

基于边界感知的复杂名词短语的识别和转换研究

刘小蝶
北京联合大学 北京 100101

摘要 为了改善专利机器翻译中复杂名词短语的翻译效果,提出了一种基于规则的复杂名词短语识别和转换方法。通过分析汉英复杂名词短语的语义块和组合单元,利用边界感知策略,抽取汉语语言特征词,为汉语复杂名词短语中组合单元边界识别编制了 57 条识别规则,设计了合并策略,得到汉语复杂名词短语的形式化结构。通过对比汉英复杂名词短语的差异,确定了汉英复杂名词短语的转换策略。最后,将识别规则、合并策略和转换策略应用到一个机器翻译系统中。测试结果表明,所提方法可以有效地实现复杂名词短语的识别和转换,提高专利文本中复杂名词短语的机器翻译效果。

关键词 名词短语;边界感知;识别;转换;机器翻译;规则;专利

中图法分类号 TP391

Recognition and Transformation for Complex Noun Phrases Based on Boundary Perception

LIU Xiao-die
Beijing Union University, Beijing 100101, China

Abstract This paper proposes a rules-based method for recognizing and transforming the complex Noun Phrases to improve the translation quality of them in patent machine translation. By analyzing the semantic chunks and the structural units of Chinese and English complex Noun Phrases, under the guide of the boundary perception, this paper extracts the feature words, builds 57 recognition rules, designs combination strategies and realizes the formalization of Chinese complex Noun Phrases. By comparing Chinese and English complex Noun Phrases, this paper summarizes the differences between them, and determines the transformation strategies based on that. At last, it applies the method to an existing machine translation system to test our work. Experimental results show that our rules and strategy are very efficient, and improve the translation quality in patent machine translation.

Keywords Noun phrase, Boundary perception, Recognize, Transform, Machine translation, Rules, Patent

1 概述

为了应对不断增长的中国专利文献带来的日益繁重的翻译压力,对专利文献的自动翻译已经成为机器翻译中一个重要的应用领域。

在专利领域,书面语句子冗长而且结构复杂,表现之一是句子名词短语可以通过修饰关系来扩展长度。本文的复杂名词短语指的是有一个或多个名词短语、名词、介宾短语、方位短语、时间短语等做修饰语的名词性短语。本文的复杂名词短语可与要素句蜕^[1]一起,与最长名词短语(简称 MNP)的所指类似。本文的复杂名词短语的范围与多元逻辑组合中串联结构^[2]的所指类似。

例 1 这个在本发明的实施例中剪式齿轮 18 于开启和关闭状态的上部立视图

参考译文: this upper elevation view of scissors gear 18 in an open and closed position respectively in an embodiment of the present invention

在例 1 中这个复杂名词短语共含汉语词语 19 个、汉字 30 个,其中心语是名词性短语“这个上部立视图”,修饰语是名词性短语“剪式齿轮 18”“本发明”、名词“实施例”和介宾短语“于开启和关闭状态”,修饰语由“的”连接,体现出复杂名词

性短语复杂性。其中,中心语“这个上部立视图”并不是线性序列,而是被其他名词性短语和介宾短语隔开,从而体现出复杂名词性短语层级性和复杂性。

复杂名词短语的机器翻译难度较高,首先,复杂名词短语本身比较复杂,修饰语可以是名词性短语、名词、介宾短语、方位短语、时间短语中的一个或多个,而且这些介宾短语、方位短语中的修饰关系亦可具有层级性。其次,复杂名词短语数量多,在专利文本中,这种语言现象非常普遍,在笔者随机抽取的 1000 个专利句子中,含有复杂名词短语的句子有 449 个,约占到句子总数的 44.90%,复杂名词短语的总数是 672 个。可以说,复杂名词短语的翻译效果将直接影响到机器翻译的整体效果。

2 相关研究

文献[2]指出复杂名词短语研究的核心问题是组合单元的确定、语言联结词的选用或习用、组合顺序及组合的优先级等几个问题。文献[3]分析了复杂名词短语的类别,探索了其各成分优先组合和语序排列的规律,提供了从局部对其进行构成分析的线索。文献[4]以《毛泽东选集》为语料对汉英复杂名词短语的转换做了较为细致的探索研究。文献[5]提出了汉英复杂名词短语变换的研究重点以及大致的处理步骤等。

基金项目:国家自然科学基金(71974095)

This work was supported by the National Natural Science Foundation of China(71974095).

