利用 Markov 网络抽取复述增强机器译文自动评价方法*

翁贞，李茂西，王明文

（江西师范大学 计算机信息工程学院，南昌 330022）

摘要： 在机器译文自动评价中，匹配具有相同语义、不同表达方式的词或短语是其中一个很大的挑战。许多研究工作提出从双语平行语料或可比语料中抽取复述来增强机器译文和人工译文的匹配。然而双语平行语料或可比语料不仅构建成本高，而且对少数语言对难以大量获取。我们提出通过构建词的 Markov 网络，从目标语言的单语文本中抽取复述的方法, 并利用该复述提高机器译文自动评价方法与人工评价方法的相关性。在 WMT'14 Metrics task 上的实验结果表明，我们从单语文本中提取复述方法的性能与从双语平行语料中提取复述方法的性能具有很强的可比性。因此，本文提出的方法可在保证复述质量的同时，降低了复述抽取的成本。

关键词： 复述；机器译文自动评价；Markov 网络；相关性

中图分类号： TP391　　**文献标识码：** A

Enhance Automatic Evaluation of Machine Translation by Markov Network-Extracted Paraphrases

Zhen Weng, Maoxi Li, Mingwen Wang

(School of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, China)

Abstract: It is a great challenge to match words or phrases which have the same meanings but different expressions in the automatic evaluation of machine translation. Many researchers proposed to enhance the matches between the words in machine translation and in human references by extracting paraphrases from bilingual parallel corpus or comparable corpus. However, the cost of constructing the bilingual parallel corpus or the comparable corpus is high; furthermore, it is difficult to obtain some language pairs on a large scale by using corpus. In this paper, the paraphrases were extracted from the monolingual texts in the target language by constructing the Markov networks of words, and applied to improve the correlation between the results of automatic evaluation and the human judgments of machine translation. The experimental results on WMT'14 Metrics task showed that the performances of the approach of extracting paraphrase from monolingual text we proposed and that of extracting paraphrase from bilingual parallel corpus have a strong comparability. Thus, the proposed approach in the paper can guarantee the quality of the paraphrases, and meanwhile lower the cost of extracting the paraphrase.

Key words: paraphrase; automatic evaluation of machine translation; Markov network; correlation

1 引言

机器译文自动评价方法不仅能给出表征翻译系统翻译质量优劣程度的定量指标，而且能在统计翻译系统开发时指导其参数优化。因此，它推动机器翻译技术的快速发展。

近几年来，许多机器译文自动评价方法相继被提出，包括被研究者广泛使用的 BLEU[1]，

*基金项目：国家自然科学基金（61163006，61203313，61462044，61272212）；国家语委“十二五”规划（YB125-99）；江西省自然科学基金（20132BAB201030，20151BAB207025）；江西省研究生创新基金（YC2014-S149）

NIST[2]，METEOR[3]，TER[4]，MAXSIM[5]等等。其中，BLEU 和 NIST 是基于 *n* 元文法匹配准确率的评价指标；METEOR 和 MAXSIM 是考虑准确率和召回率的评价指标；而 TER 是基于翻译错误率的指标。在译文评价过程中，它们均遵循 BLEU 的主要思路“机器译文越接近人工参考译文，机器译文的质量越好。”这些评价方法将机器译文中的词语与人工参考译文中的词语进行比较，词形相同的词被看作表达了同一含义，即认为是一个匹配，而词形不同的词语被看作表达不同的含义，即认为没有匹配。然而，由于语言现象的多样性，同义词、近义词和不同的表达方式等现象在评价时大量存在。因此，如何准确地进行词语匹配是机器译文自动评价时一个难题。

