北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis

一种基于流形距离的中文语块聚类分析方法

雷霖[1,†] 熊伟[1] 景宁[1] 肖建夫[2]

1. 国防科学技术大学电子科学与工程学院, 长沙 410073; 2. 长江日报报业集团, 武汉 430015; †E-mail: leihema@gmail.com

摘要 将中文语块分析看做词在句子内部聚类并标记语块类别的过程, 建立了中文语块分析的聚类模型。首先构建词的语法功能空间, 使用 ISOMAP 方法重构词空间的低维流形嵌入, 进而考察词在低维空间中的分布情况。在使用层次聚类方法分析语块时, 使用流形上的距离替代传统的欧式距离, 在算法复杂度可以接受的范围内, 提高了语块分析效果。

关键词 语块分析; 流形距离; 层次聚类; 语法功能空间

中图分类号 TP391

A Clustering Chunking Method Based on Manifold Geodesic Distance

LEI Lin[1,†], XIONG Wei[1], JING Ning[1], XIAO Jianfu[2]

1. College of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073;
2. Changjiang Daily News Group,Wuhan 430015; † E-mail: leihema@gmail.com

Abstract Regarding the Chinese chunker analysis as a procedure of inner-sentence word clustering and chunker type labeling, a grammar function space is constructed at first, and then embedding the space to a lower dimension space by applying ISOMAP to observe the distribution character of Chinese word in the embedding space. In the hierarchical clustering algorithm which is aiming at partitioning word into different clusters, the manifold geodesic distance is employed instead of Euclidean distance to measure the similarity between words. The algorithm facilitates the increment of Chinese chunker analysis performance under the condition of appropriate algorithm complexity.

Key words chunker analysis; manifold geodesic distance; hierarchical clustring; grammar function space

中文语块分析(chunking)是浅层句法分析(shallow parsing)中最主要的任务[1], 对机器翻译、信息检索等自然语言处理具有重要作用。目前对于中文语块尚未形成公认的权威解释, 沿用 Abney[2]的观点, 中文语块(下文皆称语块)是符合一定语法功能的非递归短语。每个语块都有一个中心词, 语块内的所有成分都围绕该中心词展开, 任何一种类型的语块内部不包含其他类型的语块。语块具有以下 3 个特点: 1) 全覆盖, 即将中文句子(下文皆称句子)分词之后, 每个词都属于一个语块; 2) 无嵌套, 即语块中不包含其他语块; 3) 无重叠, 即没有一个词属于两个语块。语块的划分大大降低了句子句法解析的难度, 为信息检索、自动翻译和中文智能校对等多个领域的研究提供基础。

目前对于中文语块分析主要采用有监督的序列标注方法, 包括 SVMs, CRFs, TBL 以及基于记忆的学习方法(MBL)等。文献[3]对这些方法进行了比较, 在相同的数据集(宾州中文树库 4.0)上的实验表明, SVMs 方法在语块识别效果方面优于其他方法, 而采用投票方法判定语块标记, 可以进一步降低算法的复杂度。在此基础上, 文献[4]提出基于大间隔的语块分析方法, 整体效果优于其他汉语语块分析方法。与此同时, 面向大规模真实语料库, Zhang 等[5]提出一种无监督方法, 在大规模 *N*-gram(*N* 取 2~20)中利用快速统计子串约减方法生成语块, 使得计算规模非常大的语块分析方法能够利用现有计算能力

新闻出版重大科技工程项目(1041STC40889)资助
收稿日期: 2012–05–31; 修回日期: 2012–08–15; 网络出版时间: 2012-10-26 17:04
网络出版地址: http://www.cnki.net/kcms/detail/11.2442.N.20121026.1704.004.html

实现。如果将 *N*–gram 看做一种统计意义上的词聚类，与此相类似的，文献[6]也提出利用聚类思想识别语块。这一方法目的是发现并提取词与语块之间关系的新特征，称之为词簇。该方法使用含有语法结构信息的信息熵定义簇间距离。使用了词簇信息的语块识别系统性能得到了提高。此外，文献[7]将语块识别问题描述成聚类问题，解决语料库稀疏问题和特征规模大的问题，验证了聚类方法在该问题上的有效性。

