文章编号：1003-0077(2018)01-0094-08

开放域上基于深度语义计算的复述模板获取方法

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：利用实体关系从网络大规模单语语料获取复述模板的方法可以规避对单语平行语料或可比语料的依赖，但是后期需要人工对有语义差异的关系模板分类后获取复述模板。针对这一遗留问题，该文提出基于深度语义计算的复述模板自动获取方法，首先设计基于统计特征的模板裁剪方法，从非复述语料中获取高质量的关系模板，然后设计基于深度语义计算的关系模板聚类方法获取高精度的复述模板。我们在四类实体关系数据上的实验结果表明，该方法实现了关系模板的自动获取与自动聚类，可以获得语义相近度更高、表现形式多样的复述模板。

关键词：关系模板；复述模板；深度语义计算；自动聚类

中图分类号：TP391 **文献标识码**：A

An Open Domain Paraphrasing Patterns Acquisition Based on Deep Semantic Computing

LIU Mingtong, ZHANG Yujie, XU Jinan, CHEN Yufeng

(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: This paper proposes a method of paraphrasing pattern acquisition based on deep semantic computing. We design a sentence segmentation method based on statistical features to obtain high-quality relational patterns from non-paraphrasing corpus, and then the paraphrasing patterns are detected by entity relation patterns. Finally, the patterns are automatically clustered according to their semantic similarity. Our experimental results on the four types of entity relation show that our method acquired paraphrasing patterns with good performance, with more diversity and closer semantic relation.

Key words: relational pattern; paraphrasing pattern; deep semantic computing; automatic clustering

0 引言

自然语言理解是自然语言处理的终极目标，其判定标准包括复述、翻译、问答和文摘，复述处理的复杂性和重要性可见一斑[1]。复述广泛应用于信息抽取[2-5]、机器翻译[6-8]和自动问答[9]等自然语言处理任务中。近年来，复述研究日益成为关注热点，成为自然语言处理的重要方向之一。

在自然语言处理具体任务中，复述技术包括复述识别和复述生成，而这两项任务都是以复述知识为基础，因此，复述知识的获取方法一直是复述研究的核心内容。复述知识的获取一般集中在短语、句子、模板及篇章四个级别[1]。由于复述模板具有对复述知识高度抽象的概括能力，因此，复述模板获取方法的研究成为了主要课题。例如，“曹雪芹写了《红楼梦》”和“《红楼梦》的作者是曹雪芹”是一个句子级别的复述实例。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。对上述复述实例泛化可以得到复述模板：“[X]写了[Y]”和“[Y]的作者是[X]”。

早期的方法通过对复述实例泛化获取复述模板，但复述实例资源较为匮乏，因此，难以获取多样化的复述模板。后来，从大规模单语语料中抽取复

收稿日期：2017-09-27 **定稿日期**：2017-10-20

基金项目：北京交通大学人才基金(KKRC11001532)；国家自然科学基金(61370130，61473294)；北京市自然科学基金(4172047)

述模板的自举迭代方法受到关注[11]，因为它既规避了对复述实例语料的依赖，又能不受特定语料的限制，从而可以获取多样化的复述模板。例如早期方法中需要“《红楼梦》的作者是曹雪芹”和“曹雪芹写了《红楼梦》”这样的复述实例才能获取复述模板：“[作者]写了[作品]”和“[作品]的作者是[作者]”；而之后的自举迭代方法即使没有这样的复述实例，也可以从“曹雪芹写了《红楼梦》”和“《西游记》的作者是吴承恩”这样的例子中抽取出同样的复述模板。其主要原理是利用具有特定关系的实体对从大规模语料中抽取实例进行泛化获取模板，进而利用模板抽取实例扩展实体对，如此迭代地获取具有特定关系的模板作为复述模板。但是该方法自举迭代过程中存在语义飘移问题和获取关系模板存在语义差异问题，导致复述模板的质量不高，后续需要人工按照语义进一步细分类。