针对这个问题，许多机器译文评测活动尝试提供更多的人工参考译文来提高机器译文和人工参考译文的匹配。比如在 NIST 评测中，测试集中每个待翻译的源语言句子就提供了 4 个人工参考译文，以供自动评价方法进行打分；而 IWSLT 评测中，有的翻译方向甚至提供了 16 个人工参考译文。毫无疑问，人工参考译文越多，覆盖的语言现象就越全面，机器译文中的词语就能得到更准确的匹配，但是，这也意味着构建参考译文的费用越高，而且再多的人工参考译文也不能穷尽所有的语言现象。这种方法的一个改进是自动生成参考译文以覆盖更多的语言现象，王博等通过句法结构知识来对人工参考译文进行扩展，衍生出更多的参考译文以供机器译文匹配，从而提高自动评价结果的相关性[6]。Kauchak 和 Barzilay 提出使用复述改写人工参考译文，以使参考译文接近于机器译文，提高自动评价方法的准确性[7]。

另外一种方法是，机器译文中的词和人工参考译文中的词比较时，放松词语匹配的条件，即不再限定匹配的词语仅是词形完全相同的词语，还应包括同根词、近义词、同义词和复述等等。这种方法使用语言学知识和语料资源来获取相同语义、不同表达方式的词或短语以供匹配，因此，它容易获取，便于扩展，而且构建费用低廉。它的一个典型的例子是 METEOR 工具包。METEOR 最初的版本只支持完全匹配，在后续的版本里，它相继的扩充了词干匹配、同义词匹配和复述匹配等模块，并且匹配是分阶段进行，每一阶段只匹配上一阶段没有得到匹配的词语[8]。与此相应的，TER 自动评价方法也由最初的完全匹配版本 Tercom 发展到后续的采用完全匹配和复述匹配的 Terp 版本[9]。

本文研究利用机器学习方法，词的 Markov 网络，从目标语言的单语文本中抽取复述，来替换传统的从双语文本中抽取复述，然后将抽取的复述应用在机器译文的自动评价方法 METEOR 和 TER 上增强词语之间的有效匹配，并通过实验验证我们的方法尽管只使用单语文本，但是并没有降低译文自动评价结果与人工评价结果之间的相关性。

2 相关工作

复述是指在某一种语言中，语义相同而内容和表达形式不同的词、短语、句子和段落[10]。复述知识已经成功的应用到自然语言处理的多个任务中，包括信息检索[11]、自动文摘[12]和机器翻译[13-15]等等。

在复述的抽取技术方面，Barzilay 和 McKeown 提出了利用非监督学习的方法从同一个源语言句子的不同英文译文中抽取词和短语的复述[16]。Bannard 和 Callison-Burch 提出利用统计机器翻译中的词对齐技术从双语平行语料中抽取复述，在他们方法中由于一种语言的词

或短语，被用作待抽取的另一种语言复述中的枢轴(pivot)，因此它也被称为枢轴法[17]。不同于从双语语料中抽取复述的方法，Shinyama 等提出一种使用命名实体识别特征从单语的新闻文章中抽取复述的方法，这些来源不同的新闻文章在同一时期报导了同一件新闻事件[18]。Barzilay 和 Lee 提出使用多个文本串对齐算法从未标注的可比语料库中学习句子级别的复述[19]。尽管后面两种方法从单语文本中抽取复述，但是它们对使用的单语文本语料仍然有较大的限制。而本文提出的利用词的 Markov 网络抽取复述方法对单语文本无任何限制。

在机器译文自动评价方面，Kauchak 和 Barzilay 提出使用句子级别的复述改写人工参考译文，类似于 Barzilay 和 Lee 的方法，以使参考译文中的词语与机器译文中词语最大程度的相同，并通过实验验证了使用改写的人工参考译文进行评价改善了自动评价的准确性[7]。Zhou 等提出使用词或短语的复述来增强机器译文和人工参考译文之间的匹配，他们使用 Bannard 和 Callison-Burch 提出的枢轴法从双语平行语料中抽取复述，然后通过两步法进行词语匹配，首先使用复述知识进行匹配，然后使用词形进行完全匹配[20]。沿着 Zhou 等方法的思路，Denkowski 和 Lavie 也使用枢轴法从双语平行语料中抽取目标语言的复述，并使用复述来增强 METEOR 方法中词语的匹配，但是，在他们方法中，词语的匹配顺序与前者相反[8]。与此类似，Snover 等也在 TER 最初的完全匹配的基础上增加了复述匹配[9]，他们抽取复述和匹配的顺序与 METEOR 相同，而且他们均对复述匹配和完全匹配设置了不同的权重。