语块的聚类，是一个特征发现和抽取的过程。在大规模真实语料库处理过程中，对词的信息进行加工，使用高维数据处理方法是无法避免的。利用降维方法可以发现出数据的分布规律。由于这类数据并不是全局线性的，很多研究者提出大量的非线性方法分析数据特征。其中流形学习方法在分析处理高维数据方面取得了较好的效果，如图像处理、模式识别等领域。在自然语言领域，流形学习也应用在中文词汇的语义空间分析[8]、文本分类[9]等方面。但是在语法结构学习方面，尚没有出现相关的研究成果。本文试图挖掘词根据其语法功能在数据空间中的分布，将没有规律可循的词的语法功能转换为具有几何结构的数据集，从而反映词在构成句子上下文时体现的功能，赋予词除词性特征之外的高层语法特征，并研究这种规律在浅层语法分析中的应用。

本文首先建立中文语块分析的聚类模型，然后构造词的语法功能空间，并利用流形学习的方法挖掘高维词空间的低维流形嵌入，给出三维空间中词根据其语法功能的分布。进而使用这种分布规律，使用词在流形上的距离作为聚类结果的度量，对句子中的词进行聚类，进而获得中文语块。最后，通过在宾州中文树库[10]上的实验表明，在不使用大规模统计特性的基础上，验证了其有效性。

1 中文语块分析的聚类模型

语块以核心词为中心，句子内的其他词围绕核心词构成语块的附属成分，形成具有一定语法功能的词序列，语块之间具有明显的间隔。与词的词性特征类似，语块也具有语块类别标记。本文根据宾州中文树库 CTB5 提取出的语块类别包括：ADJP, ADVP, CC, DEC, DP, ETC, IJP, LCP, NP, PP, QP, VP。其中, CC 为[CC 和]、[CC 或]、[CC 并]、[CC 且]、[CC 而]等连词; DEC 为[DEC 的]、[DEV 地]、[DER 得]等“de 字”结构; ETC 为[ETC 等]; IJP 为[SP 呢]、[SP 啊]等虚词结构; LCP 为[LC 前]、[LC 来]、[LC 上]、[LC 里]、[LC 以后]等介词结构。将句子进行划分成语块之后，形成如下结构。为了简单起见，省略了标点符号的表示。

```
[欧佩克 原油 价格] [跌破 了] [每 桶] [10 美元] [大关], [而且] [仍 在] [下滑] 。
[NR NN NN ] [VV AS] [DT M] [CD M ] [NN ], [CC ] [AD AD] [VV ] 。
[NP ] [VP ] [DP ] [QP ] [NP ], [CC ] [ADVP ] [VP ] 。
```

语块中核心词的词性决定了语块的类别。如，语块“跌破了”中，“跌破”是核心词，由于跌破的词性为动词，因而语块的类别为 VP(动词语块)。从上述例子看出，找到核心词是语块识别的首要任务。中文语块识别的聚类分析模型基本思想就是围绕核心词，考察核心词周围词与核心词之间的关系，将属于同一语块的词聚合到一起，同时确定语块之间的边界。如“欧佩克原油价格”中，“价格”是核心词，“欧佩克”和“原油”都是附着在核心词“价格”上的限定成分，与核心词关系更紧密；同时“价格”的词性是 NN 名词，因此语块的类别为 NP 名词语块。“跌破”作为描述“价格”动作的成分，在句子结构中起独立的关键作用，因而不能与“价格”划分为同一语块，而是作为独立的语块表征与相邻语块具有同等的语法地位。按照聚类的思想来考察这个句子，“欧佩克”、“原油”与“价格”之间在语法功能上“距离”更近，聚为同一语块。

根据聚类问题的一般描述[11]，语块聚类的形式化描述为：令 $W=\{w_1,w_2,...,w_p\}$ 为中文句子，由词序列组成，w_i 表示第 i 个词，i=1, …, n, $L(W)$ 为句子长度。$C_t=\{c_{t1},c_{t2},...,c_{tq}\}, L(C_t)\leqslant L(W), c_{tq}$ 为划分的语块，使得：1) $\bigcup_{i=1}^{k}c_{ti}=W$，且 $\forall c_{ti},c_{tj}\in C_t, c_{ti}\bigcap c_{tj}=\varnothing$；2) $\mathrm{Min}(\mathrm{proximity}(w_{ti,m},w_{tj,n}))>\mathrm{Max}(\mathrm{proximity}(w_{ti,x},w_{ti,y}))$，其中 $\mathrm{proximity}(X,Y)$ 是表示词间距离的函数，$w_{s,t}$ 表示语块 s 中的第 t 个词。

