NNP-TDGM: 基于最近邻提示表征的术语DEF生成模型

沈思嘉　王裴岩*　王胜任　王立帮
沈阳航空航天大学计算机学院，辽宁沈阳110136
沈阳飞机工业（集团）有限公司信息化中心，辽宁沈阳110003
wangpy@sau.edu.cn　wangsr002@avic.com
{shensijia,wanglibang}@stu.sau.edu.cn

摘要

该文研究基于HowNet的知识库描述语言语法体系的术语DEF自动生成问题，提出基于最近邻提示表征的术语DEF生成模型（NNP-TDGM），将训练集中的术语DEF构造为外显记忆集，在解码器生成（首）义原或关系时，检索与待预测术语概念结构相同或相近的术语所蕴含的核心概念，重要属性和关系类型，辅助模型完成DEF的生成，解决解码器在低频样本上训练不充分的问题。另外，通过提示预训练语言模型获得术语及术语定义内蕴涵概念信息的语义表征向量，改善编码器表征能力不足的问题。经实验验证NNP-TDGM模型生成术语DEF的义原-关系-义原三元组F1值达到31.84%、关系F1值达到53.12%、义原F1值达到51.55%、首义原F1值达到68.53%，相对于基线方法分别提升了3.38%，1.45%，1.08%，0.48%。

关键词： Hownet知网；DEF生成；最近邻；提示表征

NNP-TDGM: Nearest Neighbor Prompt Term DEF Generation Model

Sijia Shen　Peiyan Wang　Shengren Wang　Libang Wang
School of Computer Science, Shenyang Aerospace University,
Shenyang, Liaoning 110136, China
Information Center, AVIC Shenyang Aircraft Company Limited,
Shenyang, Liaoning 110003, China
wangpy@sau.edu.cn　wangsr002@avic.com
{shensijia,wanglibang}@stu.sau.edu.cn

Abstract

To address the issue that the automatic generation of term Definition (DEF) based on Knowledge Database Mark-up Language of HowNet, this paper proposes a Nearest Neighbor Prompt Term DEF Generation Model (NNP-TDGM). This model constructs the term DEF in the training set as explicit memory sets.When generating (first) sememes or roles,the decoder retrieves the core concepts,important attributes,and role types contained in terms with the same or similar conceptual structures to help the model generate DEF. This model addresses the insufficient training problem of the decoder on low-frequency samples. Furthermore, the model improves the problem of insufficient representation ability of the encoder by prompting the pre-trained language model to obtain semantic representation embeddings containing the conceptual information in terms and term definitions. Experimental evidence shows that the sememe-role-sememe triple F1 reaches 31.84%, the role F1 reaches 53.12%, the sememe F1

* 通讯作者

基金项目:全国科技名词审定委员会科研项目(YB2022015)，国家自然科学基金(U1908216)

reaches 51.55% and the first sememe F1 reaches 68.53%. These values show an improvement of 3.38%, 1.45%, 1.08% and 0.48%, respectively, compared to the baseline.

Keywords: Hownet , DEF generation , Nearest Neighbor , Prompt representation

1 引言

HowNet (董振东et al., 2007)是中文自然语言处理领域使用最为广泛的语义知识库，用于描述汉语和英语词语所代表的概念、以及概念与概念之间、概念的属性与属性之间的关系。HowNet使用义原及义原间关系描述概念，使用知识库描述语言（Knowledge Database Mark-up Language，KDML）作为概念形式化表示方法，形式化表示的概念被称为DEF（Definition）。HowNet被广泛应用于词语表征(Niu et al., 2017)、词义消歧 (Hou et al., 2020)、语言建模 (Gu et al., 2018)、反向字典构建 (Zheng et al., 2020)等任务中。但是HowNet主要面向通用领域，针对专业领域如航空领域的自然语言理解任务其支持能力尚显不足。

张等 (2014)面向航空领域的术语语义知识库构建方法进行了研究，但其采用的是人工或半自动的方法，效率有限。在自动生成义原方面，Xie等 (2017)将词和义原转化为向量，利用协同过滤和矩阵分解方法为新词推荐义原。Jin等 (2018)将字符级内部信息引入汉语词汇义原预测中，提出字符增强语义预测框架，用来缓解由于外部信息的排他性带来的问题。杜等 (2020)利用局部相关性，词语的各个义原和定义文本中的词语之间存在语义匹配关系，提出义原相关池化模型为新词预测义原。Qi等 (2020)构建基于多语言百科全书BabelNet的义原知识库，对多种语言的词语进行义原标注。上述研究将义原预测任务简化为多标签分类任务，未考虑义原的层级结构。Ye等 (2022)提出基于Transformer的义原树生成模型，该方法判断义原与义原之间是否存在关系，没有判断义原与义原间的关系类别，不是完整的词语DEF。吕 (2022)提出基于Seq2Tree (Dong and Lapata, 2016)的DEF自动生成方法，该方法能够生成完整的术语DEF，某种程度上解决了术语DEF自动生成问题。

本文以文献(吕嘉, 2022)方法为基础，进一步研究了术语DEF自动生成问题，提出基于最近邻提示表征的术语DEF生成模型（Nearest Neighbor Prompt Term DEF Generation Model, NNP-TDGM）。主要面向两方面问题：一是由于现有方法以术语为输入，采用“端到端”的编码器-解码器模型生成术语DEF。编码器较难获取术语所蕴含的核心概念及其重要属性，以及核心概念与属性之间，属性与属性之间的复杂关系，进而无法形成准确的语义表征向量。二是由于HowNet有2214种义原与89种关系，在有限训练样本的条件下，义原和关系的训练数据分布不均匀，导致解码器在低频样本下训练不充分。针对上述问题，本文提出两种解决途径：

（1）提出基于提示的术语语义信息表征方法。现有研究表明预训练语言模型蕴涵大量的语义知识 (Zhang et al., 2019)，基于此，本文通过提示预训练语言模型获得术语及术语定义内蕴涵概念信息的语义表征向量，并将其整合到解码器中，改善编码器表征能力不足的问题。

