融合外部知识的开放域复述模板获取方法

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
† 通讯作者，Email:yjzhang@bjtu.edu.cn

摘要

如何挖掘语言资源中丰富的复述模板，是复述研究中的一项重要任务。已有方法在人工给定种子实体对的基础上，利用实体关系，通过自举迭代方式，从开放域获取复述模板，规避对平行语料或可比语料的依赖，但是该方法需人工给定实体对，实体关系受限；在迭代过程中语义会发生偏移，影响获取质量。针对这些问题，我们考虑知识库中包含描述特定语义关系的实体对（即关系三元组），提出融合外部知识的开放域复述模板自动获取方法。首先，将关系三元组与开放域文本对齐，获取关系对应文本，并将文本中语义丰富部分泛化成变量槽，获取关系模板；接着设计模板表示方法，本文利用预训练语言模型，在模板表示中融合变量槽语义；最后，根据获得的模板表示，设计自动聚类与筛选方法，获取高精度的复述模板。在融合自动评测与人工评测的评价方法下，实验结果表明，本文提出的方法实现了在开放域数据上复述模板的自动泛化与获取，能够获得质量高、语义一致的复述模板。

关键词： 复述模板 ； 语义表示 ； 自动聚类

An Open Domain Paraphrasing Template Acquisition Method Based on External Knowledge

Bo Jin, Mingtong Liu, Yujie Zhang †, Jinan Xu, Yufeng Chen
School of Computer and Information Technology, Beijing Jiaotong University
Beijing 100044, China
† Corresponding author, Email:yjzhang@bjtu.edu.cn

Abstract

How to obtain paraphrase knowledge from existing online resources is an important task in paraphrase research. This paper proposes a method of mining paraphrasing templates with external knowledge from open domain. We align the relational triples with the open domain text to obtain the corresponding text of the relation; then, design the template generalization method to obtain high-quality templates from the relation text; finally, utilize the pre-training language model to design template representation and mining methods. We designed a method of both automatic evaluation and manual evaluation to judge the quality of the paraphrasing template. Experiments have shown that, the method proposed in this paper realizes the automatic acquisition of high-quality paraphrasing template.

Keywords: paraphrasing template , semantic representation , automatic clustering

国家自然科学基金（61876198,61976015,61976016）资助

1 引言

复述(paraphrases)，是指用不同的表达方式表述相同语义的语言现象，在自然语言中广泛存在。复述技术在自然语言处理领域应用广泛，其中，复述模板获取是复述领域研究热点之一。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，是对复述知识的高度抽象。应用复述模板，既可以构建复述数据，也可以提升下游任务的性能，例如文本摘要 (Zhao et al., 2018)、自动问答 (Xu et al., 2016)、机器翻译 (Du et al., 2010)等。

基于复述平行语料、可比语料或是双语语料，从复述实例泛化得到复述模板的方法简单有效，但是，这些方法均先找到存在复述关系的句子，再将句子泛化为模板，从而获取复述模板，但这些方法忽视了具备潜在复述关系的句对，举例来说，如图 1所示，“Wu Chengen wrote a Journey to the West”和“The author of Dream of the Red Chamber is Cao Xueqin”这两句话并不存在复述关系，但泛化后的模板“[Writer] wrote [Book]”和“The author of [Book] is [Writer]”之间存在复述关系。我们称这样的句对存在潜在复述关系，开放域中大量存在这样的句对，因此，应该从模板中寻找复述关系，提升复述模板的多样性。

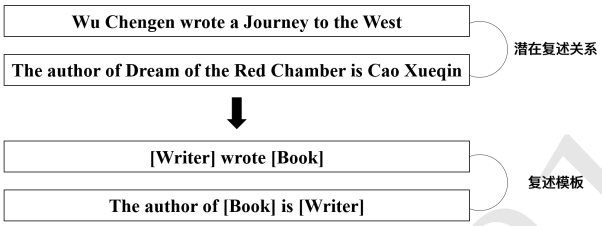

Figure 1: 存在潜在复述关系的文本与复述模板

在开放域中具有潜在复述关系的句对大量存在，但已有方法难以挖掘，例如Liu等人(2018)通过人工给定种子实体对，采取自举迭代的方式从互联网获取复述模板。此方法需人工事先给定实体对，抽取的复述模板往往集中在一些典型的实体关系上，多样性不高，并且，在迭代时，会发生语义偏移的现象，导致抽取模板质量不高；其次，该方法在模板表示时忽视对变量槽的语义约束，无法区分变量槽细微的语义差异；最后，先前复述模板研究都依赖于人工评测，效率低、主观性强，缺少统一的复述模板评测方法。