2 基于流形距离的语块聚类算法

聚类算法中，需要对数据间的距离进行度量，常用的距离度量包括欧氏距离、曼哈顿距离及 minkowski 距离等。考察词在空间中分布情况，一般的方法都是将词建模成为由多个变量组成的向量进行

分析, 构成高维空间, 并使用这些距离度量计算词间距离。这些距离度量的基本假设是数据分布在欧式空间内, 数据之间的距离代表它们在所处空间中的相似性。但是词的空间分布并不满足欧氏空间假设, 采用欧式距离难以反映它们之间的复杂结构关系。通过构造词在欧式空间的分布并计算词间距离进行聚类的方法, 无法体现词在所构造空间中的全局相似性, 甚至影响聚类效果[12]。

流形和流形学习方法为我们提供了一种考察词在高维空间分布情况及其本质特征的方法。本文通过构造词的语法功能分布空间, 不采用传统的欧氏空间假设, 而是基于流形距离的概念, 保留空间中的非线性特征, 从词的语法功能角度描述其聚集情况, 重新把握数据的结构。面对词的语法功能空间的高维度特性和词分布的稀疏性, 采用流形学习的方法, 通过对高维空间的降维来发现词在该空间中的低维流形嵌入, 从而为词距离的度量提供依据, 进而实现基于流形距离的语块聚类分析方法。

流形是可以概括地描述为局部处处为欧式空间的拓扑空间, 具有局部可坐标化的特性, 从拓扑空间的一个开集(邻域)到欧式空间的开子集的同胚映射, 使得每个局部可坐标化。流形学习作为一种高维数据处理方法, 其目的是在保持数据全局或局部特性的基础上, 从高维采样数据中恢复低维流形结构, 即找到高维空间中的低维流形嵌入, 发现数据的内在规律。2000 年《Science》发表了 3 篇论文, 从认识上讨论了流性学习。从此之后, 流形学习领域产生了大量研究成果, 其中 ISOMAP[9](等度规映射)是一种非常具有代表性的非线性降维方法。

ISOMAP 的主要思想是利用局部邻域距离近似计算数据点间的流形测地距离(geodesic distance), 通过建立原数据的测地距离与降维空间距离的对等关系完成数据降维。Isomap 算法首先输入基于数据点之间的欧氏距离, 选择近邻点并根据近邻关系构造加权邻接图 G; 然后通过计算图上两点间的最短距离估计测地距离, 得到测地距离矩阵 D^G; 最后利用 D^G 构造数据的低维嵌入表示。首先, 我们借用 ISOMAP 和宾州中文树库来研究词在语法功能空间上的分布情况, 从而揭示词在这种分布下的相似性, 获得聚类算法中距离的含义。

2.1 词在语法空间上的分布

词的语法功能一直以来是语言学家工作内容之一。与英文不同, 在汉语中, 词的使用更具有灵活性, 缺乏英语词语中明显的语法功能标识, 对其逐一提取特征并构造基于语法特征的空间是一项难以完成的任务。但是, 如果从将词放在句子上下文中, 从其地位功能方面观察, 中文句子中词的使用还是有一定规律可循的。例如, 对名词来说, 如语块[边境 NN 开放 NN 城市 NN]。

如果按照词性来划分, 3 个词都是名词, 从功能角度来看, “城市”在词组合中处于核心地位, “边境”和“开放”是对“城市”的限定成分, 形成如图 1(a)所示结构。再如语块[社会 NN 经济 NN 发展 NN]。其中, “发展”处于核心地位, “社会”和“经济”用来并列地限定“发展”, 如图 1(b)所示。

宾州中文树库 CTB5 对词性和一些语法功能进行了标记, 却并没有为我们提示词的这些功能。为了表示词的这种语法功能, 我们引入依存语法中的词功能的定义, 利用宾州中文树库提取出语块, 并分析词在语块中的功能。根据依存语法理论, 词之间的关系分为附加、修饰等 27 类[13], 其中大部分却与语块分析任务无关。因此, 在词性的基础上, 我们采用词之间的“并列”、“附加”和“修饰”关系, 进一步提取词在句子上下文中的这些特征, 使词在语法功能分类上的意义更为明确。同时, 根据中文采用“依存成分前置”的构句特点, 对词进行如下定义。

