文章编号：1003-0077(2020)09-001-08

基于高斯混合模型的现代汉语构式成分自动标注方法

黄海斌[1]，常宝宝[2]，詹卫东[1,2,3]

（1. 北京大学 中国语言文学系，北京 100871；2. 北京大学 计算语言学教育部重点实验室，
北京 100871；3. 北京大学 中国语言学研究中心，北京 100871）

摘　要：现代汉语构式成分自动标注作为文本自动标注任务之一，其最大的困难在于，当不存在标注语料作为训练数据时，如何从生语料中挖掘不同类型的构式成分相关的知识并进行标注，特别是面对构式序列在句中的边界难以判断的情况。该文试图借助高斯混合模型聚类方法，结合句中每一个字的位置特征与构式形式本身的语言学特征，融合正则表达式匹配结果信息，挖掘句子中的构式实例序列，并对构式内部成分进行自动标注。相较于仅基于正则表达式匹配和词性匹配的自动标注结果，该方法的 F_1 分别至少提高了 17.9%（半凝固型构式）、19.3%（短语型构式）、14.9%（复句型构式）。

关键词：现代汉语构式；自动标注；高斯混合模型；数据挖掘

中图分类号：TP391　　**文献标识码**：A

GMM-based Automatic Annotation of Chinese Constructions

HUANG Haibin[1], CHANG Baobao[2], ZHAN Weidong[1,2,3]

(1. Department of Chinese Language and Literature, Peking University, Beijing 100871, China;
2. MOE Key Laboratory of Computational Linguistics, Peking University, Beijing 100871, China;
3. Center for Chinese Linguistics, Peking University, Beijing 100871, China)

Abstract: The paper introduces an approach to automatic annotation of Chinese constructions. Without annotated corpora as training data, it is difficult to extract the knowledge of various constructions. To address this issue, we apply the unsupervised method based on Gaussian Mixture Model, the token position features, the linguistic features of construction as well as the regular expressions to capture the structure of the instruction, especially when the boundary is hard to be identified. Comparing to the results annotated by regular expression and part-of-speech, the proposed method achieves improvements on F_1 by 17.9% (for semi-concretionary constructions), 19.3% (for phrasal constructions) and 14.9% (for sentential constructions).

Keywords: Chinese construction; automatic annotation; Gaussian mixture model; data mining

0　引言

在自然语言处理中，针对语言的句法分析，研究者往往十分重视短语结构的处理，而忽视对语言中的构式的分析，致使研究存在局限性。比如，例 1 中的构式“改革一点是一点”，不存在中心成分，且内部成分的功能范畴并不明确，很难以常见的树结构或依存结构进行分析，这势必造成错误的句法分析。要使计算机能够更好地处理现代汉语中的构式，构建现代汉语构式语料库是十分必要的，故而需要大量的构式标注语料，由此衍生出了现代汉语构式标注任务。

对于现代汉语构式标注任务，可以分为构式形式标注和构式义标注两个层面。就前者而言，句中构式内部的成分作为目标语言单位，需要被标记为相应的构式成分标签。对于构式形式的表示，詹卫东[1]提出以常项(constant)和变项(variable)为基本单位，来记录构式的组成。例如，

例 1　一分耕耘一分收获，改革一点是一点，兰

收稿日期：2019-09-04　　**定稿日期**：2019-10-19

基金项目：教育部人文社科基地 2015 年度重大项目(15JJD740002)；国家自然科学基金(61876004)

溪市的做法就说明了这一点。

结合詹卫东[1]对构式的形式表示的研究，例1中的构式可以被描述为“v＋一＋q＋是＋一＋q”，故而符合该模式的序列“改革一点是一点”中的各个成分，需要以词为单位逐个标记为常项(记作C)与变项(记作V)，如表1所示。

表1 构式“改革一点是一点”标注示例

构式实例	改革	一	点	是	一	点
构式形式	v	一	q	是	一	q
构式成分	V	C	V	C	C	V

然而，针对同一构式的待标注语料，标注者对于句中构式实例的构式形式标注包含了大量重复的劳动，其不仅意味着类似结构的重复劳动，也意味同一构式实例在不同句子中的重复标注。因此，为了节省构建构式语料库的人工成本，需要对构式生语料进行初期标注，即构式内部成分的标注。

