文章编号: 1003-0077(2007)02-0052-06

# 基于统计抽词和格律的全宋词切分语料库建立

苏劲松[1,2],周昌乐[2],李翼鸿[2]

(1. 厦门大学 软件学院,福建 厦门 361005; 2. 厦门大学 人工智能研究所,福建 厦门 361005)

**摘 要**: 全宋词切分语料库的建立是计算机研究宋词的基础。本文对宋词中"词"的界定提出了自己的看法,并在综合考虑统计抽词方法和基于诗词格律切分方法各自优点的基础上,提出建立全宋词切分语料库的新方法。我们首先通过统计抽词来抽取结合程度较强的二字词,并结合相关资源建立词表;在此基础上,结合宋词的格律特点按照一定的规则来对全宋词进行了切分。实验证明,本文中的方法具有较好的效果。

**关键词**: 计算机应用;中文信息处理;宋词;语料库;统计抽词;格律

**中图分类号**: TP391 **文献标识码**: A

## The Establishment of the Annotated Corpus of Song Dynasty Poetry Based on the Statistical Word Extraction and Rules and Forms

SU Jin-song[1,2], ZHOU Chang-le[2], LI Yi-hong[2]

(1. Software School of Xiamen University, Xiamen, Fujian 361005, China;
2. Institute of Artificial Intelligence of Xiamen University, Xiamen, Fujian 361005, China)

**Abstract:** The annotated corpus of Song Dynasty poetry is the foundation of the computer-based study of Song Dynasty poetry. In our paper, we propose a new definition of "word" in the Song poetry and a new method for the establishment of the annotated corpus. Two available methods, statistical word extraction and segmentation based on rules and forms, are taken into consideration. The former is adopted to extract closely combined two-character words and establish word lists combining with related resources. And the latter, combined with the word lists, is used to segment Song Dynasty poetry. It is showed by the experimental results that the method applied in the paper is effective.

**Key words:** computer application; Chinese information processing; Song Proses; annotated corpus; statistics-based word extraction; rules and forms

## 1 引言

中国古代诗词是一种特殊文体的大众化文学形式,在汉语文化的成长、演变与传播中有着极重要的地位,而其中的宋词作为宋代文学的典范,赢得了众多文人骚客的青睐,成为中国古代诗词中的一颗璀璨的明珠。因此,通过对宋词进行研究进而了解宋代文化一直是语言文学研究工作者的一个研究热点。

20 世纪 80 年代以来,随着计算机应用技术的不断发展,以语料库为基础的研究在语言学和计算机科学研究中都取得了丰硕的成果。无论是在语言学研究还是自然语言处理领域,语料库都已经成为重要的基础资源,发挥了越来越重要的作用。正是基于以上认识,我们建立了全宋词语料库,结合宋词本身所具有的特点,并运用各项自然语言处理技术来进行宋词的计算机辅助研究。而这整个研究的基础就是宋词语料库的建立,可以说宋词语料库的建立具有非常重要的意义。

**收稿日期**: 2006-05-10 **定稿日期**: 2006-11-15

**基金项目**: 国家自然科学基金资助项目(60373080)

**作者简介**: 苏劲松(1982—),男,硕士生,研究方向为计算语言学。

## 2　全宋词切分语料库建立方法的提出

宋词语料库的建立必然会涉及到宋词的切分,在诗词切分方面,北大计算语言所与台湾地区元智大学都进行过相关研究并且取得了很好的研究成果。北大计算语言所通过纯统计的方法[1]将结合强度较强、使用稳定以及带有隐喻义的二字词抽取出来,为建立词表奠定了良好的基础;台湾地区元智大学罗凤珠教授则主要根据诗词格律[2]来切分诗词。经过人工切分证实,这种方法的切分点绝大部分都是正确的,有助于分词。

