面向“以A为B”构式语义场景的汉语框架识别数据集构建

杨沛渊[1,‡]，苏雪峰[1,2,‡]，李俊材[1,‡]，闫智超[1,‡]，柴清华[3,⋆]，李茹[1,4,⋆,†]
[1]山西大学 计算机与信息技术学院，山西 太原 030006
[2]山西工程科技职业大学 现代物流学院，山西 晋中 030609
[3]山西大学 外国语学院，山西 太原 030006
[4]山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006
[‡]{971859815，455375251，1251972979，751824801}@qq.com
[⋆]{charles，liru}@sxu.edu.cn

摘要

汉语中普遍存在一些语义场景，其语义核心不是以单个词语呈现，而是通过句子中的某个特定结构来表达。然而当前公开发表的数据集中，只有极少数的数据集将这种特定结构作为语义单元进行研究。汉语框架语义知识库是进行汉语深层语义分析与推理的优质资源，目前其激活框架的基本单位均为句中的一个词。本文以汉语框架语义知识库为基础，引入构式语法，使用2020《人民日报》语料库，以“以A为B”构式为例，建立了基于“以A为B”构式的汉语框架识别数据集，包含23849条例句，相应框架141个。本文使用多个汉语框架识别模型及大语言模型在该数据集上进行了实验，并针对传统框架识别模型在以构式为目标词的框架识别任务中由于目标词信息匮乏导致的识别困难问题，提出了基于目标词转化和数据增强的两种方法，使模型准确率达到了88.19%，有效提升了模型挖掘构式蕴含的深层语义信息的能力。

关键词： 框架语义学；构式语法；“以A为B”；框架识别

Dataset for Recognizing Chinese Semantic Frames based on the Semantic Scenario of the *“Yi A Wei B”* Construction

Peiyuan Yang[1,‡], Xuefeng Su[1,2,‡], Juncai Li[1,‡], Zhichao Yan[1,‡], Qinghua Chai[3,⋆], Ru Li[1,4,⋆,†]
[1]School of Computer and Information Technology，Shanxi University
[2]School of Modern Logistics, Shanxi Vocational University of Engineering Science and Technology
[3]School of Foreign Languages，Shanxi University
[4]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education
[‡] {971859815, 455375251, 1251972979, 751824801}@qq.com
[⋆] {charles, liru}@sxu.edu.cn

Abstract

In Chinese, specific sentence structures often carry semantic meaning, a characteristic underexplored in existing datasets. Leveraging the Chinese Frame Semantic Knowledge Base, we conducted deep semantic analysis, focusing on the structure *“Yi A Wei B”* found in the 2020 *“People’s Daily”* corpus. We created a frame recognition dataset with this structure, comprising 23,849 example sentences and 141 corresponding frames. Experiments featuring multiple frame recognition models revealed recognition difficulties due to insufficient target word information. Two approaches, target word transformation and data enhancement, increased model accuracy to 88.19%, thereby enhancing the model’s ability to extract construction-embedded semantics.

Keywords: Frame Semantics , Construction Grammar , *“Yi A Wei B”* , Frame Identification

1 引言

汉语框架网（Chinese FrameNet，CFN）(You and Liu, 2005)是以Fillmore (1976)的框架语义学为理论基础，以汉语真实语料为依据，参照伯克利大学的框架语义网（FrameNet，FN）(Baker et al., 1998)构建的汉语框架语义知识库，包括框架库、句子库和词元库。目前已有1397个框架，涉及18360个词元，同时也对69592条语句进行了标注。

目前汉语框架数据集中激活框架的词元均为单一词汇，然而，在某些句子中，独立的目标词难以充分体现复杂的语义场景。如图1所示，图中的语句基于现有的汉语框架语义知识库，只能将“为”当做目标词，激活【范畴化】框架，但【范畴化】框架并不适配该句子的语义场景。在此句话中，“以...为...”是一个表达语义的整体，表示“网络购物等”代表了“数字消费”。以“以...为...”为整体作为“目标词”激起【代表】框架，更能准确体现该句子的语境及含义。因此，为了丰富框架的语义表达，本文引入了构式语法，增加了以构式为“目标词”的框架语义解析数据，提升了框架语义解析能力，进一步实现对语言的深度理解。