在构式形式表示上，构式变项代表其自身的词性，比如在构式“v＋一＋q＋是＋一＋q”中，v代表动词，q代表量词，与之对应，“一、是”等汉字字符是构式中的常项成分。换言之，现代汉语构式成分标注，可以被理解为自然语言处理中的序列标注任务。

序列标注方法的基本思路是借助人工标注的语料作为训练数据，结合相关机器学习方法，得到最优模型，或者凭借训练数据构造规则库，从而实现对目标语料的标注。现在的问题是，针对现代汉语构式成分自动标注任务，并不存在人工标注的语料，这就意味着不能够继续沿着序列标注思路来研究自动标注方法。

此外，需要注意的是，借助少量人工干预，可以从未标注语料中的构式实例抽象出构式形式。比如对于例1，其构式形式表示是“v＋一＋q＋是＋一q”。凭借构式形式表示，计算机可以对生语料进行有效的模式匹配。也就是说，现代汉语构式成分自动标注任务的本质属于字符串模式匹配任务，其形式描述为：存在一个构式形式 P，给定语料实例 S，在 S 中找到符合 P 描述的字符串。

然而，凭借基本的字符串模式匹配手段，并不能很好地完成构式自动标注任务。其原因在于，构式的形式表示对于构式成分自动标注的帮助有限，无法准确地识别构式的内部成分，而这一困难主要体现在构式及其内部成分边界的确定问题。例如，

例2 我这人不太爱享受，钱赚一点是一点。

例3 你要记住，钱多赚一点是一点。

例2、例3中的构式成分十分相似，其构式形式P均为“v＋一＋q＋是＋一＋q”。然而单纯借助构式形式表示，例3相较于例2更容易误标，前者的构式边界最好落在“多”上，而后者则落在“赚”上(表2)。

表2 例2、例3构式标注示例

构式实例1	赚	一	点	是	一	点
构式实例2	多 赚	一	点	是	一	点
构式形式	v	一	q	是	一	q
构式成分	V	C	V	C	C	V

构式边界的确定会影响构式内部成分的识别。其困难不仅仅于此，除了类如例2、例3这样构式形式相同而构式实例序列不同这样的情况所造来的构式边界确定困难。对于不同的构式，确定其构式实例边界也是不同的。对此，可依照常项与变项在构式中的位置，将构式分为两类：一类是左右边界由常项成分占据，一类是左右边界由变项成分占据(后者两端都是变项，或者一端是变项)。前者容易识别边界，后者识别有困难。前一类构式如“都＋是＋n＋惹＋的＋祸”，边界容易判断。而后一类构式占大多数，如“np＋有＋我＋呢”，实际上需要识别一个语义上恰当的“np”。以现代汉语构式知识库为例，其收录构式形式总计1050条，排除凝固型构式(总计204条)，共计846条非凝固型构式。其中，首项构式成分为构式变项的构式有309条，仅尾项构式成分为构式变项，首项为常项的构式数量为438条。也就是说，左右边界由变项成分占据的构式占非凝固型构式总量的88.3%。同时，这一类构式的识别难度也有不同，如果变项有形式约束，比如两个变项要求同形，那么，构式边界亦是容易确定的。当构式变项形式约束不是特别明确时，需要去探究如何判定构式边界。

显然，正是由于句子中的构式的边界难以确定，所以仅从字符串匹配角度出发进行自动标注并不够，还需要借助数据挖掘方法，即从大量语料数据中挖掘构式及其内部成分的相关知识。[2]

詹卫东[1]将现代汉语构式分为四种类型：①凝固型构式；②半凝固型构式；③短语型构式；④复句型构式。例如，

例4 人到中年，钱赚了不少，身体却亮起了红灯。(凝固型构式：亮＋起＋了＋红灯)

例 5　在美国大发特发之后，孙正义不忘反哺故里。（半凝固型构式：大＋v＋特＋v）

例 6　不论如何，我不忍心看到他们有一餐没一餐的。（短语型构式：有＋一＋q＋没＋一＋q）

例 7　你再看那大马路上，来来往往的人，坐洋车的也有，坐汽车的也有，坐马车的也有。（复句型构式：X＋也有，Y＋也有）

对于凝固型构式，比如例 4，其本身形式不存在变项，故而借助基本的字符串模式匹配，即可完成自动标注。理论上讲，凝固型构式也有歧义的问题，即字符串匹配成功的例句，可能是凝固型构式的真实用例，也可能是伪实例。例如，

