doi:10.3772/j.issn.2095-915x.2021.05.004

开放科学（资源服务）标识码（OSID）

融合深度语义表示的开放域复述模板抽取

金波 刘明童 张玉洁 徐金安 陈钰枫

北京交通大学计算机与信息技术学院 北京 100044

摘要：[目的 / 意义] 复述知识在自然语言处理领域应用广泛，例如复述生成、自动问答、复述数据构建等，复述模板是复述知识的一种抽象表现形式。[方法 / 过程] 早期复述模板抽取方法依赖复述实例和可比语料，近年来，有研究在人工给定种子实体对的基础上，利用实体关系，通过自举迭代的方式获取复述模板，规避对特定语料依赖，但是只能获得语义受限的复述模板，且迭代过程中存在语义偏移现象。因此本文提出从英文维基百科的大规模开放域文本中自动获取复述模板的思路，针对问题，设计融合深度语义表示学习的复述模板抽取方法，从大规模开放域文本中抽取高精度多样化的复述模板。[结果 / 结论] 实验结果表明，本文所提方法可以获取高质量、语义一致、形式多样的复述模板。

关键词：复述模板；语义表示；自动聚类

中图分类号：G35; TP391

Mining Paraphrase Template with Deep Semantic Representation from Open Domain

JIN Bo LIU Mingtong ZHANG Yujie XU Jinan CHEN Yufeng

Beijing Jiaotong University, Beijing 100044, China

Abstract: [Objective/Significance] Paraphrase knowledge plays a vital role in many tasks in the field of natural language processing, such as paraphrase generation, automatic question answering, construction of paraphrase data, etc. Paraphrase

基金项目 国家自然科学基金 (61876198, 61976015, 61370130, 61473294)，北京市自然科学基金 (4172047) 和科学技术部国际科技合作计划 (K11F100010)。

作者简介 金波（1997-），硕士，研究方向为自然语言处理、复述知识获取；刘明童（1993-），博士，研究方向为自然语言处理、神经机器翻译、复述；张玉洁（1961-），教授，研究方向为自然语言处理、机器翻译、文本大数据处理，E-mail：yjzhang@bjtu.edu.cn；徐金安（1970-），教授，研究方向为自然语言处理、机器翻译；陈钰枫（1981-），教授，研究方向为自然语言处理、机器翻译。

引用格式 金波 , 刘明童 , 张玉洁 , 等 . 融合深度语义表示的开放域复述模板抽取 [J]. 情报工程 , 2021, 7(5): 40-50.

template extraction is an important way to obtain paraphrase knowledge. [Methods/Process] This paper proposes a sentence-level paraphrase template extraction method based on semantic representation and automatic clustering algorithm to extract high-precision and diverse paraphrase templates from open domain. [Results /Conclusions] Our experiments on the Wikipedia English corpus show that the proposed method can effectively obtain semantically similar and diverse forms of paraphrase templates.
Keywords: Paraphrase template; semantic representation; automatic clustering

引言

复述 (paraphrases)，是指用不同表现形式表述相同语义的语言现象，广泛存在于自然语言中。美国认知心理学家 Olson G M[1] 认为，复述是判断计算机理解自然语言的重要手段之一。复述模板在自然语言处理诸多领域有着广泛应用，例如机器翻译中利用复述生成的数据增强技术 [2]、问答系统中利用复述识别寻找相同语义的问句 [3]、以及文本摘要 [4] 等。复述研究可以分为复述识别与生成两个部分，二者都需要复述知识作为基础，由此可见，复述模板在复述识别与生成任务中扮演着重要角色，复述模板的获取一直是复述研究的核心内容。

复述模板的获取一般集中在短语、句子、模板及篇章四个级别。例如，“Wu Chengen wrote a Journey to the West” 和“The author of Journey to the West is Wu Chengen”是一组复述句对。从复述句对，可以进一步抽取出复述模板。复述模板指一组语义上等价的模板，每个模板由词语和变量槽组成， 对上述复述句对泛化可以得到复述模板“[PERSON] wrote [BOOK]”和“The author of [BOOK] is [PERSON]”。

