基于字序列标注的中文关键词抽取研究

南京大学信息管理系 王昊 邓三鸿 苏新宁

（南京大学信息管理系，江苏南京 210093）

ywhaowang@nju.edu.cn

摘 要：由于汉语组词的多变性和中文分词的不成熟，使得分词后统计处理与词序列标注等两种常用的关键词抽取方法均存在较大问题。为此，本文以某大学图书馆的所有馆藏书目为研究对象，在对图书关键词标引信息进行分析的基础上，总结了中文关键词的基本特点及其抽取规律，构建了一个基于字序列标注的中文关键词抽取模型，提出了中文关键词抽取的基础思路和实现方案，并通过实验论证了模型了合理性、正确性和实用性，认为字序列标注方法优于词序列标注，基本上解决了不分词情况下的中文关键词抽取问题。

关键字：序列标注；条件随机场；关键词抽取；机器学习；字序列；词序列

Research on Chinese Keywords Extraction based on Characters Sequence Annotation

Wang Hao, Deng Sanhong, Su Xinning

(Information Management Department of Nanjing University, Nanjing 210093, China)

[Abstract] Due to the variability of Chinese phrases and the immature of Chinese word segmentation, statistical treatment after word segmentation and words sequence annotation, the two commonly used methods are not effective enough to solve the problems about keyword extraction. Thus, based on the whole Chinese booklist of a certain university library as well as the analysis of its books indexing information, .the paper summarizes the features and extracting laws of Chinese keywords, and establishes a Chinese keywords extraction model based on characters sequence annotation, which proposes the basic idea and implementation scheme for extracting keywords. It verifies the feasibility, rationality and practicality of the model by large-scale experiments, and basically solves the case of Chinese keywords extraction problems without executing word segmentations, which shows characters sequence annotation is better than words sequence annotation.

[Keyword] sequence annotation; Conditional Random Fields; keywords extraction; machine learning; characters sequence; words sequence

1 引言

关键词抽取是指利用计算机从文本内容中自动提取出能够代表该文本主题的词汇或短语集合[1]，实现文本表示的过程，抽取出来的词汇或短语即为关键词。关键词抽取是一项基础性工作，是实现海量文本内容检索的前提，也是实现如文本标引、自动分类和聚类、自动摘要、个性化推荐等工作的核心技术[2]~[4]。然而，中文语法的特殊性以及组词的复杂性加剧了中文关键词抽取的难度。

总结前人的研究成果，目前中文关键词抽取主要有两种方法：一是传统的基于分词的统计方法，即首先对文本内容进行精确分词，在排除非用词后建立候选关键词集合，再采用各种统计算法如词频、TF-IDF 值、ATF×PDF 值等，并结合词语的词性、位置、形态等特征对候选关键词进行权重计算，进而根据权重筛选关键词[5]~[9]；另一种则是基于词序列的语言学方法，即首先对图书文本进行粗分词，使得文本转化为词汇序列，然后根据词汇的上下文语言学特征确定相邻词汇之间的语义关系，判断是否可将相邻词汇合并作为关键词[10]，或对词汇的上下文语言学特征进行机器学习，再用训练后形成的学习模型判断词汇的角色，进而根据关键词角色模板将相关词序列合并为关键词[11]~[14]。上述方法均存在明显缺陷：①目前中文分词技术还没有达到非常精确，现有的较好的中文分词系统均倾向于将文本切分为长度较小的词汇，与关键词一般为长度较大的名词性领域术语相冲突，

因此需要对分词后获得的关键词进行合并或再处理；②机器学习需要大量带有关键词的标注数据作为训练样本建立标注模型，而这些数据的建立需要大量的人力资源及领域知识[5]；③中文组词具有很强的灵活性，使得词汇数量非常庞大，特征丰富而不易学习，而且将关键词看作是词汇组合使得词汇角色非常复杂，例如关键词的组成部分可能被切分到了其他非关键词中等。

基于机器学习的方法在训练样本足够大、覆盖范围足够广的情况下，针对同领域文本的关键词抽取具有很好的准确性和召回率。因此，本文试图利用现有的图书标引数据，对词序列标注的方法进行改进，将关键词看作是汉字的组合，设计关键词的字角色空间，并在此基础上构建基于字角色标注的书目关键词抽取模型，对人工关键词标引的特征和规律进行机器学习，进而利用生成的标注模型对书目实现自动关键词抽取，以解决词序列标注稳定性差、词语转变为关键词困难等问题。笔者通过实验对比论证了字序列标注模型的正确性和合理性，为模型的进一步优化、实用提供了事实依据。

2 标引数据的统计和分析

笔者采用了某大学图书馆的馆藏书目数据作为研究对象，既解决了标引数据难获取的问题，而且由此获得的标注模型具有很强的实用价值；此外，关键词词组合的思想具有一定应用价值，但是存在切分不准确、角色多样性以及成分不稳定等缺点，需要进一步完善。因此，为了探索关键词抽取的思路和方法，有必要从总体上了解图书标引数据的分布及其特点。