为了解决上述问题，本文主要针对复述模板获取与评测方法展开研究，主要贡献如下：首先，我们融合外部知识，从开放域获取存在潜在复述关系的文本（即关系实例），利用知识库中描述相同关系的三元组，从开放域批量获取关系实例，规避了迭代过程中的语义偏移；同时，我们的方法可以应用到任意关系上，可拓展性强；接着，我们在设计模板表示方法时，利用预训练语言模型获取变量槽的语义信息，解决变量槽语义约束缺失问题；最后，我们设计了融合自动评测与人工评测的评价方法，从多个角度评价复述模板的质量。

本文剩余部分的组织结构如下：第二节介绍相关研究；第三节描述获取关系实例，并泛化为模板的方法；第四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；第五节介绍实验评测和分析结果；第六节对本文研究进行总结。

2 相关研究

针对复述模板抽取的研究起源于信息检索和信息抽取领域，历史悠久，主要用于对搜索模式、检索问题进行自动扩展。Hearst(1992)最早将模板应用到信息抽取领域，其目的是借助模板的帮助检索质量良好的数据。

随着自然语言处理的发展，挖掘复述文本成为一个重要且得到深入研究的领域，复述文本有助于问答、摘要等研究，从大规模文本中挖掘复述句对与我们的工作有相似之处，先前的工作往往利用复述平行语料(Barzilay and McKeown, 2001; Bannard and Callison-Burch, 2005; Ibrahim et al., 2003; Pang et al., 2003; Fujita et al., 2012; Regneri and Wang, 2012)、可比语料(Shinyama et al., 2002; Sekine, 2005; Shen et al., 2006; Wang and Callison-Burch, 2011)或是双语语料(Bannard and Callison-Burch, 2005; Madnani et al., 2008; Zhao et al., 2009)来提取复述句对。其中，Barzilay和Lee(2003)构造的带槽的网格（slotted lattice）与模板类似，但是该方法构造的变量槽没有类型信息，因此要依赖译本相关的复述实例库获取复述模板。Shinyama等

人(2002)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(2005)构造带类型信息的短语模板，但该方法以短语中的关键词为分类依据，且局限于存在两个实体的短语；Priyanka Das等人(2018)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。

近年来，随着深度学习以及自然语言处理技术的不断发展，出现了针对单语语料的研究(Biran et al., 2016; LIU et al., 2018)。由于单语语料缺失语义等价线索，因此，在单语语料上的复述识别与复述获取极具挑战，Biran等人(2016)选择获取相同领域的维基百科文章，利用表层字符特征作为复述模板聚类依据，但是领域这一限定条件过于宽泛，同一领域的文本混杂着过多噪音。Liu等人(2018)的方法与我们较为接近，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。

本文提出融合外部知识的复述模板获取方法，既能够规避对语料的依赖，同时也对模板语义表示方法以及聚类算法做出改进。

3 关系实例获取与模板泛化方法

怎样才能从单语语料中获得语义相近的文本？单语语料中缺少语义等价线索，这使得单语上的复述识别与复述模板获取极具挑战性。考虑到句子中实体往往蕴含丰富的语义信息，我们选择实体对作为句中语义关系的修饰特征，首先从知识库中获取大量有相同关系的三元组，将三元组与开放域文本对齐，获得大量具有相同实体关系的实例；接着对实例泛化，获取关系模板。关系模板的获取过程由以下两个部分组成。

3.1 关系实例获取

参考前人方法(Elsahar et al., 2019)，我们将知识库中的关系三元组与维基百科开放域文本对齐，从关系三元组出发获取大量描述相同关系的实例文本。关系三元组由subject、object和描述关系谓词predicate组成，当关系三元组的SPO与文本中实体以及关系谓词完全一致时，我们认为关系三元组与该文本对齐。

3.2 关系模板泛化

在获得关系实例后，我们需要对其进行泛化，获取关系模板。

变量槽位置选取 获取关系模板的第一步是要选取变量槽，本文采取多种方式：首先，我们利用正则表达式，从句中挑选出日期、百分数、货币、数字等作为变量槽，同时指定对应类型（例如“2020-01-01”对应类型为[DATE]）；接着，使用斯坦福开源工具[1]对所有的关系实例进行词性标注，选取符合以下条件的词或短语作为变量槽，一是仅包含NNP或NNPS标签的词，二是以NNP开头和结尾，同时仅包含NNP、TO、IN和DT标签的词，三是仅由大写字母组成的单词；最后，我们选取当前关系实例对应三元组中的头尾实体（subject、object）作为变量槽。考虑到变量槽过多会导致模板结构复杂，降低模板的抽象能力，本文仅保留变量槽个数在2至5之间的模板。

