文章编号：1003-0077(2013)06-0151-07

基于中英平行专利语料的短语复述自动抽取研究

李 莉，刘知远，孙茂松

（清华大学 计算机系，智能技术与系统国家重点实验室；清华信息科学与技术国家实验室（筹），北京 100084）

摘 要：短语复述自动抽取是自然语言处理领域的重要研究课题之一，已广泛应用于信息检索、问答系统、文档分类等任务中。而专利语料作为人类知识和技术的载体，内容丰富，实现基于中英平行专利语料的短语复述自动抽取对于技术主题相关的自然语言处理任务的效果提升具有积极意义。该文利用基于统计机器翻译的短语复述抽取技术从中英平行专利语料中抽取短语复述，并利用基于组块分析的技术过滤短语复述抽取结果。而且，为了处理对齐错误和翻译歧义引起的短语复述抽取错误，我们利用分布相似度对短语复述抽取结果进行重排序。实验表明，基于统计机器翻译的短语复述抽取在中英文上准确率分别为 43.20％和 43.60％，而经过基于组块分析的过滤技术后准确率分别提升至 75.50％和 52.40％。同时，利用分布相似度的重排序算法也能够有效改进抽取效果。

关键词：

中图分类号：TP391　　**文献标识码**：A

Automatically Extracting Phrase-level Paraphrases from Chinese-English Parallel Patents

LI Li, LIU Zhiyuan, SUN Maosong

(State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: Automatically extracting phrase-level paraphrases is an important research task in natural language processing (NLP), which has been applied in applications such as information retrieval, query answering and document classification. Moreover, technique patents, as an important carrier of human knowledge and technology, contain abundant information. Hence, automatically extracting phrase-level paraphrases from Chinese-English parallel patents has a positive effect on NLP tasks about technology. In this paper, we aim to extract phrase-level paraphrases from Chinese-English parallel patents automatically using method based on statistical machine translation, and use chunk parsing technology for paraphrase verification. Moreover, to dispose the errors caused by translation ambiguity and bad word alignment, we use distributional similarity to re-rank the extracted phrase-level paraphrases. In experiments, we find that the method based on statistical machine translation gets a precision of 43.20％ on Chinese patents while 43.60％ on English patents for Top-500 results. Meanwhile, after verification with chunk parsing, the precisions are raised to 75.50％ and 52.40％, respectively. Moreover, the re-ranking based on distributional similarity also improves the performance significantly.

Key words: phrase-level paraphrase; statistical machine translation; chunk parsing; distributional similarity

收稿日期：2013-06-12　**定稿日期**：2013-07-15

基金项目：国家自然科学基金资助项目(61133012)；国家 863 计划资助项目(2012AA011102)

作者简介：李莉(1990—)，女，硕士，主要研究方向为社会计算；刘知远(1984—)，男，博士后，主要研究方向为关键词抽取、社会标签分析和社会计算；孙茂松(1962—)，男，博士，教授，博士生导师，主要研究方向为自然语言处理、信息检索和社会计算。

1 引言

专利语料是人类知识和技术的载体，信息量丰富，与专利语料相关的自然语言处理研究，包括长句分割[1]、语义分词[2-3]、翻译对获取[4]和分布相似度计算[5]等，已引起学术界和工业界的广泛关注。其中，短语复述自动抽取是自然语言处理领域的重要研究课题之一，目前已经被成功应用到信息检索、自动问答、信息抽取、自动文摘和机器翻译等多个自然语言处理研究领域[6]。因而，本文希望展开基于中英平行专利语料的短语复述自动抽取研究。

复述，其英文名称是 paraphrase，有些学者也将其翻译为改写，对应的名词解释是“解释，释义等”[6]。关于复述的具体定义，最早可追溯到 20 世纪 80 年代语言学家 De Beaugrande 等人曾给出的具体定义[7]。在自然语言处理领域，“复述”研究的主要是“短语以上，句子以下”的语言单元的同义现象[6]。Bazilay 等人根据研究的语言单元粒度，将复述具体分为词汇级、短语级和句子级三类[8]。本文重点关注短语级复述的自动抽取。