至 2010 年 6 月底，该馆共有馆藏书目 187,213 种，其中给出了标引词的有 150,850 种。属于关键词标引的，即标引词能够在书目题目和内容中抽取的则有 65,482 种，涉及关键词 6,511 个；题名标引中的有 52,279 种，涉及关键词 5,746 个；提要标引的有 42,754 本，涉及关键词 5,167 个；标引词同时出现在题名和摘要中的则有 29,530 种，涉及关键词 4,159 个。本节通过对标引数据的分析，寻找书目关键词抽取的一般思路。

（1）书目关键词个数的统计和分析

在书目的关键词标引数据中，单种图书关键词个数的分布情况如表 1 所示。不难发现，单种书目关键词的个数被限制在了 1~5 个词语之间，其中绝大部分书目均只有一个关键词，部分书目有两个，两者合计占总书目的 99.8%以上。从这里笔者总结了书目关键词标引的基本特点：书目标引一般情况下只有 1~2 个关键词，较少情况下会出现 3~4 个关键词。书目关键词个数相对较少的特点决定了机器自动抽取书目关键词的方法是合适的，而且相对来说将具有较高的准确率和较大的实用性。

表 1 书目关键词个数的分布情况

关键词位置 / 关键词个数	题名或摘要	题名关键词	摘要关键词	题目摘要关键词
5	1	1	0	0
4	10	6	4	1
3	84	58	25	6
2	881	800	151	75
1	64506	51414	42575	29448
书目合计	65482	52279	42755	29530

（2）图书类目的分布情况分析

在书目的关键词标引数据中，排除错误分类现象（主要包括没有分类和大类目丢失），图书类目的分布情况如图 1 所示。从图中不难发现，该大学图书馆的馆藏书目中，F 类（经济学）图书占据了绝对领先地位，其次是 D 类（政治理论）、H 类（语言学）和 T 类（工业技术理论等）等，U 类（运输）和 V 类（航空航天）则最少；从总体上来看，图书的类目分布极不均衡，其中 B、C、D、

F、G、H、I、J、K、O、T 等 11 类的书目数量远远超过了其他 11 类目，占总数的 94%以上。关键词是图书内容的浓缩，一般具有很强的类目特征，因此在书目分类研究中，关键词在特征权重中通常占据了最大的比重。而图书分类的不均衡现象为书目特征的抽取（即关键词的抽取）增加了难度，特别是在采用机器学习方法抽取书目特征时，会导致学习不充分、标注偏差等问题的出现，最终致使模型的标引准确率下降，可以考虑将部分稀缺类目进行合并，以平衡类目数据间的差异。

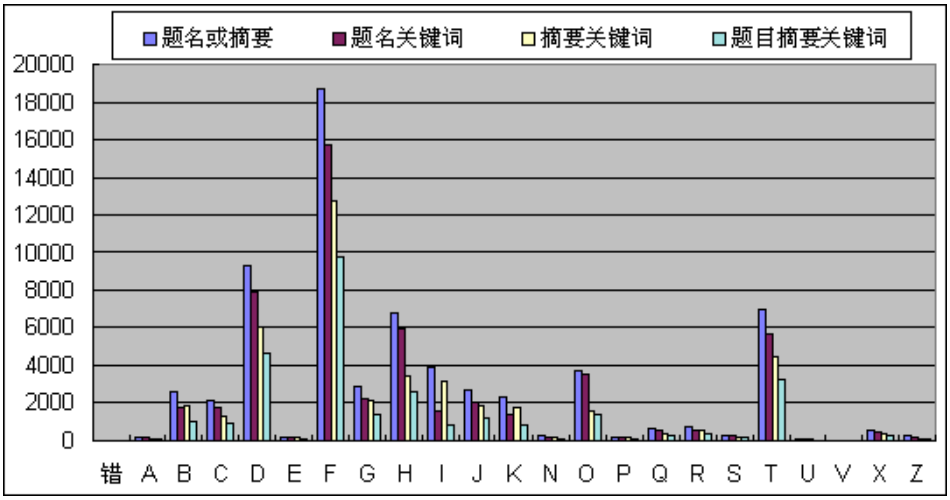


图 1 图书关键词的类目分布情况

（3）关键词组成成分分析

传统的图书自动标引方法是先对书目文本进行分词，进而统计词频，选择有意义的高频词作为书目关键词。这一方法使研究人员意识到了词汇是组成文本的基本单位，于是直接导致了中文分词技术的发展和成熟。然而，随着标引工作的逐渐深入，研究人员开始认识到：为了充分描述文本内涵，关键词通常是一个短语或词组，通过分词直接获得关键词存在较大缺陷。笔者借助中科院计算所研制的 ICTCLAS 分词系统[15]，对出现在题名和摘要等著录项中的 6,511 个关键词进行了分词并统计，结果如图 2 所示。

