中国科学: 信息科学 2017 年 第 47 卷 第 8 期: 1066–1077
SCIENTIA SINICA Informationis
从大数据到大知识工程专刊 · 论文

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

北京交通大学计算机与信息技术学院, 北京 100044
* 通信作者. E-mail: yjzhang@bjtu.edu.cn

收稿日期: 2017–04–05; 接受日期: 2017–05–26; 网络出版日期: 2017–08–14
北京交通大学人才基金 (批准号: KKRC11001532)、国家自然科学基金 (批准号: 61370130, 61473294) 和中央高校基本科研业务费专项资金 (批准号: 2015JBM033) 资助项目

摘要 在利用大规模双语语料获取复述知识方面, 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翻译关系图, 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. 本文描述了由汉英短语翻译表构建翻译关系图的方法、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, 本文提出面向多语言对的翻译关系图扩展方法. 在 NTCIR 汉英和英日双语平行语料上进行了实验与评测, 并与传统方法进行了对比. 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, 而且扩展语言对的翻译关系图能够有效获取更多潜在的复述知识.

关键词 复述获取, 短语翻译表, 翻译关系图, 随机行走, 期望步数

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. 近年来, 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, QA) [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, 对复述处理的探索起步较晚, 因而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

引用格式: 马军, 张玉洁, 徐金安, 等. 基于随机行走 N 步的汉语复述短语获取方法. 中国科学: 信息科学, 2017, 47: 1066–1077, doi: 10.1360/N112016-00303
Ma J, Zhang Y J, Xu J A, et al. Acquiring Chinese paraphrases based on random walk of N steps (in Chinese). Sci Sin Inform, 2017, 47: 1066–1077, doi: 10.1360/N112016-00303

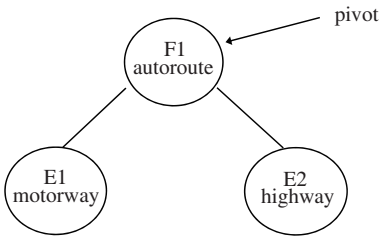


图 1 基于“枢轴”方法的复述获取

Figure 1 Paraphrases acquisition based on “pivot”

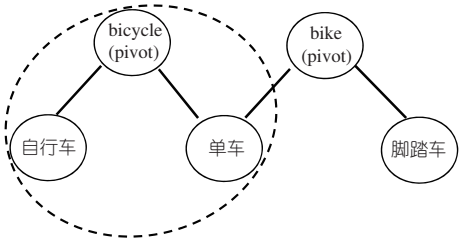


图 2 基于翻译关系图的复述获取

Figure 2 Paraphrases acquisition based on translation relation graph

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” 和 “单车” 的英语翻译都是 “bicycle”, 从而互为复述. 在随后的叙述中除非说明, 复述指复述短语. 本文针对传统的基于 “枢轴” 复述获取方法的局限性, 提出了改进方案, 包括以下 3 个方面:

(1) 由短语翻译表构建翻译关系图, 在翻译关系图上获取复述短语, 通过考虑超过两步的短语, 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 步范围内行走;

(3) 提出基于期望步数的复述可信度计算方法, 利用期望步数量化并衡量复述结果与给定短语之间互为复述的可能性.

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 4 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.

根据语料性质, 这些方法各有特点, 也有一定局限性. 单语平行语料即复述实例非常稀少, 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, 需要借助语境计算, 但是现有方法会产生过多的语境相似而语义相反的短语对. 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首先考虑的语料. 基于双语平行语料的复述获取方法也成为关键技术.

基于 “枢轴” 的大规模双语平行语料获取复述的方法由 Bannard 和 Callison-Burch [9] 提出, 基本想法是拥有同样翻译的两个短语互为复述, 如图 1 所示. 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” 可以获取英语的复述对 “motorway” 与 “highway”. 基于 “枢轴” 方法的研究集中在英语方面, 利用英语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语短语之间的复述关系, 限制了获取更多复述的可能性.

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关系. 图 2 显示了用图表示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” 分别连接起来. 在传统的基于 “枢轴” 方法中, 只能从 “自行车” 经

表 1　短语翻译表的例子

Table 1　Examples of the Chinese-English phrases translation table

Example	Chinese phrase	English phrase	Chinese-English translation probability	English-Chinese translation probability
1	充当阳极	acts as an anode	0.333333	0.25
2	用作阳极	acts as an anode	0.333333	1
3	用作阳极	serving as an anode	0.5	0.0416
4	作为阳极	serving as an anode	1	0.0313

“bicycle” 到 “单车” 这两步获取 “单车” 为复述, 而无法获取到 “脚踏车” 以及距离更远的可能存在的复述. 本文针对传统 “枢轴” 法中这种只考虑两步的局限性, 提出构建翻译关系图利用随机行走算法获取复述的方法.