传统的复述模板抽取从复述实例出发，经过泛化得到复述模板，但由于复述实例资源较为匮乏，此类方法很难获取多样化的复述模板。近年来，研究人员利用句中命名实体之间的关系，从大规模单语语料获取复述模板的方法受到关注 [13-15]，这些方法既能规避对复述实例的依赖，又能关注单语语料中隐藏的复述模板，例如，早期的方法需要“Wu Chengen wrote a Journey to the West”和“The author of Journey to the West is Wu Chengen”这样的复述实例才能够获取复述模板，而这些方法能够从“Wu Chengen wrote a Journey to the West” 和“The author of Dream of the Red Chamber is Cao Xueqin”这样的例子中获取同样的复述模板。其主要思想是利用具有特定关系的实体对，从大规模语料中抽取实体进行泛化获取模板，进而利用模板抽取实例拓展实体对，如此迭代获取具有特定关系的复述模板。但是，人工定义特定语义关系的实体对，会导致抽取的复述模板语义受限；同时，自举迭代过程存在语义偏移现象，会引入噪音；最后，获取模板语义表示时，仅利用表层字符特征，取算术平均，导致聚类精度降低，影响复述模板质量。随着深度学习的发展，表示学习方法得到长足进步，模板语义表示方法有很大的发展空间。

本文提出开放域上基于深度语义表示的复述模板自动获取方法，探索高质量地自动获取多样化复述模板的途径。针对需要人工定义种

doi:10.3772/j.issn.2095-915x.2021.05.004

子实体对以及自举迭代中语义偏移的问题，设计利用开放域语言资源，直接获取大规模潜在复述文本的方法；针对模板语义表示不精确的问题，设计融合深度语义表示的模板语义计算方法，从而提升复述模板质量。

我们的方法优点在于，一方面，本文方法不需要人工预先定义命名实体种子，可以直接从大规模单语语料中获取语义覆盖面广泛的复述模板；另一方面，本文利用最新的表示学习方法，设计基于深度学习的模板语义表示方法，提高复述模板聚类的精度。

本文剩余部分的组织结构如下：第一节介绍相关研究；第二节描述从开放域单语语料获取潜在复述文本，并泛化为模板的方法；第三节介绍融合深度语义表示的复述模板语义计算及自动聚类方法；第四节介绍实验评测和分析结果；第五节对本文研究进行总结。

1 相关研究

复述模板抽取任务起源于信息抽取领域，早期，研究者们基于复述语料库，通过泛化复述实例的方式，得到复述模板 [5-10]。其中，Barzilay 等 [5] 基于词性序列表示，利用译本相关复述实例库，泛化获取复述模板；Ali 等 [6] 以命名实体为锚点，通过计算复述实例中对应锚点间句法路径的相似性，抽取复述模板；Zhao 等 [7] 利用双语平行语料，将其中一种语言看作枢轴（pivot），利用对数线性模型计算复述概率抽取模板。

先前研究依赖复述实例，但是，获取大量复述实例较困难，因此研究者们开始在单语语料上进行复述模板抽取研究 [11-16]。其中，Shinyama 等 [11] 以日语作为研究对象，利用同一天针对同一事件的不同新闻报道，将相同实体看成锚点，通过泛化实体间依存路径得到复述模板；Sekine[12] 同样利用实体类型作为分类依据，通过分析实体对之间的短语，根据短语中的关键词，聚类得到复述模板；Priyanka Das等 [13] 专注于报道罪案的新闻，结合语法树，利用结构信息从不同犯罪类型的新闻中抽取复述模板；Biran[14] 首先利用知识库对句子泛化，得到特定语义类型模板，再通过聚类的方式获取复述模板；Liu 等 [15] 利用自举迭代的方式获取关系实例，通过深度语义计算的方式表示模板，从大规模单语语料中自动聚类得到复述模板。

