基于自动构建语料库的词汇级复述研究

赵世奇,刘 挺,李 生
(哈尔滨工业大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 本文针对词汇级复述问题提出了一种新的方法.该方法首先利用翻译引擎将双语平行语料库自动转换为单语平行语料库,以此构建复述语料库并用于候选复述的抽取.在此基础上,本文提出了一种新的统计模型.该模型根据特定的上下文为待复述词选择最为合适的复述.实验结果表明自动构建的复述语料库对于词汇级复述的抽取是有效的.同时,本文提出的模型明显优于两种传统模型,在准确率和召回率上分别提高 10 %左右.

关键词: 词汇级复述; 复述语料库; 复述模型
中图分类号: TP391 **文献标识码**: A **文章编号**: 0372-2112 (2009) 05-0975-06

Lexical Paraphrasing Based on Automatically Constructed Corpora

ZHAO Shi-qi, LIU Ting, LI Sheng
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*)

Abstract: This paper presents a new method for lexical paraphrasing. The method first constructs a paraphrase corpus by automatically translating a bilingual parallel corpus into a monolingual parallel corpus, from which candidate paraphrases for words are extracted. After that, a new statistical model is proposed for lexical paraphrasing, which selects the best paraphrase for a word in a given context sentence. Experimental results show that the automatically constructed paraphrase corpus is effective for lexical paraphrasing. In addition, the presented paraphrasing model significantly outperforms two conventional models, enhancing precision and recall by about 10 %, respectively.

Key words: lexical paraphrasing; paraphrase corpus; paraphrasing model

1 引言

复述是指对相同语义的不同表达,其研究在国外已开展多年,但在国内却相对较少[1~3].复述可分为词汇级,短语级和句子级.其中,词汇级复述(简称"词汇复述")旨在抽取词 w 在特定上下文句 S 中的复述.词汇复述在自然语言处理的诸多研究中都有应用.例如,在信息检索中,词汇复述可用于查询扩展以提高搜索引擎的召回率[4];在机器翻译中可用来解决语料稀疏问题[5];在自然语言生成中可帮助生成更加流畅生动的文本[6].词汇复述不等同于同义词,因为一个词的复述与其所在上下文紧密相关.两个同义词在特定上下文中可能不是复述,也不能相互替换.而两个非同义词却可能在特定句子中意思相同并可以相互替换.例如,"获得"和"捧走"并非同义词,但在句子"李安获得奥斯卡最佳导演奖"中却可被看作复述.另外,一个词的复述可能是词,也可能是短语.例如短语"到目前为止"便是词"至今"的复述.

现有的词汇复述研究集中在两方面,即复述语料库构建以及复述模型的研究.在复述语料库构建方面,诸如 WordNet 等类义词典被广泛应用[7,8].也有研究者尝试基于统计的方法从各种语料库中获取复述,比如从单语或双语平行语料库中获取复述资源[9,10].一般而言,从单语平行语料库中获取的复述更为准确,但规模和领域的局限性较大.与其相反,双语平行语料在规模和领域上的局限性较小,但从中抽取的复述准确率较低.

与此同时,研究者们提出了不同的复述模型[10,11].给定词 w,另一个词(或短语) w',以及上下文句 S,现有模型在识别 w 是否是 w 在 S 中的复述时通常考虑两个因素:一是 w 和 w' 的语义相似度,二是将 w' 替换到 S 中是否通顺合理.然而,两个词意思相似甚至相同并不意味着在特定上下文中是复述.例如,"逝世"和"毙命"是中文类义词典《同义词词林》(简称《词林》)[12]中的同义词,但它们在给定句子"著名作家巴金逝世."中并不能

收稿日期:2008-04-18;修回日期:2008-12-15
基金项目:国家自然科学基金(No. 60803093, No. 60675034);国家 863 高技术研究发展计划(No. 2008AA01Z144)

相互替换.类似地,词 w 在句子 S 中通顺也不意味着 w' 与 w 在该句中互为复述,如将句子"苹果是我最喜欢的水果."中的"苹果"替换成"香蕉",句子依然通顺,但替换生成的句子与原句子意思明显不同.

本文利用双语平行语料库和机器翻译引擎自动构建复述语料库.在此基础上为14000余词抽取了候选复述.进而本文提出一种新的词汇复述模型用于为每个待复述词根据特定上下文选择最优复述.实验表明,尽管自动构建的复述语料库含有噪音,但对于抽取词汇复述是有效的.另外,本文提出的模型在准确率和召回率上均明显优于两种传统模型.需要说明的是,本文研究的对象限定为动词,名词,形容词和副词.

