Interdisciplinary Science Letters 交叉科学快报, 2025, 9(4), 483-496
Published Online July 2025 in Hans. https://www.hanspub.org/journal/isl
https://doi.org/10.12677/isl.2025.94060

Hans汉斯

西夏文字的数智化：构建西夏文字组件的知识图谱

宋世超[1]，梁　循[1*]，王涵玉[1]，王　龙[2*]，杨国涛[3]，张森森[1]，王孟威[1]，李星诺[1]，葛文章[1]，杨　丽[3]

[1]中国人民大学信息学院，北京
[2]宁夏大学民族与历史学院，宁夏 银川
[3]宁夏大学管理学院，宁夏 银川

收稿日期：2025年6月9日；录用日期：2025年7月1日；发布日期：2025年7月14日

摘　要

本文研究了西夏文字的数智化问题，从构建西夏文字组件的知识图谱视角拓宽了西夏学的研究范围。本文首先给出了西夏文字的组件分解方法，然后研究了西夏文字组件的知识图谱构建方法，接着讨论了组件知识图谱的智能算法，最后搭建了西夏文字组件知识图谱的数智化平台。西夏文字偏旁部首组件通过各种排列组合后的可能结果，涵盖了超过20万个潜在的西夏文字。通过这种方式，我们可以探索未出现的西夏文字的更多可能性。本文提供的西夏文字结构和组成的系统化研究，使西夏文拆分有了方便的数智化工具，为进一步的文字解码和语言学研究奠定数智化的基础。西夏文字的组件的知识图谱，由于其组件的灵活性及计算机介入后的智能性，也许为挖掘西夏文字这个文化遗产提供更多机会。

关键词

西夏文字，西夏文字组件，知识图谱，数智化工具

Digital and Intelligent Processing of Tangut Script: Constructing Knowledge Graph of Tangut Character Components

Shichao Song[1], Xun Liang[1*], Hanyu Wang[1], Long Wang[2], Guotao Yang[3], Sensen Zhang[1], Mengwei Wang[1], Xingnuo Li[1], Wenzhang Ge[1], Li Yang[3]

[1]School of Information, Renmin University of China, Beijing
[2]School of Ethnology and History, Ningxia University, Yinchuan Ningxia
[3]School of Management, Ningxia University, Yinchuan Ningxia

[*]通讯作者。

文章引用: 宋世超, 梁循, 王涵玉, 王龙, 杨国涛, 张森森, 王孟威, 李星诺, 葛文章, 杨丽. 西夏文字的数智化: 构建西夏文字组件的知识图谱[J]. 交叉科学快报, 2025, 9(4): 483-496. DOI: 10.12677/isl.2025.94060

Received: Jun. 9th, 2025; accepted: Jul. 1st, 2025; published: Jul. 14th, 2025

Abstract

This paper explores the digital and intelligent processing of Tangut script, expanding the scope of Tangut studies from the perspective of constructing a knowledge graph of Tangut character components. We first propose a method for decomposing Tangut characters into components, followed by a systematic approach to constructing a knowledge graph based on these components. We then discuss intelligent algorithms applicable to this component-based knowledge graph and introduce a digital platform for managing and exploring the graph. The combinatorial possibilities of radicals and components in Tangut script potentially cover over 200,000 unique character forms, enabling the discovery of previously undocumented Tangut characters. This systematic study of the structure and composition of Tangut script offers a practical digital tool for character decomposition and lays a foundation for further decoding and linguistic research. The component-based knowledge graph, enhanced by its structural flexibility and computational intelligence, opens new possibilities for uncovering and revitalizing this cultural heritage.

Keywords

Tangut Script, Tangut Character Components, Knowledge Graph, Digital-Intelligent Tools

1. 西夏文字的历史及研究现状

(1) 西夏国的建立与西夏文字的创制

西夏国是 1038 年由党项民族建立的朝代，曾与北宋并立，历经 189 年。西夏文明是中华文明的重要组成部分，西夏文字是西夏文明的核心载体，富有极高的文化内涵。西夏文字诞生于 11 世纪初，属于汉藏语系的羌语支，是一种采用表意体系的合成文字，字形复杂，笔画繁多，且组成文字的组件相似度极高，被称为“人类智慧所能创造的最复杂的文字体系之一”，这些特点增加了研究的难度。

在结构上，西夏文字可分为单纯字与合体字两大类。合体字则进一步分为合成字、互换字与对称字。合成字通过多种部件组合而成，互换字可在部件上进行替换或互换，而对称字则左右结构对称。这些结构特征为西夏文字的深入研究提供了广泛的可能性，例如合成字的构造规律，组件(也称构件，为了避免与知识图谱构建的名称太相近，本文中使用组件这一名称)组成的方式和意义、利用算法判断该文字是否为对称字、文字成分的关联性分析、字形演变过程的追踪及合成造字法的独特研究。这些特性不仅具有重要的学术研究价值，也为文字学、历史学以及文化传承等领域提供了丰富的研究前景。

西夏景宗李元昊正式称帝前的公元 1036 年，命大臣野利仁荣创制西夏文字，三年后完成，共包含 6000 余字。根据西夏人梁德养(https://baike.baidu.com/item/%E6%A2%81%E5%BE%B7%E5%85%BB/23182832?fr=ge_ala)记载有 6230 个字，而俄国科兹洛夫于 1090 在原西夏黑水城遗址发现的文献有 5819 个字，日本学者统计的字为 5763 个，而中国西夏学者史金波先生认为西夏文应该在 6000 个字以上。西夏文字形体方整，笔画繁冗。西夏文字结构仿汉字，单纯字较少，合成字占绝大多数，两字合成一字居多，三字或四字合成一字者少。

西夏文字不仅记录了西夏王朝的政治、经济、文化和宗教等方面的信息，还反映了党项民族的历史和社会变迁。通过对西夏文字的研究，学者可以深入了解西夏王朝的历史背景和文化特征，从而丰富对中华文明多样性的认识。此外，西夏文字作为一种表意文字，其复杂的构造和独特的造字法也对文字学和语言学的研究具有重要参考价值。然而，西夏文字的研究同样面临挑战。由于西夏文字的字形复杂、笔画繁多，传统的手工识别方法效率低下且容易出错。大量形似字的存在进一步增加了识别的难度，给西夏文字的分类和解读带来巨大挑战。近年来，随着西夏文化研究的发展，西夏文字字符已经完成了编码和输入，其语义信息也得到了较好的归纳。但仍需要借助先进的计算机技术和深度学习方法，提升对西夏文字的识别和研究能力。

西夏文字与汉字关联紧密，“论末则殊，考本则同”，其由笔划组成，结构属性清晰明确，并且在分析构造特点时可以借鉴汉字字形的构造方式，西夏文字与中华中国传统汉字的不可分割的结构思想。