3　基于翻译关系图的复述获取方法

本节以从汉英平行语料获取汉语复述为例, 详细描述基于翻译关系图的复述获取方法和实现细节. 该方法也适用于其他语言. 首先, 利用单词对齐技术从汉英双语平行语料中抽取带有翻译概率的短语翻译表 (以下简称 “短语表”); 然后, 构建包含汉语和英语短语节点的翻译关系图, 在此技术上描述基于图的随机行走方法和基于期望步数的复述可信度计算方法.

3.1　翻译关系图的构建

首先, 我们给出几个汉英短语表中的例子, 如表 1 所示. 例 1 的汉语短语 “充当阳极” 到英语短语 “acts as an anode” 的翻译概率为 0.333333, 反方向的翻译概率为 0.25. 例 2 和 3 的汉语短语相同, 对应不同英语短语; 同样, 例 1 和 2 的英语短语相同, 对应不同汉语短语. 汉英短语间拥有多个翻译关系, 成为构建图并利用随机行走算法的基础, 通过将这些翻译关系以图的方式连接起来, 可以发现潜在的复述.

我们构建的翻译关系图是一个有向图, 包括如下定义的节点集合 V 和边集合 E.

(1) 节点表示短语, 所有汉语和英语短语构成节点集合;

(2) 图中的有向边表示短语之间的翻译关系, 如果节点 i 和 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.

如此, 构建出短语表的翻译关系图. 翻译关系图可用大小为 $|V| \times |V|$ 的矩阵 W 表示, 矩阵中元素 W_{ij} 表示边 (i, j) 对应的翻译概率. 特殊地, 值为 0 表示不存在翻译关系 [10].

假设如此构建的翻译关系图如图 3 所示, C1, C2, C3 表示汉语短语节点, E1, E2, E3, E4 表示英语短语节点, 我们要获取 C1 的复述, 下面以此为例说明复述获取原理. 图中从 C1 出发分别有边指向 E1, E2, E3, 它们又分别有边指向 C2, C2 又经 E4 与 C3 连接. 由此发现 C1 的复述不仅可能是 C2, 也可能是 C3. 本文, 以英文为 “枢轴” 获取汉语的复述.

3.2　基于翻译关系图的复述获取

下面描述在构建的翻译关系图上获取复述的方法. 在图 3 的翻译关系图中, 假设从 C1 出发, 经过一条外指的有向边行走到 E1, 称为一步行走; 继续从 E1 经过一条外指的有向边行走到 C2, 称为两

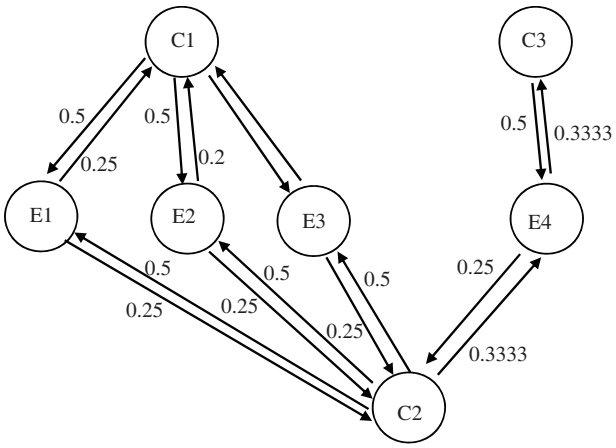


图 3 短语表的翻译关系图

Figure 3 Translation relation graph of phrases translation table

步行走. 因为 C1 和 C2 共同拥有 E1, 所以 C1 和 C2 可被视为复述. 由此可以看出传统的基于 “枢轴” 的方法就是图中两步行走的情况, 是基于翻译关系图获取复述的特例.

传统的方法从 C1 出发只行走两步, 只能找到 C2 作为复述. 实际上, C3 也是潜在的复述. 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 C3, 从 C1 出发共行走了 4 步. 由此发现通过两步以上的行走, 可以寻找更多潜在的复述. 基于图的复述获取方法本质上是通过对图的遍历, 搜索与给定短语连接的所有短语, 并从中寻找与其语义相近的短语作为复述. 如此, 可以将复述获取任务转换成图中相似节点的搜索问题. 同时, 为了避免穷尽式的搜索, 设计了基于 N 步随机行走算法以获取复述.