3 利用 Markov 网络的复述构建

3.1 Markov 网络

Markov 网络是一种进行不确定性推理的有力图形工具[11]。由于构建 Markov 网络时不考虑边的方向，因此我们可以很容易地利用 Markov 网络从实例数据中建立知识关联。一个 Markov 网络可以表示为一个二元组 $G=(V,E)$，节点 $v\in V$ 表示网络中的所有节点，边 $(x_i,x_j)\in E$ 表示一组无向边。在 Markov 网络中满足 $p(v_i\mid v_j)=p(v_i\mid v_j(v_i,v_j)\in E)$，即每个节点条件独立于其邻居节点给定的非邻居节点的任意节点子集，$E=\{(x_i, x_j)\mid x_i\neq x_{j\wedge}x_i, x_j\in V\}$，$E$ 中的边表示节点之间的关系。

通过词间相关性得出的 Markov 网络结构中，每个词为一个节点，连接两个节点的边表示这两个词之间的关系，用权重表示其相关性。有些词节点和彼此相连的边构成了一个完全子图，即任意两个节点之间都有边相连，我们把这样的多边形称作词团 C，包含 n 个词节点且含有节点词 t_i 的词团记为 $C_n(t_i)$，由 $C_n(t_i)$ 构成的集合记为 $S(C_n(t_i))$。图 1 给出了一个词的 Markov 结构图的简单示例，其中两个词团分别为 C_3(data)和 C_4(data)。我们利用两个词的词团信息来量化这两个词互为复述的可能性。

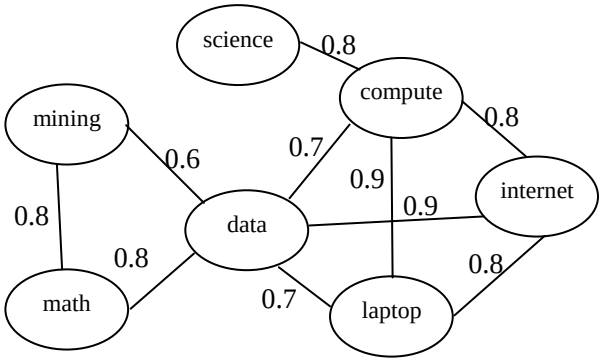

图 1 词的 Markov 网络结构图

3.2 构建词的 Markov 网络

本文采用词的共现性来计算词间的关系，计算词共现词频时一般可以以整个文档、段落或是一个固定长度的文本为窗口[11]。出于考虑效率方面的因素，本文选用固定长度的一段文本作为窗口单位，鉴于 Markov 网络的无向性，在构造词的 Markov 网络时，采用两个词的综合共现性来计算，如公式 1-公式 3 所示。

$$R(t_i,t_j)=\frac{P_{co}(t_i\mid t_j)+P_{co}(t_j\mid t_i)}{2} \tag{1}$$

$$P_{co}(t_i\mid t_j)=\frac{C(t_i,t_j)}{C(t_j)} \tag{2}$$

$$P_{co}(t_j\mid t_i)=\frac{C(t_i,t_j)}{C(t_i)} \tag{3}$$

其中，t_i 和 t_j 指两个词，$C(t_i,t_j)$ 指在训练语料中词 t_i 和 t_j 在同一个窗口中同时出现的频率，$C(t_i)$ 和 $C(t_j)$ 分别表示在训练语料中词 t_i 和 t_j 出现的频率，$R(t_i,t_j)$ 表示词 t_i 和 t_j 之间的相关性，R 值越大，两个词的相关性就越高。当 R 值大于给定的阈值时，则词 t_i 和 t_j 相互依赖，即在词的 Markov 网络中有边相连。