通信作者:刘小蝶(liuxiaodie@buu.edu.cn)

这些研究成果主要给出了初步的处理策略和步骤,点明了相应的原则,而对一些具体的问题实未过多详述,特别是针对复杂短语变换的系统分析和研究还并未全面展开。

文献[6]从词出发“从小到大”用产生式规则对名词短语(包括 MNP)的构成进行了形式化分析。文献[7]利用宾州大学汉语树库,从树库中进行组块的抽取。其中定义的名词组块大致相当于基本名词短语(BaseNP)。文献[8]利用词的信息、词性信息及上下文句法信息,提出一种混合的汉语基本名词短语(BaseNP)识别模型。文献[9]在分析简单名词短语与其他类型名词短语的异同的基础上,利用词、词性和词义信息对汉语简单名词短语进行识别。文献[10]运用半监督和主动学习相结合的方法识别包含至少 3 个词语且不含助词“的”的复杂名词短语。

本文的研究内容是是复杂名词短语内部结构的形式化问题和汉英转换问题。这对研究复杂名词短语内部结构化分析具有一定的启发。但目前的研究成果对模型或方法应用领域的适用性要求较高,有些论文(如文献[10])中所识别的“BaseNP”只是 BaseNP 的一个部分,从而降低了识别难度,且系统对上下文句法特征不明显的 BaseNP 识别的精确率较低,识别效果并不理想。

3 汉语复杂名词短语分析

3.1 汉英复杂名词短语的语义块特征

根据笔者对汉英专利文本句子中完整的名词性短语(句子的主语或宾语)的对齐语料的考查,汉英两种语言之间存在意义上等价的语义块,这是汉英复杂名词短语得以翻译的基础和前提。这些语义块的主要特征是:1)在词语序列上是连续的,是一组词串;2)词语数量不定,可以是一个词,也可以是几个词;3)如果是几个词,具有结构性,则一定是短语,而非一组随机词串;4)基本上是最小的意义单位,汉语和英语是等价的;5)主要由虚词来连接,汉英语义块连接词对应表如表 1 所列。

表 1 汉英语义块连接词对应表(部分)

Table 1 Conjunction contrast table for Chinese-English semantic chunks(part)			
类别	语义关系	汉语词语	英语词语
介词	调序说明符	的	of
	不调序说明符	的	
	单向关系	对、与、从、由、自	to, and, from
	引入对象	关于、对于	for, about
	手段	通过、采用、利用、自……	by, with
	标准说明符	如、参照、按照、作为……	as
	条件	在、于	at
介词	被动说明符	被	过去分词
方位词	方位	上、下、中、后、之间、内、里、外	on, under, in, after, between, among at, inside, outside
副词	进行说明符	正在、正、在	现在分词
	完成说明符	已经、已、后的	过去分词
	否定说明符	未、没有	否定前缀
代词	指代	等	such as

需要注意的是,汉语没有英语中的过去式(-ed)、过去分词(-ing)、否定前缀、形容词后缀等形态变化,但专利中汉语复杂名词短语中有一批具有翻译风格的“的”字组合,如“……后的”“已经……的”,例句如“清洗后的容器、已经清洗的容器(the coated vessel)”。

3.2 汉语复杂名词短语的构成分析

本文对专利文本中复杂名词短语进行汉英方向的翻译,基于这个实际,本文以汉语复杂名词短语为立足点对复杂名词短语语料的考查。

我们利用文献[2]组合单元这个术语来分析汉语复杂名词短语的构成。组合单元是汉语复杂名词短语的直接构成单位,可以根据汉语连接词和语义块的关系来确定组合单元的范围。根据汉语连接词语的特点:1)汉语中存在一批依附关系较强与英文有明显对应的连接词。这些连接词,一方面在结构上与语义块形成了汉语中普遍存在的介宾短语、方位短语、时间短语等,本文将这些短语统称为辅块;另一方面,在意义上,这些连接词具有辅助提示修饰语类的语义块与中心语类语义块的意义关系。2)两个成对出现的汉语连接词对应一个英文连接词,这可以明确提示汉语复杂名词短语的层级性。

组合单元的类型有 3 种:1)一个语义块(用 BK 表示),其包含动词性短语、名词性短语,其中动词性短语必须是中心语;2)一个或一对连接词和一个语义块的组合,即辅块(用 FK 表示);3)一个语义块的修饰部分,指的是被辅块隔离开的语义块的修饰部分,本文称为块饰(用 KU 表示)。这些被辅块分离出去的词语:1)处于短语的最左侧;2)一般由代词和动词以及表示数量、范围的词或短语充当。如“上述关于第一实施方案的内容”,该复杂名词短语是一个二元复杂名词短语,组合单元由一个辅块“关于第一实施方案”和一个名词性短语“上述内容”构成,其中“上述”是一个块饰。

我们在考查语料的基础上,从修饰语的类别上对复杂名词短语的构成进行举例式说明,如表 2 所列。

表 2 组合单元的类别及示例
Table 2 Types and examples of Chinese structural units

类别	示例
块饰	上述关于第一实施方案的内容
	改善的对烃燃料的抗渗透性
	一个或多个基于网络的服务器 1-14
名词性短语	大部分常规的交通工具布线系统的配置
动词性短语	子帧之间的不连续
辅块	方位短语
	所有网络节点之间的持续通讯
	介宾短语
	作为湿润剂的一种或多种醇类
时间短语	间歇运转模式恢复后的化学计量比
	其他
	塑料化合物等的复合树脂材料

