北京大学学报(自然科学版) 第 53 卷 第 2 期 2017 年 3 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 53, No. 2 (Mar. 2017)
doi: 10.13209/j.0479-8023.2017.030

# 机器翻译自动评价中领域知识复述抽取研究

张丽林 李茂西 肖文艳 万剑怡 王明文†

江西师范大学计算机信息工程学院, 南昌 330022; † 通信作者, E-mail: mwwang@jxnu.edu.cn

**摘要** 针对通用领域语料中抽取的复述在特定领域机器译文自动评价任务的应用中容易出现复述匹配偏差的问题, 提出采用抽取与测试领域相关的复述来提高机器译文自动评价的方法。首先将通用单语训练语料进行聚类, 并利用改进的M-L方法过滤, 得到特定领域训练语料, 然后在训练语料中利用Markov网络模型, 抽取特定领域复述表, 最后将此复述表应用在机器译文自动评价中, 以提高同义词和近义词的匹配精度。在WMT'14 Metrics task和WMT'15 Metrics task数据集上的实验结果表明, 利用领域知识抽取的复述能够增加自动评价方法METEOR和TER与人工评价的相关性。

**关键词** 复述; 机器译文自动评价; 语言模型; Markov网络; 文档聚类

**中图分类号** TP391

# Improve Automatic Evaluation of Machine Translation Using Specific-Domain Paraphrase

ZHANG Lilin, LI Maoxi, XIAO Wenyan, WAN Jianyi, WANG Mingwen†

School of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022;
† Corresponding author, E-mail: mwwang@jxnu.edu.cn

**Abstract** Since the paraphrase extracted from the general domain tends to cause paraphrase match deviation in the specific-domain automatic evaluation of machine translation, this paper proposes an approach exploited specific-domain paraphrase related to the test set to enhance automatic evaluation of machine translation. First, the *K*-means algorithm is utilized to cluster general-domain monolingual corpus, and the specific-domain training data via improved M-L approach is obtained. Then, the specific-domain paraphrase table is extracted from the training data by Markov network model. Finally, the extracted paraphrase table is applied to automatic MT evaluation metrics to improve word match. The experimental results on the dataset of WMT'14 Metrics task and WMT'15 Metrics task show that the METEOR metric and the TER metric using the specific-domain paraphrase table yield better performance than that using the general-domain paraphrase table.

**Key words** paraphrase; automatic evaluation of machine translation; language model; Markov network; document clustering

近年来, 许多机器译文自动评价方法相继提出, 包括在机器翻译评测中广泛使用的 BLEU[1]、NIST[2]、METEOR[3], TER[4]以及近期李茂西等[5]和 Li 等[6]提出的机器译文自动评价方法。这些自动评价方法通过对比机器翻译系统的输出译文和人工参考译文来定量计算机器翻译系统的质量。

BLEU 和 NIST 自动评价方法在对比时, 假设只有词形完全相同的词语才表达同一种含义, 即机器译文中的词和人工参考译文中的词在词形完全相同时才认为两者匹配。然而, 由于语言表达的多样性, 同义词和近义词在自然语言中广泛存在, 传统的仅词形相同才表达同一含义的假设显然有其局限

国家自然科学基金(61462044, 61462045, 61662031, 61562042)、江西省自然科学基金(20151BAB207025)和江西省教育厅科技项目(GJJ150352)资助

收稿日期: 2016–07–23; 修回日期: 2016–09–23; 网络出版日期: 2016–11–30

性。针对此问题，METEOR 方法和 TER 方法在词形匹配的基础上增加了同义词和近义词匹配来提高机器翻译自动评价的性能。对于欧洲语言，同义词和近义词匹配信息可以从词语的词根、WordNet 同义词典和复述知识中搜索；对于其他语言，同义词和近义词匹配信息也可以从复述知识中获取。

为了构建有效的复述知识来匹配机器译文和人工参考译文中的同义词和近义词，一种典型的方法是从训练机器翻译模型的双语平行语料中抽取复述[7−8]。但是，双语平行语料不仅构建成本高，而且对于部分语料库较小的语言，难以大量获取复述。针对这个问题，翁贞等[9]提出从目标语言的单语文本中，利用 Markov 网络抽取复述知识来提高译文自动评级中同义词和近义词匹配的准确率。

