基于随机行走 N 步的汉语复述短语获取方法*

马军，张玉洁，徐金安，陈钰枫

（北京交通大学，北京 100044）

摘要：在利用大规模双语语料获取复述知识中，传统的基于“枢轴”方法只能考虑两步以内的复述现象。本文针对已有方法的局限性，对不同语言之间互为翻译的短语对，构建基于图的复述获取模型，提出基于随机行走 N 步的复述获取算法，改进已有方法以获取更多潜在的复述知识。本文描述了以汉英短语翻译表为基础的图模型、基于 N 步的随机行走算法和基于期望步数的复述短语可信度计算方法。同时，我们在图模型基础上提出基于多语言对扩展的方法。我们在 NTCIR 汉英、英日双语平行语料上进行了实验与评测，并与已有方法进行了对比。实验结果表明本文所提出的方法能够获取更多的复述知识，而且扩展语言对的图模型能够有效获取更多潜在的复述知识。

关键词：复述获取；随机行走；图模型

中图分类号：TP391　　**文献标识码：**A

Chinese Paraphrases Acquiring Based on Random Walk N Steps

Jun Ma, Yujie Zhang, Jinan Xu, Yufeng Chen

(Beijing Jiaotong University, Beijing, 100044, China)

Abstract: Conventional “pivot” approach of acquiring paraphrases from bilingual corpus has limitations, where only candidated paraphrases within two steps are considered. In this paper, we propose a graph based model of acquiring paraphrases from phrases translation table. First, we describe a graph model based on Chinese-English phrases translation table, a random walk algorithm based on N number of steps and a confidence metric for the obtained paraphrases phrases. Furthermore, we augment the model to be able to integrate more language pairs, for instance, English-Japanese phrases translation table aiming at finding more potential Chinese paraphrases. We performed experiments on NTCIR Chinese-English and English-Japanese bilingual corpora and compared with the conventional method. The experimental results show that the proposed model acquired more paraphrases, and the performance was improved further after English-Japanese phrases translation was added into the graph model.

Key words: Paraphrases acquisition; Random walk; Graph model

1 引言

复述（Paraphrases）是指具有相同语义的不同表达[1]，是自然语言中的普遍现象，体现了语言的多样性与复杂性。近年，复述处理在自然语言处理领域日益受到关注，在机器翻译（Machine Translation: MT）[2]、自动文摘（Automatic Summarization）[3],[4]、信息检索（Information Retrieval: IR）[5]、自然语言生成（Natural Language Generation: NLG）[6]和问答系统（Question Answering）[7]中都有重要应用。复述处理在上述应用中的核心问题是复述识别与复述生成，而复述识别与生成都以复述知识作为基础。相比自然语言处理中已经积累的大规模语言资源，譬如句法标注语料库、翻译词典、平行语料库，复述可谓资源匮乏。特别是在汉语方面，复述处理探索起步较晚，复述资源构建方法的研究具有特别重要的意义。

复述资源包括复述实例（互为复述的句对）、复述短语、复述词汇。其中复述短语与复述实例相比，在复述处理中有更高的利用率，而与复述词汇相比更难以获取。所以，复述短语获取一直是研究的焦点。传统基于“枢轴”获取复述知识的方法通过将对应同一外文翻译的短语视为复述短语，比如“自行车”、“单车”通过英语“bicycle” 建立复述关系。本文

* **收稿日期：**　　**定稿日期：**

基金项目：北京交通大学人才基金（No.KKRC11001532）；国家自然科学基金（No.61370130，61473294）；中央高校基本科研业务费专项资金（No.2014RC040，2015JBM033）

针对传统“枢轴”获取复述方法的局限性，提出了改进方案，包括以下三个方面：
（1）构建基于短语翻译表的图模型，在图模型上获取复述短语，通过考虑超过两步的复述短语，获取更多的复述；
（2）设计实现基于图的随机行走算法，从开始节点（给定短语）出发，在 N 步范围内随机行走；
（3）提出基于期望步数的复述短语可信度计算方法，利用期望步数量化并衡量候选复述短语与给定短语之间互为复述的可能性。
本文的内容组织如下：第二节介绍相关工作；第三节详细阐述基于图模型的复述获取方法；第四节论述基于多语言对扩展的图模型；第五节给出实验的详细过程、实验结果和分析；第六节对工作进行总结。

