基于自然标注信息和隐含主题模型的无监督文本特征抽取[1]

饶高琦[1,2]，于东[1]，荀恩东[1]

（北京语言大学, 1.汉语国际教育技术研发中心；2.中国语言政策与标准研究所，北京，100083）

摘要：术语和惯用短语可以体现文本特征。无监督的抽取特征词语对诸多自然语言处理工作起到支持作用。本文提出了“聚类-验证”过程，使用主题模型对文本中的字符进行聚类，并采用自然标注信息对提取出的字符串进行验证和过滤，从而实现了从未分词领域语料中无监督获得词语表的方法。通过优化和过滤，我们可以进一步获得了富含有术语信息和特征短语的高置信度特征词表。在对计算机科学等六类不同领域语料的实验中，本方法抽取的特征词表具有较好的文体区分度和领域区分度。

关键词： 自然标注信息；自然语块；隐含主题模型；领域特征；文体特征

Unsupervised Text Feature Extraction based on Natural Annotation and Latent Topic Model

RAO Gaoqi[1, 2], YU Dong[1], XUN Endong[1]

(Beijing Language and Culture University, 1.International R&D Center for Chinese Education; 2.Institute for Chinese Language and Policies and Standards, 100083, China)

Abstract: Text features are often shown by its terms and phrases. Their unsupervised extraction can support various natural language processing. We proposed “Cluster-Verification” method to gain the lexicon from raw corpus, by combining latent topic model and natural annotation. Topic modeling was used to cluster strings, while we filtered and optimized its result by natural annotations in raw corpus. High accuracy was found in the lexicon we gained, as well as good performance on describing domain belonging and writing style of the texts. Experiments on 6 kinds of domain corpora showed its promising effect on classifying their domain belonging and writing style.

Key words: natural annotation; natural chunk; latent topic model; domain feature; stylistic features

1 引言

文本特征可以从两方面得到体现：领域性和文体性。前者通过术语的形式得到体现，而后者往往以惯用短语的方式出现。本文统称这两者为特征词语。对于自然语言处理而言，以词和短语的形式体现出的文本的特征，可以对分词、文本分类和自动文摘等诸多自然语言处理工作提供支持。

当前文本特征刻画的思想多来源于 BOW（Bag of Words）模型或其变种，如带有领域词典的特征袋 BOF 模型[1]，使用加入命名实体描写的 FLIC[2]，带有短语与 n-gram 描写的 STC[3]和利用词间关系进行描写[4]等。它们大多在自建或通用测试集上达到了 80-95%的精确率。但是注意到现有的方法都以词项为语义的承载单元，因而过分依赖于分词和命名实体识别的信息。中文分词虽然在通用语料上取得了较大进步，但在领域性较强的语料中，以术语为代表的未登录词依然是分词 F 值失落的重要原因。并且领域语料的标注语料十分稀少，训练十分困难。有些领域甚至连生语料也较难收集。本文基于以上困难，提出了一种无需分词与命名实体信息的无监督抽取方法。其对面向领域语料的自然语言处理具有重要的价值。

自然标注信息（Natural Annotation）来自于语料本身，本质上是语言使用者提供的一种

[1] 国家自然科学基金项目（61300081，61170162）；国家社科重大基金项目（12&ZD173）；国家语委科研基金项目（YB125-42）；北京语言大学研究生创新基金（14YCX074）

原始众包标注。在海量语料中对自然标注信息的挖掘和获取几乎不需要标注语料，极少需要先验知识。以 LDA（Latent Dirichlet Allocation，隐含狄利克雷分布）模型为代表的主题模型具有较好的无监督聚类功能，可以对词语间的隐含语义关系进行一定的描述。我们利用这一特点对文本内容进行事先聚类，可以有效的克服自然标注信息需要海量训练语料的缺陷，将自然标注信息的使用大大“轻量化”，使特征词语的整个抽取过程可以在较小规模语料上完成。所以本文将主题建模和自然标注信息相结合，提出了“聚类-验证（Cluster-Verification）”方法，以较少的信息注入在小规模语料上获取文本的领域特征和文体特征。不同于以往的研究，本文方法不需要分词和命名实体信息。而且，其提取的特征并不拘泥于传统意义上词的范畴，与阅读直觉更加相符。

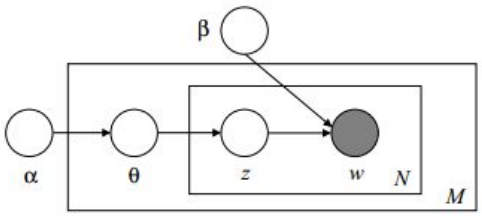
本文的组织结构如下，第二节简述本文的工作基础即 LDA 模型和自然标注信息，第三节介绍在基于 LDA 和自然标注信息的无监督“聚类-验证”方法，第四节将描述在 217 万字计算机领域语料和环境、金融等其余五类领域语料上的实验和结果。在第五节中，本文将讨论了特征词表的领域区分度和文体区分度，并实验中出现的一种术语“生长现象”，第六节是结论和未来的工作。