尽管单语文本比较容易获取，并且语料规模一般较大，但是直接从中抽取复述，并将其应用到译文自动评价中时，存在一个突出的问题。由于不同的机器翻译任务所处的领域不同(比如国际评测 IWSLT 面向的是口语翻译，NIST 评测和 WMT 评测面向的是新闻领域文本的翻译，国内 CWMT 评测的部分任务则面向科技领域文本的翻译)，在自动评价不同领域机器翻译任务的多个机器翻译系统输出译文时，所需的复述知识因领域不同而有所区别，使用通用领域的复述知识对同义词和近义词匹配存在一定的偏差，甚至因为领域不同而引入额外的噪音，降低了匹配的精度。虽然使用同一翻译任务的训练语料来抽取复述可以减少领域不一致性，但是一方面难以对语料规模进行扩充，另一方面，对于自动评价任务的领域，完全由测试集的源语言句子(或人工参考译文)来决定，训练语料与测试集所处的领域并不完全吻合。

针对此问题，本文提出利用特定领域的复述知识来提高机器译文自动评价的方法。我们从目标语言的单语文本中过滤出与人工参考译文领域完全一致的语料，用于抽取复述表，并使用抽取后的复述表，来提高机器译文和人工参考译文中同义词和近义词匹配的准确率，进而增加自动评价方法与人工评价的相关性。

## 1　相关工作

在统计自然语言处理中，训练语料的规模和质量直接影响机器学习算法的效率。在统计机器翻译中，训练语料规模越大，与机器翻译的测试集所属的领域越一致，翻译系统输出译文的质量越高。为了扩充已有特定领域训练语料的规模，Moore 等[10]通过训练一个大规模通用语料和一个特定领域语料的语言模型，计算通用语料中每个句子在不同语言模型下的交叉熵之差，并从中提取一个子语料(该子语料规模远远大于已有的特定领域语料，且提取后的子语料与特定领域语料领域非常接近)，实现为统计机器翻译提供一个与目标领域一致且规模较大的训练语料的目标。与 Moore 等的方法不同，Axelrod 等[11]提出基于交叉熵的双语平行语料选择方法，分别计算双语平行语料中每个双语句对在通用领域和特定领域下语言模型的交叉熵之差，并对双语句对的两端交叉熵之差相加，从而提取出与目标领域一致的双语平行子语料。实验表明，该方法显著提高了口语机器翻译的性能。

复述是含义相同而表达方式不同的词、短语和句子，复述现象在自然语言中大量存在[12−15]。为了自动从大规模语料中抽取复述，Barzilay 等[16]提出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。Bannard 等[17]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。由于在他们方法中，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)，因此该方法也被称为枢轴法。不同于从双语语料中抽取复述的方法，Shinyama 等[18]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道了同一件新闻事件)。Barzilay 等[19]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。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。在特定领域复述知识构建方面，Pavlick 等[20]利用 Moore 等[10]的方法过滤训练语料，并利用枢轴法，从过滤后的训练语料中抽取目标领域的复述知识。

在译文自动评价中，为了匹配机器译文和人工参考译文中的同义词和近义词，复述信息得到广泛的使用。Snover 等[8]使用枢轴法，从双语平行语料中抽取复述，并利用抽取的复述知识增强 TER 自动评价方法。Denkowski 等[7]使用双语词汇化翻译概率知识，从双语平行语料中抽取复述，用于增强 METEOR 中近义词匹配。翁贞等[9]利用 Markov 网络模型，从单语文本中抽取复述提高机器译文自动

评价方法与人工评价的相关性。由于单语文本容易获取并且数量较多，本文在翁贞等[9]工作的基础上，通过进一步过滤通用语料得到与译文自动评价领域一致的语料来提高同义词和近义词匹配的精度。

## 2 基于单语语料的特定领域复述抽取方法

本文首先抽取与机器翻译测试集中人工参考译文相关的复述，然后将其应用在机器翻译自动评价方法上。其关键技术是在通用领域语料中过滤出与参考译文这一特定领域相似的子语料。通常情况下，与双语语料或可比语料相比，通用领域单语语料不仅构建成本低，而且容易获取。因此，我们从一个大规模的通用单语语料中过滤出与特定领域相关的子语料，并从中抽取复述，最后将抽取的复述应用在机器译文自动评价中。