变量槽类型标注 在变量槽位置选取完成后，需要限定变量槽类型。由于已经获取了变量槽位置，我们可以将变量槽类型标注当做细粒度实体类别标注任务处理。本文利用Open Entity数据集(Choi et al., 2018)[2]，在细粒度实体类型标注任务上，对LUKE(Yamada et al., 2020)（Language Understanding with Knowledge-based Embeddings）模型进行微调，利用微调后的模型进行变量槽类型标注。给定模板$t_x = \{w_1, w_2, ...w_i, s_j, ..., w_l\}$，假设该模板由$m$个单词和$n$个变量槽组成，在标注变量槽$s_j$的类型时，我们先将$s_j$替换为[MASK]标签，将替换后的$t_x$输入到微调后的模型中，获得$s_j$对应的类别。

最终，在寻找复述模板之前，按照变量槽类型对关系模板进行分组处理，我们将拥有相同变量槽类型（与顺序无关）的关系模板分为同一组。

4 融合深度语义表示的复述模板获取方法

按照第三节介绍的方法，由关系三元组出发，获取修饰关系相同的关系模板，并按照变量槽类型进行分组。研究结果表明，同一组关系模板之间仍存在一些语义差异，并不能直接看做

[1]https://stanfordnlp.github.io/stanza/
[2]http://nlp.cs.washington.edu/entity_type/

复述模板。针对这一问题，本文利用预训练语言模型，融合变量槽语义，获取模板表示，接着设计自动聚类及筛选方法，从同组关系模板中获取复述模板。

4.1 模板语义计算方法

模板由词语和变量槽构成，词语部分包含语义信息，变量槽部分在包含语义信息的同时，也包含类型约束。先前研究在设计模板表示方法时，只考虑到模板中词语部分的信息，忽视了包含重要信息的变量槽。我们针对上述问题，提出了融合变量槽信息的模板语义计算方法，获取包含语义信息以及变量槽信息的模板表示。

词语部分 我们选择GloVe(Pennington et al., 2014)[3]作为词向量训练工具，上下文窗口大小设置为5，选择维基百科英文语料[4]作为训练语料，获取维度为300的词向量v_{w_i}。

变量槽部分 模板的变量槽中蕴含丰富的语义信息，先前研究忽视了对变量槽的语义约束，我们选择利用BERT获取变量槽表示。如公式 1所示，对于模板t_x，首先，我们用[MASK]标签替换模板t_x中的变量槽s_j，接着，我们将BERT后四层输出取平均，对应于[MASK]标签的输出即为变量槽s_j的表示v_{s_j}：

$$v_{s_j} = BERT(\{w_1, ...[MASK]_{s_j}, ..., w_l\}) \tag{1}$$

在分别获取词语与变量槽表示后，先前研究通常采用取算术平均的方式，获得模板表示，但是研究表明，模板中词语及变量槽的重要程度各不相同，不同部分在模板表示中应当占据不同的权重。本文根据词频以及人工给定的超参数，赋予词语以及变量槽不同权重，获取模板表示。

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 2所示。

算法 1 模板语义表示

输入： 单词语义表示 $\{v_{w_i}, w_i \in Vocab\}$，变量槽语义表示 $\{v_{s_j}, s_j \in Slot\ Types\}$，模板集合 T，超参数 a，单词出现频率 $\{p(w_i) : w_i \in Vocab\}$，变量槽出现频率 $\{p(s_j) : s_j \in Slot\ Types\}$

输出： 模板表示 $\{v_t : t \in T\}$

for all template t in T **do**

  $v_t = \frac{1}{|t|}(\sum_{w_i \in t} \frac{a}{a+p(w_i)} v_{w_i} + \sum_{s_j \in t} \frac{a}{a+p(s_j)} v_{s_j})$

end for

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

for all template t in T **do**

  $v_t = v_t - uu^T v_t$

end for

Figure 2: 模板表示方法

4.2 自动聚类获取复述模板

同一分组模板之间仍存在细微的语义差异，我们需要在每一组中分别寻找互为复述的模板。语义越接近的模板，其在同一语义空间上的距离也越接近，根据上一节获得的模板表示，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(Dasgupta and Long, 2005)为基础获取复述模板。先前研究在复述模板聚类时，根据前期实验结果，人工设置聚类簇个数(LIU et al., 2018)，这种设置方式针对特定数据，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 3所示。