2 相关工作

复述知识的获取已经积累了大量的研究成果，根据所依据的语料性质可以分为以下四类[8]：（1）从单语语料获取，譬如大规模网络数据，代表方法是基于分布假设的复述获取方法；（2）从单语可比语料获取，譬如不同新闻网站针对同一事件的报道，代表性方法是基于“锚点”的复述获取方法；（3）从双语平行语料获取，譬如英法平行语料，主要是基于“枢轴”的方法；（4）从单语平行语料获取，譬如外文文学作品不同翻译版本或者机器翻译评测用的参考译文，代表性方法有基于机器翻译技术的方法。

根据语料来源，这些方法有各自特点，也都有一定的局限性。单语平行语料即复述实例，作为复述资源本来就匮乏，利用它获取复述短语不切实际；从单语语料或单语可比语料获取复述短语，需要借助语境计算，但是现有的方法会产生过多的其上下文相似的短语对，而非语义相等的短语对。双语平行语料中的句对在语义上相等，为获取复述短语提供了优质素材。因此，各种语言对之间的大规模平行语料成为获取复述短语首先考虑的语料。基于双语平行语料的复述短语获取方法也成为关键技术。

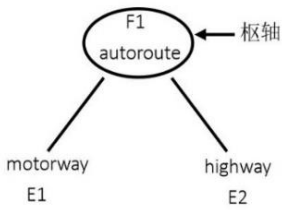


图 1 基于“枢轴”方法获取复述

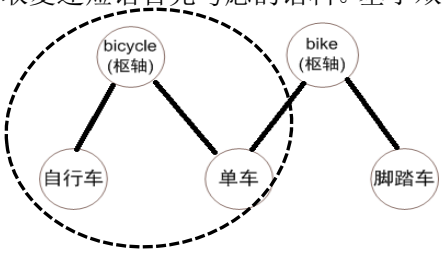


图 2 基于随机行走 N 步的复述获取原理

基于“枢轴”的大规模双语平行语料获取复述短语的方由 Bannard 和 CallisonBurch[9] 提出，基本想法是拥有同样翻译的两个短语互为复述，图 1 显示了这个关系。图中的英语单词“motorway”与“highway”有共同的法语翻译“autoroute”，那么利用法语“autoroute”作为“枢轴”可以获取英文的复述短语对 “motorway”和“highway”。基于“枢轴”方法的研究集中在英文复述获取方面，以利用英文以外的其他语言作为“枢轴”。但是，该方法只考虑了以同一个外语短语连接的英文短语之间的可能性，限制了获取更多复述短语的可能性。

实际上，如果考虑所有短语之间的翻译关系并以图的形式表示出来，会发现很多潜在的复述短语关系。图 2 显示了汉英短语之间翻译关系的一个例子， 汉语短语“自行车”、“单车”和“脚踏车”通过英语短语“bicycle”和“bike”连接起来。传统的基于枢轴的方法，只能从“自行车”经“bicycle”到“单车”这两步获取“单车”为复述短语，而无法获取到“脚踏车”以及距离更远的可能存在的复述短语。本文针对“枢轴法”这种只考虑两步的局限性，提出随机行走 N 步的复述获取方法。

3 基于图模型的复述获取方法

本节以从汉英平行语料获取汉语复述短语为例，详细描述基于图模型的复述获取方法和

实现细节。本文的方法也适用于其他语言复述短语的获取。

短语翻译表（以下简称“短语表”）是利用双语平行语料库，采用词对齐技术抽取的带有概率的短语对。首先利用汉英平行语料库抽取短语表，然后基于短语表构建包含汉英短语节点的图模型，并提出基于图的随机行走的方法，基于期望步数的复述短语可信度计算方法获取复述短语。