图 2 中横坐标表示关键词被切分后形成的词汇个数，纵坐标则表示被切分成相应个数词汇的关键词数。在 6,511 个关键词中，没有被切分的有 2,760 个，约占总数的 42.39%；大部分关键词可以被切分成 2~6 个词汇，其中超过半数被切分成了 2 个词汇；被切分成词汇的个数越多，此类关键词数量也越少。为此，笔者可以得出结论：超过半数以上的书目关键词都是有若干个词汇所构成的，对书目内容分词后直接统计词频确定关键词的方法并不合适，分词后对词汇进行角色标注进而组成关键词的抽取方法是值得探讨的。

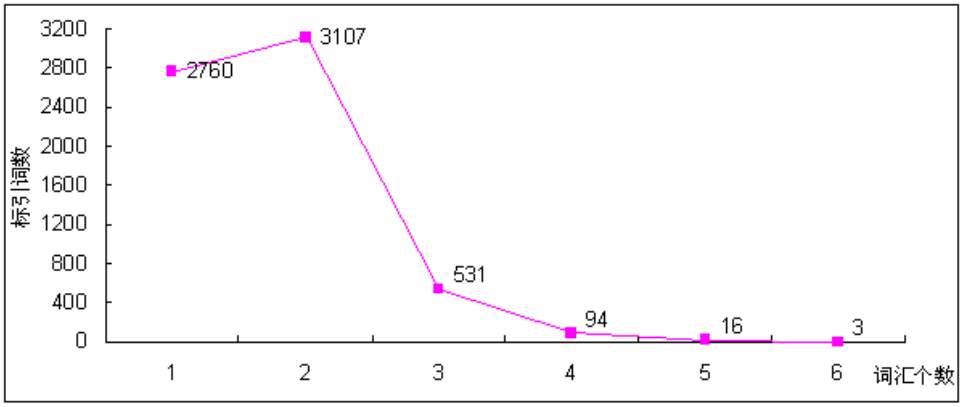


图 2 由相应词汇个数组成的关键词数量统计

（4）词组合和字组合

任一关键词均可以看成是词或字等片段的组合。为此，若能够在语言片段中识别出连续的关键词标记符号，那么这连续符号所对应的语言片段组合即为关键词。问题是：将关键词看做是词组合呢，还是字组合？关键在于何种组合的关键词抽取准确率更高。这里先从总体上对两种模型进行理论探讨。笔者对全部实验数据共 65,482 本图书的题名分别进行了词切分和字切分，前者采用 ICTCLAS 系统，对分词结果不做任何人工修正；后者则将文本切分为单个汉字，并将题名中出现的连续符号组合作为单汉字。切分结果如图 3 所示。

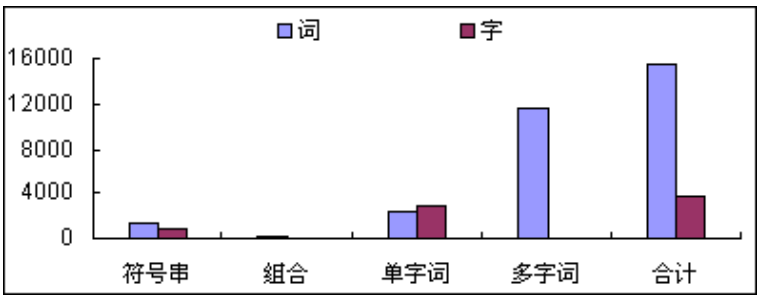


图 3 书目题名的词切分和字切分对比情况

经过词切分后，共获得了 15,442 个不同词语，包括连续符号串、符号汉字组合、单字词和多字词，其中多字词最多，占了 75.08%；而经过字切分后，获得了不同的连续符号串和单字共 3,725 个，远少于切分后的词语种类。笔者认为：①若以汉语片段作为语言学特征，那么字切分后获得的特征将比词切分少，就机器学习而言，复杂性会大大降低；②汉语的一大重要特征就是字与字之间具有灵活的组合性，于是在不同的上下文语境下，同一个字可能会出现不同的前后组合（即歧义现象），这种现象将极大增加组成关键词成分的角色数量，即关键词的组成成分很可能成为另一个词的一部分，也可能包含不是关键词的其他词语，这将给词语的角色标注带来极大不确定性和不稳定性；而字是汉语片段的最小单位，不会出现组合的多样性和分词的错误性现象，也将极大的降低角色模型的复杂性，相对的增加了字角色标注的准确性和稳定性。