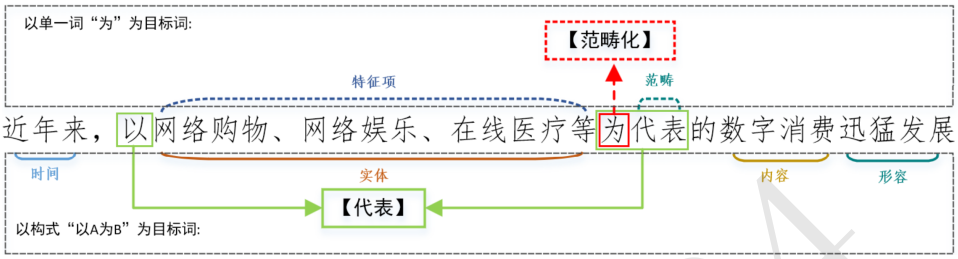

图 1: 框架语义角色标注示例

构式语法(Fillmore et al., 1988)由Fillmore教授于1988年最先提出，其主张语言是由固定的、有意义的单位组成，这些单位被称为构式，既可以是简单的词或短语，也可以是复杂的句子或话语。如，“爱买不买”、“爱说不说”、“爱要不要”对应的构式是“爱V不V”，该构式是一个表达语义的整体，表示对某行为不在意或无所谓。而本文选取的“以A为B”构式是汉语中出现较早的一种结构，它不但在文言文中的使用频率较高，而且在现代汉语中也有着很大的作用(谭世勋, 1985)，还具有一定的代表性和挑战性。一方面，该构式高度抽象，有着非常丰富的子结构，如凝固型构式：以人为本、以食为天等；半凝固型构式：以...为主、以...为例等。另一方面，该构式的语义涵盖范围广泛，能够激活多种框架，如【代表】框架、【方式】框架、【重要性】框架等。再者，该构式在句中能充当各种成分，如主语、宾语、状语等。该构式以其灵活多变的结构和丰富的语义成分，展现出了显著的代表性，是目前框架语义知识库中亟待解决的典型问题之一。

2020年《人民日报》中有新闻报道38540篇，其中有近三分之一篇报道中存在“以A为B”构式。这说明了该构式广泛存在于真实语料中，具有很高的实用价值。此外，在CFN多年的建设中，也存在许多单个目标词激活框架后无法进行语义解析的情况，其中近八分之一的问题语句中存在“以A为B”构式。因此，本文以“以A为B”构式为例，尝试引入构式语法来解决现有汉语框架语义知识库所存在的问题。

本文基于2020《人民日报》语料库，标注了构式及其对应框架的例句23849条，梳理出相应框架141个，构建了超过206万字规模的构式-框架数据集。并且以该数据集为基础，提出了一种面向构式的框架识别方法，并在多种现有模型上进行了实验。实验结果表明，本文提出的方法能有效挖掘构式所蕴含的深层语义信息，显著提高模型针对以“以A为B”构式为目标词的框架识别准确率。

2 相关工作

山西大学自2004年开始在英文框架网(FN)的基础上建立汉语框架网(CFN)，基于汉语框架

* 基金项目：国家自然科学基金重点项目（61936012）；山西省科技合作交流专项项目（202204041101016）；山西省重点研发项目（202102020101008）；国家自然科学基金面上项目（62376144）
† 通讯作者 Corresponding Author

语义资源，在框架排歧、零形式识别与填充、篇章结构生成与关系识别等任务上，取得了一系列成果，为汉语深层语义分析与推理提供支撑。如李济洪et al. (2010)等人基于CRF在CFN数据集上进行语义角色标注，屠寒非et al. (2016)等人提出一种基于主动学习的方法并取得一定的效果，Wang et al. (2020)等人提出基于自注意力机制的汉语框架语义角色标注方法以获取句子的长距离信息。2023年山西大学首次公布汉语框架网数据集(CFN2.0)并开放汉语框架语义解析评测(Li et al., 2023a)。目前汉语框架网中框架语义资源包括框架库、词元库、例句库、篇章库等都已具有一定规模，然而其词元库中的目标词均为单一词汇，无法表示构式的深层语义。

构式语法是认知语言学理论体系的重要组成部分。其理论发端于20世纪80年代末Fillmore (1988)等学者对习语的研究，是在反思、批判生成语法及其衍生的其他语法知识模型的基础上产生、发展起来的。该理论一经问世，就受到广泛关注，目前已成为国际语言学界研究的重要内容(文旭and 司卫国, 2021)。

目前的汉语构式研究较多集中在对具体构式的个案研究，且大多集中在构式的语义描写、构式的语用功能特色、构式的认知机制等方面。如孙瑜等人对“先X为敬”构式的生成机制和语用功能的研究，魏在江and 赵帮华 (2024)对隐喻式否定构式的语义生成机制的研究。而这些研究对于汉语构式知识的形式化表示以及在计算机自动分析中的应用等都还很少涉及。

北京大学为了推进构式语法理论在汉语语法领域的研究，并使构式语法理论在计算机领域发挥实际效用，建设了现代汉语构式知识描述数据库(詹卫东and 王佳骏, 2022)，采用类似词库的方式，将真实语料中实际运用的构式形式逐条收录，并详细描写每个构式的内部构成情况、构式整体的语法、语义、语用属性。

本文在北京大学现代汉语构式知识描述数据库基础上，采用其构式界定标准、分类方式和术语体系，以“以A为B”构式为例，将汉语框架网与构式相结合。一方面，丰富了框架的语义表达，解决了部分汉语框架网中存在的不足；另一方面，为构式语法理论在中文信息处理等计算机领域发挥作用提供了新的方案。

3 构式-框架语义和数据资源构建

3.1 待标注语料筛选

《人民日报》语料库以其规模宏大、主题丰富、规范性强及可靠性高的特点，成为进行各种中文自然语言处理训练的优质选择。如表1所示，本研究选取《人民日报》语句共960418条，根据是否含有构式中的词及是否表达构式含义分为三类。

类别	样例	数量
无“以”和“为”：	记者实际探访了独子山矿区。	924016
有“以”和“为”但不是构式：	...没有根本利害冲突应该完全可以成为相互需要的合作伙伴。	12553
含有构式：	近年来，以网络购物、网络娱乐、在线医疗等为代表的数字消费迅猛发展	23084