定义 1 核心词: 在类别为 CC 的语块中处于核心地位的词, 其词性与语块类别标记相同, 表示为 CC-C。

定义 2 从属词: 在类别为 CC 的语块中产生“修饰”或“附加”等功能的词, 表示为 CC-M。

定义3 首词: 在类别为 CC 的语块中处于起始位置的词, 表示为 CC-I。

定义4 独立词: 独立构成类别为 CC 的语块的词, 表示为 CC-S。

根据这些定义, 我们对中文树库中的词重新进行标记, 使用<语块类别–词功能>对来表示, 如“NP-C, NP-I, NP-M, NP-S”等, 统计词在每种标记下的数量, 从而构造描述词语法功能的数据空间。

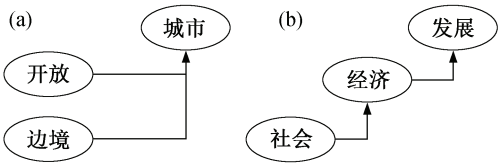

图 1 语块中词功能结构实例

Fig. 1 Example of different word grammar functions in chunk

首先从宾州中文树库中提取语块, 采用文献[14]中基于规则的语块提取方法进行数据加工。正如前述例子“边境开放城市”、“社会经济发展”等语块, 词的功能难以体现, 接下来使用依存语法分析工具, 构成词的依存结构, 采用哈尔滨工业大学 HIT_LTP 工具[15]中的 gparser 这一依存句法分析模块。

在根据宾州中文树库生成的语块中, 首先去除独立词构成的语块, 并对独立词标记为 CC-S。对两个词构成的语块, 如(QP (CD 十四) (CLP (M 个))), 将 CD 与 CLP合并为 QP, 并对词分别标记 QP-I 和 QP-C。对其他的由多词构成的语块, 使用依存句法分析工具来确定词的语法功能类别。

在去除意义相对明确的语块标记(如数词标记等)之后, 生成 33 类语法功能标签, 分别为 ADJPC, ADVPC, DPC, LCPC, NPC, PPC, VPC, CCS, QPS, ADJPM, ADVPM, DPM, LCPM, NPM, PPM, VPM, DECS, ADJPI, ADVPI, DPI, LCPI, NPI, PPI, VPI, ETCS, ADJPS, ADVPS, DPS, LCPS, NPS, PPS, VPS, IJPS。

由于词的词性、所处语块类别、语块内地位等不同, 每个词对应一个 33 维的向量, 向量的值为该词赋予的功能类别标签的数量, 表示为 $\mathrm{CCL}(w_i)=\{\mathrm{CCL}^1_{w_i},\mathrm{CCL}^2_{w_i},...,\mathrm{CCL}^{33}_{w_i}\}$, 其中 $\mathrm{CCL}^k_{w_i}$ 表示词 w_i 在第 k 类功能类别标签下的数量。令 N^k 为 k 类功能标签的总数量, 对 $\mathrm{CCL}(w_i)$ 归一化, 表示为 $\mathrm{CCL}(w_i)'=\{\mathrm{CCL}^1_{w_i}/N^1,\mathrm{CCL}^2_{w_i}/N^2,...,\mathrm{CCL}^{33}_{w_i}/N^{33}\}$。

在使用 ISOMAP 算法计算词空间的低维嵌入时, 需要首先计算词之间的距离, 这里我们采用欧式距离来计算, 然后构建词距离矩阵 D^G, 并作为 ISOMAP 算法的输入, 进而得到词在语法功能空间中的分布情况。为了直观的说明词在空间中的分布情况, 我们从三维空间和二维空间两个角度进行观察。图 2 是词空间在三维空间上的投影分布情况, 图 3(a)是图 2 在二维空间中的投影, 图 3(b)为图 3(a)的局部放大, 并进行了标注。

图 2 和图 3(a)反映的词空间分布特性表明, 高维的词的语法功能空间存在低维流形嵌入。由于词在语块中分布的稀疏性, 一般的词距离算法难以体现这种特点, 而通过流形学习, 词在语法功能上的差异被扩大, 更容易识别, 进而揭示词在语法功能空间上的聚集效果。在图 3(b)中, “以”、“有”、“到”等词由于其使用特点, 与其他词距离差异明显, 这也与通过分析这些词的使用情况得出的结论是一致的: 这些词单独成块的比例高, 与其他词之间的联系相对松散。而“投资”、“增长”、“发展”等词, 在语义空间上两者差别很大, 但是在功能上, 却具有较高的相似性, 在语块中语法功能上的统计规律也表现出这一特点。

