文章编号:1671-4598(2024)06-0198-08 DOI:10.16526/j.cnki.11-4762/tp.2024.06.029 中图分类号:TP391 文献标识码:A

机械装配工艺文本的实体关系抽取方法研究

尹昱东，王保建，李珂嘉，王紫平，刘 洁
（西安交通大学 机械工程学院，西安 710049）

摘要：机械装配过程常常需要人工阅读并理解大量装配工艺文本，从而耗费大量时间，并且由于装配工艺文本书写人员和装配人员能力的差异，可能会导致装配人员错误理解装配文本，产生零部件错装、漏装等问题；机械装配矩阵以矩阵形式存储零部件的装配实体关系，可以直接、有效表达装配关系，不仅易于工人理解装配关系，也便于计算机识别，可以显著提高装配效率。自然语言处理作为研究计算机理解人类语言的工具，在根据装配文本生成装配矩阵的任务中可以起到关键的作用；文章采用自然语言处理的方法，对装配文本进行断句、分词、词性标注等文本预处理操作，采用机械装配名词语料库辅助以提高对装配零件的分词、词性标注时的准确率；用语法依存关系分析和语法模板匹配两种方法生成每个句子的主语、谓语、宾语三元组，其中采用机械装配名词语料库进行匹配，以判断其中的装配零部件名；之后提取出主语及宾语都为装配零件的三元组作为一个装配关系，对其进行去除冗余词、实体对齐等后处理操作；最后根据零部件数量组成一个空矩阵，将装配关系填入接触矩阵，并根据零部件类型判断生成装配关系的接触一连接矩阵。

关键词：装配工艺文本；实体关系；自然语言处理；词性标注；三元组；装配关系矩阵

Research on Entity Relation Extraction Method of Mechanical Assembly Process Text

YIN Yudong, WANG Baojian, LI Kejia, WANG Ziping, LIU Jie
(School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Mechanical assembly process often requires manual reading and understanding of a large number of assembly process texts, which consumes a lot of time. Moreover, due to the differences in the abilities of assembly process text writers and assemblers, it may cause assemblers to misunderstand assembly texts, resulting in problems such as wrong assembly and missing assembly of parts. Mechanical assembly matrix stores the assembly entity relationship of parts in the form of matrix, which can directly and effectively express the assembly relationship. It is not only easy for assemblers to understand the assembly relationship, but also easy for computer recognition, which can significantly improve the assembly efficiency. Natural language processing can play a key role in generating assembly matrix from assembly text as a method for computer understanding human language. In this paper, the natural language processing method is used to preprocess the assembly texts, such as sentence breaking, word segmentation and part of speech tagging. The mechanical assembly noun corpus is used to improve the accuracy of word segmentation and part of speech tagging of assembly parts; Then, the "subject-predicate-object" triplet of each sentence is generated by two methods: the syntax dependency analysis and syntax template matching. The mechanical assembly noun corpus is used to match the assembly part names; After that, the triplet whose subject and object are assembly parts is extracted as an assembly relationship, and the post-processing operations such as removing redundant words and entity alignment are carried out; Finally, an empty matrix is formed according to the number of parts, the assembly relationship is filled into the contact matrix, and the assembly relationship matrix is generated according to the type of parts.

Keywords: assembly process text; entity relationship; natural language processing; part of speech tagging; triplet; assembly relation matrix

0 引言

在传统的机械装配过程中，通常需要纸质工艺卡片指导工人操作，这种方法不仅缺乏指导性，还有可能由于工人主观认知的错误带来装配问题。自 20 世纪 80 年代以来，学者陆续提出许多为装配关系建模的方法，包括基于图结构、树形结构的模型。而与之相对的，矩阵模型更加易于计算机识别，也更便于序列计算的实现[1]。Dini 等人[2]首先提出了几种装配关系的矩阵模型。而基于此，Yu 等人[3]从表达零件连接和功能件结构干涉信息的对象关系图中提取

收稿日期:2023-05-11； **修回日期**:2023-06-15。
基金项目:陕西省自然科学基础研究计划项目(2021M-169);陕西省自然科学基础研究计划项目(2023-JC-YB-477);2022 年西安交通大学本科实验实践与创新创业教育教学改革研究专项项目(22SJZX10)。
作者简介:尹昱东(1983-),男,博士,工程师。
引用格式:尹昱东,王保建,李珂嘉,等.机械装配工艺文本的实体关系抽取方法研究[J].计算机测量与控制,2024,32(6):198-205,219.

关系矩阵，但需要预先输入关系图。

本文利用自然语言处理中的文本处理技术，主要有分词、事件抽取、语法依存关系分析等方法，从装配文本中提取出有效的装配实体之间的关系，这对于指导装配人员正确、高效地理解装配文本有很大的辅助作用，同时，因为矩阵模型易于计算机读取、识别，也可以成为自动装配相关应用的输入模型选择之一。

随着自然语言处理的发展，分词作为其基础任务一直是研究的重点[4]。而中文分词领域的研究也有了很长时间的研究历史。最早的中文分词采用了规则匹配的方法，也就是按一定规则将字段与词典中的词条进行匹配。2001 年杨文峰等人[5]提出 PATRICIA tree 数据结构，得以实现分词词典的快速查询与快速更新。2003 年李庆虎等人[6]提出双子哈希算法，进一步提高了分词的速度与效率的同时，没有增加词典的空间复杂度。规则匹配的算法虽然简单高效，但分词准确率不高，且无法处理歧义以及未收录于词典的词，因此在此基础上基于统计的分词方法开始出现。1993 年，Derose[7]提出了 WOLSUNGA 算法，利用概率统计模型对词性进行标注。Xue[8]在 2002 年首次提出了基于字标注的分词。基于统计的分词方法对于未收录于词典的词以及有歧义的句子效果比较明显，但模型运行的速度较慢，模型较为复杂。因此神经网络的分词模型逐渐展现出了优势。2001 年，Bengio 等人[9]提出了前馈神经语言模型，将神经网络与自然语言处理领域进行了结合，但由于每个 NLP 任务都需要单独构建模型，效果较为有限。Sutskever 等人[10]提出了可以保留长距离信息的长短时记忆网络(LSTM)。随后，许多基于 LSTM 的自然语言处理方法被陆续提出。这些方法的基于的原理不同，适用范围也不同。目前相关研究人员已经开发出了许多不同的分词工具，其中可用于中文分词的包括 jieba 分词、Standford CoreNLP、LAC、HanLP。

