中$d$是上下文中的其他字符，[CLS]和[SEP]是 BERT 规定的开始和结束符号。在该模型的训练中，w随机按一定比例被替换为掩码符号[MASK]，以隐藏目标词汇的信息，使模型能够更好地学习上下文中其他词汇与对应义项的关联。本文将上下文编码器最后一层w的对应输出作为它在语境c中的高维向量表示$\boldsymbol{r}_{\boldsymbol{w}}^{\boldsymbol{c}}$。如果目标词汇是由多个汉字组成的复合词，本文将其表示为对应的汉字输出嵌入的平均值。义项编码器的输入是$g = \{[\text{CLS}], d_1, \ldots, d_m, [\text{SEP}]\}$，其中$g$是$w$对应的一个候选义项的定义文本序列，$d$是序列中的字符。本文将编码器的最后一层输出的$d_1$到$d_m$对应的$m$个词嵌入的平均值作为义项$g$的向量表示$\boldsymbol{r}_{\boldsymbol{g}}$。

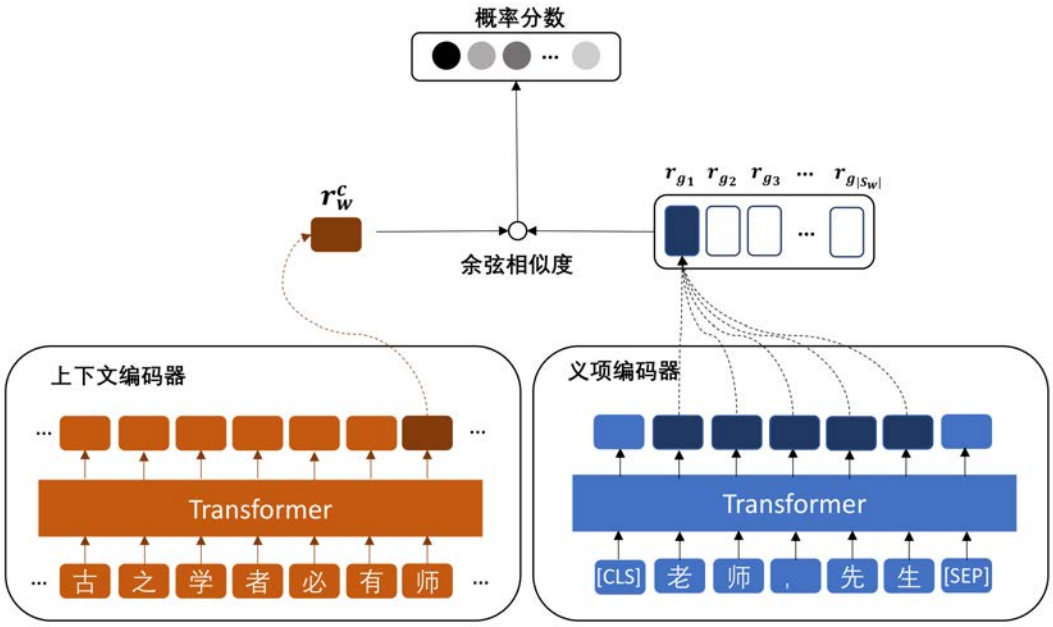


图 6. 4 汉字语境表示模型结构

本文基于对比学习[31]的思想对该模型进行训练。对于上下文编码器输入的每个上下文和目标词汇（c，w），对应批次的义项编码器的输入包含w全部候选义项的定义的文本序列$\{g_1, g_2, \ldots, g_{|S_w|}\}$。本文将语境$c$中的目标词汇$w$与候选义项$s_k$，$s_k \in S_w$的匹配分数设置为$\boldsymbol{r}_{\boldsymbol{w}}^{\boldsymbol{c}}$与$\boldsymbol{r}_{\boldsymbol{g_k}}$的余弦相似度：

$$score((c,w), s_k) = \frac{\boldsymbol{r}_{\boldsymbol{w}}^{\boldsymbol{c}} \boldsymbol{r}_{\boldsymbol{g_k}}}{\| \boldsymbol{r}_{\boldsymbol{w}}^{\boldsymbol{c}} \| \cdot \| \boldsymbol{r}_{\boldsymbol{g_k}} \|} \tag{6.1}$$

网络训练的目标是使$\boldsymbol{r}_{\boldsymbol{w}}^{\boldsymbol{c}}$接近正确义项的嵌入$\boldsymbol{r}_{\boldsymbol{g}_k}$，并远离其他候选义项的嵌入。本文使用交叉熵损失函数训练双编码器模型，定义为：

$$\ell_{(c,w)} = -log\frac{e^{score((c,w),s_k)/\tau}}{\sum_{j=1}^{|S_w|} e^{score((c,w),s_j)/\tau}} \tag{6.2}$$

其中τ是超参数，本文直接设置为默认值 0.5。义项编码器的批次大小（Batch Size）设置为 64，上下文编码器中[MASK]的概率设置为 0.2。本文在义项-上下文数据集上训练了 50 个 epoch。经过训练后，上下文编码器最后一层的输出$\boldsymbol{r}_{\boldsymbol{w}}^{\boldsymbol{c}}$可以直接作为特定语境$c$中汉字或词汇$w$的表示。

在义项识别步骤中，使用训练好的模型可以方便地识别特定上下文中词汇或文字的义项。给定目标词汇w和所处的上下文c，将c输入到模型的上下文编码器中，输出w位置对应的词汇上下文嵌入$\boldsymbol{r}_{\boldsymbol{w}}^{\boldsymbol{c}}$。对于$w$的每一个候选义项$s$，将其定义的文本序列输入到义项编码器中，输出义项的嵌入$\{\boldsymbol{r}_{\boldsymbol{g}_1}, \boldsymbol{r}_{\boldsymbol{g}_2}, \ldots, \boldsymbol{r}_{\boldsymbol{g}_{|S_w|}}\}$，本文计算$\boldsymbol{r}_{\boldsymbol{w}}^{\boldsymbol{c}}$与所有义项嵌入之间的余弦相似度，并选择相似度得分最高的义项作为词汇w在上下文 c 中的预测义项。

