# 基于 BCC 的离合词离析形式自动识别研究*

臧娇娇[1]，荀恩东[1]

（1.北京语言大学，北京市，100083）

**摘要**：本文从中文信息处理角度对动宾型离合词自动识别进行研究。通过分析离合词在实际语料中的使用特点以及离合词离析成分在大规模语料库中的表现形式，从离合词内部入手，形式化地表示离合词的离析形式，总结自动识别的规则，设计基于规则的自动识别算法。经过优化后，该算法在 20 亿字的语料中得到了 91.6%的正确率。离合词语素构词能力强，分词与词性标注错误，规则的不完整性，语料本身的错误，人工标注的疏漏等是影响实验结论的主要因素。

**关键词**：离合词；BCC；离析形式；自动识别

# Research On Automatic Recognition of Separable Words Based on BCC

ZangJiaojiao[1], XunEndong[1]

(1.Beijing Language and Culture University, Beijing 100083, China)

**Abstract:**This paper conducts a research on the automatic recognition of separable words from the perspective of Chinese information processing. It starts with separable words to summarize recognition rules and design automatic recognition algorithm with the aid of separable forms by analyzing the characteristics of different separable words and studying the forms of separable compositions in the large-scale corpus. The algorithm achieves 91.6% accuracy rate with continuous optimization of the program in the corpus of two billion word. Morpheme with strong word-building ability, wrong word segmentation and POS tagging, incomplete rules, errors in the corpus and manual annotation omissions are main factors affecting experiment results.

**Key words:**separable words; BCC; separable forms; automatic recognition

## 1 引言

所谓“离合词”，就是指汉语中一种双音节结构，意义凝固，中间可以插入其他的成分，可离可合的语言现象。陆志韦先生（1957）提出离合词的概念，认为离合词是现代汉语研究中的重要问题，并引起学界对该问题的广泛关注。此后，学者从不同的角度对离合词进行广泛的研究。

### 1.1 离合词研究综述

自上个世纪 40 年代以来，离合词的研究主要经历了两个阶段的发展：第一个是对离合词基本问题进行探索的阶段，主要涉及离合词的定义、性质、界定、类型等问题的本体研究；第二个阶段从上世纪 80 年代开始，离合词进入深入拓展研究阶段。在探讨离合词基本问题的基础上，离合词的研究逐渐由本体研究慢慢转向实际应用领域。

随着计算机技术的进步，离合词的研究在中文信息处理领域主要包括利用语料库对离合词进行统计，自动分词以及词性标注中对离合词的处理策略，汉英机器翻译中对离合词如何翻译等。

王海峰、李生等(1999)主要研究汉英机器翻译中离合词的处理策略问题，基于大规模的语料库，对离合词进行详细的统计和分析，并提出 BT863 汉英机器翻译系统中离合词的处理策略。王春霞（2001）在对大规模语料考察与分析的基础上，得到离合词的离析形式在语料中的出现情况，通过对插入成分的规律进行总结，最后获得离合词的组配模式。史晓东（2002）把离合词分为四种类型，探讨离合词在机器翻译中的句法分析、意义表示、翻译策略等问题，并作了初步实现。徐建山（2003）基于汉语长距离搭配现象，结合离合词的共同特点，实现

* **基金项目**：国家社科重点项目（大数据背景下汉语语块数据库建设与应用研究）；国家 863 计划重点项目（SQ2015AA0100074）

了一种识别离合词的算法。任海波等（2005）在大规模语料库的基础上对离散度不同的离合词进行定量分析，并尝试性地确立汉语普通话中典型离合词数量。周卫华、胡家全等人（2010）对动宾式和并列式离合词的扩展形式进行详细的分析，在考察分析这两类离合词扩展形式特点的基础上，提出在中文信息处理系统中应该建立离合词词库，并对离合词的扩展形式做出专门的符号标注。

离合词的研究与中文信息处理等领域相结合后，对离合词离析形式识别的研究成为学者们首要考虑的问题。冯向华（2009）比较系统地研究离合词的扩展形式，结合不同的扩展形式，设计了一个离合词扩展形式的自动识别程序。从识别效果来看，虽在某些类型上达到了一定的识别效果，从整体来看效果却不是很好。刘博（2015）通过分析离合词扩展形式自身的特点，依据算法设计了一个现代汉语离合词扩展形式自动识别系统，通过开放的实验测试，对数据进行测试并不断优化，但是其研究并未从整体上对识别的效果进行统计。

### 1.2 离合词自动识别的研究应用

中文信息处理应用主要涉及机器翻译、自动分词、信息检索、自动标注等领域。比如，在汉英机器翻译中，如果只能识别离合词的整体形式，而对其“离”的形式无法识别的话，可能会导致在翻译中无法从整体上理解离合词的语义，从而影响翻译的效果；又如，在信息检索时，如果离合词的离析形式不能从整体上被识别出来，计算机将会对切分后的内容进行查询，造成检索时的盲目性。另外，中文分词作为中文信息处理领域的基础技术，在离合词的自动识别中得到了具体应用，同时离合词的自动识别也对中文分词技术也起到推动作用。