表 1: 人民日报语料分类

确定含有构式中的词（“以”和“为”）后，根据是否表达构式含义分为了正负两类。其中正类是指含有“以A为B”构式的语句,例如: (1)近年来，以网络购物、网络娱乐、在线医疗等为代表的数字消费迅猛发展。(2)今年的政府工作，要以应对国际金融危机、促进经济平稳较快发展为主线。这两句话中的“以网络购物、网络娱乐、在线医疗等为代表”和“以应对国际金融危机、促进经济平稳较快发展为主线” 就是“以A为B”构式，并且句(1)符合“代表”框架所描述的场景；句(2)符合“方式”框架所描述的场景。

相对而言，负类是指含有“以A为B”样式，但不表达“以A为B”构式含义的语句，例如: (1)...没有根本利害冲突应该完全可以成为相互需要的合作伙伴。” (2)天然雪质地松软，并不适合比赛。人造雪可以打造出符合竞技场地技术规范要求的赛道为运动员提供公平竞赛条件。这两句话中都含有“以”和“为”，但是“以”和“为”不单独成词，没有直接的联系，并不属于本文研究的“以A为B”构式。

针对区分正负两类数据的问题，本文对抽取出的35637条正负类数据进行了深入的语言学分析，建立了一套符合“以A为B”构式的语言学特征的分类体系。其中主要包含了如下三条规则：

(1)包含的“以”和“为”必须在同一短句中，不允许出现跨标点的情况（顿号除外）。(2)在分词时，“以”和“为”必须被分为单独的词，不允许出现“以”和“为”和其他词粘连的情况。(3)在进行语义依存分析时，“以”和“为”之间必须存在直接关系，且“为”依存于“以”。

根据以上三条规则，本文对含有“以”和“为”的五万余条语料进行了筛选，并选取了其中的1000条数据进行了人工验证。经过测试，根据规则区分出的正负两类数据与人工筛选的吻合率达到了90.6%。证明了所提出规则的有效性，因此，研究依据这套分类体系，从35637条含有“以”和“为”的人民日报数据中选取了23849条正类数据作为下一步框架识别任务中的待标注语料。

3.2 数据标注

3.2.1 标注规范

汉语框架语义网语义标注的基本单位是由一个框架承担词和若干框架元素组成的语义结构(Li et al., 2024)。框架承担词包括动词、形容词和名词，它们是标注工作的着眼点，称之为目标词。本研究构建的数据集尝试将汉语框架语义网中现有的目标词由单一词汇扩展为构式，即以构式为激活框架的词元。每个标注实例中的目标词（构式）由系统依据位置信息给定，标注人员根据标注实例的语义信息确定待标注实例中目标词的选择是否合理。在确定标注实例的目标词选择无误后，标注人员需要遵循以下三条标注规范进行正式标注：

(1)依据标注实例的语义信息确定目标词边界。句中的“以A为B”构式在“为”之后可能存在修饰限定成分，如“茶马古道位于我国西南地区，以马帮为主要运输方式...”这句话中，“以...为主要运输方式”是一个完整的“以A为B”构式，其中“B”为“主要运输方式”，“主要运输”是对方式的修饰和限定。因此本句中的目标词边界应为“以...为主要运输方式”，而不是“以...为主要”。标注人员在进行标注时需根据语义信息，确定目标词的边界，确保标注数据的准确性。

(2)依据框架含义及框架所对应的框架元素信息确定该句中目标词所匹配的框架。如图2所示，CFN中的框架是一个树状结构，框架之间存在上下位继承关系。例如：【有意行为】框架是一个较为上位的框架，其框架含义为：有意志的生命所进行的活动；【有意创造】框架继承自【有意行为】框架，其框架含义为：创造者可能由成分创造出一个新的实体，创造物；【生产制造】又继承自【有意创造】框架，其框架含义为：生产者出于商业目的，用某种资源生产产品。可以看出，越靠上位的框架，其概念越抽象，适用范围越广；越靠下位的框架，其概念越具体，适用范围越单一。在为待标注实例匹配框架时，会遇到多个框架都匹配的情况。要求标注人员对框架语义知识库中的框架关系有深入了解，根据框架之间的继承关系，遵循“下位框架优先”的原则，选择待标注实例所匹配的框架中框架概念最具体的框架。

(3)在实际标注过程中，有些待标注句子本身存在错误，如构式位置匹配有误或句子长度过长的问题。对于这类句子，标注人员需要根据实际情况，对句子进行适当的调整或删减，以确保标注数据的准确性。