例 8　（a）他真是笨到姥姥家了。

（b）小笨笨到姥姥家了。

例 8（a）中“笨到姥姥家了”是构式实例，而例 8（b）中“笨到姥姥家了”虽然也能跟构式字符完全匹配，但并不是构式实例。不过，类似例 8 这样的凝固型构式标注歧义的问题，在真实的语料标注中，可以被忽略，原因在于，真实的语料经过预处理，能够清除噪声。因此，本文重点探讨半凝固型构式、短语型构式、复句型构式的自动标注问题。

1　相关工作

针对文本自动标注任务，大部分研究集中在用法自动识别、词性自动标注、语义自动标注等任务。其中，从研究方法的差异来看，可以分为基于规则方法的研究和基于统计方法的研究。同时，依照是否采用完备的人工标注语料作为训练语料的标准，又可以分为基于有监督学习方法的研究和基于无监督学习方法的研究。

眷红英等[3]对副词“就”的用法分别进行了基于规则和基于统计的自动识别研究，采用人民日报虚词用法标注语料作为训练语料，前者通过构造副词“就”的用法规则库，从而进行自动标注，后者则分别采用最大熵模型与条件随机场模型，构造自动识别模型，其实现效果明显优于前者。也就是说，对于文本自动标注任务而言，采用基于统计的方法属于更优选择。

大量基于统计方法的文本自动标注研究集中在词性自动标注任务上。张艳等[4]就中文词性自动标注任务，基于词性与词相结合的三元统计模型对汉语分词及标注进行一体化处理，完成对语料库的初始标注；之后用 Brill[5]的基于转换的学习方法通过转换规则完成最终的词性标注。而后者的基本思路是利用一个带词性标注的语料库来根据事先设计好的模板，通过生成的标注规则进行标注。赵海等[6]则基于高斯先验（Gaussian Prior）平滑的最大熵模型，采用人民日报标注语料库作为训练语料设计了一个基于词和字特征的汉语词性自动标注系统，取得了较好的标注效果。除却中文词性自动标注，帕提古力·依马木等[7]使用感知器训练算法和 Viterbi 算法，借助人工标注的现代维吾尔语词性标注集，利用词的上下文信息，对维吾尔语实现词性自动标注。不难发现，上述对词性自动标注方法上，基本都是沿着前文提到的序列标注的基本思路继续深入。与此同时，国外对于文本自动标注的研究集中在语义自动标注方法研究，其思路与前文基于有监督学习方法的词性自动标注研究类似，均借助专家知识和人工干预，来实现自动标注效果[8-10]。

由此可知，基于有监督的学习方法，实质上是完成序列内部成分的分类任务，而这对于现代汉语构式成分自动标注帮助有限。因此，基于无监督学习方法的相关研究可能存在更大的帮助。

孙静[11]提出了一种基于条件随机场（CRFs）模型的无监督的中文词性标注方法，其思路为利用词典对已分好词的文本进行词性预标注；借助相关规则对未登录词进行标注，从而获得初始标注语料；结合条件随机场模型对语料进行迭代标注，逐步优化标注结果。事实上，该研究在很大程度上依赖专家知识，其方法本身是通过利用上下文环境来优化词性标注。对于现代汉语构式成分自动标注来说，专家知识是缺失的。

李娇[12]从如何发现和挖掘语言的构式这一问题出发，提出了一种结合词语移动距离（word mover's distance，WMD）模型与 R&L 密度峰算法（R&L 为算法作者姓氏首字母）来进行动词模式聚类的方法，来研究动词的典型语义组合模式。林江豪等[13]针对大规模语料手动标注困难的问题，提出利用概率潜在语义分析（PLSA）模型的新闻评论自动标注方法。利用 PLSA 计算获得语料集的“文档—主题”和“词语—主题”概率矩阵；基于情感本体库和“词语-主题”概率矩阵，认为某一类情绪词汇出现的概率最高的主题与词汇的情绪类别相同，对主题进行情绪类别标注；最后，基于“文档—主题”概率矩阵，认为出现在某一主题概率最高的文档与主题的情绪类别相同，通过“词汇—主题—文档”三者的关系，达到自动标注的效果。两者的研究均未采用