根据组合单元的个数,可以确定汉语复杂名词短语主要包括二元复杂名词短语、三元复杂名词短语、四元复杂名词短语、五元复杂名词短语。

(1)语义块作修饰语

这类复杂名词短语的类别及示例如表 3 所示。其中 MK 分割了不同的语义块,一般由汉语连接词“的”承担,在实际文本中也存在隐现的“的”,即无“的”分割的情况。

表 3 语义块作修饰语的复杂名词短语类别及示例
Table 3 Types and examples for complex NPs which semantic chunks are used as modifiers

类别	线性结构	示例
二元	BK1 MK BK2	拖尾校正电路 6 的详细配置
三元	BK1 MK BK2 MK BK3	示例(6)的未归一化离散频谱的绘图
四元	BK1 MK BK2 MK BK3 MK BK4	示例(2)的对比频谱的测试集的绘图
五元	BK1 MK BK2 MK BK3 MK BK4 MK BK5	本发明的比较例 2 的流体动压轴承装置的顶表面的照片

(2)辅块作修饰语

一个辅块可以作修饰语,连用的辅块也可以作修饰语,有的辅块的连接词具有方向性,如示例中的“从……向……”。这类复杂名词短语的类别及示例如表 4 所列。

表 4 辅块作修饰语的复杂名词短语类别及示例
Table 4 Types and examples for complex NPs which auxiliary chunks are used as modifiers

类别	线性结构	示例
二元	FK BK	模具部件上的密封件 从燃料气体供给通路 42 向燃料气体循环通路 43 的燃料气体的逆流
三元	FK1 FK2 BK	三维中的平面上的区域
四元	FK1 FK2 FK3 BK	上述闸板部件上的除了上述凸部及上述凹部以外的上述对置面与上述外周面之间的间隙

(3)辅块和语义块作修饰语

此类复杂名词短语中的修饰语是辅块和语义块的组合,对于三元、四元复杂名词短语,根据辅块和语义块两者排列组合,可穷尽其组合形式,如表 5 所列。

表 5 辅块和语义块作修饰语的复杂名词短语类别及示例
Table 5 Types and examples for complex NPs which auxiliary chunks and semantic chunks are used as modifiers

类别	结构	示例
三元	BK1 FK BK2	滑门构件主体 21 的 X 方向上的大致中心位置 l
	FK1 BK1 MK BK2	矩阵成像器芯片上的像素的通常尺寸
四元	FK1 BK1 MK BK2 MK BK3	Y 方向上的两个斜面部 21b 的外侧端的距离
	FK1 FK2 BK1 MK BK2	从燃料气体供给通路 42 向燃料气体循环通路 43 的燃料气体的逆流
	FK1 BK1	本发明的实施例中剪齿齿轮于开启状态的上部立视图
	FK2 BK2	

4 汉语复杂名词短语形式化识别

本节主要关注 3 点:识别组合单元的边界、合成组合单元及判断组合单元的类型。

4.1 组合单元的识别策略

汉语复杂名词短语形式化是利用边界感知策略来实现的,即先利用边界词语集识别出组合单元的边界,再根据边界词语集的语义特点、切分特点和组合特点合成不同类型的组合单元,完成汉语复杂名词短语内部语义结构的切分和形式化。

汉语虚词有 3 方面的作用:1)提供组合单元的类型信息,“在、以、用、通过”是辅块标识符,“的”是名词性短语标识符;2)提供组合单元的分界信息,也就是说它们前后的词语不属于同一个组合单元,而属于不同的组合单元;3)提供搭配信息,这包括两个方面,一是可以搭配,如“在……中”可以搭配使用,二是不可以搭配,如“对……内”就不能搭配。因此,一个汉语复杂名词短语可以被分解为若干个组合单元。汉语连接词的边界特征和示例如表 6 所列。

对于词串中显现的“的”,系统需要对“的”判断其是语义块外的“的”(即需要调序的外分型“的¹”)还是语义块内的“的”(即不需要调序的内聚性型“的²”)。因此,与“的”相关的规则分为两类:一类是调序类规则,即将“的”识别为调序的“的¹”的切分标记;一类是排除类规则,即将“的”识别为“的²”。

对于隐性的“的”,系统需要判定在一定语言环境下的切分规则,来增加一个调序“的”的切分标记。

对于前连接词、后连接词,我们需要构建切分规则来识别切分标记,对于成对出现的前连接词和后连接词,我们需要通过属性赋值予以区分。

表 6 汉语连接词的组合特点和示例
Table 6 Types and examples for conjunction in Chinese structural units

类型	汉语连接词
左连接词	从、自、自从、由、往、至、到、于、临、沿、顺着、为了、为、因为、因、对、对于、关于、按、除了、如、诸如
	间、之间、上、之上、以上、上面、中、下、之下、以下、下面、前、之前、以前、前面、后、之后、以后、后面、内、之内、以内、里、外、以外、之外、外面、旁、旁边
右连接词	时等
	外分型 的 ¹
内聚型	的 ²