3 中文关键词抽取模型的分析和设计

根据中文书目关键词标引的特点和规律，笔者构建了一个基于字角色序列标注的中文书目关键词抽取模型，其基本思路是：首先建立一个角色空间模型，用于标识在图书内容中出现的所有汉字（包括符号串），使得每个汉字都对应一个符号角色；将图书内容转化为字观察序列作为第一种语言特征，并对特征进行衍生，扩展观察序列以强化汉字的上下文语境规律；然后根据角色空间模型，将汉字映射成角色符号作为标注序列，观察序列和标注序列一起构成了学习样本（或称训练语料）；采用机器学习算法（如最大熵模型、隐马尔科夫模型、条件随机场等）对学习样本的序列规律（纵向，或称上下文特征）和标注规律（横向，或称标注特征）进行学习，形成学习模型；最后将学习模型作用于仅由观察序列构成的测试样本（或称测试语料）进行序列标注，自动获得字序列所对应的角色序列，则符合关键词角色模板的相应的字序列组合即为关键词。整个过程如图 4 所示。

定义 1：集合 Ω={R, P}被定义为字角色空间，其中 R 为字角色集合，用于机器学习阶段完成对汉字的标注；P 为关键词角色组合模板集合，用于书目标引阶段完成对应关键词的抽取。

定义 2：集合 R={B, M, E, S, X, F}被定义为字角色集合。其中，单字关键词中汉字标注为 S，多字关键词则根据其组成汉字在词中位置分别标注为 B、M、E；非关键词中的汉字则标注为 X；在中文书目中出现的符号串则标注为 F。

根据 R，可以将任意汉字转化为角色，完成字空间（数千个汉字）到角色空间（6 种角色）的映

射，从而极大地降低了序列的复杂度。需要说明的是，在书目内容中可能出现多个使用了相同语言片段的关键词，那么这些共享的语言片段就可能出现多种角色。例如在书目“可持续发展战略”中，存在两个关键词“可持续发展”和“发展战略”，那么在题名中“发”和“展”就均存在两种角色，分别为“M”、“B”和“E”、“M”。本文把这种可能有多种角色的汉字用符号组合进行标记，例如把上例中“发”和“展”分别标记为“MB”和“EM”。

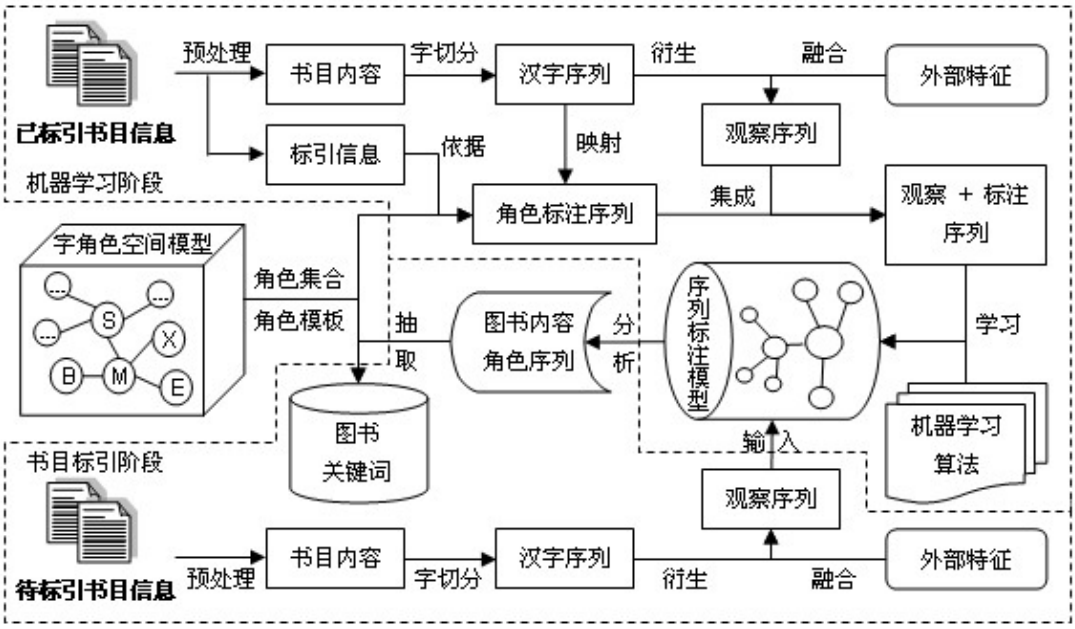


图 4 基于字角色标注的中文书目关键词抽取模型

定义 3：集合 $P=\{S, BE, BM_nE \mid n=1, 2, \ldots, N\}$被定义为关键词角色组合模板集合。其中，S 表示单汉字关键词模板，BE 为两个汉字关键词模板，BM_nE 为多汉字关键词模板（n 为大于 1 的整数）。P 中模板所对应的汉字序列组合即为关键词。

定义 4：字序列[16]$B=\{B_1, B_2, B_3, \ldots, B_n \mid n>0\}$，其中 B_n 为汉字或符号串。字序列用于描述语言片段的上下文语境特征。

例如题名“中国增值税转型可行性实证研究”对应的字序列即为{中, 国, 增, 值, 税, 转, 型, 可, 行, 性, 实, 证, 研, 究}。字序列是最基本的观察序列，正是由于汉字的自身特点及其固定排列，使得序列中的每个汉字都表现出一定的角色特征。本文仅以字序列作为机器学习的观察序列。