离合词是现代汉语中比较特殊的语言现象，从语言学本体角度进行的研究较为丰富，而在自然语言处理角度对离合词的研究逐渐起步，并受到越来越多学者的重视。目前离合词的研究现状主要集中在以下几个方面：在研究内容上，本体研究多于应用研究；在研究方法上，定性分析多于定量分析；在研究深度上，统计研究多于识别研究。本文的主要工作是，设计出一种识别的算法，将这种算法应用于某种程序语言，通过编写程序实现对离合词离析形式的自动识别研究，从而有利于计算机在自动分词、统计识别、机器翻译等方面的应用研究。

## 2 离合词离析形式的统计与分析

### 2.1 离合词词表的确定与语料选择

#### 2.1.1 离合词词表的确定

不同学者对离合词的界定标准不一样，其数量没有固定的统计。本文在前人的研究基础上，以相关论文和著作为依据，所研究的离合词主要来自《现代汉语词典（第五版）》（以下简称“现汉”）和《汉语水平词汇与汉字等级大纲》（以下简称“大纲”），并根据《现汉》（第六版）对所提取的离合词进行修订。

《现汉》对离合词做了形式标记。离合词的注音在中间加双斜线“//”，表示中间可以插入其他成分，如“洗澡 xǐ//zǎo”。本文借助注音中的“//”共提取出 3487 个离合词，然后把《大纲》的词与《现汉》提取的 3487 个离合词进行交集合并，共得到 402 个离合词。本文又对 402 个离合词进行细化分析。首先删减了一些动补型离合词，比如“提高”、“出来”、“看见”、“抓紧”、“起来”等；其次删减了一些歧义的词语，比如“点心”作离合词时，在《现汉》中是动词“吃东西”的含义，但在《大纲》中却是名词，表示“一种食品”，与此类似的还有“运气”、“制服”、“入口”等；最后根据《现汉》（第六版）中离合词标记的变化，又删减了一些在第五版中存在形式标记“//”，但在第六版中已经取消标记的离合词，比如“出席”、“登陆”、“关心”、“突出”、“作文”等；增加了一些在第五版中没有形式标记，但是在第六版中存在形式标记的词，比如“游泳”、“贬值”等。本文最终确定 140 个离合词作为识别的对象（见附录 1）。

#### 2.1.2 语料的选择

本文的语料主要来自于北京语言大学大型语料库 BCC 中的综合频道语料。综合频道是一

个平衡语料库，其中包括文学、科技、微博、报刊不同的语体，约 20 亿字。本文利用综合频道的语料，通过 BCC 的检索模式，得到离合词离合现象的语言实例。

### 2.2 离合词训练集和测试集的确定

#### 2.2.1 人工标注

通过 BCC“A*B”的检索模式，对 140 个离合词进行穷尽式检索。鉴于语料的复杂性和检索模式的局限性，每个离合词检索的语料都包含正确的离析形式和错误的离析形式。通过人工标注的方法，对检索的每个离合词例句进行标记。人工标记的规则如下：正确的形式在后面标记“1”，错误的形式在后面标记“0”。例如：

他刚洗完澡、刮完胡子，身上还残留着淡淡的芳香。1 （湍梓《相逢不恨晚》）

在这种时候，千万不能回家睡觉，一睡便觉得万念俱灰。0（亦舒《城市故事》）

#### 2.2.2 预处理

人工标注之后，对待识别的文本文件进行预处理。离合词中间插入中间成分的现象，一般在分句中。所以先对语料进行预处理，包括词性标注和分句处理。在词性标注的基础上，再借助“/w”（北大的词性标注体系，“/w”表示标点符号）词性符号的标识对语料进行分句。在对原始语料进行分句的时候，主要依据标点符号，不仅要对整句进行分句，对小句也要分句，即在遇到逗号、句号、问号、冒号、分号、顿号、感叹号、省略号等标点符号时要进行分句处理。

#### 2.2.3 离合词离析形式正反例频率统计

结合每个离合词所包含的正确和错误离析形式的语料，统计离合词正确和错误离析形式的例句数。正例数是离合词正确离析形式的数量，错例数是离合词错误离析形式的数量；正例率是离合词正确离析形式所占的比例，错例率是离合词错误离析形式所占的比例。以下是两个计算频率的公式：