人工标注语料作为训练语料，而是侧重于数据内部相关知识的挖掘。故而，借助概率方法和聚类方法，对现代汉语构式自动标注可能存在帮助。

2 模型与框架

前文提到，现代汉语构式成分自动标注任务需要借助数据挖掘方法，即从大量语料数据中挖掘构式及其内部成分的相关知识。具体来说，就是从包含构式的句子中挖掘出构式实例序列。对此，存在一个基本的假设：在句子中，作为构式的序列与其他短语构成的序列是不同的。因此，挖掘句子中的构式实例序列的基本思路是，结合各种特征，比如句中每一个字的字形特征、位置特征及其所处位置的词的词性特征、正则表达式匹配的序列位置区间等，寻找句子中的构式实例。

在此之前，还需要讨论如何处理形式相似的复句型构式。对于形式相似的复句型构式，以“X＋也＋有，Y＋也＋有，(Z＋也＋有……)”为例，其内部以标点为界可以分为两到三个小句，两两之间的差异仅仅体现在变项的具体内容。在现实语料中，其小句出现的数量可能更多，因此，对于此类形式相似的复句型构式，构式自动标注系统倾向于将其压缩为一个通用形式，比如“X＋也＋有”，从而比较准确地识别语料中的构式，而将打通内部界限的工作交由后续的人工标注。

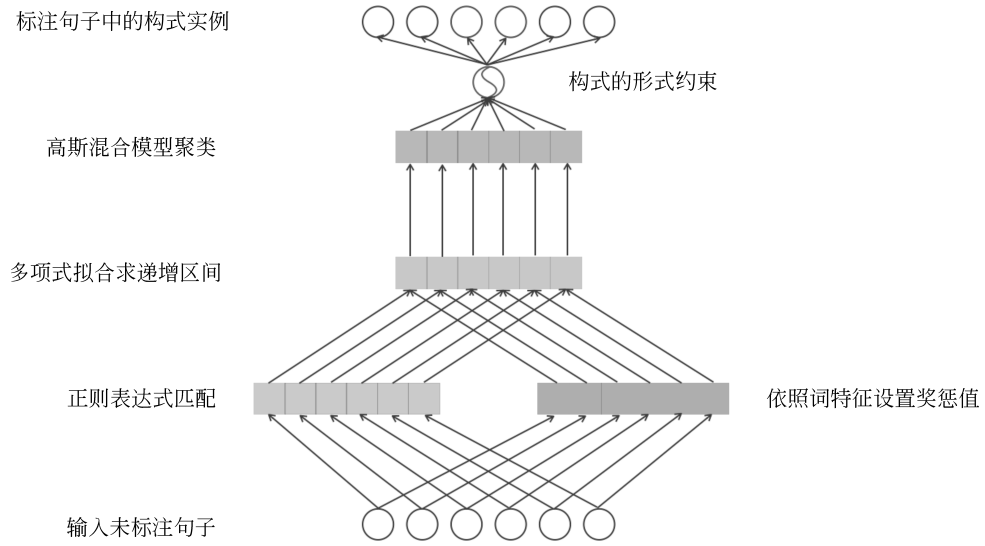


图 1 现代汉语构式成分自动标注方法流程

2.1 文本的数据表示

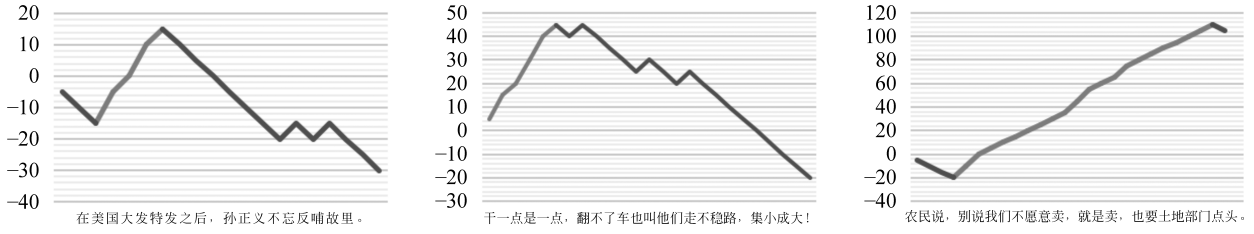
图 1 展示了现代汉语构式成分自动标注方法的基本流程。首先需要考虑的是文本的数据表示，即将一个句子转换为向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$ 和 $\boldsymbol{y}=(y_1,y_2,\cdots,y_n)$，使之能够进行聚类。其中，向量 $\boldsymbol{x}$ 表示句中每一个字的位置特征；而向量 $\boldsymbol{y}$ 则可以看成句子中每一个字除却位置特征外其他特征共同影响的结果。要知道，句子中的每一个字，其位置是独特而有序的，同时，自动标注方法需要对句中的每一个字提取各类特征，故而对于每一字，都具备一系列特征值，表示为 (x_n,y_n)，由此能够对其进行聚类。根据聚类结果，获得句子中每一个字的标签，并识别出句中的构式实例及其不同的成分。