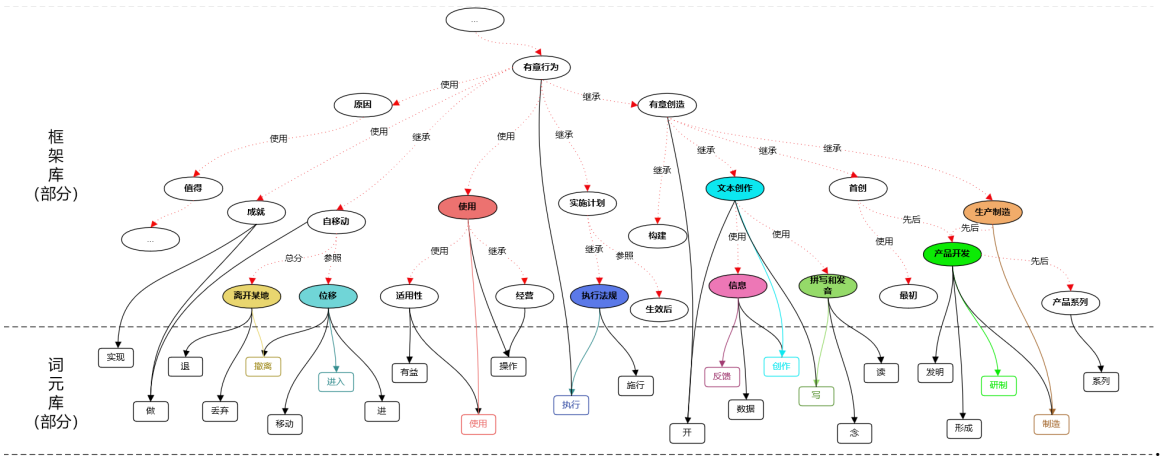

图 2: 框架树

3.2.2 待标注实例分配

由于“以A为B”构式所表达的含义与“B”有着密切的关系。为了确保人工标注的一致性及准确性，尽量让每个标注人员及审核人员所标注、审核的数据都具有相同或相近的含义。因此在进行人工标注前，根据“B”的词性对“以A为B”构式进行了分类，并将归类后的数据按类别分配于不同的标注人员。如表2所示，分类后确保了每个标注人员分配到的待标注实例均为同一类别。

“B”类别	待标注实例
“B”为名词	...要**以**应对国际金融危机、促进经济平稳较快发展**为**主线。 **以**规范制度和制约权力**为**核心，针对腐败现象易发多发的领域和环节...
“B”为形容词	可是，现在的人们不但不复**以**窄窄金莲为美，反异口同韵的诋**为**丑恶。 ...对什么也有主张，而且**以**扯谎**为**荣。
“B”为动词	我们的敌人，哼，只**以**流血**为**享受，而毫无禁忌。 这是由心中平静而然，并不是**以**退**为**进。

表 2: 构式分类

此外，在分配任务时使用了贪心算法，该算法给每个标注人员和审核人员分配数据时，会尽可能分配标注人员和审核人员曾经标注和审核过的“以A为B”构式，充分的保证了数据标注的准确、高效。

3.3 标注流程

本研究基于汉语框架语义知识库及新构建的“以A为B”构式数据集，开展了语料标注工作。整个标注分为两个阶段。第一阶段组织标注人员使用少量数据进行预标注，根据标注结果对标注人员进行讲解及培训。确保标注人员对任务有较深的理解。依据第一阶段的标注经验，对标注流程和系统进行优化，第二阶段展开大规模的标注工作。为了使人工标注更加便捷高效，研究还构建了一个轻量级的标注网站。

参与标注的人员分为标注人员，审核人员及答疑人员，本文将对一个句子的标注称为一个“标注实例”，一个标注实例包含以下基础信息：(1)含有待确定框架的“以A为B”构式完整语句。(2)经过筛选的所有待选择框架。每个标注实例都会被分配给两名不同的标注人员，标注人员根据语义信息及框架含义进行框架选择。在标注人员标注完成后，标注实例进入审核流程。一审人员会对两个标注人员的标注结果提出意见，然后交由二审人员进行最终的审核。二审人员给出最终答案后，经审核的实例结果会返回给标注人员，如果标注人员无异议，则根据最终结果进行学习。如果标注人员有异议则由答疑人员进行解答并给出最终答复。答疑老师会针对标注遇到的问题及时对数据或知识库进行更新或修改。详细标注流程如图3所示。

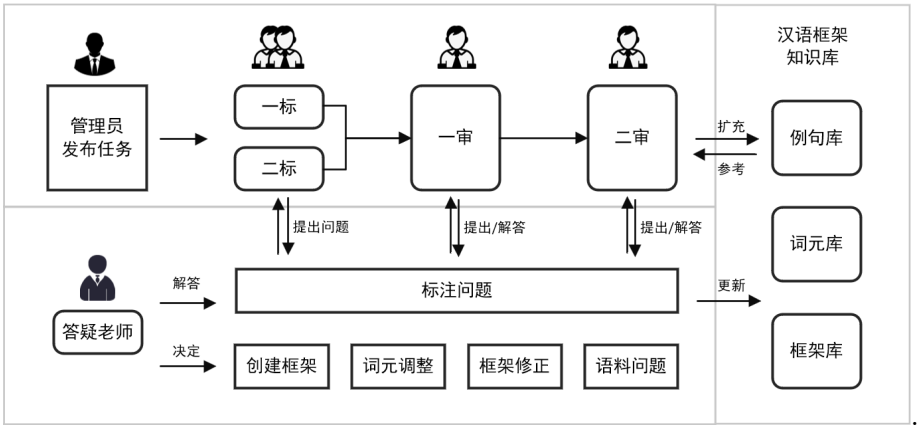

图 3: 标注流程

3.4 数据集规模及分布

本研究共标注了23849条数据。其中有138个句子长度过长，243个句子中构式的位置未被正确识别，还有384个句子本身存在错误。最终获得了23084条高质量的标注语料，每条语料都被匹配了唯一的框架，该构式-框架数据集共涉及到141个框架，其中131个框架已存在于汉语框架语义知识库中，还有10个框架待创建。数据集中各框架的分布如图4所示。