离合词正例率=离合词正例数/离合词总标注实例数；

离合词错例率=离合词错例数/离合词总标注实例数。

按照计算公式，得到 140 个离合词的正例率。表 1 是 140 个离合词正例率的分布情况：

表 1 离合词正例率的分布情况

| 正例率 | 数量 | 所占比例 | 正例率 | 数量 | 所占比例 |
|---|---|---|---|---|---|
| 0-10% | 34 | 24.3% | 50%-60% | 11 | 7.8% |
| 10%-20% | 18 | 12.9% | 60%-70% | 6 | 4.3% |
| 20%-30% | 9 | 6.4% | 70%-80% | 9 | 6.5% |
| 30%-40% | 9 | 6.4% | 80%-90% | 14 | 10% |
| 40%-50% | 9 | 6.4% | ＞90% | 21 | 15% |

从表 1 可以看出离合词的正例率分布情况有很大的差别，有些离合词正例率高；有些正例率低，甚至有些离合词正例率为 0，即在本文所使用的语料中没有出现正确的离析现象。在统计结果中，正例率在 90%以上的有 21 个离合词，所占比例为 15%，包括“鞠躬、洗澡、吵架、吃亏、叹气”等词；正例率在 80%-90%之间的有 14 个离合词，所占比例为 10%，包括“泄气、拼命、散步、鼓掌、告状”等词；正例率在 10%以下的离合词有 34 个，所占比例为 24.3%，包括“配套、出神、罢工、探亲、出院”等词。从 140 个离合词正例率分布情况可以看出，本文所选择的 140 个离合词具有代表性，每个频率段的离合词都有所涉及，并且分布相对均衡。

在确定测试集和训练集的过程中，要考虑一些特殊情况。比如“鞠躬”，在语料中只有正确的实例，所以只能选择做训练集；而“集邮”在语料中只有错误的实例，所以只能用来做测试集。除了这些特殊的离合词，本文根据每个离合词的正例率所占比例和分布情况，选取 20 个离合词作为训练集，既包含一些正例率高且离析形式多的离合词，以便于总结离析

形式；也包括一些正例率低且离析形式少的离合词。训练集包括“鞠躬、吃亏、冒险、帮忙、洗澡、打仗、倒霉、分红、开幕、遭殃、动身、就业、迎面、到期、听话、着急、出差、及格、握手、报名”，剩下的 120 个离合词作为测试集。

### 2.3 离合词离析成分的统计

通过对训练集中 20 个离合词语言实例的分析，对离合词的插入成分进行提取和统计，然后再结合人工筛选的过程，总结离合词的离析形式。比如表 2 是“帮忙”中间插入成分的统计情况。鉴于其复杂性，只列出其中间插入成分频数在前 30 的情况：

表 2 “帮忙”中间成分插入情况

| 插入成分 | 频次 | 插入成分 | 频次 | 插入成分 | 频次 | 插入成分 | 频次 |
|---|---|---|---|---|---|---|---|
| 个 | 637 | 上 | 184 | 点 | 101 | 了我的 | 63 |
| 不上 | 461 | 什么 | 168 | 我个 | 96 | 不了你的 | 52 |
| 你的 | 367 | 这个 | 148 | 上什么 | 78 | 帮我的 | 52 |
| 我的 | 319 | 我 | 140 | 我们的 | 78 | 你这个 | 51 |
| 得上 | 296 | 我这个 | 126 | 你什么 | 74 | 一个 | 48 |
| 我一个 | 252 | 不了 | 115 | 的 | 72 | 不了什么 | 45 |
| 他的 | 189 | 她的 | 110 | 了 | 66 | | |
| 你 | 188 | 不上什么 | 108 | 他 | 66 | | |

本文对 20 个离合词的中间成分进行提取并按照在语料中的词频排序，每个离合词的统计结果诸如“帮忙”表 2 的统计模式。鉴于提取的方便和人工筛选的复杂性，在对离合词中间成分自动提取的过程中，只获取其中间成分，直接根据中间成分对离析形式进行总结。

### 2.4 离合词离析形式的分析

对训练集中的 20 个离合词的插入形式进行归纳，主要分为以下几种类型：

1.插入助词“了”、“着”、“过”成分

离合词中间插入“了”、“着”、“过”是最普遍的情况。如：鞠了躬、吃过亏、冒着险等。

2.插入补语

离合词中间插入补语的情况比较复杂，鉴于本文主要从形式入手，所以本文不把数量短语列入补语的范畴，而是把数量短语单独列出来。插入补语中间成为多为“上、完、起、成、不成、得、不得、不了、不到”等词。如：帮不上忙、洗完澡、打起仗来等。

3.插入量词“个”

插入量词“个”，“个”在量词中比较特殊，使用比较广泛。如：报个名、冒个险等。

4.插入数词

在离合词的前后语素间插入数词大多是插入“一”的情况，也有插入其他数词的情况，如“两”、“几”等。如：鞠一躬、打几仗等。

5.插入量词

在离合词的前后语素间插入量词，可以用来补充说明动作次数，主要是动量词，包括“次、回、下、遍”等。如：出次差、帮回忙等。

6.插入数量短语

在离合词的前后语素间插入数量短语，用来补充说明动作的数量或者持续的时间。如：吃一分亏、洗一趟澡等。

7.插入代词

离合词的前后语素间插入代词，一般分为三种类型：插入人称代词、指示代词、插入疑问代词“什么”。如：听我的话、报这个名、着什么急等。

8.插入名词/形容词

在离合词前后语素之间插入名词、形容词作定语，修饰后面的名语素，是比较常见的情况。如：洗冷水澡、倒大霉等。

9.插入结构助词“的”