2 相关研究

概括地讲,主要有三类复述资源,即类义词典,单语平行语料库及双语平行语料库.WordNet等类义词典被广泛用作词汇复述资源.其中有研究者将词典中的同义词看作复述[7],有的则放宽限制,将上位/下位词等也看作复述[8].但总的来说,究竟词典中定义的哪些词汇关系可看作复述却无定论[9].在中文复述研究中,有人使用《词林》作为资源抽取复述[13].

也有研究者基于统计方法从语料库中抽取复述,Barzilay和McKeown尝试从单语平行语料库中获取复述[9].他们利用同一部外文小说的不同英文译本作为平行语料并从中抽取复述.Pang等人也使用过类似方法[14].此类方法的一个显著缺陷是在文学之外的其他领域很难找到规模较大的平行译本,因此该方法在规模和领域上的局限性较大.另外,也有研究者利用对同一事件的不同新闻报道来获取复述[15,16].这类语料相对比较容易获得,但主题相同的不同新闻报道通常仅是相似文本,而非平行文本.因此从此类语料中抽取复述比较困难.

Bannard和Callison-Burch利用双语平行语料获取复述[10].该方法基于的基本假设是如果两个词或短语的外文翻译相同,这两个词或短语的意思便相同.这种方法的弊端在于两个意思不同的词也可能对应同一个外文翻译(比如该外文翻译是一个多义词).除上述资源之外,我们在之前的研究中也曾尝试过基于互联网抽取词汇级复述[17].

在复述模型的研究方面,Bannard和Callison-Burch基于双语平行语料提出了一种复述模型[10],该模型利用外文翻译作为"枢轴"来计算短语 e_2 是 e_1 的复述的概率 $P(e_2|e_1)$.具体地,设 f 是 e_2 和 e_1 共有的外文翻译,则该模型通过计算 $P(f|e_1)$ 和 $P(e_2|f)$ 的乘积来得到 $P(e_2|e_1)$.同时,该模型还结合语言模型来计算 e_2 出现在给定上下文中的概率.Zukerman等人提出的模型通过计算两个词基于词典的语义相似度来近似二者的复述概率.同时该模型基于"共现(co-occurrence)"计算 e_2 出现在给定上下文中的概率[11].

3 复述语料库自动构建

如上所述,已有的单语平行语料库在规模和领域上都有较大局限性.而若采用人工构建的方法,其过程又十分耗时耗力.本文利用一个双语平行语料库*和一个机器翻译引擎** 自动构建复述语料库.该双语语料库含有英文部分 E 和中文部分 C.本文利用翻译引擎将 E 自动翻译成中文 C',这样 C 和 C' 便构成了单语平行语料库(如图1所示).

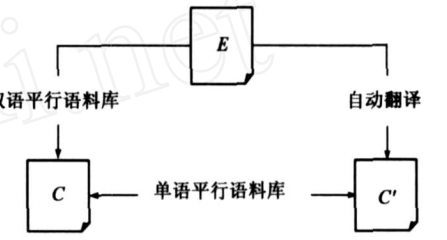

图1 利用双语平行语料库和机器翻译引擎自动构建单语平行语料库

这里,我们可以将机器翻译引擎看作是一个间接的复述引擎.由它将 C 复述成了 C'.尽管我们发现如此自动翻译得到的句子很多都存在翻译错误,但却可以用于词汇复述的抽取.下图展示的便是利用上述过程得到的平行句对以及其中的词汇复述.

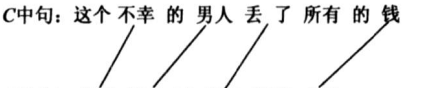

图2 自动构建得到的单语平行句对示例

这种自动构建的语料库结合了单语平行语料库和双语平行语料库的优点.一方面,作为一个单语平行语料库,我们可以直接从中抽取词汇复述,而不用像双语平行语料库那样利用另外一种语言作为枢轴.另一方面,该语料库和双语平行语料库一样其规模和领域的局限性较小.从后续实验中我们发现尽管该语料库中含有噪音,但对于抽取词汇复述却是有效的.