本文利用基于统计机器翻译的复述抽取技术[9]实现中英平行专利语料的短语复述自动抽取。该方法的基本思想是将对齐到同一目标语言短语的两个源语言短语视为互为短语复述。该方法的主要优点是基于目前互联网上大量存在的双语平行语料，可以同时实现双语短语复述自动抽取。例如，对于我们的中英平行专利语料，清华大学计算机系可以同时实现中文短语复述自动抽取和英文短语复述自动抽取。同时，该方法由于是基于短语的统计机器翻译模型的扩展，天然适用于短语级复述自动抽取任务。但是，该方法也存在以下两点不足之处。1）该方法依赖基于短语的统计机器翻译，但目前基于短语的统计机器翻译模型中的短语并不是语言学意义上的短语概念[10]，因而抽取的短语复述中存在大量非语言单元，例如，“network device is”和“网络设备为”等；2）该方法会受到对齐错误和翻译歧义的限制，经常无法区分短语复述的抽取质量[11]。

针对该方法的两点不足，我们分别引入基于组块分析的过滤技术和基于分布相似度的重排序技术来改进。组块（Chunk）是一种高于词序列，低于短语的语法结构[12]。组块分析即将输入句子中的所有词都划分到若干相应的组块中[13]。本文通过对中、英文专利语料分别进行组块分析，构建中、英专利组块表，并基于这两个组块表过滤短语复述结果中的非语言单元，提高短语复述抽取的准确率。而为了解决第二个问题，我们利用分布相似度对抽取的短语复述结果重排序。基于分布相似度进行复述抽取也是短语复述抽取的常用方法之一，基本思想是认为出现在相同或相似上下文的两个短语倾向于互为短语复述[14]。该方法借助大规模语料，可以较好区分短语复述的抽取质量，但是却容易将反义短语误判定为复述结果[15]。幸运的是，基于统计机器翻译的方法得到的候选结果中较少包含反义短语[11]。因而基于分布相似度对候选短语复述结果重排序，在解决基于统计机器翻译的方法无法区分短语复述的抽取质量的不足的同时，也回避了自身容易将反义短语误判定为复述结果的不足。所以，本文基于分布相似度对短语复述结果重排序，以改进基于统计机器翻译的方法经常无法区分短语复述的抽取质量的不足。

文章接下来的组织结构如下：第 2 节介绍算法设计；第 3 节介绍实验设计；第 4 节介绍实验结果，第 5 节介绍相关工作，最后进行总结。

2 短语复述自动抽取算法

本文基于中英平行专利语料实现短语复述自动抽取的算法流程如图 1 所示。首先借助基于统计机器翻译的短语复述抽取技术，实现短语复述候选结果抽取，然后利用组块分析技术过滤候选结果中的非语言单元。最后，基于分布相似度对过滤后的短语复述结果重排序，以解决对齐错误和翻译歧义引起的错误。下面逐一介绍各流程的基本思想。

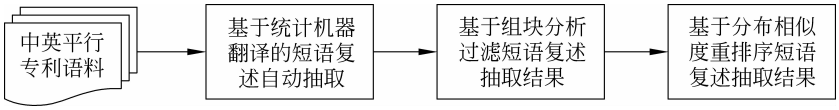

图 1 算法流程图

2.1　基于统计机器翻译的短语复述自动抽取

基于统计机器翻译的短语复述自动抽取技术是基于短语的统计机器翻译模型[16]的扩展，该方法的核心思想是将双语平行语料进行短语对齐后，将对齐到目标语言（如英语）下相同短语的源语言（如中文）下不同的短语视为互为短语复述。例如，图 2 所示的例子，在该思想的引导下会将“网络装置”和“网络设备”作为一对中文短语复述抽取出来。考虑到该方法包括短语对齐和复述抽取，下面依次介绍。

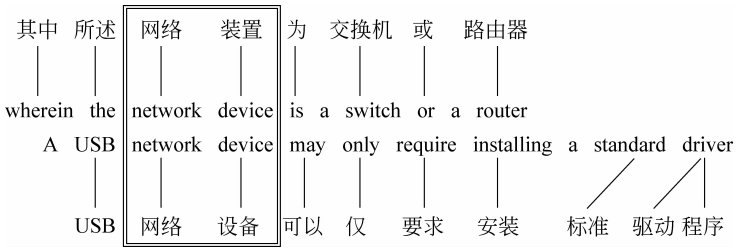

图 2　基于双语平行语料抽取短语复述

2.1.1　短语对齐

最早的统计机器翻译模型是基于词的[17]。在基于词的统计机器翻译模型下，一个源语言句子 e 翻译为目标语言句子 f 的翻译概率是通过将所有可能的词对齐（a，alignment 首字母）条件下句对的翻译概率叠加计算的。具体如公式(1)所示。