2 工作基础

2.1 隐含主题模型

隐含狄利克雷分布模型 LDA 最早由 Blei、Ng 和 Jordan 在 2003 年提出[5]，用以发掘文本中的隐含主题。LDA 模型是一种完全的生成模型，其假设一个文本集中存在以狄利克雷分布为先验的隐含主题分布，而对于任一主题存在一个隐含的词选择分布。它的概率图表示如图 1 所示，M 为文档集中的文档数目，N 为文档中的词数。整个过程中的外显参数为词 w 和超参数 α 与 β 。其经验性的选择一般为 α =50/T， β =0.01. T 为主题个数。

图 1 隐含主题模型 LDA 的概率图表示



2004 年，Griffiths 与 Steyvers[6]开始采用吉布斯采样（Gibbs Sampling)学习 LDA 模型。本文使用的工具也采用了该采样方法。

2.2 自然标注信息

自然标注信息的概念来自互联网应用中的用户生成信息(User Generated Content)。其作为概念在国内最早由孙茂松在 2011 年提出[7]，主要用来从海量互联网数据中提取对自然语言处理可用的信息。其后饶高琦和黄志娥[8,9]将其发展，用于无监督发掘语料库中的词汇信息。不少学者也在中文分词和博客信息挖掘中使用自然标注信息进行尝试[10-12]。近几年，关于自然标注信息的研究日益广泛，逐渐扩展到信息检索[13]，社会计算[14],情感分析[15]，信息抽取[16]。然而总体而言，该领域的研究都需要较大的训练语料，并且方法仍处于起步阶段。

人类语言中蕴含有丰富的自然标注信息。其中以标点符号为代表的显性自然标注信息对词边界的探测具有重要意义。如下式所示，Pi 为一处自然标注（如标点符号），如果其在中文里绝不与其他符号构成词，则其自身就形成了一处天然的词边界。饶高琦[3]的工作发现，大规模语料中仅通过显性自然标注信息（主要包括标点符号、拉丁字母和阿拉伯数字）对字

符串的切分就可以获得现汉词典中几乎所有的词项。而且即便仅仅使用 1998 年人民日报的语料也可以获得现汉 87.84%的词项。这样出现在显性自然标注信息之间的汉字字符串被称作“自然语块”（Natural Chunk），其边界为是词边界的子集。

$$C_1C_2\ldots..C_iP_1C_{i+2}C_{i+3}\ldots..C_nP_kC_{n+2}C_{n+3}\ldots..C_m$$

因此，本文假设在领域语料中通过自然标注信息对字符串的切分也可以无监督的获得具有领域性的词语或语块。本文使用了来自 2002 年《计算机学报》的文本 220 篇。将标点符号、运算符号、拉丁字母和阿拉伯数字视作标记词边界的自然标注信息，并替换为标记‘SPACE’。这样整个语料仅存留汉字字符和‘SPACE’标记（替换后约 217 万字）。将由此形成的自然语块进行统计可以获得表 1 结果。

表 1. 计算机科学语料上的自然语块举例

自然语块	频次	语块	频次	自然语块	频次
和	1974	图	580	为	475
其中	940	所示	508	中	464
一	847	因此	485	若	458
在	630	则	481	与	449
……		……		……	
国家自然科学基金	94	计算机学报	58	中国科学院计算技术研究所	38

注意到，由自然标注信息标识的词边界具有很高的正确率。饶 2013 和黄 2013 都报告了由显示自然标注信息而来的词边界识别在通用语料上具有较高正确率[3,4]。而在本文所使用的计算机科学领域语料中，显示标注信息和汉字字符结合成词的现象同样少见。这样的现象多为如“x^2 检验”这样处于半译写状态的外来术语。

但是从上表所示的现象中还可以注意到，该抽取结果并不能体现出其作为科技论文的文体特征。另一方面在领域特性上，语块频次也无法显示其作为计算机语料的特性，排位最高的术语仅占到第 80 位。其他技术性术语排名更加靠后。其原因在于文本所具有的领域性并不完全由字词的频次体现。领域性的短语和词汇往往隐藏在文本所述的众多主题之中。因此有必要使用主题建模的方法对其进行挖掘。

3“聚类-验证”方法

3.1LDA 聚类方法

本文假设，如果一个字符串可以形成稳定使用的词或惯用短语，则其内部成分（字或字组）出现的相对位置，上下文环境，甚至概率都趋于相近。又因为稳定使用的词（或短语）的子串共同参与了该词（或短语）的语义表达，则它们也倾向于出现在同一个主题之中。基于统计方法的主题建模通常以词簇来表现主题。在只存在字符边界（没有分词信息）和显性自然标注信息的语料中，构成一个词（或短语）的字（或字组），也倾向于被 LDA 模型聚于同一主题内，如下例。