6.3.4 语素义项挖掘方法

图 6.5 展示了 “师”字的语素义项挖掘示例：

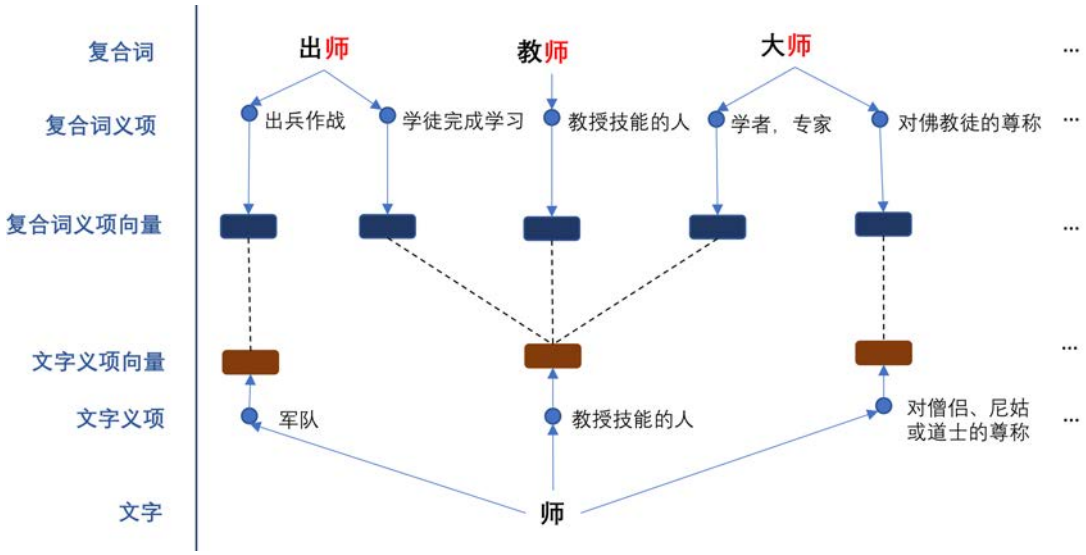


图 6.5 语素义项挖掘方法

该方法首先从数据集中检索出由“师”组成的所有复合词，然后将复合词的每一个义项与“师”字的义项进行匹配，例如，复合词“出师”的第一个义项表

示“出兵作战”，应匹配到的“师”字“军队”的义项中；第二个义项表示“学徒完成学习”，应匹配到“师”的 “教师”义项中。本文遵从一个假设，即汉字如果作为语素组成复合词，那么该语素表达的含义与复合词的含义具有相关性。

具体来说，对于目标文字w，假设w在历史文献语料库中存在m个上下文，我们可以将其输入模型的上下文编码器中获得w的语境表示$\{\boldsymbol{r}_{\boldsymbol{w}}^{\boldsymbol{c1}}, \boldsymbol{r}_{\boldsymbol{w}}^{\boldsymbol{c2}}, \ldots, \boldsymbol{r}_{\boldsymbol{w}}^{\boldsymbol{cm}}\}$，接下来执行义项识别任务，识别出每个上下文中$w$对应的义项，义项$s$的向量表示$\boldsymbol{r}_{\boldsymbol{s}}$被定义为识别为该义项的所有语境向量的平均值。

对于每一个由w构成的复合词，本文也以同样的方式获得其所有义项的表征。对于复合词的一个义项cs，本文计算其向量$\boldsymbol{r}_{\boldsymbol{cs}}$与$w$的所有义项向量$\{\boldsymbol{r}_{\boldsymbol{s_1}}, \boldsymbol{r}_{\boldsymbol{s_2}}, \ldots, \boldsymbol{r}_{\boldsymbol{s_{|S_w|}}}\}$之间的余弦相似度，将复合词义项与相似度最高的文字义项进行匹配，如果最大的相似度得分仍小于 0.01，则该复合词义项不与任何文字义项匹配，因为本文发现当相似得分<0.01 时，大多数的匹配是不正确的。在此步骤之后，对于文字w的每个义项，都可以得到一组对应的复合词义项，当匹配到的复合词义项数大于 3 时，将义项加入到w的语素义项集合MS_w中，$MS_w \subseteq S_w$。

6.3.5 历时语义变化表示方法

本文基于 Hu 等人[11]提出的方法对汉字义项的历时频率分布进行表示，但是将该方法扩展到语素层面。具体来说，文字w的每一个义项s_i的频率分布的历时变化可以表示为：

$$T_{word}(s_i) = \{P_{t_1}^{s_i}, P_{t_2}^{s_i}, \ldots, P_{t_n}^{s_i}\} \tag{6.3}$$

$$T_{\text{morpheme}}(s_i) = \{PM_{t_1}^{s_i}, PM_{t_2}^{s_i}, \ldots, PM_{t_n}^{s_i}\} \ if \ s_i \in MS_w \tag{6.4}$$

其中，$T_{word}(s_i)$表示当w作为单音节词汇使用时s_i的历时频率分布，$T_{\text{morpheme}}(s_i)$是w作为复合词的语素使用时s_i的历时频率分布表示。$\{t_1, t_2, \ldots, t_n\}$是连续的 n 个历史阶段（6.3.2 节）。$P_t^{s_i}$和$PM_t^{s_i}$分别定义为：

$$P_t^{s_i} = \frac{N_t^{s_i}}{\sum_{k=1}^{|S_w|} N_t^{s_k}} \tag{6.5}$$

$$PM_t^{s_i} = 0.5 \times \frac{NM_t^{s_i}}{\sum_{k=1}^{|MS_w|} NM_t^{s_k}} + 0.5 \times \frac{NS_t^{s_i}}{\sum_{k=1}^{|MS_w|} NS_t^{s_k}} \tag{6.6}$$

其中$N_t^{s_i}$是被分类到义项s_i的时间阶段t的汉字语境数量，$NM_t^{s_i}$是经过义项识别和语素义项挖掘过程后，在时间阶段t与s_i匹配的复合词的语境数量，$NS_t^{s_i}$是在t时间阶段与s_i匹配的复合词的义项数量。最后该方法对$T_{word}(s_i)$和$T_{\text{morpheme}}(s_i)$进行四次多项式曲线拟合，得到s_i义项的连续频率分布表示，通过对w每个义项进行相同的操作，可以得到汉字所有义项随时间的使用频率变化的可视化分析。

6.4 实验与评估

6.4.1 实验设计

实验分别对义项识别和语素义项挖掘的效果进行定量评估，并为汉字语义跟踪可视化提供定性分析案例，具体包括：（1）义项识别评估：给定目标汉字或词汇和相应的语境上下文，模型从该词汇的义项列表中选择正确的义项，以精确度（Accuracy）为评价指标。（2）语素义项匹配评估：给定目标汉字的义项，模型为它匹配复合词义项，本文使用准确率（Precision）、召回率（Recall）和 F1 得分作为评价指标。（3）本文提供对汉字“师”的语义跟踪的可视化案例，作为定性分析。以上内容将分别在 6.4.2-6.4.4 节介绍。

6.4.2 义项识别实验结果

本文从数据集中随机抽取 12000 个上下文、目标词汇和对应的正确义项标签构建测试数据集，测试数据集不参与模型训练。本文将测试集按照目标词汇属于文字或者复合词分为两组，因为复合词的候选义项数量（平均 1.35 个）明显少于汉字的候选词义项数量（平均 5.26 个）。本文进一步根据义项标签的训练样本数量将测试集进一步划分为三种情况：只有一个训练样本的目标词汇、训练样本数量大于等于 7 的目标词汇，以及训练样本数量小于 7 大于 1 的目标词汇。最终，12000 个测试样本被分为六组，每组 2000 个样本：多样本的文字（C-multi）；小样本的文字（C-less）；单样本的文字（C-one）；多样本的复合词（W-multi）；少样本的复合词（W-less）以及单样本的复合词（W-one）。

实验将两个原始的 BERT 模型 bert-base-Chinese 和 bert-ancient-Chinese 作为基线，对于基线模型，本文使用训练集中样本上下文中目标词汇的平均嵌入作为对应的义项标签的表示，计算测试目标词的嵌入与它的候选义项嵌入之间的余弦