在此,本文参考以上两种方法,提出了结合以上两种方法的优点来建立宋词切分语料库。首先,对如何界定"词"提出了自己的看法;然后,分析宋词切分和格律之间的关系,建立词体格律数据库;再次,根据词体格律数据库把词句细分为子句,对子句字串进行统计,从中抽取结合强度较强的二字字串,并结合各种已有的词典资源来建立宋词词表;最后,根据古代诗词切分规则再对子句进行进一步细分,并根据词表来切分最后的子句。本文方法的主要框架流程如下:

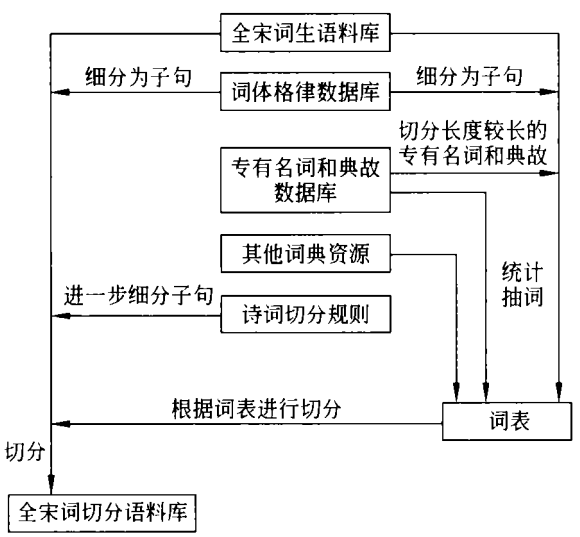


图 1　方法框架流程图

## 3　全宋词中"词"的概念界定

要对全宋词进行分词,首先要明白如何对"词"进行界定。在研究中国古代诗词和现代汉语中对词汇的定义之后,可以发现以下 4 种类型的字串可以切分为词:

(1) 宋词中大量采用领字。领字是指在词意转折处,使上下句相连接,起过渡或联系作用的字。而其中的单字领字具有独立的意义,可以单独切分出来独立成词。例如:"过沙溪急,霜溪冷,月溪明"(作者: 苏轼 词牌: 行香子)中的"过"就为单字领字,可以单独切分出来为词。

(2) 诗词中含有大量的专有名词。这些字串都可以直接作为一个词汇单位,收录进词表。例如:"补天又笑女娲忙"(作者: 辛弃疾 词牌: 归朝欢 标题: 题晋臣积翠岩)中的"女娲"为人名专有名词。

(3) 诗词中含有大量的典故。这些字串也可以直接作为一个词汇单位,收录进词表。例如:"骑鹤上扬州"一句,来源于南朝梁殷芸《小说》卷六:"有客相从,各言所志,或愿为扬州刺史,或愿多赀财,或愿骑鹤上升。其一人曰:'腰缠十万贯,骑鹤上扬州。'欲兼三者。"后以此比喻欲集做官、发财、成仙于一身,或形容贪婪、妄想。

(4) 在语言发展过程中,有些字串也许刚开始不是作为一个词,但是由于它们结合紧密,使用稳定,并且往往有特定的含义,因而人们也把它们当作一个词了。对此,引入统计手段来衡量两个字之间的结合强度。如果这些相邻的二元字串结合强度足够大,则对其进行判断。例如:"落花、春水"等在现代汉语中通常不被认为是词,但它们使用频度很高,具有明显的统计特征。分析发现,这类词汇一般都具有较为明确的隐喻义,完全可以将其收入进词表。另外一些词如"牙床、小槽、代北"等由于社会环境的变化,在现代汉语中已经不是词或词义发生根本变化,但在古汉语中的确是词,也要将其收入进词表[1]。

## 4　全宋词切分和格律的关系及词体句法数据库的建立

通过对部分宋词进行手工切分,可以发现宋词切分和格律存在一定关系: 词句可以根据对应的句法来细分成子句。以"舍北烟霏舍南浪。雪倾篱、雨荒薇涨。问小桥、别后谁过,惟有迷鸟羁雌来往。重寻山水问无恙。扫柴荆、土花尘网。留小桃、先试光风,从此芝草琅玕日长。"(作者: 范成大 词牌: 宜男草)为例,根据词谱可以知道"宜男草"上阙和下阙的最后一句都为上二下六的句法,于是可以将其细分为:"{惟有}{迷鸟羁雌来往}"、"{从此}{芝草琅玕日长}"(其中{}表示根据句法细分成的子句)。