3.1 基于短语表的图模型构建

首先，我们给出一个汉英短语表中的例子，如表 1 所示。汉语短语“充当 阳极”到其所对应的翻译短语“acts as an anode”的翻译概率为 0.333333，反方向的翻译概率为 0.25。可以看到同一个汉语短语可对应多个英语短语，如表 1 中的“用作 阳极”；同样，同一个英语短语也可有多个汉语短语与之对应，如表 1 中的“acts as an anode”。短语之间拥有多个翻译关系，成为我们建立随机行走图模型的基础，通过将这些不同的汉英短语之间的翻译关系用图的方式关联起来，从而发现潜在的复述短语。

表 1 翻译短语表示例

汉语	英语	概率	概率
		汉-英	英-汉
充当 阳极	acts as an anode	0.333333	0.25
用作 阳极	acts as an anode	0.333333	1
用作 阳极	serving as an anode	0.5	0.0416
作为 阳极	serving as an anode	1	0.0313

我们构建的有向图包括节点集合 V 和边集合 E。

（1）用节点表示短语，所有汉语和英语短语构成节点集合；

（2）图中的有向边代表短语之间的翻译关系，如果节点 i 和节点 j 所对应的短语在短语表中有翻译关系，就有边（i,j）属于集合 E，同样也存在边（j,i）属于 E 集；

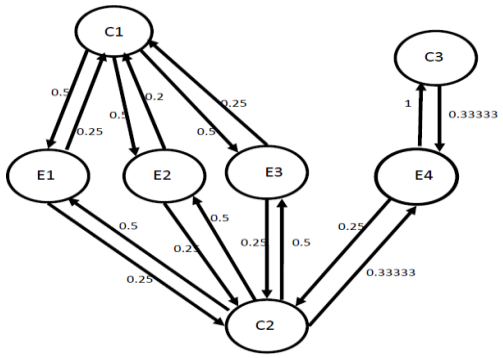


图 3 基于短语表的图模型

如此，构建出基于短语表的图模型。图模型的信息可以用大小为|V| × |V|的矩阵 W 表示，矩阵中的元素w_{ij}表示边（i,j）的权重。特殊地，权重为 0 表示不存在翻译关系[9]。

假设如此构建的图模型如图 3 所示，我们以此为例说明复述获取原理。如果我们要获取汉语 C1 的复述短语，那么从 C1 节点指出的有向边有 C1-E1、C1-E2、C1-E3，从 E1、E2、E3、E4 四个英语节点又都有边指向 C1、C2、C3，由此发现 C1 的复述短语不仅可能是 C2，也可能是 C3。本论文中，我们以英文短语为枢轴获取汉语短语的复述短语。

3.2 基于随机行走的复述获取

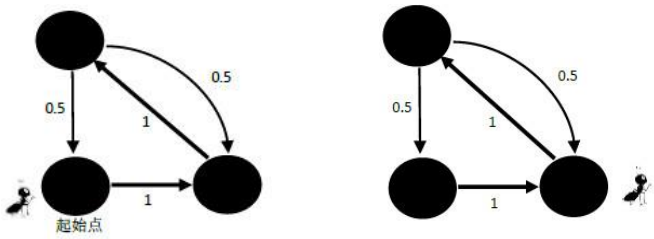
下面我们描述在基于短语表的图模型上获取复述短语的方法。在图 3 的图模型中，假设从汉语短语 C1 出发，经过一条外指的有向边行走到英语短语 E1，我们称为一步行走，继续从英语短语 E1 经过一条外指的有向边行走到汉语短语 C2，我们称为两步行走。因为 C1 和 C2 共同拥有 E1，所以 C1 和 C2 可被看作互为复述。由此可以看出传统的基于枢轴的方法就

是基于图的两步行走来获取复述短语。

但是，传统的方法从 C1 只行走两步，只能找到 C2 作为复述短语。实际上，C3 也是潜在的复述短语，因为如果从 C2 继续行走到 E4，再从 E4 行走到 C3，这样就可获取到另一个复述短语 C3，从 C1 出发共行走了四步。由此，我们发现通过两步以上的行走，可以寻找更多潜在的复述短语。但是，我们不可能穷尽式地搜索，为解决这个问题我们设计了基于 N 步随机行走的复述获取方法。