定义 5：标注序列 $L=\{L_1, L_2, L_3, \ldots, L_n \mid n>0 \text{ and } L_n \in R\}$，其中 L_n 为汉字所对应的角色。

例如题名“中国增值税转型可行性实证研究”对应的标注序列即为{X, X, B, M, E, X, X, X, X, X, X, X, X}。标注序列仅出现在训练语料中，作为学习样本，即让机器学习某汉字之所以被标注为某角色的上下文特征，并将其作为一种知识（模型）存储。

定义 6：特征模板 $TMPT=\{B_n(n=-2,-1,0,1,2), B_{n-1}B_n(n=-1,0,1,2), B_{n-2}B_n(n=0,1,2), B_{n-2}B_{n-1}B_n(n=0,1,2), L_{-1}L_0 \mid B_n \in B \text{ and } L_n \in R\}$，其中 B_n 为一元模板，描述的汉字本身的特征；$B_{n-1}B_n$ 和 $B_{n-2}B_n$ 为二元模板，描述的是前后汉字之间以及字与前前字之间的关系特征；$B_{n-2}B_{n-1}B_n$ 为三元模板，描述的是字与前后汉字之间的 3 元关系特征；$L_{-1}L_0$ 描述的则是前字的标注角色对后字标注结果的影响；n 极大与极小值之间的间隔被称为字长窗口[16]，用于描述上下文的约束距离，这里采用 5 字长窗口。

在图 4 中，笔者将整个图书关键词抽取过程分为了两个阶段：先学习，后分析。即先学习已人工标引的书目字序列上下文特征，再分析出未标引书目字序列对应的角色序列，进而抽取关键词。

所谓机器学习，就是根据对已知情况状态及其可能原因的分析和学习，来判断未知情况的可能状态的方法和过程。常用的机器学习方法大致可以分为两类，①根据对象自身特征判断对象状态，也称为分类，常用的算法包括决策树、人工神经网络、支持向量机、朴素贝叶斯等；②根据对象自身及其上下文环境判断对象状态，不仅需要将对象用特征向量来描述，而且还需要设定当前对象和前后对象之间存在的语义关系类型，因此也称为序列标注，常用的算法包括隐马尔科夫模型（Hidden Markov Model，HMM）[17]、最大熵模型（Max Entropy Model，MEM）[18]、条件随机场（Conditional Random Fields，CRFs）[19]等等。在这两类机器学习方法中，前者适用于对象的特征较多而且特征值具有较强的区分度的情况下，而后者则适用于上下文语境对当前对象具有较强影响的环境下，一个汉字被标注为什么角色不仅是由其自身特征所决定的，而且与其所在的上下文环境，例如前后字特征、前字的标注角色等都有关系。因此在本文构建的关键词抽取模型中采用 CRFs 序列标注算法，并以开源软件 CRF++0.51[20]作为运行平台。

4 面向题名的书目关键词抽取实验分析

题名作为对图书内容的高度浓缩，能够在很大程度上反映图书主要内容。在本文使用的所有图书标引数据中，共有关键词 156,772 个，其中 53,217 个来自题名，约占总数 1/3 强。可见从题名中抽取关键词是合理的，也是目前常见的人工标引方式。本节以题名关键词标引数据（53,217 个关键词共来自 52,279 本图书）作为实验对象进行分析，从训练样本规模以及标注方法等方面来论证基于字序列标注抽取关键词模型的正确性和合理性。

在实验分析中，笔者采用正确率（P）、召回率（R）和 F1 值对关键词抽取模型进行性能评价。上述测评指标的计算公式如下所示：

$$P=\frac{RC}{C}$$ （其中：C为关键词识别数，RC为正确识别数） （1）

$$R=\frac{RC}{N}$$ （其中：RC为正确识别数；N表示人工标引关键词的个数） （2）

$$F1=\frac{2PR}{P+R}$$ （3）

4.1 基于不同训练样本规模的关键词抽取结果比较

一般认为，基于机器学习的方法在训练样本越充分的情况下，获得学习模型的识别效果越佳。为此，笔者分别选择 27,000，32,000，37,000，42,000，47,000 本图书的标引信息作为训练数据，并以其他 5,279 本图书（含有关键词 5,279 个）作为测试数据，来探讨随着训练样本规模的逐渐扩大，关键词抽取效果的变化情况。结果如图 5 所示。