字形结构是西夏文字的核心，掌握了字形结构就掌握了识读西夏文字的重要途径。西夏文字中也存在偏旁部首的概念，但与汉字不同的是，一些文字无法单纯地用偏旁部首进行分类。所以在分析其构造特点时，要掌握和运用西夏文字自身的特点和规律，从字形和字义两方面同时入手，二者均为识别相似西夏文字的关键。

研究西夏文字不仅有助于解读古代西北社会人们生活面貌，也有助于理解中华民族文字的演化和传承过程。西夏文作为中国最早的成熟文字系统之一，在记录中华民族西北人民生活的同时，也有少数描述了中原和新疆的资料，为研究中国古代历史和文化提供了宝贵的资料。

(2) 西夏文字的研究现状

日本西夏学学者，西田龙雄研究了西夏文字的基本组成要素(即构成字形的最小单位，本文也称之为组件)，认为通过这些基本要素，按照一定的组合样式，如“上下结构”“左中右结构”等，组成了西夏文字[1] [2]。他归纳列出了 44 种组合样式。此外，西田龙雄还总结出西夏字的文字要素共有 350 种，这个工作后来成为了西夏文字学领域头等重要的基础研究。几十年后，他的这项研究成了人们设计西夏文电脑字符的先期依据。

我国西夏学学者、“中央研究院”院士龚煌城，探索了构词法和西夏文字制作上的特征，建立了西夏文字衍生过程的理论[3] [4]。

我国著名西夏学学者、中国社会科学院民族学与人类学研究所研究员、中国民族古文字研究会副会长聂鸿音先生，研究了西夏文字的构成和辅音韵尾问题[5]。我国著名西夏学学者、中国社会科学院学部委员史金波先生，从语义的角度，把西夏字形结构归纳为单纯字、会意合成字、反切合成字、间接音意合成字、左右互换近义字等五类[6] [7]，他系统地研究了西夏文字构造，提出西夏文 90%以上的合体字是依据部位合成法原则构成的。

事实上，如果结合史金波先生和西田龙雄的研究，应该不止上述的 44 种组合样式了。如果一个字包含左中右 3 个基本要素(组件)，其左中右排序有 3!=6 种情况，对上中下也有 3!=6 种情况，再复杂的组合样式其排列数会更多，例如 3 个组件(左、右上、右下)又会形成 6 种结构，(左上、左下、右)又会形成 6 种结构，所以我们粗略估计 44 种组合数会形成排列数 2000 种以上，这远远超过了人类大脑能记忆和分辨掌控的范围。

到了 20 世纪，为了满足西夏文字电脑处理的需要，学者们分析西夏文字的角度发生了根本的转移[5] [8] [9]。针对西夏文字的信息处理研究主要集中在手抄古籍文献中的文字识别问题。随着西夏文化研究的发展，西夏文字字符已经完成了编码和输入，其语义信息也得到了较好的归纳。

(3) 西夏文字的计算机数据库

尽管西夏文字常让人想到“古老”二字，但在西夏文的研究历程中，各种前沿科技的运用屡见不鲜，

并为西夏学进步做出了很大贡献。西夏学与计算机的结合就是这样的例子。早在 20 世纪 90 年代，西夏文字的计算机编码研究就已展开，通过创建专门的字库、设计专用的输入法，西夏文字得以进入数据库，不仅丰富了西夏文字的信息存储和管理方式，也促使西夏文研究模式开始从纸质化向电子化转变。这些基础性先导工作，在西夏学和计算机之间创建了学科协同的无限可能。

现时的人们更加重视对西夏“正字”的认定以及对文字组件本身的拆分和归纳，而不再留意这些组件在组字时的作用。这虽然仅仅是技术方面的问题，称不上正统的学术研究，但学术界就此对西夏字正确写法的认识则是越来越清晰了——每一笔的形态和位置都不容有丝毫含糊。

计算机具有运算速度快、存储容量大等特点，利用好这些特点，发挥计算机在大数据处理方面的优势，计算西夏学便能获得创新性发展。过去依靠人力很难保证全面性和穷尽性，而随着人工智能技术的成熟和进步，通过系统而全面的整理便可展现出创新的前景。过去由于数据量太大，做不好或不能做的题目，目前随机计算机算力越来越强大，这些题目也都将成为计算西夏学可以重点发力的方向。

西夏文字反映了党项民族在历史上的独特地位和贡献，展现了中华文明的多样性和丰富性。通过研究构建电子数据库，能够为学术研究和文化传承提供重要的基础数据，促进古文字的数字化保护和传播。

目前，国内外对西夏文字的研究主要集中在基础工作上，如文字的摹写和识别、字典和工具书的编纂、历史文献的整理和分析等。近年来，随着计算机技术和深度学习方法的发展，一些学者开始尝试将这些技术应用于古文字的识别和研究中。例如，通过图像识别技术，对西夏文字进行自动化识别和分类，提高了研究效率和准确性。

通过构建全面的电子数据库，西夏文字研究者可以更方便地访问和分析文字数据，从而揭示文字演变的规律和特点。此外，借助大数据和人工智能技术，可以对大量的西夏文字进行快速识别和分类，解读更多的古文字信息，推动中华古文字研究的深入发展。总之，西夏文字的研究具有重要的历史和文化价值，通过引入先进的计算机技术和深度学习方法，可以提高研究效率，推动相关领域的发展。研究这些古文字，不仅可以解读古代社会的面貌，还可以揭示文字的演化和传承过程，为中华文明的研究提供新的视角和方法。

通过研究西夏文字的识别方法，可以为古文字研究提供新的工具和手段，提高文字识别的效率和准确性，推动相关领域的研究进展。同时，通过构建电子数据库，能够为学术研究和文化传承提供重要的基础数据，促进古文字的数字化保护和传播。

2. 西夏文字的数智化

(1) 人工智能在古文字研究上的应用

当今，社会科学和人文科学已经走进电子化的时代。近几年 AI 已经成为牵引科技发展的火车头，NLP 又是 AI 皇冠上的明珠，所以，知识图谱作为 NLP 一种典型分支，一直在科技发展列车的头部，出现了很多高效模型，发展迅速。现如今，知识图谱不论是从其构建，还是利用其进行知识推理，都是在 AI (Artificial Intelligence, AI)加持下进行的。

近年来，人工智能技术突飞猛进，推动西夏学与计算机深度结合。人工智能使得计算机功能更强大，与西夏学结合的领域更多样化，涌现了一批引人注目的成果。在西夏学与计算机深度结合的背后，还有着更深的学科发展背景。放眼世界，“人工智能 + 古文字”的浪潮正席卷而来。以色列发现的“死海古卷”曾被称为 20 世纪最重要的考古发现，人工智能通过“虚拟拼图”帮助残卷复原，又通过辨析古卷上的笔迹类型，推测出“死海古卷”的撰写人数。著名的人工智能团队 DeepMind 与古希腊文专家合作，不仅使得古希腊文字的修复准确率不断提高，而且还确定了这些文字的书写时期。所以，西夏学与计算机的深度结合，也是这一波浪潮中的一朵浪花。