[3]https://github.com/stanfordnlp/GloVe

[4]https://dumps.wikimedia.org/

给定拥有n个模板的分组$T = \{t_1, t_2, ..., t_n\}$，我们利用上一小节提出的模板表示方法获得模板表示$v_T = \{v_{t_1}, v_{t_2}, ...v_{t_n}\}$，通过计算模板表示之间的欧几里得距离，来衡量模板间的语义相似度，作为自动聚类的标准。

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
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
  end if
end while

Figure 3: 复述模板自动聚类算法

在获得对应的聚类标签后，我们舍弃只有一个模板的聚类簇，当聚类簇中模板数量大于等于3时，我们对该聚类簇进行筛选处理。在筛选时，首先计算该聚类簇中模板间余弦相似度矩阵，对每一个模板，寻找距离最接近的模板，将这两个模板看做复述模板，作为最终的抽取结果。

5 实验评价与结果分析

为了验证本文所提方法的有效性，我们在开放域数据上进行评测实验。

5.1 实验数据

本文在六种语义关系实例上进行实验，关系实例的描述以及数据规模如表 1所示。

关系	描述	数据规模
Sport	体育活动	31598
Capital of	国家、州、省或其他行政区划的首府	10237
Director	电影、电视剧、舞台剧、视频游戏的导演	9850
Author	书面作品的作者	6368
Member of	组织、音乐团体或俱乐部的成员	4298
Conflict	人员、装备参与的军事行动、地区冲突	2798

Table 1: 关系实例描述及数据规模

我们选择Wikidata(Vrandečić and Krötzsch, 2014)[5]作为知识库，从Wikidata中获取上述关

[5]www.wikidata.org

系对应的三元组，关系三元组的部分示例如附录A中表 6所示，接着，我们将关系三元组与英文维基百科[6]对齐获取关系实例文本。抽取关系实例文本的部分示例如附录A中表 7所示。

5.2 复述模板自动评价的实验结果

在获取关系模板后，我们利用第四节描述的方法获取复述模板，先前研究在评价复述模板时，依赖人工评测方法，效率低。我们设计了自动评价方法：对于每个模板对，我们随机选择其中一个模板，使用其原始实体填入模板变量槽中，还原成复述句对，接着评价复述句对的质量。我们选择BLEU、ROUGE-L以及BERT-Score作为评测指标，其中，BLEU和ROUGE是基于词重叠率的方法，这两项指标能够体现句对之间形式上的相似性，BLEU评价句对之间n-gram的共现度，ROUGE-L基于最长公共子序列计算贡献概率；但是这两项指标也存在一定的缺陷，它们仅对词汇变化敏感，并不能够识别语义、语法的变化，无法正确处理字面表达不同但语义相同的现象，因此我们选择BERT-Score作为补充，BERT-Score利用预训练语言模型，建模句对之间的语义相近度，从而能够从语义角度评价复述模板质量。

关系	模板数量	BLEU	ROUGE-L			BERT-Score		
			P /%	R /%	F /%	P /%	R /%	F /%
Sport	4465	58.70	75.19	75.58	75.28	63.86	63.69	63.78
Capital of	1908	67.78	79.89	80.13	80.24	70.48	70.43	70.45
Director	1516	57.14	74.42	76.01	74.07	68.79	66.69	67.72
Author	1431	67.97	83.44	83.48	83.81	78.96	79.16	79.07
Member of	626	61.83	76.47	77.31	76.25	68.47	68.46	68.47
Conflict	582	62.08	79.02	79.43	79.12	72.67	71.80	72.24

Table 2: 复述模板自动评测结果

实验结果如表 2所示，在表现最好的Author关系上，从词语重叠率评价角度看，BLEU值为67.97，ROUGE-L对应的F1值为83.81%，从语义相似度来看，BERT-Score对应的F1值为79.07%；即使是在指标最差的Sport关系上，BLEU、ROUGE-L以及BERT-Score也能分别达到58.70、75.28%、63.78%。自动评测的结果能够充分证明，本文提出的从关系三元组出发，融合外部知识的开放域复述模板获取方法，能够有效抽取出质量优秀的复述模板。