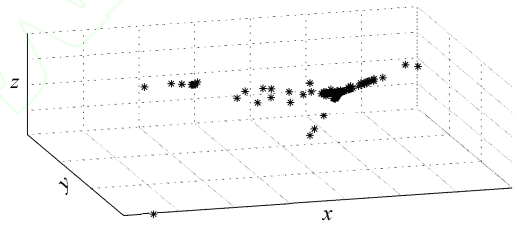

图 2 词根据语法功能在三维空间中的分布

Fig. 2 Word distribution in 3-D space according to grammar function

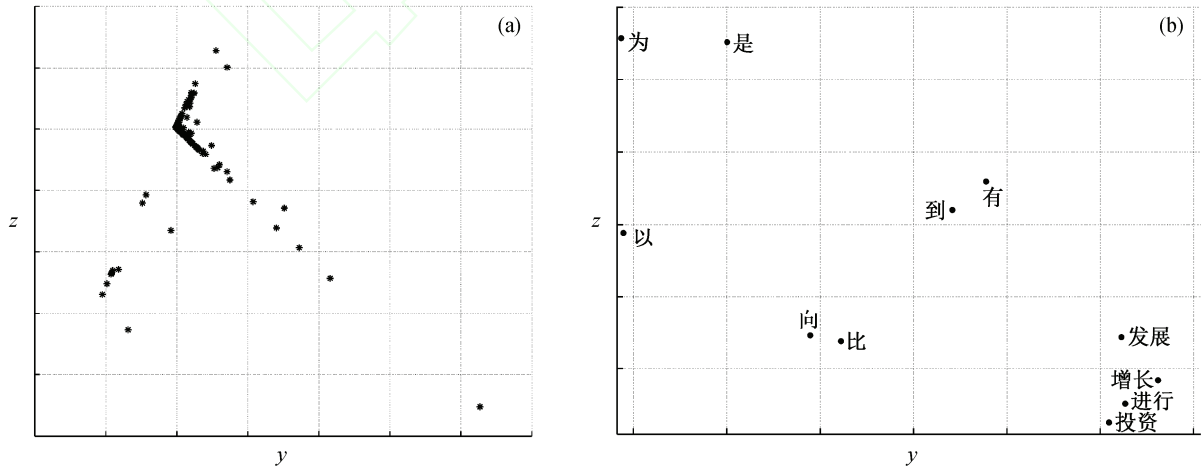

图 3 在二维空间上的投影以及局部放大的词语法功能分布

Fig. 3 Word distribution in 2-D space and its detailed local magnification

2.2 基于流形距离的语块聚类算法

词的低维流形嵌入进一步揭示了词之间的关系。这种关系不以词性这种硬分类准则为基础, 模糊化兼类词在词性上的差别, 着重体现词在构成语块时的功能, 加入了“修饰”、“依存”、“附着”等更高层次的语法概念, 解决了统计学习方法中特征数量庞大、提取困难的问题, 同时较好地保留了词的差异性。因此, 使用流形距离可以作为距离度量。

令 D_{ji} 为点 x_i, x_j 之间路径的集合, 则 $\mathrm{d}(x_i, x_j) = \min\{D_{ij}\}$ 为两点间的距离测度。显然满足距离测度的 4 个条件: 对称性、非负性、三角不等式和自反性。两个词之间的距离越小, 它们之间的关联度就越高。我们将这种关联度定义为

$$r_{ij} = 1/\mathrm{d}(x_i, x_j) + \delta, \tag{1}$$

δ为调节因子。

使用聚类方法进行语块分析的目标是使簇内距离越小越好, 而簇间距离越大越好。为了衡量聚类效果, 需要衡量类内距离和类间距离差异的度量准则。当词聚类到同一个簇中时, 该簇与另一个簇之间的相似度使用两个簇之间的距离之和表示:

$$S_{pq} = \sum_{w_i \in C_p} \sum_{w_j \in C_q} r_{ij},$$

归一化为

$$S_{pq} = \sum_{w_i \in W_p} \sum_{w_j \in W_q} r_{ij} \Big/ \sum_{w_i \in W} \sum_{w_j \in W} r_{ij},$$

W 表示全体实例的集合, W_p, W_q 表示划分的词簇。

我们采用文献[16]提出的聚类评价准则:

$$J = \sum_{p=1}^{m} \left[S_{pp} - \left(\sum_{q=1}^{m} S_{pq} \right)^2 \right]。 \tag{2}$$