3.3 限定 N 步的随机行走

假设图模型上有蚂蚁沿着边从一个节点行走到另一个节点。图上的随机行走是指一只蚂蚁从给定的起始点出发，随机地选择一个邻居节点，行走到邻居节点上，然后把当前节点作为起始点，重复上述过程，直至随机行走的步数达到限定步数 N，就结束随机行走的过程[10]。



（a）起始点出发　　（b）到达第二个节点后选择下一目标

图 4 显示了蚂蚁的随机行走过程。假设有 M 只蚂蚁，从同一起始点开始随机行走，每一只蚂蚁的行走过程是独立的，不受其他蚂蚁的影响。一定步数之后，到达某一节点的蚂蚁个数越多，说明该节点与起始点的相关性越大。

蚂蚁随机选择路径的原则是更倾向于选择权重较大的边作为行走的路径。在基于短语表的图模型中，汉语短语与英语短语以翻译概率的边作为连接，表示两个短语在语义上相近，概率越大，语义越相近。如此随机行走一定步数后，如果更多的蚂蚁到达某个节点，表示该节点和起始节点在语义上越相近；如果同为汉语短语，表示互为复述可能性越大。

3.4 基于 N 步的随机行走算法

通过在图模型上的随机行走这种启发式搜索，可以帮助我们寻找潜在的复述短语，避免穷尽式的搜索。随机行走需要限定在一定步数之内，称为最大步数，用 N 表示。为了实现 M 只蚂蚁在图模型上随机行走，我们需要解决两个问题，一个是记录行走状态，另一个是路径选择。下面，我们描述限定步　数的随机行走算法，流程图如图 5 所示。

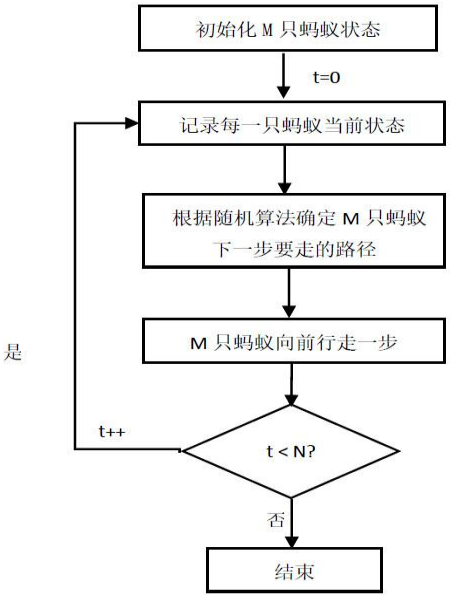


图 5 基于 N 步的随机行走算法流程图

（1）初始化 M 只蚂蚁状态：每只蚂蚁处于起始点，行走步数 t = 0；
（2）记录每只蚂蚁当前状态：包括蚂蚁当前所处节点，蚂蚁是否是第一次到达该节点，如果是第一次到达，则做标记并记录蚂蚁行走的步数；
（3）根据随机算法确定下一步要走的路径：找出蚂蚁所处节点外指的所有边，对所有边上的权重求和；根据每条边的权重大小按比例划分区间，标上对应边的标号；然后产生一个随机数，判断随机数落在哪个区间，就将所对应的边作为下一步要走的路径；
（4）M 只蚂蚁向前行走一步：根据上一步骤确定的行走路径，每只蚂蚁向前行走一步，t = t+1；
（5）判断每只蚂蚁是否已经行走了最大步数 N。
本文采用 C++实现了基于 N 步的随机行走算法，算法的伪代码描述如下：

```
Algorithm Random_Walk(Parameter    Graph,
AntNumber, MaxStep)
  Initialize AntState
    for t = 1 to MaxStep do
      for j = 1 to AntNumber do
            RandomDetermineTheNextPath(j,Graph)
            GoForwardOneStep
            Record（AntState）
        end for
    end for
return (AntState)
```