4 词汇复述获取

本文将词汇复述获取分为两步,即候选复述抽取和复述选择.其中,第一步将待复述词 w 的所有可能复述从上面构建的复述语料库中抽取出来作为候选.第二步则利用本文提出的模型根据给定的上下文句 S 对候选复述进行排序并选出 w 在 S 中的最优复述.

---

* 该语料库含有616402英-汉双语平行句对.

** http://translate.google.cn

4.1 候选复述抽取

复述可被看作单语机器翻译[18]. 因此机器翻译中诸如词对齐等技术可被应用到复述研究中来. 本文利用词对齐工具 Giza++[19]对自动构建的复述语料库进行词对齐,进而从中抽取候选复述. 我们将词对齐之后得到的所有的对齐对及其对齐概率抽取出来,并将对齐频数小于3的对齐对过滤掉,过滤之后剩下的对齐对便构成了候选复述. 按照上述方法,我们共抽取出14541个词的64329个候选复述. 例如,对于"充足"一词,我们抽取出的候选复述包括"足够","充裕","大量","充分","很多"等. 同时,对于每个词 w 与其候选复述 w',我们记录下了在语料库 C 的哪些句子中 w 与 w 对齐. 这些句子的重要性在于,我们可以认为是在这些句子中 w 可被复述为 w',从而统计学习其上下文特征.

4.2 用于复述选择的统计模型

本文提出的复述选择模型直接计算在给定上下文句 S 中 w 被复述成 w 的概率 $P(w\rightarrow w'|S)$. 其中,"$w\rightarrow w'$"表示将 w 复述成 w';进而,w 在 S 中的最优复述 $\hat{w}'$ 需满足:

$$\hat{w}'=\arg\max_{w'} P(w\rightarrow w'|S) \tag{1}$$

基于贝叶斯法则,上述公式可被改写为*:

$$\hat{w}'=\arg\max_{w'} P(w\rightarrow w')P(S|w\rightarrow w') \tag{2}$$

这里,$P(w\rightarrow w')$ 表示 w 在不考虑上下文的情况下复述成 w 的概率. $P(S|w\rightarrow w')$ 则是在 w 复述成 w 的情况下观察到 S 的概率. 为方便起见,我们将语料库中那些使 w 复述成 w 的上下文句的集合表示为 $Set(S_{w\rightarrow w})$. 显然,S 与该集合中的句子越相似,则在 S 中 w 被复述成 w 的可能性越大. 本文基于语言模型来度量 S 与 $Set(S_{w\rightarrow w})$ 中句子的相似性,即:

$$P(S|w\rightarrow w')\approx P(S|LM_{w\rightarrow w}) \tag{3}$$

其中,$LM_{w\rightarrow w}$ 表示利用 $Set(S_{w\rightarrow w})$ 中的句子训练得到的语言模型. 显然 $P(S|LM_{w\rightarrow w})$ 估计了语言模型 $LM_{w\rightarrow w}$ 生成 S 的概率. 将公式(3)代入公式(2),我们可得到:

$$\hat{w}'=\arg\max_{w'} P(w\rightarrow w')P(S|LM_{w\rightarrow w}) \tag{4}$$

我们使用上述公式进行候选复述的排序和选择. 这里,$P(w\rightarrow w')$ 通过极大似然估计得到:

$$P(w\rightarrow w')=\frac{c(w\rightarrow w')}{\sum_{w''}c(w\rightarrow w'')} \tag{5}$$

其中,分子为 w 被复述成 w 的次数,分母为 w 被复述成不同词的次数总和,这些均直接从 w 的候选复述集中统计得到. 另一方面,为计算概率 $P(S|LM_{w\rightarrow w})$,我们首先分别为 w 的每个候选复述 w 利用其上下文句训练一个语言模型**,并进而基于该语言模型计算当前给定的上下文句 S 的概率:

$$P(S|LM_{w\rightarrow w})=\prod_{w_i\in S}P(w_i|LM_{w\rightarrow w}) \tag{6}$$

这里,$P(w_i|LM_{w\rightarrow w})$ 利用下面公式计算得到:

$$P(w_i|LM_{w\rightarrow w})=\frac{c_{w\rightarrow w}(w_i)}{\sum_{w_j}c_{w\rightarrow w}(w_j)} \tag{7}$$

其中,分子表示词 w_i 出现在 w 复述成 w 所对应的上下文句中的次数,分母表示出现在 w 复述成 w 的上下文句中所有词的词数之和. 特别地,我们假设上下文句中那些与 w 距离较远的词对于决定 w 的复述没有作用,故而我们限定只有位于 w 的左右各3个词以内的上下文词被用于训练语言模型.