事件抽取是指从自然语言文本中抽取用户感兴趣的事件信息，以结构化的方式表达出来。事件抽取任务依赖于命名实体识别、关系抽取等多个自然语言处理任务的结果。事件抽取任务有基于模式匹配的方法和基于机器学习方法两大类。基于模式匹配事件抽取又分为有监督和半监督两种方式。相比于有监督模型，半监督模型中并不是所有训练样本都进行标注，而是依靠少部分标注样本和大部分无标注样本共同参与模型学习，自动对所有样本进行标注。这种学习方法可以大大减少标注样本的时间，并且也能保证足够的准确度。有监督的事件抽取依赖于人工标注的语料进行学习。Riloff 等人[11]在 1993 年通过建立触发词词典和 13 种事件匹配模式进行事件识别与抽取。Kim 等人[12]在 1995 年引入 WordNet 语义词典，利用语义框架和短语结构进行事件抽取。Rilofe 等人[13]在 1995 年开发出了不需提前标注的语料库的事件自动获取系统。模式匹配事件抽取虽然准确率高，但制作模板是一个非常费时费力的过程，且模板通常不具有通用性，只能适用特定领域，因此使用的便利性不足。而基于机器学习的事件抽取方法将事件抽取作为分类任务，是目前研究的重点方向。2013 年 Li 等人[14]提出基于结构预测的事件抽取联合模型，避免了误差传递的问题。Feng 等人[15]在 2016 年提出语言独立的事件检测神经网络。基于弱监督的模式匹配的目的是通过少量标注数据和大量未标注数据，生成大量标注数据。Chen 等人[16]2009 年针对中文在触发词标记上的问题，提出了更高性能的模型。目前事件抽取技术在许多领域都有着广泛的应用。2022 年 Jacobs 等人[17]提出从财经新闻中提取公司相关事件的模型。2021 年 Fei 等人[18]提出从知识图谱中建立富语境化语言模型用于生物医学信息抽取。2020 年崔晴洋等人[19]提出利用关键词抽取技术对卫星装配工艺文本进行分类。

依存关系分析，即分析句中词与词之间的依存关系，由法国语言学家 Tesniere 提出[20]，对于提取装配动作信息起到相当重要的作用。其理论的核心是将谓语作为句子的核心成分，从谓语出发与其它成分建立联系，最终将一个句子以一棵语法生成树的形式表示，其中父节点对子节点起支配作用，可以清晰地表示出句中各个语法成分之间的关系。对于依存树的形式需要进行一定的约束，以使其成为合法的、可操作的依存树。目前主流的依存句法分析大致有 4 种模型：生成式的句法分析模型、判别式的句法分析模型、决策式的句法分析模型和约束满足的句法分析模型。生成式的句法模型的方法是首先生成一系列依存句法树，再找到其中概率最大的一棵，而生成依存句法树则主要利用了词性和词汇信息，使用 prim 算法搜索最大生成树。

目前机械领域的自然语言处理相关的研究还相对较少，特别是机械装配领域中对装配文本的信息抽取任务的研究更是还处于起步阶段。本文对自然语言处理中的事件抽取方法展开研究，探索机械装配工艺文本的实体间的相互关系。研究目标为根据给定的装配文本，在预先不知道装配零件个数及名称的情况下，抽取出装配文本中所有零部件名词，并根据装配文本描述的装配关系，以接触－连接矩阵的形式输出。

1　文本预处理

在生成接触－连接矩阵之前，应首先对装配工艺文本进行一些预处理操作，以提取出重要的信息，便于下一步的提取操作。在自然语言处理领域，根据语言的不同，文本处理的方式也是不同的。本文使用中文文本作为处理对象，因为中文语料中词与词之间没有分隔符，因此需要单独的分词处理。然后基于分词的结果，标注每个词的词性。

本文的预处理流程为先导入文本，采用 jieba 分词进行分词、词性标注。但是待处理的机械装配文本中包含了大量的专有名词，如“齿轴盖”“紧定螺丝”等，而这些名词往往是没有被主流词库所收录的，这就会导致分词的结果会产生错误，从而影响最终结果的生成。

同时，对于装配文本中的机械零部件名词和其它名词，直接进行分词是不会做区分的。但是只有知道一个名词是

不是机械零部件，才能正确地识别一句话是否在描述一个装配动作，以及判断这个动作的主体是哪些零部件。基于以上原因，需要引入一个机械装配零部件语料库，以对分词、词性标注步骤进行修正，使其能正确识别机械装配零件，并对其做出标记。

1.1 分词

针对机械装配工艺文本的分词问题，使用 jieba 分词作为分词工具，它所采用的是基于统计词频的分词方法，主要使用了 N-gram 语言模型和隐式马尔科夫模型（HMM）。