离合词的前后语素间可以插入结构助词“的”。如：洗的澡、吃的亏等。

10.重叠

重叠的情况主要包括以下五种形式：“AAB、A 一 AB、A 了 AB、A 没 AB、A 不 AB”。如：鞠鞠躬、帮一帮忙等。

11.离合词前后语素之间插入复杂成分

以上 10 种形式主要是离合词中间插入单一成分的情况，另外还有插入多种成分的情况。当插入成分为“了/着/过+其他成分”这种类型时，例如，“洗了个热水澡、打了一场辛苦的仗、倒了一次大霉”等。在自动识别过程中将对插入多种形式现象进行详细的归纳总结。

## 3 基于 BCC 的离合词离析形式自动识别

### 3.1 离合词离析形式的规则总结

根据识别的难度和在大规模语料中的实际使用情况，对上面所总结的离合词离析形式的识别规则进行总结归纳，并转换成机器可以识别的程序化语言。离合词的插入成分主要有两种情况：一是插入单一成分；二是插入多种成分。插入多种成分的情况比较复杂，会根据离析长度和离析成分进行总结。下面先分析插入单一成分的情况：

3.1.1 插入单一成分的规则总结

（1）插入成分为固定汉字：A + u + B（u=汉字集合）

根据上面总结的 10 种离析形式，总结规则时只考虑语法形式。先把具有明显特征的汉字提取出来，作为一个集合。比如“了”、“着”、“过”、“个”、“什么”、“的”等。

（2）插入成分为词性：A + p + B(p=词性集合)

通过对离析形式的总结，插入词性的情况有以下几种：r：代词，n：名词，v：动词，a：形容词，m：数词，q：量词，d:副词等。

（3）重叠的形式

重叠形式包括“AAB、A 一 AB、A 了 AB、A 没 AB、A 不 AB”，主要是前面动语素的重叠。

3.1.2 插入多种成分的规则总结

离合词插入多种成分数量比较多，其形式不易总结，而且没有太多的规律可遵循，下面先根据离合词的离析长度对插入多种成分的长度进行限定。

1 离合词的离析长度

本文根据中间长度来确定规则，表 3 是 140 个离合词的离析长度统计与分析：

表 3　离合词离析形式的长度

| 插入字数 | 实例条数 | 插入字数 | 实例条数 | 插入字数 | 实例条数 |
|---|---|---|---|---|---|
| 1 | 43312 | 6 | 2615 | 11 | 149 |
| 2 | 45422 | 7 | 1300 | 12 | 100 |
| 3 | 20526 | 8 | 808 | 13 | 69 |
| 4 | 11272 | 9 | 416 | 14 | 41 |
| 5 | 5438 | 10 | 266 | 15 | 31 |

从 140 个离合词离析长度的分布情况可以看出，离合词的离析长度主要集中在 12 个字以内，多于 12 个字的出现很少，并且中间修饰的成分比较多。由表 3 可以看出离合词中间插入长度为 5 个字以内的所占比例最多。根据离合词的这一特点，本文在自动识别中对规则的总结主要限制在三个成分内。对于个别例句超过三个成分的情况，在自动识别过程中一律用符号“*”处理，对规则进行总结时不做细化归类。如：“那/r 晚/Tg 我/r 睡/v 了/u 一/m 个/q 特别/d 舒服/a 的/u 觉/Ng”，在总结规则分别划分到“睡/v 了/u 一/m 个/q”

这个层面，对于后面的成分一律用“*”表示。

2 插入多种成分的总结

本文在离析长度的基础上，充分考虑可行性和有效性两个方面，从自动识别的角度，在对规则进行总结时只限定在三个成分以内。

（1）A+r+m/q/r/的+B，中间插入成分为代词，后面加数词、量词、代词、结构助词“的”，比如“帮这么点忙”、“沾别人的光”等；

（2）A+n+m/q/的+B，中间插入成分为代词，后面加数词、量词、结构助词“的”，比如“生爸爸的气”、“见老师一面”等；

（3）A+着/了/过+m/q/r/a/n+B，中间插入成分为“了、着、过”，后面加数词、量词、代词、形容词、名词，比如“吃了这个亏”、“发着高烧”等；

（4）A+m/q+r/a/n+B，中间插入成分为数词或量词，后面加代词、形容词、名词，比如“洗个温水澡”、“吃一大惊”等；

（5）A+d+v/d/u+B，中间插入成分为副词，后面加副词、助词、动词，比如“出不了院”、“帮不到忙”等；

（6）A+m+q+B，中间插入成分为数量短语，比如“冒一次险”、“沾一回光”等；

（7）A+m+q+a/n+B，中间插入成分为数量短语，后面加名词、形容词，比如“听一次妈妈话”、“睡一个好觉”等；

（8）A+了/过+m+q+B，中间插入成分为“了、过”，后面加数量短语，比如“见了一次面”、“叹了一口气”等。

根据对离合词中间插入成分的总结，本文将其分为 4 个集合，放在 4 个文本文件中，分别是：汉字集合、词性集合、重叠集合、插入多种成分的集合。（具体集合的规则见附录 2）