3.3 限定 N 步的随机行走

解决图中相似节点搜索问题的典型方法是基于随机行走的排序算法 [11]. Sarkar 等 [12] 在排序算法中融合取样技术与剪枝方法, 提出一种高效的图中节点间相似度的量化计算方法. 我们采用该方法实现基于随机行走的复述获取, 具体方案描述如下.

图上的随机行走是指给定一个图和一个出发点, 随机地选择一个邻居结点并移动到该结点, 然后把该结点作为新的出发点, 重复以上过程. 我们用一只蚂蚁的爬行形象化地表示一个随机行走过程, 多只蚂蚁代表多个随机行走过程. 从给定的出发点出发, 每只蚂蚁随机地选择一个邻居节点并行走到邻居节点上, 然后把所到邻居节点作为新的出发点, 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. 图 4 显示了蚂蚁的随机行走过程, 图 4(a) 表示蚂蚁位于起始点, 图 4(b) 表示经过一步随机行走之后到达邻居节点.

假设有 M 只蚂蚁, 从同一起始点开始随机行走, 每只蚂蚁的行走过程独立, 不受其他蚂蚁影响. 一定步数之后, 到达某一节点的蚂蚁个数越多, 说明该节点与出发点相关的概率越大. 蚂蚁随机选择路径的原则是倾向于选择权重较大的边. 在基于短语表的翻译关系图中, 连接短语节点之间的边以翻译概率作为权重, 表示两个短语在语义上相近, 而且概率越大, 语义越相近. 蚂蚁在翻译关系图上的行走, 意味着蚂蚁作为信使将出发点的语义信息传递到其他节点.

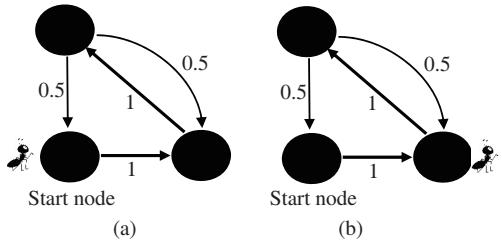


图 4 蚂蚁的随机行走

Figure 4 The random walk of an ant. (a) The ant at the start node; (b) after one step of random walk

3.4 N 步随机行走算法

通过图上基于随机行走算法的搜索, 可以帮助我们避免穷尽式搜索寻找潜在的复述. 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 只蚂蚁在图上的随机行走, 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, 描述 N 步随机行走算法.

(1) 初始化 M 只蚂蚁状态: 每只蚂蚁处于出发点, 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) 根据随机算法确定每只蚂蚁下一步要走的路径: 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, 若已行走最大步数 N, 结束; 否则跳转到第 (2) 步.

本文采用 C++ 实现了 N 步随机行走算法, 算法的伪代码如下, 算法的复杂度为 O $(M \times N)$, 其中, M 表示蚂蚁总数, N 表示随机行走最大步数.

算法 1 基于 N 步的随机行走算法

```
Input: Graph, M, N
Output: AntState
while t < N do
    j ⇐ 1;
    while j < M do
        RandomDetermineTheNextPath(j, Graph);
        GoForwardOneStep;
        Record(AntState);
        j ⇐ j + 1;
    end while
    t ⇐ t + 1;
end while
```

3.5 基于期望步数的复述可信度计算

基于 N 步随机行走之后, M 只蚂蚁所到达节点与出发点之间的相似度, 决定节点分别表示的短语互为复述的可能性. 为此, 本文提出基于期望步数的复述可信度计算方法.

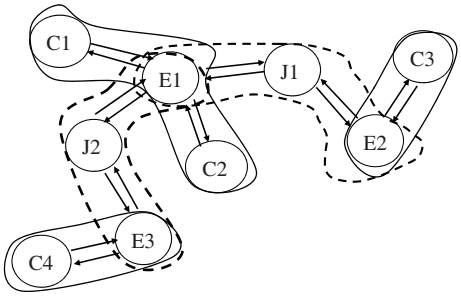


图 5 面向汉英和英日多语言对的翻译关系图

Figure 5 Translation relation graph of Chinese-English phrase translation table integrated with that of English-Japanese

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 $\hat{h}_{ij}^N$ 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, 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, 则对 m 只蚂蚁第一次到达 j 的行走步数求和; 对于 $M-m$ 只没有到达节点 j 的蚂蚁, 认定其行走步数为 N.