相比于复述实例、或是双语平行语料，单语语料研究的最大挑战在于缺少语义等价线索。先前研究大多从文本着手，利用高精度的句法分析 [13]、大规模的知识库 [14] 或是互联网搜索引擎 [15] 获取语义相近的文本，从而泛化获取复述模板。本文研究开放域上英文复述模板的抽取方法，通过改进模板语义表示方法，探索结合深度语义计算的复述模板获取方法，提升模板表示精度，复述模板质量。

2 基于实体识别的模板获取方法

在单语语料中挖掘复述知识，最大的挑战在于，单语语料中缺乏语义一致线索。考虑到同一领域内的文章描述内容相关性高，存在复述文本的可能性更大，首先，我们从维基百科获取同一领域的文本；由于句子中实体蕴含丰

富语义信息，依照通常做法[14]，我们以实体对作为实体之间语义关系的分布特征，对句子中的命名实体进行泛化，作为复述模板变量槽，并将对应的命名实体类别作为变量槽类型，获得模板。模板获取过程如图 1 所示，由以下两个部分组成。

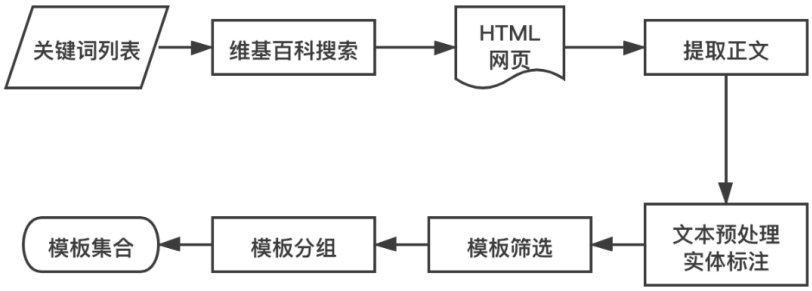


图 1 从开放域获取同领域文本并泛化获得模板的处理过程

2.1 同领域文本获取

本文选择维基百科英文文本作为开放域单语语料来源，从维基百科获取相关文章，获取大量描述同一领域的实例。我们通过给定关键词，收集同一领域内文本。利用维基百科搜索引擎，爬取领域相关关键词对应的维基百科正文。例如，在获取 NBA 篮球联赛领域文本时，我们根据 NBA 相关关键词列表，利用搜索引擎获取对应的英文维基百科网页，从网页中提取文本，得到 NBA 领域文本。

2.2 模板获取

我们对 2.1 节获取的文本进行处理，泛化得到模板，由于网页内容存在较多噪音，我们按照以下策略对文本进行预处理：

1）对文本进行数据清洗，去除文本中的 HTML 标签、特殊标记，并进行语句切分。

2）考虑到句子过短，泛化得到的模板过短，缺乏实用价值；而句子过长，结构复杂，会影响模板的泛化程度。因此，保留句长大于 10、小于 40 的句子。

3）进行实体标注，我们使用斯坦福开源的 Stanza 工具[17]作为实体标注工具。

4）实体标注完成后，考虑到实体过多，模板结构复杂，会降低模板抽象能力，所以我们根据模板中的实体个数对模板进行筛选，保留变量槽个数在 2~5 之间的模板。

在完成上述预处理步骤后，我们将文本中的命名实体泛化为变量槽，并将变量槽类别标注为对应的实体类别，得到模板集合。随后，我们根据变量槽个数和类型对模板进行分组，将变量槽个数相同，而且类型一致的模板归为一组，作为潜在复述模板组合，将分组后的模板集合交给自动聚类处理。

3 基于深度语义表示的复述模板获取

根据第二节介绍的方法，我们从开放域获取领域对应的模板实例，并且按照变量槽类型

doi:10.3772/j.issn.2095-915x.2021.05.004

对模板实例进行分组。研究结果表明，分组后，同一组的模板之间仍存在一些语义差异，不能直接作为复述模板。针对这一问题，本文提出基于深度语义表示的复述模板自动聚类方法，希望将表达相同语义的模板分到同一类。