为什么人工智能会与古文字碰撞出如此多火花呢？因为二者有很多天然的契合点。人工智能近期取得突破的重要领域，恰好能与古文字密切联系起来。比如，运用广泛的语音识别、机器翻译，本质上是人工智能的自然语言处理技术，而古文字是古代语言的一种书面形式，当然也可以成为自然语言处理的对象。还有我们经常用到的搜索引擎，背后都有知识图谱的加持，利用其中蕴含的知识，搜索引擎才越来越智能。而古文字问题的解决，也离不开语言、历史、文化等知识的应用。结合知识图谱技术，便可以让这些知识服务于古文字研究。可以说，人工智能取得重要突破的这些领域，自然而然会让人与古文字研究发生联想。

在人工智能的助推下，古文字与计算机已经从初步连接开始走向深度结合，“计算古文字学”正在路上。

(2) 人工智能中的知识图谱的应用：基于人工智能知识表示学习的文字知识图谱

知识图谱本身是一种网状数据结构，它基于图的数据结构，是一种语义网络，它将复杂的知识领域以可视化的方式展示出来，更灵活、更多角度的揭示事物之间的关系，揭示事物之间有价值的信息。知识图谱的构建和使用涉及数据的组织和管理，它可以对数据进行结构化存储和高效检索，并赋予知识管理更方便的可视化度功能。

知识表示学习由于其便于计算、提高知识获取和推理的效率等优点，逐渐成为知识图谱领域的突破点之一。知识表示学习通过不断学习知识图谱中的实体和关系并将其映射到低维稠密的向量空间，实现知识的有效表示。知识表示学习按照技术应用可以划分为基于翻译的方法[10]，基于语义匹配的方法[11]，基于神经网络的方法[12]和基于图神经网络的方法[13]。在基于翻译的方法中，Santoro 等(2016) [10]提出的 TransE 方法因为具有模型高效简单的优势而广受青睐，并在此基础上提出很多改进和拓展的模型。考虑到 TransE 模型处理 1-N、N-1、N-N、对称等复杂关系的局限性，RotatE 模型，将知识图谱中的关系定义为从头实体到尾实体的二维旋转变换，并在建模和推断对称/反对称等各种关系模式中取得了较好的效果。

针对文字知识表示学习的问题，在汉字方面，结合汉字的笔画属性提出字向量表示方法，该方法具有向量维度低、空间消耗小的优点。Wang 等(2014) [14]提出了汉字的字形-语义融合嵌入方法，这种方法考虑到了汉字字形与语义因素的关联并充分利用了这一特点。在其它文字方面，Wang 等(2019) [15]设计了涵盖多种中国民族古文字知识图谱，并运用深度学习方法实现了智能问答的功能。Wen 等(2016) [16]构建了基于知识图谱的蒙古族信息检索系统，系统具备知识查询、基于 Word2Vec 的查询扩展等多个功能。

(3) 基于人工智能的汉字偏旁部首组件的知识图谱

刘梦迪和梁循(2021) [17]提出并构建了汉字拆分部首的知识图谱，并提出了基于汉字结构特点的 2CTransE 模型，进行汉字信息的表示学习并将输出的向量应用于相似度计算。该工作首先从汉字组成部件的角度出发，完成了偏旁部首组件的知识图谱建设，然后针对汉字核心结构进行分析，建立 2CTransE 模型，完成汉字语义信息的表示学习，最后将得到的实体向量用于汉字之间的相似度计算，进行目标汉字的形似字候选集的推理和讨论。

由于目前没有较为完整的汉字部件库，在实现汉字相似度判定时存在一定困难，该工作为了解决这一问题，制定汉字拆分规则及主次要部件划分规则，人工筛选汉字及其组成部件，构建了 5000 个常用汉字的部件组成库。该工作构建了偏旁部首知识图谱，横向拓展了知识图谱的应用领域。事实上，即使是对汉字，目前也鲜有研究将汉字解析的“粒度”“下沉”到“偏旁部首”组件的层面。

该工作提出的方法对于不同结构汉字的形似字查找均具有一定的参考价值，同时也说明了在构建常用汉字部件组成库时所制定的规则具有一定的合理性和有效性，为之后的研究提供了行之有效的汉字部件数据集，也为形似字的研究提供了新的思路和方法。

3. 构建西夏文字组件的知识图谱

目前已有一些西夏学研究人员开始了这方面的尝试，本文提出的西夏文字组件知识图谱，就是走向数智化的努力之一。

(1) 西夏文字知识图谱

社科院学者安波(2022) [18]曾发表研究过中国古文字知识图谱的论文《文字知识图谱构建及应用》，在该论文的表 2 中，作者完成了包括西夏文字在内的 11 种文字的知识图谱。

(2) 西夏文字偏旁部首知识图谱

在《说文解字》中，许慎对“文”和“字”的解释为，独体为“文”，合体为“字”，“字”是在“文”的基础上通过组合等方式形成的。由于本文对西夏文字的组件组成和字形结构进行研究，为了概念的清晰，借用《说文解字》中的解释，将西夏文字中的独体字称为“文”，合体字称为“字”。

汉字研究界常使用组件这一概念，它是一种由笔画组成，具有相对独立的形体，介于笔画和整字之间的形体和构造单位。在同为方块型表意文字的西夏文领域中，目前学者在研究过程也使用了西夏文字“组件”的概念，如“部件”、“偏旁部首”、“文字要素”等，但尚未有明确的界定。本文将西夏文字组件定义为具有表义、表音、标示等构造功能的形体单位。

西夏文字组件是构字的基本单位，结合学者的研究成果，且为了避免“构件”与“部件”及“组件”过于相近的命名方式混合使用造成混淆，所以本文只使用“组件”这一名词定义，并且组件分为“部首”和“部尾”两种类型。其中，“部首”大致沿用了汉字学中的意义，是参与构成文字的直接组件，具备表义或表音功能，且大部分已经被学者归纳出固定意义或在编著的西夏文字典中被列为检索部首。“部尾”则是除“部首”外，从单字中拆分出的其它文字单位，且目前没有发现可解释意义的组件。

由于本文将构建西夏文字知识图谱，与现有的中文知识图谱不同，本文基于西夏文字以“偏旁部首”作为构字“积木”的特点，降低了知识图谱的“粒度”，在“偏旁部首”层次上进行构建。可以说，这项工作将西夏文字深度嵌入了计算机的“思维”中。