（2）提出基于最近邻的DEF生成方法。将训练集中的术语DEF分解为首义原、义原和关系，首义原代表术语的核心概念，义原细化描述术语的重要属性，关系揭示术语核心概念与属性及属性与属性之间的复杂关系，分别构造首义原外显记忆集，义原外显记忆集和关系外显记忆集。在解码器生成（首）义原或关系时，通过检索与待预测术语概念结构相同或相近的术语中蕴含的核心概念、重要属性和关系类型，辅助模型完成DEF生成，从而弥补低频样本下模型训练不充分问题。

本文在航空术语数据集上进行了实验，NNP-TDGM模型生成术语DEF三元组（义原-关系-义原）$F1$值达到31.84%、关系$F1$值达到53.12%、义原$F1$值达到51.55%、首义原$F1$值达到68.53%，相对于基线方法分别提升了3.38%，1.45%，1.08%，0.48%。实验结果验证了所提出方法的有效性。

本文第2节介绍文本表示与检索增强的相关研究。第3节详细阐述基于最近邻提示表征的术语DEF生成模型，包括整体框架与各模块实现细节。第4节为实验结果与分析，介绍实验数据集，对比模型、评价指标、参数设置、生成DEF实验、消融实验以及对各模块影响分析实验。最后为结论与展望。

2 相关研究

2.1 文本表示

文本表示是将文本转化为连续稠密向量 (曾骏et al., 2024)。预训练语言模型BERT (Devlin et al., 2019)在文本表示方面得到了广泛的应用。Yan等 (2021)和Gao等 (2021)通过对比学习的方法优化预训练语言模型的句表示，同时也缓解了句表示的奇异值快速下降的问题。Su等 (2021)发现传统机器学习中的白化操作可以增强BERT句子表征的各向同性。Jiang等 (2022)通过提示和对比学习相结合的方法提取预训练语言模型的句表示来减小词嵌入带来的偏置。

在生成术语DEF中，可利用的信息仅为术语及术语定义，并且较难构造出大量的训练数据，因此无法使用需要大量训练语料或丰富上下文的文本表示方法。因此，本文提出基于提示的术语语义信息表征方法，提示预训练语言模型获得术语及术语定义内蕴涵概念信息的语义表征向量，能够充分利用预训练语言模型具有的大量语义知识。

2.2 检索增强

最近，检索增强方法被应用于自然语言处理的各类任务中。其方法是构建外显记忆模块，在模型预测过程中检索外显记忆，获得实例，通过实例与模型联合预测，改善预测结果，如Khandelwal等 (2020)提出的K近邻机器翻译，Guu等 (2020)提出的K近邻语言模型。Wang等 (2020)在人机对话中检索相似历史对话并结合知识图谱，提升对话的流畅性与信息丰富度。Fan等 (2021)提出通过检索增强为对话生成任务提供先验知识。Liu等 (2023)提出一种用于低资源神经机器翻译的K最近邻迁移学习方法，该方法提高了低资源神经机器翻译的性能。

现有研究主要面向翻译与语言模型等序列生成任务，在树形结构生成任务中还鲜有研究。本文在树形结构的术语DEF生成任务上研究检索增强方法，可为结构化数据生成提供研究借鉴。

3 基于最近邻提示表征的术语DEF生成模型

3.1 问题定义

本文目标是输入术语e，自动生成术语e的DEF。举例说明，术语“主燃区”的DEF为“{place|地方:RelateTo={burn|焚烧},modifier={primary|主}}”，其DEF的义原树结构如图1所示。“place|地方”为首义原，代表“主燃区”的核心概念为某个特定位置，而“burn|焚烧”和“primary|主”为义原，分别指明了“主燃区”的具有的主要功能和属性。“RelateTo”和“modifier”为关系，揭示了属性与核心概念存在相关和修饰关系。

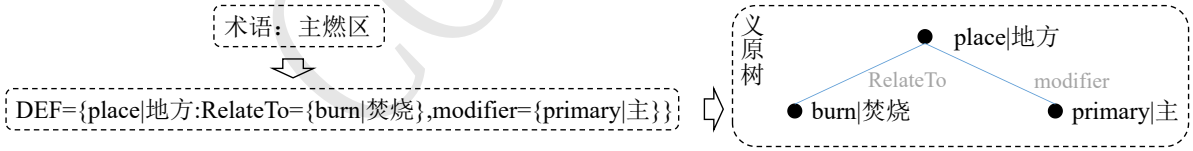

Figure 1: DEF和义原树示例

DEF生成问题可转化为义原树生成问题 (Ye et al., 2022; 吕嘉, 2022)。一个义原树可以分为从根节点到叶子节点的多条路径。假设不同的路径都是独立的，则生成义原树的概率如式(1)$p(T|e)$所示。每个路径的概率如式(1)$p(S|e)$所示。利用马尔可夫假设，进一步将路径分解为父-子路径，基于父节点生成子节点的概率如式(1)$p(s_t|e)$所示。

$$p(T|e)=\prod_{S\in T}p(S|e);\quad p(S|e)=\prod_{t=1}^{N_S}p(s_t|e,S_{0:t-1});\quad p(s_t|e)=\prod_{t=1}^{N_S}p(s_t|e,s_{t-1}); \tag{1}$$

其中，T表示的是术语e的义原树，S表示T中的路径，由义原及关系构成，是“义原-关系-义原-关系-...-义原”序列。N_S是S的长度，t为当前时间点，s_t是S当前的路径状态，对于义原树路径，s_t为义原或是关系，$S_{0:t-1}$是从根节点到S的$t-1$个时间步的路径。

3.2 模型总框架

本文提出基于最近邻提示表征的术语DEF生成模型（NNP-TDGM）。本模型沿袭了SRS（吕嘉，2022)模型所用的义原框架填充思想，得到形如“义原-关系-义原”形式的输出。NNP-TDGM主要由基于提示的术语语义信息表征模块、树形解码器、KNN模块、基于最近邻的术语DEF生成模块构成。NNP-TDGM模型结构如图2所示。