$$\mathrm{p}(f \mid e) = \sum_{a} p(f, a \mid e) \tag{1}$$

随着统计机器翻译的发展，句子之间的翻译概率逐渐使用更大的对齐文本块（如短语，这里的短语仅仅指词序列，而不是语言学意义上的短语）来进行计算，其中基于对齐短语计算句对翻译概率的统计机器翻译模型即为基于短语的统计机器翻译模型，也是目前主流的统计机器翻译模型之一。实际上，基于短语的统计机器翻译模型中的短语对齐也是基于词对齐实现的，其基本思想是递归地将句对中词对齐点相邻的词序列作为短语对齐结果[18]。例如，图 2 所示的例子，因为词对齐点相邻，会将“网络设备”和“network device”作为一对短语对齐结果。类似地，“网络设备为”和“network device is”也会被作为一对短语对齐结果。基于短语对齐结果，可以很方便地实现双语短语对抽取从而获得短语表以用于下面的复述抽取。

2.1.2　复述抽取

基于统计机器翻译的短语复述自动抽取技术本质上是将目标语言的短语作为锚点，找到短语对齐结果中与该锚点对齐的所有源语言短语作为候选复述结果。为了计算各个候选复述结果的概率值，Chris 等人引入 $p(e_2|e_1)$ 表示短语对 $<e_1, e_2>$ 互为复述的概率，引入 $p(e|f)$ 表示目标语言短语 f 翻译到源语言短语 e 的翻译概率，反过来即为 $p(f|e)$。之后 Chris 等人通过将所有锚点对应的翻译概率叠加计算短语对互为复述的概率，并挑选对应概率值最大的短语 e_2 作为 e_1 的短语复述结果抽出。具体计算公式如式(2)所示：

$$\begin{aligned} e_2 &= \arg \underset{e_2 \neq e_1}{\mathrm{MAX}} p(e_2 \mid e_1) \\ &= \arg \underset{e_2 \neq e1}{\mathrm{MAX}} \sum_{f} p(f \mid e_1) * p(e_2 \mid f) \end{aligned} \tag{2}$$

其中短语对翻译概率 $p(e|f)$ 基于最大似然概率计算，如式(3)所示。

$$p(e \mid f) = \frac{count(e, f)}{\sum_{e} count(e, f)} \tag{3}$$

引言部分提到，因为基于短语的统计机器翻译模型中的短语并不是源自任何深层次语言知识的短语概念，即并不一定是严格语言学意义的短语。例如，从图 2 例子得到的短语“网络设备为”和“network device is”，这样就导致了我们基于统计机器翻译进行短语复述抽取得到的候选结果中，包含着大量非语言单元。所以，我们期望通过基于组块分析的技术过滤其中的非语言单元。下面具体介绍。

2.2　基于组块分析过滤短语复述抽取结果

组块是一种语法结构，是符合一定语法功能的非递归短语，任何一种组块内部都不包含其他类型的组块[13]。例如，“网络设备”就是一个名词组块。而组块分析即将一段输入文本划分成一组互不重叠、非递归的组块片段[19]。

本文中，我们希望基于组块分析技术，限制抽取的短语复述结果是语言单元，而不仅仅是词序列。即我们期望通过中、英组块分析工具分别得到中文组块列表和英文组块列表，然后通过这两个列表对

上一步得到的短语复述结果进行确认，仅仅保留在这两个组块列表中出现的短语复述结果。

2.3 基于分布相似度重排序短语复述抽取结果

引言部分提到，基于统计机器翻译的短语复述抽取受限于对齐错误和翻译歧义，有时并不能很好的区分好、坏复述。特别在应用于中英平行专利语料时，翻译歧义造成的中文短语复述抽取错误比较明显。例如，因为“程序”和“节目”都被翻译为“program”，导致“程序类型”和“节目类型”被作为一对短语复述抽取出来。为了消除翻译歧义和对齐错误等造成的影响，我们基于分布相似度对上一步过滤后的短语复述抽取结果进行重排序以优化短语复述抽取结果。

