基于词聚类特征的统计中文组块分析模型

孙广路[1,2], 王晓龙[1], 刘秉权[1], 关 毅[1]
(1. 哈尔滨工业大学 计算机科学与技术学院, 黑龙江哈尔滨 150001; 2. 哈尔滨理工大学 计算机科学与技术学院, 黑龙江哈尔滨 150080)

摘 要: 提出了一种基于信息熵的层次词聚类算法, 并将该算法产生的词簇作为特征应用到中文组块分析模型中. 词聚类算法基于信息熵的理论, 利用中文组块语料库中的词及其组块标记作为基本信息, 采用二元层次聚类的方法形成具有一定句法功能的词簇. 在聚类过程中, 设计了优化算法节省聚类时间. 用词簇特征代替传统的词性特征应用到组块分析模型中, 并引入名实体和仿词识别模块, 在此基础上构建了基于最大熵马尔科夫模型的中文组块分析系统. 实验表明, 本文的算法提升了聚类效率, 产生的词簇特征有效地改进了中文组块分析系统的性能.

关键词: 词聚类; 信息熵; 中文组块分析; 句法功能
中图分类号: TP391.2 文献标识码: A 文章编号: 0372-2112 (2008) 12-2450-04

Statistical Chinese Chunking Model Based on Word Clustering Features

SUN Guang-lu[1,2], WANG Xiao-long[1], LIU Bing-quan[1], GUAN Yi[1]
(1. *School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*;
2. *School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, Heilongjiang* 150080, *China*)

Abstract: An entropy-based hierarchical word clustering algorithm is proposed. Word clusters generated by the clustering algorithm were used as features in Chinese chunking model. Based on words' chunk tags and the theory of entropy, a binary hierarchical clustering algorithm was applied to the words in Chinese chunking corpus. An accelerating algorithm was employed to save the clustering time. With the recognition of name entity and factoid, the new Chinese chunking system was constructed based on maximum entropy Markov models, while part-of-speech features were replaced with the entropy-based word clustering features. Experimental results show that the algorithm increases the efficiency of the word clustering, and the entropy-based word clustering features improve the performance of Chinese chunking effectively.

Key words: word clustering; information entropy; Chinese chunking; syntactic function

1 引言

组块分析(chunking), 也称作部分句法分析(partial parsing)或浅层句法分析(shallow parsing), 在利用句子的词法信息(包括分词标记和词性标记)基础上, 提供句法级的标记.

现有的中文组块分析研究一般将分词标记和词性标记作为分析的基本特征, 利用不同的机器学习模型(如: 隐马尔科夫模型(HMM), 最大熵模型(MEM)[1]和条件随机域模型(CRF)[2])来识别句子中组块信息. 我们对于现有方法进行研究, 发现了以下两方面的问题:

(1) 现有方法是在金本位* (gold-standard)的分词和词性的基础上进行, 组块分析的输入数据不包含任何错误信息, 而在网络上的自然语句经过由机器自动分词和词性标注后, 会产生一定比率的错误分词和词性标记. 而这些错误标记会及联地影响到组块分析的准确率, 导致将近 10 个百分点的下降[3].

(2) 现有方法的研究重点放在了如何应用更合适的机器学习模型来解决中文组块分析问题上, 而忽视了引入新的对于组块分析更具预测能力的特征.

基于上述两点, 发现金本位的词性对于组块分析有很大的帮助, 但是带有一定错误标记的词性大大削弱了这种帮助. 于是, 我们试图找到一种新的特征, 其对于组块分析的预测能力优于机器自动标记的词性特征. 根据齐普夫定律(Zipf's law), 数据稀疏问题普遍存在于语言处理当中, 而不会随着语料库规模的扩大而消失. 对于大量的稀疏词, 词性具有给予每一个词一个词法类的平滑功能, 使得机器可以根据它们的词法类来预测它们可能属于的组块. 但从本质上讲, 词性与组块标记并没有直接的联系. 根据奥卡姆剃刀原理(Occam's Razor), 我

收稿日期: 2007-06-04; 修回日期: 2007-12-20
基金项目: 国家自然科学基金(No. 60435020; No. 60673037); 国家 863 项目(No. 2006AA01Z197; No. 2007AA01Z172)
* 金本位是指由人工标记的被认为没有标记错误的语料.