通过观察构式的形式表示，比如例 6 中的短语型构式“有＋一＋q＋没＋一＋q”，不难发现常项和变项能够成为自动标注时的重要参考，前者直接指定了构式成分，后者则规定了构式成分的词性。构式成分自动标注系统在扫描待标注句子时，当句中的字与常项一致或其词性与变项一致，则判断其更有可能是构式成分，前者比后者可靠性更高。因此，设计一个奖惩机制，当句中的字与常项一致时，或其所处位置上的词的词性与变项的词性一致时，赋予正向奖励，且前者的奖励(＋10)高于后者(＋5)，为其他成分赋予惩罚(－5)，并进行累加，例如，对于短语型构式“n＋归＋n”，对句中的出现的常项“归”的奖惩值加 10，而对句中词性为“n”的字的奖惩值加 5，否则奖惩值减去 5。例如，“意见归意见，钱还是要筹的”，其奖惩值如表 3 所示。

表 3　短语型构式“n＋归＋n”奖惩值

Token	POS	Reward	Score
意	/n	＋5	－10
见	/n	＋5	－5
归	/v	＋10	5
意	/n	＋5	10
见	/n	＋5	15
，	/w	－5	10
钱	/n	＋5	15
还	/c	－5	10
是	/c	－5	5
要	/v	－5	0
筹	/v	－5	－5
。	/w	－5	－10

表 3 中，针对包含短语型构式“n＋归＋n”的句子，通过设置奖惩，来表示其常项文本特征和变项的词性特征。结合向量 $\boldsymbol{x}$，将句内成分的奖惩值表示为折线图（图 2），不难发现句子中的构式实例



(a) 半凝固型构式“大+v+特+v”　(b) 短语型构式“v+一+q+是+一+q”　(c) 复句型构式“别说+X+就是+Y+也+Z”

图 2　半凝固型构式、短语型构式、复句型构式奖惩值曲线

序列“大发特发”“干一点是一点”和“别说我们不愿意卖，就是卖，也要土地部门点头”处于折线的持续递增区间。但仅仅如此，并不足以准确而广泛地识别句中的构式实例序列，为此，可以考虑挖掘更多特征，从而增加相应的权重，增强其与其他短语组成的序列差异。

2.2　正则表达式匹配

前文提到，现代汉语构式成分自动标注任务并非纯粹的模式匹配任务，但这并不意味着模式匹配方法不能产生效果。相反，对于半凝固型构式和部分短语型构式而言，其识别正确率同样比较高。因此，可以考虑依照构式的形式表示，生成相应的正则表达式，并对匹配得到的序列，赋予一定的权重 w_r。其中，若奖惩值为负，则 w_r 由人为任意设置为(0，1)之间的值，实验设置为 0.5；反之，则人为任意设置为(1，2)之间的值，实验设置为 1.5，从而使匹配序列得到更高的奖惩值，其默认值为 1。同样以表 3 中的句子为例，经过正则表达式匹配，得到结果“意见归意见”，故其权重如表 4 所示。

表 4　短语型构式“n＋归＋n”正则表达式匹配权重

Token	Reward	w_r
意	－10	0.5
见	－5	0.5
归	5	1.5
意	10	1.5
见	15	1
，	10	1
钱	15	1
还	10	1
是	5	1
要	0	1
筹	－5	1
。	－10	1