3.3 词团的提取

构成词团的词彼此相互依赖，即存在某种语义关联，可以认为他们表达了同一个概念或主题。如图 1 中的词“compute”，“data”，“internet”，“laptop”构成了一个 4 阶词团。根据离散数学中定理：C 是一个团，那么必存在一个最大团 $C_{\max}$ 使得 $C\subseteq C_{\max}$。假设在一个 Markov 网络中的节点集合 $V=\{t_1,t_2,...t_n\}$ 构造团序列 $C_0\subset C_1\subset C_2\subset...$ 其中 $C_0=C$ 且 $C_{i+1}=C_i\cup\{t_j\}$，j 满足 $t_j\notin C_i$，t_j 与 C_i 中各节点都有边相连。由于 T 的词节点个数 $|T|=n$，所以最多经过 $n-|C|$ 步，就使得这个过程终止，此序列的最后一个团，就是所要找的最大团。根据上述思想，本文从词的 Markov 网络中提取词的词团，即在 $C_n(t_i)$ 的基础上获取 $C_{n+1}(t_i)$。实现算法如算法 1 所示，其中 $S(C_k(t_i))$ 表示词 t_i 的 k 阶词团集合，$S(C_{k+1}(t_i))$ 表示 t_i 的 $k+1$ 阶词团集合，set1、set2、set3、set4 是定义的四个集合，算法 1 第 6 行表示取出 $S(C_k(t_i))$ 中的一个词团，算法 1 第 10-12 行说明 set1 和 set2 这 2 个词团只有 2 个不同的词，且这两个词有

边相连。我们用公式（4）计算每个词团的权重。其中，n 表示词团中的节点个数，$R(t_i,t_j)$ 表示词 t_i , t_j 的相关性。

$$w_n\{t_1,t_2,...t_n\}=\frac{\sum_{1\le i,j\le n} R(t_i,t_j)}{\frac{1}{2}n(n-1)} \quad (4)$$

算法 1 通过 $C_k(t_i)$ 提取 $C_{k+1}(t_i)$ 的算法

```
1.  输入： S(C_k(t_i))
2.  输出： S(C_{k+1}(t_i))
3.  BEGIN
4.  INITIALIZE: m:=0 ;     S(C_{k+1}(t_i)) := ∅ ;
5.  DO WHILE m< | S(C_k(t_i)) |
6.      set1= S(C_k(t_i)) .get(m) ;
7.      FROM x=m+1 TO x= | S(C_k(t_i)) |
8.          set2 := S(C_k(t_i)) .get(x) ;
9.          set3 :=set1 U set2 ;
10.         IF |set3|==k+1
11.             set4=set3-(set2 I  set1) ;
12.             IF t1,t2 ∈ set4 and  < t_i,t_j > ∈ E
13.                 IF set3 ∉ S(C_{k+1}(t_i))
14.                     S(C_{k+1}(t_i)) .add(set3) ;
15.                 END IF
16.             END IF
17.         END IF
18.     END FROM
19. END DO
20. END
```

3.4 复述构建

传统词的Markov网络节点的粒度都是单词级别的[11]。本文为了抽取不同粒度的复述对，首先统计每个句子中的 n 元文法在整个语料中出现的次数，次数超过预先设置阈值的语块视为该句子中的短语（并非语言学意义上的短语），将这些短语看成一个整体，并以它们为粒度对该句子进行切分，得到词或短语用于构建 Markov 网络的节点，利用这种方法抽取单词或短语级别的复述实例。在后续的实验中，我们把在语料中出现次数超过 3 次的语块视为短语，并经验性的设置短语抽取长度不超过 2 个单词。