3.1 模板语义计算方法

模板由变量槽和词语两部分组成，变量槽包含类型信息，词语部分包含更丰富具体的语义信息，因此我们利用词语部分计算模板语义表示。先前研究在表示模板时，仅仅基于词向量，取算术平均，获得模板表示，但是研究表明，模板中词的重要程度各不相同，不同词在模板表示中应当占据不同的权重。针对这一问题，我们基于 Sanjeev Arora 等 [18] 提出的无监督句向量算法，设计模板语义计算方法，根据词频，给模板中每个词赋予权重，得到模板语义表示向量。

模板语义计算方法主要由两个步骤组成，第一步是基于词向量，进行加权求和取平均，得到初步的模板向量；第二步是利用主成分分析，移除模板向量中的无关部分，具体计算算法如图 2 所示。

算法 1 模板语义表示

输入： 单词语义表示 $\{v_w, w \in V\}$，模板集合 T，超参数 a，单词出现频率 $\{p(w) : w \in V\}$

输出： 模板表示 $\{v_{t_i} : t_i \in T\}$

1: **for all** template t_i in T **do**

2: $\quad v_{t_i} = \frac{1}{|t_i|} \sum_{w \in t_i} \frac{a}{a+p(w)} v_w$

3: **end for**

4: 将每个模板表示作为一列，构成矩阵，计算该矩阵的第一个主成分 u

5: **for all** template t_i in T **do**

6: $\quad v_{t_i} = v_{t_i} - uu^T v_{t_i}$

7: **end for**

图 2 模板语义表示算法

其中，我们利用 Word2Vec[19] 获取词向量，采用 skip-gram 模型，上下文窗口设置为 5，在维基百科英文语料①上训练得到 300 维词向量，作为每个单词的语义表示 v_w。

3.2 基于自动聚类的复述模板获取

语义越接近的模板，其在语义空间上的距离也越接近，因此，本文基于模板表示，在同一语义空间计算模板间的欧几里得距离，利用聚类算法，在每一组模板集合中寻找互为复述的模板。我们以单链接自底向上层次聚类算法为基础获取复述模板，先前研究在聚类时，根据前期实验结果设置聚类簇个数 [15]，这种设置方式需要针对特定数据，对聚类簇数量进行提前评估预测，针对这一问题，我们利用聚类簇内最大平方误差和作为判断聚类是否终止的依据，具体算法如图 3 所示。

① https://dumps.wikimedia.org/

算法 2 基于层次聚类的模板聚类算法

输入： 模板表示集合 $\{v_{t_i} : t_i \in T, i \in [0, n]\}$，聚类簇距离度量函数 d，聚类阈值 τ

输出： 模板划分 $\mathcal{C} = \{C_1, C_2, ..., C_k\}$

for all template t_i in T **do**
  $C_i = \{v_{t_i}\}$
end for
for $i = 1, 2, ..., n$ **do**
  for $j = 1, 2, ..., n$ **do**
    $M(i, j) = d(C_i, C_j)$；$M(j, i) = M(i, j)$
  end for
end for
当前聚类簇个数：$q = m$
while True **do**
  寻找距离最近的两个聚类簇 C_{i^*} 和 C_{j^*}，合并 C_{i^*} 和 C_{j^*}：$C_{i^*} = C_{i^*} \cup C_{j^*}$
  计算合并后聚类簇 C_{i^*} 内部的平方误差和 $SSE = log(\sum_{v_{t_i}}^{C} d(v_{t_i}, \mu_C)^2)$
  其中，μ_C 是指当前聚类簇的中心
  if $SSE \geq \tau$ **then**
    终止聚类
  else
    for $j = j^* + 1, j^* + 2, ...,$ **do**
      将聚类簇 C_j 重新编号为 C_{j-1}
    end for
    更新距离矩阵 M，删除 M 的第 j^* 行和第 j^* 列
    for $j = 1, 2, ..., q - 1$ **do**
      $M(i^*, j) = d(C_{i^*}, C_j)$；$M(j, i^*) = M(i^*, j)$
    end for
    $q = q - 1$
  end if
end while