0 号主题：	型 模 化 的 简 细 存 原 表 簇 角 顶 三 们 外 量 向 二 我
1 号主题：	概 的 念 格 所 而 则 更 应 了 及 称 被 本 某 名 前 当 优
2 号主题：	信 息 中 据 来 获 确 基 的 对 相 地 是 通 部 标 首 目 三
3 号主题：	存 储 问 访 一 之 和 方 共 享 可 为 表 冲 完 执 指 比 或

例 1. 无词边界语料上的 LDA 聚类结果举例

形成上例的语料为《计算机学报》生语料，标点符号为停用词，处理单元是汉字，参数为 α =0.23，β =0.01，迭代次数 1000。注意到，虽然理论上 LDA 模型在生语料上体现出了较好的字聚类性能，但是构成某词语的汉字也可能构成其他主题的其它词语，因此一个词的内部构件间的概率并非完全相等。加之 LDA 模型的采样方法，这些都决定了一个主题虽倾向于包含构成一个词的众多子串，但其相对位置和词内原来的字序很少相同。对此，本文选取每个主题中出现概率最高的 N 个字，对其进行 N x N 的两两匹配，形成每个主题的候选词集 S。又因为自然标注信息标记词边界具有高正确率的特性。我们使用它对 S 中的成员进行过滤和确认，经过优化打分（即自然标注信息的验证过程，3.2 节里的公式 e）之后形成筛选词表。

在生语料中，经过一次主题成员的两两匹配，所获得的候选词显然都是二字串。我们选取其中高置信度的成员，回标原始语料。从而增加了原始语料内的词边界信息，形成结构更加丰富的“字-词”混合语料。这一过程改变了 LDA 的聚类对象和概率空间，使得主题成员得到改变，以便进行下一轮迭代，获取更多特征词语。

3.2 自然标注信息验证过程

对 LDA 聚类产生的候选词表 S 中的成员，我们可以使用其在原始语料中与自然标注信息的相对位置来判断其成为词（或短语）的可能性。因为本工作采用了显性的自然标注信息如标点符号和数字，则可以认为它们直接表达了作者的切分意图。在语料中，自然标注信息被替换为‘SPACE’符号。两个‘SPACE’标记之间的字符串（自然语块）$C_{i+1}\cdots\cdots C_{i+n}$ 可以被认为是一个独立且分单元，其左右边界为词边界。它未必总是语言直观上的词，然而语块 $C_{i+1}\cdots\cdots C_{i+n}$ 与语言直观上的词 Word 之间必然存在如下四种包含关系：

$$\text{Word}= C_{i+1}\cdots\cdots C_{i+n} \quad ; \quad \text{(a)}$$

$$\text{Word}= C_{i+m}\cdots\cdots C_{i+n}\ , n>m>1; \quad \text{(b)}$$

$$\text{Word}= C_{i+1}\cdots\cdots C_{i+m}\ , n>m>1; \quad \text{(c)}$$

$$\text{Word}= C_{i+k}\cdots\cdots C_{i+m}\ , n>m>k>1; \quad \text{(d)}$$

即 Word 与自然语块的两个边界同时邻接（a，Word 等于语块本身），与自然语块左边界邻接（b），与自然语块右边界邻接（c）和成为自然语块的子串（d）。

对于待验证的词（或短语），其是否稳定使用则可以用其在原始语料中出现四种蕴含关系的频次来衡量。因此使用如下公式来对候选词集成员打分。

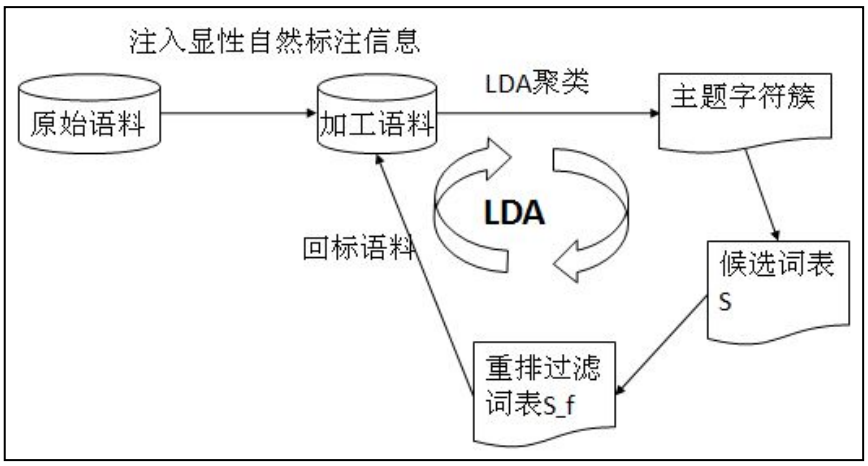
$$Score = \lambda_b f_b + \lambda_s (f_l + f_r) + \lambda_n f_n \ ; \text{(e)}$$