们可以寻找一种直接映射到组块标记上的词类, 使其对组块标记具有更强的预测能力.

在上面的研究分析基础上, 本文首先提出了基于信息熵理论的层次词聚类算法以获得具有句法功能的词簇, 该算法利用组块语料中的词及其组块标记作为基本信息, 计算它们的条件熵作为量度, 采用信息增益的方法确定收敛准则, 使用了二元层次聚类的方法对训练语料库中的词进行聚类, 设计了优化算法以提高聚类效率; 其次, 说明了中文组块的定义和类别, 在不影响验证特征有效性的条件下, 采用了训练时间相对较短的最大熵马尔科夫(MEMM)模型进行中文组块分析的训练; 最后在对命名实体(name entity)和仿词(factoid word)的识别基础上, 融入词簇特征, 建立基于 MEMM 的中文组块分析系统, 获得了更好的组块分析性能.

2　基于信息熵的词聚类算法

聚类算法在机器学习范畴中属于无指导学习方法, 我们采用了层次聚类的方法, 这种方法对给定的数据集合进行层次划分.

2.1　词聚类算法的基本粒度, 量度和收敛准则

在确定了基本聚类方法后, 我们将研究重点放在了聚类的基本粒度, 量度和收敛准则上. 布朗利用语言模型的迷惑度作为聚类的量度, 用迷惑度降至最低作为聚类收敛的准则[4]. 但是, 无论是互信息还是迷惑度都与组块分析任务不相关[5], 根据它们产生的词簇不适合组块分析任务. 我们试图找到一种聚类量度, 在它基础上生成的词簇能够适合组块分析任务.

设 $W=w_1, w_2, \cdots, w_k$ 代表词序列, $T=t_1, t_2, \cdots, t_k$ 代表对应的组块标记序列, $\{t_i\}$ 代表所有在训练语料中出现过的组块标记. 由此, 我们得到了具有映射关系的词与组块标记的二元对 (w_i, t_i). 我们对这些二元对进行统计, 生成了一种新的聚类量度:

$$-\sum_j\sum_i \mathrm{Count}(t_j, c_i)\log P(t_j/c_i) \qquad (1)$$

其中, c_i 代表一个词簇, $\{c_i\}$ 代表聚类生成的所有词簇, $\mathrm{Count}(t_j, c_i)$ 代表二元对 (t_j, c_i) 的同现次数, $P(t_j/c_i)$ 代表 t_j 在 c_i 下的条件概率. 词的每层聚类的收敛准则就是最小化公式(1). 而 $P(t_j/c_i)$ 的条件熵公式为:

$$-\sum_j\sum_i \frac{\mathrm{Count}(t_j, c_i)}{N}\log P(t_j/c_i) \qquad (2)$$

N 代表二元对的总数. 由于 N 为常数, 最小化公式(1)就等于最小化 $P(t_j/c_i)$ 的条件熵. 下文的熵值公式和计算都忽略了常数 N.

从信息熵的观点来解释, 这种收敛准则使得组块标记集 $\{t_i\}$ 在聚类的结果词簇 $\{c_i\}$ 下的条件熵最小, 也就是使得 $\{c_i\}$ 映射到 $\{t_i\}$ 的迷惑度最小. 由此, 通过公式(1)的量度进行词聚类, 我们能够得到与组块分析任务直接相关的词簇.

2.2　二元层次聚类优化算法

层次聚类具体又可以分为"自顶向下"和"自底向上"两种方法. 因为在聚类过程中前者花费较少的计算时间[6], 所以我们采用自顶向下的层次聚类方法来进行词聚类.