相似度，选择相似度得分最高的义项作为最终结果。本文的模型遵循 6.3.3 节介绍的方法，分别将编码器初始化为bert-base-Chinese和bert-ancient-Chinese模型，并在训练集上进行微调训练，生成 bert-base-Chinese (ours)和 bert-ancient-Chinese (ours)模型，对测试上下文中的目标词汇进行义项识别，四个模型的结果如表 6.2 所示，最佳结果由粗体显示。

表 6.2 义项识别结果

模型	C-Multi	C-Less	C-One	W-Multi	W-Less	W-One
bert-base-Chinese	52.68%	45.15%	23.61%	69.90%	69.20%	43.65%
bert-ancient-Chinese	55.08%	46.10%	25.46%	73.15%	70.70%	42.30%
bert-base-Chinese (ours)	68.08%	63.25%	58.18%	76.15%	76.55%	65.65%
bert-ancient-Chinese (ours)	**74.19%**	**66.55%**	**60.03%**	**78.70%**	**78.10%**	**67.10%**

结果表明，直接使用原始 BERT 模型进行义项识别的效果非常有限，尤其是对于汉字来说，在只有一个训练样本的情况下（C-One），bert-ancient-Chinese 的精确度仅为25.46%，在训练样本大于等于 7 的情况下（C-Multi），也只有 55.08%。复合词测试组的效果显著高于汉字组，W-One 为 42.30%，W-Multi 可以达到 73.15%。本文的模型 bert-ancient-Chinese (ours)可以将 C-One 组的精确度从 25.46%提升到 60.03%，C-Multi 从 55.08%提升到 74.19%，证明了该模型可以显著提高文字和复合词的义项识别效果。

6.4.3 语素义项匹配实验结果

语素义项匹配的评估的数据由三位母语者进行注释，其过程如下：本文随机选择常用汉字的 200 个义项样本进行评价，给定待评价的汉字义项、汉字对应的复合词、及其复合词的全部义项信息，要求三个注释者为每一个汉字义项提供与之匹配的所有复合词义项，如果两个以上的注释者为目标汉字义项注释了同一个复合词义项，则将其添加到该汉字义项的匹配列表中。最后本文得到了 200 个汉字义项的匹配列表，并将其与模型输出的匹配列表进行比较，结果如表 6.3 所示。可以看出，本文的模型（bert-ancient-Chinese (ours)）的 F1 得分为 75.14%，而原始 BERT 模型（bert-ancient-Chinese）的 F1 得分仅为 59.21%，这说明本文的模型可以提升语素义项挖掘任务的效果。

表 6.3 汉字-复合词义项匹配结果

模型	准确率	召回率	F1
bert-ancient-Chinese	67.71%	61.23%	59.21%
bert-ancient-Chinese (ours)	**77.25%**	**79.12%**	**75.14%**

6.4.4 汉字历时语义跟踪可视化案例

图 6.6 展示了汉字“师”的语义跟踪可视化分析，不同颜色和形状的曲线代表“师”字的不同义项，实线表示词汇义项，虚线表示应用于复合词的语素的义项：

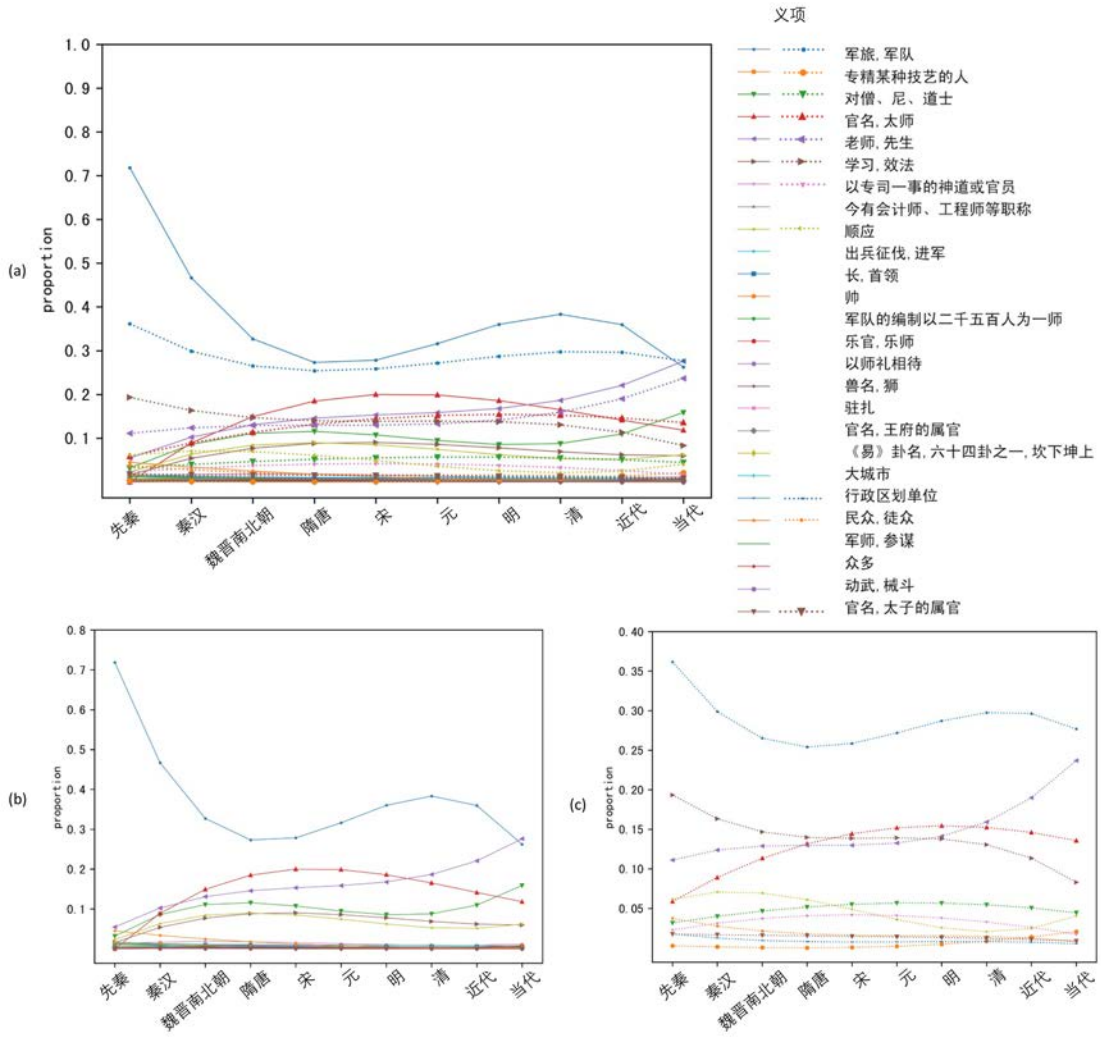


图 6.6 汉字“师”的语义跟踪可视化案例，实线表示词汇义项，虚线表示语素义项。子图（a）为词汇和语素义项的历时语义变化；子图（b）为词汇义项的历时语义变化；子图（c）为语素义项的历时语义变化