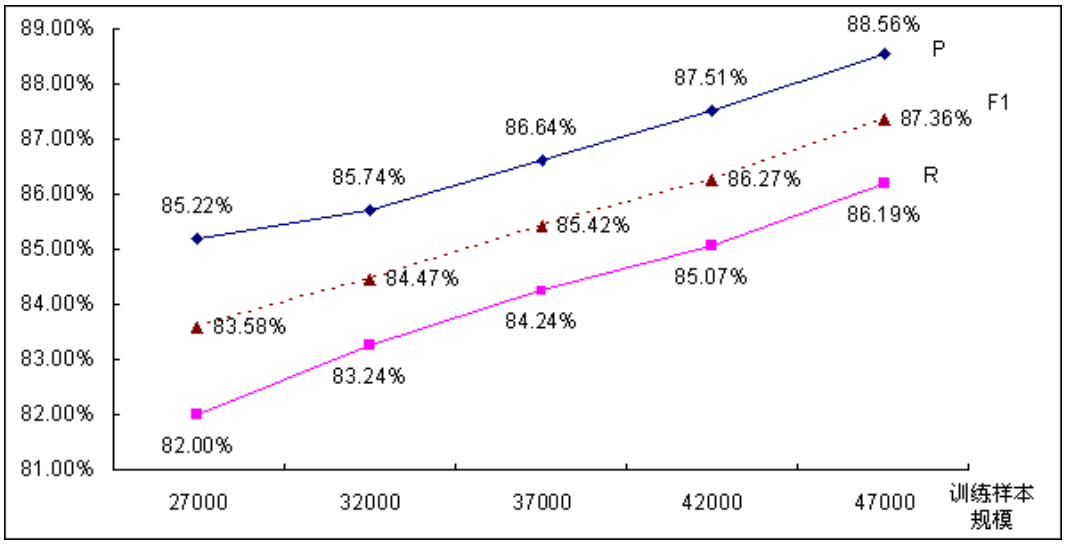


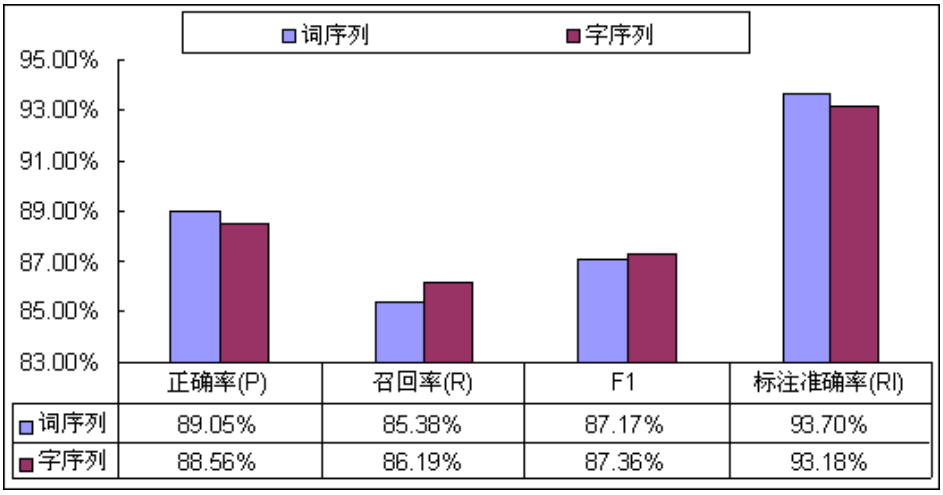
图 5 基于不同训练样本规模的关键词抽取结果比较

从图中笔者发现，①在以单字序列作为观察对象，即仅考虑汉字自身以及上下文语境特征（字长为 5），并在包含 6 种单字角色{B, M, E, S, X, F}和 4 种组合角色{EM, EB, MM, MB}的角色空间模型的约束下，书目题名关键词抽取的 P、R 和 F1 值最高分别达到了 88.56%、86.19%和 87.36%，可见基于字序列标注的书目关键词抽取模型具有一定的实用价值；②随着训练样本规模的逐渐扩大，关键词抽取的效果增长明显，在书目样本量为 27,000 时，关键词抽取的 F1 值仅为 83.58%，当将样本量扩大到 47,000 后，F1 值增加了近 4 个百分点，进一步验证了对于基于机器学习的方法，在学习足够充分的情况下，能够发挥其最大的优势。可以推测：在本文构建的实验环境中，如果进一步加大训练样本，可以使关键词抽取的效果达到最佳；③当训练样本量达到 47,000 时，从测试数据（5,279 个关键词）中可以抽取 5,138 个关键词，其中正确的为 4,550 个，与完全抽取还有相当一段差距，除了可以通过扩充训练样本规模的方式加强机器学习之外，还可以增加机器学习样本的特征来提高样本区分度，包括扩展观察序列和增加上下文字长窗口等。可以想象，对象的特征越明显，越容易被学习和识别；④本文将关键词抽取的应用限定在了书目标引中，而目前在各级图书馆中存在大量可利用的关键词人工标引数据，可以保证充分的机器学习，因此在书目关键词抽取中引入序列标注机器学习的方法不仅是可行的，而且也是合理的。

4.2 基于词序列和字序列标注的关键词抽取结果比较

由于中文分词的不准确以及汉语组词的多变性，使得将文本作为词序列组合可能会来带词角色的复杂性和特征学习的不稳定性，为此本文提出了修正思路，即使用字序列来代替词序列，以提高关键词识别的综合效果。本节对采用词序列和字序列标注的关键词抽取进行了全方位的比较。