5.3 复述模板人工评价结果

本节通过人工评测方法，对复述模板质量进行评测，作为对自动评价方法的补充。我们从每一类关系类别抽取结果中随机选取200对复述模板用于人工评价。先前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，评价指标分为1 5得分，具体评分标准如表 3所示。

得分	评价标准
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

Table 3: 人工评测指标

人工评测的结果如表 4所示，我们在统计平均得分的同时，统计了三分、四分以上模板的占比。得分大于三分的复述模板占比均在70%以上，这说明大多数模板对之间语义相似；从人工评测的结果我们能看出，本文提出的抽取方法，在不同的领域、关系上的表现差异不大，这充分说明，该方法是具有拓展性的，能够应用到开放域上，从而获取大规模的复述数据。

[6]https://en.wikipedia.org/

关系	平均得分	%3+	%4+
Sport	3.45	79	61
Capital of	3.52	75	66
Director	3.65	76	67
Author	3.6	78	68
Member of	3.74	82	64
Conflict	4.2	89	72

Table 4: 人工评测结果

5.4 消融实验分析

为了验证变量槽约束以及聚类后筛选步骤的效果，我们选择在“Capital of”关系上进行消融实验，实验结果如表 5所示。

Capital of	模板数量	平均得分	BLEU	ROUGE-L	BERT-Score
原方法	2146	3.32	60.59	71.83	61.56
+变量槽类型信息约束	2077	3.61	63.89	76.84	68.46
+聚类后筛选	1939	3.75	66.62	77.03	65.41
+变量槽类型信息约束 +聚类后筛选	1908	3.92	67.78	80.24	70.45

Table 5: 消融实验

实验结果表明，当我们在模板表示中融入变量槽信息时，BERT-Score指标得到了显著提升，BERT-Score指标能够有效地衡量语义相似程度，这表明对变量槽类型加以约束，能够有效地提升模板表示的精度，进而提高抽取出的复述模板的质量。而聚类后在聚类簇中增加筛选步骤，是对聚类获取复述模板的一次择优。通过增加筛选步骤，人工评分以及各项自动评测指标均得到一定提升，且人工评分提升较大。这表明，筛选能够在有效剔除噪音的同时，选出语义最接近的模板对，丰富复述模板的实际应用价值。

6 总结

本文提出融合外部知识的开放域复述模板获取方法，针对先前研究在关系实例获取中语义偏移、关系受限等问题，设计利用知识库，从关系三元组出发的关系实例获取方法；针对模板表示中变量槽的语义约束缺失等问题，设计利用预训练语言模型获取变量槽表示的方法；最后，为了更好的评价复述模板的质量，我们设计融合自动评测与人工评测的评价方法，实验结果表明，本文提出的方法可以获取到语义一致，质量高的复述模板。在未来的研究中，我们将尝试更细粒度的变量槽类型，继续改进变量槽类型标注方法；同时，改进模板语义表示方法，利用更精准的变量槽表示提高复述模板自动获取的精度和复述模板的质量。

参考文献

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. *arXiv preprint cs/0304006*.

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 50–57.

Or Biran, Terra Blevins, and Kathleen McKeown. 2016. Mining paraphrasal typed templates from a plain text corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1923.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. *arXiv preprint arXiv:1807.04905*.

Priyanka Das and Asit Kumar Das. 2018. An unsupervised approach of paraphrase discovery from large crime corpus. In *2018 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6. IEEE.

Sanjoy Dasgupta and Philip M Long. 2005. Performance guarantees for hierarchical clustering. *Journal of Computer and System Sciences*, 70(4):555–569.

Jinhua Du, Jie Jiang, and Andy Way. 2010. Facilitating translation using source language paraphrase lattices. Association for Computational Linguistics.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Elena Simperl, and Frederique Laforest. 2019. T-rex: A large scale alignment of natural language with knowledge base triples.

Atsushi Fujita, Pierre Isabelle, and Roland Kuhn. 2012. Enlarging paraphrase collections through generalization and instantiation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 631–642.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Coling 1992 volume 2: The 15th international conference on computational linguistics*.

Ali Ibrahim, Boris Katz, and Jimmy Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the Second International Workshop on Paraphrasing*, pages 57–64, Sapporo, Japan, July. Association for Computational Linguistics.

Mingtong LIU, Yujie ZHANG, Jinan XU, and Yufeng CHEN. 2018. An open domain paraphrasing patterns acquisition based on deep semantic computing. *Journal of Chinese Information Processing*, page 02.