图 3 复述模板自动聚类算法

我们可以根据预先设置的聚类阈值，获取聚类结果。最终，具有相同标签的模板看作复述模板集合。舍弃只有一个模板的类；同时，在大于十个模板的类中，存在噪音的可能性较高，我们选择放弃这样的类。

4 实验评价与结果分析

为了验证本文所提方法的有效性，本文一共进行两项实验，分别对模板语义计算方法，以及复述模板质量进行评测分析。

4.1 模板语义计算方法评价

模板语义计算方法能够获取模板表示，为复述模板自动聚类获取提供支撑。模板表示的精确性，对复述模板质量有较大影响。我们利用 Takase 等 [20] 提出的 relPatSim 数据集，对模板语义计算方法进行评测分析。

4.1.1 实验数据

relPatSim 数据集由 5555 对模板构成，每

doi:10.3772/j.issn.2095-915x.2021.05.004

一对模板都有五个人工标注的相似度得分，从 1 到 7，得分越高表示模板间语义越相近。数据示例如表 1 所示。

表 1 relPatSim 数据示例

Pattern1	Pattern2	Average similarity score
Turn into	Lead to	5.8
Be open from	Close at	1.6

表 1 中的两个实例分别代表存在复述关系的模板以及不存在复述关系的模板，Turn into 与 Lead to 存在复述关系。因此这对关系模板平均标注得分较高，为 5.8；而 Be open from 与 Close at 不存在复述关系，因而得分较低，为 1.6。

4.1.2 评价指标

本文选择斯皮尔曼相关系数作为评价指标，比较人工标注语义相似度本文提出的模板语义计算方法算出的语义相近度之间的相关性。相关性越高，说明模板表示与人工标注结果越接近，说明利用本文提出的模板计算方法能够准确获得模版的语义表示，从而准确衡量模版间的语义相近程度。

斯皮尔曼相关系数计算如公式（3）所示：

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (3)$$

其中，x_i 和 y_i 是由原始数据 X_i 和 Y_i 转换而来的等级得分。

4.1.3 评测结果与分析

我们主要比较算术平均（Average）、主流的深度语义计算方法（RNN、LSTM、GRU）和本文提出的语义表示计算方法在 relPatSim 数据集上的表现，实验结果如表 2 所示。

表 2 relPatSim 评测结果

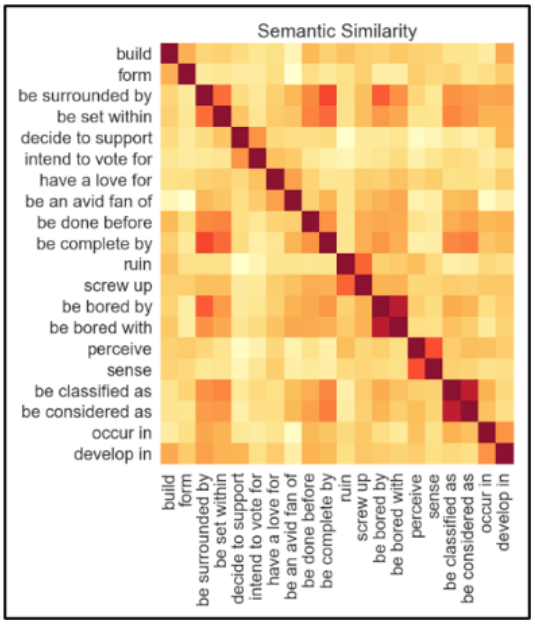
模板长度	模板数量	Average	LSTM	GRU	RNN	Ours
1	636	0.324	0.324	0.324	0.324	0.324
2	1018	0.319	0.257	0.274	0.274	**0.321**
3	2272	0.386	0.344	0.370	0.387	**0.411**
4	1206	0.306	0.314	0.329	0.319	**0.387**
> 5	423	0.315	0.369	0.384	0.394	**0.396**
All	5555	0.340	0.336	0.356	0.362	**0.378**