然而，正则表达式的缺点同样是明显的。句中构式的边界是多样的，正则表达式很难去覆盖这一多样性。例如，对于构式“连＋X＋都＋Y，更别说＋Z”，存在构式实例：

例9 (a) 这题连数学老师都不会做,更别说一个高二学生了。

(b) 这道题连高考出题老师都不会很轻松地做出来,更别说这一群成天不上课,只知道玩的的高二学生了。

例9中,两者所表达的构式义相近,但构式边界是开放的。这就对正则表达式匹配造成了一定困难,比如匹配有误。正是因为这一困难的存在,使得汉语构式成分自动标注任务不能够仅依靠字符串模式匹配来完成。因此,仅仅借助正则表达式匹配并对句中的每一个字赋予相应权重,使之成为自动标注系统识别构式的特征之一,并输出到下一层。

2.3 文本数据的多项式拟合

由图2可知,句子中的构式实例处于折线上的持续递增区间。结合正则表达式匹配权重,这一趋势会更加凸显。因此,处于递增区间的序列,可能是构式实例序列。基于这一推论,考虑对文本数据进行多项式拟合,并求递增区间。

经过预处理,句子文本转换为向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$,其中向量 $\boldsymbol{y}$ 表示句子中每一个字的位置特征,而向量 $\boldsymbol{x}$ 此时则表示每一个字奖惩值与其正则表达式匹配权重,如式(1)所示。

$$y_i = \text{reward}_i \cdot w_{ri} \tag{1}$$

基于向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$,并对之进行多项式拟合,可以得到一条反映句子内序列类型的曲线,处于持续递增区间的序列更有可能判定为构式实例序列,反之则判定为非构式序列(图3)。

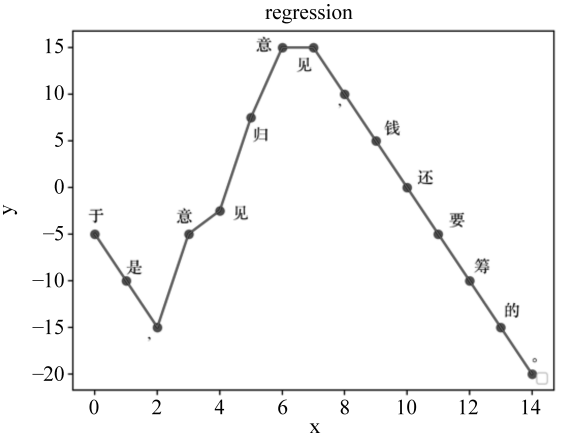


图3 短语型构式"n+归+n"多项式拟合:"于是,意见归意见,钱还是要筹的"

如图3所示,构式实例"意见归意见"整体处于递增区间。然而,不同的句子,多项式拟合效果并不一致。对于拟合偏差较大的句子,递增区间对于判断构式实例序列的帮助有限。因此,将多项式拟合的结果同样作为每一个字的权重 w_p,其赋值区间同样依照每一个字的奖惩值人为任意指定,即奖惩值为负,则设置为(0,1)之间的值,实验设置为0.2;反之,则设置为(1,2)之间的值,实验设置为1.2,并输出到下一层继续处理,由此来规避拟合所得递增区间偏差较大时带来的误差。

2.4 高斯混合模型聚类

从经由多项式拟合的文本数据中判断构式实例序列的位置区间,其根本依据在于构式形式的语言学特征,即构式常项的文本特征与构式变项的词性特征。然而,经过实际测试,发现所判断的构式实例序列与正确的构式实例序列所在的位置仍然有可能产生了较大的差异,故而考虑利用聚类方法对其进行聚类,并最终获取构式实例序列。结合文本数据多项式拟合新增的权重 w_p,向量 $\boldsymbol{y}$ 中的每一个值可以表示为式(2)。

$$y_i = \text{reward}_i \cdot w_{ri} \cdot w_{pi} \tag{2}$$

上式表示句中第 i 个字被表示为该字的奖惩值、正则表达式匹配权重与多项式拟合所得权重的积。接下来考虑的是聚类方法的选择。构式自动识别要求将构式实例序列与其他序列区别开来。对此,经典聚类算法包括K-Means算法、层次聚类算法和具有噪声的基于密度的聚类方法(density-based spatial clustering of applications with noise,DBSCAN)均存在各自的缺陷。K-Means算法对于非球形数据聚类存在局限性,而由句子产生的文本数据并非球形数据;层次聚类的局限性则体现在不能有效抗噪声,事实上,选择聚类算法的初衷就是确定正确的构式边界;DBSCAN算法对于噪声点的处理是删除或忽略,不符合构式自动识别的要求。故而,选择理论上能够拟合任意分布的高斯混合模型。