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

其中, $\hat{h}_{ij}^N$ 表示出发点 i 和节点 j 之间的期望步数; M 表示蚂蚁总数; N 表示随机行走最大步数; t_j^k 表示第 k 只蚂蚁第一次到达节点 j 所行走的步数; M 必须满足的约束条件是 $M \geqslant 1/2\epsilon^2 \log(\frac{2n}{\delta})$, n 是图中节点的个数, δ 和 ϵ 是预先设定的调节蚂蚁个数 M 的参数, 满足条件 $0 \leqslant \delta, \epsilon \leqslant 1$.

如果到达节点 j 的蚂蚁越多, 即 m 越大, 并且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, 那么该节点和出发点在图中的距离越近, 相似度越高. 在获取复述的任务中, 意味着两个节点所表示的短语在语义上距离越小, 互为复述的可信度越高. 蚂蚁在翻译关系图上的行走, 意味着蚂蚁作为信使将出发点的语义信息传递到其他节点. 假设 M 只蚂蚁作为信使从出发点 i 出发, 通过彼此独立的随机行走将出发点短语的信息向外传播. 随机行走一定步数后, 如果更多的蚂蚁到达某个节点, 意味着有更多信使携带出发点的信息到达该节点. 那么, 出发点的语义信息传递到该节点的概率越大, 从而该节点与出发点在语义上相近的概率越高. 另一方面, 蚂蚁从出发点出发沿着赋予翻译概率的边走到邻居节点, 通常翻译概率小于 1, 意味着蚂蚁每行走一步都有信息损失, 即出发点的语义信息只有一部分被传递到邻居节点. 如此, 随机行走步数越多意味着起始点的语义信息损失越多. 相反, 行走步数越少意味着出发点的语义信息保留得越多, 到达节点的语义与出发点的语义越相近, 成为复述的可能性越大.

4 面向多语言对的翻译关系图扩展

在图 3 构建的翻译关系图中, 两个汉语短语通过与同一个英语短语的翻译关系建立了复述联系, 但是这样的图中会出现很多孤立的子图. 为了解决这个问题, 本文进一步提出引入新的语言对扩展翻译关系图的方法. 我们以引入英日短语表为例进行说明, 扩展后的翻译关系图如图 5 所示. 从图 5 中可以看出, 由实线包围的子图是由汉英短语表构建的翻译关系连接, 形成了 3 个孤立的子图. 其中, C1 和 C2 有连接关系, 但是与 C3 和 C4 之间没有连接关系, C3 与 C4 也没有连接关系. 引入新的语言对

后, 借助日语短语连接起孤立子图中的英语短语, 进而连接了汉语短语. 图 5 中, J2 连接起 E1 和 E3, 从而 C1 和 C4 建立了连接关系. 从图 5 的示例中看到, 当引入英日短语表后建立了虚线包围的英日翻译关系子图, 进而连接了孤立的 3 个汉英翻译关系子图, 从而在 C1 与 C3 和 C4 之间分别建立起新的复述关系.

通过对第 3 节构建翻译关系图的算法进行修改, 使构建的翻译关系图中可同时引入多语言对的翻译关系, 具体包括:

(1) 在翻译关系图中增加表示语言种类的标记, 使得增加的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) 针对多种语言的短语, 采用基于广度优先的方式构建翻译关系图;

(3) 在扩展后的翻译关系图中, 仍然利用前面描述的随机行走算法, 对图进行搜索获取复述.

5 评测实验与结果分析

前面描述了基于翻译关系图获取复述的方法, 以及利用期望步数计算复述可信度的方法; 并对翻译关系图进一步扩展, 加入了英日短语表. 下面, 通过 3 个实验验证所提方法的有效性: 实验 1, 首先利用汉英短语表实现传统的基于 “枢轴” 方法作为基线系统获取复述; 实验 2, 利用汉英短语表构建翻译关系图, 利用限定 N 步的随机行走算法获取复述; 实验 3, 加入英日短语表, 在翻译关系扩展图上获取复述. 随后, 分析比较 3 个实验的评测结果.