本文提出开放域上基于深度语义计算的复述模板获取方法，针对自举迭代过程中的语义飘移问题，设计基于统计特征的模板裁剪方法；针对需要人工细分类的问题，设计基于语义组合计算的模板自动聚类方法，从而提升复述模板的质量。

本文剩余部分组织如下：第一节介绍相关研究；第二节针对自举迭代方法中的语义飘移问题，描述基于统计特征的模板裁剪方法；第三节针对需要人工细分类的关系模板语义差异问题，描述基于深度语义计算的模板自动聚类方法；第四节介绍实验评价和结果分析；第五节对本文研究进行总结。

1　相关研究

复述模板的获取起源于信息抽取的需求，早期的方法通过对复述实例泛化获取复述模板。Barzilay 等人[10]使用词性序列泛化获得复述模板，李维刚[11]使用语义分类标识表示模板槽变量。

由于获取大量复述实例较为困难，研究者开始在单语语料展开研究，利用单语语料获取复述模板的方法主要依赖分布式假设。Lin 等人[12]提出以英语为对象的 DIRT 方法，利用句子依存路径两端的词语作为分布特征，寻找相近路径生成复述模板；Shinyama 等人[13]以日语为对象，收集依存路径两端的命名实体作为分布特征，然后将相近路径泛化得到模板；Biran[14]探索了基于知识库的复述模板获取方法，首先利用知识库对句子泛化得到特定语义类型的模板，然后通过聚类寻找复述关系；RELLY 系统[15]利用知识库中的上下位关系作为特征获得复述模板。随着深度学习在自然语言处理上的应用，在模板语义特征的计算中，Takase 等人[16-17]通过分布式表示学习模板的语义向量，然后对模板间的语义相关性进行排序。

因为单语语料没有语义等价的线索，使得复述识别和复述获取极具挑战性，已有研究方法使用了高精度的句法分析和大规模知识库，主要工作集中在英语和日语，而汉语上的复述模板研究很少。本文研究开放域上汉语复述模板的获取方法，采用自举迭代方法，针对其中语义飘移问题和模板语义差异问题，探索结合统计特征和深度语义计算的复述模板获取方法。Biran 的方法与我们较为接近，但他们仅利用了表层字符特征作为复述模板聚类依据，同时，他们利用了英语丰富的知识库获取复述模板，所以这一方法难以扩展到语言学资源不丰富的语言中。

2　基于统计特征裁剪的关系模板获取

根据分布式假说（distributional hypothesis）[18-19]，分布特征越相近的词语在语义上越接近，因而成为复述的可能性也越大。我们以实体对（entity pairs）作为两个实体之间语义关系的分布特征，首先利用种子实体对获取反映实体关系的句子（关系实例）；然后通过循环迭代扩展实体对，以获取大量具有相同实体关系的实例；最后对实例泛化获取关系模板。给定实体对意味着两个实体之间具有指定的语义关系，因此包含实体对的句子也具有指定的语义关系。自举迭代获取关系模板的过程由以下三个部分组成。

2.1　实体对扩展

参考前人的方法[20]，我们将一个种子实体对中的两个实体（E_1，E_2）和关系关键词（Key）以“E_1＋Key＋E_2”的形式作为查询输入搜索网络数据，从实体对出发自举迭代获取大量具有相同关系的实体对。例如，我们将“百度＋CEO＋李彦宏”三个元素输入搜索引擎，通过自举的方法可以获取如＜腾讯，马化腾＞，＜阿里巴巴，马云＞等具有相同关系的实体对。