基于分布相似度进行短语复述抽取的基本思想是将上下文分布相同或相似的短语作为短语复述结果抽出。目前比较常用的分布相似度计算方法包括词袋方法(bag-of-words approach)和上下文窗口(context window approach)方法等[20]。这两种方法的基本思想都是计算短语的上下文矩阵，然后通过余弦相似度计算对应上下文矩阵的相似度作为两个短语的分布相似度。其中，词袋方法对于每个短语 p，收集以 p 为中心的上下文窗口中的每个词，并将这些词的频度加入上下文矩阵。而上下文窗口方法则收集每个以 p 为中心的上下文窗口，并将这些上下文窗口的频度加入上下文矩阵。考虑到上下文窗口方法更适用于海量语料，而我们是基于已有的中英平行专利语料计算短语复述的分布相似度，所以我们选择词袋方法，并基于该分布相似度对短语复述抽取结果进行重排序。

3 实验设计

本文中，我们基于的中英平行专利语料包含 5 867 组中英平行专利文档，每组包括一篇中文专利文档和对应的英文专利文档。我们利用句子对齐工具[21]从中抽取中英平行句对 252 790 对，并从中随机抽取了 46 543 对句对进行人工检验，得到对齐准确率达到 98.4%。

下面我们按照图 1 所示的算法流程依次介绍我们具体的实验设计。

3.1 基于统计机器翻译的短语复述自动抽取

我们借助基于短语的统计机器翻译开源工具 Moses[22]实现短语对齐和双语短语对抽取，得到包含 9 933 939 对双语短语对的短语表。正如前文提到的，这一步我们抽取得到的短语表中大部分是无意义的词序列（其中大量词序列甚至包括标点符号），而不是实际语言学意义上的短语。这些词序列的存在不仅对于短语复述抽取无意义，而且增加了计算复杂度。所以在基于式(2)和式(3)进行短语复述抽取之前，我们基于简单规则对短语表进行过滤。我们定义的过滤规则如下：

1) 过滤短语表中包含中英标点符号的短语对；

2) 过滤短语表中在中文部分包含数字、英文字符的短语对；

3) 过滤短语表中在英文部分包含非英文字符的短语对；

4) 将中文部分相同，英文部分在忽略大小写时相同的短语对合并。

经过这一步简单的基于规则过滤，我们保留下来的短语表仅包含 2 850 237 对双语短语对，规模约为原来短语表的 28.69%。

之后，我们通过式(2)和式(3)，利用基于统计机器翻译的短语复述抽取技术，分别实现中文短语复述自动抽取和英文短语复述自动抽取。

3.2 基于组块分析技术过滤非语言单元

因为基于短语的统计机器翻译模型中的短语概念并不是实际语言学意义上的短语，更多的是无意义的词序列，而简单的基于规则过滤并不能保证保留下来的短语是语言学意义上的短语。所以，我们接下来基于组块分析技术过滤中英短语复述结果中的非语言单元。

在这一步骤中，我们使用 CRFTagger[23]对英文专利语料进行词性标注，使用 THULAC[24]对中文专利语料进行分词和词性标注。对英文专利语料我们借助开源工具 CRFChunk[25]进行组块分析，而对于中文专利语料，我们基于 CRF 模型，借助清华中文树库(Tsinghua Chinese Treebank)[26]训练了一个中文组块分析器，并在清华中文树库上检验了该分析器的效果，如表 1 所示。可以看到，我们设计的中文组块分析器在组块识别上 F1 值基本都在 85% 以上，效果较好。

我们通过中、英组块分析工具分别得到中文组块列表和英文组块列表，然后通过这两个列表对上一步得到的短语复述结果进行确认，仅仅保留在这两个组块列表中出现的短语复述结果。

表 1　中文组块分析器在清华中文树库评测效果

组块类型	Precision/%	Recall/%	F1-Measure/%
地点	84.71	84.48	84.59
形容词	88.87	84.72	86.75
名词	86.23	87.68	86.95
时间	80.19	74.65	77.32
介词	97.85	98.00	97.92
动词	91.22	94.64	92.90
副词	89.01	65.85	75.70

3.3　基于分布相似度重排序短语复述抽取结果

考虑到基于统计机器翻译的短语复述抽取受限于对齐错误和翻译歧义，有时并不能很好地区分短语复述质量。特别在应用于中英平行专利语料时，翻译歧义造成的中文短语复述抽取错误比较明显。所以我们基于现有的中英平行专利语料，借助分布相似度中的词袋方法对上一步过滤后的短语复述抽取结果进行重排序。在具体实验中，我们重点关注中文短语复述抽取，并对比分析了不同上下文窗口大小时重排序的效果。具体见实验结果部分。