### 3.2 离合词自动识别的具体过程

在识别过程中，读入的文本经过分词和词性标注的预处理，已经被切分为相对独立的成分。下面将离合词离析形式自动识别的步骤进行具体阐述。

（1）将测试集中的 120 个离合词放入文本文件中，而离合词是实验前已事先准备的词表。由于实验过程仅对离合词的离析形式进行自动识别，而对离合词本身在语料中的使用情况不作识别标注。离合词词表文件读入程序；并且 120 个离合词的人工标注语料，包括正确和错误的离析形式，也被读入程序。

（2）离合词的四个规则文本文件依次读入程序中，当分词和词性标注的语句经过正则表达式时，依据规则的判断进行自动标注。如果匹配到规则，则机器自动标注为“1”，输出到一个新的文件里；如果没有匹配到规则，则进入到下一个正则表达式中进行匹配。

（3）识别的基本顺序。当人工标注的文本进入正则表达式中时，识别的顺序是先识别具有明显形式标记的汉字，识别不到的话则进入词性规则的匹配中；再进入重叠规则的匹配中；最后进入插入多个成分的规则中。没有匹配到的语言实例，机器会自动标记为“0”，输出到一个新的文件里。

比如：“不是/c 怒目而视/n 就是/v 和/c 他/r 大/d 吵/v 一/m 架/q”，符合“A+m+B”的规则，被自动标注为“1”；“睡梦/n 中/f 一/m 觉/Ng 醒来/v”不符合所有的规则，便被自动标注为“0”。

## 4 实验结论及分析

### 4.1 正确率和召回率的计算结果

本文运用 Perl 程序语言进行自动识别标注，并通过人工标注与自动识别标注的结果来计算正确率与召回率，以下是计算公式：

正确率 = 自动识别正确的信息条数 / 提取出的信息条数；

召回率 = 自动识别正确的信息条数 / 样本中的信息条数。

以下是 120 个离合词的正确率和召回率：

表 5　120 个离合词的正确率

| 正确率 | 标 1 的正确率 | 标 0 的正确率 |
|---|---|---|
| 79.3% | 66.8% | 91.8% |

表 6　120 个离合词的正确率与召回率

| 标 1 的正确率 | 标 1 的召回率 |
|---|---|
| 66.8% | 84.3% |

如表 5 所示，120 个离合词总的正确率在 79.3%左右，其中自动识别为 1 的正确率为 66.8%，自动识别为 0 的正确率为 91.8%。相比来看，自动识别为 1 的正确率比较低。从表 6 的数据结果来看，自动识别为 1 的召回率比正确率要高得多，可能是提取出的语言实例过多导致正确率偏低。下面本文分别从正确率和召回率两个方面对每个离合词自动标注的情况进行分析，部分离合词的实验数据如表 7 所示。

表 7 部分离合词的正确率和召回率

| 离合词 | 自动标注句数 | 标注正确句数 | 人工标注句数 | 正确率 | 召回率 |
|---|---|---|---|---|---|
| 睡觉 | 4711 | 4658 | 5628 | 98.4% | 82.8% |
| 搞鬼 | 1860 | 1811 | 1841 | 97.4% | 98.3% |
| 沾光 | 628 | 599 | 699 | 95.8% | 85.7% |
| 告状 | 606 | 564 | 684 | 93.1% | 82.5% |
| 散步 | 245 | 225 | 246 | 91.8% | 91.4% |
| 操心 | 580 | 523 | 561 | 90.2% | 93.2% |
| 毕业 | 295 | 263 | 263 | 89.2% | 100% |
| 跳舞 | 2018 | 1713 | 2108 | 84.9% | 81.3% |
| 见面 | 8716 | 5534 | 6342 | 63.5% | 87.3% |
| 丢人 | 618 | 231 | 286 | 37.3% | 80.8% |

从表 7 的部分统计数据可以看出，其中“睡觉”、“搞鬼”、“沾光”、“告状”、“散步”、“操心”这 6 个离合词的正确率均达到 90%以上，但是只有“搞鬼”、“操心”两个词的召回率在 90%以上。而“见面”、“丢人”的这两个词的正确率比较低，分别为 63.5%和 37.3%。离合词“毕业”的召回率达到 100%，但是正确率却只有 89.2%。