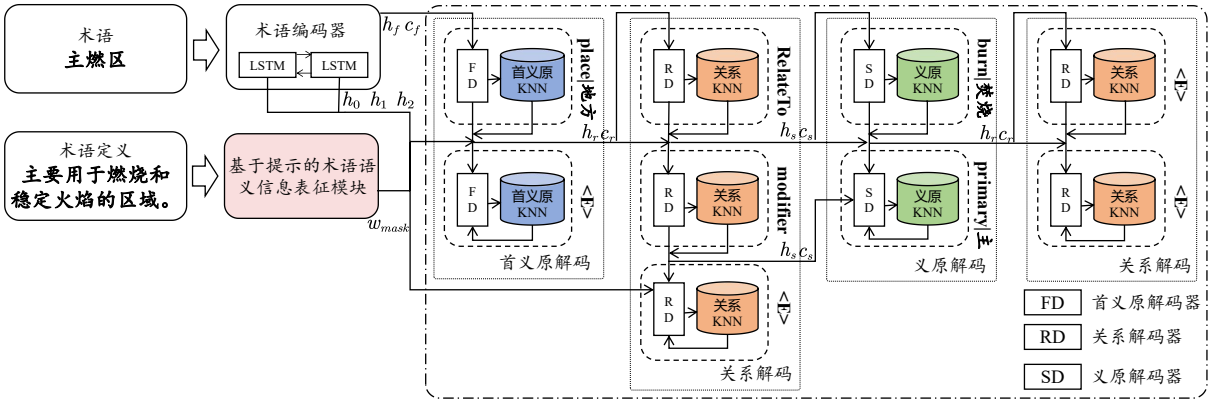

Figure 2: NNP-TDGM模型结构图

3.3 基于提示的术语语义信息表征模块

本文提出一种基于提示的术语语义信息表征方法。通过提示从预训练语言模型中获得术语语义的向量化表征，作为补充信息辅助DEF生成。以术语及术语定义为基础，构建3种提示信息：术语、定义、术语和定义。其中，术语定义是对术语的解释说明文本，例如术语“主燃区”的定义是“主要用于燃烧和稳定火焰的区域。”。每种提示信息包括4类模板，分别命名为CLS、MASK、PMASK和LMASK。模板中“[术语]”代表术语，“[定义]”表示术语定义，“[MASK]”表示掩码提示符。每类模板与3类提示信息组合形成12种提示模板，如TCLS表示使用术语信息的CLS类模板。12种提示模板列入表1中。CLS类模板是将提示信息$\{x_{CLS}, x_0, \cdots, x_n, x_{SEP}\}$输入到预训练语言模型中，获取输出层CLS向量作为语义表征向量$\boldsymbol{w}_{CLS}$，如式(2)所示。MASK类模板将输入（术语、定义、术语及定义）与提示符“[MASK]”拼接，并将拼接之后的提示语句$\{x_{CLS}, x_0, \cdots, x_{MASK}, x_{SEP}\}$输入到预训练语言模型中，获取输出层“[MASK]”提示符输出的向量作为语义表征向量$\boldsymbol{w}_{MASK}$，如式(3)所示。PMASK和LMASK类模板都将输入放到更加平滑的语境中，同样添加提示符“[MASK]”，获取预训练语言模型“[MASK]”提示符输出的向量作为语义表征向量$\boldsymbol{w}_{MASK}$，如式(3)所示。

$$\boldsymbol{w}_{CLS}, \boldsymbol{w}_0, \cdots, \boldsymbol{w}_n, \boldsymbol{w}_{SEP} = BERT\left(x_{CLS}, x_0, \cdots, x_n, x_{SEP}\right) \tag{2}$$

$$\boldsymbol{w}_{CLS}, \boldsymbol{w}_0, \cdots, \boldsymbol{w}_{MASK}, \boldsymbol{w}_{SEP} = BERT\left(x_{CLS}, x_0, \cdots, x_{MASK}, x_{SEP}\right) \tag{3}$$

模板名称	模板	模板名称	模板
TCLS	[术语]	DCLS	[定义]
TMASK	[术语][MASK]	DMASK	[定义][MASK]
TPMASK	这句话“[术语]”的意思是[MASK]	DPMASK	这句话“[定义]”的意思是[MASK]
TLMASK	“[术语]”这句话的意思是[MASK]	DLMASK	“[定义]”这句话的意思是[MASK]
ACLS	[术语][定义]		
AMASK	[术语][定义][MASK]		
APMASK	这句话“术语:[术语]的定义是:[定义]”的意思是[MASK]		
ALMASK	“术语:[术语]的定义是:[定义]”这句话的意思是[MASK]		

Table 1: 12类提示模板

3.4 树形解码器

树形解码器包含首义原解码器（First Sememe Decoder，FD）、关系解码器（Role Decoder，RD）和义原解码器（Sememe Decoder，SD）3部分。3个解码器都采用长短期记忆神经网络（LSTM）模型。3个解码器的结构相同，以首义原解码器为例，模型结构如图3中首义原解码器所示。将语义表征向量$\boldsymbol{w}_{MASK}$和解码器当前时刻的隐藏层向量$\boldsymbol{h}_t$以及术语编码器上下文向量$\boldsymbol{g}_t$进行拼接得到拼接向量$\boldsymbol{o}_t$，生成的路径状态概率如式(4)所示。由于树形解码器包含三个解码器，FD得到的概率是$p_{MF}(s_t|e)$，RD得到的概率是$p_{MR}(s_t|e)$，SD得到的概率是$p_{MS}(s_t|e)$。

$$
\begin{aligned}
\boldsymbol{h}_t, \boldsymbol{c}_t &= \text{LSTM}\left(\boldsymbol{h}_{t-1}, \boldsymbol{c}_{t-1}\right) \\
score_i^t &= \frac{\exp\left\langle \boldsymbol{h}_i \cdot \boldsymbol{h}_t \right\rangle}{\sum_{j=1}^{l} \exp\left\langle \boldsymbol{h_j} \cdot \boldsymbol{h}_t \right\rangle} \\
\boldsymbol{g}_t &= \sum_{i=1}^{l} score_i^t \boldsymbol{h}_i \\
\boldsymbol{o}_t &= [\boldsymbol{h}_t; \boldsymbol{g}_t; \boldsymbol{w}_{MASK}] \\
p_M\left(s_t|e\right) &= \text{softmax}\left(W \cdot \boldsymbol{o}_t\right)
\end{aligned}
\tag{4}
$$

其中$\boldsymbol{h}_{t-1}$和$\boldsymbol{c}_{t-1}$分别是$t-1$时刻解码器的隐藏层向量和细胞状态，$\boldsymbol{c}_t$是t时刻解码器的细胞状态，W是参数矩阵，$\boldsymbol{h}_i$为术语编码器端第i个字符对应的隐藏层向量，l为术语e的字符数量。