3.5 基于期望步数的复述可信度计算

基于 N 步行走之后，M 只蚂蚁所到达节点与起始节点之间的关联性到底有多大，决定节点分别代表的短语互为复述的可能性大小。为此，我们提出基于期望步数的复述短语可信度计算方法。

在基于随机行走的搜索算法结束后，需要对搜索结果排序，期望步数是排序的依据，其定义 $\hat{h}_{ij}^{N}$描述如下：在限定最大随机行走步数为 N 的情况下，M 只蚂蚁从节点 i 开始经过随机行走，第一次抵达节点 j 的平均行走步数，公式（1）给出了期望步数的计算方法。M 只蚂蚁随机行走过程结束之后，若有 m 只到达节点 j，则记录 m 只蚂蚁第一次到达 j 的行走步数；对于（M-m）只没有到达节点 j 的蚂蚁，认定其行走步数为 N。

$$\hat{h}_{ij}^{N} = \frac{\sum_{k=1}^{m} t_j^k + (M-m)N}{M} \qquad (1)$$

$\hat{h}_{ij}^{N}$：起始点 i 和节点 j 之间的期望步数
M：蚂蚁总数
N：行走最大步数
m：到达节点 j 的蚂蚁总数
t_j^k：蚂蚁 k 第一次到达节点 j 所行走的步数

其中 M 必须满足的约束条件 $\frac{1}{2\epsilon^2}\log(\frac{2n}{\delta})$ $M \geq n$是图中节点的个数，δ和ε是设定的调节蚂蚁个数 M 的参数，满足条件 $0 \leq \delta, \varepsilon \leq 1$。

如果到达节点 j 的蚂蚁越多，即 m 越大，同时到达该节点的每只蚂蚁的行走步数t_j^k越小，则 $\hat{h}_{ij}^{N}$的值越小，该节点和起始节点在图中的距离越接近。在我们获取复述短语的任务中，意味着两个节点所代表的短语在语义上的距离越小，互为复述的可信度越高。

4 基于多语言对扩展图模型

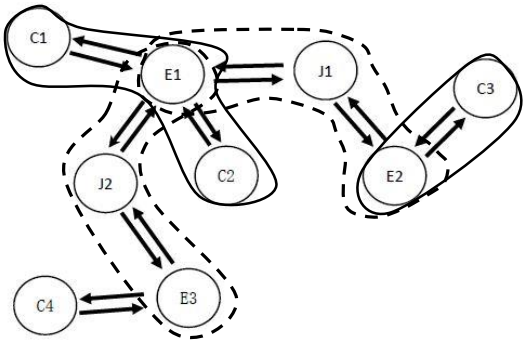


图 6 汉英、英日扩展图模型

在基于短语表的图模型中，同一语言的两个短语通过与另一语言的翻译关系建立了复述联系，如果引入多语言对的翻译关系，将会使更多短语之间建立联系，从而发现更多复述。为此，我们考虑了多语言对短语表。为了表示并计算多语言短语间翻译关系，我们对图模型进行了扩展，使之可以加入多语言短语表。通过对第 3 节构建图模型的算法进行修改，使构建的图模型中可同时引入多语言对翻译关系，具体包括：

（1）在图模型中增加了表示语言种类的标记，使得增加语言的短语可以同原有的语言一样作为节点存储和处理；

（2）针对多种不同语言的短语，采用基于广度优先的方式构建图模型；在扩展后的图模型中，仍然利用前面介绍的随机行走算法通过对图进行遍历，获取复述短语。

我们以汉英短语表加入英日短语表为例，介绍基于多语言对的扩展模型。如此扩展的图模型如图 6 所示。从图中可以看出，由实线包围的子图是基于汉英短语表构建的联系，是三个彼此独立的子图，之间没有联系。当引入英日短语表后，建立了虚线包围的子图，从而联结了原来孤立的三个汉英子图，最终建立了 C1 与 C3、C4 之间的复述联系。