4.2 边界识别规则的构建

本文中边界是用规则来识别的。

4.2.1 规则说明

规则是产生式规则,由条件和操作两部分组成,“ $A \Rightarrow B$ ”是规则表达式的一般模式,表示当满足 A 时,执行 B 操作。

(1)规则特征集

CHN[x],表示终结符,其中 x 的取值是中文字符。
LC_CC[x],表示非终结符,其中 x 的取值是词性集合,如表示名词的“n”、表示动词的“v”等。

LC_TREE (CHK_VALUE, $start, end$)生成节点函数,生成一个非终结符作为边界,其中 CHK_VALUE 值为 FK, KU,分别表示后边界和前边界, $start$ 和 end 是规则左边的编号,程序自动对应到条件串中的绝对位置。如 LC_TREE (KU, $-1, -1$)在 1 号位置生成右边界 KU。

PUT($obj, feature, value$)赋值函数,给已经生成或已经存在的节点赋值。 obj 表示被赋值对象,可以是非终结符,可以是父节点 fp ,也可以是位置符号, $feature$ 表示属性特征名称, $value$ 表示属性值。如 PUT($fp, REPEAT_CC$)给父节点打上 REPEAT_CC 权值。

(2)边界节点

L1 是内部边界左边界的节点, L1H 是内部边界右边界的节点。MK 是外部边界,由“的”充当;MK_Q 是外部边界,是通过规则在右边界前增加的节点;MK_H 是外部边界,是通过规则在左边界前增加的节点;GBK_B%是词串开始的位置节点;GBK_E%是词串结尾的位置节点。

(3)节点属性

LEVEL 是 MK 标记上的属性值,有 3 个值,其 0 和 -1 是调序, -2 是保持位置不动。

CHANGE 是块饰 KU、特殊 FK 的属性值,有 2 个值, Y 是调序, N 是保持位置不动。

4.2.2 节点权值

考虑到复杂名词短语的层级性问题,我们对节点的权值大小进行了排列,这个设计也为合并策略的实现提供便利,当两个节点相邻时,以权值大者为准。L1 和 L1H 比 MK 大,块首位置和块尾位置最小。

4.2.3 规则集及示例

本文共构建了 39 条组合型连接词识别规则、17 条排除类规则、1 条调序类规则。

(1)组合型连接词识别规则

组合型连接词包括前连接词和后连接词。我们利用二者的语言特征,将它们分别识别为前边界和后边界。规则 1 表示前连接词集合 l1 被识别为 L1 前边界。规则 2 表示如果前面紧邻的词不是“的”的后连接词集合 l1h 被识别为 L1H 后边界。规则示例如下。

规则 1 (0)LC_CC[l1]=>LC_TREE(L1,0,0)

规则 2 (−1)!CHN[的]+(0)LC_CC[l1h]=> LC_TREE(L1H,0,0)&.\$

(2)识别“的”的排除型规则

在“的”的前面是形容词,这样的“的”表示单纯的修饰关系,比如“简明的目的”,规则示例如下。

规则 3 (−1){LC_CC[a]&!LC_CHK[L1H]}+(0){CHN[的,之]&!LEVEL[−2]}=>PUT(0,LEVEL,−2)\$

(3)识别“的”的调序型规则

经过了排除规则的排除,LEVEL 特征值不是 −1 的“的”就可以在调序“的”的识别中作为激活点。识别调序“的”规则是:(0){CHN[的,之]&!LEVEL[−1]}=>LC_TREE(MK,0,0)\$。

(4)增加“的”的调序型规则

在数词的左边是一个数字或数字与字母的组合,则在数词前面增加一个调序“的”。下面是增加调序“的”规则示例:(−1){LC_CC[NUM]+}(0){LC_CC[S]}=>LC_TREE(MK,0,−1)\$。

4.3 合并边界的策略

根据连续词串中出现的边界节点 FK,L1_FK,L1H_FK,L1,L1H,MK,GBK_B% 和 GBK_B% 合并组合单元边界。根据不同边界节点的组合特点,设计出相应的组合模式,其组合关系可用(,),[,]4 种符号¹⁾表示。

(1)合并辅块的策略

如果连续但不紧邻两个连接词是 L1“在”和 L1H“中”等,这两个边界优先捆绑为一个包含左边界 L1 和右边界的辅块 L1H。

如果非紧邻的两个左边界连接词 L1 是成对出现的连接词“从”和“到”,这两个辅块捆绑为一个只有左边界 L1 的辅块。

非紧邻节点按照以下 7 个组合模式(L1,L1H),(L1,L1],[L1H,L1H),(L1,MK],[MK,L1],[GBK_B%,L1H),[MK,L1H)合并,生成辅块。