2.2　关系实例获取

我们利用第 2.1 节获取的实体对作为查询输入

从网络获取关系实例，获取过程如图1所示。由于自举迭代扩展实体对的过程会导致语义飘移问题，我们采用如下策略进行数据预处理。

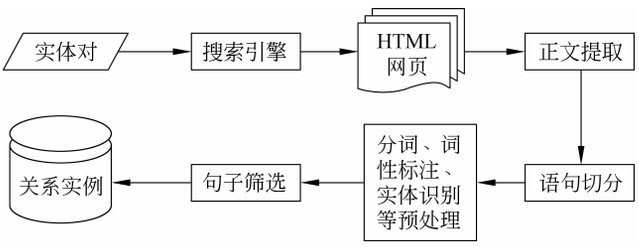

图1 关系实例获取流程图

(1) 对搜索返回结果中排名靠前的 $N(N=30)$ 个网页，我们只收集页面标题和摘要作为数据，并进行语句切分。

(2) 筛选句子，只保留完全包含查询实体对的句子。

(3) 进行分词、词性标注，以及命名实体识别。为了保证检索到的句子具有指定语义关系，我们进一步利用命名实体标注的词性类别，只保留和种子实体对词性类别一致的句子。

(4) 计算实体对间的距离，去掉距离超过一定长度的句子，本文取5。因为实体间距离过大，包含词汇信息较多，会降低模板的抽象能力。

过滤后的句子构成关系实例集合，我们利用以上方法筛选句子以获取高质量的关系实例。

2.3 关系模板获取

首先将实例中的实体对泛化成变量槽，并以实体的词性类别作为语义限制获取关系模板，如从实例“李彦宏/nr 担任/v 百度/nt CEO/nx”可以获得关系模板“PERSON/nr 担任/v ORGANIZATION/nt CEO/nx”。

模板泛化的主要问题是，从较长实例中获取的模板也较长，而较长模板的泛化能力较弱，为此我们参考前人方法[11]，采取如下模板裁剪策略：

(1) 基于长度的裁剪：以实体对包围的单词序列为中心，向左右分别扩展一定长度的窗口，作为初始模板，以此限制模板的长度。

(2) 基于统计特征的裁剪：为了对左右窗口部分进一步裁剪，对初始模板左右窗口中的单词分别计算与实体对的语义相关度作为边界可信度，分别选择边界可信度最大的单词作为左右边界。如此，保留和实体语义最相关的词语，删除不相关的词语，以提高模板泛化能力。我们基于统计特征设计边界可信度计算函数，词语 W_i 的边界可信度计算如下：

$$\mathrm{Confi}(W_i)=\frac{\mathrm{Pos}(W_i)\times\mathrm{Idf}(W_i)}{P(W_{i+j})\times\log(D(W_i)+1)} \quad (1)$$

其中：

$$\mathrm{Pos}(W_i)=\begin{cases}1 & \text{if } \mathrm{pos}(W_i)\in\{n,v,a\}\\ 0 & \text{other}\end{cases} \quad (2)$$

$$\mathrm{Idf}(W_i)=\frac{\mathrm{tfAsContext}(W_i)}{\mathrm{tfTotal}(W_i)} \quad (3)$$

$$P(W_{i+j})=\frac{\mathrm{Pos}(W_{i+j})\times\mathrm{Idf}(W_{i+j})}{\mathrm{Idf}(W_i)}+\lambda \quad (4)$$

$$D(W_i)=\mathrm{abs}(\mathrm{Loc}(W_i)-\mathrm{Loc}(W_{\mathrm{nearNE}})) \quad (5)$$

式(2)表示当单词词性为名词、动词和形容词时，函数取值为1，否则为0，因为这类词性的词语通常具有实质性的语义信息；式(3)计算单词出现在左右窗口的概率，其中 $\mathrm{tfAsContext}(W_i)$ 是 W_i 在左右窗口出现的次数，$\mathrm{tfTotal}(W_i)$ 是 W_i 在整个实例集合中出现的次数，如此可以降低“的”这类功能性词汇作为模板边界的可能性；式(4)主要考察了相邻候选词对当前候选词边界可信度的影响，确定左边界时 j 取 -1，右边界时 j 取1，其中 λ 是一个常数，本文取1，主要是对 $P(W_{i+j})$ 作为分母其值为0时做一个平滑处理；式(5)考察了候选词与最近实体之间的距离对边界可信度的影响，$\mathrm{Loc}(W_i)$ 表示 W_i 在句子中的位置，距离实体越近的词在语义上和实体越相关。