### 2.1 语料过滤方法

我们采用并扩展了 Moore 等[10]提出的 M-L 语料过滤方法。M-L 语料过滤方法分别在通用领域语料和特定领域语料中训练不同的语言模型，并通过这两个语言模型，计算通用领域语料的每个句子语言模型概率，最后从大规模通用领域语料中抽取一个与特定领域相似的子语料。M-L 方法通过计算同一句子在不同语言模型下的交叉熵之差来进行语料过滤。为了增强机器译文自动评价方法，在特定领域语料选择上，我们选用人工参考译文。由于在机器译文自动评价中，通过对比机器翻译系统的输出译文与人工参考译文的相似度来定量计算机器翻译系统的质量，评测任务的领域完全由测试集的源语言句子(或人工参考译文)来决定，因此，我们将机器译文自动评价中每个子任务的人工参考译文作为特定领域的语料，分别训练通用领域语料和测试集人工参考译文的语言模型，通过计算同一句子的交叉熵之差来度量该句子与人工参考译文的相似性，过滤出与人工参考译文相关的子语料。计算公式如下:

$$\delta_i = H_{\text{ref}}(S_i) - H_{\text{gen}}(S_i), \tag{1}$$

其中，$\delta_i$ 是通用领域语料中第 $i$ 个句子得分，$H_{\text{ref}}$ 是人工参考译文语言模型下的交叉熵，$H_{\text{gen}}$ 是通用领域语料语言模型下的交叉熵。

### 2.2 复述抽取方法

过滤出特定领域的单语训练语料后，我们利用 Markov 网络模型，从其中抽取复述[9]。从构建好的词项 Markov 网络中抽取复述基于以下假设: 两个词项共同出现的词团越多，则这两个词项的语义越相似，并认为这两个词项互为复述。首先，我们利用词项在文档集中的共现性来计算词项之间的相关关系，构建一个词项 Markov 网络。利用顶点词项在文档集中的联合条件概率，计算网络中边的权重，即两个词之间的相关度。然后，给每个词项建立一个 $n$ 阶词团集合，计算两个词项共同出现的词团个数占所有出现这两个词项中任意一个词项的词团个数之和的比值，将这个比值视为这两个词项互为复述的概率，计算方法见式(2)~(4)。其中，式(4)为 $n$ 阶词团权重的计算方法。

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

$$W_3(t_i, t_j) = \sum_{k \neq i \wedge k \neq j \wedge t_k \in \text{clique}(t_i, t_j, t_k)} W_3(t_i, t_j, t_k), \tag{3}$$

$$W_n\{t_1, t_2, ..., t_n\} = \frac{\sum_{1 \leqslant i, j \leqslant n} R(t_i, t_j)}{\frac{1}{2}n(n-1)}。 \tag{4}$$

式(2)~(4)中，$\text{prob}(t_i, t_j)$为词项 $t_i$, $t_j$ 的复述概率，$W_3(t_i, t_j)$为同时包含词项 $t_i$ 和 $t_j$ 的所有三阶词团权重和，$W_3(t_i)$为包含词项 $t_i$ 的所有三阶词团权重和，$W_3(t_j)$表示包含词项 $t_j$ 的所有三阶词团权重和，$n$ 为词团中的节点个数，$t_i$, $t_j$ 和 $t_k$ 构成一个三阶词团，$R(t_i, t_j)$为词项 $t_i$ 和 $t_j$ 的相关性。

### 2.3 SD-Markov

在 Markov 网络模型中，我们以特定领域单语文档为单位抽取复述，简称 SD-Markov (extract the specific-domain paraphrase tables using Markov network)。与以句子为单位抽取复述相比，本文方法考虑了文档级的信息，并且以文档为单位进行词频统计，减少了数据的稀疏性，更有利于 Markov 网络的构建。

利用 Markov 网络模型自动抽取复述时，采用词的共现性计算词语间的关系，统计词共现频率时一般以整篇文档为单位[21]，而翁贞等[9]仅将一段连续固定长度的文本视为一篇文档进行统计计数，没有考虑到文档内部句子的相关性。为了引入文档的信息，我们把通用领域的单语语料划分成不同的文

档集，以文档为单位进行词频统计。利用哈希技巧(hashing trick)，将通用领域语料中的句子向量化，获取语料中每个句子对应的特征向量，然后利用 *K*-means 算法将领域接近的句子进行聚类，聚类后的同类句子汇总成一篇文档。

将通用领域语料拆分为不同文档后，利用 M-L 方法从聚类出的文档集中抽取与目标领域(即机器译文自动评价任务中的人工参考译文)接近的文档子集。在语料过滤过程中，不同于其他方法中以句子为最小单元，本文方法以一篇文档为最小过滤单元。在 M-L 方法中，通过比较句子的交叉熵之差，计算句子与目标领域的相似度，本文则通过比较文档的得分，衡量各文档与目标领域的相似度。利用 *K*-means 聚类算法，将一个大规模通用领域语料划分成文档集的过程中，每篇文档中包含的句子个数不相等，因此，我们对文档中每个句子的交叉熵之差相加后再求均值，得出每篇文档的得分:

$$\delta_{D_i}=\frac{\sum_{j=1}^{n}(H_{\text{ref}}(S_j)-H_{\text{gen}}(S_j))}{n}, \tag{5}$$