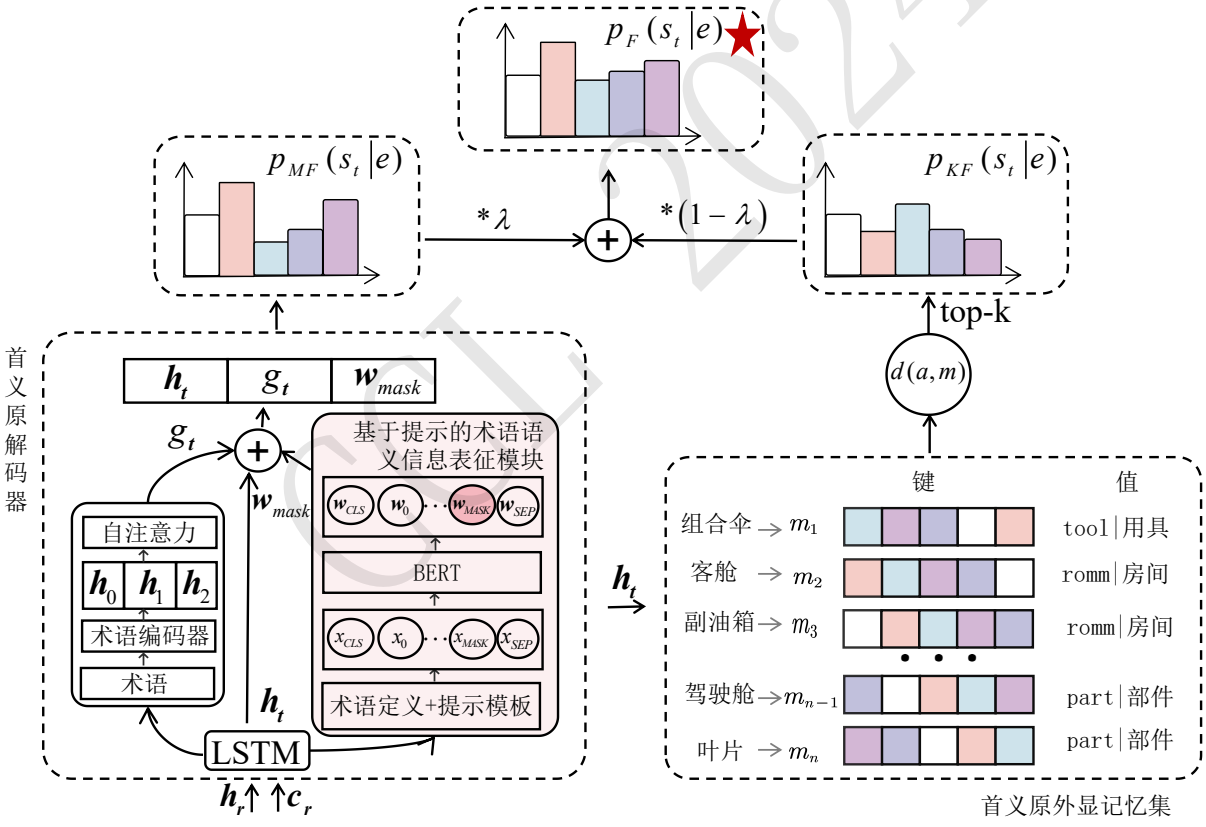

Figure 3: 首义原解码器和首义原KNN联合解码

3.5 KNN模块

本文针对3个解码器分别设计3个外显记忆集，分别为首义原外显记忆集，关系外显记忆集和义原外显记忆集。外显记忆集是由键-值(K,V)对构成，构建方式如式(6)所示。“键”K有三种选择：$t-1$时刻解码器的隐藏层向量$\boldsymbol{h}_{t-1}$，t时刻解码器的隐藏层向量$\boldsymbol{h}_t$和t时刻解码器的拼接向

量$\boldsymbol{o}_t$，如式(5)。首义原外显记忆集和义原外显记忆集的“值”V是义原集合Y_s，关系外显记忆集的“值”V是关系集合Y_r。

$$H = \{\boldsymbol{h}_{t-1}, \boldsymbol{h}_t, \boldsymbol{o}_t\} \tag{5}$$

$$(K, V) = \left\{(\boldsymbol{m}, v) \,\middle|\, \boldsymbol{m} \in H, v \in Y_r \text{或} Y_s\right\} \tag{6}$$

在解码时，$\boldsymbol{a} \in H$为查询向量，$\boldsymbol{m}$和v为t时刻当前解码器的外显记忆集中的键和值。计算$\boldsymbol{a}$与$\boldsymbol{m}$的余弦相似度$d(\boldsymbol{a}, \boldsymbol{m})$，取与$\boldsymbol{a}$余弦相似度最大的$k$个键-值对构成$t$时刻当前解码器的最近邻集合$U_t$，通过式(7)得到$t$时刻KNN概率$p_K(s_t|e)$。由于本文设置三个解码器，所以首义原KNN概率$p_{KF}(s_t|e)$、关系KNN概率$p_{KR}(s_t|e)$、义原KNN概率$p_{KS}(s_t|e)$。

t时刻解码器概率$p_M(s_t|e)$和KNN概率$p_K(s_t|e)$加权求和得到t时刻路径状态概率$p(s_t|e)$，如式(8)所示，其中λ是概率权值。

$$p_K(s_t|e) \propto \sum_{(\boldsymbol{m}, v) \in U_t} \mathbb{I}_{s_t = v} \exp(d(\boldsymbol{a}, \boldsymbol{m})) \tag{7}$$

$$p(s_t|e) = \lambda p_M(s_t|e) + (1-\lambda) p_K(s_t|e), \ \lambda \in (0,1) \tag{8}$$

3.6 基于最近邻的术语DEF生成模块

DEF具体生成过程按照如下步骤:

步骤1：$\boldsymbol{h}_f$和$\boldsymbol{c}_f$是术语编码器输出的最终隐藏状态和最终细胞状态，用于生成术语对应的首义原。

步骤2：解码生成首义原。解码生成首义原过程如图3所示。调用首义原解码器，得到首义原解码器概率$p_{MF}(s_t|e)$并调用首义原KNN得到首义原KNN概率$p_{KF}(s_t|e)$将二者加权求和如式(8)获取首义原概率$p_F(s_t|e)$，最终输出概率最大的首义原。有些术语的首义原有多个，比如“故障检测与定位”，该术语的首义原是“decide|决定”和“check|查”。所以每次解码时会得到一个首义原或者是终结符“<E>”，当生成结果是首义原时保存本次输出的隐藏状态$\boldsymbol{h}_r$和细胞状态$\boldsymbol{c}_r$用于生成首义原所对应的关系。

步骤3：解码生成关系。将步骤2的$\boldsymbol{h}_r$和$\boldsymbol{c}_r$输入关系解码器和关系KNN中，得到关系的概率$p_R(s_t|e)$，概率最大的值即为义原对应的关系。当生成终结符“<E>”或者达到最大数量后停止。如果解码结果是关系，保存本次输出的隐藏状态$\boldsymbol{h}_s$和细胞状态$\boldsymbol{c}_s$用于生成此关系对应的义原。

步骤4：解码生成义原。将步骤3输出的$\boldsymbol{h}_s$和$\boldsymbol{c}_s$输入到义原解码器和义原KNN中，得到义原的概率$p_S(s_t|e)$，输出义原后，保存输出的隐藏状态$\boldsymbol{h}_r$和细胞状态$\boldsymbol{c}_r$用于生成义原对应的关系。

步骤5：重复步骤3与4，直到步骤3生成了终结符“<E>”或者生成结果所对应的树形结构已经达到了预设的最大深度。

4 实验结果与分析

4.1 数据集

本文实验数据集取自航空术语语义知识库 (张桂平et al., 2014)，标注了3864个航空术语，在其中选取能在《中国航空百科词典》 (《中国航空百科词典》编辑部, 2000)查找到完整定义的术语。最终本实验数据集细分为17种航空领域类型，共计3146条术语，包含900种义原，105种关系，术语平均长度4.78字符，术语定义平均长度36.54字符。本文将数据随机分为三部分：训练集（70%）、验证集（10%）和测试集（20%）。本数据集的统计信息如表2所示。

数据类型	术语数	义原数	关系数	三元组数
训练集	2431	11224	8765	8765
验证集	347	1610	1260	1260
测试集	694	3173	2471	2471

Table 2: 航空术语语义知识库数据集统计信息

4.2 对比模型

本文选择LLaMA-7B、ChatGLM-6B、Seq2Tree、SSR、SRS、TSTG和TaSTG模型作为对比模型:

LLaMA-7B (Touvron et al., 2023)，基于Transformer架构的大型语言模型。ChatGLM-6B (Du et al., 2022)，基于General Language Model（GLM）架构的对话语言模型。Seq2Tree (Dong and Lapata, 2016)，一种变种的Seq2Seq方法。在解码时，将Seq2Seq一轮顺序解码变为多批次的顺序解码，每批次解码出树形结构中一个节点的所有子节点。SSR (吕嘉, 2022)，使用添加注意力机制的树形解码器生成父和子义原以及两个义原之间的关系，得到“义原-义原-关系”三元组形式的输出，最终得到术语对应的DEF。SRS (吕嘉, 2022)，交替使用添加注意力机制的义原解码器和关系解码器，得到“义原-关系-义原”三元组形式的输出，最终得到术语对应的DEF。TSTG (Ye et al., 2022)，专为义原树生成任务设计的Transformer模型。其将任务限定为仅生成义原，不考虑生成义原之间的关系。TaSTG (Ye et al., 2022)，为了更好捕捉数据的结构，设计树注意力，分为语义注意力和位置注意力。语义注意力捕获树中两个节点的语义相似性，位置注意力捕获树中节点的拓扑关系。

4.3 评价指标

NNP-TDGM模型在测试集上得到的三元组（义原-关系-义原）、关系、义原和首义原的F1值作为评价指标。F1值越高代表识别效果越好。具体计算公式如式(9)，(10)，(11)所示。

$$F1 = \frac{2 \times P \times R}{P + R} \tag{9}$$

$$P = \frac{\text{生成正确}\,(\text{首})\,\text{义原、关系、三元组数}}{\text{所有生成}\,(\text{首})\,\text{义原、关系、三元组数}} \times 100\% \tag{10}$$

$$R = \frac{\text{生成正确}\,(\text{首})\,\text{义原、关系、三元组数}}{\text{真实样本}\,(\text{首})\,\text{义原、关系、三元组数}} \times 100\% \tag{11}$$

4.4 模型参数设置

NNP-TDGM模型的超参数设置如表3所示。

超参数	取值范围	超参数	取值范围
训练轮次	500	优化器	AdamW
训练批量	500	Dropout	0.5
字向量维度	400	LSTM层数	1
隐藏层维度	400	随机种子	1
学习率	1e-3	注意力头数	1
k	5	λ	(0,1)

Table 3: NNP-TDGM模型参数设置

4.5 术语DEF生成实验

表4展示各模型的实验结果。TSTG和TaSTG仅生成义原，不生成义原之间的关系，所以只有义原$F1$值和首义原$F1$值的实验结果。

模型	三元组F1	关系F1	义原F1	首义原F1
TSTG	-	-	44.26	31.95
TaSTG	-	-	39.90	30.88
ChatGLM-6B	10.26	34.59	29.42	51.36
LLaMA-7B	11.15	36.51	31.67	51.29
Seq2Tree	18.97	42.98	35.41	65.22
SSR	19.76	39.70	48.22	**71.37**
SRS	28.46	51.67	50.47	68.05
NNP-TDGM(ours)	**31.84**	**53.12**	**51.55**	68.53

Table 4: 各模型实验结果（%）

由表4可见，NNP-TDGM好于对比模型，在三元组、关系、义原和首义原的生成上都有所提高，与对比方法中DEF生成效果最好的SRS模型相比，三元组、关系、义原和首义原$F1$值分别提高了3.38%，1.45%，1.08%，0.48%。

4.6 消融实验

4.6.1 NNP-TDGM消融实验

本节进行消融实验，验证基于提示的术语语义信息表征模块（Prompt）与KNN模块（KNN）的有效性。表5展示消融实验结果，其中"-"表示模型不使用该模块，"√"表示模型使用该模块。