在测试集的 120 个离合词中正确率在 90%以上的只有 21 个，所占比例仅为 17.5%；而召回率在 90%以上的有 82 个词，所占的比例为 68.3%。下面对具体离合词的标注结果和数据语料进行研究，以分析自动识别正确率低的原因。

### 4.2 离合词正确率和召回率低的原因

#### 4.2.1 提取规则过于宽泛

从表 7 的统计数据，本文还发现一种特殊的现象，就是自动识别的句子数量远远多于人工标注的句子数量。如“见面”“丢人”。比如：“见面”人工标注为 1 的语言实例为 6342 句，而自动提取的结果却有 8716 句，所以导致两个词的正确率比较低。结合表 6 的数据，分析出正确率偏低的原因之一是在自动识别过程中，存在很多把错误实例标成正确实例的情况，导致自动提取的数量大大增加。为了验证不是个例的现象，本文对其他离合词进行统计。

在 120 个离合词中有 97 个离合词，自动识别为 1 的句子数量多于人工标注的句子数量，大约占 80.8%，也就是说 80%以上的离合词出现识别错误的情况。分析原因，一方面是语料本身可能出现错误；另一方面通过分析发现，对插入多种成分时所总结的规则过于宽泛。当离合词离析形式出现多于两个或三个成分时，超出规则长度的成分用“*”代替，导致自动识别过程中标注为 1 的数量自然大大增加。

本文对 120 个离合词自动识别错误的数量进行统计，人工标注为 1 而自动识别标注为 0 的例句数是 9436 句，人工标注为 0 而自动识别标注为 1 的例句数为 38621 句。自动标注为 1 的句子数量所占的比重较大，是导致自动识别正确率比较低的一个重要原因。

4.2.2 离合词前后语素构词能力强

通过分析发现，例如“毕业”、“沾光”等这些词识别标注的正确率比较高。因为离合词的前后语素中包含粘着语素，它们的动语素“毕”、“沾”为粘着语素，由于其自身的粘着性，它们在实际语言运用多与名语素“业”、“光”构成离合词。通过程序中的规则验证，其离析形式就很容易被程序自动识别标注出来。但是，像“干杯”“当面”等词，它们的前后语素均为自由语素，而且有些语素还是多音字。由于语素自由性比较大，构词能力比较强，因而比较容易构成新词。在自动识别过程中，只识别单个语素，并未作任何限定。从统计数据可以看出，包含粘着语素离合词的识别正确率要高于包含自由语素离合词的识别正确率。

**4.3 程序优化的数据分析**

针对上面的两个原因，对程序进行优化，以提高其自动识别的正确率。针对提取规则过于宽泛，将正则表达式的规则读取限定在四个成分以内。具体优化过程为：之前总结的规则不变，在规则读入正则表达式时，对离析长度的限定做了改变。当离合词中间插入三个成分时，后面再加上一个词表符号“/”，不对插入成分做具体词性的处理；而当插入成分为两个时，要在后面加上两个词性符号“/”。

另一方面，优化过程中对离合词的前后语素做了限定，要求分词结果独立，不得与其他语素组合成词，借用词性符号“/”对离合词的语素做了限定。比如“吹/v 什么/r 牛/n”会被规则“A+什么+B”提取，而“吹/v 什么/r 牛皮/n”则不会被提取。

程序优化之后，正确率和召回率得到很大的提升。下面是优化之后 120 个离合词的正确率和召回率：

表 8　120 个离合词的正确率

| 正确率 | 标 1 的正确率 | 标 0 的正确率 |
|---|---|---|
| 91% | 91.6% | 90.6% |

表 9　120 个离合词的正确率与召回率

| 标 1 的正确率 | 标 1 的召回率 |
|---|---|
| 91.6% | 89.3% |

通过上面表 8 和表 9 与之前的表 5 和表 6 对比，可以看出，正确率和召回率都有了很大的提高，尤其是自动识别为 1 的正确率，由优化之前的 66.8%提升到优化之后的 91.6%。

本文又统计了自动标注与人工标注的对比情况。正确的标成错误的共有 8246 句，而错误的标成正确的有 6725，相对之前的 38621 句，其数量大大下降。所有离合词的语料实例为 23 万多句。加入限定条件对程序进行优化后，在一定程度上使很多非离合词的离析形式被过滤掉。以下是部分离合词优化之后的结果，具体见表 10：