N-gram 语言模型的原理主要是利用了条件概率，通过贝叶斯公式将一种分词结果的概率转化为一系列条件概率的乘积，最终比较出概率最大的分词方式作为最终的结果。根据如下条件概率公式：

$$p(s) = p(w_1)p(w_2 \mid w_1)\cdots p(w_n \mid w_{n-1}, w_{n-2}, \cdots, w_1) \tag{1}$$

式中，s 为一种分词方式，w_n 为句中第 n 个词。

但是这种方法计算量太大，不便于执行。俄国数学家马尔可夫提出一种假设：每个词的出现概率只与它的前一个词有关。这个假设被称为马尔科夫假设。这种假设虽然会影响最终计算出的概率值，但这种影响相对较小，却可以大大降低计算量。这样该公式就可简化为如下：

$$p(s) = p(w_1)p(w_2 \mid w_1)\cdots p(w_n \mid w_{n-1}) \tag{2}$$

注意这里的概率是由大量的文本样本统计得出的，因此可以很大程度地确保概率的准确度。在计算概率之前，需要先将所有的分词结果列出来。为了快速找出所有分词可能的情况，我们需要用到我们预先构建的词典中的词频，这时可以构建一个有向无环图（DAG），以图的形式表示所有可能的分词方式。

生成图的过程如下：首先每个字都为一条边，每两个字之间有一个节点，句首句尾各有一个节点。我们对于每个字，在字典中查询以这个字为首的词的词频，将这个字加上之后的 n 个字分别与词典中的词进行匹配，一旦匹配到一个词频不为 0 的词，就从这个词的词头前的节点向词尾后的节点添加一条有向边。遍历完每个节点后，这个图的边就构建完毕。而在寻找最大概率的路径时，由于图中只有词频数据，因此需要计算出每一个词出现的概率，即：

$$p(w_n) = \frac{freq[w_n] + 1}{total} \tag{3}$$

之后对每个词概率取对数，使得概率的相乘变为相加，这样就得到了每条边的权重。使用动态规划的算法，即可求解出整个句子最大概率的分词结果。如对于“他说的确实在理”这句话，便可以匹配到其中所有的词“他”“说”“说的”“的确”“确实”“实在”“在理”“理”，并成功构建出如图 1 所示的 DAG。

N-gram 模型可以根据统计数据快速找到最大概率的分词结果，但这种方法显然并不适用于词典中没有收录的词语，因此，还需要用到 HMM 算法对分词系统进行补充。

HMM 模型是一种经典的机器学习模型，它在自然语言处理领域起到较为重要的作用，可以通过初始状态的概率分布、状态转移概率矩阵以及观测概率 3 个要素描述马尔可夫模型，这基于与之前相同的假设，即当前状态仅依赖于前一个时刻的状态，而与更早的状态无关。

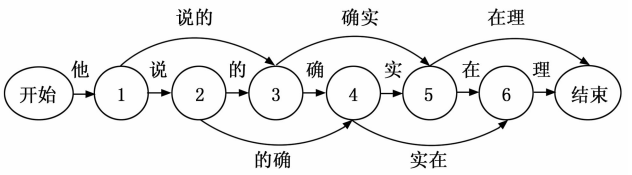


图 1 一个句子的有向无环图

由于在隐藏马尔可夫模型中，并不能直接看到状态序列，因此隐藏马尔可夫模型则具有两个序列：观测序列和隐藏的状态序列，而在中文分词中，观测序列即为句子本身，状态序列就是分词的结果。

这时需要事先训练出 HMM 模型，然后用 Viterbi 算法进行求解，得到最优的状态序列。其中，状态序列中每个字都为以下 4 种状态之一：B 为词首的字，M 为词中间的字，E 为词尾的字，S 为单字词语。

Jieba 分词中可以加载预先训练好的初始概率参数、发射概率和状态转移概率。如句子“小明是学生”，句中小明是词典中未收录的词，这时使用 viterbi 算法，也就是篱笆型图的最短路径求解问题。本例中由于“学生”一词已经被字典匹配，所以只需对“小明是”3 个字进行求解即可。图 2 中表示出了算法所用的篱笆型图。

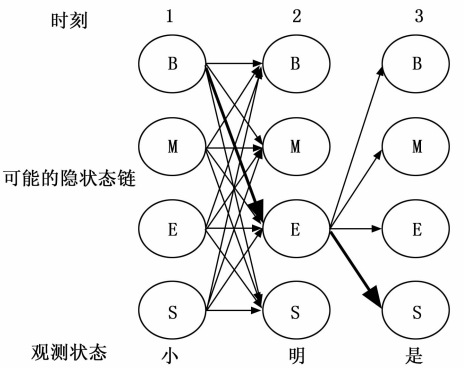


图 2 Viterbi 算法篱笆图

通过状态转移概率，可以得到每条边的长度，然后从起点出发，每走一步，都计算所有可能的路径的距离，如图 3 中所有的箭头所示。

根据最短路径的局部最优性的原理，每到达一个节点，只需保留当前最短的路径即可。这样，假如一共有 N 个字，每个字有 S 种状态，只需 $O(N * S^2)$ 的时间就可以求解出最短路径，远快于暴力搜索的时间复杂度 $O(S^N)$。最终可以计算出“小明是”这 3 个字的状态为“B，E，S”，这样就可以得到最终的分词结果“小明/是/学生”，如图中加粗箭头路径所示。

1.2　构建机械装配词典

Jieba 分词从大量新闻文本等文本库中统计词语，因此默认词库储存了大量日常使用的词语，并且统计词频也能较好地反映词语出现的频率。但是本文处理的装配文本中包含了大量装配相关的术语，这些术语中有部分并未收录在默认词库中，这就使得 jieba 分词在装配文本分词时出现错误。由于这是文本预处理的第一步，这时产生错误会非常不利于后续的处理，因此保证对于装配文本分词的准确度极其重要。