宋词格律比较复杂,词牌下的词体句法都有固定的总字数、总句数,每一句的字数也是固定的。最短的词句是一字句,只出现在词牌“哨遍”和“钗头凤”,显然可以独立成词。二字句出现在“调笑令”、“如梦令”和“醉翁操”等词牌,能否成词可以根据两个字之间的结合强度来判定。十字句只有词牌“摸鱼儿”前阕第六句和后阕第七句,为上三下七句法。十一字句只有词牌“水调歌头”根据不同的词体句法有上六下五和上四下七两种句法。三字句至九字句占的数量最多,其句法分别是:1)三字句:上二下一、上一下二;2)四字句:上二下二;3)五字句:上二下三、上三下二、上一下四;4)六字句:上二下四、上四下二、上三下三;5)七字句:上三下四、上四下三、上一下六;6)八字句:上三下五、上四下四、上一下七、上二下六;7)九字句:上三下六、上四下五、上五下四、上六下三。这些上下句之上句若是奇数字句,多数句子之第一个字是单字领字,这种情况在五字句之上一下四,七字句之上一下六是最多的[2]。

本文以潘慎等人编著的《词律辞典》[3]、《钦定词谱》[4]、龙榆生编著的《唐宋词格律》[5]、王兆鹏等人编著的《宋词大辞典》[6]和陆辅之编著《词旨》[7]为基础,建立词体句法数据库,该数据库含有不同词体的句法 2 415 种,标注了各词体的单字领字位置和句法。一个词牌对应一种或者多种词体,大部分词体的句法都是惟一确定,但是,同一种词体也有可能存在不同句法,经过统计,有 37 种词体含有不同句法。对此,本文把同种词体的不同句法全部收录进数据库,并在生语料库中对该类词体的宋词所属词体句法类别进行人工标注(同种词体下的宋词,字数分布是一样的,因而计算机不能自动识别出拥有多种句法的词体所对应的宋词是属于哪种句法,而且该类词体所包含的词并不多,所以采用人工标注的方法)。

## 5 基于统计抽词的全宋词词表建立

### 5.1 专有名词和典故数据库的建立

宋词中存在有大量的专有名词和典故,如人名、地名或者特殊含义等。但是有时这类词或者为长度大于 2 的多字词,或者出现次数非常少而没有显示出很强的结合强度,为了保证词表能有较高的覆盖率,就需要建立专门的数据库来收集该类词语。

以王兆鹏等人编著的《宋词大辞典》[6]为基础,结合前人所作的一些归纳[2],本文建立专有名词数据库,该数据库共分为人名、地名、天文、时令、音乐、人伦、人事、闺阁、形体、文事、珍宝、建筑、服饰、饮食、草木百花 15 大类;以金启华等人编著的《全宋词典故考释辞典》[8]和台湾地区元智大学罗凤珠教授的诗词典故资料数据库为基础,建立明典数据库。以上两个数据库共含有词条 6 873 条。

### 5.2 统计抽词

诗词中除了单字领字、专有名词和典故之外,有些字串结合紧密,使用稳定,并且代表特定的含义,因而也把它们当作一个词。对此,本文运用统计学的方法来抽取全宋词中的二字词。为了提高抽词的准确率,在统计抽词之前,需要对语料库进行预处理,主要包括:

**步骤 1** 查询词体句法数据库,根据对应词体句法的句法将词句细分为子句,同时标出词句中的单字领字,不参与统计抽词的计算。

**步骤 2** 查询专有名词数据库和典故数据库,与词句进行匹配,标出其中含有的长度大于等于 3 的专有名词、典故,不参与统计抽词的计算。

经过抽样统计,标出单字领字和长度大于等于 3 的专有名词、典故,正确率可以达到 97 %。通过以上两个步骤,为统计抽词奠定了良好的基础。

(1) 频率与互信息

在九种常用的抽词统计量中[9],选择用互信息来进行二字词自动抽取,并结合频率来改进互信息的提取效果,这种方法即简单又有效。

$$I(xy) = \ln \frac{P(xy)}{P_1(x) \times P_2(y)}$$

其中,$P_1(x)$ 表示字 $x$ 在语料库所有相邻二元字串中作为前字出现的概率;$P_2(y)$ 表示字 $y$ 在语料库所有相邻二元字串中作为后字出现的概率;$P(xy)$ 是汉字 $x, y$ 在语料库中同现的概率。需要说明的是,互信息是一种衡量二元字串之间出现信息增加的统计量,但是它并不是衡量二元字串之间依赖性的好方法。由低频率字组成的相邻二元字串的互信息要大于高频率字组成的二元字串,而这会影响互信息的使用效果,在此使用频率截断的方法,只考虑出现频率大于某个阈值的相邻二元字串,以此来改进互信息提取的效果。

(2) 共现度

在汉语中,词往往都是由不同字组合而成的。

考虑到互信息在衡量二元字串之间依赖性上的不足,本文采用共现度[11]作为补充,用于衡量相邻二元字串中字的相互依赖性。

$$C(x,y) = R_1(xy,x) + R_2(xy,y)$$

$$R_1(xy,x) = \frac{f(xy) \times \ln(f(xy))}{f_1(x) - \sum f(xz_i)}$$

$$P_2(xy,y) = \frac{f(xy) \times \ln(f(xy))}{f_2(y) - \sum f(u_j y)}$$

其中,$f(xy)$为相邻二元字串 $xy$ 在语料中的出现次数, $f_1(x)$为字 $x$ 在语料库作为前字出现的次数,$f_2(y)$为字 $y$ 在语料库中作为后字出现的次数,$f(xz_i)\ (i=1,2,\cdots,n_1)$为已经抽取的以 $x$ 为前字的二元字串的出现次数,$f(u_j y)\ (i=1,2,\cdots,n_2)$为已经抽取的以 $y$ 为后字的二元字串的出现次数。

### 5.3　全宋词词表的建立

至此,可以结合利用已有的三个统计量来统计抽词,并用建立好的专有名词和典故数据库进行补充,以此来构造初步词表。词表建立过程主要由以下三个步骤组成:

**步骤 1**　对所有出现频率大于等于 3 的相邻二元字串,计算其互信息,提取互信息大于等于 3.5 的所有相邻二元字串,该部分字串具有较高可能性为词。

经统计,频度大于 3 且互信息大于等于 3.5 的相邻二元字串一共有 17 353 条,其中可以成词的字串为 10 825 条。

**步骤 2**　对剩下频率大于等于 3 的相邻二元字串进行共现度迭代计算,每次提取共现度最大的 500 个相邻二元字串,然后递进地计算剩下的相邻二元字串的共现度,当连续 3 次迭代计算抽词准确率都低于 20 %(共现度迭代计算的抽词准确率不是一条严格的递减函数,在逐渐递减的过程中,会产生上下波动,该准确率阈值和语料库大小有关),停止迭代计算。

实验中,经过 53 轮迭代计算,共现度迭代计算抽词达到预先设置的阈值,抽词停止。经统计,抽出的相邻二元字串一共有 26 501 条,其中可以成词的字串为 9 044 条。迭代计算抽词实验过程中,抽出的词条数目、抽词准确率和平均迭代次数的变化过程如表 1 所示:

其中,第 $t$ 次迭代抽词准确率和到第 $t$ 次迭代的平均迭代次数如下定义:

表 1　共现度迭代计算过程变化

| 迭代计算次数 t | 抽出的词条个数 | 抽词准确率 | 平均迭代次数 |
|---|---|---|---|
| 1 | 429 | 85.8 % | 1.00 |
| 2 | 391 | 78.2 % | 1.48 |
| 3 | 355 | 71.0 % | 1.94 |
| 4 | 329 | 65.8 % | 2.39 |
| 5 | 317 | 63.4 % | 2.84 |
| 6 | 299 | 59.8 % | 3.29 |
| 7 | 272 | 54.4 % | 3.71 |
| 8 | 257 | 51.4 % | 4.13 |
| 9 | 239 | 47.8 % | 4.53 |
| 10 | 241 | 48.2 % | 4.95 |
| 11 | 225 | 45.0 % | 5.36 |
| 12 | 200 | 40.0 % | 5.73 |
| 13 | 207 | 41.4 % | 6.13 |
| 14 | 201 | 40.2 % | 6.53 |
| 15 | 175 | 35.0 % | 6.89 |
| 16 | 178 | 35.6 % | 7.26 |
| 17 | 171 | 34.2 % | 7.64 |
| 18 | 164 | 32.8 % | 8.00 |
| 19 | 179 | 35.8 % | 8.41 |
| 20 | 152 | 30.4 % | 8.76 |
| 21 | 171 | 34.2 % | 9.17 |
| 22 | 145 | 29.0 % | 9.52 |
| 23 | 150 | 30.0 % | 9.89 |
| 24 | 167 | 33.4 % | 10.31 |
| 25 | 151 | 30.2 % | 10.70 |
| 26 | 151 | 30.1 % | 11.09 |
| 27 | 131 | 26.2 % | 11.43 |
| 28 | 131 | 26.2 % | 11.78 |
| 29 | 126 | 25.2 % | 12.13 |
| 30 | 129 | 25.8 % | 12.48 |
| 31 | 144 | 28.8 % | 12.89 |
| 32 | 117 | 23.4 % | 13.22 |
| 33 | 121 | 24.2 % | 13.58 |
| 34 | 144 | 28.8 % | 14.00 |
| 35 | 131 | 26.2 % | 14.39 |
| 36 | 136 | 27.2 % | 14.79 |
| 37 | 138 | 27.6 % | 15.21 |

续表

| 迭代计算次数 t | 抽出的词条个数 | 抽词准确率 | 平均迭代次数 |
|---|---|---|---|
| 38 | 119 | 23.8% | 15.57 |
| 39 | 121 | 24.2% | 15.94 |
| 40 | 117 | 23.4% | 16.31 |
| 41 | 112 | 22.4% | 16.66 |
| 42 | 112 | 22.4% | 17.02 |
| 43 | 108 | 22.6% | 17.37 |
| 44 | 95 | 19.0% | 17.68 |
| 45 | 109 | 21.8% | 18.04 |
| 46 | 97 | 19.4% | 18.36 |
| 47 | 104 | 20.8% | 18.72 |
| 48 | 107 | 21.4% | 19.08 |
| 49 | 101 | 20.2% | 19.43 |
| 50 | 108 | 21.6% | 19.81 |
| 51 | 90 | 18.0% | 20.12 |
| 52 | 87 | 17.4% | 20.45 |
| 53 | 78 | 15.6% | 20.77 |

第 $t$ 次迭代计算抽词准确率

$$= \frac{\text{第 } t \text{ 次迭代计算抽出来的词条数目}}{\text{第 } t \text{ 次迭代计算抽出来的字串数目}} \times 100\%$$

到第 $t$ 次的平均迭代次数

$$= \frac{\sum_{i=1}^{t} \text{第 } i \text{ 次迭代计算抽出来的词条数目} \times i}{\sum_{i=1}^{t} \text{第 } i \text{ 次迭代计算抽出来的词条数目}}$$