模型	KNN	Prompt	三元组F1	关系F1	义原F1	首义原F1
NNP-TDGM	√	√	31.84	53.12	51.55	68.53
NN-TDGM	√	-	30.88	52.96	52.10	69.25
P-TDGM	-	√	30.39	53.31	51.52	69.10
SRS	-	-	28.46	51.67	50.47	68.05

Table 5: 消融实验结果（%）

表5实验结果表明，基于提示的术语语义信息表征模块和KNN模块的加入能够提升模型DEF生成效果。与SRS模型相比，加入KNN模块使三元组、关系、义原和首义原$F1$值分别提高了2.42%，1.29%，1.63%，1.20%。加入基于提示的术语语义信息表征模块使三元组、关系、义原和首义原$F1$值分别提高了1.93%，1.64%，1.05%，1.05%。结果证明将基于提示的术语语义信息表征模块和KNN模块应用于生成术语DEF的有效性。

4.6.2 KNN模块消融实验

为研究KNN模块对术语DEF生成效果的影响。设计了K_F：仅首义原解码器加入KNN,K_R：仅关系解码器加入KNN和K_S：仅义原解码器加入KNN的实验。采用的模型是P-TDGM。实验效果如表6所示。

模型	三元组F1	关系F1	义原F1	首义原F1
P-TDGM	30.39	53.31	51.52	69.10
$+K_F$	30.53	53.31	51.69	**69.80**
$+K_R$	**30.74**	**53.59**	51.72	69.10
$+K_S$	30.66	53.31	**52.00**	69.10

Table 6: 解码器加入KNN的实验结果（%）

在解码器下加入KNN（K_F、K_R和K_S）对各解码器性能产生积极效果。K_F对首义原F1值提升明显，同样，K_S对义原F1值提高明显，K_R对三元组和关系F1值提高明显。因此，解码器

引入KNN模块是一种有效方法，可以提高模型性能，而且不同解码器加入KNN能有针对性地提升模型在不同任务下的性能。

4.7 提示模板影响实验

为验证3.3节12种提示模板对术语DEF生成的影响，设计了P-TDGM使用各类提示模板的实验。模型对12种不同提示模板进行实验得到12种模型。如P-TDGM$_{DLMASK}$模型是使用表1中DLMASK模板的P-TDGM。实验结果如表7所示，同时列出了SRS结果作为参照。

模型	三元组F1	关系F1	义原F1	首义原F1
SRS	28.46	51.67	50.47	68.05
P-TDGM$_{TCLS}$	26.61	50.83	49.54	67.09
P-TDGM$_{TMASK}$	25.47	50.86	47.06	65.22
P-TDGM$_{TPMASK}$	26.80	49.93	48.47	65.80
P-TDGM$_{TLMASK}$	27.01	50.41	48.89	65.94
P-TDGM$_{ACLS}$	29.09	52.06	51.40	69.35
P-TDGM$_{AMASK}$	29.64	52.79	51.12	68.77
P-TDGM$_{APMASK}$	30.10	52.67	51.18	67.48
P-TDGM$_{ALMASK}$	30.75	53.20	51.07	67.76
P-TDGM$_{DCLS}$	28.53	53.66	51.24	70.40
P-TDGM$_{DMASK}$	29.00	51.87	52.02	70.97
P-TDGM$_{DPMASK}$	29.58	52.48	51.03	69.10
P-TDGM$_{DLMASK}$	30.39	53.31	51.52	69.10

Table 7: P-TDGM使用不同提示模板的实验结果（%）

由表7可见，采用不同提示模板对术语DEF生成效果具有明显的影响，这与预训练语言模型提示敏感的性质相符，进一步分析数据可发现:

(1)提示中仅使用术语信息的P-TDGM$_{TCLS}$、P-TDGM$_{TMASK}$、P-TDGM$_{TPMASK}$与P-TDGM$_{TLMASK}$模型的$F1$值与SRS相比均有小幅度下降。本文认为其原因是，一方面术语仅为一个词或短语，有效提示信息较少，很难从预训练语言模型中提取出有效的语义表征向量。另一方面，术语作为模型输入，在模型内形成了表征向量，该表征向量与预训练语言模型提示生成的向量不同构，也可能产生负面影响。

(2)相比于基于术语信息的提示，使用具有更丰富信息的术语定义的提示实验效果更好。P-TDGM$_{DCLS}$、P-TDGM$_{DMASK}$、P-TDGM$_{DPMASK}$与P-TDGM$_{DLMASK}$的各项指标皆好于对比方法SRS。这表明在提示中加入更多描述术语含义的信息是有必要的。此外，还发现同时使用术语及定义提示与仅使用定义提示的效果差别不大。另外，PMASK类及LMASK类模板与CLS及MASK类模板相比在三元组上F1值有所提升，表明增加更为平滑语言环境是有效的。

综上，实验表明提示设计的必要性，对于利用提示生成的术语语义表征向量问题，应考虑如何在提示语句中加入描述术语含义的信息，并且设计平滑语境使得提示语句更符合语言规律。

4.8 检索向量影响实验

在KNN模块中，3个外显记忆集是由键-值对构成，如式(6)所示。每个外显记忆集中的键可使用3种向量：（1）$\boldsymbol{h}_F$、$\boldsymbol{h}_R$与$\boldsymbol{h}_S$分别代表输入到首义原解码器、关系解码器与义原解码器的隐藏层向量，这些向量携带了截至当前时间步之前的所有信息。（2）$\boldsymbol{h}_{FT}$、$\boldsymbol{h}_{RT}$与$\boldsymbol{h}_{ST}$分别代表首义原解码器、关系解码器与义原解码器的隐藏状态，这些向量携带了解码器输出的首义原、关系或义原的信息。（3）$\boldsymbol{o}_{FT}$、$\boldsymbol{o}_{RT}$与$\boldsymbol{o}_{ST}$分别代表首义原解码器、关系解码器与义原解码器的拼接向量。将这3类向量分别作为外显记忆集中的键，研究其对术语DEF生成效果的影响。实验结果如表8所示，同时将SRS模型作为参照，列出了SRS模型的实验结果。表中FD表示首义原解码器、RD表示关系解码器、SD表示义原解码器。