4　实验结果

为了验证基于统计机器翻译的短语复述抽取，以及后面的两个改进策略（包括基于组块分析的过滤技术以及基于分布相似度的重排序技术）的效果。我们对排名前 500 的短语复述结果进行人工标注。以“E”标注对应的短语复述结果并不是语言学意义上的短语；以“N”标注虽然对应的短语复述结果是语言学意义上的短语，但是两个短语并不互为复述；以“Y”标注正确的短语复述结果。

我们分别统计了基于统计机器翻译的短语复述抽取和两个改进策略在前 500 个结果中的 Precision、Recall 和 F1 值。需要特别说明的是 Recall 值的计算，因为很难计算准确的 Recall 值，所以我们基于 Pooling[27] 方法。即我们将 3 个方法对应的前 500 个结果中的所有正确结果作为结果池（pool），然后基于这个结果池统计每个方法对应的 Recall 值。下面我们依次分析两个改进策略的表现。

4.1　基于组块分析过滤非语言单元

表 2 和表 3 分别展现了中文短语复述抽取结果的准确率和英文短语复述抽取结果的准确率。可以看到，基于统计机器翻译的短语复述抽取技术（表中简称为复述抽取）在中、英文上的准确率分别为 43.20% 和 43.60%，而经过基于组块分析过滤非语言单元（表中简称为组块过滤）后，准确率分别上升至 75.00% 和 52.40%，准确率均有大幅提升，由此验证了基于组块分析过滤非语言单元确实能够改进基于统计机器翻译的短语复述抽取效果。这一点也可以通过改进前后标注为“E”的结果数大幅减少看出。

表 2　中文短语复述抽取结果人工评测（前 500）

	#Y	#N	#E	Precision/%	Recall/%	F1-Measure/%
复述抽取	216	71	213	43.20	35.94	39.24
复述抽取＋组块过滤	375	117	8	75.00	62.40	68.12

表 3　英文短语复述抽取结果人工评测（前 500）

	#Y	#N	#E	Precision/%	Recall/%	F1-Measure/%
复述抽取	218	101	181	43.60	43.60	38.93
复述抽取＋组块过滤	282	99	139	52.40	42.26	46.79

但是，我们同时注意到，基于组块分析过滤非语言单元的改进策略在中文上的表现要优于英文。关于这点，我们经过分析数据发现，拼写错误以及英文语言环境中丰富的词性变化是错误率较高的原因之一。如“filer coefficients”和“filter coefficients”被作为一对短语复述抽取出来，但是其实前者是后者的错误拼写之一。再比如，虽然“alteration”和“modified”都是“改变”的意思，但是因为词性不同，并不能作为一对短语复述结果。而中文语言环境中几乎不存在这样的拼写错误、词性变化，所以基于组块分析过滤非语言单元的改进策略在中文上的表现要优于英文。

4.2 基于分布相似度重排序短语复述结果

考虑到该改进策略在英文短语复述抽取中的有效性已经被Chan等研究者证明[11]。下面我们重点分析该改进策略在中文短语复述抽取上的效果。

我们选用词袋方法(表中简记为Bow)对4.1改进后的前500个复述抽取结果进行重排序,并对比分析不同上下文窗口大小时该改进策略的效果。我们依次统计重排序前后短语复述抽取结果Top100、Top200、Top300、Top400的对应的Precision、Recall和F1值,并在图3中展示(重排序技术并不会影响Top500对应的准确率)。图3中Baseline是重排序前的准确率,而Bow_N是利用上下文窗口大小为N时的词袋方法重排序后的准确率。

从图3中Precision、Recall和F1值的对比中,我们可以明显看到利用分布相似度对短语复述结果重排序改进了抽取效果。同时,从Precision值的对比中可以较明显的看到上下文窗口大小较小时的改进效果要优于上下文窗口大小较大时的改进效果。我们猜测是因为当选定的上下文窗口大小较大时,模糊了不互为短语复述的两个短语的相似度差值。