**步骤 3** 对于步骤 2 中剩余的字串和频率小于 3 的字串,实验发现统计抽词的方法效果并不明显,结合《辞源》《现代汉语词典》等词典来对该类字串进行人工判断;并且用前面建立好的专有名词和典故数据库对词表进行补充。

通过以上 3 个步骤建立了初步词表,该词表包含专有名词、典故等,共含有词条 43 387 条,为宋词的机器自动切分奠定了良好的基础。

## 6 全宋词的切分步骤及实验结果分析

### 6.1 切分步骤

本文通过以上步骤初步建立了词表,但是要使得词表覆盖整个语料库并不容易。在此,如前所述,本文结合已生成的词体句法数据库、专有名词数据库、典故数据库,再根据古代诗词切分规则[2],以词句为单位来对全宋词进行切分,以此来提高分词正确率。在此,以"二十四桥仍在,波心荡、冷月无声。念桥边红药,年年知为谁生。"(作者:姜夔 词牌:扬州慢)为例来说明宋词的切分步骤。(其中粗体字表示每个步骤切分出来的词,/ 表示分词符号,{}表示根据句法设置的句法切分点,| 表示根据古代诗词切分规则设置词结构切分点)

**步骤 1** 对比词体句法数据库,设置句法切分点,同时将单字领字切分出来。

"念"为单字领字,领后面两句:"{二十四桥仍在},/{波心荡}、/{冷月无声}。/{念}/{桥边红药},/{年年知为谁生}。/"

**步骤 2** 经过步骤 1 设置句法切分点后,切分成的各子句字数为一到七之间,在这里根据古代诗词切分规则[2]来设定词结构切分点:规则(1),字数为一的字串单字成词;规则(2),剩下字数为二、三的字串不设置词结构切分点;规则(3),剩下字数为四、五的字串,可在第二字后面设置词结构切分点;规则(4),字数为六、七的字串,可分别在第二字和第四字后面设置词结构切分点。

设置词结构切分点如下:"{二十| 四桥| 仍在},/{波心荡}、/{冷月| 无声}。/{念}/{桥边| 红药},/{年年| 知为| 谁生}。/"

**步骤 3** 对比专有名词数据库和典故数据库,将长度大于等于 3 的专有名词和典故切分出来。这类词语应该优先切分,它们切分的优先等级高于步骤 2 设置的词结构切分点。对于字句中切分剩下的字串,结合句法切分点和词结构切分点考虑,如果长度为一,则单字成词。

"二十四桥"为扬州的古桥名,是一个表示地点的专有名词。根据上述切分规则,它的切分优先等级高于步骤 2 设置的词结构切分点,将其先切分出来成词:"{二十四桥/ 仍在},/{波心荡}、/{冷月| 无声}。/{念}/{桥边| 红药},/{年年| 知为| 谁生}。/"

**步骤 4** 根据词结构切分点,结合词表对字串进行进一步切分:规则(1),对于长度为二的字串 XY,如果 XY 有收入进词表,则切分成词 XY/,否则进一步切分为 X/ Y/。规则(2),对于长度为三的字串 XYZ,如果 XY 有收入进词表而 YZ 没有收入进词表时,则将其切分为 XY/ Z/;如果 XY 没有收入进词表而 YZ 有收入进词表时,则将其切分为 X/ YZ/;如果 XY、YZ 同时没有被收入进词表时,则切分为 X/ Y/ Z/;如果 XY、YZ 同时被收录进了词表,则计算 XY 和 YZ 的互信息,如果 I(XY) < = I

(YZ) ,则切分为 X/ YZ/ ,反之,切分为 XY/ Z/ 。

查询词表,“波心”、“无声”、“桥边”、“年年”具有较高的结合强度而被收录进词表;“红药”是专有名词,被收录进词表;“冷月”不具有明显的统计特征,但代表有特殊意义,在其他词典中有收录,因而也被收录进词表。将句中的句法切分点、词结构切分点去除:“二十四桥/仍/在/ ,/波心/荡/ 、/冷月/无声/ 。/念/桥边/红药/ ,/年年/知/为/谁/生/ 。/ ”,全句切分完毕。

### 6.2　结果分析

本文以唐圭璋的《全宋词》[10]为基础,建立了全宋词生语料库,该生语料库收集了宋词 20 162 首,以前面所阐述的方法,对全宋词生语料库进行抽词,建立了全宋词词表并实现了对宋词的初步机器自动分词。从已经人工校对的 3 318 首宋词统计来看,分词正确率(分词正确率为正确切分的词中所含汉字数与总汉字之比)为 83.92 %,说明该方法是有效的。