如图 6.6 所示，该方法发现了 11 个语素义项，“军队”义项在最早期所占的比例最高，说明“军队”是该字的本义，虽然该义项在后世的使用频率不断减少，但现在仍然是一个常用的含义。“教师”和“专精某种技艺的人”义项逐渐增多，尤其是“教师”取代了“军队”成为在现代最常用的含义。“对僧、尼姑、道士的尊称”的义项在中古时期有所增加，但在近代有下降的趋势。与这些常用义项相比，大多数其他义项占比很小（<2%）。

本文还对“军队”这一概念的用词频率变化进行了可视化分析。图 6.7 是“军队”概念的一个例子，该概念可以通过“师”、“军”、“兵”、“旅”、“戎”、“赋”和“伍”等文字进行词汇化表示，通过可视化分析可以看出，“军”和“兵”是各个历史时期最常用来表示该概念的两个字，它们呈现出增长趋势。还可以发现“伍”字作为词汇使用的频率与作为语素的使用频率相差较大，这说明“伍”字是构成复合词的常用语素，但是它较少直接用于单音节词。

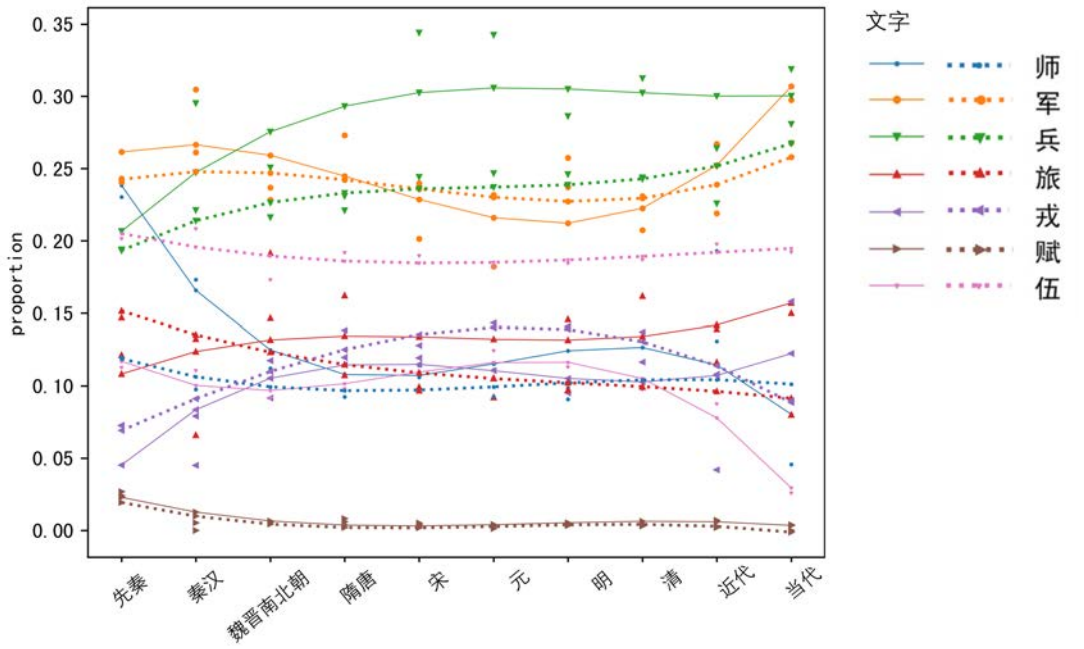


图 6.7“军队”概念用字频率的可视化跟踪案例

6.5 词汇与语素义项历时变化相关性分析

本文对词汇与语素义项变化的相关性进行统计分析，具体来说，本文关注以下三个问题：

（1） 汉字词汇与语素两种表达的历时比例变化。

（2） 词汇义项与对应的语素义项的使用频率是否具有相关性，例如，如果

汉字的某个含义经常在其作为单音节词汇时表达，那么它是否也经常用于语素？

（3） 词汇义项与对应的语素义项的变化趋势是否具有相关性，例如，如果汉字的词汇义项的使用频率有增加趋势，那么该义项在语素中是否也有同样的趋势？

本文选取100个古今常用汉字作为样本。每个文字的平均义项数量为22.19，但平均常用的义项数量仅为 3.57 个，在至少一个历史时间阶段内使用频率大于0.1的义项被定义为常用义项。这三个问题的分析分别见第6.5.1－6.5.3节。

6.5.1 单音节词汇与语素的历时比例变化

本文分别统计不同时间阶段目标汉字作为单音节词汇和构成复合词语素使用的上下文数量，对样本平均后得到历时比例变化。如图6.8所示，随着时间的推移，语素的比例显著增加，而单音节词汇则呈现减少趋势。统计结果符合对汉语词汇发展过程的直觉：随着人类认知观念的增加，汉字数量的无限增加会增加记忆负担，因此形成复合词来表达新的概念。尤其是在人类认知发生剧烈变化的历史时期，如近代时期，复合词的比例呈现出快速上升的趋势。

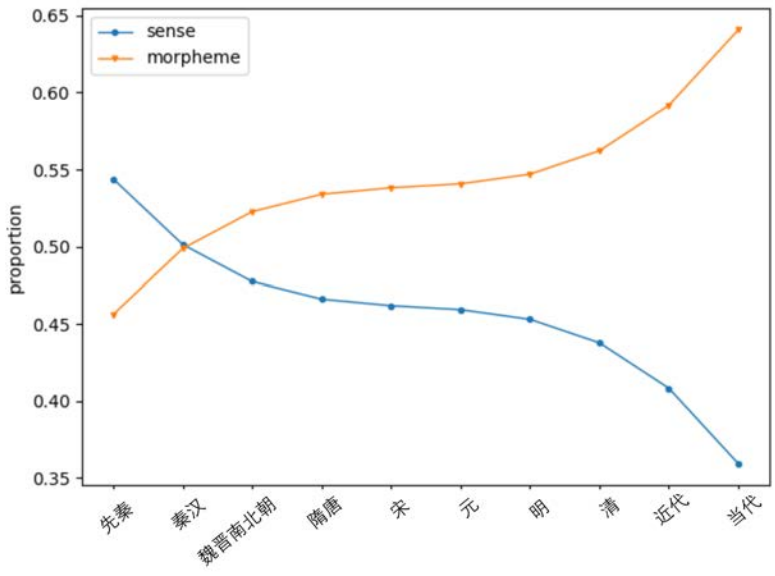


图 6.8 单音节词汇与语素的历时比例变化

6.5.2 词汇与语素义项的使用频率相关性

本文使用义项的分布函数（如图 6.6 中的函数曲线）和x轴包围区域的面积来量化义项的使用频率得分：

$$P(X) = \int f(x)dx \tag{6.7}$$

义项使用频率越高，$P(X)$值越大，相反，$P(X)$将趋向于接近 0。本文分别计算 100 个常用字的 1279 个义项样本分别作为词汇义项（如图 6.6 中实线函数）和语素义项（如图 6.6 中虚线函数）时的使用频率分数，最后得到所有词汇义项的频率得分序列：$Seq_w\big(P_w(X_1), P_w(X_2), \dots, P_w(X_{1279})\big)$，以及语素义项的频率分布序列：$Seq_m(P_m(X_1), P_m(X_2), \dots, P_m(X_{1279}))$。最后，计算出两个序列之间的 Spearman 相关性为 0.751，p<0.01。该结果表明，词汇义项与对应的语素义项的使用频率存在较强的正相关关系。