5 实验

上面我们提出了基于图模型获取复述短语的方法，以及利用期望步数实现对候选复述短语可信度的计算，并对图模型进一步扩展，加入了英日短语翻译表。下面，我们通过实验验证所提方法的有效性。

5.1 实验数据

本文使用的平行语料来自 NTCIR 汉英平行语料 100 万句对和英日平行语料 300 万句对。首先我们采用 GIZA++进行单词对齐，然后使用 grow-diag-final 启发式算法优化单词对齐结果，抽取汉英、英日短语表；然后，对噪音数据进行过滤，具体过滤规则如下：

（1）过滤掉含有特殊字符的短语对及含有停用词的短语对；

（2）过滤掉以翻译概率排序在 15 位以后的低翻译概率短语对；

最终，经过过滤后的短语表包括 21610194 条汉英短语对、34465517 条英日短语对。在图模型构建的过程中，如果将所有的短语对载入内存，将非常耗时。为此我们提出以下解决方案：

（1）只需考虑与给定短语有翻译关系的短语，即从给定短语出发经过翻译关系扩展的短语将被载入内存，以此为基础构建图模型；

（2）对大规模的短语经过排序之后建立短语对的索引，根据索引将与给定短语有翻译关系的短语载入内存。

5.2 实验参数设定

根据公式（1）的条件限制，同时为保证实验的可行性，实验结果的正确性。本次实验的参数设定如下。

（1）在初期试验中，我们发现在建立图的过程中，随着节点的不断扩展，从给定短语节点

依照翻译关系依次向不同分支扩展得到的图中，会引入过多的噪音数据，从而影响获取复述短语的效果。所以，为了有效获取潜在的复述短语，在建立图的过程中避免引入过多的干扰数据，同时确保实验内数据的可计算性，设定与给定短语节点最长距离，即图的层数 d=8；

（2）图中包含的最大节点数目 n = 50000;

（3）随机行走过程中蚂蚁的个数 M = 1000000；

（4）最大随机行走步数 N = 12 （$N \geq d$）；

（5）δ = 0.05，$\varepsilon \leq 0.03$。

5.3 输出结果过滤

在输出结果中我们发现了以下两种情况：（1）输出短语和给定短语互相包含，例如对于给定短语“含药物”，输出短语为“含药物的”；（2）多个输出短语互相包含，例如对于测试短语“充当阳极”，输出短语为“作为阳极”、“作为阳极的”。我们认为这两种情况输出的短语只是后缀变化改变词性的短语不是真正意义上的复述短语。因此根据下面规则对其进行过滤。针对情况（1）不会将其作为复述短语输出；情况（2）只保留和测试短语长度最接近的短语。

5.4 实验结果

为了验证本文提出的方法以及加入英日短语表扩展后的图模型获取复述短语的效果，我们从汉英短语表中随机选择 100 条汉语短语作为测试数据（部分测试短语如表 2 所示）。

表 2 部分测试汉语短语例子

编号	测试短语	编号	测试短语
1	传送 给 远程	6	不断 提高
2	船体	7	家庭 用品
3	微不足道	8	船体
4	发送 图像	9	我们 期望
5	医药 材料	10	专属

对于每一条测试短语，因为根据翻译表构建的图非常复杂（翻译表保留概率排序 15 位以内的短语对，图的层数为 8），所以我们无法得知正确的复述短语的个数，无法通过召回率和 F-值对结果进行评测。对此，我们采取下面评测方式。我们设置最大随机行走的步数分别为 2 步（枢轴法）、N 步、加入英日短语表后 N 步，这三种情况下分别获取的候选复述短语，然后人工（1 人）判断并统计其中包含的正确复述短语个数，最后计算复述短语的准确率。计算结果如表 3 所示。

表 3 实验评价结果

行走步数 N	平均候选复述短语个数	平均准确复述短语	平均准确率
N = 2（枢轴法）	3.79	2.88	75.99%
N = 12	11.56	7.18	62.11%
N = 12（英日短语表）	15.32	9.04	59.01%