聚类算法:

(1)将待聚类的词任意均分为两个词簇 A、B.

(2)根据公式(1), 分别计算两个词簇 A、B 对应到组块标记的条件熵 $E(A)$, $E(B)$, 相加计算总的熵值 $E^*=E(A)+E(B)$.

(3) $\forall w\in A$, 将 w 移动到 B 中, 重新计算 $E(A)$, $E(B)$, 及 $E'=E(A)+E(B)$. 如果 $E'<E^*$, 那么 w 在 B 中符合收敛准则, 不进行其他操作; 如果 $E'>E^*$, 那么 w 在 B 中是不符合收敛准则, 则将 w 移动回 A 中.

(4) $\forall w\in B$, w 是否移动到 A 中的方法类似(3).

(5)重复执行步骤(3)、(4), 直到没有任何一次移动 w, 使得总的熵值降低. 从而得到两个固定的词簇 A^*、B^*.

(6)对于词簇 A^*、B^*, 分别执行步骤(1)~(5), 形成各自的下一层的两个子词簇.

在上述的聚类算法中, 步骤(3)、(4)的计算是非常耗时的. 为了降低计算量, 减少每一层聚类的收敛时间, 我们提出了一种优化算法. 因为一个词对应的平均组块标记数远小于组块标记总数, 在每一次移动 w 后重新计算熵值时, 我们可以利用前一次的熵值以及词和组块标记的同现次数来计算新的熵值, 以节省计算时间. 当 w 从 A 移动到 B 中时, 我们用 B' 表示 $B+w$, 用 A' 表示 $A-w$. 新的熵值 $E(B')$, 可以根据 w 是否对应 t_j 被分解为:

$$\begin{aligned}E(B') &= -\sum_j\sum_i \mathrm{Count}(t_j, B')\log P(t_j/B')\\ &= -\sum_{j/\mathrm{Count}(t_j, w)\neq 0} \mathrm{Count}(t_j, B')\log P(t_j/B')\\ &\quad -\sum_{j/\mathrm{Count}(t_j, w)=0} \mathrm{Count}(t_j, B')\log P(t_j/B') \qquad (3)\end{aligned}$$

公式(3)右边的第一项可以直接进行计算获得. 第二项可以被分解为:

$$\begin{aligned}&\sum_{j/\mathrm{Count}(t_j, w)=0} \mathrm{Count}(t_j, B')\log P(t_j/B')\\ &= \sum_j\sum_i \mathrm{Count}(t_j, B)\log P(t_j/B)\\ &\quad -\sum_{j/\mathrm{Count}(t_j, w)\neq 0} \mathrm{Count}(t_j, B)\log[P(t_j/B')]\\ &\quad +\left[\mathrm{Count}(B)-\sum_{j/\mathrm{Count}(t_j, w)=0}\mathrm{Count}(t_j, B)\right]\\ &\quad \times\log(\mathrm{count}(B)/\mathrm{count}(B')) \qquad (4)\end{aligned}$$

公式(4)的右边第一项就是前一次的熵值. 其它各项可以直接求出. 同理, 新的熵值 $E(A')$ 可以通过推导求得. 通过上述优化算法, 可以在很大程度上节省新的熵值计算时间. 我们采用的这种层次聚类的方法最终可以将每个词都分成一类, 此时层次聚类停止. 但在实验中, 我们只聚类到所有词簇中的词数都小于 50 就停止, 并通过剪枝的方法得到相应层数的词簇.

3 中文组块分析系统

参考 CoNLL-2000 的英文组块定义形式, 我们采用了微软亚洲研究院(MSRA)的中文组块分析定义形式[7]. MSRA 组块定义包含 11 种组块类型.

3.1 中文组块分析模型

MEMM 模型是由 McCallum 在 2000 年提出的有指导的机器学习模型[8], MEMM 模型的公式如下:

$$P(t_i \mid t_{i-1}, O) = P(t_i \mid t_{i-1}) \frac{\exp\left(\sum_j \lambda_j f_j(t_i, h)\right)}{\sum_t \exp\left(\sum_j \lambda_j f_j(t_i, h)\right)} \quad (5)$$

h 和 O 代表当前标记 t_i 的上下文. 我们用 Viterbi 算法来求得最优的标记序列.