从表 2 可以看出，在较短的模板上，相比于深度神经网络，本文提出的方法有较大提升。以主流的 LSTM 为例，在长度为 2 的模板上，提升了 0.64，在长度为 3 的模板上，提升了 0.67，这说明句长较短时，本文提出的方法能够更好得表示模板；即使在神经网络擅长的长序列上，我们的方法也能取得与 RNN 相当的效果。实验结果表明，本文提出的模板语义计算方法相较于传统的算术平均以及其他的深度语义计算方法更加有效。

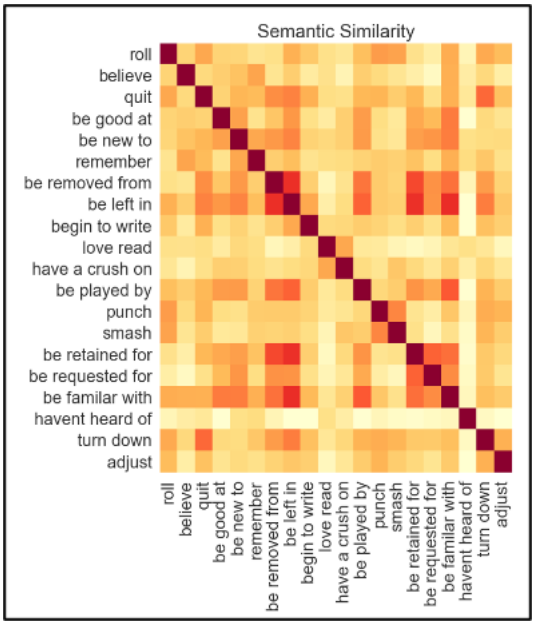
我们从 relPatSim 数据集中分别随机选取了十对平均得分 5 以上、十对平均得分 4 以下的关系模板，对语义相似度进行可视化分析，分析结果如图 4 所示。

图 4 中，颜色越深的方块表示对应的模板对相似度越高，我们主要关注模板对之间的结果，即主对角线上大小为 2*2 的方块的颜色。图 4（a）主要统计了人工标注得分较高的模板对，例如“be classified as”和“be considered as”这一对模板，人工得分为 6.5 分，在我们的可视化结果中，对应方块颜色较深；而在图 4（b）主要统计标注得分较低的模板对，例如“turn down”与“adjust”之间不存在复述关系，对应方块颜色也较浅。可视化结果从一个直观

的角度表明本文提出的模板语义计算方法是有效的。



（a）得分高于5的模板



（b）得分低于4的模板

图 4 语义相似度可视化

4.2 复述模板分析与评价

本节通过人工评测方法，对复述模板质量进行评测与分析。

4.2.1 实验数据

我们选择维基百科作为评测数据来源，利用第 4 节描述的方法获取复述模板，表 3 显示了实验数据信息，我们选择三个不相关领域的维基百科英文语料进行实验评测分析。

4.2.2 评价指标

先前的研究缺乏对人工评测指标的准确定义，我们定义了清晰明确的人工评测指标，通过人工评测来衡量复述模板的质量，具体评价标准如表 4 所示。

表 3 维基百科语料信息

领域	描述	规模（句子数）	源链接（https://en.wikipedia.org/wiki/）
NBA	美国职业篮球联赛相关队伍信息	6556	National_Basketball_Association
MARVEL	漫威电影相关信息	3918	Category:Marvel_Comics_superheroes
Chinese Province	中国各个省份信息	4621	Provinces_of_China

表 4 复述模板评价指标

得分	描述
1	模板之间语义完全不相关，描述不同话题
2	模板描述同一话题，但是模板间不存在复述关系，语义不一
3	模板语义相似，仅存在一些细微差别；或者复杂模板中某些部分存在复述关系
4	模板语义几乎等价，只允许存在一个细微差别（单词、时态不同）
5	模板语义一致，构成复述

doi:10.3772/j.issn.2095-915x.2021.05.004

4.2.3 复述模板评测

我们从每一个领域聚类结果中，随机选取 100 个类，对于每一类模板，我们随机选择其中一个模板，使用其原始实体，将同类中所有模板，还原成复述句，对填充后的句子进行人工评测，评测结果如表 5 所示。