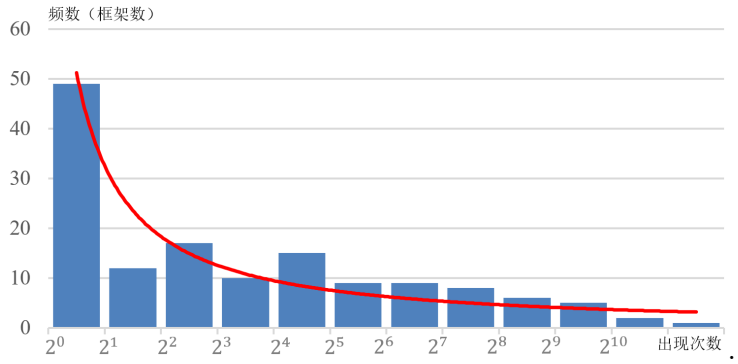

图 4: 框架数量频数分布直方图

由图可见，存在大量框架仅具有少数例句的情况，甚至超过半数的框架仅具有20条以下的例句，而例句数最多的框架则具有2256条例句，虽然呈现长尾分布现象，但符合人类在进行自然语言描述时的现实规律，这种现象增加了数据的复杂性。这些特点对框架语义分析模型提出了较高的要求。

此外，本文对“以A为B”构式进行了梳理，按照“B”的内容对数据集进行了分类，整理出不同类型的“B”内容共计413种。如图5所示，存在一种构式类型对应多种框架的情况，也存在一种框架对应多种构式类型的情况，即构式类型与框架不是简单的一对一的情况。这样的分布情况使得构式-框架数据集具有较高的复杂性，对框架识别模型的识别能力提出了更高的要求。模型不能简单的通过构式的匹配来确定框架，必须学习到构式在句中的深层次含义。

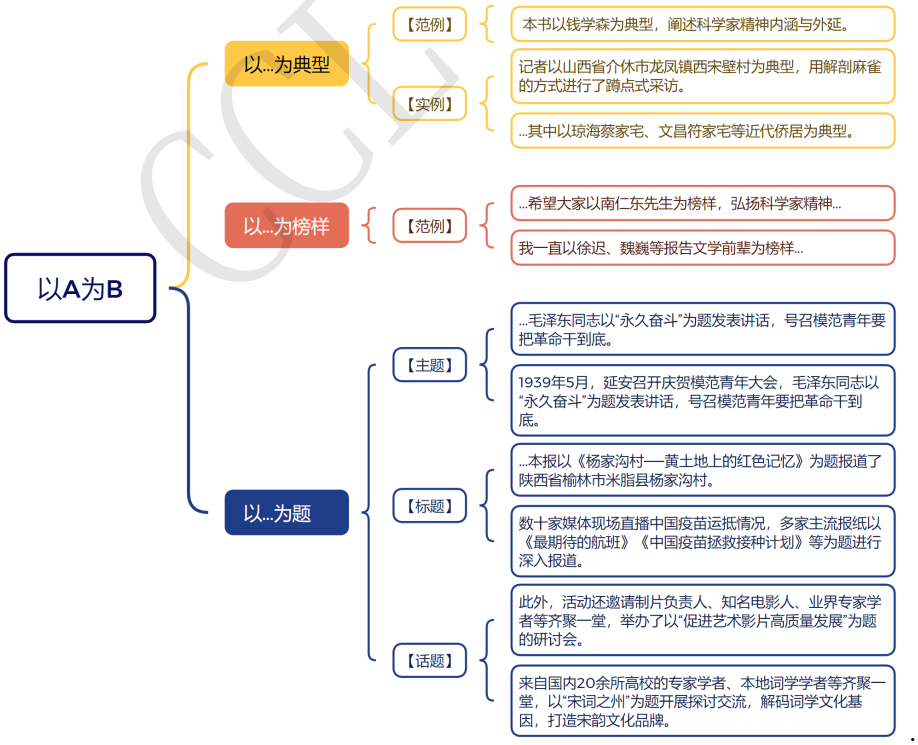

图 5: 构式类型与框架对应关系

4 以构式为目标词的框架识别

框架语义解析(Frame Semantic Parsing, FSP) 最早由Gildea and Jurafsky (2002)基于英文框架网数据FrameNet提出，是一种细粒度语义分析任务，能表示特定词（目标词）在句子中所能激活的语义场景（框架）以及该场景对应的语义角色。因此，FSP被广泛地应用于机器阅读理解(Guo et al., 2020b; Guo et al., 2020a; 王智强et al., 2016)、文章摘要抽取(Guan et al., 2021a; Guan et al., 2021b)、关系抽取(Zhao et al., 2020)及文本生成(谭红叶et al., 2018) 等多种自然语言处理任务。而框架识别（Frame Identification 以下简称FI）是进行FSP的前提(Su et al., 2021b)，该任务是给定可激活框架的目标词，根据上下文语境，从多个所属框架中选取最符合该目标词语境的语义框架的任务。

该任务的形式化定义如下：给定包含目标词的句子S ，记为$S = (w_1, w_2, ..., w_n)$，其中w_i为组成句子的第i个词，其中$1 \leq i \leq n$。待识别目标词记为w_t，$w_t \in S$。要求通过上下文的语义场景从给定的框架库$F = \{f_1, f_2, ...f_n\}$中选择出合适的框架f_t，记为式(1)：

$$f_t = \underset{f_i \in F, w_t \in S}{\operatorname{argmax}} P(f_i | S, w_t) \tag{1}$$