如“将滚针轴承压入齿轴盖孔内”这句话，由于“滚针轴承”“齿轴盖”两词未被收录于词库中，使用默认词库分词的结果为“将/滚针/轴承/压/入/齿轴/盖孔/内”，这将影响后续统计文本中包含的零部件时，“齿轴盖”这一零件会丢失。为了解决这个问题，需要构建一个机械装配零部件术语库，其中收录所有可能用到的零部件名词，这样就可以极大程度提高分词结果在机械装配文本上的准确度。

本文从《机械零部件名词术语图解词典（第二版）》一书中找出所有零部件名词。该书详细记述了机械通用零部件，传动系统零部件超重机械零部件，机械设备结构件等多个类别的零部件。如螺栓这一类别就包含六角头螺栓、六角头铰制孔用螺栓等等。这样如果零部件属于一个门类中具体的类型，我们可以将其完整名称分为一个名词，如果文本中只是单纯提到了一个门类的零件，就直接以门类名作为该零件名称。如“将 4 颗螺钉拧到箱体上”这句话中，由于未指定螺钉的类型，所以就将螺钉作为零件名。

Jieba 分词支持自定义字典中自定每个词的词频，但为了结果的准确性，所以不对词频进行指定，对于所有零部件词库中出现的词都直接匹配，这样才能不遗漏零部件，表 1 为零部件词库的部分。

表 1　部分零部件库

零件名	词性	零件类型(功能件 2/螺钉 螺栓 3/螺母 4/键 5/其它≧ 6)
螺栓	*nz*	3
六角头螺栓	*nz*	3
六角头头部带槽螺栓	*nz*	3
螺母	*nz*	4
六角形螺母	*nz*	4
六角形开槽螺母	*nz*	4

Jieba 分词有 3 种分词模式：精确模式、全模式和搜索引擎模式，针对当前问题，采用精确模式。分词前先对装配文本进行断句，以句号作为基本的断句单元，这样能确保一个句子中包含的信息是完整的，不会出现代词指代对象不明的情况。分词的结果保存为列表的形式，为下一步处理做好准备。

从运行结果来看，添加自定义字典的分词结果基本符合预期，特别是由于自定义词典的硬匹配机制，所有出现在零部件语料库中的词的分词结果一定是正确的。而且由于装配文本是一种较为书面的、严谨的语言，句子描述中产生歧义的可能性很低，所以不需要额外的标注样本进行学习，只使用默认的分词方式就可以取得足够好的效果。结果也显示，在 10 个测试用的装配文本内，分词结果的准确度都达到了 100%。图 3 为分词输出结果的示例。

```
['将', '清洗', '好', '的', '调速阀', '装', '上', '4', '×', '1', '：', '5', '密封圈']
['将', '装有', '密封圈', '的', '调速阀', '装入', '清洗', '好', '的', '壳体', '上', '4', '个', '调速阀', '孔', '中', '的', '3', '个']
['并', '预留', '1', '个', '调速阀', '孔', '在', '加油', '工序', '加油', '用']
['将', '适量', '的', '厌氧胶', '滴入', '壳体', '紧定螺丝', '孔', '内']
['将', '紧定螺丝', '旋入', '螺纹', '孔', '内']
```

图 3　分词结果

1.3　词性标注

词性标注（Part-of-speech Tagging）也是文本预处理的一个重要环节，即描述一个词在句中的成分、作用，主要的词性包含形容词，副词，名词，介词，动词等。

中文词性的特点是中文不像印欧语言一样具有丰富的词形变化，所以无法从词形直接辨别出词形，并且常用词常常具有多种词性，不结合上下文根本无法判断词性。因此，需要用到与分词时类似的方法，通过计算有向无环图最大路径的方式计算概率最大的词性标注结果。

在 jieba 分词中，词性标注的流程与分词类似。首先，加载词性词典，通过词性词典并构建出前缀词典；基于前缀词典，构建出词性的有向无环图；如果存在前缀词典中不存在的词，就利用隐式马尔科夫模型对其进行标注。

首先把分词结果的列表中的每一个元素作为一个节点，而不再是以字为最小单元；然后列出每个词可能的词性，构建出有向无环图。这里每个节点的特征就是词性，而转移状态概率就是已知前一个词的词性，后一个词是某一词性的概率。这样就可以使用 Viterbi 算法，从句首出发，一步步探寻局部的最优解，并最终构建出完整的词性标注结果。在使用该算法前，应先对句中的所有非中文词进行标注，这里可以用正则表达式匹配的方法，将英文单词和数字等其它语言匹配出来，预先进行标注。

图 4 是一个词性标注的示例。其中对于每个词，将其所有可能的词性列举出来，然后根据最大概率寻找出最可能的状态转移路径，如图中加粗箭头线所示。

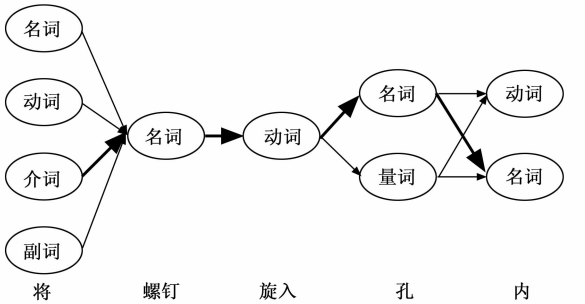


图 4　词性标注算法图

与分词时的情况类似，虽然 jieba 分词对常用词的标注结果相对理想，但面对机械装配零部件术语时，即使分词的结果正确，也不一定能正确标注出词性。而且由于一般的名词在 jieba 分词的标注结果都为“n”，即一般名词，因此无法区分一般名词与装配零部件名词。比如“将适量的厌氧胶滴入壳体紧定螺丝孔内”这句话，通过一般的词性标注得到“将（d）适量的（a）厌氧胶（n）滴入（v）壳体（n）紧定螺丝（n）孔（n）内（n）”，但事实上，由于这句话中的主语“厌氧胶”并不是一个装配零部件，这句话并不是描述装配关系的句子，在后续处理中该句子应当被忽略。