其中，$\delta_{D_i}$ 是第 $i$ 个文档的得分，$H_{\text{ref}}(S_j)$是文档 $D_i$ 中第 $j$ 个句子在参考译文训练出的语言模型下得出的交叉熵，$H_{\text{gen}}(S_j)$是文档 $D_i$ 中第 $j$ 个句子在通用领域语料中训练出的语言模型上的交叉熵，$n$ 是文档 $D_i$ 的句子数。

通过式(5)计算出文档得分 $\delta_{D_i}$，然后对所有分 $\delta_{D_i}$ 由小到大排序，得分越低表明文档与人工参考译文越相似，是需要提取的目标文档。通过改进 M-L 方法，本文在通用领域语料中过滤出与人工参考译文相似的文档集。

# 3　实验

为了比较基于单语语料的特定领域复述抽取方法与其他利用 Markov 网络抽取通用领域复述方法的性能，我们将抽取的复述表分别应用在机器译文自动评价开源工具 terp-v1[8]和 meteor-1.5[7]中，并在 WMT'14 Metrics task[22]和 WMT'15 Metrics task[23]上进行句子级别和系统级别的对比实验。

## 3.1　实验数据集

在 WMT'14 和 WMT'15 评测任务中，用目标端的单语语料来抽取复述，每个任务的人工参考译文作为特定领域语料。表 1 和 2 表示相关语料的统计数据。W14-corpus 表示 WMT'14 通用领域语料，分别选用 WMT'14 的 Europarl v7、Common Crawl corpus 和 News Crawl: articles from 2013。W15-corpus 表示 WMT'15 通用领域语料，分别选用 WMT'15 的 Europarl v7, Europarl v8, Common Crawl corpus 和 News Commentary v10。

在 Markov 网络的构建中，为了降低词频统计数据的稀疏性，本文方法将上述语料用 *K*-means 算法进行文本聚类，聚类后的文本称为文档，这些文档考虑了其内部句子之间的相关性，有利于 Markov 网络模型自动抽取复述时采用词的共现性来计算词语间的关系。

## 3.2　实验设置

本文将 W14-corpus 和 W15-corpus 聚类成文档，并组成新的通用领域文档集语料，使用 4-gram 语言模型和 Kneser-Ney 平滑方法，分别对不同翻译方向的通用领域文档集语料和特定领域语料训练相应的语言模型，计算不同语言模型下通用领域文档集语料每个句子的交叉熵之差，将聚类后文档内

**表 1　WMT'14 语料统计数据**

Table 1　Statistics of the WMT'14 Corpus

| 语料 | en-cs | en-de | en-fr | en-hi | en-ru | cs-en | de-en | fr-en | hi-en | ru-en |
|---|---|---|---|---|---|---|---|---|---|---|
| W14-corpus | 647 | 1920 | 2007 | 1084 | 878 | 2218 | 2218 | 2218 | 2218 | 2218 |

**表 2　WMT'15 语料统计数据**

Table 2　Statistics of the WMT'15 Corpus

| 语料 | en-cs | en-de | en-fr | en-fi | en-ru | cs-en | de-en | fr-en | fi-en | ru-en |
|---|---|---|---|---|---|---|---|---|---|---|
| W15-corpus | 1000 | 1920 | 2007 | 1926 | 1074 | 2218 | 2218 | 2218 | 2218 | 2218 |

句子的交叉熵之差相加并归一化。通过这种计分模型，计算每个文档对应的分值，分值越小越接近特定领域，将这些文档按得分从小到大排序，同时给定一个阈值，过滤出与特定领域相似的子语料。

为了确定阈值的大小，需要过滤出与特定领域相似的子语料。本文通过上述的排序模型，抽取排名前 $N$ 的句子，训练不同阈值下过滤语料的语言模型，然后计算它们与特定领域语料的困惑度。以德语为例，过滤出德语在 WMT'14 和 WMT'15 中排名前 $N$ 的文档，$N$ 的数值为 4000~36000 篇，计算过滤语料与相同的特定领域语料的困惑度，结果如图 1 所示。

从图 1 看出，在 WMT'14 Metrics task 任务中，使用本文方法最好的得分是选择阈值为 16000 篇文档，困惑度为 660.498；在 WMT'15 Metrics task 中，最好得分是选择阈值为 18000 篇文档，困惑度为 705.724。