5.1 实验数据

本文使用 NTCIR100 万句对的汉英平行语料和 300 万句对的英日平行语料. 首先采用 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, 抽取汉英和英日短语表, 并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以及含有停用词的短语对;

(2) 过滤掉以翻译概率排序的 15 位以后的低翻译概率短语对.

最终, 经过过滤后的汉英短语表包括 21610194 条短语对, 英日短语表包括 34465517 条短语对. 在翻译关系图构建的过程中, 如果将所有的短语对载入内存, 将非常耗时. 为此提出以下解决方案:

(1) 只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才被载入内存, 并以此为基础构建翻译关系图;

(2) 对大规模短语排序并建立索引, 根据索引快速找到与给定短语有翻译关系的短语并载入内存.

5.2 实验参数设定

考虑式 (1) 的条件限制, 同时为保证实验的可行性和实验结果的正确性, 本文实验的参数设定如下.

(1) 图的层数 d =8. 在初期试验中, 我们发现在图的构建中, 从给定短语节点出发依照翻译关系逐步向不同分支扩展过程中, 会逐渐引入噪音数据, 从而影响获取复述的质量, 为了避免噪音数据的干扰, 获取高质量的复述, 同时为了防止图的过于庞大影响计算的可行性, 在图的构建中, 设定与给定短语节点最长的距离, 即图的层数 d =8.

表 2 部分测试短语的例子

Table 2 Some phrases for experiment

Number	Given phrase	Number	Given phrase
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	充当阳极
4	发送图像	9	我们期望
5	医药材料	10	专属

(2) 图中包含的最大节点数目 $n = 50000$. 为了避免构建的翻译关系图过于庞大难以计算, 设置图的节点个数. 另一方面, 图的节点个数过小会无法支持翻译关系图的构建. 本文通过分别对汉英和英汉短语表中所有短语拥有的翻译短语的个数进行统计, 发现拥有翻译短语的平均个数分别为 1.7 和 1.6. 考虑到图的层数 d =8, 并且平均翻译短语个数小于 2, 故设置图的最大节点数为 50000.

(3) 随机行走的蚂蚁个数 $M = 1000000$. 蚂蚁的个数由式 (1) 的参数约束条件 $M \geqslant 1/2\epsilon^2 \log(\frac{2n}{\delta})$ 确定, 参考 Sarkar 的工作 [12], 将 δ 和 ϵ 的值分别设置为 $\delta = 0.05$, $\epsilon = 0.03$.

(4) 最大随机行走步数 $N = 12$. 考虑到图的层数 $d = 8$, 在初期实验中, 分别考察了小于、等于和大于 d 的随机行走步数, 分别为 $N = 4$, 8 以及 12. 实验结果表明, $N = 4$ 时获取的复述数量最少; $N = 8$ 时获取的复述数量多于 $N = 4$ 的情况; $N = 12$ 时获取的复述数量最多. 经过分析, 我们认为当 $N = 4$ 时蚂蚁实际上无法走到图中最远的那些节点, 故获取的复述有限; N =8 时有可能走到那些最远的节点, 所以数量多于 $N = 4$ 的情况; 相比于 $N = 8$, $N = 12$ 时有可能走到更多最远的节点, 所以获取的复述数量最多. 我们也调查了 N 大于 12 的情况, 发现复述数量没有明显增加, 故将最大随机行走步数设为 12. 由于下一步路径既可能向前也可能后退, 需要将随机行走步数设置为 $N \geqslant d$, 以保证随机行走有可能走到最远的节点. 如果某个节点的期望步数为 12, 意味着没有蚂蚁走到该节点或者所有蚂蚁以 12 步第一次到达该节点. 鉴于此, 本文只输出期望步数小于 12 的结果作为复述, 并进行评测.

5.3 输出结果过滤

在输出结果中发现了以下两种情况: (1) 输出短语和给定短语互相包含, 例如给定短语为 “含药物”, 输出短语为 “含药物的”; (2) 多个输出短语互相包含, 例如给定短语为 “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引起词性改变, 不是真正意义上的复述, 因此根据下面规则对其进行过滤. 针对情况 (1) 不会将其作为复述输出; 情况 (2) 只保留与给定短语长度最接近的短语.

5.4 实验结果

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.

由于构建的翻译关系图都非常庞大, 无法人工从中寻找出所有复述作为标准答案, 也就无法用召回率对结果进行评测. 对此, 采取下面评测方式. 对于 3 个实验中分别输出的复述结果, 人工 (1 人) 判断并统计其中包含的正确复述个数, 并计算平均错误率. 100 条测试数据上的平均评测结果如表 3 所示, 分析结果如下.