所以需要将名词中的零部件和非零部件区别开，而在词性标注时可以较容易地实现这个目标，只要在上文提到的零部件语料库中将所有零部件名词标注为“nz”，也就是专有名词，就能在词性标注这一环节将装配零部件与非零部件进行区分，从而便于在后续提取装配动作时，可以直接通过判断主语和宾语是否都为装配零部件的方式，将非装配语句去除掉。其中由于装配文本的内容只会局限于装配体本身，所以不会出现除零部件外其它的专有名词，也就不会存在其它专有名词干扰词性标注结果的问题。

实际操作的流程如下：先在表格中所有词标记为“nz”，然后载入词典，以分词后的列表结果作为输入，运行词性标注函数。得到的结果为一个两层嵌套列表。第一层列表列举所有的词语，第二层列表中第一个元素为词字符串，第二个元素为词性字符串。这时词性标注任务已经完成。在测试用的装配文本中，所有的装配文本的词性标注并不完全正确，但对零部件名词的特殊标注均没有错误。其中比较关键的动词，名词等标注结果是重点。图 5 为词性标注输出结果的示例。

```
['将', '清洗', '好', '的', '调速阀', '装', '上', '4', '×', '1', '.', '5', '密封圈']
['d',  'v',  'a',  'uj',  'nz',  'v',  'f','eng', 'x', 'eng', 'x', 'eng', 'nz']

['将', '装有', '密封圈', '的', '调速阀', '装入', '清洗', '好', '的', '壳体', '上', '4', '个', '调速阀', '孔', '中', '的', '3', '个']
['d',  'b',   'nz',   'uj',  'nz',   'v',   'v',  'a',  'uj',  'nz',  'f', 'eng', 'q',  'nz',   'nr',  'f',  'uj',  'eng', 'q']

['并', '预留', '1', '个', '调速阀', '孔', '在', '加油', '工序', '加油', '用']
['c',  'v',   'eng', 'q',  'nz',   'nr', 'p',  'v',   'n',   'v',   'p']

['将', '适量', '的', '厌氧胶', '滴入', '壳体', '紧定螺丝', '孔', '内']
['d',  'v',   'uj',  'n',    'v',   'nz',   'nz',    'nr',  'n']
['将', '紧定螺丝', '旋入', '螺纹', '孔', '内']
['d',  'nz',    'v',   'n',   'nr',  'n']
```

图 5 词性标注结果

图中每个分词结果列表和下面的词性标注列表一一对应。标注的第一句中出现了“4×1.5”这个非中文词，这里被分为了非语素字，而“孔”作为普通的名词被标注为了人名，可见词性标注的结果并不完全准确。但是由于后续关注的主谓宾三元组主要是名词、动词相关的，因此在下一步处理之前，可以先将不关心的词性的词去除掉，再进行下一步处理。经过处理的结果如图 6 所示。

通过分析分词、词性标注两个文本预处理的方法，其中包括各种算法的原理及其对比，证明了构建装配零部件术语库的方法以及其使用的效果。

```
['清洗', '好', '调速阀', '装', '415', '密封圈']
['v',    'a',  'nz',    'v',  'eng',  'nz']

['密封圈', '调速阀', '装入', '清洗', '好', '壳体', '4', '调速阀', '孔', '3']
['nz',     'nz',     'v',    'v',    'a',  'nz',   'eng', 'nz',   'nr', 'eng']

['预留', '1', '调速阀', '孔', '加油', '工序', '加油']
['v',    'eng', 'nz',   'nr', 'v',    'n',    'v']

['适量', '厌氧胶', '滴入', '壳体', '紧定螺丝', '孔', '内']
['v',    'n',     'v',    'nz',   'nz',     'nr',  'n']

['将', '紧定螺丝', '旋入', '螺纹', '孔', '内']
['d',  'nz',     'v',    'n',    'nr',  'n']
```

图 6 去除指定词性后结果列表

2 装配关系抽取

经过文本预处理后，已经获得了经过词性标注的分词结果，而接下来需要从分词结果中提取出装配动作，再生成装配矩阵。这其中包括依存语法分析，生成三元组，生成接触－连接矩阵 3 个分步骤。在进行操作前，首先要清楚事件抽取的概念。

2.1 依存关系分析

生成式的句法模型的方法是首先生成一系列依存句法树，再找到其中概率最大的一棵，而生成依存句法树则主要利用了词性和词汇信息，使用 prim 算法搜索最大生成树。

这时需要先创建一个虚拟的根节点，当前有 6 个节点，每个节点都有指向其它 3 个节点的一条有向边，每条边的权值由这两个词之间各种可能的依存关系中最大概率的关系的权值构成。使用 prim 算法寻找最小生成树，其中根节点为起始节点，且只有一条出边，最终就可以生成语法依存关系。

生成式模型的参数为；通过找到使得联合概率最大的值，也就是依存分析结果概率最大的值。

通过对模型加入一些规则限制可以减少模型的复杂度，但是以联合概率中变量独立作为假设前提的，而实际上由于上下文之间存在关联，各个词之间的依存关系并不是独立的。因此就产生了另一种以条件概率为判断依据的判别式依存句法分析模型。