提取出特定领域的单语语料后，利用 Markov 网络构建一个词项 Markov 网络模型，结合两个词团信息的相似性，计算这两个词互为复述的可能性。我们分别提取 WMT'14 和 WMT'15 多种语言的复述表，分别是捷克语、德语、法语、芬兰语、海地语、俄罗斯语和英语。根据 WMT'14 Metrics task 和 WMT'15 Metrics task 的 20 个任务，分别抽取 20 张对应的复述表。

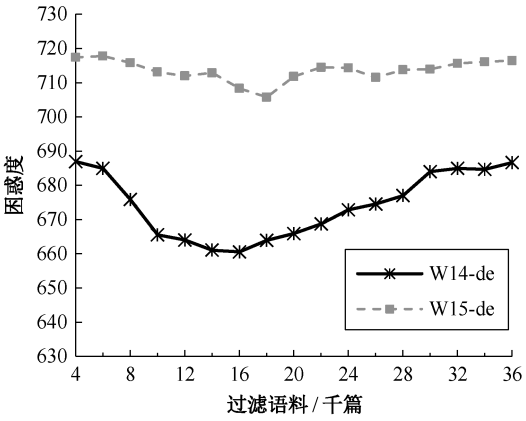


**图 1 数据选择结果**

Fig. 1 Results of the data selection

### 3.3 实验结果与分析

采用 Pearson 相关系数和 Kendall's $\tau$ 相关系数，分别计算自动评价结果与人工评价结果的系统级别相关性和句子级别相关性。表 3 和 4 分别给出不使用复述知识及分别使用不同复述知识的 METEOR 和 TER 自动评价方法，在 WMT'14 Metrics task 上与人工评价的句子级别和系统级别的相关性。表 5 和 6 给出它们在 WMT'15 Metrics task 上，与人工评价的句子级别和系统级别的相关性。"TER"和"METEOR" 表示 TER 和 METEOR 不使用复述知识进行同义词和近义词匹配；TER-Markov 和 METEOR-Markov 表示 TER 和 METEOR 使用基于

**表 3 使用特定领域复述和通用领域复述的 METEOR 和 TER 在 WMT'14 Metrics task 上评价英语译文的结果与人工评价的相关性**

Table 3 Correlation between automatic metrics METEOR and TER, which use specific-domain paraphrase or general domain paraphrase, and human judgments on into-English translation evaluation of WMT'14 Metrics task

| 评价级别 | 方法 | de-en | cs-en | fr-en | fi-en | ru-en | Average |
|---|---|---|---|---|---|---|---|
| 系统级别 | TER | 0.775 | 0.989 | 0.952 | 0.629 | 0.809 | 0.831 |
| | TER-Markov | 0.775 | 0.989 | 0.952 | 0.629 | 0.809 | 0.831 |
| | TER-SD-Markov | 0.784 | 0.989 | 0.955 | 0.629 | 0.802 | **0.832** |
| | METEOR | 0.885 | 0.952 | 0.971 | 0.515 | 0.789 | 0.822 |
| | METEOR-Markov | 0.913 | 0.955 | 0.971 | 0.488 | 0.804 | 0.826 |
| | METEOR-SD-Makov | 0.926 | 0.951 | 0.975 | 0.488 | 0.804 | **0.829** |
| 句子级别 | TER | 0.270 | 0.218 | 0.384 | 0.326 | 0.270 | 0.294 |
| | TER-Markov | 0.270 | 0.218 | 0.383 | 0.326 | 0.270 | 0.294 |
| | TER-SD-Makov | 0.295 | 0.233 | 0.392 | 0.342 | 0.281 | **0.308** |
| | METEOR | 0.302 | 0.253 | 0.397 | 0.378 | 0.297 | 0.325 |
| | METEOR-Markov | 0.325 | 0.272 | 0.399 | 0.400 | 0.313 | 0.342 |
| | METEOR-SD-Makov | 0.333 | 0.285 | 0.406 | 0.417 | 0.330 | **0.354** |

**表 4　使用特定领域复述和通用领域复述的 METEOR 和 TER 在 WMT'14 Metrics task 上评价英语到其他语言翻译的结果与人工评价的相关性**

Table 4　Correlation between automatic metrics METEOR and TER, which use specific-domain paraphrase or general domain paraphrase, and human judgments on out-of-English translation evaluation of WMT'14 Metrics task