该方法是计算在聚类合并时, 根据新的数据簇重新计算 J', 当 ΔJ 的变化超过一定阈值, 则表明簇合并是有效的, 且是最优的。

常用的聚类算法中, 基于“质心”的方法, 如 K 均值方法等, 假设数据分布较为规则, 或者划分的数据子集都符合球形分布[12]。但是在语块分析任务中, 由于数据分布并不规则, 使用基于“质心”的方法容易产生聚类稀疏问题, 对聚类效果造成影响。针对语块聚类任务中, 实例(也就是词) 数量规模小、距离分布稀疏的特点, 我们采用一种凝聚层次聚类方法。层次聚类方法可以发现任意形状的簇, 且具有较好的处理噪声的能力, 在处理小到中等规模的数据时, 其复杂度也是可以接受的。使用层次聚类方法分析语块的迭代过程如下。

算法 1 语块聚类算法。

1) 将每个词视为一类, 计算初始合并时准则的变化量。

2) 找出变化量中的最大值 $\max(S_{fg})$, 如果大于 0, 转到 4。

3) 合并导致变化量最大的两个簇, 重新计算所有准则变化量, 并执行 2。

4) 算法停止并输出聚类结果。

与序列化标注这种语块划分与类别标注一体化方法不同, 采用聚类方法分析语块, 首先用聚类算法将词划分为不同的簇。如“驻京外资金融机构有意为河南省的经济建设和中国中西部地区的开发做出贡献”, 分为如下几个簇: {京}{外资 金融 机构 经济}{建设 开发 贡献}{河南省 中国}{中西部 地区}{驻 有意 做出}{和 为 的}。

词簇的划分结果与词性类别有一定的联系, 如“外资”、“机构”等, 都是一般名词, 而“中国”、“河南省”是实体名词。而在词的使用上, “和”、“为”、“的”词性不同, 却由于它们距离其他词在功能空间上较远, 划分为一个簇。这样的划分结果并不是说明“和”、“为”、“的”在空间中的绝对距离更近, 而是表明这类词与其他词的距离都较远。因此, 使用聚类算法, 一方面体现了词之间的共性, 同时还可以体现出词的差异性。

将簇内的词按照句子中的顺序, 形成语块。如上例中, 将簇中词按照句子顺序排列, 形成词串: {驻}{京}{外资 金融 机构}{有意}{为}{河南省}{的}{经济}{建设}{和}{中国}{中西部 地区}{的}{开发}{做出}{贡献}。其中, {外资 金融 机构}构成一个子簇。对于独立的簇, 很自然地, 单独构成语块。而对于子簇, 需要明确其边界, 解决语块的嵌套问题, 如示例中的“外资金融机构”, “金融机构”本身可以作为独立的语块。本方法利用最长组合原则, 将子簇合并为一个语块, 来体现语块的无嵌套性特点。最终划分成的语块如下: [驻 京 外资金融机构 有意 为 河南省 的 经济 建设 和 中国 中西部地区 的 开发 做出 贡献]。

2.3 语块类别标记

在确定了语块的边界之后, 为了为划分出的语块标记类别, 我们采用统计与规则相结合的方法。

1) 如果语块由一个词构成, 如[有意 VV]、[贡献 NN], 将由独立词构成的语块根据词性规则进行标记, 如[有意 VV]标记为[VP], [贡献 NN] 标记为

[NP]。

2) 对于多个词组成的语块, 根据〈词性–语块〉统计值确定其类别。〈词性–语块〉统计值表明了不同词性下, 语块类别标签的概率。我们选择概率高的值作为标签类别。其含义是, 在某类语块中出现频率最高的词性在统计意义上决定了语块的类别。

3 实验结果与分析

3.1 数据集及评价指标

由于语块识别方面, 目前缺乏标准的数据集, 我们使用基于宾州中文树库 CTB5.0 作为基本语料库, 使用 SVM 与规则相结合的方法, 从中抽取语块。所使用的基本数据见表 1。测试的结果采取了常用的 3 个评测指标, 即准确率(P)、召回率(R)和综合指标 F 值来评测语块识别的结果[3]。

因为由于词的语法功能分布非常稀疏, 而 ISOMAP 并不是根据词与词、词性与词性之间的共现等需要更多样本才能获得的统计信息来计算距离, 增加或者减少有限的训练集, 对于 ISOMAP 估算测地线距离影响不大。而为了测试方法效果, 我们也将树库划分为训练集和测试集。训练集用来计算词的分布特性, 即计算词间距离, 为语块的聚类分析提供基础。