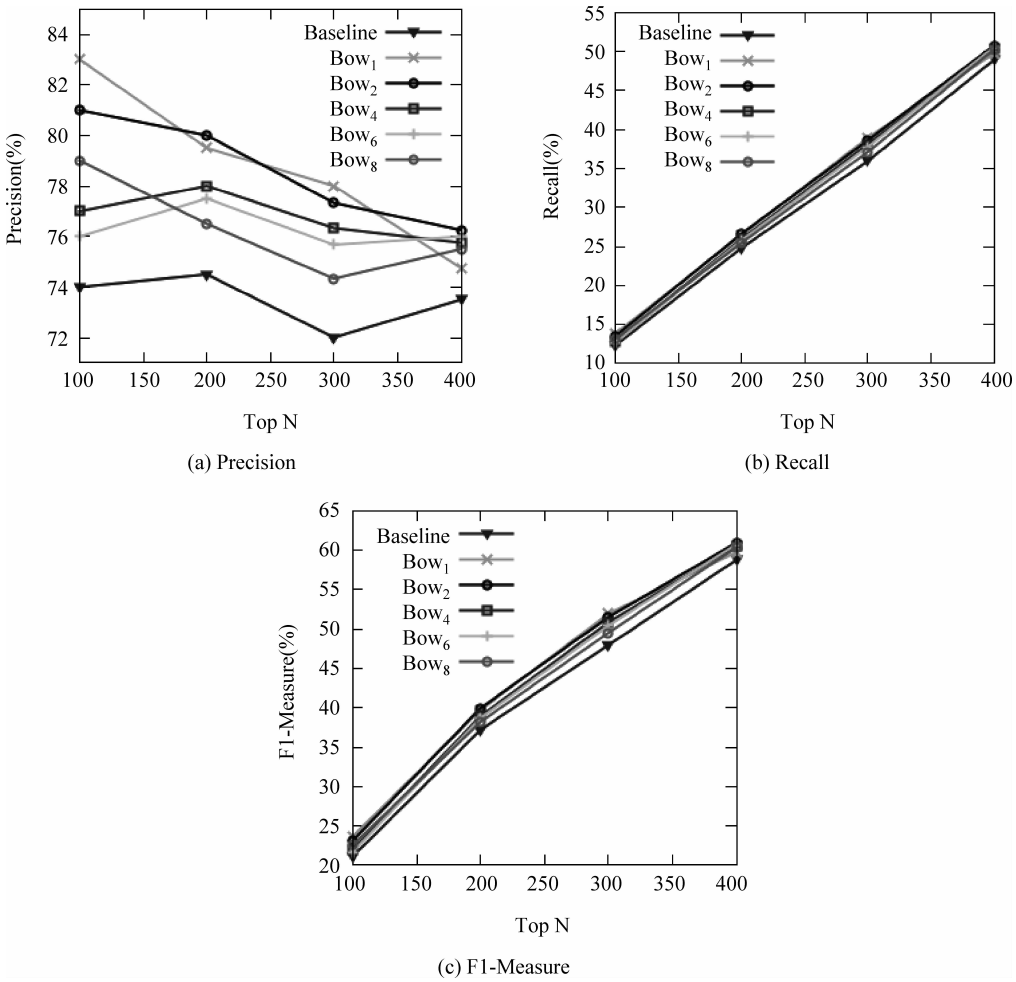

图3 不同上下文窗口大小下基于分布相似度重排序短语复述结果的效果

5 相关工作

目前复述抽取的相关工作主要分为四大类,包括基于已有语言标注资源(如WordNet[28]、HowNet[29]等)的复述抽取[30-32];基于分布相似度的复述抽取[33-34];基于译文语料的复述抽取[8,35]和基于统计机器翻译的复述抽取[9]。

基于现有语言学资源的复述抽取精度较高,但是受到现有语言学资源的规模、主题甚至语言的限制,而且比较适用于词汇级复述自动抽取任务。

基于分布相似度的复述抽取利用了目前海量的互联网语料资源,计算方便直观,但是缺点是容易将反义词或反义短语作为复述结果抽取出来。

基于译文语料的复述抽取借助目前网络存在的关于具体文章的多种译文版本，通过句子对齐工具构建复述抽取语料，之后借助上下文模版实现复述自动抽取。该方法精度较高，同时适用于词汇级、短语级和句子级复述自动抽取，但是受限于译文资源较少，对于缺乏多版本译文资源的专利领域难以开展。