| 评价级别 | 方法 | en-de | en-cs | en-fr | en-hi | en-ru | Average |
|---|---|---|---|---|---|---|---|
| 系统级别 | TER | 0.322 | 0.979 | 0.955 | 0.828 | 0.934 | 0.803 |
| | TER-Markov | 0.322 | 0.979 | 0.955 | 0.828 | 0.934 | 0.803 |
| | TER-SD-Makov | 0.337 | 0.976 | 0.954 | 0.828 | 0.934 | **0.806** |
| | METEOR | 0.240 | 0.979 | 0.939 | 0.924 | 0.932 | 0.803 |
| | METEOR-Markov | 0.226 | 0.979 | 0.939 | 0.924 | 0.913 | 0.796 |
| | METEOR-SD- Makov | 0.263 | 0.976 | 0.940 | 0.924 | 0.923 | **0.805** |
| 句子级别 | TER | 0.210 | 0.292 | 0.261 | 0.183 | 0.392 | 0.268 |
| | TER-Markov | 0.210 | 0.292 | 0.261 | 0.183 | 0.392 | 0.268 |
| | TER-SD-Makov | 0.217 | 0.292 | 0.270 | 0.183 | 0.392 | **0.271** |
| | METEOR | 0.212 | 0.310 | 0.274 | 0.303 | 0.407 | 0.301 |
| | METEOR-Markov | 0.222 | 0.310 | 0.275 | 0.303 | 0.422 | 0.306 |
| | METEOR-SD- Makov | 0.238 | 0.318 | 0.277 | 0.303 | 0.427 | **0.313*** |

注：* 表示此结果在 WMT'14 Metrics task 评价英语到其他语言翻译的结果和人工评价相关性的方法中排名第二，参见 http://www.statmt.org/wmt14/pdf/W14-3336.pdf。

**表 5　使用特定领域复述和通用领域复述的 METEOR 和 TER 在 WMT'15 Metrics task 上评价英语译文的结果与人工评价的相关性**

Table 5　Correlation between automatic metrics METEOR and TER, which use specific-domain paraphrase or general domain paraphrase, and human judgments on into-English translation evaluation of WMT'15 Metrics task

| 评级级别 | 方法 | de-en | cs-en | fr-en | fi-en | ru-en | Average |
|---|---|---|---|---|---|---|---|
| 系统级别 | TER | 0.890 | 0.914 | 0.980 | 0.878 | 0.910 | 0.914 |
| | TER-Markov | 0.888 | 0.926 | 0.977 | 0.885 | 0.912 | 0.918 |
| | TER-SD- Makov | 0.907 | 0.914 | 0.977 | 0.865 | 0.932 | **0.919** |
| | METEOR | 0.726 | 0.973 | 0.979 | 0.929 | 0.959 | 0.953 |
| | METEOR-Markov | 0.950 | 0.974 | 0.978 | 0.929 | 0.965 | 0.959 |
| | METEOR-SD-Makov | 0.959 | 0.974 | 0.979 | 0.939 | 0.963 | **0.963** |
| 句子级别 | TER | 0.362 | 0.391 | 0.359 | 0.278 | 0.330 | 0.348 |
| | TER-Markov | 0.358 | 0.394 | 0.357 | 0.297 | 0.333 | 0.348 |
| | TER-SD-Makov | 0.375 | 0.391 | 0.352 | 0.315 | 0.340 | **0.355** |
| | METEOR | 0.389 | 0.406 | 0.375 | 0.385 | 0.358 | 0.378 |
| | METEOR-Markov | 0.421 | 0.429 | 0.386 | 0.393 | 0.367 | 0.400 |
| | METEOR-SD-Makov | 0.431 | 0.434 | 0.376 | 0.404 | 0.383 | **0.406** |

Markov 网络模型提取的通用复述表进行同义词和近义词匹配; TER-SD-Markov 和 METEOR-SD-Markov 表示 TER 和 METEOR 使用本文方法提取的特定领域复述表进行同义词和近义词匹配。

从表 3 可以看出, TER-SD-Markov 和 METEOR-SD-Makov 机器译文自动评价方法在 WMT'14 Metrics task 目标语言是英语的评测任务上，与人工评级的句子级别和系统级别相关系数的均值分别高于 TER, TER-Markov 和 METEOR, METEOR-Markov。在系统级别相关性上，相应的均值提高幅度为 0.1%~0.7%；在句子级别相关性上，相应的均值提高幅度为 1.2%~2.9%。上述结果表明，特定领域的复述