解码器	模型	三元组F1	关系F1	义原F1	首义原F1
	SRS	28.46	51.67	50.47	68.05
FD	$+\boldsymbol{h}_F$	28.51	51.67	50.48	68.10
	$+\boldsymbol{h}_{FT}$	28.51	51.67	50.41	67.95
	$+\boldsymbol{o}_{FT}$	28.51	51.67	50.68	69.10
RD	$+\boldsymbol{h}_R$	29.57	52.57	51.16	68.05
	$+\boldsymbol{h}_{RT}$	29.33	52.43	50.73	68.05
	$+\boldsymbol{o}_{RT}$	29.49	52.81	51.05	68.05
SD	$+\boldsymbol{h}_S$	29.04	51.65	50.86	68.05
	$+\boldsymbol{h}_{ST}$	29.13	51.65	51.00	68.05
	$+\boldsymbol{o}_{ST}$	29.17	51.65	50.86	68.05

Table 8: λ=0.8时不同解码器选择不同向量的实验结果（%）

实验结果表明，对于各解码器采用不同向量均能提高模型性能。综合多方面拼接向量$\boldsymbol{o}$的效果更好。在首义原解码器中使用$\boldsymbol{o}_{FT}$效果最佳，首义原F1值提升1.05%。在关系解码器中使用$\boldsymbol{o}_{RT}$对关系$F1$值提升最高，提升了1.14%。在义原解码器中使用$\boldsymbol{h}_{ST}$使义原$F1$值提升0.53%。

4.9 λ值影响实验

本节研究KNN中概率权值λ对NNP-TDGM模型效果的影响。λ在区间$(0,1)$内取值，NNP-TDGM模型各$F1$值如图4所示。从图4中可以观察到，$F1$值随着λ的增加呈现上升趋势，在达到峰值后有所回落。表明λ对$F1$值是产生影响的。λ小于0.5时，KNN概率占比较大，实验结果低于基准值，表明KNN仅能提供辅助作用，不能作为主导。当$\lambda = 0.6$时三元组$F1$值最高，并且随着λ的增加，三元组$F1$值变化趋于平稳。

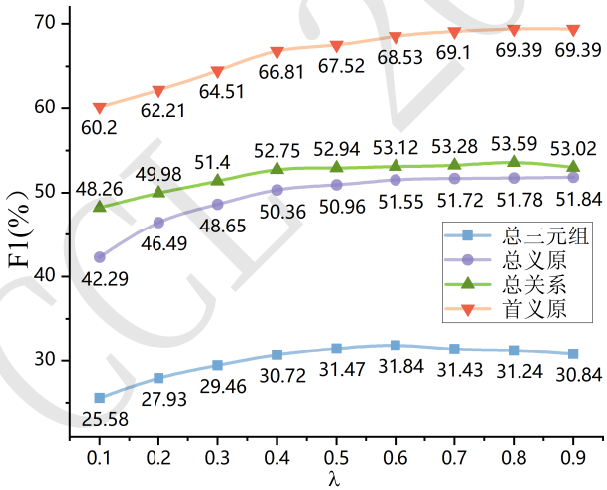

Figure 4: $k = 5$时NNP-TDGM模型各$F1$值随λ变化曲线

4.10 实例分析

本节对KNN模块的作用进行了分析，附录中展示了术语、正确的DEF、SRS和NNP-TDGM生成的DEF，阐述了首义原KNN、关系KNN和义原KNN的检索情况。

从检索实例来看，首义原KNN能检索出与待预测术语概念结构相同或相近的术语中蕴含的核心概念。如实例1中，预测术语“壁温试验”的首义原时，首义原KNN检索到概念结构相近的术语“高温试验”和“耐久性试验”，并且得到这些术语对应的核心概念“fact|事情”。义原KNN能检索出与待预测术语概念结构相同或相近的术语的重要属性。如实例2中，预测术语“自动驾驶仪”的义原时，义原KNN能检索出术语“积分式自动驾驶仪”和“驾驶仪”的重要属性“drive|驾驶”。关系KNN能检索出与待预测术语概念结构相同或相近的术语的关系类型。如实例3中，

预测术语“超临界机翼”的关系时，关系KNN检索出术语“有限翼展机翼”和“层流机翼”的关系类型“whole”。

结果表明，KNN通过检索与待预测术语概念结构相同或相近的术语中蕴含的核心概念、重要属性和关系类型，辅助模型完成DEF的生成，从而弥补低频样本下模型训练不充分的问题。

5 结论与展望

本文提出一种基于最近邻提示表征的术语DEF生成模型（NNP-TDGM）。提出基于提示的术语语义信息表征方法，基于术语、定义、术语和定义3类信息构建4类共12种模板，通过提示预训练语言模型获得术语及术语定义内蕴含概念信息的语义表征向量，将其整合到解码器中。提出基于最近邻的DEF生成方法，将训练集中的术语DEF构造成3种外显记忆集。在解码器生成（首）义原或关系时，通过检索与待预测术语概念结构相同或相近的术语中蕴含的核心概念、重要属性和关系类型，辅助模型完成DEF生成。通过实验验证了基于提示的术语语义信息表征方法及基于最近邻的DEF生成方法能够改善术语DEF生成效果。对于利用提示生成的术语语义表征向量问题，应考虑如何在提示语句中加入描述术语含义的信息，同时应设计平滑语境使得提示语句更符合语言规律。对于首义原、关系和义原解码器分别加入KNN方法皆有效果。

在未来工作中，将考虑如何动态选取KNN的参数，减少噪声的引入，充分发挥KNN检索性能以及如何获得并加入更多术语信息，例如含该术语的文本片段，相关其他术语等。

参考文献

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, pages 33–43.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2021. Augmenting transformers with knn-based composite memory for dialog. *Transactions of the Association for Computational Linguistics*, 9:82–99.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910.

Yihong Gu, Jun Yan, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, and Leyu Lin. 2018. Language modeling with sparse product of sememe experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4642–4651.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 3929–3938.

Bairu Hou, Fanchao Qi, Yuan Zang, Xurui Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Try to substitute: An unsupervised chinese word sense disambiguation method based on hownet. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1752–1757.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. Promptbert: Improving BERT sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8826–8837.

Huiming Jin, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, and Leyu Lin. 2018. Incorporating chinese characters of words for lexical sememe prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2439–2449.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Nearest neighbor machine translation. In *Proceedings of the 9th International Conference on Learning Representations, ICLR 2021,Virtual Event, Austria, May 3-7, 2021*.