仔细分析经过人工校对的宋词,发现切分错误主要由以下四个方面造成:

1. 专有名词语料库和典故数据库的不够完善。该类词条一般在统计衡量时都没有显示出明显的统计特征,因此,文章中专有名词语料库和典故数据库的建立对于宋词切分就具有非常重要的意义了,这两个数据库的好坏直接影响着分词正确率的高低。

2. 词体句法数据库的不够完善。词体句法数据库中的单字领字标注直接影响的词句切分正确与否,因此要进一步完善该数据库。

3. 专有名词本身表示的多样性。例如:“广寒”、“广寒宫”都表示月宫。但是,根据文章的切分方法,长度长的专有名词优先切分,因而容易将“广寒宫阙”切分为“广寒宫/阙/ ”,而正确的切分结果应该为“广寒/宫阙/ ”。

4. 包孕型的切分错误。例如:“三月”有可能是表示三个月,也可能表示十二月份中的三月,而两种情况下的切分是不同的。

从实验结果来看,相对于现代汉语的切分,全宋词的切分正确率还有一定差距的。而由于宋词构词的特殊性,现代汉语的切分方法并不适用于宋词的切分,因而需要进一步完善各个相关数据库,才能进一步提高分词的正确率。

## 7　结语

运用计算机来对中国古代诗词进行研究是一个全新的领域,而这所有的研究都是要建立在诗词切分标注语料库的基础上的。本文参考北大和台湾地区元智大学切词方法,提出了一种新方法来建立全宋词切分语料库。实验证明,本文的方法是有效的。

古代诗词的切分既有和现代汉语切分相似的一面,也有他独特的一面。而本文所采用的方法较好地将两方面特点都结合了起来,建立了全宋词切分语料库,为下一步研究奠定了良好的基础。当然,仅仅对语料进行切分是不够的,只有以词汇为单位对进行标注加工,才能对整首宋词的情感、风格、节律等诸多方面进行深入研究。

## 参考文献:


[1] 俞士汶,胡俊峰. 唐宋诗之词汇自动分析及应用[J]. 语言暨语言学,2000,4(3):631-647.

[2] 罗凤珠. 诗词语言切分与语意分类标记之系统设计及应用[A]. 第四届数位典藏技术研讨会[C]. 2005.

[3] 潘慎. 词律辞典[M]. 山西:山西人民出版社,1982.

[4] 钦定词谱[M]. 北京:北京人民出版社,1983.

[5] 龙榆生. 唐宋词格律[M]. 上海:上海古籍出版社,1978.

[6] 王兆鹏,刘尊明. 宋词大辞典[M]. 南京:凤凰出版社,2003.

[7] 陆辅之. 续修四库全书·词旨[M]. 上海:上海古籍出版社,1997.

[8] 金启华. 全宋词典故考释辞典[M]. 吉林:吉林文史出版社,1991.

[9] 罗盛芬,孙茂松. 基于字串内部结合紧密度的汉语自动抽词实验研究[J]. 中文信息学报,2003,17(3):9-14.

[10] 唐圭璋. 全宋词[M]. 上海:中华书局,1997.

[11] 张敏,马少平. 用于信息检索的古文统计分析[J]. 中文信息学报,2001,15(6):41-46.

[12] 俞士汶,段慧明,等. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报,2002,16(5):49-64.

[13] 俞士汶,段慧明,等. 北京大学现代汉语语料库基本加工规范(续)[J]. 中文信息学报,2002,16(6):58-64.

[14] 郑家恒. 二字词词义组合推理方法的研究[J]. 中文信息学报,2001,15(6):1-26.

[15] 孙茂松,邹嘉彦. 汉语自动分词研究评述[J]. 当代语言学,2001,1:22-32.

[16] 王力. 诗词格律概要[M]. 北京:北京出版社,2002.