**表 6 使用特定领域复述和通用领域复述的 METEOR 和 TER 在 WMT'15 Metrics task 上评价英语到其他语言翻译的结果与人工评价的相关性**

Table 6 Correlation between automatic metrics METEOR and TER, which use specific-domain paraphrase or general domain paraphrase, and human judgments on out-of-English translation evaluation of WMT'15 Metrics task

| 评价级别 | 方法 | en-de | en-cs | en-fr | en-fi | en-ru | Average |
|---|---|---|---|---|---|---|---|
| 系统级别 | TER | 0.557 | 0.918 | 0.946 | 0.617 | 0.890 | 0.786 |
| | TER-Markov | 0.557 | 0.916 | 0.946 | 0.616 | 0.890 | 0.785 |
| | TER-SD-Makov | 0.584 | 0.909 | 0.944 | 0.617 | 0.890 | **0.789** |
| | METEOR | 0.680 | 0.957 | 0.951 | 0.713 | 0.864 | 0.833 |
| | METEOR-Markov | 0.705 | 0.954 | 0.949 | 0.712 | 0.845 | 0.833 |
| | METEOR-SD-Makov | 0.735 | 0.938 | 0.955 | 0.714 | 0.851 | **0.839** |
| 句子级别 | TER | 0.289 | 0.358 | 0.326 | 0.215 | 0.357 | 0.309 |
| | TER-Markov | 0.289 | 0.358 | 0.326 | 0.216 | 0.357 | 0.309 |
| | TER-SD- Makov | 0.301 | 0.354 | 0.330 | 0.215 | 0.357 | **0.311** |
| | METEOR | 0.319 | 0.389 | 0.335 | 0.251 | 0.373 | 0.333 |
| | METEOR-Markov | 0.332 | 0.389 | 0.339 | 0.251 | 0.381 | 0.338 |
| | METEOR-SD-Makov | 0.342 | 0.385 | 0.341 | 0.251 | 0.381 | **0.340** |

知识不仅提高了机器译文和人工参考译文中的同义词与近义词的匹配, 而且在机器译文自动评价方法 METEOR 和 TER 上的性能比不使用复述知识和使用基于 Markov 网络模型提取的通用复述表效果好。

从表 4 可以看出, TER-SD-Markov 和 METEOR-SD-Makov 机器译文自动评价方法在 WMT'14 Metrics task 上, 评价英语到其他语言翻译的结果与人工评价的句子级别和系统级别相关系数的均值也分别高于 TER, TER-Markov 和 METEOR, METEOR-Markov。在系统级别相关性上, 相应的平均提高幅度为 0.3%~0.9%; 在句子级别相关性上, 相应的平均提高幅度为 0.3%~1.2%。METEOR-SD-Markov 方法与人工评价的句子级别相关性在 WMT'14 Metrics task 所有参赛方法中排名第二, 说明本文提出的特定领域的复述知识可以有效地提高 METEOR 和 TER 与人工评价的相关性。

从表 5 可以看出, TER-SD-Markov 和 METEOR-SD-Makov 机器译文自动评价方法在 WMT'15 Metrics task 目标语言是英语的评测任务上, 与人工评级的句子级别和系统级别相关系数的均值分别高于 TER, TER-Markov 和 METEOR, METEOR-Markov。在系统级别相关性上, 相应的平均提高幅度为 0.1%~1.0%; 在句子级别相关性上相应的平均提高幅度为 0.6%~0.8%。以上结果说明抽取单语语料的特定领域复述表在 METEOR 和 TER 上的性能比不使用复述知识和使用基于 Markov 网络模型提取的通用复述表效果好。

从表 6 可以看出, TER-SD-Markov 和 METEOR-SD-Makov 机器译文自动评价方法在 WMT'15 Metrics task 源语言是英语的评测任务上, 与人工评级的句子级别和系统级别相关系数的均值分别高于 TER, TER-Markov 和 METEOR, METEOR-Markov。在系统级别相关性上, 相应的平均提高幅度为 0.3%~0.6%; 在句子级别相关性上, 相应的平均提高幅度为 0.2%~0.7%。以上结果说明抽取单语语料的特定领域复述表可以有效地提高 METEOR 和 TER 与人工评价的相关性。

为了进一步定量说明本文方法抽取的特定领域复述比通用领域复述更能增强机器译文和人工参考译文的匹配程度, 我们从 Illiois.4083 翻译系统输出的 2600 句机器翻译中抽取前 300 句, 将其与相应的人工参考译文进行词语匹配人工标注。结果表明, 56%的词语可以通过词形进行匹配, 仅有 5%的词语需要进行复述匹配。在复述匹配部分, 我们比较了基于特定领域的复述抽取方法与通用领域的复述抽取方法在 METEOR 复述匹配时的准确率、召回率和 F1 值, 结果如表 7 所示。从表 7 可以看出, 使用特定领域的复述显著提高了复述匹配准确率、召回率和 F1 值。