(2)合并语义块的策略

从左到右依次扫描边界节点,非相邻节点按照 4 组合模式[L1H,GBK_E%),(MK,MK),(GBK_B%,MK),(MK,GBK_E%)合并,生成语义块。

4.4 组合单元类型的判定

辅块的判定在合并操作中完成。

块饰利用规则判定,识别规则是:(b){(−1)CHN[一种,

一个,许多,一系列]&GBK_B%)+(0){CHN[的,之]&LC_CHK[MK]}=>LC_TREE(KU,−1,−1)&PUT(−1,LC_EXP,NOT_CHANGE)。

5 汉语复杂名词短语转换

本节首先分析英语复杂名词短语的格式,然后进行汉英复杂名词短语的对比,最后根据对比结果制定汉语复杂名词短语的转换策略。

面向汉英机器翻译的复杂名词短语转换处理就是将汉语的复杂名词短语转换成相应的英语顺序,转换的前提和基础是复杂名词短语识别处理得到的形式化的结果,转换的依据是通过分析汉英双语对齐语料得到的对应规律。

汉语复杂短语的识别是转换处理的前提,转换的条件是识别得到的汉语复杂短语形式化的语义结构,转换的内容是汉语复杂短语的组合单元。需要明确的是,转换得到的只是一个中间过渡结果,而非最终的翻译结果。对复杂名词短语的汉英转换来说,系统的主要任务是根据识别结果生成的组合单元进行调序、节点增删和节点替换。

5.1 汉英复杂名词短语对比分析

汉语复杂名词短语翻译成英文时有两种情况:对应的情况和不对应的情况,其在语料中的占比如表 7 所列。前者指的是汉英复杂名词短语的意义完整,只是表现形式不同;而后者指的是汉英复杂名词短语转换时意义不完整,有可能出现汉语复杂名词短语的一个辅块或语义块在英语复杂名词短语中未得到体现,这种情况主要是承前省略或翻译融合等因素导致的。

表 7 汉英复杂名词短语的对应情况的比例

Table 7 Proportion for the English matching of Chinese complex Noun Phrases

汉语复杂名词短语	比例/%
英语复杂名词短语	91.33
英语句蜕	1.69
英语小句	0.23
其他	7.75

5.2 汉英复杂名词短语转换分析

汉语复杂名词短语识别结果是 3 种组合单元“KU,BK,FK”和调序标记“MK”的组合。在组合单元确定后,汉英复杂名词短语的转换研究只需要根据汉语组合单元的结构关系来确定其英语组合单元顺序。为了设计有效的转换策略,我们考查了二元复杂名词短语、三元复杂名词短语、四元复杂名词短语、五元复杂名词短语。为了方便表述,我们将多个同类型的组合单元用阿拉伯数字表示其从左到右出现的顺序;英语中不同的连接词都可以对应汉语的一个“的”,但对应“of”的现象比较突出,我们将英语连接词暂定为“of”。

5.2.1 二元复杂名词短语的汉英转换分析

我们采取穷尽式列举二元复杂名词短语中所有可能的结构关系来推理英语的组合单元顺序。我们可以列出所有的有两个组合单元的汉语复杂名词短语的全部构成。从理论上分析,名词性短语必须是中心语,所以其修饰语可以是辅块、名词性短语,组合单元的类型有且仅有这两种,如表 8 所列。

¹⁾ 组合模式中前小括号“(A”和后小括号“B)”表示包含 A 节点和包含 B 节点,前中括号 “[A”和后中括号“B]”表示不包含 A 节点和不包含 B 节点

表 8 二元复杂名词短语的汉英转换

Table 8 Chinese-English transformation for Complex NPs with two structural units

复杂名词短语 识别结果	汉语线性 序列	汉语结构 关系	英文线性 序列
BK MK BK	BK1 MK BK2	BK1 MK BK2	BK2 MK BK1
FK BK	FK BK	FK BK	BK FK

5.2.2 三元复杂名词短语的汉英转换分析

我们采取穷尽式列举三元复杂名词短语中所有可能的结

表 9 三元复杂名词短语的汉英转换

Table 9 Chinese-English transformation for Complex NPs with three structural units

汉语复杂名词短语识别结果	汉语线性序列	汉语结构关系	英文线性序列
BK MK BK MK BK	BK1 MK BK2 MK BK3	(BK1 MK(BK2 MK BK3)) ((BK1 MK BK2)MK BK3) (BK1)MK(BK2)MK BK3)	BK3 of BK2 of BK1
FK BK MK BK	FK BK1 MK BK2	(FK(BK1 MK BK2)) ((FK BK1)MK BK2) ((FK)(BK1 MK BK2)	BK2 of BK1 FK
FK FK BK	FK1 FK2 BK	(FK1(FK2 BK)) ((FK1)(FK2) BK)	BK FK2 FK1
BK FK BK	BK1 FK BK2	(BK1(FK BK2)) ((BK1)(FK)BK2)	BK2 FK BK1
KU FK BK	KU FK BK	KU FK BK	KU BK FK

5.2.3 四元、五元复杂名词短语的汉英转换分析

上述都是常见且容易穷尽式分析的复杂名词短语,但是四元、五元的结构关系不容易穷尽式分析,所以我们只能对语料中常见的四元、五元复杂名词短语进行举例式分析。

(1)四元复杂名词短语示例

如“示例(2)的对比(challenge)频谱的测试集的绘图(a plot of a test set of challenge spectra of example (2))”。此例的汉语复杂名词短语的表达式为“BK1 MK BK2 MK BK3 MK BK4”,英语复杂名词短语的表达式为“BK4 of BK3 of BK2 of BK1”。