本文中所有框架识别实验均使用accuracy作为评价指标，其中N'为所有样本总数，y_i为每一个样本的原始标签，$\hat{f}(x_i)$为模型预测结果。具体定义如公式(2)所示：

$$Acc = \frac{1}{N'} \sum_{i=1}^{N'} I\left(y_i = \hat{f}(x_i)\right) \tag{2}$$

4.1 模型介绍

EMARoberta(Huang et al., 2023) 一种基于多优化策略的框架识别方法，采用多种优化策略以解决模型不稳定的问题，提高模型鲁棒性。采取的优化策略主要有以下三种：(1)指数滑动平均(Exponential Moving Average,EMA)(Klinker, 2011)：给予近期数据更高权重，对模型参数做平均，使得模型参数的更新与一段时间内的历史取值有关，可以提高模型在测试集上的鲁棒性。(2)Warm-up策略：训练开始前先使用一个较小的学习率进行一定的迭代次数，使得模型逐渐适应数据集的特征，让模型的权重更新更加平稳，减少训练时的震荡和不稳定性。(3)FGM对抗训练(Miyato et al., 2021)：对embedding层在梯度方向添加扰动，引入噪声，既提高了模型的泛化能力，又提高了模型的鲁棒性。

RoPEBert(Li et al., 2023b) 一种基于旋转位置编码的框架识别方法(Su et al., 2021a)，针对在框架识别中会丢失目标词与整体句子之间的位置信息关系的问题，使用了一种基于旋转式位置编码的方法来计算实体之间注意力信息，然后进行分类和抽取，该方法使用旋转矩阵对绝对位置进行编码，同时将显式的相对位置依赖性纳入自注意公式中。

4.2 数据集划分

本研究将之前构建的标注语料数据集按照7：1：2的比例划分为训练集，验证集，测试集。具体情况如表3所示。

数据集	框架数	标注句子数	平均长度
训练集	82	16242	92.7
验证集	68	2367	83.8
测试集	80	4675	84.1

表 3: 数据集划分

本文在划分数据集时遵循以下三条规则：(1)训练集中的框架应该在验证集和测试集中出现。(2)相同的框架，但是“B”不相同，一些“B”在测试集中出现，但是在训练集中不出现。(3)相同的“B”但是框架不相同，则有些框架在训练集中出现，有些框架在测试集中出现。

采用以上三条规则对本文构建的构式-框架数据集进行划分，测试模型在进行框架识别时是否具备一定的泛化能力和排歧能力，在面对复杂数据集时能否正确的识别出目标框架。

4.3 实验结果及分析

本文使用CCL2023-Eval汉语框架语义解析评测任务所提供的基线模型(Li et al., 2023a)，EMARoberta模型及RoPEBert模型作参考，三种模型均采用原论文给定的参数设置。

在本文构建数据集的测试集上，最终实验结果如表4所示。其中Acc是模型所预测的所有测试集数据整体的准确率，AccR2是所有根据划分规则二划分后的数据准确率，AccR3是所有根据划分规则三划分后的数据准确率。

Models	Acc(%)	AccR2(%)	AccR3(%)
Bert	13.63	8.48	6.57
EMARoberta	38.26	**38.29**	12.39
RoPEBert	**42.81**	17.49	**32.62**

表 4: 实验结果

实验结果表明，在目标词为单一词汇的框架识别任务中，表现较好的模型在直接迁移到以构式为目标词的框架识别任务中时表现欠佳。本文选取了部分三种模型都存在的错误进行案例分析，如表5所示。其中，依据划分规则(2)，在划分时将所有“以A为代表”的被标注为【代表】框架的标注实例放入了训练集中，但是没有放入“以A为标志”的被标注为【代表】框架的标注实例。这一规则旨在考察模型在进行框架识别时是否具备泛化能力。实验结果表明，三种模型的泛化能力较弱，导致了模型在寻找“以A为标志”的数据所匹配的框架时，出现了较大的错误率。没有将“以A为标志”的被标注为【代表】框架的数据正确的识别为【代表】框架。而根据划分规则(3)，将“以A为典型”的标注为【实例】框架的标注实例放入了训练集中，但是没有放入“以A为典型”的标注为【范例】框架的数据。这一规则旨在考察模型在进行框架识别时是否具备排歧的能力。实验结果表明，三种模型的排歧能力也较为薄弱，使模型将绝大部分“以A为典型”的数据都笼统的识别为了【实例】框架，没有根据上下文语境对“以A为典型”构式和【范例】框架、【实例】框架进行深入理解，导致了错误的产生。这也与传统的框架识别模型在进行框架识别时，过分依赖目标词所蕴含的语义信息有关。在以构式为目标词的框架识别任务中，构式所蕴含的信息不像单一词汇那样丰富、明显，需要模型根据大量的数据进行学习、挖掘，而这也是现有的框架识别模型所欠缺的。

构式类型	标注实例	模型识别框架	正确框架
代表	近年来，以网络购物、网络娱乐、在线医疗等为代表的数字消费迅猛发展。	代表	代表
标志	我们相信，以北京冬残奥会的成功举办为标志...	目标	代表
典型	记者以山西省介休市龙凤镇西宋壁村为典型，用解剖麻雀的方式进行了蹲点式采访。	实例	实例
典型	本书以钱学森为典型人物，阐述科学家精神内涵与外延。	实例	范例