西夏文字可以分为独体字和合体字。独体字在西夏学研究中也常称为单纯字，独体结构也称单体结构[3] [7]。西夏文字独体字很少，绝大部分都是组合字，非常适合以偏旁部首组件的“粒度”，建立知识图谱。由于西夏文中存在作为部首使用的独体字，即笔画简单、不可再分、字形单纯的文字，这些字既可以独立地转化为组件，也是构成其它新字的基础，为了保持知识图谱的完整性，本文的西夏“偏旁部首”知识图谱中也包括独体字，并将其构建为其中一个节点。

(3) 具有超边的西夏文字知识图谱

在网络中的超边，用于描述高阶关系，超边可以连接任何数量的节点。具有超边的图，称为超图。例如，在 10 个作者组成的作者网络中，某篇文章有 4 个作者(网络节点)共同完成，则这篇文章就形成了连接 4 个节点的超边。我们在图 1 中作一个一般性说明。图 1 中红色的线 S_1 即为超边，因为它连接了 v_1、v_2、v_6，绿色的线 S_2 也为超边，因为它连接了 v_2、v_3、v_4，兰色的线 S_3 也为超边，因为它连接了 v_1、v_4、v_5。

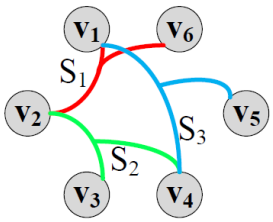

Figure 1. Hyperedge and hypergraph
图 1. 超边和超图

如果将 6 个节点填入表中(见图 2)， $v_i, i=1,\cdots,6$，如果有连接则表中填为 1，否则为 0 (为了清晰，表中空着，填 0 是为了计算机运行方便)。由于是一个对称矩阵，所以我们只需要考虑一半，即下三角即可。这个 350 × 350 的对称矩阵展示了所有超边，称为超边矩阵。

	v_1	v_2	v_3	v_4	v_5	v_6
v_1		1		1	1	1
v_2	1		1	1		1
v_3		1		1		
v_4	1	1	1		1	
v_5	1			1		
v_6	1	1				

Figure 2. Hyperedge matrix corresponding to Figure 1 (symmetrical matrix, each forming one super-edge with one color)
图 2. 对应图 1 的超边矩阵(对称阵，同 1 种颜色形成各自形成 1 个超边)

我们设想 $v_i, i=1,\cdots,6$，代表知识图谱实体，例如西夏文组件，则组件之间形成了超边的关系。

但超边中如果没有顺序和位置的表述，仅凭已知哪几个西夏文字组件参与构建某西夏文字，计算机一样无法生成该西夏文字。所以，本文需要首先根据西夏文字组件知识图谱的需要，扩展超边理论。

(4) 具有超边的西夏文字的偏旁部首组件的知识图谱

正如美国东方学家、人类学家贝特霍尔德·劳费尔(1874~1934)所说的，“西夏文字大概是人类大脑能发明出来的最复杂的系统了。”(https://baike.baidu.com/item/%E8%B4%9D%E7%89%B9%E9%9C%8D%E5%B0%94%E5%BE%B7%C2%B7%E5%8A%B3%E8%B4%B9%E5%B0%94/4651245?fr=ge_ala)。其偏旁部首组件的前后上下中位置排列组合之复杂程度，即使连计算机和高度智能的知识图谱都不得不使用超图超边的概念和模型来解决。

但是西夏文还不是目前知识图谱理论中简单的超边就能解决的。首先目前知识图谱理论中简单的超边实际是一个集合，也就是说，超边内是无序的。但是即使这个集合是有序的，也无法解决西夏文字偏旁部首知识图谱建设，因为这个超边必须还是有方位的(即表明左中右、上中下、及其互相嵌入的情况)，也就是说，西夏文字偏旁部首知识图谱建设的技术瓶颈是：基于有位有序的超边西夏文字偏旁部首知识图谱。

(5) 基于有位有序超边的西夏文字偏旁部首组件知识图谱

如果说美国东方学家劳费尔曾比喻西夏文字是“人类大脑”的产物，那么本文的工作则构建了一个“机器大脑”——它以“有位有序超边”的数学结构为支撑，模拟了西夏大臣野利仁荣的构字活动，不同之处在于，本文使用的是计算机语言和形式系统。

所谓“重新构建”，是指我们并不存储全部西夏文字的字形图像或字形编码(除少量独体字外)，而是抽象出文字中的“偏旁部首”组件，并以**有位有序超边**的方式组织这些构件。具体而言，每一个字符被建模为一个带有顺序与位置语义的超边，即一个元组 $e=\left((c_1,p_1),(c_2,p_2),\cdots,(c_n,p_n)\right)$，其中 c_i 是构件，p_i 是它在字中的语义/空间位置(如“左部”、“下部”、“音旁”等)。这种结构不仅保留了构形的层级关系，更使得图谱具备可计算的生成能力。

在此基础上，基于我们构建的《基于有位有序超边的西夏文字偏旁部首知识图谱》系统，用户可以通过简单的指令在线生成任何一个已知的西夏文字；进一步地，系统还支持组合未曾记载的构件，推演出可能存在的“新西夏文字”，从而赋予古文字体系以全新的再创造能力。

4. 西夏文字偏旁部首组件的拆解理论

我们的研究对象是西夏文字偏旁部首组件的知识图谱。该领域的研究难度较大，主要因为现有的西夏文信息化水平较低，研究资源和工具生态尚未完善。

西夏文字的分解是奠定整个工作基础的一环。通过对文字部首和组件的详细拆分，能够更好地理解西夏文字的构造逻辑和演变规律。拆分时需要依据现有的字典和文献资料，例如《夏汉字典》等，确保每个组件的合理性和准确性。我们将部首的拆分作为核心步骤，结合现代计算方法，尽可能多地保留文字的文化背景和语义关联性，这为后续的知识图谱构建提供了基础数据支持。

(1) 数据收集和处理

数据收集和处理是工作的基础，也是最具挑战性的环节之一。古西夏文字数据分散在各类文献、考古报告和实物资料中，收集这些数据需要大量的时间和精力。同时，数据的全面性和准确性也是一个重大挑战。

(2) 西夏文字的拆分原则

为了得到西夏文字组件拆分数据集并且将拆解方式合理化，制定以下三条拆分规则。

部首拆分方式。参考 Unicode 部首、《夏汉字典》、《夏俄英汉词典》等较为权威的西夏文字字典，查找对应的部首，优先作出拆分。例如“□”可以拆分成{□，□，□}，也可以拆分成{□，□}。根据这一规则，查找“□”部首得到“□”，所以选择{□，□}的拆分方式。