3.2 聚类结果和分析

使用 ISOMAP 算法估计词之间距离时, 由于 ISOMAP 会产生“短路”现象, 即无法测算两点间距离, 因此在式(1)中加入了平滑因子δ。从表 2 可以看出, 词在语法功能空间中, 彼此距离的分布非常稀疏, 导致计算相似性时, 结果对于δ取值的敏感。我们通过实验分析经验值δ对聚类效果的影响, 方式是检验不同δ取值对语块分析效果的影响。

当δ取 10^{-4}~10^{-5}之间的值时, F 值接近理想的效果; 而取 1~10^{-3} 时, 由于δ值已经大大超过词间距离, 使原本距离较近的词其相似度变化超过一个数量级, 将词间距离平均化, 导致聚类算法无法很好地区别词的特征, 精度大大降低; 而使用过小的δ值, 在发生短路时, 导致距离无穷小的词, 也就是发生短路的词之间相似性过大, 从而导致召回率的下降。因此, 在使用聚类算法进行语块分析时, 我们将δ的值设置为 0.0006, 这一数据更符合词在空间分布中的统计结果。

表 1 宾州中文树库语料库统计

Table 1 Statistics of CTB5.0

语料库	句子数	语块数	语块类别	词
宾州中文树库	10099	156954	12	240770
训练集	9999	155283	12	220356
测试集	1000	1671	12	20414

表 2 δ 取值对总 F 值的影响

Table 2 Impact of δ on F value

δ	F/%
1	41.71
0.1	49.9
0.001	51.04
0.0001	81.18
0.00001	71.45

我们将聚类结果与有监督的SVM方法、文献[6]中使用词聚类特征进行语块分析结果以及文献[7]中的改进 K 均值方法进行对比。这 3 种方法分别属于有监督方法、基于无监督聚类特征的混合方法以及无监督方法。我们重新在宾州中文树库语块库基础上进行了测试, 比较 F 值。文献[7]只识别了 7 种语块, 本文对其中的语块库进行了扩展。文献[6]中, 语块分析过程本质是在 MEMM 方法的特征中加入预处理的通过聚类获得的词特征。本文在试验中, 省略了其中的命名实体识别、仿词识别等步骤, 对宾州树库中的词进行聚类, 剪枝参数设置为 6, 即 64 个类别, 然后使用 MEMM 方法进行语块识别。结果见表 3。

与之前的无监督方法相比, 我们的方法在精度提升上效果并不明显, 但在召回率方面取得了 1.7%~4%的提升, 并使 F 值得到了提升。与 SVM 相比, 聚类方法整体性能还有一定差距, 但是在某些类别语块中, 也取得了不差于统计学习方法的结果, 如表 4 所示。由于流形距离包含词的语法功能信息, 对于特征明显的词具有较高的识别率和召回率, 如连词“和, 而”、介词“以, 由”等。如词串“和计算方法等”, 我们划分的结果为[和 计算方法 等], 而没有划分为[计算方法等]。

在不同类别语块分析结果中, CC, DEC, ETC, LCP, PP 等类别的语块, 分析结果较好, 而 ADJP, IJP, VP 等类别语块的识别能力却低于平均水平。根据语料库中语块的统计结果, 构成 CC, ETC 等语块的词数量少, 词性构成相对单一, 且一般单独构成

表 3 与现有方法的实验对比结果

Table 3 Comparasion results of different methods

方法	P/%	R/%	F/%
SVM	91.91	90.19	91.04
改进的 K 均值算法	89.11	83.06	85.98
层次聚类算法	86.75	87.98	87.36
使用词聚类特征	85.04	86.88	85.95

表 4　不同类别语块识别 *F* 值结果(与 SVM 方法比较)

Table 4　*F* values on various kinds of chunk(compared with SVM)

语块类别	聚类方法	SVM
ADJP	67.92	84.81
ADVP	78.58	83.14
CC	99.84	98.22
DEC	99.11	97.16
DP	99.93	99.70
ETC	94.56	89.11
IJP	67.16	63.68
LCP	97.68	98.7
NP	86.63	91.36
PP	95.23	98.21
QP	91.10	98.33
VP	70.85	90.1
总和	87.36	91.04