表 5: 部分框架识别结果

4.4 面向构式目标词的框架识别方法

传统的利用单一目标词框架识别任务中，目标词为句子中的单一词汇，含有一定的词义信息帮助模型进行识别，但是在以构式为目标词的框架识别任务中，目标词为一种结构，如本文的“以A为B”构式，目标词为“以”，“为”这两个字，本身蕴含的信息较少，不能辅助模型进行框架识别。为了解决这一问题，研究尝试将目标词除“以”，“为”这两个字外，再加入“以A为B”构

式中的核心信息“B”。因为“以A为B”构式中的“B”往往与构式的含义有着密切的关系。于是，目标词从仅有“以”和“为”两个字转变为“以”，“为”和“B”，通过这种方法来增强模型对构式的理解，提高模型的排歧能力。

针对解决模型泛化能力不足的问题，本文尝试以“B”为重要提示信息，进行数据增强，向训练集中增加了一定数量“B”的近义词信息，即使用近义词替换“B”中的词汇，不改变其他信息。增强方式如图6所示。

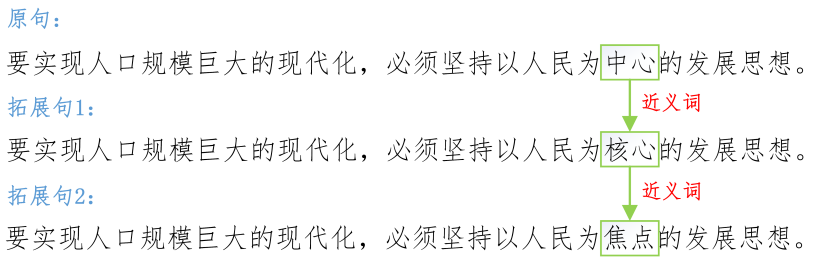

图 6: 数据增强方式

经实验证明，在经过数据增强（Data Augmentation）及目标词转化（Target Word Conversion）后，三种模型的性能均有显著的提高，实验结果如表6所示。这证明了我们提出的两种方法的有效性。

Models	Acc(%)	AccR2(%)	AccR3(%)
Bert	78.53	83.69	38.46
EMARoberta	**88.19**	**88.17**	80.76
RoPEBert	82.46	80.69	**88.46**

表 6: 实验结果

4.5 消融实验

为了证明数据增强及目标词转化的有效性，我们进行了消融实验，将数据增强及目标词转化的方法从模型中剔除，最终的实验结果如表7所示。可以得出：(1)将数据增强模块去掉后，模型准确率在Bert模型上下降了27.81%；在EMARoberta模型上下降了12.7%；在RoPEBert模型上下降了4.82%。说明了数据增强方法在以构式为目标词的框架识别任务中的有效性。(2)将目标词转化模块去掉后，模型准确率在Bert模型上下降了45.91%；在EMARoberta模型上下降了28.66%；在RoPEBert模型上下降了17.34%。说明了目标词转化方法能够有效增强模型对句中构式的理解，也证明了目标词转化模块在以构式为目标词的框架识别任务中的有效性。

Models	-DA(%)	-TWC(%)
Bert	50.72	32.62
EMARoberta	75.49	59.53
RoPEBert	77.64	62.15

表 7: 消融结果

4.6 在大模型上的实验

在本文所构建的数据集的基础上，选取了1000条数据进一步在GPT3.5(Ouyang et al., 2022)和Gemini(Team et al., 2023)上进行了测试。实验主要针对Zero-shot和Few-shot两种不同场景设定，利用如图7所示的方法构建了提示模版。在Zero-shot场景下，提示模板中不提供任何带有答案的信息，仅说明任务需求，评估大语言模型自身是否具有框架语义的相关知识，并分析其能否利用相关知识解决框架识别问题。在Few-shot场景下，提示模板中提供了三条带有

正确答案的示例样本，使得大语言模型能更好的理解任务需求，从而评估大语言模型能否有效利用少量的示例信息来提升框架识别能力。

图 7: 提示模版构建

实验结果如表8所示。由此可见，面对以构式为目标词的框架识别任务，大语言模型在Few-shot场景下的框架识别能力要显著优于Zero-shot场景，这表明大语言模型具备的上下文学习能力在框架语义分析任务中能够发挥一定作用，能够根据提示样例更好的理解任务需求。然而，无论是在Zero-shot还是Few-shot场景下，大语言模型在以构式为目标词的框架识别任务中的表现与传统模型相比仍存在一定的差异。

Models	zero-shot(%)	few-shot(%)
Gemini	23.5	37.5
GPT-3.5	12.8	21.5

表 8: 大模型框架识别结果

5 结论

本文以2020年《人民日报》语料库为基础，选取了23849条具有“以A为B”构式的语句作为待标注对象，依据汉语框架标注规范进行了人工标注。最终建立了具有141个框架，超过206万字规模的构式-框架数据集。同时利用了多种框架识别模型及大语言模型进行测试，既验证了数据集的质量，也为框架识别模型的改进提供了一定的参考。验证了大语言模型在以构式为目标词的框架识别任务中距传统模型还有一定差距。文章进一步丰富了基于语言认知的汉语框架语义知识库构建理论，为框架语义知识库扩充和复杂语义场景分析打下一定基础。

然而，本文构建的构式-框架数据集仍然存在一定的局限性，之后将会从以下几点继续改进现有的资源：(1)目前数据集中所覆盖的构式单一，拓展构式种类，增加数据集的多样性；(2)数据集中有较多框架包含的标注实例较少，使用《人民日报》语料库以外的其他语料库，如高考阅读理解语料库等,增加框架下的标注实例数量。

参考文献

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Charles J Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, pages 501–538.