汉语复杂名词短语的表达式“BK1 MK BK2 MK BK3 MK BK4”中 BK4 是中心语,BK1,BK2,BK3 是修饰语。从理论上分析该类结构关系有更多种,但是基于二元、三元复杂名词短语的分析,我们假设该类只有一种调序方式——组合单元逆序排列,即 BK4 BK3 BK2 BK1。

(2)五元复杂名词短语示例

如“本发明的比较例 2 的流体动压轴承装置的顶表面的照片(a photograph of a top surface of a fluid dynamic pressure bearing device of comparative example 2 of this invention)”。此例的汉语复杂名词短语的表达式为“BK1 MK BK2 MK BK3 MK BK4 MK BK5”,英语复杂名词短语的表达式为“BK5 of BK4 of BK3 of BK2 of BK1”。

汉语复杂名词短语的表达式“BK1 of BK2 of BK3 of BK4 of BK5”中 BK5 是中心语,BK1,BK2,BK3,BK4 是修饰语。从理论上分析该类结构关系有更多种,但是基于二元、三元复杂名词短语的分析,我们假设该类只有一种调序方式——组合单元逆序排列,即 BK5 of BK4 of BK3 BK2 of BK1。又如“Z 方向上的卡搬运路 13 的中心的上侧或下侧的形态”对应的英语复杂名词短语是“a position either higher or lower than the center of the card path 13 in the direction Z”。此例的汉语复杂名词短语的表达式为“FK MK BK1 MK BK2 MK BK3

构关系来推理英语的组合单元顺序。从理论上分析,三元复杂名词短语中,名词性短语 BK 必须是中心语,但修饰语比较丰富,可以是块饰、辅块和名词性短语。按照语言学知识和上述分析可知,辅块不能修饰辅块,且块饰不能跟名词性短语 BK 紧邻。根据调研 30 篇语料中复杂名词短语的情况,再加上合理演绎,我们可以列出所有的具有 3 个组合单元的复杂名词短语的全部构成,然后以修饰语之间的 3 种关系来推理线性排列的三元复杂名词短语的汉英转换情况,如表 9 所列。

MK BK4”,英语复杂名词短语的表达式为“BK4 of BK3 of BK2 of BK1 FK”。

5.3 复杂名词短语的汉英转换策略

我们可以根据上述二元和三元复杂名词短语的转换的情况大胆假设一个汉英复杂名词短语转换的策略:

- (1)如果有块饰 KU,则按原有顺序保持不动;
- (2)如果 FK 在两个或多个 BK 中间,所有组合单元按原来顺序逆序排列;
- (3)否则,将由 LEVEL 为 0 的 MK 连接语义块 BK 按照原有顺序的逆序排列,将辅块 FK 逆序位列其后;
- (4)将有 LEVEL 为-1 的 MK 连接语义块 BK 按照原有顺序的逆序排列;
- (5)删除汉语标记有 MK 的节点“的”;
- (6)增加上节点英文“of”,删除汉语辅块标记节点,加上对应的英语辅块标记节点。

6 复杂名词短语处理的应用

我们将识别规则和转换策略应用到现有的汉英机器翻译系统^[1,11-12]中,该系统融合了模板、统计、规则、实例等多种机器翻译方法,通过分析、转换、生成 3 个阶段实现对专利文本的汉英翻译,其中分析和转换都分为句间、句子和短语 3 个层面。复杂名词短语的识别和转换处理分别属于短语分析层面和短语转换层面,如图 1 所示。

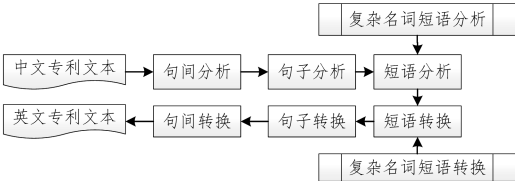


图 1 复杂名词短语在已有系统中的应用

Fig. 1 Application for the existing patent machine translation with processing module of complex NPs

为了测试处理效果,本文选择进行开放测试。开放测试语料来自 NTCIR-9 PatentMT 专利机器翻译项目中的 2 000 句专利语料,含有 1 469 个复杂名词短语。为了检验复杂名词短语的处理是否改进了机器翻译效果,我们针对包含复杂名词短语的句子,分别测试了其在基线系统和添加了复杂名词短语处理后的测试系统中的 *BLUE* 值,如表 10 所列。

表 10 添加复杂名词短语前后系统的 *BLUE* 值对比
Table 10 *BLUE* results for patent machine translation with and without processing module of complex NPs

<i>BLUE</i> 值	1-gram	2-gram	3-gram	4-gram
基线系统	0.4836	0.3415	0.2644	0.1846
测试系统	0.4859	0.4417	0.3605	0.2551