判别式依存句法分析模型采用条件概率模型，找到使概率最大的模型参数。最大生成树模型与生成式模型类似，不过将寻找概率最大的生成树变为一种新的评分机制，由每个节点的特征和边的权值共同决定每条边的分数。决策树模型将查询问答作为一棵树，通过一系列决策寻找最优结果，判别式句法模型虽然运用了统计原理提高了准确率，但却使算法的事件复杂度变得较高。

决策式的句法分析模型与人阅读句子的习惯类似，从左到右依次读入每个词，每读入一个词，就对其与之前读入的词的依赖关系进行一次决策。移进－归约状态转移模

型由一个储存依存子树的堆栈和一个储存未分析词的队列组成。从队列中一次取出一个词，从移动（shift）、左支配（Left-Arc）、右支配（Right-Arc）和规约（reduce）中选择一个动作执行，直到队列为空，同时所有词都合并为一个依赖树。

决策式句法分析的时间复杂度为，相比其它算法来说效率很高，但无法处理非投射语言，且准确率稍低。约束满足的句法分析模型根据预先设定好的规则对依存关系进行裁剪，将所有不符合约束的依赖关系去除，直到留下一个满足约束条件的依存关系树。这种方法有可能无法生成满足所有约束条件的依赖树，也可能产生多个依赖树，所以准确度无法保证，可以辅以其它算法使用。

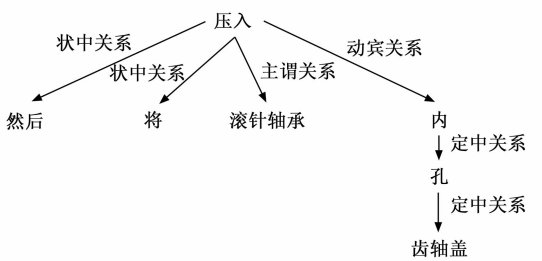
2.2　抽取三元组测试

考虑到机械装配文本中句子结构相对严谨，因此本文使用哈工大语言技术平台 LTP、百度 DDParse 分别进行了依存关系分析，另外使用一种不进行依存关系分析，而是采用词性标注进行模式匹配的方法直接获取三元组，并对这 3 种方法进行对比测试。

经过依存语法分析后，就能获得句子中各成分的关系。这时，由于一个装配动作一般是将一个或多个零部件与一个或多个零部件连接的形式，所以基本是以主谓宾形式出现的。

因此，只需从依存关系树中的根节点出发，通过预先标注的依存关系找到需要的主谓宾成分即可。首先，从根节点出发，通过核心关系（Head）找到句中的核心谓语。然后从谓语出发，用主谓关系（Subject-verb）找到句子主语。用动宾关系（Verb-object）、间宾关系（Indirect-object）或者前置宾语关系（Fronting-object）找到句子的宾语。

但此时所获得的只是主语和宾语的核心词，还需要通过其它的一些依赖关系，分别从主语和宾语出发，将修饰主语和宾语的成分也全都找出来。例如使用定中关系找出修饰核心主语、宾语的词，如“齿轴盖孔”一词中“齿轴盖”与“孔”之间的关系就是定中关系，使用并列关系可以处理包含多个主语或宾语成分的句子，最终成功找到所有句子成分。图 7 为从一句装配语句中提取主谓宾的案例。



[滚针轴承，压入，齿轴盖孔内]

图 7　装配语句中提取主谓宾

除了这种方法以外，还可以使用直接用三元组词性模板进行匹配的方法，跳过依存关系分析直接生成三元组。由于装配文本处理问题只是要提取主谓宾 3 个语言成分，对其它不重要的句子成分并不关心，因此可以直接利用主谓宾成分词性规则来从句中直接匹配出主谓宾成分。

基于语言学的基础知识，将不同的短语类型（如 NP，VP）用正则表达式的形式表示出来，然后遍历整个句子，将符合主语、谓语、宾语表达式的部分直接提取出来，两种方法都可以成功提取出主谓宾三元组，测试选用从互联网搜索的装配工艺卡，将其转化为文本格式传入系统中。

测试一共选用了 10 篇装配文本，去除没有完整主谓宾 3 个成分的句子，总句数为 105 句，一个句子中主谓宾提取均正确视为该句子整体正确，否则视为整句错误，经测试，3 种方法的准确率如表 2 所示。

表 2　算法正确率对比图

算法	正确句数	正确率/%
LTP	73	69.5
DDParse	78	74.3
模板匹配	80	76.2

从测试结果可以看出，使用词性模板匹配的方法进行三元组抽取的准确率是最高的，除了算法本身的识别错误外，对产生错误的原因进行分析，主要有以下几个问题：

1）当超过 4 个主语并列时，依存关系分析出现错误。如“依次将键、蜗轮、垫圈、锥齿轮、带翅垫圈和圆螺母装在轴上”一句中有 6 个并列主语，超过了主语数量判断的范畴，所以造成整句话的主谓宾均判断错误。

2）主语以代词形式给出时，无法联系上文。虽然在断句处理时以句号作为分割标准一定程度上可以实现指代消解，即遇到代词时寻找前文提到的最近的名词作为其指代对象，但指代消解的结果未必正确，因为代词可能指代更前文提到的名词。而且还有一些代词是跨句子进行指代的，由于断句隔绝了上下文信息故无法进行指代消解。

3）文本中某些句子省略了主语或宾语，或其中的一部分，导致信息无法弥补的缺失，造成错误。根据结果的准确率，采用第三种方法作为系统所使用的方法。基于词法模式的三元组抽取速度较快，且可以得到更多的三元组信息。