(1) 实验 1 (传统基于 “枢轴” 方法) 平均获取 3.79 个复述结果, 其中平均正确复述数量为 2.88. 实验 2 (随机行走步数 $N = 12$) 平均获取 11.56 个复述结果, 其中平均正确复述数量为 7.18, 与实验 1 的

表 3　实验评价结果

Table 3　The experimental results

Experiment	Average number of paraphrases	Average number of correct paraphrases	Average error rate (%)	Average penalty (The error rate of single paraphrases)
1	3.79	2.88	24.01	8.34
2	11.56	7.18	37.89	5.28
3	15.32	9.04	40.99	4.53

传统方法相比发现了更多的正确复述, 是实验 1 结果的 2.49 倍, 说明了本文所提方法具有挖掘更多复述的能力, 验证了构建翻译关系图并利用随机行走算法获取复述的有效性.

(2) 实验 3 (多语言对的翻译关系图扩展) 平均获取 15.32 个复述结果, 其中平均正确的复述数量为 9.04, 与实验 2 的结果相比进一步发现了更多的复述, 数量提升了 25%, 验证了增加语言对进行翻译关系扩展的方法的有效性.

(3) 同时, 在实验 2 和 3 中, 随着正确复述数量的增加, 平均错误率也有所增加. 如前所述, 由于无法得知正确复述数量, 所以无法计算召回率并使用 F 值对召回率和正确率进行综合评价. 为了对复述数量和错误率进行统一考量, 从获取复述的代价这一角度出发, 进一步调查了复述数量增加与错误率上升之间的关系. 用平均错误率与平均正确复述数量的比例表示发现正确复述的平均惩罚, 惩罚越高预示后续处理利用该结果的代价越高, 比如, 人工选择正确复述的代价. 平均惩罚的计算结果显示在表 3 的最后一列. 结果表明基于本文所提方法的实验 2 和 3 获取正确复述的平均惩罚分别为 5.28 和 4.53, 远远低于基于传统方法实验 1 的 8.34. 通过分析比较实验结果, 发现与传统方法相比, 本文所提方法以更小代价获取到更多数量的复述, 进一步验证了本文所提方法的有效性.

式 (1) 定义期望步数作为复述的可信度评测指标, 期望步数越小, 表示成为给定短语的复述的可能性越大. 表 4 给出了部分测试短语的复述结果以及相应的期望步数, 结果按照期望步数的递增顺序排列. 本文观察分析其中编号为 1, 2, 3 的 3 个测试短语的结果.

(1) 编号为 1 的测试短语 “传送到远程” 的结果, 发现期望步数小于 12 的排在前面的结果在语义上接近测试短语, 比如排在第一位的 “发送给远端” 和第二位的 “传送到远处”; 而期望步数等于 12 (表 4 中下划线标示部分) 的排在后面的结果在语义上与测试短语相差较大, 比如排在第十一位的 “命令发送到” 和第十二位的 “距离该远端” 在语义上与测试短语差异较大. 我们只输出期望步数小于 12 的结果作为复述, 此处为了分析比较列出了几个期望步数等于 12 的结果, 用下划线标识. 这些比较结果表明, 利用期望步数作为可信度指标并将期望步数小于 12 作为复述的选择标准是可行的.

(2) 观察编号为 2 的测试短语 “船体” 的结果, 发现排在第一位 “船壳” 比排在第四位的 “轮船” 在语义上更接近测试短语.

(3) 观察编号为 3 的测试短语 “微不足道” 的结果, 发现排在第一位的 “无足轻重” 比排在第三位的 “忽略不计” 在语义上更接近测试短语.

这些分析结果表明, 用期望步数作为可信度指标对结果排序的确将语义更接近测试短语的结果排在了前面, 而语义稍微偏离的排在了后面. 由此得出结论, 用期望步数作为复述可信度指标是合理有效的, 期望步数越小, 对应的短语成为复述的可能性越大.