含有最长公共子序列拆分方式。当文字存在不同部首版本时，例如，“□”在 Unicode 部首中的部首为“□”，在《夏汉字典》和《夏俄英汉词典》中的部首分别为“□”和“□”。且与之存在紧密关系的“□”的拆分序列为{□，□}。根据该规则，对“□”选择{□}的拆分方式。这样便于其和“□”建立联系，具有很强的关联性。

当文字在不同版本字典中具有不同的部首时，且拆分部首后的剩余部分无法通过 Unicode 编码进行输入，优先选取部首进行下一步拆分。例如“□”，其在 Unicode 部首、《夏汉字典部首》中的部首均为“□”，但在《夏俄英汉词典》中的部首为“□”。所以在拆解部首“□”后，其右半部分存在多个拆解方式，如{□，□，□}、{□，□}、{□，□，□}。根据此规则，选取第三种拆解方式，即将“□”拆分成{□，□，□，□}的形式。

(3) 西夏文字的拆分方法

为了将西夏文字拆分成组件，需要按照一定逻辑将其拆分。

如上所述，西夏文字可以分为为独体字、左右结构、上下结构、圈围结构。

在计算机程序进行操作时，如果是左中右结构(或左中中右结构……)，则归为左右结构，并不断递归下去，直至全部拆分。对上下结构、圈围结构也同样操作。

如果遇上左右为主型半包围机构，例如，左包右(□，□)、左上包右下(□，□，□)、右上包左下(□，□)，则归为左右结构。

如果是上下为主型半包围机构，例如，上包下(□，□)、下包上(□)，则归为上下结构。

经过上述处理后，所有西夏文字均被拆分成组件，包括但不限于目前西夏学研究中已经发现的 350 种文字元素。

5. 西夏文字偏旁部首组件的重建理论

在完成西夏文字的分解之后，重建是对其结构和组件关系的深入理解与优化。通过结合知识图谱技术，我们能够将西夏文字的各个组件重新组合，识别出文字的演化规律。这一过程不仅帮助我们发现了部分未知的文字构字规律，还为新字的推测与创造提供了理论依据。

(1) 西夏文的组件的形式化

西夏文字之所以复杂，一定程度上就在于其具有大量形体相似度极高的组件。一方面，在西夏文字中本身存在一些形近义同的同义组件和具有密切关系的关联组件，如“□”与“□”，二者经考证均为“言”旁，为同义组件，“□”与“□”均口、齿有关，为关联组件。另一方面，因为很多组件差异甚微，在书写过程中逐渐产生笔形、构型上的变异，且这种变异多为双向变异，如“□”与“□”、“□”与“□”等(Yang 等，2022) [19]。

(2) 关联组件、形近组件和成字组件进行定义

关联组件：如果组件 a 与组件 b 意义相近且相互关联，则二者互为关联组件，记作 a~b。如“□~□”。当二者意义完全相同时，则称之为同义组件，记作 a~b，如“□~□”。

形近组件：如果组件 a 与组件 b 形成双向变异，则二者互为形近组件，记作 a↔b。如“□↔□”、“□↔□”等。特别地，西夏文中存在作为部首使用的独体字，即笔画简单、不可再分、字形单纯的文字。这些字既可以独立地转化为组件，也是具有音义属性的文字，是构成其它新字的基础，有着重要意义。所以对其做出如下定义：

成字组件：所以如果西夏文字 A 中包含组件 a，且组件 a 可以作为单独的文字来使用，则 a 为成字组件。如“□”本身为“虫”字，也是“□”字中的部首组件，所以“□”是“□”的成字组件。

任何西夏文字都可以按照书写顺序拆分成组件组合。记西夏文字为 T，则该文字 T 的组件拆分序列为 $T=\{t_1,t_2,\cdots,t_n\}$，其中 n 为文字 T 的组件数量，t_i 为该字的第 i 个组件，$i\in[1,2,\cdots,n]$，t_i 为西夏文字组件库数据集的元素。

(3) 西夏文字组件序列的基本性质

西夏文字组件序列具有有限性、有序性、元素多样性三个基本性质：

有限性：任何西夏文字的组件数都不是无限大的，即不可以被无限拆分，所以西夏文字组件序列是有限序列。

有序性：西夏文字组件序列的有序性体现在两个方面：一方面，西夏文借鉴了汉字的书写顺序，所以组件的排序也是按照从上到下、从左到右的方式；另外一方面，由于西夏文构字中的互换合成方法，即存在两个西夏文字的左右或上下两部分正好相反且字义有密切的联系。如“□(地)”与“□(坤)”、“□(差)”与“□(别)”，所以组件元素的顺序至关重要，在西夏文字组件列举展示时加以区分，如□ = {□，□}、□ = {□，□}。

元素多样性：西夏文字组件序列的组件元素分类为部首元素和部尾元素，但序列中两种类型元素的数量并不是固定的，特别是部首元素。因为一个西夏文字可以包含多个部首和部尾，所以在一个序列中可能存在多个部首元素或部尾元素。如“□(一种草名)”字中，“□”表草义，“□”表音，所以二者均为部首，□ = {□，□}这一组件序列中包含两个部首元素，不包含部尾元素。

(4) 西夏文字组件及其之间关系的建模

西夏文字组件知识图谱以西夏文字部首作为基础，展示了西夏文的组件组成以及文字之间的内在联系[20]。知识图谱中的实体有三种，分别为西夏文字、西夏文字部首和其它部尾。

我们选用 Neo4j 作为生成西夏文字组件知识图谱的图形知识图谱软件[21]。Neo4j 是一个嵌入式的 Java 持久化引擎，具备高性能和轻量级的优势。实体表现为 Neo4j 中的节点，关系表现为知识图谱中的边。我们选取 Python 中的 pyneo2 工具包作为数据的传输方式。其优点为，可以灵活地从 Python 中处理 Neo4j，实现数据的批量导入和节点、关系的增删改查。但是在最开始的导入过程中，出现了在 Neo4j 中无法正常显示西夏文字的问题。经过多方面的原因查找，最终发现是运行 Neo4j 的浏览器的设置和选择导致了这一问题的出现。当使用 FireFox 浏览器运行 Neo4j 平台时，西夏文字可以直接正常地显示出来。

而使用其它浏览器，如 Google Chrome 等需要在设置中的自定义字体选项中，修改标准字体为安装的 Tangut N4694 字体后，才可以在知识图谱中正常的显示。

知识图谱中的实体的 Lable 包含三类，分别为文字 Character、部首 Radical 和部尾 Component，总计 6693 个。关系包括三类，分别为“部首组成”、“部尾组成”和“近义字”。其中近义字关系共 362 对，为对称关系。知识图谱包含三元组 8113 个。在此知识图谱中，以部首“□”为中心进行查找，得到 24 个西夏文字实体，存在两组近义字关系。