3　装配接触－连接矩阵生成测试

在装配文本进行预处理、依存句法分析以及抽取三元组之后，通过实体对齐方法消除多种说法的同一实体，特别是让零部件的部分向零部件整体对齐。然后消除非零部件名词，并以此为依据排除掉非装配动作的句子。最后将装配动作中的主语、宾语加入一个集合，作为零部件集合，将［主语，宾语］二元组作为图的一条边，主语、宾语分别为图的一个节点，绘制出零部件连接图。最后，通过零部件语料库中的标注，根据零部件类别的不同，用不同数字描述不同的装配关系，生成接触－连接矩阵，算法流程如图 8 所示。

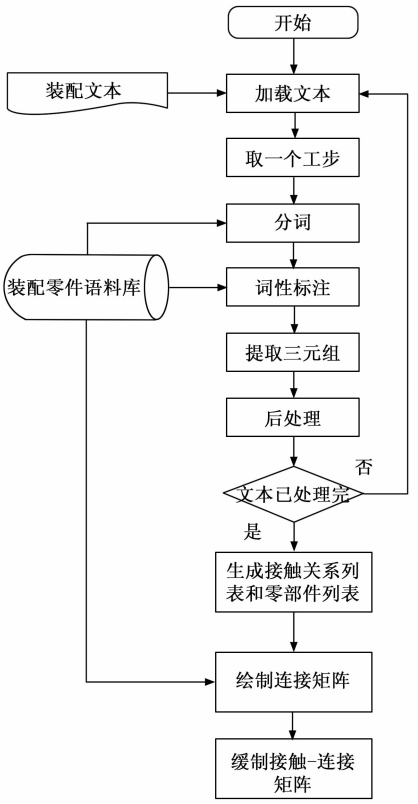


图 8 算法流程图

接触－连接矩阵是 2010 年于嘉鹏等人[1]提出的表示功能件和连接件间的连接关系的矩阵，本文研究对其概念进行了一定修改，设定矩阵将零件分为满足特定要求的功能件（如轴，齿轮）和保持结构稳定性的连接件（如螺栓，螺母），考虑到连接件功能的不同，装配顺序也不同，因此将其再按类别区分开，以不同的数字表示，具体规则见表 3。

表 3 连接类型与连接标号的关系

连接类型标号	零件 p_i 和 p_j 的连接状态
0	p_i 和 p_j 无接触
1	p_i 为功能件，与功能件 p_j 一般接触
2	p_i 为功能件，与功能件 p_j 直接紧固连接
3	p_i 为连接件，且为螺钉、螺栓类型
4	p_i 为连接件，且为螺母类型
5	p_i 为连接件，且为键类型

由于重点在于找到零部件之间的连接关系，而不考虑接触关系，因此不存在接触但不连接的情况，所以两零件连接的数字不会为 1。基于上述规则，首先对提取出的主谓宾三元组先进行一次筛选，将主语或宾语中有一个以上不包含装配动作的句子剔除，在此之前，先对提取出的主语和宾语进行处理，让它们只保留零部件名词。

根据词性标注结果，将所有除了名词（n）和专有名词（nz）外所有其它词性的词删除。如“紧定螺丝孔内”一个短语中“内”不属于这两种词性，会干扰后续结果，所以需要予以删除。

之后需要进行实体对齐操作，实体对齐是将对同一实体的不同描述方法予以统一的操作。如“将适量的厌氧胶滴入壳体紧定螺丝孔内，将紧定螺丝旋入螺纹孔内”这句话中，从思考的逻辑来看，“螺纹孔”一词指的就是前文提到的“壳体紧定螺丝孔”，而“壳体紧定螺丝孔”实际上是“壳体”的一部分，因此要找到“壳体”这个实体，将这两个短语都向这个实体进行对齐。虽然实体对齐都是基于文本相似性评估进行的，但这种方法对于装配文本却无法取得好的效果，因为装配文本中经常出现文本相似度极高但并非同一实体的词，如“紧定螺丝孔”和“紧定螺丝”，就分属于两个实体。所以只能使用结构相似性匹配的方法进行识别。如“壳体紧定螺丝孔”经过词性标注的结果为“nz，nz，n”，这种情况需要将实体对齐到最先出现的专有名词，也就是“nz”上。而对于“螺纹孔”一词，词性标注的结果为“n，n”，这时从提取最后一个名词“孔”开始，往前文寻找是否出现过同样以“孔”字结尾的短语，如果有，将其替换为前一个出现的短语，再对两者同时进行前面的操作对齐到一个专有名词“壳体”，这样就实现了同一个句子的实体对齐。

经过这一步处理，剩下的主语和谓语都只可能由一个装配零件名词或一个其它名词组成，只要进行一次判断，保留所有主语、宾语都为装配零部件名词的三元组，并将其变为［主语，宾语］的列表即可。

下一步，将所有主语宾语合为一个列表，并将列表转化为一个集合。由于集合中不会出现重复元素，所以这个集合中的元素就是零部件名词的集合，因此就得到了装配文本中出现的所有零部件。提取出的零部件名词列表如图 9 所示。

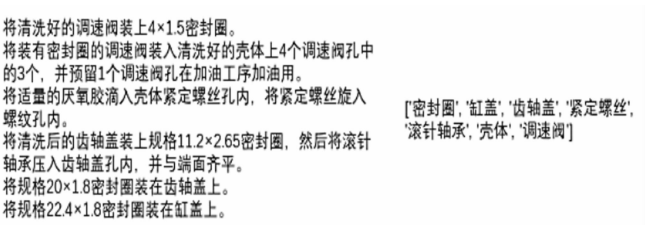


图 9 装配文本生成零部件示意图

有了零件列表就可以构建接触－连接矩阵，由于机械装配的连接可以用图的形式表示，可以将一组［主语，宾语］视为连接图的一条边，根据节点信息和边的信息生成完整的图，连接矩阵是图的一种表示形式，如果第个零件与第个零件相连，就将矩阵的第行第列和第行第列置为 1，这时得到的是初步的连接矩阵。