利用基于统计特征的方法，我们选取边界可信度最高的单词作为边界，进行模板裁剪。对裁剪后的模板统计数量，设定阈值，保留一定频度以上的模板。我们认为出现次数越多的模板质量越高。

3 基于深度语义计算的复述模板获取

按照第二节介绍方法，由给定实体对获取的关系模板应具有指定语义关系，但是研究结果表明，其中依然存在细微语义差异，导致关系模板不能直接作为复述模板，还需要人工进行细分类[11]。例如，给定实体对<百度，李彦宏>，在我们获取的关系模板结果中，模板中的词语包括“创新”“创业”“发言”“致辞”等，这些模板在语义上有一定差异。针对这一问题，本文提出的基于深度语义计算的关系模板自动聚类方法，希望能将上面例子中的“发言”和“致辞”这些模板分为一类，而将表达“创新”和“创业”语义的模板分为另一类。

3.1　模板语义计算方法

模板由变量槽和单词两部分构成，其中单词具有更加具体的语义信息，因此，我们利用单词部分计算模板的语义表示。如对模板 p＝“ORGANIZATION/nt 董事长/nnt PERSON/nr 表示/v”，我们抽取单词集合{董事长、表示}计算模板的语义。首先，计算每个单词的语义表示，我们利用 Word2Vec① 获取单词的词向量，采用 skip-gram 模型，上下文窗口设置为 5，利用中文维基百科语料② 作为训练数据，设置维度大小为 200。然后，利用单词的语义表示组合计算得到模板的语义表示。模板的语义组合计算方式有算术平均和几何平均两种，根据已有研究结果[14,16-17]，本文采用效果较好的算术平均方法进行语义组合计算，具体公式如(6)所示。

$$v(p)=\frac{1}{n}\sum_{i=1}^{n}v_i \tag{6}$$

其中，n 表示模板中单词个数，v_i 表示第 i 个单词对应的词向量。

3.2　自动聚类获取复述模板

语义越相近的模板在语义空间上的距离也越接近，由此可以在同一语义空间寻找复述模板。我们以 K-means 聚类算法[21] 为基础获取复述模板，针对聚类效果不稳定的初始中心点选择问题进行了改进，改进算法如图 1 所示。我们对聚类个数 K 值的设置基于如下考虑，从前期实验结果发现具有相同语义关系的复述模板一般为 5 个左右，即聚类后每个类的元素为 5，故设置聚类后的复述集合个数 K＝待聚类的模板个数/5。

1. K-means 初始质心选择算法：
2. 从输入模板集合中随机选择一个点作为第一个聚类中心 u_1；
3. 对于集合中的每一个点 p_i，计算它与已选择的聚类中心中最近聚类中心的距离：

$$D(p)=\operatorname{argmax}\sum_{r=1}^{k_{\text{selected}}}\|p_i-u_r\|_2$$

4. 选择一个新的数据点作为新的聚类中心，选择的原则是：$D(p)$较大的点被选取作为聚类中心；
5. 重复步骤 3 和 4，直到选择出 K 个聚类质心；
6. 利用这 K 个质心作为初始化质心运行 K-means 算法。

1. 基于 K-means 的模板聚类算法：
2. 输入：模板向量表示集合 $S(p)=\{p_1, p_2, p_3, \cdots, p_N\}$；$N$ 为模板总数；最大迭代次数为 T；设置聚类个数为 K
3. 输出：模板聚类集合$\{C_1, C_2, C_3, \cdots\cdots C_K\}$
4. 初始化：按照上述初始质心选择方法，从数据集 $S(p)$中选择 K 个样本点作为模板聚类的质心向量$\{u_1, u_2, u_3, \cdots, u_K\}$
5. For $t=1$ to T：初始化模板集合：将模板划分 C 初始化为 $C_k=\varnothing$，$k=1, 2, \cdots, K$
6. For $i=1$ to N：计算样本 p_i 和各个质心向量 $u_j(j=1,2,\cdots,K)$的距离：
   $d_{ij}=\|p_i-u_j\|_2$，将 p_i 标记为最小的为 d_{ij} 所对应的类别 λ_i，更新 $c_{\lambda_i}=c_{\lambda_i}\cup p_i$
7. End For
8. For j=1 to K：更新聚类中心