Nitin Madnani, Philip Resnik, Bonnie Dorr, and Richard Schwartz. 2008. Applying automatically generated semantic knowledge: A case study in machine translation. In *NSF Symposium on Semantic Knowledge Discovery, Organization and Use*, pages 60–61. Citeseer.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–188.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Michaela Regneri and Rui Wang. 2012. Using discourse information for paraphrase extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 916–927, Jeju Island, Korea, July. Association for Computational Linguistics.

Satoshi Sekine. 2005. Automatic paraphrase discovery based on context and keywords between NE pairs. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Siwei Shen, Dragomir Radev, Agam Patel, and Gunes Erkan. 2006. Adding syntax to dynamic programming for aligning comparable texts for the generation of paraphrases. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 747–754.

Yusuke Shinyama, Satoshi Sekine, Kiyoshi Sudo, and Ralph Grishman. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of HLT*, volume 2, page 1. San Diego, US.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Rui Wang and Chris Callison-Burch. 2011. Paraphrase fragment extraction from monolingual comparable corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 52–60.

Ying Xu, Pascual Martínez-Gómez, Yusuke Miyao, and Randy Goebel. 2016. Paraphrase for open question answering: New dataset and methods. In *Proceedings of the Workshop on Human-Computer Question Answering*, pages 53–61.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online, November. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 834–842.

Sanqiang Zhao, Rui Meng, Daqing He, Saptono Andi, and Parmanto Bambang. 2018. Integrating transformer and paraphrase rules for sentence simplification. *arXiv preprint arXiv:1810.11193*.

附录A.实验数据示例

关系	三元组实例
Sport	{cricket, Victoria}, {bowler, batsman}, {footballer, Simon Grayson}
Capital of	{South Australia, Adelaide}, {Turkey, Ankara}, {Kenya, Nairobi}
Director	{Tsui Hark, New Dragon Gate Inn}, {Hou Hsiao-Hsien, Flowers of Shanghai}, {Peter Greenaway, $8\frac{1}{2}$ Women}
Author	{Priestley, An Inspector Calls}, {William Blake, Jerusalem}, {Shakespeare, Macbeth}
Member of	{USSR, RSFSR}, {Zyklon, Samoth}, {FRCP, Heinz Wolff}
Conflict	{World War I, Austro-Hungarian Army}, {Second World War, AMC 34}, {American Civil War, James William Boyd}

Table 6: 关系三元组部分示例

关系	关系实例
Sport	Alfredo bodoira was an italian professional football player, who played as a goalkeeper.
	Yury berezhko is a volleyball player from russia.
Capital of	He still would attend gafcon ii, in nairobi, kenya.
	Porta cavalleggeri was one of the gates of the leonine wall in rome.
Director	Thark is a 1932 film farce, directed by tom walls, with a script by ben travers.
	The episode was written by bill oakley and josh weinstein and directed by wes archer.
Author	Emmanuel goldstein is a key character in george orwell's novel nineteen eighty-four.
	Gaal dornick is a fictional character in isaac asimov's foundation series.
Member of	Corke is a life member of kappa alpha psi fraternity.
	He is a corresponding member, member of the hungarian academy of sciences.
Conflict	Erich handke was a highly decorated oberfeldwebel in the luftwaffe during world war ii.
	Following the outbreak of world war ii he served as a captain in the u.s.

Table 7: 关系实例文本部分示例

附录B.复述模板示例

表 8 分别给出了各个关系中正确的复述模板实例,

关系	复述模板示例
Sport	[**Person**] was an [**Country**] professional football player, who played as a [**Profession**].
	[**Person**] is a retired [**Country**] professional football player, who played in the position of [**Profession**].
Capital of	[**Person**] established the first newspaper of [**City**], in [**Country**] in [**Date**].
	the first newspaper in [**Country**] was published by [**Person**], in [**City**] in [**Date**].
Director	the episode was written by [**Person1**] and [**Person2**] and directed by [**Person3**] .
	[**Person1**] and [**Person2**] wrote the episode; it was directed by [**Person3**] .
Author	it was first depicted on a celestial atlas by [**Person**] in his [**Work**].
	its first depiction in a celestial atlas was in [**Person**]'s [**Work**].
Member of	[**Person**] was one of the founding members of the [**Organization**] in [**Date**].
	[**Person**] was one of the original participators of the [**Organization**], founded in [**Date**].
Conflict	following the outbreak of [**War**] [**Person**] served as a captain in the [**Country**].
	during [**War**] , [**Person**] was a captain in the [**Country**].

Table 8: 复述模板抽取结果的部分示例