Shudong Liu, Xuebo Liu, Derek F. Wong, Zhaocong Li, Wenxiang Jiao, Lidia S. Chao, and Min Zhang. 2023. knn-tl: k-nearest-neighbor transfer learning for low-resource neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1878–1891.

Yilin Niu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. Improved word representation learning with sememes. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 2049–2058.

Fanchao Qi, Liang Chang, Maosong Sun, Sicong Ouyang, and Zhiyuan Liu. 2020. Towards building a multilingual sememe knowledge base: Predicting sememes for babelnet synsets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8624–8631.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020. Improving knowledge-aware dialogue generation via knowledge base question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9169–9176.

Ruobing Xie, Xingchi Yuan, Zhiyuan Liu, and Maosong Sun. 2017. Lexical sememe prediction via word embeddings and matrix factorization. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4200–4206.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5065–5075.

Yining Ye, Fanchao Qi, Zhiyuan Liu, and Maosong Sun. 2022. Going "deeper": Structured sememe prediction via transformer with tree attention. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 128–138.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451.

Lei Zheng, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2020. Multi-channel reverse dictionary model. In *Proceedings of the AAAI conference on artificial intelligence*, pages 312–319.

《中国航空百科词典》编辑部. 2000. 中国航空百科词典. 航空工业出版社.

吕嘉. 2022. 面向复合型术语的知网DEF自动生成方法研究. 硕士学位论文, 沈阳航空航天大学.

张桂平, 刁丽娜, and 王裴岩. 2014. 基于HowNet的航空术语语义知识库的构建. 中文信息学报, 28:92–101.

曾骏, 王子威, 于扬, and 文俊浩. 2024. 自然语言处理领域中的词嵌入方法综述. 计算机科学与探索, 18:24–43.

杜家驹, 岂凡超, 孙茂松, and 刘知远. 2020. 基于局部语义相关性的定义文本义原预测. 中文信息学报, 34:1–9.

董振东, 董强, and 郝长伶. 2007. 知网的理论发现. 中文信息学报, 4:3–9.

附录.实例分析

术语	壁温试验				
正确 DEF	{fact|事情:CoEvent={experiment|实验:content={Temperature|温度:host={part|部件:PartPosition={skin|皮},whole={building|建筑物}}},means={try|尝试}}}				
SRS	{**experiment|实验**:CoEvent={experiment|实验:content={Temperature|温度:host={physical|物质}}}}				
NNP-TDGM	{**fact|事情**:CoEvent={experiment|实验:content={Temperature|温度},means={try|尝试:host={physical|物质}}}}				
首义原	p_{SRS}	p_{FKNN}	检索出的核心概念	检索出的核心概念对应的术语和 DEF	
				术语	DEF
fact|事情	0.45	1.0	fact|事情	高温试验	{**fact|事情**:CoEvent={experiment|实验:PatientAttribute={Temperature|温度:host={physical|物质},modifier={GreaterThanNormal|高于正常}}}}
			fact|事情	耐久性试验	{**fact|事情**:CoEvent={experiment|实验:content={Quality|质量:host={physical|物质}},means={try|尝试}}}

Figure 5: 实例分析1

预测术语“壁温试验”的首义原时，首义原KNN检索到与“壁温试验”概念结构相近的术语“高温试验”和“耐久性试验”，并且得到这些术语对应的核心概念“fact|事情”。

术语	自动驾驶仪				
正确 DEF	{tool|用具:#instrument={drive|驾驭:modifier={automatic|自动}}}				
SRS	{tool|用具:#instrument={**automatic|自动**},modifier={automatic|自动:modifier={automatic|自动}}}				
NNP-TDGM	{tool|用具:#instrument={**drive|驾驭**:modifier={automatic|自动}}}				
义原	p_{SRS}	p_{SKNN}	检索出的重要属性	检索出的重要属性对应的术语和 DEF	
				术语	DEF
drive|驾驭	0.59	0.66	drive|驾驭	积分式自动驾驶仪	{tool|用具:#instrument={**drive|驾驭**:manner={automatic|自动}},RelateTo={Pattern|样式:host={knowledge|知识:domain={math|数学}}}}
			drive|驾驭	驾驶仪	{tool|用具:#instrument={**drive|驾驭**}}

Figure 6: 实例分析2

预测术语“自动驾驶仪”的义原时，义原KNN能检索到与“自动驾驶仪”概念结构相近的术语“积分式自动驾驶仪”和“驾驶仪”，并且得到这些术语对应的重要属性“drive|驾驶”。

<table>
<tr><td>术语</td><td colspan="5">超临界机翼</td></tr>
<tr><td>正确 DEF</td><td colspan="5">{part|部件:PartPosition={wing|翅},whole={aircraft|飞行器},modifier={special|特别}}</td></tr>
<tr><td>SRS</td><td colspan="5">{part|部件:PartPosition={wing|翅},modifier={aircraft|飞行器},whole={aircraft|飞行器}}</td></tr>
<tr><td>NNP-TDGM</td><td colspan="5">{part|部件:PartPosition={wing|翅},whole={aircraft|飞行器},modifier={special|特别}}</td></tr>
<tr><td rowspan="2">关系</td><td rowspan="2">p_{SRS}</td><td rowspan="2">p_{RKNN}</td><td rowspan="2">检索出的关系类型</td><td colspan="2">检索出的关系类型对应的术语和 DEF</td></tr>
<tr><td>术语</td><td>DEF</td></tr>
<tr><td rowspan="2">whole</td><td rowspan="2">0.45</td><td rowspan="2">1.0</td><td>whole</td><td>有限翼展机翼</td><td>{part|部件:PartPosition={wing|翅},whole={aircraft|飞行器},modifier={few|少},RelateTo={Length|长度:host={part|部件:PartPosition={wing|翅},whole={aircraft|飞行器}}}}</td></tr>
<tr><td>whole</td><td>层流机翼</td><td>{part|部件:PartPosition={wing|翅},whole={aircraft|飞行器},modifier={polished|光:degree={most|最}}}</td></tr>
</table>

Figure 7: 实例分析3

预测术语“超临界机翼”的关系时，关系KNN能检索到与“超临界机翼”概念结构相近的术语“有限翼展机翼”和“层流机翼”，并且得到这些术语对应的关系类型“whole”。