Score 为候选词集成员成为一个稳定使用词的可能性打分，而 λ_b、λ_s、λ_n 是四种蕴含状况所占有的权重。当一个候选词 Word 首尾与自然语块一致，都是词边界的时候，其成为一个语言直觉上的词的可能性最大。我们朴素的认为与左邻接与右邻接的情况在语言使用上有略小且相等的权重。考虑到使用语料的规模较小，有很多词没有机会出现在含有自然标注信息的上下文中，但其在单纯汉字字符上下文中出现的频率也应被考虑。因而其参数有 $\lambda_b >$

$\lambda_s > \lambda_n$（f），且 $\lambda_b+2\lambda_s+\lambda_n=1$（g）的关系。

候选词表成员经过打分排序，形成重排过滤词表 S_f。利用该词表，使用最大正向分词方法，对语料进行回标，使得原始语料的结构得到改变，词边界更加丰富，进而优化下一轮迭代中的聚类结果。整个聚类-验证的迭代过程如图 2 所示。

图 2. 聚类–验证方法的工作流程



4 实验

4.1 自然标注信息的注入

本文选取了来自 2002 年《计算机学报》的文本 220 篇，汉字字符约 217 万个。将标点符号、运算符号、拉丁字母和阿拉伯数字视等显性自然标注信息替换为标记‘SPACE’后（语料样例见例 2），共形成自然语块 87348 个（举例见表 1）。

SPACE 在 处 理 器 内 部 有 SPACE 个 开 关 控 制 这 SPACE 个 端 口 之 间 的 连 接 关 系 SPACE 如 图 SPACE 所 示 SPACE 这 SPACE 个 端 口 之 间 共 有 SPACE 种 连 接 方 式 SPACE 如 图 SPACE 所 示 SPACE 处 理 器 内 部 的 这 些 开 关 可 以 在 算 法 的 执 行 过 程 中 动 态 地 置 成 开 或 关 SPACE 从 而 将 整 根 总 线 分 成 一 些 相 互 独 立 的 子 总 线 SPACE

例 2. 引入显性自然标注信息后的原始语料举例

本文使用马萨诸塞大学的开源工具 Mallet 实现 LDA 模型[17]，并根据经验选择主题数目为 220 个，α =50/220，β =0.01，迭代次数 1000。第一轮主题训练结果的举例见例 1。对每个主题我们选取出现概率最高的 20 个字进行 N x N 组合，形成每个主题的候选词集 S。在自然标注信息验证过程中，使用 3.2 中的打分公式 e。并在约束条件 f 和 g 的限制下，根据经验选取了 e 的系数λ_b=0.5, λ_s=0.2, λ_n=0.1。

表 2. 过滤重排词表中打分前十五的词语举例

候选词	打分	候选词	打分	候选词	打分
算法	620.8	模型	580	问题	250.2
一个	443.1	可以	508	图像	237.6
方法	404.3	计算	485	中的	232.7
我们	390.1	数据	481	网络	221.3

系统	353.2	进行	261.5	时间	215.9

第一次迭代共得到非零分候选词 4708 个，得分最高的 15 个如表 2 所示。因为原始语料没有词边界，则聚类对象均为单字，故得到的候选词都是二字词。与表 1 相比，其对领域性的表达得到了很大增强。如果将单字词的组合视作词组而判为正确（因为其并未打破词边界），得分最高的 600 个候选词中正确率为 92.7%。

并且注意到 600 个候选词中 44 个错例里有 43 个是和“的”字的组合，如“的对”、“的数”、“义的”等。唯一一个不包含“的”的错例是“务器”。其原因在于“的”字是现代汉语各类语料中出现频率最高的汉字。虽然很少与显性自然标注信息邻接出现，但是其自身过高的频率也拉高了它和自己邻接汉字组成的候选词的得分。可以观察到，“的”字组合错例中的另一个字都是计算机领域中高频词的首字或末字，如“的对（话、象）”和“的网（络、关、口、端、卡）”等。

“的”、“着”、“也”、“是”与“和”等在自然标注信息的研究中通常被称作隐性自然标注信息[3]。本文参考了饶 2012 中在大规模通用语料中的统计结果，从选取词边界标记置信度较高的隐性自然标注信息 11 个[2]，对候选词集 S 进行过滤，大大的提升了正确率（99.8%）。并且为了在语料回标过程中减少减小交搭型歧义的出现，我们将词语长度加入打分公式以获得更长的切分单元。修正后公式为 Score’ =Length(Word)*Score。

4. 2 迭代实验

对于经过显性和隐性自然标注信息处理的重排过滤词表 S_f，本文取打分最高的 5%作为词表，通过最大正向分词的方法重新标入原始语料。例 2 中的语料则成为例 3 的状态。

在 处理器 内 部 有 SPACE 个 开 关 控制 这 SPACE 个 端 口 之 间 的 连接 关系 SPACE 如图 SPACE 所示 SPACE 这 SPACE 个 端 口 之 间 共 有 SPACE 种 连接 方式 SPACE 如图 SPACE 所示 SPACE 处理器 内 部 的 这些 开 关 可以 在 算法 的 执行 过程 中 动态 地 置 成 开 或 关 SPACE 从而 将 整 根 总 线 分 成 一些 相 互 独 立 的 子 总 线 SPACE