高斯混合模型指的是多个高斯分布函数的线性组合,其模型的数学表示如式(3)所示。

$$p(x) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k) \tag{3}$$

其中,K 表示分量个数(类簇数量),$N(x|\mu_k,\Sigma_k)$ 表示混合模型中的第 k 个分量,π_k 表示各个分量的权重。各项参数由EM算法计算获得。

就现代汉语构式成分自动标注而言,对文本数据进行聚类前,需要指定分量个数 K。理论上,句子中存在构式实例序列与非构式序列两种类型,在

实际测试中，将分量个数指定为 3，能够取得更佳的效果，即句子序列被分为了构式实例前序列、构式实例序列和构式实例后序列。借助高斯混合模型聚类方法，自动标注系统最终得到了句中每一个字的标签，并对其进行最后的标注。

2.5　构式的形式约束

在进行最后一步标注之前，还需要考虑构式本身具备的形式约束，比如构式变项的形式约束、构式形式的字符串长度约束。

通过观察部分构式形式表示，不难发现，这一类构式存在一个共同点，即构式中存在部分常项和变项同形且音节数相同，比如短语型构式“有＋一＋q＋是＋一＋q”，半凝固型构式“大＋v＋特＋v”，暂且称之为构式成分一致性。同时，构式变项本身的音节数存在一定约束，比如构式“a＋牌＋n”（耐心牌老公，放心牌干部），a 一般为 2 音节，n 一般为 2～3 音节。

此外，构式形式的字符串长度本身规定了构式实例的最小长度，比如短语型构式“v＋一＋q＋是＋一＋q”，其最小长度为 6，但凡识别的构式实例长度小于 6，即可标记为误判，并进行改正。

因此，借助构式的形式约束，构式成分自动标注系统能够有效确定相关构式实例的边界。其不仅可以去除被误判为构式成分的字符，同样可以将未被识别的构式成分正确地识别出来。

3　实验

现代汉语构式成分自动标注系统主要针对半凝固型构式、短语型构式和复句型构式的成分标注。因此，分别选择三种类型的语料进行实验，总计 2 524条包含构式实例的完整句子，其中，半凝固型构式 874 条，短语型构式 850 条，复句型构式 800 条。而后将实验数据输入到自动标注系统，得到标注数据，并通过人工校对检查其正确率与召回率。同时，基于构式形式中的常项对待标注句子进行简单的字符串匹配，匹配得到的结果作为基线实验，与现代汉语构式成分自动标注系统产生的结果进行比较，实验结果见 3.1 节。

3.1　实验结果

经过测试，实验结果如表 5 所示，由上至下分别是半凝固型构式、短语型构式、复句型构式语料的测试结果。

表 5　半凝固型构式、短语型构式、复句型构式语料测试结果

方法	类型	P	R	F_1
基于 GMM＋正则表达式＋词性及字形匹配方法	半凝固型	0.870	0.814	0.841
	短语型	0.855	0.775	0.813
	复句型	0.773	0.444	0.564
基于简单字符串匹配的方法	半凝固型	0.660	0.664	0.662
	短语型	0.671	0.577	0.620
	复句型	0.531	0.340	0.415

由表 5 可知，基于高斯混合模型，融合了正则表达式匹配所得信息及构式形式的语言学特征匹配所得信息的现代汉语构式标注方法的效果，比之简单字符串匹配的自动标注结果更优。此外，就不同类型的构式之间进行比较，不难发现半凝固型构式标注难度要远远低于复句型构式，原因在于复句型构式本身更加复杂，使得其构式实例的边界更加难以确定。

3.2　讨论

基于高斯混合模型的汉语构式成分自动标注方法在一定程度上解决了构式生语料的自动标注问题，但其仍然存在不少有待改进之处。首先，句子中每一个字的奖惩值初始是依照常项的文本特征与变项的词性特征给定的。这就使得词性标注的准确性会影响奖惩值，从而影响最终的标注效果。故而，数据预处理阶段提高词性标注的准确性能够改善最终的标注效果。

其次，在同一个语料文件中，同一个构式实例在不同的上下文中会出现标注不一致的情况，这同样会降低自动标注的准确率。一种可能的处理办法是，记录标注过的构式实例序列，当再次遇到同样的序列时，直接标注其为构式序列。然而，这一思路存在一个问题，即如何在标注过程中判断已标注序列的准确性。如果能够解决这一问题，最终的自动标注效果会得到更大的改善。