在 Markov 网络中构成词团的词项存在的语义相关包括语义相同和主题相关但语义不同。如果直接用词团的权重度量两个词项互为复述的可能性，会存在大量相关而不相似的词对。本文采用两个词项的 n 阶词团集合的相似性度量这两个词项互为复述的可能性，其本质是通过除这两个节点以外的其他邻居节点来计算这两个节点的关系。因此，词团的节点个数必须大于 2，考虑到可以通过合并词项的三阶词团得到该词项的任何一个更高阶的词团，本文的后续实验中通过计算两个词项的 3 阶词团集合的相似性，来度量这两个词项互为复述的可能性。如图 2 中每个词项的 3 阶词团集合分别是 $S(C_3(t_1))$={t_1， t_2，t_4 }，$S(C_3(t_2))$={{ t_1，t_2，t_4}，{t_2，t_3，t_4}}，$S(C_3(t_3)$={ t_2，t_3，t_4}，$S(C_3(t_4))$={{ t_1，t_2，t_4}，{ t_2，t_3，t_4}}因此，词

t_2 和 t_4 更有可能互为复述。我们用公式（5）和（6）来计算词 t_i, t_j 互为复述的可能性 $prob(t_i, t_j)$，其中 $W_3(t_i, t_j)$ 表示所有同时包含词项 t_i 和 t_j 的三阶词团的权重和，$W_3(t_i)$ 表示所有包含词项 t_i 的三阶词团的权重和，$W_3(t_j)$ 表示所有包含词项 t_j 的三阶词团的权重和：

$$prob(t_i, t_j) = \frac{W_3(t_i, t_j)}{\frac{1}{2}(W_3(t_i) + W_3(t_j))} \tag{5}$$

$$W_3(t_i, t_j) = \sum_{k \neq i \wedge k \neq j \wedge t_k \in clique(t_i, t_j, t_k)} w_3(t_i, t_j, t_k) \tag{6}$$

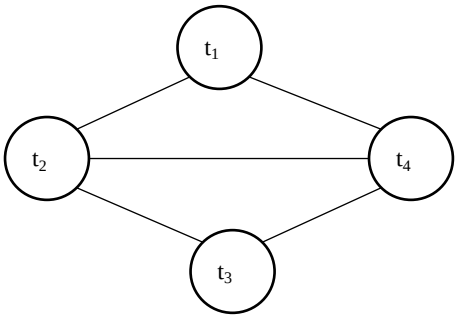

图 2 3 阶词团

4 实验

4.1 实验数据

为了比较利用 Markov 网络提取的复述和基于枢轴法提取的复述在机器译文自动评价方法上的性能，我们将提取的复述应用在机器译文自动评价开源工具包 terp-v1[1]和 meteor-1.5[2]，并在 WMT'14 Metrics task 上进行对比实验。该评测包含 10 种不同的翻译方向的任务，其中包含 5 种目标语言是英语的任务，5 种源语言为英语，目标语言是其他欧洲语言的任务，每个任务的人工参考译文只有 1 个，所提交的机器翻译系统一共有 110 个。为了提取 5 种语言的复述，我们选用 5 个不同语言对的双语平行语料进行提取复述表，其中本文的方法只用双语平行语料的目标语言端文本，而基于“枢轴法”则需要包含源语言端和目标语言端的双语平行语料。实验中我们选用 WMT'14 和 WMT'15 的机器翻译训练语料 Europarl v8、NewsCommentary[3] 和 Europarl v7[4]进行提取复述表。语料的统计数据见表 1。

[1] http://www.umiacs.umd.edu/~snover/terp/
[2] http://www.cs.cmu.edu/~alavie/METEOR/
[3] http://www.statmt.org/wmt15translation-task.html
[4] http://www.statmt.org/wmt14/translation-task.html

表 1：提取复述的语料

语料名称	语言对	句子数量
Europarl v8	芬兰语，英语	约 200 万
Europarl v7	德语，英语	约 200 万
	捷克语，英语	约 70 万
News Commentary	法语，英语	约 10 万
	俄语，英语	约 20 万