此外, 对错误的结果也进行了分析. 通过观察表 4 中编号为 4 的测试短语 “不断提高” 的结果, 发现排在第四位和第六位的结果 “然后增加” 和 “以及提高” 并不是 “不断提高” 的复述, 但是得到这些结果的期望步数较小被排在靠前的位置. 经过分析我们认为, 因为在使用期望步数计算复述可信度

表 4 部分测试短语的结果及对应的期望步数

Table 4 Some examples of the acquired paraphrase results and the corresponding hitting time

Number	Given phrase	The result of paraphrases	Hitting time	Number	Given phrase	The result of paraphrases	Hitting time
1	传送到远程	发送给远端	9.44202	4	不断提高	继续增大	11.2896
		传送到远处	10.6529			连续升高	11.3157
		传输到远端	10.8738			日益提高	11.8252
		传送到远端	11.3424			然后增加	11.8338
		传输给远端	11.452			连续增加	11.8834
		朝向该远程	11.8415			以及提高	11.9046
		传输到远程	11.8629			继续增加	11.9507
		传递到远端	11.9333			相继增加	11.9507
		传递给远程	11.9999			日益增多	11.9636
		运输至遥远	11.9999			然后增大	11.9724
		命令发送到	12			连续增大	11.9935
		距离该远端	12			持续上升	11.9943
		向远端延伸	12			持续增长	11.9947
		溶液传送至	12			然后升高	11.9964
		给远端位置	12			不断增大	11.9970
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

时, 期望步数与短语表中的翻译概率有间接关系, 所以对齐结果中估算有偏差的翻译概率值对其产生影响, 导致个别非复述结果的期望步数较小, 排位比较靠前.

实验 3 中通过加入英日短语表扩展翻译关系图, 获取了更多的复述, 其中一个例子如图 6 所示. 图 6 中基于汉英翻译关系建立的子图用实线包围, 基于英日翻译关系建立的子图用虚线包围. 从图 6 中可以看到两个实线子图通过虚线子图连接建立了联系. 具体地, 通过日语短语 “整合” 与两个汉英子图的 “the matching”, “matching”, “alignment” 的翻译关系, 建立了 “匹配” 与 “对准” 和 “校正” 的连接, 从而获取到新的复述 “匹配”. 如果只利用汉英短语表构建翻译关系图, 汉语短语 “匹配” 无法与 “对准” 或 “校正” 建立联系, 复述结果只有 “对准” 和 “校正”.

6 结论及未来工作

在利用双语平行语料获取复述知识方面, 本文针对传统的基于 “枢轴” 方法的局限性, 提出了构建翻译关系图并利用随机行走算法获取复述的方法和基于期望步数的复述可信度计算方法, 并进一步提出面向多语言对的翻译关系图扩展方法. 本文在 NTCIR 汉英和英日平行语料上进行了评测, 并与

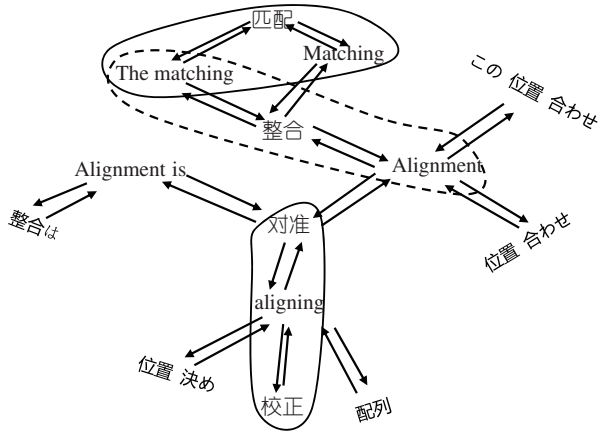


图 6　加入英日短语表获取的复述的例子

Figure 6　An example acquired by adding English-Japanese phrase table

传统方法进行了对比实验, 实验结果表明本文所提方法能够有效提高获取复述的数量.

今后, 准备在本文工作的基础上, 扩大实验测试数据集, 进一步调查实验参数对复述获取结果的影响; 并针对单词对齐带来的噪音问题, 探索在翻译关系图中引入新的特征节点提高复述结果正确率的方法.