语块。这种特征明显的词, 在空间中分布较为离散。宏观上, 词在空间的分布聚集在一个相对小的范围内, 这个范围内的词彼此关系更为密切。而离散分布的各点, 往往表示一种独立的语法功能特点。因此, 这些语块的精度和召回率相对较高。反过来, 构成比较复杂的语块, 由于受到稀疏性和距离估算方法的影响, 在整体上, 落后于 CC, ETC 等语块的识别效果。这些结果表明, 基于聚类的语块识别方法在一定程度上具有可行性, 并且在发现数据内在规律方面也具有一定的效果。

4　结语

本文基于聚类思想处理中文语块分析。为了考察词在句子构成中的特征, 构建词的语法功能空间, 并利用 ISOMAP 的方法获得空间的低维流形嵌入, 通过对二维、三维空间的观察, 可以发现词在低维流形嵌入中的分布情况, 并获得词的流形距离。词的语块分析过程中, 使用词的流形距离作为相似性度量, 在使用基于簇内/簇间距离测度的聚类准则下, 特征明显的中文语块类别, 识别情况令人满意。同时, 也证明利用聚类方法分析语块的可行性, 并为今后的工作提供改进优化的基础。目前, 在名词语块中, 由于未登陆词、命名实体等语法特征不明显、难以提高聚类分析效果, 本文采用了平滑因子来解决该问题。但是平滑因子值采用经验值, 不能完全反映词的分布特性, 因此设置合适的平滑因子成为今后研究的一个重要任务。同时, 在面对大规模数据处理时, ISOMAP 算法还需要进一步优化。目前的算法仅考虑了聚类的复杂度, 而由于 ISOMAP 在利用图上最短距离估计流形距离时复杂度还比较高, 并且出现短路的情况。因此, 如何更好地重构词的语法空间的低维流形嵌入、更直观地表达词的功能特点也是下一步努力的方向。

参考文献

[1] 宗成庆. 统计自然语言处理. 北京: 清华大学出版社, 2008

[2] Abney S. Part of speech tagging and partial parsing // Church K, Young S, Bloothooft G. Proceedings of the Corpus-Based Mathods in Language and Speech, An ELSNET Volume. Dordrecht: Kluwer Academic Publishers, 1996: 119–136

[3] Chen W, Zhang Y, Isahara H. An empirical study of Chinese chunking // Proceedings of the COLING/ACL on Main Conference Poster Sessions. Sydney: Association for Computational Linguistics, 2006: 97–104

[4] 周俊生, 戴新宇, 陈家骏, 等. 基于大间隔方法的汉语组块分析. 软件学报, 2009, 20(4): 870–877

[5] Zhang L, Lü X Q, Shen Y A, et al. A statistical approach to extract Chinese chunking candidates from large corpora // Proceeding of the 20th International Conference on the Computer Processing of Oriental Languages (ICCPOL 2003). Shenyang, 2003: 109–117

[6] 孙广路, 王晓龙, 刘秉权, 等. 基于词聚类特征的统计中文组块分析. 电子学报, 2008, 36(12): 2450–2454

[7] 梁颖红, 赵铁军, 于浩, 等. 基于改进 K-均值聚类的汉语语块识别. 哈尔滨工业大学学报, 2007, 39(7): 1106–1109

[8] 杨震, 范科峰, 雷建军, 等. 基于语义的文本流形研究. 电子学报, 2009, 37(3): 557–561

[9] 王自强, 钱旭. 基于流形学习和 SVM 的 WEB 文档分类算法. 计算机工程, 2009, 35(15): 38–40

[10] Xue N, Xia F, Huang S Z. The bracketing guidelines for the Penn Chinese Treebank [M/OL]. (2000) [2012–08–30]. http://www.cis.upenn.edu/~chinese/parseguide.3rd.ch. pdf

[11] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究. 软件学报, 2008, 19(1): 48–61

[12] 公茂果, 王爽, 马萌, 等. 复杂分布数据的二阶段聚类算法. 软件学报, 2011, 22(11): 2760–2772

[13] 冯志伟. 特思尼耶尔的从属关系语法. 国外语言学, 1983(1): 57: 63–65

[14] 邹宏梅. 组块识别技术的研究与实现[D]. 长沙: 国防科技大学, 2006

[15] Che W, Li Z, Liu T. LTP: a Chinese language technology platform // Proceeding of the 23rd International Conference on Computational Linguistics: Demonstration Volume. Beijing, 2010: 13–16

[16] 王娜, 杜海峰, 王孙安. 一种基于流形距离的迭代优化算法. 西安交通大学学报, 2009, 43(5): 76–79