(5) 西夏文字知识表示学习模型

西夏文字作为一个尚未被完全破解的学科领域，其结构、组件及其关联关系中仍有许多未解之谜[22]。因此，如何预测这些未知组件的含义及未确定的关联关系，成为学术界关注的重要课题之一。随着知识图谱的构建，这类预测任务可以巧妙地转化为知识图谱的表示学习。作为处理复杂关系的最先进方法之一，受欧拉公式的启发，RotatE 将头实体和尾实体映射到复杂的嵌入中，并且将关系 r 看作是头实体 h 到尾实体 t 的旋转，除了对知识图谱中的一对多、多对一、多对多的复杂关系进行知识表示学习时效果较好，还能够解决对称、反对称等关系。

我们对 RotatE 模型构建负三元组的方式进行了修改，即负三元组是由正确的三元组随机替换头实体、尾实体或关系类型得到的。因为保留了 RotatE 模型的旋转属性，且西夏文字的英文为 Tangut，所以修改后的模型称为 RotatE-T (RotatE for Tangut)。例如，在正确的三元组(□，部首组成，□)，随机替换关系类型，得到负三元组(□，部尾组成，□)，利用欧氏距离计算得到了两个实体的相似度。

6. 西夏文字组件知识图谱模型

西夏文字组件知识图谱模型的构建，旨在通过图谱的形式全面展现西夏文字及其组件的复杂关系。我们采用有位有序的超边理论，以确保知识图谱中的组件具有明确的结构和位置信息。通过对文字的部首和其他组件的有序排列，模型能够精准复现已知的西夏文字结构，并为未来新文字的发现或创造提供依据。该模型的动态扩展性使得知识图谱能够不断更新，适应新数据的融入。

(1) 基于有位有序超边的西夏文字组件的知识图谱构建

前面的西夏文字组件的知识图谱实际上是无序的，更是无位的(没有位置指定)，也就是说，其构建的知识图谱其实只是(大)部分反映了西夏文字的信息[23]。本节将进一步扩展，讨论怎样用西夏文字组件的知识图谱完整的反西夏文字的信息，即借助本节的知识图谱，计算机可以全方位的复现所有已知的西夏文字，并且也可以给出未发现的西夏文字，甚至创制新的西夏文字。

我们遇到的第一个问题是：西夏文字是以一个平面呈现的，目前知识图谱中的超边理论的研究，还没有进展到容许超边是有序的，更不要说是有位的了。

我们先以一个例子说明，然后再过渡到抽象符号。我们采用了以下方案。

先说标准方案(表位矩阵)：在西夏文字组件中，设横向排开位置分别为 i = 1, 2, 3, 4, 5, 6, 7，对上述任何一个 i，从上向下位置分别为 j = 1, 2, 3, 4, 5, 6, 7, 8 (为了和横向有区别，我们这里写 8)，所以任何一个平面的西夏文字的组件，我们做出一个 7×8 维的平面矩阵，矩阵中大部分元素为 0。下面的矩阵是一个例子。

字的结构一般有独体结构、左右结构、上下结构和包围结构。我们把独体结构直接归为偏旁部首的一个基本元素，包围结构，我们先不考虑。等描述完左右或上下结构之后，我们再返回来讨论包围结构。如果一个字是左右或左中右或左中……中右排开，我们称该字首结构为左右结构。如果一个字是上下或上中下或上中……中下排开，我们称该字首结构为上下结构。

下面讨论 7 × 8 结构矩阵(也称 7 × 8 结构表)的问题。结构矩阵也称表位矩阵(表位表)或示位矩阵(示位表)。使用表位矩阵，把对应位置直接放，都统统左上顶格放 1，其余为 0。计算机程序从横向、纵向去

遍历表位表，遇 0 则止(知道到头了)，然后遍历下一行，直至遇 0。

显然，只有包围结构会涉及到多层表，其它结构一层表就能完成表示。

不考虑有位，仅是有序的超边知识图谱，上面 6 × 7 维矩阵退化成 1 维的矩阵(即向量)，而传统的超边为无序的集合，即为向量的特例(无序向量)。

扩展方案(辅助式表位矩阵)：不过，上面标准方案也有例外，假如对 i 和 i + 1，有一个组件横跨它们，怎么解决呢？显然，规规矩矩的 7 × 8 矩阵方格表示不了，因为等于这个组件跨了两个矩阵的方格。解决方法是：在计算机算法中，只要在矩阵后面列加若干个特例，即计算机先执行标准方案。我们可以称之为扩展矩阵(或：辅助式表位矩阵)。用辅助式表位矩阵作为(西夏文偏旁部首知识图谱的)的超边，那么任意一个西夏文字就可以表示出来了。

我们知道，诸如图 2 的超边矩阵只反映了几个组件的存在于该字中，但具体位置怎么摆放，超边矩阵并没有给出，所以需要结合结构矩阵(示位矩阵)。在示位矩阵中，我们将超边矩阵给出的超边集合，按某西夏文字的结构(偏旁部首位置)，放入一个以二维坐标(x, y)为元素的示位矩阵中，其中 x 和 y 是坐标，也就是说在示位矩阵中，每一个元素不再是 1 或 0，而是一个二元数。

显然，超边矩阵只是一个辅助矩阵，示位矩阵才是核心，超边矩阵是为示位矩阵服务的，超边矩阵辅助下的示位矩阵，矩阵中的元素为从 1 至 350 的西夏文字组件。

(2) 基于人工智能预训练语言模型与组件构建策略进行西夏文字的发现和创造

西夏文字拥有独特的偏旁部首和组件，这些组成部分不仅可以用来推测尚未被发现的西夏文字结构，还可以作为创造新西夏文字的基础。基于人工智能预训练语言模型的强大能力，通过现有西夏文字的已知结构和语义信息，可以生成未知的西夏文字，并创造出新的西夏文字。

预训练语言模型在自然语言处理中的广泛应用显著提升了多种语言任务的性能。这些模型通过在大规模无监督文本上进行预训练，能够捕捉语义和上下文信息，从而在下游任务中展现出优异的性能。模型采用自注意力机制，使得模型能够并行处理长序列并捕获长距离依赖关系。这一架构极大地提升了自然语言处理任务的效率和效果，成为当前主流语言模型的基础架构。基于 Transformer 架构衍生出两类预训练语言模型，一类是以 BERT 为代表的嵌入模型，另一类是以 GPT 和 BART 为代表的生成式语言模型。BERT 通过双向编码器捕捉每个词在上下文中的双向依赖关系。BERT 通过掩码预测和句间预测学习文本的精确语义表示。BGE [24]是另一种应用于文本嵌入的开源模型，具有精确的语义表征能力。GPT 则采用自回归的生成策略，主要用于文本生成任务。与 BERT 专注于文本理解不同，GPT 在文本生成任务中具备出色的生成能力，并广泛应用于对话生成、文本补全等任务中[25]。BART 结合了 BERT 的双向编码和 GPT 的自回归解码特性，通过去噪的预训练任务自回归地生成文本，能够在自然语言生成、翻译等任务中表现优异。