例 3. 第二轮迭代后回标形成的语料样例

对重新注入过自然标注信息的语料进行新的 LDA 聚类。过程同前节所述，总的试验流程如图 2 所示。

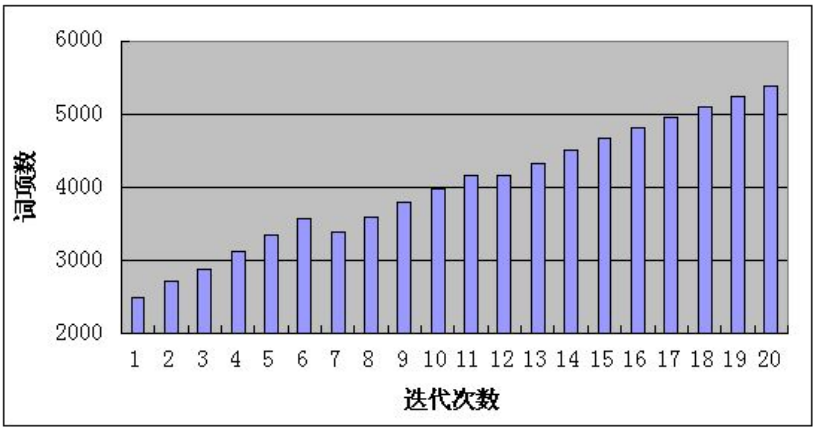
0 号主题：的 超顶点 图 邻 中 有 最小 量 通 分 树 孔 连 矩阵 含 维 次 所 相
1 号主题：区间 值 离散化 概率 属性 模型 参数 叶 斯 样本 学习 贝 数目 合 督 处理 数 监 混
2 号主题：时间 本 次 可 新 并 复 图 所 重 在 及 若 将 可以 号 结果 多 过
3 号主题：图形 相似 尺 寸 几何 相似性 特征 识别 图 中 元素 结构 约束 对 其 性 方法 模式 连接

例 4. 第三轮迭代中经过 LDA 聚类后的主题举例

随着自然标注信息的注入，原始语料的边界信息更加丰富。抽取出的“总词表”规模随迭代次数增长（图 3）。原始语料的字表规模为 2495 个，即语料共使用不同的汉字 2494 个，外加代表显性自然标注信息的‘SPACE’标记，在第 20 次迭代后词表则为 5376 个。

图 3. 重排过滤词表的规模随迭代次数的变化

[2] '是','和','的','也','着','与','个','在','之','有','为'



通用词语的领域性和文体性特征均不明显。因此为了进一步提高重排过滤词表中术语和特征词组所占的比例，降低通用词语的排名，本文使用了可公开获取的 1998 年 1 月的人民日报词表对重排过滤词表进行剪枝。在诸次迭代所产生的词表中，剪枝率为 5.2-21.3%。

表 5、6 和图 5-7 分别为在计算机领域语料上迭代 20 次过程中，特征词表、通用词、术语、特征短语和抽取规模的变化。还可以看到词表正确率（既抽取出的字符串是词或词组，下同）基本稳定，术语和短语数量稳步上升。术语比例在 7 次迭代前后收敛。类似的，本文也在其它领域语料上进行了相同的实验。表 3 为在环境科学、金融学、医学和土木工程四个领域各选取论文 100 篇，迭代 9 次形成的结果。它们与在计算机科学语料上第 9 次迭代形成的结果具有可比性。

表 3. 环境、金融、医学和土木工程类语料上的“聚类-验证”实验结果

领域	词数	正确率	术语数	术语比例	短语	短语比例	迭代轮次
环境科学	612	.8758	308	.503	79	.129	9
金融学	724	.9144	196	.271	52	.072	9
医学	657	.8828	395	.601	108	.164	9
土木工程	692	.8974	420	.607	67	.097	9

5 实验分析和讨论

5.1 领域区分度

通过观察不同语料中特征词表，我们也验证了特征词表和术语对领域性的充分刻画。如表 4 所示，学科间的差异可以得到较好体现。其中我们对语料分词后，不同领域语料的最高频 1000 个词（已去停用词）形成的词表之间的重合度远远大于特征词表（术语+惯用短语）和其中术语的重合度。这表明了抽取出的特征词表对不同领域文本具有很强的区分度。

表 4. 不同领域语料特征词表中的术语重合度（左栏为分词最高频 1000 词的重合度，中栏为特征词语表的重合度，右栏为术语的重合度）

	环境科学			金融			医学			土木工程			计算机科学		
环境科学	1			.32	.085	.003	.38	.092	.093	.44	.098	.048	.36	.136	.115
金融	.32	.072	.005	1			.28	.043	.005	.44	.084	.01	.34	.091	.001
医学	.38	.085	.073	.28	.047	.003	1			.33	.059	.005	.28	.13	.068
土木工程	.44	.087	.035	.32	.088	.005	.33	.056	.003	1			.46	.13	.066
计算机科学	.36	.093	.064	.34	.074	.004	.28	.06	.048	.46	.101	.05	1		