4.2 实验设置

在实验中，分别用本文的方法和基于枢轴法的方法提取五种语言的复述，分别是英语、法语、德语、俄语、捷克语。由于利用枢轴法提取复述必须在双语平行语料上完成，为了更准确的比较两个方法，本实验在提取复述时选用同一个双语平行语料，但本文的方法只用双语平行语料中的目标语言端。获取到了复述后，我们将其应用在 METEOR 和 TER 上，对 WMT'14 Metrics task 的 10 个任务进行评测。

为了比较不同的复述抽取方法在机器译文自动评价上的性能，我们利用皮尔森相关系数计算自动评价结果和人工评价结果的系统级别相关性：

$$r=\frac{\sum_{i=1}^{n}(H_i-\bar{H})(M_i-\bar{M})}{\sqrt{\sum_{i=1}^{n}(H_i-\bar{H})^2}\sqrt{\sum_{i=1}^{n}(M_i-\bar{M})^2}} \quad (7)$$

在公式（7）中，H_i，M_i 分别是第 i 个系统的人工评价得分和自动评价得分，$\bar{H}$，$\bar{M}$ 分别是人工评价得分和自动评价得分的平均值。

同时，我们利用 Kendall's τ 相关系数计算自动评价结果和人工评价结果的句子级别相关性：

$$\tau=\frac{|Concordant|-|Discordant|}{|Concordant|+|Discordant|} \quad (8)$$

在公式（8）中，*Concordant* 表示人工评价与自动评价排名一致的集合，*Discordant* 表示人工评价与自动评价排名不一致的集合。

4.3 实验结果

我们提取了 5 种语言的复述表，表 2 是用本文的方法获取的英语复述表的一部分实例。其中第一列为两个词互为复述的可能性，第二列和第三列分别是互为复述的两个词。本文提取的复述包含单词级别和短语级别的复述。

表 2 基于本文方法提取的英文复述表的一些例子

…

0.0880852061391501|||allow us|||enable us

0.0937897393987293|||such reports|||these reports

0.085853721363|||essential for|||fundamental for

0.08335206139702|||publishers|||the publishing

0.08617136303955722|||military|||warplanes

0.0937985660236197|||eu|||european

0.0922552473361782|||features|||characteristics

0.08909795820836298|||court|||justice

0.08450979808463252|||countries||| states

0.0937852061391501|||respond||| answer

…

表 3 和表 5 给出了机器译文自动评价方法 METEOR 和 TER 在 WMT'14 Metrics task 目标语言为英语的任务上与人工评价的系统级别相关性和句子级别相关性，表 4 和表 6 给出了其在目标语言为其他欧洲语言的任务上与人工评价的系统级别相关性和句子级别相关性。这四张表的第一列表示使用不同复述资源的 METEOR 和 TER，其中“METEOR”和“TER”表示只做词形上的完全匹配，不做复述匹配，“METEOR-Markov”和“TER-Markov”表示 METEOR 和 TER 使用基于 Markov 网络模型提取的复述表进行复述匹配，“METEOR-Pivot”，“TER-Pivot”表示 METEOR 和 TER 使用基于枢轴法提取的复述表进行复述匹配。

从表 3 给出的数据可以看出，METEOR 和 TER 在源语言分别为法语和德语的任务上，“METEOR-Markov”与人工评价的系统级别相关系数最大；“TER-Markov”在源语言为印度语的任务上，与人工评价的系统级别相关系数最大，且五个任务的系统级别相关系数的平均值与基于枢轴法的相等。从表 5 给出的数据可以看出，“METEOR-Markov”与人工评价的句子级别相关系数的平均值最大。这可以说明利用 Markov 网络模型构造的英语复述表不仅可以增强除词形完全匹配外的有效匹配，而且在机器译文自动评价方法 METEOR 和 TER 上的性能比基于枢轴法提取的复述表略好。