由于我们使用的语言模型数据稀疏问题比较严重,所以平滑处理十分必要. 换言之,就是要给"未见词"赋予一个概率值. 本问题的特殊性在于每个候选复述都对应一个语言模型,这样倘若分别对每个语言模型进行平滑的话,各个语言模型赋予未见词的概率将是不同的,甚至相差很大. 这对后续的复述排序和选择将有很大影响. 我们对此采取的策略是,在 w 的所有候选复述所对应的语言模型中,为未见词赋予一个相同的概率值:

$$P(w_i|LM_{w\rightarrow w})=\frac{1}{c_w(all)} \tag{8}$$

这里,w_i 表示一个未见词,$c_w(all)$ 表示 w 的所有候选复述对应的上下文句的词数总和. 该平滑方法会使一个语言模型中的概率和不等于1. 但该方法却可避免上面提到的问题. 另外,该方法还能保证当一个给定上下文句 S 中的所有上下文词在 w 的全部候选复述的语言模型中均未见时,$P(S|LM_{w\rightarrow w})$ 的值对各个候选复述将完全相同,这时将只根据 $P(w\rightarrow w')$ 对候选复述排序.

5 实验

5.1 对候选复述集的评测

利用第3节中介绍的方法,本文自动构建了一个复述语料库并进而从中抽取出了候选复述集. 该候选复述集含有14541个词的共计64329个候选复述(平均为每个词抽取4.4个候选复述). 如上所述,我们为每个词的候选复述记录了上下文句. 通过统计发现平均每个候选复述对应27.3个上下文句.

由于候选复述集是利用自动翻译和对齐得到的,因此有必要评测候选复述集的准确率. 具体地,给定词 w,上下文句 S 以及 w 在 S 中的一个候选复述 w',我们需要判断 w 是否是 w 在 S 中的正确复述. 本文将其表

---

* 分母 $P(S)$ 对于各候选复述是一个常数,因此省略.

** 这里我们采用一元语言模型.

示为三元组 < w, w', S > . 在本实验中,我们从候选复述集中随机抽取 1000 个三元组并通过人工标注来评价候选复述集的准确率. 本实验借鉴 Bannard 和 Callison-Burch 的评测方法[10],定义两种准确率 $P1$ 和 $P2$. 在 $P1$ 的定义里,只有 w 与 w 在 S 中意思相同且 w 在 S 中句法正确,该三元组才被判定为正例. 在 $P2$ 的定义里,只要 w 与 w 在 S 中意思相同,该三元组就被判定为正例. 标注者根据上述标准对 1000 个三元组进行标注,准确率如表 1 所示.

表 1 从候选复述集中随机抽取的 1000 个三元组的准确率

	$P1$	$P2$
1000 个三元组	54.9 %	60.6 %

另外,本实验还统计了在正确的三元组中有多少候选复述无法在《词林》中找到. 结果表明,在 606 个正确的三元组中(根据 $P2$),有 413 个(68 %)候选复述不是《词林》中定义的同义词. 这表明本方法可以抽取出许多非同义词的复述.

5.2 对复述方法的评测

从表 1 可以看出候选复述集中含有很多噪声,这主要是由自动翻译和对齐的错误导致的. 本文提出的复述模型正是基于这样的含噪声的候选复述集来进行复述的排序和选择. 本节将对基于上述候选复述集的复述方法进行评测. 为构建测试集,本实验从一个中文词表中随机选取了 1000 个中文词作为待测试词,并为每个词在上面提到的语料库 C 中随机选取一个上下文句.

本文提出的复述模型针对测试集中每个待复述词 w 及给定的上下文句 S 对 w 的所有候选复述进行排序并选出最优的复述 $\hat{w}'$. 因此我们得到 1000 个形如〈w, $\hat{w}'$, S〉的三元组. 这些三元组同样由标注者来判定 $\hat{w}'$ 是否是 w 在 S 中的正确复述. 这里采用的评测指标是准确率和召回率,其定义如下:设 C_T 为待测试词的数目, C_M 为本方法能够选出最优复述的测试词个数, C_1 为本方法选出的最优复述意思和句法均正确的测试词个数, C_2 为待测试方法选出的最优复述意思正确的测试词个数. 在此基础上我们定义两种准确率 $P1$, $P2$ 和两种召回率 $R1$, $R2$, 分别如下: $P1 = C_1/C_M$, $P2 = C_2/C_M$, $R1 = C_1/C_T$, $R2 = C_2/C_T$. 基于上述指标,我们对本文提出的复述模型进行了评测,结果如表 2 所示.