5.2 特征短语与文体区分度

对过滤后的特征词表进行标注和统计可以观察到随着迭代次数的增加，词表规模、术语和通用词的绝对数量都在增加，整体正确率基本稳定（如表 6 和图 6）。比例和绝对数量增长最为明显的是一类“特征短语”。而文体特征可以由这类短语来进行刻画。

本文将特征短语分为两类：术语增生而形成的和表示习惯用法的。如“服务器上”、“满足约束条件”和“基于斐波那契数列”这样的短语包含有术语，属于术语增生型。“一种基于”、“我们提出了”、“下面给出”和“如图”等则属于惯用短语。 对 20 次迭代后产生的短语进行统计发现 18.9%的短语为术语增生型，惯用短语占 81.1%，后者则普遍具有学术写作的文体特点。由于上一部分实验中选取的五种语料均为科技论文，语体相同。本文以《圣经》马太福音为语料进行实验。其与计算机科学语料短语的重合度仅为 1.4%。例 5 为两者惯用短语的举例。

注意到，特征短语与陈文亮[1]的工作中所提出的特征关联词有一定的相似性，但它的粒度超过复合词，多为短语。例如本文方法提取的“本文采用”比“本文”和“采用”两个词更能体现科技论文的文体性。

《马太福音》	《计算机学报》
我告诉你们	本文采用
所以你们要	当且仅当
记着说	实验表明
不要怕	我们给出了
他们回答说	定义如下

例 5. 马太福音与计算机学报的特征短语举例

在特征短语中，术语增生而得的短语比例相对较少，而且集中于一些极高频术语的周围，如“在算法”、“算法中”、“由算法”和“算法进行”等，其依然带有较强的领域性。

5.3 复杂术语生长

在诸次迭代中，LDA 聚类后形成的主题由“字簇”变为“字-词簇”，并逐渐向“词簇”变化（例 4 为第三次迭代中 LDA 聚类产生的簇）。类似的，在候选词表与后续的过滤重排词表中，词语长度也在逐渐变长，呈现出一种生长的态势。如上一部分提到的第一次迭代中的错例“务器”，在第二次迭代中就生长完全成为“服务器”。更长的术语“服务器网络”则是在第 17 次迭代出现。其重要组分“网络”则是在第一次迭代中形成的二字术语。

又如“自组织隐马尔可夫模型”这一术语，它的生长过程如图 4 所示，数字为该字符串第一次出现时的迭代轮数。其中“自组织”和“模型”在语言学上都是该术语的子成分。而且它们出现的迭代轮次间隔巨大。这是因为“自组织”和“模型”频率很大，而且本身可以出现在大量的其他术语中，因而其作为“自组织隐马尔可夫模型”的组分不如其他组分（如“隐马尔可夫”）的结合程度高。在未来的工作中可以利用这样的信息对诸如此类的长术语进行内部结构的分析。

图 4. “自组织隐马尔可夫模型”的生长过程

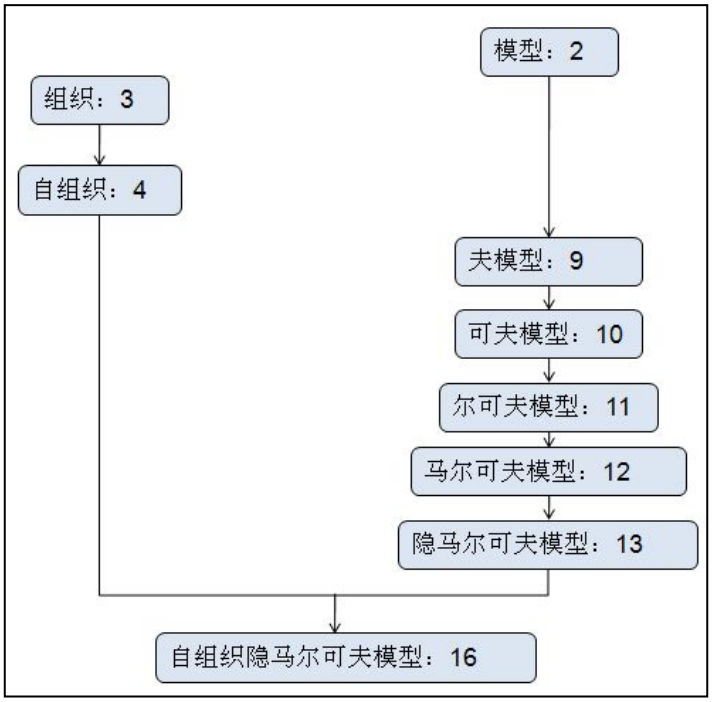


表 5 抽取词数和通用词比例

词语数	正确率	通用词比例	迭代轮数	词语数	正确率	通用词比例	迭代轮数
227	0.991	.670	1	970	0.933	.191	11
295	0.932	.251	2	1017	0.938	.180	12
480	0.952	.235	3	1096	0.935	.172	13
617	0.961	.214	4	1162	0.927	.164	14
706	0.950	.120	5	1234	0.922	.159	15
680	0.950	.257	6	1309	0.918	.160	16
739	0.953	.214	7	1377	0.908	.157	17
800	0.960	.200	8	1449	0.894	.153	18
888	0.910	.190	9	1523	0.870	.146	19
942	0.899	.182	10	1583	0.872	.140	20