表 5 不同领域模板得分

领域	抽取复述模板组数	抽取复述模板总数	平均得分	%3+	%4+
NBA	1624	4595	3.46	83	58
MARVEL	707	1983	3.3	78	52
Chinese Province	990	2707	2.96	56	37

我们可以从表 5 的结果中看到，不同领域的模板抽取难度各不相同，对于 NBA 和 MARVEL 领域，模板得分较高，均分在三分以上，这意味着提取出来的复述模板大多语义相似，有着或部分有着复述关系。中国省份领域的文本中存在大量句式复杂的模板，且较多实体存在歧义，虽然模板平均得分低于三分，但是仍有超过一半的模板得分在三分以上，这证明我们的方法在处理长句较多、歧义实体较多的情况时，有一定效果。表 5 结果表明本文提出的方法可以有效从不同领域开放域文本抽取复述模板，并且模板质量较高。

4.2.4 实例分析

为了进一步验证本文提出复述模板自动获取方法的有效性，我们进行实例分析，表 6 给出了不同领域的一些例子，以及对应得分，我们选择了得分 1~5 的十组实例进行分析。

首先我们对错误的例子进行分析，从低分的一些例子中，我们能发现，变量槽类型粒度太粗，会导致分组时，引入一些噪音文本，例如得分为 1 的 Chinese Province 示例，这对模板将“大学”和“机构”都标注成了“ORG”类型，没有更细粒度的标注，导致语义不相关的模板出现了同一组中，影响了后续聚类。另一种常见的情况是实体类型标注错误，例如得分为 3 的 MARVEL 示例，实体识别工具错误的将完整人名“Maria Hill”识别成“[PERSON] Hill”，在实体识别错误时，如果模板词语部分的语义信息足够丰富，相关模板还是能够被聚到同一类，但是由于变量槽类型标注的错误，得到的模板丧失了下游任务中的应用价值。

从表 6 中正确实例中可以看出，在实体类型标注正确的前提下，即使不能得到完美的复述模板（得分为 5），至少也能够保证话题的相关性。

5 总结

本文提出一种从网络开放域文本自动抽取复述模板的方法，针对早期研究依赖复述实例和可比语料的问题，设计实现了基于深度语义表示学习和自动聚类算法的复述模板抽取方法，从开放域文本中抽取高精度多样化的复述模板。同时，我们改进了先前方法中的模板语义计算方法，以希望得到更高质量的复述模板。实验结果表明，本文提出的方法可以获取到质量较高的复述模板。未来，我们将尝试改进模板泛化方法，寻找更精确的变量槽选取以及标注方法；同时，在模板语义计算过程中融合变量槽的信息以及句法信息，以进一步提高复述模板的质量。

表 6 复述模板实例

得分	领域	模板
5	NBA	[PERSON] was hired as head coach in [DATE].
		[DATE], they hired [PERSON] as head coach.
	MARVEL	[ORG] released the album digitally on [DATE], and in physical formats on [DATE].
		[ORG] released the film soundtrack digitally on [DATE], and physically on [DATE].
4	NBA	The officials estimated that the play took [TIME], and since there were [TIME] remaining when it began, the field goal was counted.
		Because of the rule, the officials then estimated that the play took [TIME], and because there were [TIME] remaining when play began, the field goal was allowed to be counted.
	Chinese Province	[GPE] hare is a species of hare endemic to [GPE].
		[GPE] partridge is a species of bird endemic to [GPE] Island.
3	NBA	After [PERSON] early retirement in [DATE], the [ORG] entered a period of rebuilding, completely dismantling and retooling their roster.
		After the retirement of [PERSON] in [DATE], the [ORG] entered a period of rebuilding.
	MARVEL	[PERSON]as [PERSON] Hill: A high-ranking [ORG] agent who works closely with [PERSON].
		[PERSON]as [PERSON] Hill: A former high-ranking [ORG] agent who now works for [PERSON].
2	Chines Province	[GPE] is [CARDINAL] of the most important commodity grain bases in [GPE].
		[GPE] is [CARDINAL] of the richest and most developed provinces in [GPE].
	MARVEL	The wife of [PERSON], queen of [PERSON], mother of [PERSON], and adoptive mother of [PERSON], based on the mythological deity of the same name.
		The king of [PERSON], father of [PERSON]and [PERSON], and adoptive father of [PERSON], based on the deity of the same name.
1	Chinese Province	Many of [GPE] [CARDINAL] universities consistently rank among the best in [GPE], such as [ORG] and [ORG].
		[GPE] is ranked the lowest among [GPE] [CARDINAL] provinces on [ORG] according to [ORG] data.
	NBA	However, [PERSON] was willing to spend money and urged [PERSON] to make draft-day trades.
		[PERSON] initially agreed to seek a trade, but also worked to try to change [PERSON] mind.