最后，当音节数目不确定的构式变项处在构式形式的收尾末端时，由于边界难以确定，故而会极大地影响到自动标注效果。例如，复句型构式“别说＋X，就是＋Y＋也＋Z”，构式变项 Z 的边界，往往决定了句中构式实例标注正确与否。存在一种处理策略，即将 Z 的边界定到所在小句的标点之前。然

而，对于短语型构式“一＋q＋都＋没有＋X”，例如，

例 10 瞧瞧你洗碗洗了多久？整整一天都没有好。

例 11 整整一夜都没有睡着觉的他心情很烦躁。

应用上文提到的策略，自动标注系统能够正确地标注例 10 中的构式实例，对例 11 则会误标注。显然，这一策略存在很大的优化空间，优化的同时亦会改善自动标注效果。

4 结论

现代汉语构式成分自动标注作为文本自动标注任务之一，其最大的困难在于，当不存在标注语料作为训练数据时，如何从句子中确定构式的边界信息。本文通过设置奖惩值，结合正则表达式匹配与多项式拟合所得的权重，借助高斯混合模型聚类方法对语料进行自动标注，发现相较于单纯基于正则表达式匹配和基于词性匹配的自动标注结果，其效果更优。

然而，需要注意的是，当前对构式生语料的标注，构式形式表示是已知的。如果能够直接从语料中挖掘得到构式形式表示，构式成分自动标注的效率会得到更大的提高，而这有待进一步研究。

参考文献

[1] 詹卫东. 从短语到构式：构式知识库建设的若干理论问题探析[J]. 中文信息学报，2017，31(1)：230-238.

[2] Jiawei Han，Micheline Kamber，Jian Pei. Data mining：Concepts and techniques[M]. SF：Margan Karfmann，2012.

[3] 昝红英，张军珲，朱学锋，等. 副词“就”的用法及其自动识别研究[J]. 中文信息学报，2010，24(5)：10-16.

[4] 张艳，徐波. 基于转换的错误学习方法的汉语词性自动标注研究[C]. 中国中文信息学会二十周年学术会议，2011：147-154.

[5] Brill E. Transformation-based error-driven learning and natural language processing：A case study in part-of-speech tagging [J]. Computational Linguistics，1995，21(4)：543-565.

[6] 赵伟，赵法兴，王东海，等. 一种基于改进的最大熵模型的汉语词性自动标注的新方法[C]. 第二十三届中国数据库学术会议，2006：185-189.

[7] 帕提古力·依马木，买合木提·买买提，吐尔根·依布拉音，等. 基于感知器算法的维吾尔语词性标注研究[J]. 中文信息学报，2014，28(5)：187-191.

[8] Djioua B，et al. EXCOM：An automatic annotation engine for semantic information[C]//Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference，2006：285-290.

[9] Alexiei Dingli，Fabio Ciravegna，Yorick Wilks. Automatic semantic annotation using unsupervised information extraction and integration[C]//Proceedings of the KCAP 2003 Workshop on Knowledge Markup and Semantic Annotation，2003.

[10] Kiryakov A.et al. Semantic annotation，indexing，and retrieval[J]. Journal of Web Semantics，2004，2(1)：49-79.

[11] 孙静. 基于平行语料库的无监督中文词性标注研究[D]. 苏州：苏州大学硕士学位论文，2010.

[12] 李娇. 面向认知构式语法的英语动词模式的识别[D]. 哈尔滨：哈尔滨工业大学硕士学位论文，2016.

[13] 林江豪，顾也力，周咏梅等. 基于 PLSA 的新闻评论情绪类别自动标注方法[J]. 计算机系统应用，2019，28(1)：207-211.

黄海斌(1997—)，硕士研究生，主要研究领域为汉语语言知识工程、中文信息处理。
E-mail：huanghaibin@pku.edu.cn

詹卫东(1972—)，通信作者，博士，教授，主要研究领域为现代汉语形式语法、中文信息处理、汉语语言知识工程。
E-mail：zwd@pku.edu.cn

常宝宝(1971—)，博士，副教授，主要研究领域为自然语言处理、计算语言学。
E-mail：chbb@pku.edu.cn