表 6. 特征词表中术语比例和特征短语比例

词语数	术语比例	短语比例	迭代轮次	词语数	术语比例	短语比例	迭代轮次
227	0.308	0.013	1	970	0.645	0.097	11
295	0.634	0.047	2	1017	0.635	0.123	12
480	0.675	0.042	3	1096	0.627	0.137	13
617	0.682	0.065	4	1162	0.608	0.156	14
706	0.674	0.076	5	1234	0.606	0.157	15
680	0.657	0.035	6	1309	0.600	0.159	16
739	**0.677**	0.062	7	1377	0.568	0.183	17
800	0.683	0.079	8	1449	0.556	0.186	18
888	0.631	0.089	9	1523	0.540	0.184	19
942	0.635	0.083	10	1583	0.541	**0.191**	20

图 5. 短语数量随迭代次数变化

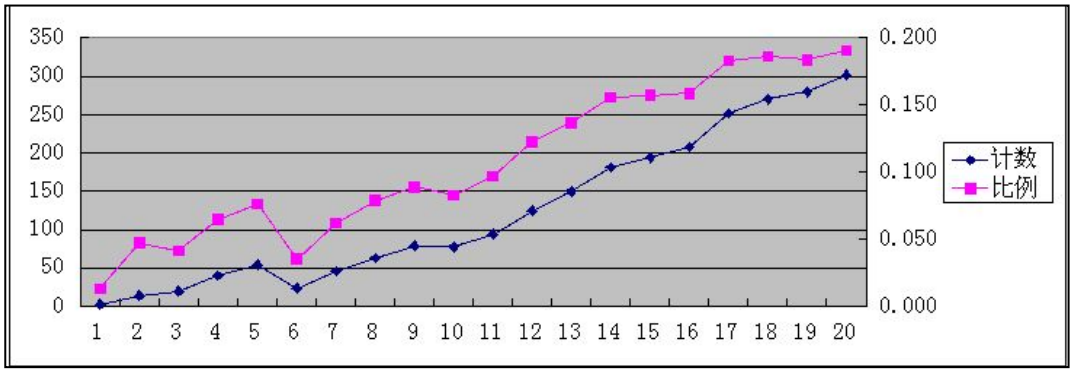


图 6. 术语数量随迭代次数变化

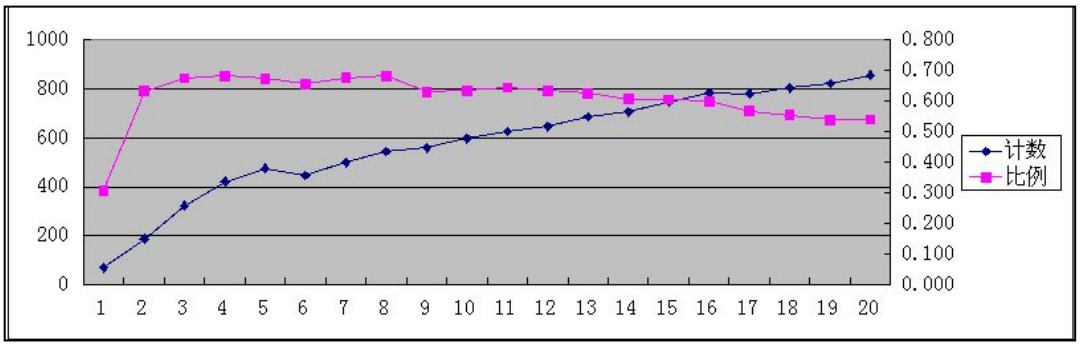
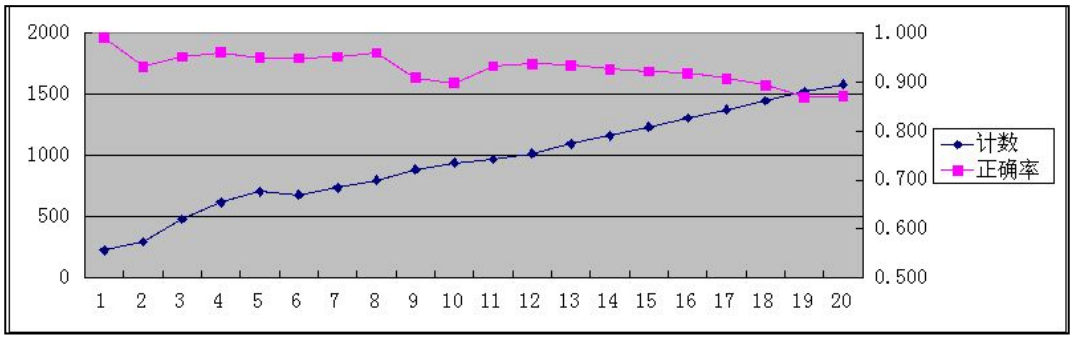