表 2 对本文提出的复述方法的评测

	$P1$	$P2$	$R1$	$R2$
本模型	56.2 %	62.1 %	56.2 %	62.1 %

从表 2 中可以看出,含有噪音的候选复述集并没有使得复述选择无法实现. 同时,这也说明了本文提出的复述模型在利用含有噪声的数据进行复述选择方面是有效的. 另外,对比表 1 和表 2 我们会发现,在测试集上的复述准确率和召回率比候选复述集还要略高. 通过分析我们认为这主要是因为在候选复述集上评测时我们是随机选取若干三元组进行标注评测;而在测试集上我们是为每个待测试词在给定上下文句中选出一个最优复述,因此后者的效果自然更好些.

5.3 与传统方法的比较

在传统的复述模型中复述概率被表示为 $P(S'|S)$,其中, S' 为将 S 中的 w 替换为 w' 所生成的句子. 因此最优复述应满足:

$$\hat{w}' = \arg\max_{w'} P(S'|S) = \arg\max_{w'} P(S|S')P(S') \quad (9)$$

这里, $P(S|S')$ 可被看作是 S' 到 S 的翻译概率,根据 IBM 模型 1,该概率可被简化为 $\prod_i P(w_i|w_i')$,其中, w_i 和 w_i' 分别是 S 和 S' 中的第 i 个词. 设待复述词 w 为 S 中的第 j 个词,由于 S 和 S' 的区别仅是将 w 变成 w',而其它的词对是相同的. 这些相同的词对对于候选复述的排序和选择均无影响,可忽略不计. 因此, $P(S|S')$ 可被近似为 $P(w|w')$,从而得到:

$$\hat{w}' = \arg\max_{w'} P(w|w')P(S') \quad (10)$$

上述公式即为基本的传统复述模型,不过在计算概率的方法上,各个具体的模型又有所不同. 这里我们将介绍两种有代表性的具体模型,并与本文提出的模型进行比较.

模型 1 模型 1 与 Zukerman 等提出的模型[11]基本相同. 该模型用 w 和 w' 基于词典的相似度来近似概率 $P(w|w')$. 在本实验中,我们基于本实验室的《同义词词林》(扩展版)计算词的相似度. 具体地,若 w 与 w' 是《词林》中定义的同义词,则 $P(w|w')=1$,否则 $P(w|w')=0$. 同 Zukerman 等使用的方法一样,模型 1 也是基于共现模型来计算概率 $P(S')$,即:

$$P(S') \propto \prod_{i=1}^{n} P(w'|w_i) \quad (11)$$

其中, w_i 为句子 S 中 w 的一个上下文词, n 为 S 中 w 的上下文词的个数. $P(w'|w_i)$ 的计算方式如下: $P(w'|w_i) = c(w', w_i)/c(w_i)$. 这里, $c(w', w_i)$ 为词 w' 和 w_i 的共现次数, $c(w_i)$ 为 w_i 的词频. $c(w', w_i)$ 和 $c(w_i)$ 均是利用双语语料库的中文部分 C 估计得到.

模型 2 模型 2 与 Bannard 和 Callison-Burch 提出的模型[10]基本相同. 在模型 2 中,复述概率 $P(w|w')$ 利用 w 和 w' 共有的英文翻译作为"枢轴"来计算:

$$P(w|w') = \sum_e P(w|e)P(e|w') \quad (12)$$

其中, e 表示 w 和 w' 共有的英文翻译. 概率 $P(w|e)$ 和 $P(e|w')$ 利用我们的双语平行语料库并基于极大似然估计计算得到. 在模型 2 中, $P(S')$ 基于二元语言模型

来计算,公式为:

$$P(S') = \prod_{i=1}^{n} P(w_i | w_{i-1}) \quad (13)$$

其中, $P(w_i | w_{i-1})$ 根据极大似然估计计算得到. 需要注意的是,模型 1 和模型 2 没有事先抽取候选复述. 在具体实现时,中文词表中的所有词均被作为词 w 的候选复述,进而再利用两个模型分别排序并选择最优复述. 通过上面两个模型的定义可以看出,对于词 w,若其在《词林》中没有同义词,则利用模型 1 将无法得到 w 的复述;同样,倘若 w 在语料库中没有与之对应相同英文翻译的词,则利用模型 2 无法得到 w 的复述.