从表 3 的结果可以看出本文提出的基于随机行走 N 步复述获取方法能够较传统的“枢轴法”获取到更多的复述短语；而且，本文的方法在加入英日短语表进行扩展之后，获取到更多复述。同时我们看到，图模型构建过程中会伴随着节点扩展引入一些噪音数据，导致准确率下降。在我们发现复述的任务中，更偏重所发现复述数量的增加。

表 4 给出了基于本文提出的方法获取的部分复述短语结果以及相应的期望步数。由前

文所述，本文使用期望步数对复述短语可信度进行评价，期望步数越小，表明该短语成为给定短语的复述可能性越大。在表 4 中，对于给定的测试短语，输出的候选复述短语中，其期望步数越小排位越靠前。我们观察发现，排在前面的结果比排在后面的结果，在语义上更接近测试短语。由此，我们得出结论，用期望步数作为候选短语复述可信度的评价依据是合理有效的，期望步数越小，对应候选短语成为复述短语的可能性的确越大。第 3 节提出的基于短语表构建图模型的算法、基于期望步数计算复述短语可信度的方法同样适用于扩展模型中复述的获取。

表 4 部分复述短语实例及期望步数

编号	测试短语	复述短语	期望步数	编号	测试短语	复述短语	期望步数
1	传送 给 远程	发送 给 远端	9.44202	4	不断 提高	继续 增大	11.2896
		发送 到 远端	10.1433			连续 升高	11.3157
		传送 到 远处	10.6529			日益 提高	11.8252
		传输 到 远端	10.8738			然后 增加	11.8338
		传送 到 远端	11.3424			连续 增加	11.8834
		传输 给 远端	11.452			以及 提高	11.9046
		朝向 该 远程	11.8415			继续 增加	11.9507
		传输 到 远程	11.8629			相继 增加	11.9507
		传递 到 远端	11.9333			日益 增多	11.9636
		传递 给 远程	11.9999			然后 增大	11.9724
		运输 至 遥远	11.9999			连续 增大	11.9935
						持续 上升	11.9943
						持续 增长	11.9947
						然后 升高	11.9964
						不断 增大	11.9970
						继续 上升	11.9985
						继续 升高	11.9999
2	船体	船壳	11.8173	5	医药 材料	医疗 器具	9.59756
		船身	11.9981			医学 材料	9.62103
		船侧	11.9999			医学 物质	9.87545
		轮船	11.9999			医疗 材料	11.5707
3	微不足道	无足轻重	11.9998	6	发送 图像	传输 图像	11.3969
		无关紧要	11.9998			图像 传输	11.9899
		忽略不计	11.9999			传送 图像	11.9999

我们对人工判断为错误的结果进行了分析。通过观察表 4 中编号为 6 的测试短语“不断 提高”的候选复述短语，我们发现排在第二位和第四位的结果 “然后 增加”、“以及 提高”并不是“不断 提高”的复述短语，但是得到该结果的期望步数较小并被排在靠前的位置。经过分析我们认为，在使用期望步数与计算候选复述短语可信度时，期望步数与短语表中的翻译概率有间接关系。因此，对齐结果中翻译概率值估算有偏差的结果对其有影响，导致个别非复述短语的期望步数较小，排位比较靠前。

图 7 是我们加入英日短语后扩展模型获取的实例。基于汉英翻译关系“匹配”的子图与“对准”、“校正”的子图用实线包围，英日语短语“整合”的子图用虚线包围。两个实线子图通过虚线子图连接建立了联系，这样就能够获取汉语复述短语对“匹配”、“对准”和“校