为了具有可比性，笔者在两次实验中采用了相同的训练和测试样本，其中训练样本为 47,000 种书目中包含 47,938 个关键词，测试样本为 5,279 种图书含有人工标引关键词 5,279 个；均以字或词本身作为观察序列；字序列标注中采用了 10 角色（包括 6 种单角色和 4 种组合角色）空间，而在词序列标注中，为了能够从粗分词中进一步分离出关键词，笔者采用了 35 种角色（包括 5 种单角色和 30 种组合角色）对词汇进行区分，汉语组词的多变性可见一斑，比字序列标注的角色空间要复杂的多。两次实验的结果对比如图 6 所示。



	正确率(P)	召回率(R)	F1	标注准确率(RI)
词序列	89.05%	85.38%	87.17%	93.70%
字序列	88.56%	86.19%	87.36%	93.18%

图 6 基于词序列和字序列标注的关键词抽取结果比较

从图中可以发现，①两种方法的关键词抽取效果均较佳，F1 值均超过了 85%，可见基于序列标注机器学习的方法来实现关键词的抽取具有一定的合理性和实用性；②基于字序列的关键词抽取方法在 F1 值上比词序列高出了 0.81 个百分点，可见前者的抽取效果比后者更佳；③字序列方法优于词序列的主要原因是高召回，即前者比后者能够识别更多的正确的关键词；④词序列在识别正确率(P)和标注正确率(RI)两项上均优于字序列，一方面是由于词序列方法总体上识别的关键词数较少，使得其 P 值较高，这可能是由于其具有较多特征所造成的；另一方面由于相同文本的词总数远远小于字总数，因此在 RI 一项上因为基数较小而使得 PI 值偏高。

为了进一步对比两种方法，图 7 列出了相关的其他实验参数，包括关键词识别数、正确识别数、CRFs 算法特征数以及训练时间等。①在同等实验环境下，字序列标注方法在关键词识别数以及正确识别数两项指标上均列前茅，这也是其高召回率的直接原因，因此如果为了尽可能多的识别出文本关键词，字序列标注方法更为可行；②从特征选择上来看，词序列标注能够采用的特征明显高于字序列标注，一方面在汉语中，词语数量远远高于字数量，参见图 3，使得词语本身表现出来的特征明显多于汉字；另一方面，在词序列标注中使用了 35 种标注角色，远多于字序列，这也使得词序列表现出了更强烈的特征复杂性。特征数的增多虽然加大了学习的复杂性，同时也使得学习更有效果，加强了标注模型的区分度，P 值升高；③训练样本特征多，同时也使得训练的时间变长，图中清晰显示词序列标注学习所需时间远远高于字序列标注，当然角色空间的复杂化也是训练时间变长的一个重要因素。

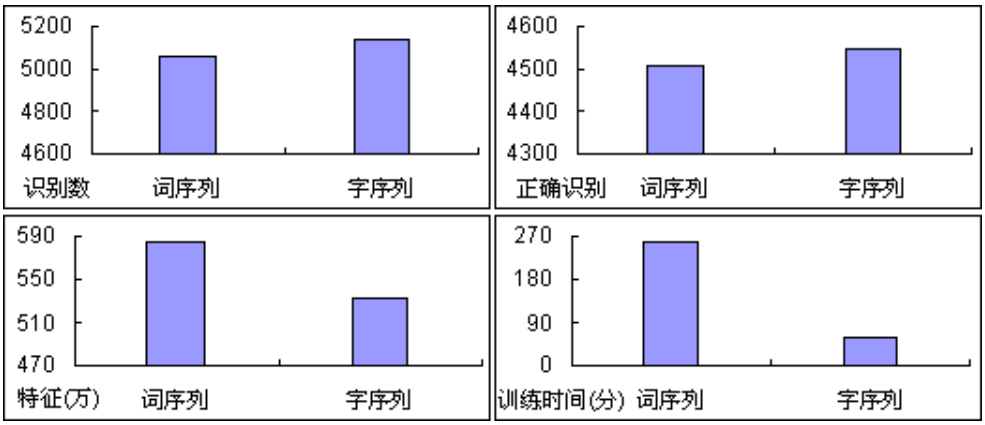


图 7 基于词序列和字序列标注的相关实验参数比较

两种关键词抽取方法均具有较强的实用性。笔者认为，字序列标注方法的角色空间较为简单，训练速度较快，可利用的语言特征变少，能够召回更多的关键词，具有较高的综合识别效果；而词

序列标注方法能够利用更多的语言特征约束，需要更长的训练时间，关键词识别准确率较高，但是召回的关键词变少，综合识别效果不如字序列标注，同时由于汉语组词的复杂性，关键词的一部分或全部有可能被歧义的成为另一个非关键词汇的一部分，这也是导致其角色空间相当复杂的主要原因，相应的从词语中分离出关键词组成成分的算法也变得非常复杂。