从表 4 给出的数据可以看出，“TER-Markov”在目标语言为德语的任务上，与人工评价的系统级别相关系数最大，且五个任务的系统级别相关系数的平均值最大。从表 6 给出的数据可以看出，“METEOR-Markov”在目标语言为德语和俄语的任务上，与人工评价的句子级别相关性最大，且与人工评价的句子级别相关系数的平均值与“METEOR-Pivot”相等。在这五个任务中目标语言为德语的任务参加翻译的系统数量最多。这可以说明利用 Markov 网络模型构造的其他欧洲语言的复述表在机器译文自动评价方法 METEOR 和 TER 上的性能并没有低于基于枢轴法提取的复述表，甚至还略有提升。

总之，实验结果表明：我们提出的利用 Markov 网络构建复述表的方法不仅降低了对训练语料的要求，而且还保证了所提取的复述表在机器译文自动评价方法上的性能不低于前人的方法。

表 3 各自动评价方法在 WMT2014 上目标语言是英文的任务上的评价结果与人工评价的系统级别相关系数

Meterics	系统级别相关系数					
	fr-en	de-en	hi-en	cs-en	ru-en	Average
TER	0.952	0.775	0.618	0.976	0.809	0.826
TER-Pivot	0.958	0.784	0.719	**0.990**	**0.811**	**0.852**
TER-Markov	0.957	0.775	**0.729**	0.988	**0.811**	**0.852**
Meteor	0.969	0.889	0.484	0.985	0.786	0.823
Meteor-Pivot	0.972	0.908	0.459	0.975	0.800	0.823
Meteor-Markov	**0.975**	**0.918**	0.451	0.969	0.798	0.822

表 4 各自动评价方法在 WMT2014 上源语言是英文的任务上的评价结果与人工评价的系统级别相关系数

Meterics	系统级别相关系数					
	en-fr	en-de	en-hi	en-cs	en-ru	Average
TER	0.957	0.399	0.772	0.973	0.930	0.806
TER-Pivot	**0.959**	0.422	0.772	0.969	**0.934**	0.811
TER-Markov	0.958	**0.440**	0.772	0.964	0.928	**0.812**
Meteor	0.939	0.240	**0.924**	**0.979**	0.932	0.803
Meteor-Pivot	0.942	0.261	**0.924**	0.977	0.933	0.807
Meteor-Markov	0.941	0.280	**0.924**	0.975	0.931	0.810

表 5 各自动评价方法在 WMT2014 上目标语言是英文的任务上的评价结果与人工评价的句子级别相关系数

Meterics	句子级别一致性					
	fr-en	de-en	hi-en	cs-en	ru-en	Average
TER	0.371	0.253	0.265	0.192	0.266	0.269
TER-Pivot	0.379	0.260	0.274	0.198	0.273	0.277
TER-Markov	0.371	0.253	0.265	0.192	0.266	0.269
Meteor	0.401	0.319	0.398	0.267	0.311	0.339
Meteor-Pivot	**0.414**	**0.330**	0.416	0.265	0.326	0.350
Meteor-Markov	0.404	0.324	**0.421**	**0.274**	**0.328**	**0.351**

表 6 各自动评价方法在 WMT2014 上源语言是英文的任务上的评价结果与人工评价的句子级别相关系数

Meterics	句子级别一致性					
	en-fr	en-de	en-hi	en-cs	en-ru	Average
TER	0.246	0.206	0.146	0.280	0.358	0.247
TER-Pivot	0.247	0.215	0.146	0.285	0.392	0.257
TER-Markov	0.242	0.216	0.146	0.282	0.395	0.256
Meteor	0.275	0.212	**0.303**	0.310	0.407	0.301
Meteor-Pivot	**0.280**	0.227	**0.303**	**0.319**	0.423	**0.310**
Meteor-Markov	0.276	**0.232**	**0.303**	0.314	**0.426**	**0.310**

5 总结和展望

利用 Markov 网络在语义推理方面的优势，本文提出了一种从单语文本中抽取复述的方法，并将其成功应用在机器译文自动评价中，以有效的进行语义相同表达不同的词或短语的匹配。与从双语平行语料和单语可比语料抽取复述方法相比，该方法抽取复述时对单语文本没有任何限制，因此它有很好的推广性，在以后的研究中，我们将尝试将其应用在机器翻译短语表的扩展、自动文摘中近义词的生成以及信息检索中相关搜索的构建上，以丰富复述的研究。

参考文献

[1] Papineni K, Roukos S, Ward T, et al. BLEU: a Method for Automatic Evaluation of Machine Translation[C]// Proce- -edings of the 40th Annual Meeting on Association for Computational Linguistics, 2002: 311-318.