Charles J Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Charles J Fillmore. 1988. The mechanisms of" construction grammar". In *Annual Meeting of the Berkeley Linguistics Society*, pages 35–55.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Yong Guan, Shaoru Guo, Ru Li, Xiaoli Li, and Hongye Tan. 2021a. Frame semantic-enhanced sentence modeling for sentence-level extractive text summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4045–4052.

Yong Guan, Shaoru Guo, Ru Li, Xiaoli Li, and Hu Zhang. 2021b. Integrating semantic scenario and word relations for abstractive sentence summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2522–2529.

Shaoru Guo, Yong Guan, Ru Li, Xiaoli Li, and Hongye Tan. 2020a. Incorporating syntax and frame semantics in neural network for machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2635–2641.

Shaoru Guo, Ru Li, Hongye Tan, Xiaoli Li, Yong Guan, Hongyan Zhao, and Yueping Zhang. 2020b. A frame-based sentence representation for machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 891–896.

Shutan Huang, Qiuyan Shao, and Wei Li. 2023. CCL23-eval 任务3系统报告:基于多任务pipeline策略的汉语框架语义解析(system report for CCL23-eval task 3: Chinese frame semantic parsing based on multi task pipeline strategy). In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 105–112, Harbin, China, August. Chinese Information Processing Society of China.

Frank Klinker. 2011. Exponential moving average versus moving exponential average. *Mathematische Semesterberichte*, 58:97–107.

Juncai Li, Zhichao Yan, Xuefeng Su, Boxiang Ma, Peiyuan Yang1, and Ru Li. 2023a. CCL23-eval 任务3总结报告:汉语框架语义解析评测(overview of CCL23-eval task 1:Chinese FrameNet semantic parsing). In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 113–123, Harbin, China, August. Chinese Information Processing Society of China.

Zuoheng Li, Xuanzhi Guo, Dengjian Qiao, and Fan Wu. 2023b. CCL23-eval 任务3系统报告:基于旋转式位置编码的实体分类在汉语框架语义解析中的应用(system report for CCL23-eval task 3: Application of entity classification model based on rotary position embedding in chiness frame semantic parsing). In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 94–104, Harbin, China, August. Chinese Information Processing Society of China.

Ru Li, Yunxiao Zhao, Zhiqiang Wang, Xuefeng Su, Shaoru Guo, Yong Guan, Xiaoqi Han, and Hongyan Zhao. 2024. A comprehensive overview of cfn from a commonsense perspective. *Machine Intelligence Research*, pages 1–18.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2021. Adversarial training methods for semi-supervised text classification.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback, 2022. *URL https://arxiv. org/abs/2203.02155*, 13:1.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021a. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

Xuefeng Su, Ru Li, Xiaoli Li, Jeff Z. Pan, Hu Zhang, Qinghua Chai, and Xiaoqi Han. 2021b. A knowledge-guided framework for frame identification. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5230–5240, Online, August. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Xiaohui Wang, Ru Li, Zhiqiang Wang, Qinghua Chai, and Xiaoqi Han. 2020. 基于self-attention的句法感知汉语框架语义角色标注(syntax-aware Chinese frame semantic role labeling based on self-attention). In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 616–623, Haikou, China, October. Chinese Information Processing Society of China.

Liping You and Kaiying Liu. 2005. Building chinese framenet database. In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on*.

Hongyan Zhao, Ru Li, Xiaoli Li, and Hongye Tan. 2020. Cfsre: Context-aware based on frame-semantics for distantly supervised relation extraction. *Knowledge-Based Systems*, 210:106480.

屠寒非, 李茹, 王智强, and 周铁峰. 2016. 一种基于主动学习的框架元素标注. 中文信息学报, 30(4):44–55.

文旭and 司卫国. 2021. 中国构式语法研究20年：回顾与展望. 解放军外国语学院学报, 44(05):43–51+160–161.

李济洪, 王瑞波, 王蔚林, and 李国臣. 2010. 汉语框架语义角色的自动标注. 软件学报, 21(4):597–611.

王智强, 李茹, 梁吉业, 张旭华, 武娟, and 苏娜. 2016. 基于汉语篇章框架语义分析的阅读理解问答研究. 计算机学报, 39:795–807.

詹卫东and 王佳骏. 2022. 面向计算的构式研究：现状、问题与展望. 语言学研究, pages 39–51.

谭世勋. 1985. 试论"以a为b"结构的发展. 华南师范大学学报（社会科学版）, pages 90–96.

谭红叶, 闫真, 李茹, and 敬毅民. 2018. 迈向创造性语言生成:汉语幽默自动生成的探索. 中国科学:信息科学, 48:1497–1509.

魏在江and 赵帮华. 2024. 隐喻式否定构式的语义生成机制研究. 外国语文, 40:21–32.