在此基础上，本文进一步分析 1279 个义项对$(P_w(X_i),\ P_m(X_i))$的使用频率的相似度分布，使用频率相似度的计算公式为：

$$fre\_sim\big(P(X_1), P(X_2)\big) = 1 - \frac{|P(X_1) - P(X_2)|}{P(X_1) + P(X_2)} \tag{6.8}$$

两个义项的使用频率越接近，fre_sim越接近 1，否则接近 0。本文对 1279 个对应的词汇与语素义项组合（$P_w(X_i), P_m(X_i)$）的fre_sim分数进行了统计。为了进行比较，本文还计算了 1279 个随机组合（$P_w(X_i), P_m(X_k)$） $i \neq k$的fre_sim得分，随机组合中的词汇义项和语素义项分别来自两个不同的义项。结果如表 6.4 所示，频率相似度的中位数为 0.628，随机组合的中位数仅为 0.354，Wilcoxon 符号秩和检验（Wilcoxon Signed-Rank Test）表明义项匹配组合的得分显著高于随机组合得分。以上分析可以得出，词汇与相应的语素义项的使用频率有很强的正相关性。

表 6.4 1279 个匹配和随机义项组合的使用频率相似度分布

义项组合类别	中位数(P25, P75)	Wilcoxon 符号秩和检验	
		z	p
$fre\_sim(P_w(X_i), P_m(X_i))$	0.628 (0.340, 0.851)	17.185	< 0.01
$fre\_sim(P_w(X_i), P_m(X_k)), i \neq k$	0.354 (0.145, 0.657)		

6.5.3 词汇与语素义项的变化趋势相关性

本文以 0.1 为区间为义项的频率分布函数进行插值，得到每一个义项函数的离散序列Seq，对所有 1279 个义项样本的词汇义项函数的离散序列进行拼接，得

到词汇义项序列：$Seq_w = Seq_{w_1} \oplus Seq_{w_2}, \ldots, \oplus Seq_{w_{1279}}$，相同地，对 1279 个义项样本的语素义项函数的离散序列进行拼接，得到语素义项序列：$Seq_m = Seq_{m_1} \oplus Seq_{m_2}, \ldots, \oplus Seq_{m_{1279}}$。本文对$Seq_w$和$Seq_m$进行 Spearman 相关性分析，得到了词汇义项与语素义项的变化趋势之间的相关性为 0.731，p<0.01，这说明二者呈较强的正相关关系。

本文使用公式 6.9 来量化两个义项函数之间的趋势相似性，最终得到 1279 对词汇与语素义项组合（Seq_{w_i}, Seq_{m_i}）的趋势相似性。为了比较，本文还计算了 1279 个随机义项组合（Seq_{w_i}, Seq_{m_k}），$i \neq k$的趋势相似性得分，随机组合中的词汇义项和语素义项分别来自两个不同的义项。结果如表 6.5 所示。

$$tr\_sim(Seq_1, Seq_2) = \text{Spearman}(Seq_1, Seq_2) \tag{6.9}$$

表 6. 5 1279 个匹配和随机义项组合的变化趋势相似度分布

义项组合类别	中位数(P25, P75)	Wilcoxon 符号秩和检验	
		z	p
$tr\_sim(Seq_{w_i}, Seq_{m_i})$	0.555 (0.051, 0.862)	17.308	< 0.01
$tr\_sim(Seq_{w_i}, Seq_{m_k}), i \neq k$	−0.082 (−0.587, 0.510)		

结果表明，趋势相似度的中位数为 0.555，不同义项随机组合的中位数为−0.082，Wilcoxon 符号秩和检验显示，配对组合的得分显著高于随机组合。基于以上分析，本文得出词汇与语素义项的变化趋势存在着较强的正相关。

6.6 本章小结

本章对汉字的语义跟踪框架和方法进行了扩展，包括构建汉字语境表示模型、进行语素义项挖掘，最终得到义项的历时频率变化表示。实验结果证明，与原始的 BERT 模型相比，本文的模型在义项识别和语素义项匹配两个任务中都取得了更好的效果。义项识别的精确度从 55.08%提高到 74.19%，语素义项挖掘的 F1 得分从 59.21%提高到 75.14%。可视化和定性分析的案例证明该方法能够平滑、连续地表示词汇和语素义项的历时频率分布变化。最后，本文通过案例分析发现词汇义项与语素义项的频率和变化趋势有很强的正相关关系。

第 7 章　总结与展望

7.1 工作总结

本文面向古文字字形与字义的知识表示与整合、字形相似度计算、古文字考释与异体字形识别、字义历时变化检测等古文字研究领域的实际问题展开人工智能交叉研究，主要完成工作如下：

第一，本文对古文字领域实体与关系的数据模式进行了详细定义，并与领域专家协作，对古文字多模态数据进行半自动化的处理与标注，构建了一个专业、可扩展的历时汉字知识图谱。目前为止，该知识库收录了包含甲骨文、金文、楚简在内的约 1800 余个部首，16000 余个古文字字形，550000 余个义项，并应用于智慧古文字检索平台。

第二，本文提出了古文字字形相似度的计算方法，该方法引入古文字的部件领域知识，不仅针对古文字视觉相似性，也计算古文字的部件语义相似性。实验证明，该方法与人类专家提供的相似度分数具有更高的正相关性，斯皮尔曼相关系数为 0.8422，$p<0.01$。案例分析表明，该方法能够捕捉字源以及部件语义逻辑相关联的古文字，符合古文字学者的专业需求。

第三，本文面向古文字考释研究，提出了古文字字形识别的创新性任务，并为此设计了一个融合古文字图像、部件描述文本与字形知识图谱特征的多模态编码器模型。实验证明，与基线方法相比，本文的方法能够取得更好的识别效果，共时测试字形的准确率为 73.7%，历时组的准确率为 56.5%，并证明了图像与部件模态特征对于字形识别任务均具有贡献。

第四，本文提出了一个汉字语素含义挖掘与历时语义跟踪方法，设计了一个联合优化的双编码器汉字语境表征模型，并构建用于模型训练的数据集。该方法对现有的词汇语义跟踪方法扩展到语素层面，从词汇与语素两方面对汉字的语义变化进行检测和可视化分析，并从统计分析中发现汉字的词汇义项与语素义项的历时变化具有正相关性。

综上所述，在理论方面，本文最大的创新是提出了新颖的汉字知识图谱，并将部件语义知识融入古文字字形表示中，为古文字形近字推荐、未知字形识别等任务提出了基于部件语义知识的解决方案，并证明了古文字的多模态特征的有效性。在应用方面，尽管 AI 已经被用于古文字识别、文物修复等多个方向，但是

古文字知识表示、古文字考释的相关应用研究还很少，且具有挑战性，本文的研究成果可以为领域专家提供知识检索、形近字推荐、字形识别、字义分析等服务，辅助专家进古文字考释研究，对于我国古文字的发展与传承具有重要意义。

7.2 工作展望

在 AI 与古文字交叉研究领域，未来的最终目标是实现 AI 驱动的未知古文字考释，该智能系统需要具有对古文字领域知识和多模态数据的理解能力、推理能力、解释能力、以及与古文字领域专家的交互能力，这是一项非常具有挑战性的研究。作为对该领域的初步探索，工作尚有很多局限性，未来的展望如下:

第一，未来将对 AI 驱动的古文字考释方法进行更多维度的探索。目前古文字智能考释工作局限在字形层面，本文引入部件语义知识，将古文字考释建模为字形分类的单一的任务，方法的局限性在于难以识别那些视觉形态和部件语义特征均已经发生巨大变化的字形，并且该方法依赖于汉字字形训练集的完整性，不适用于那些缺乏异体字形训练样本的未知字形。在古文字考释中，除字形外，人类专家还会依据历史文献上下文、前人研究等重要数据，例如，将目标字形所处的出土文献上下文与相似的传世文献进行对读，从而根据语义对该位置的文字进行比较与预测。因此本文认为，未来 AI 驱动的古文字考释不会局限于单一的任务，需要研发由多种任务组成的智能系统，该系统利用古文字的多模态特征，能够从字形、字音、字义、语用多个角度对未知字形进行考释并对结果进行解释。

第二，由于对出土文物数据进行修复、知识标注与抽取的专业性与复杂性，汉字知识图谱的持续扩展与更新模式仍有待完善。未来需要进一步扩展和改进数据预处理、知识抽取与知识融合的算法框架，提升知识图谱构建的效率，持续整合汉简等更多类型的古文字以及多模态数据。

第三，未来将构建古文字领域集识别、理解与推理等多种智能任务为一体的多模态大语言模型，并研究与领域专家的交互机制。大模型是 AI 发展的趋势，未来工作将研究古文字多模态大模型预训练和应用方法，在模型中引入知识图谱的领域知识，实现文本、图像、语音等多模态数据的语义理解和细粒度对齐，提供统一的古文字知识检索、语义理解、推理考释、可视化分析等服务。同时研究多模态大模型的指令调优方法，使人类专家可以利用自然语言问答方式与模型进行交互，构建 AI 驱动的古文字研究的新模式。

参考文献

[1] XIONG J, LIU G, LIU Y, LIU M, et al. Oracle Bone Inscriptions information processing based on multi-modal knowledge graph[J]. Computers and Electrical Engineering, 2021, 92: 107173.

[2] YANG X, ARORA A, JHENG S, et al. Quantifying character similarity with vision transformers[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 13982-13996.

[3] 祁俊辉，龙华，邵玉斌，等. 基于特征向量和笔顺编码的字形相似算法研究[J]. 重庆邮电大学学报，2019，31(6): 886-891.

[4] 胡浩，李平，陈凯琪. 基于汉字固有属性的中文字向量方法研究[J]. 中文信息学报，2017，31(3): 33-40.

[5] 王东，熊世桓. 一种新颖的汉字字形相似度计算方法[J]. 计算机应用研究，2013，30(8): 2396-2397.

[6] SU T, LEE H. Learning Chinese word representations from glyphs of characters[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 264–273.

[7] 刘梦迪，梁循. 基于偏旁部首知识表示学习的汉字字形相似度计算方法[J]. 中文信息学报，2021，35(12): 48-59.

[8] CHANG X, CHAO F, SHANG C, et al. Sundial-GAN: A cascade generative adversarial networks framework for deciphering Oracle Bone Inscriptions[C]//Proceedings of the 30th ACM International Conference on Multimedia, 2022: 1195-1203.

[9] GUAN H, YANG H, WANG X, et al. Deciphering oracle bone language with diffusion models[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024: 15554-15567.

[10] WANG P, ZHANG K, WANG X, et al. Puzzle Pieces Picker: Deciphering ancient Chinese characters with radical reconstruction[C]// Proceedings of the International Conference on Document Analysis and Recognition, 2024: 169-187.

[11] HU, R, LI S, LIANG, S. Diachronic sense modeling with deep contextualized word embeddings: An ecological view[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 3899–3908.

[12] 舒蕾，郭懿鸾，王慧萍，等. 古汉语词义标注语料库的构建及应用研究[J]. 中文信息学报，2022，36(5): 21-30.

[13] SHI D, DIAO X, TANG H, et al. RCRN: Real-world character image restoration network via skeleton extraction[C]//Proceedings of the 30th ACM International Conference on Multimedia, 2022: 1177-1185.

[14] LI H, DU C, JIANG Z, et al. Towards automated Chinese ancient character restoration: A diffusion-based method with a new dataset[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2024: 3073-3081.

[15] LI B, DAI Q, GAO F, et al. HWOBC-A handwriting oracle bone character recognition database[C]//Proceedings of the 2020 second International Conference on Artificial Intelligence Technologies and Application, 2020: 012050.

[16] WANG M, DENG W, LIU C. Unsupervised structure-texture separation network for oracle character recognition[J]. IEEE Transactions on Image Processing, 2022, 31: 3137-3150.

[17] WANG M, DENG W, SU S. Oracle character recognition using unsupervised discriminative consistency network[J]. Pattern Recognition, 2024, 148: 110180.

[18] ZHANG Y, ZHANG H, LIU Y, et al. Oracle character recognition by nearest neighbor classification with deep metric learning[C]//Proceedings of the 2019 International Conference on Document Analysis and Recognition, 2019: 309-314.

[19] LI J, WANG Q, HUANG K, et al. Towards better long-tailed oracle character recognition with adversarial data augmentation[J]. Pattern Recognition, 2023, 140: 109534.

[20] WANG W, ZHANG T, ZHAO Y, et al. Improving oracle bone characters recognition via a CycleGAN-based data augmentation method[C]//Proceedings of the 2022 International Conference on Neural Information Processing, 2022: 88-100.

[21] LI J, WANG Q, ZHANG R, et al. Mix-Up augmentation for oracle character recognition with imbalanced data distribution[C]//Proceedings of the 2021 International Conference on Document Analysis and Recognition, 2021: 237-251.

[22] YUE X, LI H, FUJIKAWA Y, et al. Dynamic dataset augmentation for deep learning-based oracle bone inscriptions recognition[J]. ACM Journal on Computing and Cultural Heritage, 2022, 15(4): 1-20.

[23] XU Y, ZHANG X, ZHANG Z, et al. Large-scale continual learning for ancient Chinese character recognition[J]. Pattern Recognition, 2024, 150: 110283.

[24] ZHANG J, ZHU Y, DU J, et al. Radical analysis network for zero-shot learning in printed Chinese character recognition[C]//Proceedings of the 2018 IEEE International Conference on Multimedia and Expo, 2018: 1-6.

[25] CAO Z, LU J, CUI S, et al. Zero-shot handwritten Chinese character recognition with hierarchical decomposition embedding[J]. Pattern Recognition, 2020, 107:107488.

[26] CHEN J, LI B, XUE X. Zero-shot Chinese character recognition with stroke level decomposition[C]//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021: 615-621.

[27] HUANG G, LUO X, WANG S, et al. Hippocampus-heuristic character recognition network for zero-shot learning in Chinese character recognition[J]. Pattern Recognition, 2022, 130:108818.

[28] DIAO X, SHI D, TANG H, et al. RZCR: Zero-shot Character recognition via radical-based reasoning[C]//Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, 2023: 654-662.