$$u_j=\frac{1}{|C_j|}\sum_{p\in C_j}p$$

9. End For
10. 如果所有的 K 个质心向量都没有发生变化，则转到步骤 12
11. End For
12. 输出聚类后的模板集合$\{C_1, C_2, C_3, \cdots, C_K\}$

图 1　K-means 聚类算法获取复述模板

4　实验评价与结果分析

为了验证本文所提方法的有效性，我们在网络开放域的数据上进行了评测实验。

4.1　实验数据

本文选取四类语义关系的实体对作为初始种子

① http://word2vec.googlecode.com.
② https://dumps.wikimedia.org.

进行实验。这四类关系分别是CEO关系、发明关系、病理关系、作品作者关系。我们利用百度搜索引擎[①]扩展实体对，扩展结果的部分示例见表1所示。然后利用这些实体对和百度搜索引擎获取实例，获取结果的部分示例见表2所示。本文使用HanLP[②]自然语言处理工具进行分词、词性标注、实体识别处理。下面是以这些实体对和实例作为模板泛化和聚类获取复述模板的实验数据。

表1 实体对扩展结果的部分示例

关系	实 体 对
CEO	＜百度，李彦宏＞；＜京东，刘强东＞；＜腾讯，马化腾＞
发明	＜莱特，飞机＞；＜爱迪生，电灯＞；＜张衡，地动仪＞
病理	＜肥胖，糖尿病＞；＜吸烟，冠心病＞；＜劳累，肾炎＞
作品	＜红楼梦，曹雪芹＞；＜边城，沈从文＞；＜人生，路遥＞

表2 关系实例结果的部分示例

关系类别	关 系 实 例
CEO关系	百度的CEO是李彦宏
	李彦宏担任百度首席执行官
	专访阿里巴巴董事长马云
发明关系	张衡制造了地动仪
	蔡伦改进了造纸术
	火车的发明者是斯蒂芬森
病理关系	高血压患者更易发生冠心病
	认为高血压是冠心病的重要引发因素
	高血压是冠心病的危险因素
作品关系	《老人与海》成为海明威的巅峰之作
	美国作家海明威写出了中篇小说《老人与海》
	《四世同堂》是老舍的代表作品

4.2 关系模板评价与分析

利用第二节描述的方法对上面获得的关系实例泛化、裁剪，获得关系模板。本文采用人工评测的方法，并使用下面的评测指标。

(1) 准确率(P)：如果一个模板能够正确表达“某人(PERSON)是某组织(ORGANIZATION)的CEO”的含义，则判断这个模板是正确的，准确率计算公式如式(7)所示。

(2) 覆盖率(C)：给定若干具有指定语义关系的实例，使用模板匹配这些实例，以能够匹配的实例个数计算覆盖率，覆盖率计算如式(8)所示。

$$\text{准确率} = \frac{\text{正确模板数}}{\text{总的关系模板数}} \times 100\% \quad (7)$$

$$\text{覆盖率} = \frac{\text{匹配实例数}}{\text{总的关系实例数}} \times 100\% \quad (8)$$

本文只在CEO关系的数据上进行了覆盖率评测，我们按以下方法构建评测数据：对一个给定的实例，如果实例中包含有一对实体，且实体对应的词性分别为人名和机构名，若这样的实例表达了CEO语义关系，就选择该实例作为标准评测数据。最终，我们构建500个关系实例集合用于覆盖率评测。

我们主要评价模板裁剪中不同阈值和窗口大小对关系模板准确率和覆盖率的影响，实验结果分别如表3和表4所示。

表3 频率阈值对模板获取的影响

频率阈值	模板总数	正确模板数	P/%	C/%
2	325	204	62.77	70.46
3	128	80	74.22	52.67
4	82	65	75.58	48.67
5	51	40	78.43	48.33