在在许多语言研究领域独特问题的研究中，预训练语言模型已被应用，不仅提升了任务解决能力，还为创造新语言形式提供了可能性[26]。BERT 模型通过强化语境和语义信息来提升对话幽默识别能力，为复杂对话任务提供了新的视角。多通道 BERT，用于跨语言属性级情感分类，通过不同语言的 BERT 模型相互交互，提升了模型的语义理解能力[27]。开发了事件预训练方法，通过跨语言事件对比预训练和事件要素掩码预训练，显著提升了跨语言事件检索的性能[28]。

西夏文字结构的构造主要利用了其复杂形态和多样的构字方式。西夏文字形态的复杂性在字谜中得到了充分利用。西夏文字的构形不仅包含独立的部件，还包括这些部件的组合方式，如上下结构、左右结构等。在字谜解答过程中，结构预测者需要识别这些部件之间的关系，理解它们如何组合成完整的西夏文字[29]。

西夏文字结构是一类通过拆解西夏文字的部件和结构来完成结构预测，结构预测者需要识别西夏文

字的内部构造，如上下、左右等结构，从而推测谜底[30]。这种类型的结构预测不仅要求模型具备语言理解能力，还需要能够处理西夏文字的复杂形态和构造逻辑。

在整个结构预测过程中，模型首先通过字谜结构分析模型预测西夏文字结构的结构(如“左右”)，然后利用微调生成模型提取关键部件。基于字谜的结构和关键字信息，映射生成最终的谜底[31]。如“左岸□波涛汹涌，右岸□波澜不惊”解析为“[左右]□□”)，并映射为“□”字。这种方法不仅考虑了字谜的语义信息，还有效地整合了西夏文字的结构特性，从而提升了模型对复杂结构预测的理解和解析能力。

在西夏文字结构中，结构预测的语义往往与谜底西夏文字的结构特性直接相关[32]，因此准确预测这些结构特性对解答西夏文字结构至关重要。具体来说，文字结构的语义线索通常暗示西夏文字的结构。例如，结构预测中如果提到“高天”和“大地”，通常表明谜底西夏文字具有上下结构。类似“一面……一面……”的描述，则可能意味着谜底西夏文字的结构是左右关系。这样的语义与结构的关联为深度学习模型提供了丰富的学习信号，使其能够在大规模训练数据中捕捉这种隐含模式[33]。

通过构建和训练基于句子嵌入的文本表示模型，可以将结构预测文本有效地转化为向量空间中的点。这些向量表示捕捉了结构预测中的语义和结构特征[34]。接着，通过将这些句向量输入到微调后的分类器中，从而预测西夏文字结构的结构类型。

具体而言，通过将结构预测文本转化为向量空间中的点，可以利用这些向量来捕捉结构预测的语义和结构特点。对于每条结构预测 r，其向量表示为 v_r，通过微调的 BAAI/bge-small-zh-v1.5 (hugging face.co/BAAI/bge-small-zh-v1.5)模型获取，该模型是一个经过大量中文文字文本预训练的文本嵌入模型，能够跨语言生成高质量的句向量[35] [36]。通过构建西夏文字词表并且使用西夏文字文本对该模型进行微调，可以使其更好地适应西夏文字结构的特定任务需求。为了将结构预测的向量表示映射到其结构类型，我们设计了一个基于神经网络的分类器。分类器首先通过一个线性变换层提取特征，然后通过一个输出层进行结构类型的分类。

对有的(偏旁部首)组件是否能够再分，学术界有不同观点的，本节两种观点都算作知识图谱的实体(基本单元)，并主要采用主流观点的分法[37]。

7. 结语

通过对西夏文字组件的分解和重建，我们探索了西夏文字的基础理论和知识表示学习方法，通过知识图谱，充分挖掘西夏文字组件的内涵、以及关系。此外，我们还进行了关联性研究，充分利用西夏文字组件在知识图谱结构上的鲜明特点，使用实体相似度、Eureka 模型等完成了自动化的知识发现，例如，推断出一些西夏文化和文字的构字规律，如部首组件的含义、文字演化的特点和组件字形的变异情况等。在这些基础上，我们完成了西夏文字、组件、关系及知识的知识图谱的构建和动态获取，使西夏文拆分有了方便的数智化工具。

本文实现了西夏文字的系统化研究和知识图谱的构建。首先，西夏文字组件的分解对西夏文字进行了系统拆解，将复杂的西夏文字分解为基础的部件和组件，为后续研究奠定了数据基础。其次，西夏文字组件的重建通过知识图谱技术，将这些分解后的组件重新组合，探索组件之间的构字规律和复现西夏文字的结构。再次，西夏文字组件知识图谱模型基于有位有序的超边理论，构建了西夏文字的知识图谱模型，进一步细化和展示组件之间复杂的排列方式和关联关系。第三，西夏文字组件关联研究利用相似度分析和 Eureka 模型，深入研究组件之间的关系，推断出未记录的字形及其关联，推动了新字的发现与创造。最后，本文构建了一个西夏文字组件知识图谱，为未来的学术研究和文化保护提供重要的技术支持。

本文从数字化视角拓宽了西夏学的研究范围，涵盖了更多可能未被深入探讨的问题；其次，本文丰

富了西夏学的研究手段，引入了现代信息技术和数据分析方法，以提升研究的精准度和深度；再次，本文能够给出一大组穷举式的偏旁部首排列组合后的西夏文字可能结果，初步估计将涵盖超过 20 万个潜在的西夏文字，通过这种方式，我们可以探索西夏文字的更多可能性，为该领域提供新的数据支持；最后，本文提供了西夏文字结构和组成的系统化研究，为进一步的文字解码和语言学研究奠定数字化的基础。

基金项目

国家社会科学基金重大项目“大数据驱动的社交网络舆情主题图谱构建及调控策略研究”(18ZDA309)、教育部哲学社会科学研究后期资助项目“西夏文《圣立义海》译释与研究”。

参考文献

[1] 西田龙雄. 西夏文字解读[M]. 陈健铃, 译. 银川: 宁夏人民出版社, 1998.

[2] 西田龙雄. 西夏语研究新论[M]. 兰州: 甘肃文化出版社, 2022.

[3] 龚煌城. 西夏文字衍生过程的重建[J]. 国立政治大学边证研究所年报, 1984(15): 63-80.

[4] 龚煌城. 西夏文的意符和声符及其衍生过程[J]. 中央研究院历史语言研究所集刊, 1985(56): 719-758.

[5] 聂鸿音. 打开西夏文字之门[M]. 北京: 国家图书馆出版社, 2014.

[6] 史金波. 从文海看西夏文字构造的特点[M]. 北京: 中国社会科学出版社, 1983.