我们利用 5.2 节介绍的测试词及上下文句对模型 1 和模型 2 进行了评测,即由这两种模型分别选出最优复述再由标注者进行人工标注,其结果如表 3 所示.

表 3　两种传统复述模型的实验结果

	$P1$	$P2$	$R1$	$R2$
模型 1	45.1 %	51.1 %	40.8 %	46.2 %
模型 2	50.0 %	53.2 %	46.3 %	49.3 %
本模型	56.2 %	62.1 %	56.2 %	62.1 %

从表 3 中可见,模型 1 和模型 2 的准确率和召回率均明显低于本文提出的模型. 我们认为有两点原因. 第一,在估计 $P(w | w')$ 时,模型 1 和模型 2 都没有直接估计两个词互为复述的概率. 相反,模型 1 计算两个词基于词典的相似度,模型 2 通过计算翻译概率来近似得到复述概率. 第二,在计算 $P(S')$ 时,无论是模型 1 采用的共现模型还是模型 2 采用的语言模型都旨在度量候选复述 w 出现在上下文句 S 中是否合适. 但 w 出现在 S 中是合适的并不意味着 w 在 S 中可以被复述成 w!

5.4　结果分析

根据对实验结果的观察我们发现,与两种传统模型相比,本文提出的模型可获取更多"不明显"的复述. 例如将"说实话"复述为"老实说",将"十足"复述为"不折不扣"等. 此外,本模型还可以将词复述为短语,例如将"国共"复述为"国民党和共产党". 我们认为,本文提出的方法之所以在效果上优于两种传统方法,一个重要的原因是本方法不仅利用了双语平行语料库所含的知识,而且利用了翻译引擎所含的知识. 这使得本方法可以抽取出更多更丰富的复述.

通过对错误复述结果的分析,我们发现导致错误的两个主要原因,即翻译/对齐错误以及数据稀疏. 如上所述,本文提出的模型基于含噪声的候选复述集,而该噪声是由自动翻译和对齐引入的. 本文虽已证明这种噪声不会使方法失效,但其的确会对复述效果造成影响. 我们对本方法获取的 379 个错误复述进行了分析,发现其中有 246 个是由翻译错误导致的,另有 68 个是由对齐错误导致的. 二者相加接近错误总数的 83 %. 例如,我们的方法将"董事会"错误地复述为"板子",因为翻译引擎在构建复述语料库时将一些句子中意为"董事会"的多义词"board"翻译成了"板子". 而由对齐错误带来的复述错误例如将"果树"复述为"种植果树"等. 为解决上述问题,我们将在未来工作中增加对候选复述集的过滤. 例如,我们可以基于词典层次结构计算词的词义距离[20],并以此过滤掉一部分翻译错误. 具体地,若 w 与其候选复述 w 的词义距离大于指定阈值则将 w 过滤掉.

6　结论

本文提出了一种面向词汇复述的新方法,其贡献主要包括以下两点:第一,本文利用双语平行语料库和翻译引擎自动构建了一个复述语料库,并进而从中抽取了候选复述集. 第二,本文提出了一种新的模型用于复述的排序和选择. 实验结果证明,尽管自动构建的复述语料库含有噪声,但对于抽取候选词汇复述却是有效的. 同时,本文提出的复述模型对于词汇复述的排序和选择也是行之有效的,在准确率和召回率方面均优于两种传统模型. 本方法的缺陷在于自动翻译和对齐的性能对本方法的效果影响较大. 此外,复述模型的数据稀疏问题比较明显. 在未来工作中,我们将尝试使用更大规模的语料库并采用基于词义的语言模型,从而减轻数据稀疏对复述效果的影响. 另外,在今后的工作中,我们会考虑使用多翻译引擎的策略,通过对多翻译引擎的翻译结果进行比较和统计来过滤噪声,从而提高复述效果.

参考文献:

[1] Zong C, Zhang Y, Yamamoto K, Sakamoto M, Shirai S. Approach to spoken Chinese paraphrasing based on feature extraction[A]. Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS)[C]. Tokyo, Japan, 2001. 551 - 556.