[29] ZHANG C, ZONG R, CAO S, et al. Ai-powered oracle bone inscriptions recognition and fragments rejoining[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 2020: 5309–5311.

[30] ZHANG G, LIU D, SMYTH B, et al. Deciphering ancient Chinese oracle bone inscriptions using case-based reasoning[C]//Proceedings of the International Conference on Case-Based Reasoning, 2021: 309–324.

[31] GAO T, YAO X, CHEN D. SimCSE: Simple contrastive learning of sentence embeddings[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021: 6894–6910.

[32] XU Y, FENG Y, LIU J, et al. Conf-UNet: A model for speculation on unknown oracle bone characters[C]//Proceedings of Knowledge Science, Engineering and Management, 2023: 89–103.

[33] ZHOU R, WEI J, ZHANG Q, et al. Multi-Granularity archaeological dating of Chinese bronze dings based on a knowledge-guided relation graph[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023 3103-3113.

[34] 王东波，刘畅，朱子赫，等. SikuBERT 与 SikuRoBERTa: 面向数字人文的《四库全书》预训练模型构建及应用研究[J]. 图书馆论坛，2022, 42(6): 31-43.

[35] CUI Y, CHE W, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing[J]. arXiv preprint arXiv:2004.13922, 2020.

[36] JI Z, WANG X, SHEN Y, et al. CANCN-BERT: A joint pre-trained language model for classical and modern Chinese[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021: 3112-3116.

[37] ASSAEL Y, SOMMERSCHIELD T, SHILLINGFORD B, et al. Restoring and attributing ancient texts using deep neural networks[J]. Nature, 2022, 603(7900): 280-283.

[38] LUO J, CAO Y, BARZILAY R. Neural decipherment via minimum-cost flow: from Ugaritic to Linear B[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 3146-3155.

[39] DEVLIM J, CHANG M, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.

[40] 莫伯峰，邱炜琦，谢泽澄. 人工智能模拟辞例归纳的初步测试[J]. 汉语言文学研究，2021，12(3): 128-135.

[41] LI X, ZHAO X, XU J, et al. Imf: Interactive multimodal fusion model for link prediction[C]//Proceedings of the ACM Web Conference, 2023: 2572–2580.

[42] WANG P, WU J, CHEN X. Multimodal entity linking with gated hierarchical fusion and contrastive training[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022: 938-948.

[43] WISHART R, PROKOPIDIS P. Topic modelling experiments on hellenistic corpora[C]//CDH@ TLT, 2017: 39-47.

[44] KESTEMONT M, STOVER J, KOPPEL M, et al. Authenticating the writings of Julius Caesar[J]. Expert Systems with Applications, 2016, 63: 86-96.

[45] OCHAB J, ESSLER H. Stylometry of literary papyri[C]//Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, 2019: 139-142.

[46] RONG X. Word2vec parameter learning explained[J], arXiv:1411.2738, 2014.

[47] YOUSEF T, PALLADINO C, WRIGHT D, et al. Automatic translation alignment for ancient Greek and Latin[C]//Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, 2022: 101-107.

[48] FANG A, LO F, CHINN C. Adapting NLP and corpus analysis techniques to structured imagery analysis in classical Chinese poetry[C]//Proceedings of the Workshop on Adaptation of Language Resources and Technology to New Domains, 2009: 27-34.

[49] HE J, ZHU Q, CHEN Y, et al. Bronze inscriptions classification algorithm on imbalanced dataset[C]//Proceedings of the 5th International Conference on Mechanical, Control and Computer Engineering, 2020: 1715-1718.

[50] ISAAC A, HASLHOFER B. Europeana linked open data-data.europeana.eu[J]. Semantic Web, 2013, 4(3): 291-7.

[51] MERILLAS OF, RODRIGUEZ MM. An analysis of educational designs in intangible cultural heritage programmes: the case of Spain[J]. Int J Intang Herit, 2018,13:190-202.

[52] CARRIERO VA, GANGEMI A, MANCINELLI ML, et al. Pattern-based design applied to cultural heritage knowledge graphs[J]. Semantic Web, 2021,12(2):313-57.

[53] BAI B, HOU W. The application of knowledge graphs in the Chinese cultural field: the ancient capital culture of Beijing[J]. Heritage Science, 2023, 11:77.

[54] FAN T, WANG H, HODEL T. CICHMKG: a large-scale and comprehensive Chinese intangible cultural heritage multimodal knowledge graph[J]. Heritage Science, 2023, 11:115.

[55] LU L, LIANG X, YUAN G, et al. A study on the construction of knowledge graph of Yunjin video resources under productive conservation[J]. Heritage Science, 2023, 11:83.

[56] YANG S, HOU M. Knowledge graph representation method for semantic 3D modeling of Chinese grottoes[J]. Heritage Science, 2023, 11:266.

[57] WAN J, ZHANG H, ZOU J, et al. WuMKG: a Chinese painting and calligraphy multimodal knowledge graph[J]. Heritage Science, 2024, 12:159.

[58] Quan H, LI Y, LIU D, et al. Protection of Guizhou Miao batik culture based on knowledge graph and deep learning[J]. Heritage Science, 2024, 12:202.

[59] HAN X, BAI Y, QIU K, et al. Isobs: An information system for oracle bone script[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020: 227-233.

[60] GEORGE M. Wordnet: A lexical database for English[J]. Communications of the ACM, 1995, 38(11):39-41.

[61] ZHANG Y, LI B, DAI X, et al. PQAC-WN: constructing a wordnet for Pre-Qin ancient Chinese[J]. Lang Resources & Evaluation, 2017, 51: 525-545.

[62] BOND F, FOSTER R. Linking and extending an open multilingual wordnet[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013: 1352-1362.

[63] NAVIGLI R, PONZETTO S. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network[J]. Artificial Intelligence, 2012, 193: 217-250.

[64] NAVIGLI R, BLLOSHMI R, LORENZO A. BabelNet meaning representation: A fully semantic formalism to overcome language barriers[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2022: 12274-12279.

[65] GIUNCHIGLIA F, BATSUREN K, BELLA G. Understanding and exploiting language diversity[C]//Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017: 4009-4017.

[66] BATSUREN K, BELLA G, GIUNCHIGLIA F. Cognet: A large-scale cognate database[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2020: 3136-3145.

[67] LOUREIRO D, JORGE A. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 5682-5691.

[68] MAO R, LIN C, GUERIN F. Word embedding and WordNet based metaphor identification and interpretation[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 1222-1231.

[69] LORENZO A, MARU M, NAVIGLI R. Fully-semantic parsing and generation: the BabelNet meaning representation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 1727-1741.

[70] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C]//Proceedings of the 1st International Conference on Learning Representations, 2013: 1-12.

[71] KIM Y, CHIU Y, HANAKI K, et al. Temporal analysis of language through neural language models[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2014: 61-65.

[72] HAMILTON W, LESKOVEC J, JURAFSKY D. Cultural shift or linguistic drift? Comparing two computational measures of semantic change[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 2116-2121.

[73] ROSENFELD A, ERIK K. Deep neural models of semantic shift[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 474-484.

[74] YIN Z, SACHIDANANDA V, PRABHAKAR B. The global anchor method for quantifying linguistic shifts and domain adaptation[C]//Proceedings of the 32nd Conference on Neural Information Processing Systems, 2018: 9412-9423.