表 10 部分离合词的正确率和召回率

| 离合词 | 自动标注句数 | 标注正确句数 | 人工标注句数 | 正确率 | 召回率 |
|---|---|---|---|---|---|
| 睡觉 | 5378 | 5374 | 5628 | 99.93% | 95.49% |
| 搞鬼 | 1824 | 1809 | 1841 | 99.02% | 98.37% |
| 沾光 | 672 | 664 | 699 | 98.80% | 95.00% |
| 告状 | 642 | 626 | 684 | 97.50% | 91.50% |
| 散步 | 232 | 230 | 246 | 99.14% | 93.50% |
| 操心 | 537 | 519 | 561 | 96.65% | 92.51% |
| 毕业 | 263 | 263 | 263 | 100% | 100% |
| 跳舞 | 1989 | 1934 | 2108 | 97.23% | 91.75% |
| 见面 | 5924 | 5201 | 6342 | 87.80% | 82.24% |
| 丢人 | 261 | 219 | 286 | 83.91% | 76.57% |

与表 7 的数据结果相比，这几个离合词自动识别的正确率得到显著的提高，特别是离合词“见面”和“丢人”。从 120 个离合词离析形式的数据结果来看，正确率在 90%以上的离合词有 66 个，所占比例为 55%；其中正确率为 100%的离合词有 13 个，如“吵架”、“叹气”、

“碍事”等。正确率在 80%以上的离合词有 92 个，所占比例为 76.67%。正确率在 70%以上的离合词有 98 个，所占比例为 81.67%。而正确率在 50%以下的仅为 10 个，所占比例为 8.3%，其中“集邮”识别的正确率为 0。“集邮”在人工标注中没有正确的例句，所以标 1 的正确率和召回率均为 0，但是标 0 的正确率却为 100%。这从反面也印证自动识别算法的有效性。

正确率和召回率均为 100%的离合词有 6 个词，包括：“毕业”、“贬值”、“延期”、“减产”、“行贿”、“执勤”。有些离合词正确率高，但是召回率却比较低，比如“拨款”正确率为 100%，召回率只有 35.3%；“碍事”正确率为 100%，召回率只有 78.5%。分析发现，“拨款”语料中很多出现词性标注错误的情况，如“拨/v 救灾/vn 款/n”中“救灾”标成动名词“vn”，所占的比例很大，导致很多例句没有被提取出来，召回率较低；“碍事”的语料多使用“事儿”，由于对前后语素做了限定，很多例句没有被提取出来，召回率较低。

正确率在 50%以下的离合词有 10 个。如：“吹牛”、“行军”、“起哄”等。除了“集邮”之外，剩下的 9 个词在实际语言生活中很少出现离析形式。在它们前后语素中间插入其他成分，很可能不是其离析形式，而又符合提取的规则，所以导致识别的正确率比较低。如：

（1）“配/v 一/m 套/q 红宝石/n 钻/v 饰/v”，符合“A+ m +B”的规则，不是“配套”的离析形式；

（2）“给/p 杨局长/nr 行/v 了/u 个/q 军/n 礼/Ng ”，符合“A+ 了 +q +B”的规则，不是“行军”的离析形式。

这些离合词的召回率比较高，说明这些词符合规则的正确语料都被提取出来。有些离合词的召回率达到 100%，如“行军”正确率为 46.2%，而召回率为 100%；“起哄”正确率为 44%，召回率为 100%；“配套”正确率为 32.6%，召回率为 93.3%。（120 个离合词的正确率和召回率见附录 3）

从识别效果来看，本文的识别结果与冯向华（2009）的结果相比得到了很大的提升。冯向华设计的程序对主谓型及动补型离合词的识别效果好于动宾型离合词，对插入成分封闭的离合词的识别效果好于插入开放成分的离合词，对插入一个和多个成分的离合词扩展形式的识别效果区别不大。而本文主要是对动宾型的离合词进行自动识别研究，相比动宾型和主谓型的离合词，其在实际语言生活中更易出现离析形式。从整体来看，本文的识别效果达到 91.6%的正确率；并且对插入多个成分识别也有较高的准确率。比如“睡觉”、“吃惊”等。

### 4.4 影响正确率和召回率的主要因素

对程序进行优化之后，自动识别的正确率和召回率都得到很大的提升，但是有些离合词自动识别的正确率仍然比较低。下面是影响正确率和召回率的主要因素：

1 影响正确率一个主要原因是离合词的前后语素构词能力比较强

虽在程序优化中对离合词的前后语素做了限定，但仍有个别离合词存在歧义的情况。正确率在 50%以下的离合词大多是因为前后语素构词能力强。

2 中文分词或词性标注错误对自动识别的影响

本文自动识别的方法主要是依据词性，如果在分词处理上出现差错或是歧义切分，词性标注错误则直接导致自动标注的错误。如:“如果/c 当/v 着/u 她/r 的/u 面谈/vn 话/n”。

3 对插入成分的规则限制单一

在程序的设计上，插入成分限定在四个以内，由于插入成分复杂，不宜对超出的成分进行具体词性的限制，这可能就导致很多正确的离析形式没有被提取出来。

4 语料本身存在的问题

本文所用的语料来自 BCC。由于 BCC 语料内容比较多，来源比较广泛，语料本身不可避免地存在很多错误，也可能导致识别错误的情况。比如：“什么”写成“甚么”，便不能被正确识别。