如果要得到最终的接触－连接矩阵，需要结合每个零部件的类型信息，在制作机械零部件名词词库的时候，根据零件类型的不同，对其进行预先的标注，标注的结果储存在另一个字典中。

在生成接触－连接矩阵时，调用该字典，以零部件名称作为键值索引到零部件的类型，并将类型数值直接乘到该零件所在行的每一个元素上，如第个零件对应的类型为，就让矩阵第行的所有元素自乘。生成的接触矩阵和接触－连接矩阵如图 10 所示。

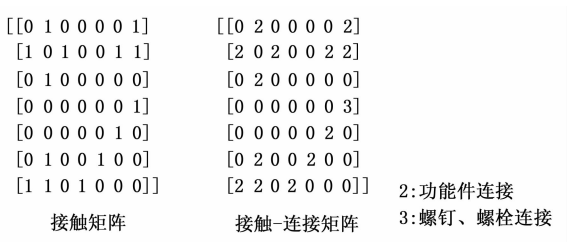


图 10　接触矩阵和接触连接矩阵对比

本文仍然使用 10 个装配文本对接触－连接矩阵生成的准确率进行评估。评价的标准包括零部件检测的准确率和连接关系的准确率。准确率结果见表 4。

表 4　零部件正确率和连接关系正确率表

文本序号	零部件正确率/%	连接关系正确率/%
1	100	100
2	78.2	44.4
3	90	75
4	85.2	62.5
5	80	57.1
6	100	60
7	85.7	69.2
8	87.5	66.7
9	70	76.9
10	66.7	53.8

可以看出，总体来看，零部件检测的准确率普遍较高，平均准确率达到了 84.33%。这得益于零部件词库的硬匹配机制使得零部件不会被漏掉。但是在句法模式匹配的过程中零部件可能会因为未被正确分到主语、宾语短语中从而产生遗失。而接触－连接矩阵方面，由于装配文本之间描述方式的差异，生成矩阵的准确率方差较大，平均准确率为 66.56%。错误的产生主要源于上一个环节句法模式匹配误差的传递，以及指代消解、实体对齐两个环节中产生的错误。

总体来看，这套流程基本实现了装配文本到接触－连接矩阵的转化，其准确率已具有一定程度的指导意义，但还需要人工进行校验。在装配文本的描述较为标准，代词指示符合匹配规则时，该系统的表现比较稳定，可通过规范化装配语言的方法显著提高准确率。

4　结束语

本文实现了机械装配文本自动生成接触－连接矩阵的功能，从系统准确率来看，在文本预处理，也就是分词和词性标注环节，由于装配文本的严谨性，再结合零部件词库的匹配，在测试用的 10 个装配文本中准确率达到了 95%以上。而在三元组生成环节，由于部分文本并列零件数过多，代词无法还原，零部件被省略等多种原因，导致准确率相对较低。在生成零部件名词列表任务中，由于引入词典的缘故，准确率达到了 84%以上。最后在接触－连接矩阵生成的任务中，由于装配文本的描述方式、断句方式等差别较大，所以导致生成的矩阵准确率参差不齐，平均准确率在 65%左右。对于句式简单，描述准确，代词指示符合规则的文本，生成的准确率相对较高。而有些文本由于描述上的歧义等问题，准确率相对较低。但总体来看，该方法的效果在目前技术看来还是比较令人满意的，但还有一些可以改进的空间。

首先，在分词和词性标注环节，使用词库虽然在匹配时准确度可以达到 100%，但遇到词库中未出现的零部件时就一定无法识别，所以可以采用深度学习的方式进行，采用 BERT 模型使用标注后的文本进行预训练，预计可以提高系统面对未收录词时的准确率，同时考虑到领域适应性问题，可以采用相关专业领域的知识，提取与机械领域相关度更高的关联词。

在语法依存关系分析环节，使用句法匹配的方法虽然复杂度低，但结果受限于规则，对于规则外的语法或一些不标准的语句，分析结果就会很容易错误。因此可以使用图循环网络来进行语义角色标注，使用预先标注的数据，将句子视作图进行边权值的更新，预计能将准确率大幅提高。

在生成装配矩阵中，指代消解任务的表现不够稳定，特别是对于零指代消解（零元素没有显示给出，需要根据上下文识别出隐式成分）没有进行相应的处理。可以采用多注意力机制，分析上下文语境并挖掘词序列之间的依赖关系，以提高指代消解的效果。并且由于使用词库，不仅只能识别词库中出现的词，还无法对零部件的前缀型号进行识别，如“将清洗好的调速阀装上 4×1.5 密封圈。”一句，经处理后只剩［调速阀，密封圈］，导致密封圈的型号信息缺失，所以需要采用命名实体识别任务解决，可以采用 BERT 模型对零部件进行识别，使零部件包含参数信息，更加易于阅读，也可以避免不同零件因仅型号不同而被归为同一零件。

当前，机械装配正逐渐从传统的人工装配向着自动化装配的目标转变，而这个过程中，装配文本的数字化、标准化是十分关键的研究方向。本文通过自然语言处理任务的组合，在 Python 环境中开发了功能完整的机械装配文本自动生成接触－连接矩阵系统，使得装配文本中蕴含的装配信息可以更直观地进行展示。同时，这对于计算机理解装配文本也有很大帮助。通过接触－连接矩阵，计算机能快速了解整个装配的情况，并可根据矩阵进一步实现自动计算装配序列，自动匹配装配零部件，为实现全自动装配构建理论支撑。

本文受限于样本量不足的原因，没有构建出足以训练的

（下转第 219 页）