**表 7　对 Illiois.4083 翻译系统前 300 句译文复述匹配的准确率、召回率和 F1 值**

Table 7　Precision, recall, and F1-measure for the paraphrase matching of the top 300 translations of the Illiois. 4083 translation system

| 方法 | 准确率/% | 召回率/% | F1 |
|---|---|---|---|
| METEOR-Markov | 55 | 65 | 0.63 |
| METEOR-SD-Makov | 63 | 82 | 0.80 |

通过比较不使用复述知识和分别使用不同复述知识的 METEOR 和 TER 自动评价方法在 WMT'14 Metrics task 和 WMT'15 Metrics task 上与人工评价的句子级别和系统级别的相关性，说明了领域复述知识的重要性。同时，通过对比不同的复述知识在 METEOR 中的匹配情况，解释了领域复述知识的有效性。实验结果表明，我们提出的方法优于实验比较的基线方法，说明特定领域的复述知识能够提高机器译文和人工参考译文中的同义词与近义词的匹配，进而增加自动评价方法与人工评价的相关性。

## 4　总结

针对通用领域语料中抽取的复述在特定领域机器译文自动评价任务的应用中容易导致复述匹配偏差的问题，本文提出利用特定领域复述知识，增强机器翻译自动评价中特定领域的同义词与近义词的匹配精度。我们将通用领域单语语料进行聚类，并过滤出特定领域语料，然后在过滤后的语料中抽取特定领域复述表，最后将抽取的复述表应用在机器译文自动评价中。实验结果表明本文方法提高了自动评价结果与人工评价结果的相关性，使用特定领域的复述知识显著地提高了特定领域同义词与近义词匹配的准确率和召回率。在最近官方公布的 WMT'2016 Metric task 评测结果中，我们提出的方法取得很好的成绩[24]。在将来的工作中，我们会尝试将特定领域的复述知识应用到信息检索、自动文摘和机器翻译等自然语言处理任务中。

## 参考文献


[1] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, 2002: 311–318

[2] Doddington G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics // Proceedings of the second international conference on Human Language Technology Research (HLT'02). San Diego, 2002: 138–145

[3] Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments // Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, 2005: 65–72

[4] Snover M, Dorr B, Schwartz R, et al. A study of translation edit rate with targeted human annotation // Proceedings of Association for Machine Translation in the Americas. Cambridge, 2006: 223–231

[5] 李茂西，江爱文，王明文. 基于 ListMLE 排序学习方法的机器译文自动评价研究. 中文信息学报，2013, 27(4): 22–29

[6] Li M, Wang M, Li H, et al. Modeling monolingual character alignment for automatic evaluation of Chinese translation. ACM Transactions on Asian and Low—Resource Language Information Processing, 2016, 15(3): 1–16

[7] Denkowski M, Lavie A. Meteor universal: language specific translation evaluation for any target language // Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT). Baltimore, 2014: 376–380

[8] Snover M, Madnani N, Dorr B, et al. TER-Plus: paraphrase, semantic, and alignment enhancements to translation edit rate. Machine Translation, 2009, 23(2): 117–127

[9] 翁贞，李茂西，王明文. 利用 Markov 网络抽取复述增强机器译文自动评价方法. 中文信息学报，2015, 29(5): 136–142

[10] Moore R C, Lewis W. Intelligent selection of language model training data // Proceedings of the ACL 2010 Conference. Uppsala, 2010: 220–224

[11] Axelrod A, He X, Gao J. Domain adaptation via pseudo in-domain data selection // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, 2011: 355–362

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

[13] 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. 中文信息学报，2013,

27(6): 151–157

[14] 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. 智能系统学报, 2013, 8(3): 199–207

[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. 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 // Proceedings of 39th Annual Meeting of the Association for Computational Linguistics. Toulouse, 2001: 50–57

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

[18] Shinyama Y, Sekine S, Sudo K. Automatic paraphrase acquisition from news articles // 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 // Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2003: 16–23

[20] Pavlick E, Ganitkevitch J, Chan T P, et al. Domain-specific paraphrase extraction // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, 2015: 57–62

[21] 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. 中文信息学报, 2013, 27(5): 122–128

[22] Bojar O, Buck C, Federmann C, et al. Findings of the 2014 workshop on statistical machine translation // Proceedings of the Ninth Workshop on Statistical Machine Translation. Baltimore, 2014: 12–58

[23] Bojar O, Chatterjee R, Federmann C, et al. Findings of the 2015 workshop on statistical machine translation // Proceedings of the Tenth Workshop on Statistical Machine Translation. Lisbon, 2015: 1–46

[24] Zhang L, Weng Z, Xiao W, et al. Extract domain-specific paraphrase from monolingual corpus for automatic evaluation of machine translation // Proceedings of the First Conference on Machine Translation. Berlin, 2016: 511–517