5 人工标注的疏漏

进行人工标注时，有些离合词离析形式较多，在 BCC 中的例句数达到上万句。人工标注难免会有疏漏，正确率不可能达到百分之百，因此自动识别的正确率和召回率也会受到影响。

## 5 总结与展望

本文在大规模语料的基础上，对离合词离析形式进行自动识别，一方面对离合词的本体研究进行补充和完善，另一方面，同时也为离合词在中文信息处理方面的研究提供一定的借鉴意义。另外，在识别中所使用的规则在一定程度上也验证了离合词的不同离析形式。本文只选取 140 个离合词进行研究，再加上自动识别程序的局限性，对于一些特殊的离合词还不能进行有效地识别。下一步的工作希望扩展到汉语所有离合词的研究；其次，在研究方法上可以考虑从离合词外部因素入手，借助离合词的上下文进行自动识别研究。

## 参考文献：


[1]陆志韦.汉语的构词法[M]. 北京：科学出版社，1957：38-40.

[2]王海峰，李生等.汉英机器翻译中汉语离合词的处理策略[J].情报学报，1999，04：303-305.

[3]王春霞. 基于语料库的离合词研究[D]. 北京：北京语言文化大学，2001.

[4]史晓东.汉英机器翻译中离合词的处理[C].黄河燕.全国机器翻译研讨会论文集.北京：电子工业出版社，2002：69-72.

[5]徐建山. 汉语离合词和长距离搭配的研究[D].哈尔滨：哈尔滨工业大学，2003.

[6]任海波，王刚.基于语料库的现代汉语离合词形式分析[J].语言科学，2005，04：81-84.

[7]周卫华，胡家全. 中文信息处理中离合词的处理策略[J]. 三峡大学学报，2010，06：41-44.

[8]冯向华. 现代汉语文本中离合词扩展形式的自动识别[D]. 北京：北京师范大学，2009.

[9]刘博.基于语料库的离合词扩展形式自动识别研究[D].保定：河北大学.2015.

[10]荀恩东，饶高琦，臧娇娇等. 大数据背景下 BCC 语料库的研制[J].语料语言学，2016，01：91-106.


## 附录

### 附录 1　140 个离合词

碍事 罢工 拜年 帮忙 保密 报仇 报名 毕业 闭幕 贬值 变形 变质 拨款 补课 参军 操心 插嘴 吵架
吵嘴 称心 吃惊 吃苦 吃亏 抽空 出差 出神 出院 吹牛 辞职 打架 打猎 打仗 打针 带头 担心 当面
捣蛋 捣乱 倒霉 到期 道歉 登记 定性 丢人 懂事 动身 发烧 放假 放心 放学 分红 干杯 搞鬼 告状
鼓掌 挂钩 挂号 拐弯 害怕 害羞 狠心 化妆 怀孕 灰心 集邮 及格 加油 剪彩 减产 见面 讲理 接班
结果 结婚 敬礼 就业 鞠躬 决口 绝望 开刀 开课 开幕 考试 旷工 旷课 劳驾 离婚 理发 聊天 留意
埋头 满月 冒险 纳闷 配套 拼命 破产 起草 起床 起哄 请假 请客 让步 入学 散步 伤心 上当 生气
升学 失学 失业 失约 睡觉 探亲 叹气 提醒 跳舞 听话 投标 完蛋 握手 洗澡 献身 泄气 行贿 行军
宣誓 延期 要命 移民 迎面 游泳 遭殃 沾光 站岗 照相 争气 执勤 注册 着急

### 附录 2　规则集合

**插入汉字的集合：**了、过、过了、着、个、什么、的、上、不上、完、不完、好、不好、起、成、不成、得、不得、不了、不到、一、大、高、闷、透、尽、碎、足

**插入词性的集合：**m q a v n r d f

**重叠集合：**AAB、A 一 AB、A 了 AB、A 没 AB、A 不 AB

**插入多种成分的集合：**A+r+m/q/r/的+B、A+n+m/q/a/的+B、A+着/了/过+ m/q/r/a/n+B、A+m/q+r/a/n+B、A+d+v/d/u+B、A+m+q+B、A+m+q+a/n+B、A+了/过+m+q+B

### 附录 3　120 个离合词的正确率和召回率

120 离合词的正确率和召回率详见网址：https://pan.baidu.com/s/1c13zAak

**作者简介：**

臧娇娇（1990—），女，文学硕士，主要研究领域为计算语言学，北京语言大学大数据与语言教育研究所，Email:qiaolidiefei528@163.com，电话：18813151602。

荀恩东（1967—），男，教授，主要研究领域为自然语言处理、计算机教育技术，北京语言大学大数据与语言教育研究所，Email: edxun@126.com，电话：010-82300316。