表4 窗口大小对模板获取的影响

窗口大小	模板总数	正确模板数	P/%	C/%
2	138	85	61.59	42.33
3	126	91	72.22	51.67
4	105	86	81.90	74.33
5	96	79	82.30	75.00

从表3中可以看出，随着阈值的增加，模板准确率随之提升，覆盖率随之下降，但覆盖率下降趋势较为缓慢。由此可以发现设置较高阈值获取的模板具有更好的泛化能力，可以覆盖更多的实例，这表明我们的方法可以获取到高质量的模板。同时，我们也发现随着阈值的不断增加，关系模板的数量下降较快。

从表4中可以看出，随着窗口的增大，模板的准确率和覆盖率都不断提升。分析其原因，当窗口增大时，有更多的上下文信息可以帮助判断模板的边

① https://www.baidu.com

② http://hanlp.linrunsoft.com

界。当窗口增加到 4 时，准确率和覆盖率的提升变得缓慢。

4.3　复述模板评价和分析

针对获取的关系模板，我们利用第 4 节描述的自动聚类方法获取复述模板，表 5 显示了从 CEO 关系模板获得的复述模板的六个聚类结果。本文对聚类结果按如下方法进行过滤：如果某类中只有一个模板，舍弃该类；如果某类中的模板数量超过 10，舍弃该类，我们假设复述模板的数量最多为 10，若超过 10 个，则认为该类中包含过多不正确的复述模板，将这样的类过滤。然后，我们对过滤后的聚类结果进行评测，评测方法如下：对每一类中的模板按照语义人工进行分组，每一组内的模板互为复述，然后选取模板数量最多的一组作为正确的复述模板集合，并计算正确复述模板的个数占类中模板总数的比例，得到每一类的正确率(P)，并对所有类计算平均正确率。

表 5　聚类获取复述模板的部分示例

ORGANIZATION/nt CEO/nx PERSON/nr 认为/v ORGANIZATION/nt 总裁/nnt PERSON/nr 看来/v	PERSON/nr 接任/v ORGANIZATION/nt CEO/nx PERSON/nr 担任/v ORGANIZATION/nt CEO/nx
ORGANIZATION/nt PERSON/nr 下台/vi ORGANIZATION/nt PERSON/nr 下课/vi 解读/v ORGANIZATION/nt PERSON/nr 离职/vi	专访/vn ORGANIZATION/nt 总经理/nnt PERSON/nr ORGANIZATION/nt 总裁/nnt PERSON/nr 采访/v 采访/v ORGANIZATION/nt CEO/nx PERSON/nr
ORGANIZATION/nt 缔造者/nnt 兼/v CEO/nx PERSON/nr ORGANIZATION/nt 创始人/nnt 兼/v 总裁/nnt PERSON/nr	ORGANIZATION/nt CEO/nx PERSON/nr 现身/v ORGANIZATION/nt 创始人/nnt PERSON/nr 出现/v

我们对 CEO 关系模板聚类后的结果进行评测，将聚类结果中的复述模板数量和正确率的关系列在表 6 中，PC 表示复述模板正确率。从表 6 的结果来看，当类中模板数量为 2 和 3 时，互为复述模板的可能性很高，随着类中模板数量的增多，正确率有所下降。

表 6　CEO 关系模板自动聚类后的复述模板

类的编号	模板总数	复述模板数量	正确率 PC
1	7	6	0.8574
2	10	7	0.7
3	2	2	1
4	2	2	1
5	5	3	0.6
6	3	3	1
7	2	2	1
8	2	2	1
9	2	2	1
10	6	4	0.6667
11	3	3	1
12	2	2	1
13	8	8	1
14	2	2	1
15	4	4	1

续表 6

类的编号	模板总数	复述模板数量	正确率 PC
16	4	3	0.75
17	7	4	0.5714
18	9	4	0.4444

表 7 给出了四类关系模板的聚类结果，其中 PR 表示关系模板正确率，PAC 表示复述模板平均正确率。

表 7　4 类关系模板的聚类评测结果

关系	模板总数	正确率 PR/%	平均正确率 PAC/%
CEO	211	72.04	86.61
发明	135	68.89	74.59
病理	303	69.31	76.52
作品	258	74.42	81.57