[2] Doddington G. Automatic Evaluation of Machine Transla- tion Quality Using N-gram Co-occurrence Statistics[C]// HLT '02 Proceedings of the second international confe- -rence on Human Language Technology Research, 2002:138-145.

[3] Banerjee S, Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments[C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005:65-72.

[4] Snover M, Dorr B, Schwartz R, et al. A Study of Translation Edit Rate with Targeted Human Annotation[C]//Proce- -edings of Association for Machine Translation in the Americas, 2006:223-231.

[5] Chan Y S, Ng H T. MAXSIM: A Maximum Similarity Metric for Machine Translation Evaluation[C]// Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, 2008: 55-62.

[6] Wang B, Zhao T, Yang M, et al. References Extension for the Automatic Evaluation of MT by Syntactic Hybridization[C]// Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation, 2009: 37-44.

[7] Kauchak D, Barzilay R. Paraphrasing for automatic evaluation[C]// Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2006: 455-462.

[8] Lavie M D A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language[J]. Proceedings of the Ninth Workshop on Statistical Machine Translation, 2014: 376-380.

[9] Snover M G, Madnani N, Dorr B, et al. TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate[J]. Machine Translation, 2009, 23(2-3): 117-127.

[10] 赵世奇,刘挺,李生. 复述技术研究[J]. 软件学报,2009, 20(8):2124-2137.

[11] 洪欢,王明文,万剑怡, 等. 基于迭代方法的多层 Markov 网络信息检索模型[J]. 中文信息学报,2013, 27(5):122-128.

[12] Zhou L, Lin C Y, Munteanu D S, et al. ParaEval: Using Paraphrases to Evaluate Summaries Automatically [C]// Proceedings of the Human Language Technology Conference of the NAACL, 2006: 447-454.

[13] 胡金铭,史晓东,苏劲松, 等. 引入复述技术的统计机器翻译研究综述[J]. 智能系统学报,2013, 8(3):199-207.

[14] 李莉,刘知远,孙茂松. 基于中英平行专利语料的短语复述自动抽取研究[J]. 中文信息学报,2013, 27(6):151-157.

[15] 苏晨,张玉洁,郭振，等. 使用源语言复述知识改善统计机器翻译性能[J]. 北京大学学报(自然科学版),2015, 51(2):342-348.

[16] Barzilay R, McKeown K R. Extracting Paraphrases from a Parallel Corpus[C]// Proceedings of 39th Annual Meeting of the Association for Computational Linguistics, 2001: 50-57.

[17] Bannard C, Callison-Burch C. Paraphrasing with Bilingual Parallel Corpora[C]// Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, 2005: 597-604.

[18] 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.

[19] Barzilay R, Lee L. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment[C]// Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2003: 16-23.

[20] Zhou L, Lin C Y, Hovy E. Re-evaluating Machine Translation Results with Paraphrase Support[C]// Proce--edings of the 2006 Conference on Empirical Methods in Natural Language Processing, 2006: 77-84.

[21] 甘丽新. 基于 Markov 概念的信息检索模型 [D]. 江西师范大学, 2007.

作者简介：翁贞（1991—），女，硕士研究生，主要研究方向为机器翻译。E-mail: wengzhen186@hotmail.com

李茂西（1977—），男，博士，副教授，主要研究方向为自然语言处理和机器翻译。E-mail: mosesli@jxnu.edu.cn

王明文（1964—），男，博士，教授，主要研究方向为信息检索、数据挖掘和机器学习。E-mail: mwwang@jxnu.edu.cn