如表 10 所列,从两个结果对比可以看到,添加了复杂名词短语处理的测试系统的 *BLUE* 值要好于基线系统的 *BLUE* 值,尤其是前者的 2-gram,3-gram 和 4-gram *BLUE* 值明显好于后者。这说明本文的复杂名词短语处理可以明显改善含有复杂名词短语句子的机器翻译效果。

为了分别测试复杂名词短语形式化识别和转换的处理效果,本文又对 2 000 句专利语料中的 1 469 个复杂名词短语进行识别测试和转换测试。测试的标准是正确率, T 是正确识别的数量, N 是总数。正确率的计算公式为:

$$P=T / N \times 100 \%$$

需要说明的是,本文在测试中使用了测试句“本发明提供了……”,这样保证了复杂名词短语本身的正确识别。

6.1 复杂名词短语的识别处理结果与分析

复杂名词短语组合单元识别结果如表 11 所列。从表中的结果可以看到,识别的正确率高达 93.17%,这说明本文的识别规则和合并策略还是很有效的。

表 11 汉语复杂名词短语组合单元识别结果
Table 11 Recognition results for structural units of Chinese complex NPs

	二元复杂 名词短语	三元复杂 名词短语	四元复杂 名词短语	五元复杂 名词短语	<i>P</i> /%
应识别	1469	214	28	2	
识别	1312	209	28	2	
正确识别	1254	168	21	2	93.17

我们对其中的错误进行了分析,认为有以下几个因素影响到了识别的结果。

(1)组合单元不容易确定

汉英两种语言的词汇之间的对应关系很复杂,语义块不容易确定。

汉语的一个名词可以对应一个英语词组,而在汉英专利文献中,名词作修饰语的情况比较普遍。如复杂名词短语“闸板部件 14 的轴向中央部(a center part in the axial direction of the gate member 14)”,按照英语逻辑连接词,英语中有 3 个语义块,但按照汉语逻辑连接词,汉语中只能识别两个语义块。

汉语的一个辅块对应英语的一个单词。如复杂名词短语“细胞内胰岛素信号转导机制(intracellular insulin signal transduction mechanism)”,按照英语逻辑连接词,英语中有一个语义块,按照汉语逻辑连接词,汉语能识别两个语义块。其中汉语辅块“细胞内”对应一个英语单词“intracellular”。

汉语的一个词对应英语的不同形式。如复杂名词短语

“40 只哺乳期的母猪(the forty lactating sows)”和“哺乳期的最后 3 天(the 3 last days of lactation)”,其中汉语“哺乳期”是名词,它对应不同英语形式,一个是动词的现在分词形式,一个是动词的名词形式。

(2)歧义结构

歧义结构一般是由兼类词造成的。例如复杂名词短语“喷涂的时间和喷涂的区域(the timing of spray delivery and the area of spray application)”。根据预处理分词和加载的词典,“喷涂”是个动词和名词的兼类词,系统预处理阶段无法确定“喷涂”是动词还是名词。如果是名词,一般认为“喷涂的时间”和“喷涂的区域”都是一个组合单元;如果是名词,一般认为“喷涂的时间”和“喷涂的区域”都由两个组合单元构成的。

(3)单边界连接词

由于多个修饰语具有层级性,人工亦不容易正确判定只要一个连接词的辖域,机器只能按照大概率执行操作,正确识别所有只有一个连接词的组合单元是不可能完成的事情,导致组合单元识别错误。

如复杂名词短语“器件的皮肤接触侧上的凹痕中的粘合剂体(the adhesive body in the dent at the skin contacting side of the device)”,该复杂名词短语中的连接词“上”的辖域是“器件的皮肤接触侧”,根据切分规则所得的切分结果是:器件的/皮肤接触侧/上/的/凹痕/中/的/粘合剂体。对照译文,语义块切分是正确的。按照目前组合单元的合并策略,合并结果是:(器件)的(皮肤接触侧上)的(凹痕中)的(粘合剂体)。对照译文,组合单元的合并是错误的,应该是:(器件)的(皮肤接触侧上))的(凹痕中)的(粘合剂体)。造成这个错误的原因是单个边界“上”的辖域不明确。

6.2 复杂名词短语的转换分析与处理

同一个汉语复杂名词短语可以有不同的表达形式,可以对应要素句蜕,也可以对应复杂名词短语。结合计算机的特点和目前的转换策略,系统只能选择一种翻译方式,我们确定本文的转换是汉语复杂名词短语向英语复杂名词短语的转换。如果因为某些因素导致找不到完全对应语义块的情况,则我们仍然按照分析的结果进行转换。根据复杂名词短语识别正确的数据,复杂名词短语汉英转换的正确率如表 12 所列。

表 12 复杂名词短语汉英转换的正确率
Table 12 Accuracy for Chinese-English transformation for complex NPs

	二元复杂 名词短语	三元复杂 名词短语	四元复杂 名词短语	五元复杂 名词短语	<i>P</i> /%
应转换	1254	168	21	2	
转换	1254	168	21	2	
正确转换	1251	163	19	2	97.33