从表 7 的分析结果来看，本文基于深度语义计算的聚类方法可以有效过滤噪声模板，实现关系模板在细微语义层面上的深入分类，最终获得质量较高的复述模板。

关系模板的正确率与复述模板的平均正确率之间的关系如图 2 所示。从图 2 中可以看出，关系模板的质量对聚类结果有较大影响，当关系模板的质量提高时，自动聚类的效果就变好，获取复述模板的质量也随之提高，说明了本文关系模板裁剪方法的有效性和重要性。

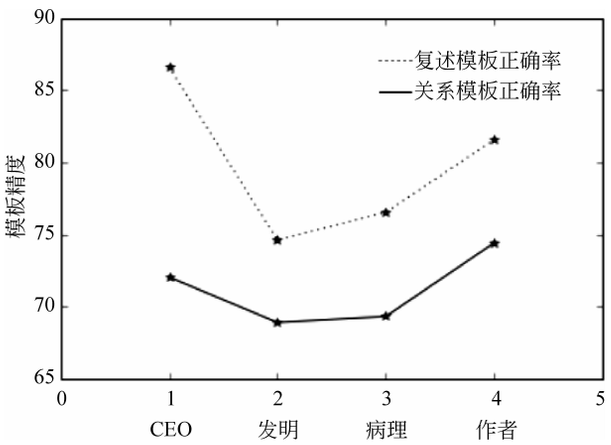

图2 关系模板正确率和复述模板正确率的关系

5 总结

本文提出开放域上基于深度语义计算的复述模板获取方法，针对自举迭代过程中的语义飘移问题，设计基于统计特征的模板裁剪方法；针对需要人工细分类模板语义的问题，设计基于语义组合计算的模板自动聚类方法。实验结果显示，本文可以获取到高质量的复述模板。针对未来的研究工作，我们需要进一步提高模板语义组合计算的精度，包括更精准的词向量学习方法，以及更有效的自动聚类算法，以提高自动获取复述模板的准确度。

参考文献

[1] 赵世奇，刘挺，李生. 复述技术研究[J]. 软件学报，2009(08)：2124-2137.

[2] Zhang Y, Yamamoto K. Paraphrasing of Chinese utterances[C]//Proceedings of COLING, 2002: 1163-1169.

[3] Ravichandran D, Hovy E. Learning surface text patterns for a question answering system[C]//Proceedings of ACL, 2002:41-47.

[4] Rahul Bhagat, Deepak Ravichandran. Large scale acquisition of paraphrases for learning surface patterns// Proceedings of ACL, Columbus, OH,2008: 674-682

[5] Chen B, Sun L, Han X, et al. Sentence rewriting for semantic parsing[C]//Proceedings of Meeting of the Association for computational linguistics, 2016:766-777.

[6] Nitin Madnani, Necip Fazil Ayan, Philip Resnik, et al. Using paraphrases for parameter tuning in statistical machine translation[C]//Proceedings of the Workshop on Statistical Machine Translation, Prague, 2007: 120-127

[7] Su J S, Dong H L, Chen Y D, et al. Improved statistical machine translation model with topic-based paraphrase[J]. Zhejiang Daxue Xuebao (Gongxue Ban)/ Journal of Zhejiang University (Engineering Science Edition, 2014, 48(10):1843-1849.

[8] Zhang L, Weng Z, Xiao W, et al. Extract domain-specific paraphrase from monolingual corpus for automatic evaluation of machine translation[C]//Proceedings of Conference on Machine Translation: Volume 2, Shared Task Papers. 2016:511-517.

[9] Zhao S, Zhou M, Liu T. Learning question paraphrases for QA from encarta logs[C]//Proceedings of IJCAI,2007:1796-1800.

[10] Barzilay R, McKeown K. Extracting paraphrases from a parallel corpus[C]//Proceedings of Meeting of the Association for Computational Linguistics, 2001: 50-57.