[7] 史金波. 西夏文化[M]. 长春: 吉林教育出版社, 1986.

[8] 马希荣. 夏汉字处理系统[M]. 北京: 清华大学出版社, 1999.

[9] 景永时, 贾常业. 西夏文字处理系统[M]. 银川: 宁夏人民出版社, 2007.

[10] Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D. and Lillicrap, T. (2016) Meta-Learning with Memory-Augmented Neural Networks. 2016 *Proceedings of Machine Learning Research*, New York, 23-26 June 2016, 1842-1850.

[11] Sheng, J., Guo, S., Chen, Z., Yue, J., Wang, L., Liu, T., *et al*. (2020) Adaptive Attentional Network for Few-Shot Knowledge Graph Completion. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Online, November 2020, 1681-1691. https://doi.org/10.18653/v1/2020.emnlp-main.131

[12] Shi, B. and Weninger, T. (2018) Open-World Knowledge Graph Completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**, 1957-1964. https://doi.org/10.1609/aaai.v32i1.11535

[13] Song, T.W. and Luo, J. (2021) Modeling Transitivity by Projection in Knowledge Graph Embedding. 2021 *Conference on Neural Information Processing Systems*, Vancouver, 6-14 December 2021, 351-362.

[14] Wang, Z., Zhang, J., Feng, J. and Chen, Z. (2014) Knowledge Graph Embedding by Translating on Hyperplanes. *Proceedings of the AAAI Conference on Artificial Intelligence*, **28**, 1112-1119. https://doi.org/10.1609/aaai.v28i1.8870

[15] Wang, Z., Li, L., Li, Q. and Zeng, D. (2019) Multimodal Data Enhanced Representation Learning for Knowledge Graphs. 2019 *International Joint Conference on Neural Networks* (*IJCNN*), Budapest, 14-19 July 2019, 1-8. https://doi.org/10.1109/ijcnn.2019.8852079

[16] Wen, J.F., Li, J.X., Mao, Y.Y., Chen, S.N. and Zhang, R.C. (2016) On the Representation and Embedding of Knowledge Bases beyond Binary Relations. 2016 *International Joint Conference on Artificial Intelligence*, New York, 9-15 July 2016, 1300-1307.

[17] 刘梦迪, 梁循. 基于偏旁部首知识表示学习的汉字字形计算[J]. 中文信息学报, 2021, 35(12): 47-59.

[18] 安波. 文字知识图谱构建及应用[J]. 中国科技资源导刊, 2022, 54(1): 76-82.

[19] Yang, L., Fan, J., Xu, S., Li, E. and Liu, Y. (2022) Vision-Based Power Line Segmentation with an Attention Fusion Network. *IEEE Sensors Journal*, **22**, 8196-8205. https://doi.org/10.1109/jsen.2022.3157336

[20] 杨志高. 建国五十年来我国西夏文献整理研究简述[J]. 古籍整理研究学刊, 2002(2): 40-44.

[21] Zhang, S., Liang, X., Tang, H., Zheng, X., Zhang, A. and Ma, Y. (2023) DuCape: Dual Quaternion and Capsule Network-Based Temporal Knowledge Graph Embedding. *ACM Transactions on Knowledge Discovery from Data*, **17**, 1-19. https://doi.org/10.1145/3589644

[22] 王龙. 西夏文写本《瑜伽师地论》卷五十九考释[J]. 西夏研究, 2020(3): 9-20.

[23] 杨志高. 1999 年西夏学研究综述[J]. 宁夏大学学报(人文社会科学版), 2000, 20(4): 90-93+98.

[24] Rebuffi, S., Kolesnikov, A., Sperl, G. and Lampert, C.H. (2017) iCaRL: Incremental Classifier and Representation Learning. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 5533-5542. https://doi.org/10.1109/cvpr.2017.587

[25] 徐洋, 蒋玉茹. 强化语境与语义信息的对话幽默识别模型[J]. 中文信息学报, 2022, 36(4): 73-80.

[26] 陈潇, 王晶晶. 基于多通道 BERT 的跨语言属性级分类[J]. 中文信息学报, 2022, 36(2): 121-128.

[27] 胡星武, 桂韬, 张奇. 基于对比学习的表达式偏序关系建模[J]. 中文信息学报, 2023, 37(4): 166-174.

[28] 杜建录. 西夏学[M]. 银川: 宁夏人民出版社, 2007.

[29] 李星诺, 梁循. 基于西夏文字组件知识图谱的实体相似度研究[J]. 中文信息学报, 2024(37): 54-169.

[30] Zhang, S.S., Liang, X., Wang, L. and Guan, Z.Y. (2025) Integrating Large Language Models and Mobius Group Transformations for Temporal Knowledge Graph Embedding on the Riemann Sphere. *AAAI Conference on Artificial Intelligence* (*AAAI*), Philadelphia, 25 February-4 March 2025, 23667-23676.

[31] Liang, X., Song, S., Niu, S., Li, Z., Xiong, F., Tang, B., *et al*. (2024) UHGEval: Benchmarking the Hallucination of Chinese Large Language Models via Unconstrained Generation. *Proceedings of the* 62*nd Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Bangkok, August 2024, 5266-5293. https://doi.org/10.18653/v1/2024.acl-long.288

[32] Liang, X., Wang, H., Song, S., Hu, M., Wang, X., Li, Z., *et al*. (2024) Controlled Text Generation for Large Language Model with Dynamic Attribute Graphs. *Findings of the Association for Computational Linguistics ACL* 2024, Bangkok, August 2024, 5797-5814. https://doi.org/10.18653/v1/2024.findings-acl.345

[33] Fan, A., Lavril, T., Grave, E., *et al*. (2020) Addressing Some Limitations of Transformers with Feedback Memory. https://arxiv.org/abs/2002.09402

[34] Galkin, M., Trivedi, P., Maheshwari, G., Usbeck, R. and Lehmann, J. (2020) Message Passing for Hyper-Relational Knowledge Graphs. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Online, November 2020, 7346-7359. https://doi.org/10.18653/v1/2020.emnlp-main.596

[35] Glorot, X., Bordes, A. and Bengio, Y. (2011) Deep Sparse Rectifier Neural Networks. 2011 *International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, 11-13 April 2011, 315-323.

[36] Guan, S., Jin, X., Guo, J., Wang, Y. and Cheng, X. (2020) NeuInfer: Knowledge Inference on N-Ary Facts. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, 6141-6151. https://doi.org/10.18653/v1/2020.acl-main.546

[37] Guan, S., Jin, X., Wang, Y. and Cheng, X. (2019) Link Prediction on N-Ary Relational Data. *The World Wide Web Conference*, San Francisco, 13-17 May 2019, 583-593. https://doi.org/10.1145/3308558.3313414