图 7. 特征词表规模与正确率



6 结论与展望

本文提出了无监督提取文本特征的“聚类-验证”方法：使用隐含主题模型在领域语料中进行无监督聚类，并采用隐性和显性的自然标注信息对提取出的候选字串进行验证，从而获得特征词表。统计显示该词表具有较高的正确率。通过对原始语料进行回标，我们改变主题模型的概率空间和字词分布。迭代多次后可以获得较好体现语料领域特征和文体特征的词语表。实验从 217 万字的计算机领域语料中获得了可表征其领域特性和文体特征的词语表，并在和环境、金融等语料上实验的比较中体现出了其领域性差异。我们还通过科技论文和《圣经》语料实验结果的对比，验证了它对语体差异的描写。

本文方法使用主题模型对候选字符串进行预聚类，有助于加速通过自然标注信息发现词语的过程。相较于以往的自然标注信息使用，本方法所需背景语料少。全过程中待处理语料的信息注入仅限于显性自然标注信息（标点符号、运算符号、字母和数字）与 11 个隐性自然标记，在过滤优化过程中也仅使用了 1998 年 1 月人民日报词表。

不同于以往的研究，该方法不需要分词语料和命名实体信息。因而对缺乏资源的语种和语料处理具有较好的借鉴意义。然而本文只是无监督聚类和自然标注信息相结合的一次尝试。主题模型本身的优化、求优打分的调参和自然标注信息的灵活应用都有待未来更深入的

研究。尤其是在词语生长这一现象中，如何使用不同无监督学习策略来控制和发掘词语的组分和生长过程，将对更深入的研究构词，实现词法自动分析带来巨大帮助。

参考文献

[1]陈文亮, 朱靖波, 朱慕华, 姚天顺. 基于领域词典的文本特征表示，计算机研究与发展 42 卷 12 期 2005. P2154-2160

[2]赵世奇, 刘挺, 李生. 一种基于主题的文本聚类方法，中文信息学报 21 卷 2 期 2007.P59-62

[3]Zamir O and Etzioni O. Web Docu ment C l us t er i n g: A Feasibility Demonstration [A] . In : Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval [C].1998.46-54.

[4]吴双, 张文生, 徐海瑞. 基于词间关系分析的文本特征选择算法，计算机工程与科学 34 卷 6 期 2012. P140-145

[5]Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, 3:993 – 1022

[6]Griffiths T, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(90001):5228 – 5235.

[7]孙茂松: 基于互联网自然标注资源的自然语言处理，中文信息学报 25 卷 6 期 2011.P26-32.

[8]饶高琦, 修驰, 荀恩东. 语料库自然标注信息与中文分词应用研究, 北京大学学报（自然科学版）2013 Vol49. Nr.1

[9]HUANG Zhie, XUN Endong, RAO Gaoqi, YU Dong. Chinese Natural Chunk Research based on Natural Annotations in Massive Scale Corpora -- Exploring Work on Natural Chunk Recognition using Explicit Boundary Indicator ， Lecture Notes in Artificial Intelligence,Vol.8202

[10]Zhongguo Li, Maosong Sun : Punctuation as Implicit Annotations for Chinese Word Segmentation. Computational Linguist. 2009 Vol.35 Nr.4

[11]Xiance Si, Zhiyuan Liu and Maosong Sun. Modeling Social Annotations via Latent Reason Identification. IEEE Intellingent Systems, 2010,25(6):42-49

[12 刘知远, 司宪策, 郑亚斌. 中文博客标签的若干统计性质[C]//第七届中文处理国际会议 (ICCC). 2007.

[13]Jeremy Ginsberg, Matthew. H. Mohebbi, Rajan S, Patel, Lynnette Brammer, Mark S. Smolinski and Larry Brilliant, Detecting Influenza Epidemics with Using Search Engine Query Data [J] Nature 2009.(457).

[14]Sepandar D. Kamvar and Jonathan Harris. We Feel Fine and Searching the Emotional Web. Proceeding s of the Fourth ACM International Conference on Web Search and Data Mining. 2011

[15]Qu and Liu. Interactive Group Suggesting for Twitter. ACL'11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics

[16]Wu and Weld. Open Information Extraction using Wikipedia.ACL'10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics

[17]McCallum, Andrew Kachites. “MALLET: A Machine Learning for Language Toolkit.” http://mallet.cs.umass.edu. 2002.