从表 12 可以看出,本文的转换策略是很有效的,正确率是 97.33%。需要说明的是,转换结果是按照分析正确的情况下进行的。

汉语复杂名词短语翻译为英语时非常复杂,可以有多种表现形式。如“基于 JSTOF 的滤波器 10 的多信道结构(the multi-channel structure of the JSTOF-based filter 10)”“基于 MIMO 的 JSTOF 电路 30 (the MIMO-based JSTOF circuit 30)”。辅块标记“基于”在翻译时作为构词成分融合到英语复杂名词短语内部。对于这种情况,我们采取权宜之计,服从大

局的处理策略。

单个连接词的组合单元跟隐现的“的”同时存在时,系统在转换上会出现错误。如复杂名词短语“膜两侧之间的晶体量之差(the difference in crystal amount between both sides of the film)”,按照目前的识别规则和合并策略,此复杂名词短语的识别结果是“膜/两侧之间/的/晶体量/之/差”。按照目前的调序策略,复杂名词短语的结果是“差 of 晶体量 of 膜 in 两侧”。对照译文,识别的组合单元没有错误,但是调序在“膜”和“两侧”处发生了错误。

结束语 本文提出了基于规则的复杂名词短语的识别和转换的处理方法。通过分析汉英双语对齐的专利文本中复杂名词短语语料的共性,以语义块为切入点,根据汉语语言特征的语义特点、切分功能和组合特点,采用边界感知原则为汉语复杂名词短语中组合单元边界构建了 57 条识别规则并设计了合并策略,得到汉语复杂名词短语的形式化结构。通过对比汉英复杂名词短语的差异,确定了汉英复杂名词短语的转换策略。最后,将识别规则、合并策略和转换策略应用到一个机器翻译系统中。测试结果表明,这种方法可以有效地实现复杂名词短语的识别和转换,提高专利文本中复杂名词短语的机器翻译效果。

在下一步的工作中,我们将通过分析更多的语料,深化对复杂名词短语的识别规则和转换规律的总结,并进一步修改和完善规则,加强对复杂名词短语形式化识别和转换的处理,以提高复杂名词短语识别和转换的准确率,进一步改善机器翻译效果。

参 考 文 献

[1] 张冬梅,晋耀红.面向专利机器翻译的要素句识别和转换研究[J]. 计算机科学,2014,41(S1):67-71.

(上接第 298 页)

[12] XING F Y. A Research on Chinese Compound Sentences[M]. Beijing:Commercial Press,2003.

[13] VASWANI A,SHAZEER N. Attention Is All You Need[C] // 31st Conference on Neural Information Processing Systems, 2017.

[14] IOFFE S,SZEGEDY C. Batch Normalization:Accelerating Deep Network Training by Reducing Internal Covariate Shift[EB/OL]. [2015-03-02]. <https://arxiv.org/abs/1502.03167>.

[15] MIKOLOV T,SUTSKEVER I,CHEN K,et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013,26:3111-3119.

[16] SRIVASTAVA N,HINTON G,KRIZHEVSKY A,et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[17] CARUANA R,LAWRENCE S,GILES L. Overfitting in Neural

[2] 池毓焕.多元逻辑组合的汉英对比初探.第二届 NNC 与语言学研讨会论文集[M]. 北京:海洋出版社,2004:308-312.

[3] 熊亮.非句蛇广义对象语义块的分析与处理[D]. 北京:中国科学院声学所,2006.

[4] 李千驹.基于 HNC 理论的汉英逻辑组合变换研究[D]. 北京:北京师范大学,2008.

[5] 李颖,王侃,池毓焕.面向汉英机器翻译的语义块构成变换[M]. 北京:科学出版社,2009:91-124.

[6] 詹卫东.面向中文信息处理的现代汉语短语结构规则研究[M]. 北京:清华大学出版社出版,2000:61-75.

[7] 李素建,刘群.汉语组块的定义和获取.语言计算与基于内容的文本处理[M]. 北京:清华大学出版社,2003:110-115.

[8] 胡乃全,朱巧明,周国栋.混合的汉语基本名词短语识别方法[J]. 计算机工程,2009,35(20):199-201.

[9] 田雪,黄德根.一种混合的汉语简单名词短语识别方法[J]. 小型微型计算机系统,2017,38(4):749-754.

[10] 姜亚辉,姬东鸿.结合半监督与主动学习的复杂名词短语识别[J]. 计算机工程与设计,2015,36(2):498-501,506.

[11] ZHU Y,JIN Y H. A Chinese-English patent machine translation system based on the theory of hierarchical network of concepts [J]. The Journal of China Universities and Telecommunications,2012,19:140-146.

[12] 刘小蝶,朱筠,晋耀红.中文专利中有标记并列结构的自动识别研究[J]. 计算机工程,2018,44(6):162-168,175.



LIU Xiao-die, born in 1984, Ph.D, lecturer. Her main research interests include Chinese information processing and corpus linguistics.



YANG Jin-cai, born in 1967, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include advanced database and information system, Chinese information processing, artificial intelligence and natural language processing.