[75] KAISER J, KURTYIGIT S, KOTCHOURKO S, et al. Effects of pre- and post-processing on type-based embeddings in lexical semantic change detection[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021: 125-137.

[76] QIU L, TU K, YU Y. Context-dependent sense embedding[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 183-191.

[77] LEE G, CHEN Y. MUSE: Modularizing unsupervised sense embeddings[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 327-337.

[78] GIULIANELLI M, TREDICI M, FERNANDEZ, R. Analysing lexical semantic change with contextualised word representations[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2020: 3960-3973.

[79] KURTYIGIT S, PARK M, SCHLECHTWEG D. Lexical semantic change discovery[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 6985-6998.

[80] ROSIN G, GUY I, RADINSKY K. Time masking for temporal language models[C]//Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022: 833-841.

[81] TEODORESCU D, OHE S, KONDRAK G. UAlberta at LSCDiscovery: Lexical semantic change detection via word sense disambiguation[C]//Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, 2022: 180-186.

[82] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised cross-lingual representation learning at scale[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2020: 8440-8451.

[83] SUCHANEK F, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge[C]//Proceedings of the 16th international conference on World Wide Web, 2007: 697-706.

[84] 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.

[85] WU W, LI H, WANG H, et al. Probase: A probabilistic taxonomy for text understanding[C]//Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 2012: 481-492.

[86] AUER S, BIZER C, KOBILAROV G, et al. Dbpedia: A nucleus for a web of open data[C]//Proceedings of the semantic web, 2017: 722-735.

[87] XU B, XU Y, LIANG J, et al. Cn-dbpedia: A never-ending Chinese knowledge extraction system[C]//Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2017: 428-438.

[88] FANG J, WANG X, MENG Z, et al. Manner: A variational memory-augmented model for cross domain few-shot named entity recognition[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 4261-4276.

[89] CHEN J, LU Y, LIN H, et al. Learning In-context Learning for Named Entity Recognition[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 13661-13675.

[90] MA Z, DU J, ZHOU S. Noise-robust training with dynamic loss and contrastive learning for distantly-supervised named entity recognition[C]//Proceedings of the Findings of the Association for Computational Linguistics, 2023: 10119-10128.

[91] IMAI S, KAWAHARA D, ORITA N, et al. Theoretical linguistics rivals embeddings in language clustering for multilingual named entity recognition[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 139-151.

[92] CUI W, CHEN X. Exploring automatically perturbed natural language explanations in relation extraction[C]//Proceedings of the Findings of the Association for Computational Linguistics, 2023: 3454-3467.

[93] HU X, GUO Z, TENG Z, et al. Multimodal relation extraction with cross-modal retrieval and synthesis[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 303-311.

[94] SUN Q, HUANG K, YANG X, et al. Uncertainty guided label denoising for document-level distant relation extraction[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023: 15960-15973.

[95] VEYSEH A, DERNONCOURT F, MIN B, et al. Generating labeled data for relation extraction: A meta learning approach with joint GPT-2 training[C]//Proceedings of the Findings of the Association for Computational Linguistics, 2023: 11466-11478.

[96] PEREZ E, STRUB F, VRIES H, et al. Visual reasoning with a general conditioning layer[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2018: 3942-3951.

[97] WIATRAK M, ARVANITI E, BRAYNE A, et al. Proxy-based zero-shot entity linking by effective candidate retrieval[C]//Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis, 2022: 87-99.

[98] YANG D, XIAO Y. Multi-task entity linking with supervision from a taxonomy[J]. Knowledge and Information Systems, 2023, 65(10): 4335-4358.

[99] HOU F, WANG R, NG S, et al. Exploiting anonymous entity mentions for named entity linking[J]. Knowledge and Information Systems, 2023, 65(3): 1221-1242.

[100] LIU Y, TIAN Y, LIAN J, et al. Towards better entity linking with multi-view enhanced distillation[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 9729-9743.

[101] SHANG B, ZHAO Y, LIU J. Learnable convolutional attention network for knowledge graph completion[J]. Knowledge-Based Systems, 2024, 285: 111360.

[102] ZHONG H, LI W, ZHANG Q, et al. A unified embedding-based relation completion framework for knowledge graph[J]. Knowledge-Based Systems. 2024, 289: 111468.

[103] PEI S, KOU Z, ZHANG Q, et al. Few-shot low-resource knowledge graph completion with multi-view task representation generation[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023: 1862-1871.

[104] REN H, DAI H, DAI B, et al. SMORE: Knowledge graph completion and multi-hop reasoning in massive knowledge graphs[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022: 1472-1482.

[105] SHANG B, ZHAO Y, LIU J, et al. Mixed geometry message and trainable convolutional attention network for knowledge graph completion[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2024: 8966-8974.

[106] MARCHEGGIANI D, BASTINGS J, TITOV I. Exploiting semantics in neural machine translation with graph convolutional networks[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 486-492.

[107] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning[C]//Proceedings of the International Conference on Machine Learning, 2017: 1243-1252.

[108] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning[C]//Proceedings of the International Joint Conference on Artificial Intelligence, 2016: 2873-2879.

[109] KIPF T, Welling M. Semi-supervised classification with graph convolutional networks[J], arXiv:1609.02907v4, 2016.

[110] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph Attention Networks[J]. arXiv:1710.10903v3, 2017.

[111] PENNINGTON J, SOCHER R, MANNING C. Glove: Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1532-1543.

[112] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. The Journal of Machine Learning Research, 2020, 21(1): 5485-5551.

[113] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J]. 2018.

[114] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized BERT pretraining approach[J]. CoRR, 2019, abs/1907.11692.

[115] CLARK K, LUONG M, LE Q, et al. ELECTRA: pre-training text encoders as discriminators rather than generators[C]//Proceedings of the 8th International Conference on Learning Representations, 2020.

[116] HE P, LIU X, GAO J, et al. Deberta: decoding-enhanced bert with disentangled attention[C]//Proceedings of the 9th International Conference on Learning Representations, 2021.

[117] LEWIS M, LIU Y, GOYAL N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 7871-7880.

[118] ZAHEER M, GURUGANESH G, DUBEY K, et al. Big bird: Transformers for longer sequences[C]//Proceedings of the Advances in Neural Information Processing Systems, 2020.

[119] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[120] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional networks[C]//Proceedings of the Advances in Neural Information Processing Systems, 2012: 1106-1114.

[121] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale visual recognition[C]//Proceedings of the 2015 International Conference on Learning Representations, 2015.

[122] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015:1-9.

[123] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[124] HUANG G, LIU Z, MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2261-2269.

[125] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16X16 words: Transformers for image recognition at scale[C]//Proceedings of the 9th International Conference on Learning Representations,2021.

[126] KIM W, SON B, KIM I. ViLT: Vision-and-Language Transformer without convolution or region supervision[C]//Proceedings of the 38th International Conference on Machine Learning, 2021: 5583-5594.

[127] LEE K, CHEN X, HUA G, et al. Stacked cross attention for image-text matching[C]//Proceedings of the European Conference on Computer Vision, 2018: 201-216.

[128] RADFORD A, KIM J, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the 38th International Conference on Machine Learning, 2021: 8748-8763.