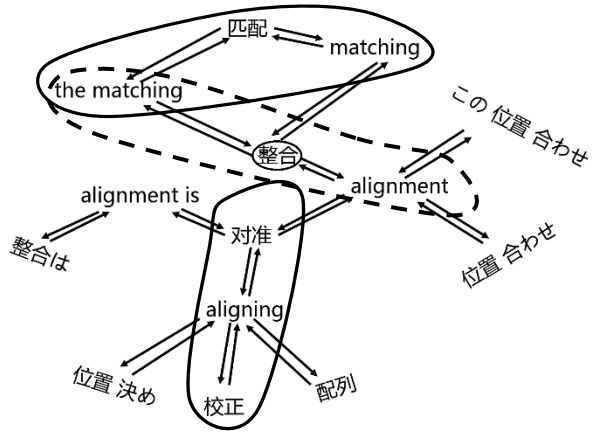


图 7 加入英日短语获取复述实例

正”。如果只利用汉英短语表构建图模型，汉语短语“匹配”无法获取到潜在的复述短语“对准”、“校正”。通过扩展多语言对即英日短语表，日语短语“整合”与两个汉英子图的“the matching”、“matching”、“alignment”有翻译关系，从而连接了两个子图，汉语短语“匹配”就可以获取到复述短语“对准”和“校正”。

6 结论及未来工作

针对传统的基于枢轴获取复述知识方法的局限性，本文提出了图模型的基于随机行走的获取复述方法，和基于期望步数的复述短语可信度计算方法，进一步提出基于多语言对的扩展图模型。我们在 NTCIR 汉英、英日平行语料上进行了实验评测，并与已有方法进行了对比，实验结果表明本文所提方法能够有效提高获取复述短语的数量。

今后，我们准备在本文工作的基础上，变换实验参数，扩大实验，特别是探索随机行走步数 N 的不同取值，对输出候选复述短语的数量及准确率的影响。

参考文献

[1] R. Barzilay and K. R. McKeown. Extracting Paraphrases from a Parallel Corpus. Procee-dings of ACL/EACL. 2001:50-57.

[2] Callison-Burch C,Koehn P’Osborne M.Improved statistical machine translation using paraphrases[C].In:Proc.Of the HLT-NAACL Morristown:Association for Computational Linguistics.2006:1 7—24.

[3] R.Barzilay.Information Fusion for Multi-document Summarization: Paraphrasing and Generation. Ph.D. thesis, Columbia University. 2003.

[4] Zhou L, Lin CY Munteanu DS, Hovy E. ParaEval: Using paraphrases to vealuate summaries automatically[C]. In: Proc. of the HLT-NAACL. Morristown: Association for Computational Linguistics. 2006: 447-454.

[5] Zukerman I, Raskutti B. Lexical query paraphrasing for document retrieval[C]. In: Proc of COLING. Morristown: Association for Computational Linguistics. 2002: 1-7.

[6] Iordanskaja L, Kittredge R Polguere A. Lexical selection and paraphrase in a meaning—text generation model[C]. In: Paris CL, Swartout WR,Mann WC,eds. Natural Language Generation in Artificial Intelligence and Computational Linguistics. 1991: 293-3 1 2.

[7] McKeown KR. Paraphrasing using given and new information in a question-answer system[C]. In: Proc. of the ACL. Morristown: Association for Computational Linguistics. 1979: 67-72.

[8] 赵世奇.基于统计的复述获取与生成技术研究[学位论文].哈尔滨.哈尔滨工业大学.2009.1-9

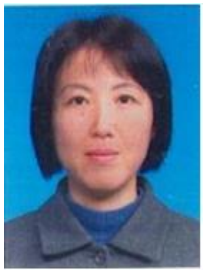
[9] Stanley Kok, Chris Brockett. Hitting the Right Paraphrases in Good Time. ACM.2010. 45-153.

[10] 徐晓华. 图上的随机游走学习[学位论文].南京.南京航空航天大学.2008.

作者简介：马军（1991——），男，在读研究生，主要研究领域为自然语言处理。Email:void_sky@163.com；张玉洁（1961——），女，教授，主要研究领域为自然语言处理、机器翻译，通讯作者。Email: yjzhang@bjtu.edu.cn；徐金安（1970——），男，副教授，主要研究领域为自然语言处理和机器翻译。Email: jaxu@bjtu.edu.cn；陈钰枫（1981——），女，副教授， 主要研究领域为自然语言处理

和机器翻译。Email: chenyf@bjtu.edu.cn。

马军



张玉洁

徐金安

陈钰枫