参 考 文 献

[1] Olson G M, Olson J S. Human-computer interaction: Psychological aspects of the human use of computing[J]. Annual Review of Psychology, 2003, 54(1): 491-516.

[2] Du J, Jiang J, Way A. Facilitating translation using source language paraphrase lattices[C]. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2010: 420-429.

[3] Xu Y, Martínez-Gómez P, Miyao Y, et al. Paraphrase for open question answering: New dataset and methods[C]. Proceedings of the Workshop on Human-Computer Question Answering. 2016: 53-61.

[4] Zhao S, Meng R, He D, et al. Integrating transformer and paraphrase rules for sentence simplification[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 3164-3173

[5] Barzilay R, McKeown K. Extracting paraphrases from a parallel corpus[C]. In Proceedings of the 39th annual meeting of the Association for Computational Linguistics. 2001:50-57.

[6] Ibrahim Ali, Katz Boris, Lin Jimmy. Extracting structural paraphrases from aligned monolingual corpora[C]. In Proceedings of the Second International Workshop on Paraphrasing, Sapporo, Japan, July. 2003:57-64.

[7] Zhao S, Wang H, Liu T, et al. Pivot approach for extracting paraphrase patterns from bilingual corpora[C]. Proceedings of ACL-20008: HLT. 2008: 780-788.

doi:10.3772/j.issn.2095-915x.2021.05.004

[8] Pang B, Knight K, Marcu D. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences[C]. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. 2003:181-188.

[9] Fujita A, Isabelle P, Roland Kuhn R. Enlarging paraphrase collections through generalization and instantiation[C]. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.2012:631-642.

[10] Regneri M, Rui Wang R. Using discourse information for paraphrase extraction[C]. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, July. 2012: 916-927.

[11] Shinyama Y, Sekine S, Sudo K. Automatic paraphrase acquisition from news articles[C]. Proceedings of the second international conference on Human Language Technology Research. 2002: 313-318.

[12] Sekine S. Automatic paraphrase discovery based on context and keywords between ne pairs[C]. Proceedings of the Third International Workshop on Paraphrasing (IWP2005). 2005.

[13] Das P, Das A K. An Unsupervised Approach of Paraphrase Discovery from Large Crime Corpus[C]. 2018 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2018: 1-6.

[14] Biran O, Blevins T, McKeown K. Mining paraphrasal typed templates from a plain text corpus[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 1913-1923.

[15] Liu M, Zhang Y, Xu J, et al. An Open Domain Paraphrasing Patterns Acquisition Based on Deep Semantic Computing[J]. Journal of Chinese Information Processing, 2018 (2):12.

[16] Wang R, Callison-Burch C. Paraphrase fragment extraction from monolingual comparable corpora. In Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web. 2011:52-60.

[17] Qi P, Zhang Y H, Bolton J., et al. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. 2020.

[18] Arora S, Liang Y, Ma T. A simple but tough-to-beat baseline for sentence embeddings[C]. The 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017.

[19] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]. Advances in neural information processing systems. 2013: 3111-3119.

[20] Takase S, Okazaki N, Inui K. Composing distributed representations of relational patterns[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2017: 2276-2286.