基于统计机器翻译的复述抽取基本思想是将目标语言的短语视为锚，将对齐到同一锚的两个源语言短语视为短语复述结果。该方法依赖基于短语的统计机器翻译模型中的短语对齐和短语对抽取技术，优点包括 1）对于每个短语都提供了可能的复述列表并且包含相应的概率值，这样在具体的任务中，用户可以根据具体上下文选择最合适的短语复述；2）该方法天然适用于短语级复述自动抽取；3）该方法适用于任何双语或多语平行语料，对于语料资源限制极少。但是，该方法也有着自己的不足之处，包括以下两点：1）其中的短语概念并不是严格语言学意义上的短语概念，大部分是无意义的词序列，这样就导致抽取的短语复述中存在大量非语言单元；2）该方法因为受限于对齐错误和翻译歧义，因此有些时候并不能很好地区分好、坏短语复述。对于第二点不足，Chan 等研究者尝试利用分布相似度对复述抽取结果重排序，并在英文短语复述抽取中验证了有效性[11]。

6　结论

本文利用基于统计机器翻译的短语复述抽取技术从中英平行专利语料中抽取短语复述，并利用基于组块分析的技术过滤短语复述抽取结果。而且，为了处理对齐错误和翻译歧义引起的短语复述抽取错误，我们利用分布相似度对过滤后的短语复述结果进行重排序。实验表明，基于统计机器翻译的短语复述抽取在中英文上准确率分别为 43.20％和 43.60％，而经过基于组块分析的过滤技术后准确率分别提升至 75.50％和 52.40％。同时，利用分布相似度的重排序算法也能够有效改进中文短语复述抽取效果。

参考文献

[1] 张西龙，季铎，王岩，等. 英汉专利语料中长句的分割[J]. 沈阳航空航天大学学报. 2011，28(5)：67-70.

[2] 张桂平，刘东生，尹宝生，等. 面向专利文献的中文分词技术的研究[J]. 中文信息学报. 2010，24(3)：112-116.

[3] 岳金媛，徐金安，张玉洁. 面向专利文献的汉语分词技术研究[J]. 北京大学学报：自然科学版. 2013(1)：159-164.

[4] 刘颖，铁铮，余畅. 汉英短语翻译对的自动抽取[J]. 计算机应用与软件. 2012，29(7)：69-72.

[5] 郭丽. 基于上下文的词语相似度计算及其应用 [D][D]. 沈阳航空工业学院，2009.

[6] 刘挺，李维刚，张宇，等. 复述技术研究综述[J]. 中文信息学报，2006，20(4)：25-33.

[7] De Beaugrande R，Dressler W. Introduction to text linguistics[Z]. London：Longman，1981.

[8] Bazilay R，Mckeown K R. Extracting paraphrases from a parallel corpus[C]//2001.

[9] Bannard C，Callison-Burch C. Paraphrasing with bilingual parallel corpora[C]//2005.

[10] 宗成庆，张宵军. 统计机器翻译[M]. 电子工业出版社，2012.

[11] Chan T P，Callison-Burch C，Van Durme B. Reranking bilingually extracted paraphrases using monolingual distributional similarity[C]. 2011.

[12] 周强，孙茂松，黄昌宁. 汉语句子的组块分析体系[J]. 计算机学报. 1999，22(11)：1158-1165.

[13] 徐中一，胡谦，刘磊. 基于 CRF 的中文组块分析[J]. 吉林大学学报：理学版. 2007，45(3)：416-420.

[14] Katz J J. The philosophy of linguistics[M]. Oxford University Press，1985.

[15] Lin D，Pantel P. Discovery of inference rules for question-answering[J]. Natural Language Engineering. 2001，7(4)：343-360.

[16] Koehn P，Och F J，Marcu D. Statistical phrase-based translation[C]. 2003.

[17] Brown P F，Pietra V J D，Pietra S A D，et al. The mathematics of statistical machine translation：Parameter estimation [J]. Computational linguistics. 1993，19(2)：263-311.

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

[19] 李珩，朱靖波，姚天顺. 基于 SVM 的中文组块分析[J]. 中文信息学报. 2004，18(2)：1-7.

[20] Agirre E，Alfonseca E，Hall K，et al. A study on similarity and relatedness using distributional and wordnet-based approaches[C]. 2009.

[21] Li P，Sun M，Xue P. Fast-Champollion：A Fast and Robust Sentence Alignment Algorithm[C]//Proceedings of Beijing，China：Coling 2010 Organizing Committee，2010.

（下转第 174 页）