5 结语

本文对前人研究成果进行了总结，认为采用序列标注机器学习方法实现关键词抽取是可行而且是合理的做法，并指出了目前该方法存在大规模学习语料难以获取、传统的词序列标注存在不稳定性和复杂性等弊端。为了消除障碍，笔者认为可将序列标注方法应用于存在大量人工标引数据的中文书目关键词抽取领域，并通过对现有标引数据的详细分析，对词序列标注方法进行了简化和修正，构建了基于字序列标注的关键词抽取模型，提出了该模型的基本思路和实现方案，并以某大学图书馆馆藏书目作为实验对象，论证了采用字序列标注从图书题名中抽取关键词的合理性、正确性和实用性。在仅以单汉字作为观察序列的情况下，F1 值达到了 87.36%，可见该模型具有很强的实用价值。

然而，本文仅仅提出了字序列标注方法的基本思路并验证了可行性，并没有对影响该模型的特征因素以及实验参数进行详细论证，没有计算出最佳的标注模型，因此还有待于今后进一步开展研究，具体包括训练样本规模的最合理化、融合多特征的观察序列扩展、特征模板的选择、字角色空间模型的优化、CRFs 参数的控制以及不同序列标注机器学习算法的比较等；此外，序列标注是一种实用的机器学习方法和自然语言处理技术，它同样可以用于从文本中抽取其他语言片段，例如从专利文献中抽取领域术语、标注领域术语间关系等等。

参考文献：

[1] Hulth A. Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction[D]. Stockholm: Stockholm University, 2004.

[2] 王昊, 严明, 苏新宁. 基于机器学习的中文书目自动分类研究[J]. *中国图书馆学报*, 2011(5): 28 - 39

[3] 章成志, 苏新宁. 基于条件随机场的自动标引模型研究[J]. *中国图书馆学报*, 2008(5): 89 - 94, 99

[4] Chu C. M., O'Brien A.. Subject Analysis: the First Critical Stages in Indexing[C]. *Journal of Information Science*, 1993, 19(6): 439 - 454.

[5] 邓箴, 包宏. 改进的关键词抽取方法研究. 计算机工程与设计, 2009(20): 4677-4680, 4769.

[6] 张雪英, Jürgen Krause. 中文文本关键词自动抽取方法研究[J]. *情报学报*, 2008(4): 512 - 520.

[7] 徐文海. 温有奎. 一种基于 TFIDF 方法的中文关键词抽取算法[J]. *情报理论与实践*, 2008(2): 298-302.

[8] 张庆国, 薛德军, 张振海, 等. 海量数据集上基于特征组合的关键词自动抽取[J]. *情报学报*, 2006(5): 587 - 593.

[9] 杨洁, 季铎, 蔡东风, 等. 基于联合权重的多文档关键词抽取技术[J]. *中文信息学报*, 2008(06): 75-79.

[10] 王灿辉, 张敏, 马少平, 等. 基于相邻词的中文关键词自动抽取[J]. *广西师范大学学报：自然科学版*, 2007(2): 161 - 164.

[11] 李素建, 王厚峰, 俞士汉等. 关键词自动标引的最大熵模型应用研究[J]. *计算机学报*, 2004: 1192 - 1197.

[12] Frank E., Paynter G. W., and Witten I. H.. Domain-Specific Keyphrase Extraction[C]. In

Proceedings of the 16th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, Morgan Kaufmann, 1999: 668 - 673

[13] 章成志. 基于集成学习的自动标引方法研究[J]. *情报学报*, 2010(1): 3 - 8

[14] Zhang K., Xu H., Tang J., etal. Keyword Extraction Using Support Vector Machine[C]. In: Proceedings of the Seventh International Conference on Web-Age Information Management (WA IM2006) , Hong Kong, China, 2006: 85 - 96.

[15] ICTCLAS 分词系统[OL]. [2011-8-13]. http://ictclas.org/.

[16] 黄昌宁, 赵海. 由字构词——中文分词新方法[C]. 中国中文信息学会二十五周年学术会议报告, 2006: 53 - 63.

[17] Zhou Guodong, Su Jian. Named Entity Recognition using an HMM-based Chunk Tagger[C]. In: Proceedings of the 40th Annual Meeting of the ACL, Philadelphia, July 2002: 473 - 480.

[18] Olover B., Franz J. O.. Hermann N.. Maximum Entropy Models for Named Entity Recognition[C]. Proceedings of the Conference on Natural Language Learning at HLT-NAACL. Edmonton, Canada, 2003: 148 - 151.

[19] Settles B. Biomedical Named Entity Recognition using Conditional Random Fields and Rich Feature Sets[C]. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Application (NLPBA). Geneva, Switzerland, 2004: 104 - 107.

[20] Kudo T.. CRF++: Yet another CRF Toolkit[OL]. [2011-8-7]. http://crfpp.sourceforge.net/