参考文献

1 Barzilay R, McKeown K. Extracting paraphrases from a parallel corpus. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, 2001. 50–57

2 Callison-Burch C, Koehn P, Osborne M. Improved statistical machine translation using paraphrases. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, New York, 2006. 17–24

3 Barzilay R. Information fusion for multi-document summarization: paraphrasing and generation. Dissertation for Ph.D. Degree. New York: Columbia University, 2003

4 Zhou L, Lin C Y, Munteanu D S, et al. ParaEval: using paraphrases to vealuate summaries automatically. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, New York, 2006. 447–454

5 Zukerman I, Raskutti B. Lexical query paraphrasing for document retrieval. In: Proceedings of the 19th International Conference on Computational Linguistics, Taipei, 2002. 1–7

6 Iordanskaja L, Kittredge R, Polguere A. Lexical selection and paraphrase in a meaning-text generation model. In: Proceedings of Natural Language Generation in Artificial Intelligence and Computational Linguistics. New York: Springer, 1991. 293–312

7 McKeown K. Paraphrasing using given and new information in a question-answer system. In: Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics, La Jolla, 1979. 67–72

8 Madnani N, Dorr B J. Generating phrasal and sentential paraphrases: a survey of data-driven methods. Comput Linguist, 2010, 3: 341–387

9 Bannard C, Callison-Burch C. Paraphraseing with bilingual parallel corpora. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Michigan, 2005. 597–604

10 Kok S, Brockett C. Hitting the right paraphrases in good time. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Los Angeles, 2010. 145–153

11 Brand M. A random walks perspective on maximizing satisfaction and profit. In: Proceedings of the 2005 SIAM

International Conference on Data Mining, Irvine, 2005. 12–19
12 Sarkar P, Moore A W, Prakash A. Fast incremental proximity search in large graphs. In: Proceedings of the 25th International Conference on Machine learning. Helsinki: Finland Ages, 2008. 896–903

Acquiring Chinese paraphrases based on random walk of N steps

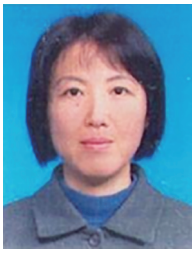
Jun MA, Yujie ZHANG*, Jin'an XU & Yufeng CHEN

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044*, China*
* Corresponding author. E-mail: yjzhang@bjtu.edu.cn

Abstract The conventional "pivot" approach of acquiring paraphrases from bilingual corpus has certain limitations where only candidate paraphrases within two steps are considered. In this paper, we propose a graph-based model of acquiring paraphrases from a phrase translation table. First, we describe a graph-based model representing Chinese-English phrase translation relations, a random walk algorithm based on N number of steps and a confidence metric for the obtained paraphrases. Furthermore, with the aim of finding more potential for Chinese paraphrases, we augment the model so that it is able to integrate other language pairs, such as English-Japanese phrase translation relations. We performed experiments on NTCIR Chinese-English and English-Japanese bilingual corpus and compared the results to those of conventional methods. The experimental results show that the proposed approach acquires more paraphrases. In addition, the performance was improved further after the English-Japanese phrase translations were added to the graph-based model.

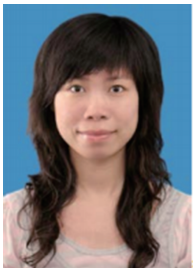
Keywords paraphrase acquisition, phrase translation table, translation relation graph, random walk, hitting time

Jun MA was born in 1991. He received a B.E. degree in computer science and information technology from Beijing Jiaotong University, Beijing, China, in 2015. He is pursuing a master's degree at Beijing Jiaotong Univerisity, majoring in computer science and technology. His main research is paraphrase acquisition and application.



Yujie ZHANG was born in 1961. She received a B.E. degree in computer science and information technology from Beijing Jiaotong University, Beijing, China, in 1983 and a Ph.D. in computer science from the University of Electro-Communications, Tokyo, Japan, in 1999. She joined Beijing Jiaotong University in April 2011 and is currently a professor at the School of Computer and Information Technology. Her current research interests include machine translation, multilingual information processing, and natural language processing.

Jinan XU was born in 1970. He received a B.E. degree from the Communication and Control Engineering Department from Northern Jiaotong University, Beijing, China, in 1992 and M.S. and Ph.D. degrees from Hokkaido University, Japan, in 2003 and 2006 respectively. From 2006 to 2009, he worked as an expert researcher at the Central Research Institute of Nippon Electronic Company (NEC). He became an associate professor in Beijing Jiaotong University at the School of Computer and Information Technology in 2009. His research interests include natural language processing, machine translation, AI, and affective computing.



Yufeng CHEN was born in 1981. She received an undergraduate degree in mechanical electrical engineering from Beijing Jiaotong University in 2003, and her Ph.D. in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, in June 2008. From July 2008 to September 2014, she joined NLPR and worked as an associate professor. Since September 2014, she has been working as an associate professor at the School of Computer and Information Technology, Beijing Jiaotong University. Her research interests include natural language processing, machine translation, and information retrieval.