[11] 李维刚. 中文复述实例与复述模板抽取技术研究[D]. 哈尔滨：哈尔滨工业大学博士学位论文，2008：1-139.

[12] Lin D, Pantel P. Discovery of inference rules for question answering[J]. Natural Language Engineering,2001, 7(4): 343-360.

[13] Yusuke Shinyama, Satoshi Sekine. Paraphrase acquisition for information extraction[J]. Spinal Cord, 2003, 52(4):264-267.

[14] Biran O, Blevins T, Mckeown K. Mining paraphrasal typed templates from a plain text corpus[C]//Proceedings of Meeting of the Association for Computational Linguistics, 2016:1913-1923.

[15] Grycner, Adam and Weikum, Gerhard and Pujara, Jay and Foulds, James and Getoor, Lise, RELLY: Inferring hypernym relationships between relational phrases[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 971-981.

[16] Takase S, Okazaki N, Inui K. Composing distributed representations of relational patterns[C]//Proceedings of Meeting of the Association for Computational Linguistics, 2016: 2276-2286.

[17] Takase S, Okazaki N, Inui K. Modeling semantic compositionality of relational patterns[J]. Engineering Applications of Artificial Intelligence, 2016(50): 256-264.

[18] Harris Z. Distributional structure[J]. Word,1954,10 (23): 146-162.

[19] Firth J R. A synopsis of linguistic theory 1930-1955. Studies in Linguistic Analysis[M]. Oxford: Philological Society,1957: 1-32.

[20] 李维刚，刘挺，李生. 基于网络挖掘的实体关系元组自动获取[J]. 电子学报，2007，35(11):2111-2116.

[21] 王千，王成，冯振元，等. K-means 聚类算法研究综述[J]. 电子设计工程，2012，20(7)：21-24.

刘明童(1993—)，博士研究生，主要研究领域为自然语言处理、神经机器翻译、复述。
E-mail：16112075@bjtu.edu.cn

徐金安(1970—)，副教授，主要研究领域为自然语言处理和机器翻译。
E-mail：jaxu@bjtu.edu.cn

张玉洁(1961—)，通信作者，教授，主要研究领域为自然语言处理和机器翻译。
E-mail：yjzhang@bjtu.edu.cn

(上接第 93 页)

[18] Gorman K，OpenFst Library，http://www.openfst.org/twiki/bin/view/FST/WebHome，2017-07-05/2017-10-18.

[19] Allauzen C，Riley M，Schalkwyk J，et al. OpenFst：A general and efficient weighted finite-state transducer library[C]//Proceedings of International Conference on Implementation and Application of Automata. Springer-Verlag，2007：11-23.

[20] Povey D，Hannemann M，Boulianne G，et al. Generating exact lattices in the WFST framework[C]//Proceedings of IEEE International Conference on Acoustics，Speech and Signal Processing. IEEE，2012：4213-4216.

[21] Novak J R，Dixon P R，Minematsu N，et al. Improving WFST-based G2P Conversion with Alignment-Constraints and RNNLM N-best Rescoring[J]. Booklist，2013.

[22] 信德麟，张会森，华劭.《俄语语法》(第二版)[M]. 北京：外语教学与研究出版社，2009.

[23] Ронжин А.，Карпов А.，Лобанов Б.，Et al. Фонетико-морфологическая разметка речевых корпусов для распознавания и синтеза русской речи[J]. Информационно-управляющие системы，2006，(6)：24-35.

[24] Важенина Д. А.，Кипяткова И. С.，Марков К. П.，et al. Методика выбора фонемного набора для автоматического распознавания русской речи[J]. Труды СПИИРАН，2014，5(36)：92-113.

冯伟(1993—)，硕士研究生，主要研究领域为自然语言处理。
E-mail：303203093@qq.com

马延周(1977—)，博士，主要研究领域为计算语言学和语言信息处理。
E-mail：myz827@126.com

易绵竹(1964—)，通信作者，教授，博士生导师，主要研究领域为计算语言学、语言信息处理。
E-mail：mianzhuyi@gmail.com