[2] 宗成庆,张玉洁,山本和英,坂本仁,白井谕. 面向口语翻译的汉语语句改写方法[J]. 汉语语言与计算学报,2002,12(1):63 - 77.
Zong C., Zhang Y., Yamamoto K., Sakamoto M., Shirai S. Chinese utterance paraphrasing for spoken language translation [J]. Journal of Chinese Language Computing, 2002, 12(1): 63 - 77. (in Chinese)

[3] 刘挺,李维刚,张宇,李生. 复述技术研究综述[J]. 中文信息学报,2006,40(4):25 - 33.
Liu T, Li W, Zhang Y, Li S. Survey on paraphrasing technology [J]. Journal of Chinese Information Processing, 2006, 40(4): 25 - 33. (in Chinese)

[4] Sparck Jones K, Tait J I. Automatic search term variant genera-

tion[J]. Journal of Documentation, 1984, 40(1): 50 - 66.

[5] Callison-Burch C, Koehn P, Osborne M. Improved statistical machine translation using paraphrases [A]. Proceedings of NAACL [C]. New York, US, 2006. 17 - 24.

[6] Iordanskaja L, Kittredge R, Polguere A. Lexical selection and paraphrase in a meaning-text generation model [J]. Natural Language Generation in Artificial Intelligence and Computational Linguistics, 1991. 293 - 312.

[7] Langkilde I, Knight K. Generation that exploits corpus-based statistical knowledge [A]. Proceedings of the COLING/ACL [C]. Montreal, Canada, 1998. 704 - 710.

[8] Barzilay R, Elhadad M. Using lexical chains for text summarization[A]. Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization [C]. Madrid, Spain, 1997. 10 - 17.

[9] Barzilay R, McKeown K R. Extracting paraphrases from a parallel corpus [A]. Proceedings of ACL [C]. Toulouse, France, 2001. 50 - 57.

[10] Bannard C, Callison-Burch C. Paraphrasing with bilingual parallel corpora[A]. Proceedings of ACL [C]. Ann Arbor, Michigan, US, 2005. 597 - 604.

[11] Zukerman I, George S, Wen Y. Lexical paraphrasing for document retrieval and node identification[A]. Proceedings of IWP [C]. Sapporo, Japan, 2003. 94 - 101.

[12] 梅家驹, 竺一鸣, 高蕴琦, 殷鸿翔. 同义词词林[M]. 上海: 上海辞书出版社, 1996.
Mei J, Zhu Y, Gao Y, Yin H. Tong Yi Ci Ci Lin[M]. Shanghai: Shanghai Lexicographical Publishing House, 1996. (in Chinese)

[13] Li W, Liu T, Zhang Y, Li S, He W. Automated generalization of phrasal paraphrases from the web[A]. Proceedings of IWP [C]. Jeju Island, South Korea, 2005. 49 - 56.

[14] Pang B, Knight K, Marcu D. Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences[A]. Proceedings of HLT-NAACL [C]. Edmonton, Canada, 2003. 181 - 188.

[15] Shinyama Y, Sekine S, Sudo K. Automatic paraphrase acquisition from news articles [A]. Proceedings of HLT [C]. San Diego, U. S., 2002. 40 - 46.

[16] Dolan B, Quirk C, Brockett C. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources[A]. Proceedings of COLING [C]. Geneva, Switzerland, 2004. 350 - 356.

[17] Zhao S, Liu T, Yuan X, Li S, Zhang Y. Automatic acquisition of context-specific lexical paraphrases[A]. Proceedings of IJCAI[C]. Hyderabad, India, 2007. 1789 - 1794.

[18] Quirk C., Brockett C., and Dolan B. Monolingual machine translation for paraphrase generation [A]. Proceedings of EMNLP[C]. Barcelona, Spain, 2004. 142 - 149.

[19] Och F J, Ney H. A systematic comparison of various statistical alignment models [J]. Computational Linguistics, 2003, 29 (1): 19 - 51.

[20] Lin D. An information-theoretic definition of similarity[A]. Proceedings of ICML [C]. San Francisco, U. S., 1998. 296 - 304.

作者简介:

赵世奇 男, 1981 年生于辽宁抚顺. 哈尔滨工业大学计算机科学与技术学院博士生. 研究方向为自然语言处理.
E-mail: zhaosq@ir.hit.edu.cn

刘 挺 男, 1972 年生于黑龙江哈尔滨. 哈尔滨工业大学计算机科学与技术学院教授, 博士生导师. 研究方向为自然语言处理、信息检索.