我们将中文组块分析作为序列化分析和标记的任务来进行处理. 通过 MEMM 模型的处理, 词序列被划分成若干个组块, 并被标记了组块标记.

3.2 词簇在模型中的应用

我们对词序列 W 中的每一个词 w_i 标记一个 c_i, c_i 是聚类结果 $\{c_i\}$ 中的一个词簇且 c_i 中包含词 w_i. 在对 W 进行标记之后, 从新的词序列 W 和对应的标记序列 $C = c_1, c_2, \cdots, c_k$ 中抽取 MEMM 模型需要的特征. 我们选取宽度为 5 的窗口, 抽取当前词 w_i 和前后各两个词的词特征、词簇标记特征. 另外, 我们还抽取了 w_i 的词缀特征. 我们将这些特征称原子特征, 而将它们的组合称为复合特征. 由于低频的特征对模型有不良的干扰, 次数高于 3 的特征被保留.

3.3 中文组块分析系统的构成

图 1 描述了中文组块分析的训练过程. 由于词簇是通过对训练语料库中的词进行聚类得到的, 在处理网络文本时, 会有一部分词不出现在词簇中. 我们对于其中的名实体和仿词进行处理. 在已建立的人名地名词典的基础上, 我们采用基于 ME 的名实体标注器对其进行识别. 此外, 我们采用人工书写规则结合有限自动机(FSA)的方法进行识别.

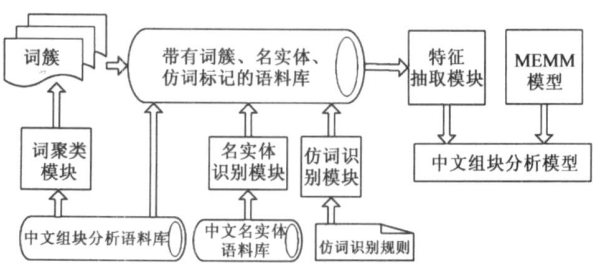


图1 中文组块分析系统训练流程图

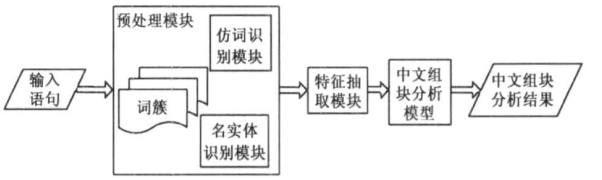


图2 中文组块分析系统标注流程图

图 2 描述了利用训练好的各个模块进行处理和标注的过程.

4 实验结果和分析

4.1 MSRA 中文组块分析语料库

MSRA 中文组块分析语料库是在北京大学开发的"98 年 1 月人民日报分词和词性标注公开语料库"的基础上, 人工进行组块标记而成的. 表 1 列举了语料库的几类统计结果.

表 1 MSRA 中文组块分析语料库统计

语料库	句子数量	组块数量	词数
训练集	17, 253	229, 989	444, 777
开发集	1000	13, 992	28, 645
测试集	986	13, 879	28, 382

表 2 词聚类的部分结果

类别号	词典词
1	三百零五十万 许久 许许多多 一半 一辈子 一笔笔 一大半 一丛丛 一幢幢 一百亿 …
2	必 不 不大 不断 不仅仅 不禁 不要 不再 常常 彻底 沉 成 持续 从不 从中 大胆 大都 毫不 呼 互相 集中 绝不 可惜 可想而知立刻 连连 略 满意 盲目 没 没有 密切 明白 莫 难以 排 频频 颇 其乐融融 奇怪 岂 认真 仍 仍然 稍 少不了 深深为什么 未 未必相继 也 一度 一律 依法 引人注目 永远 勇敢 再度 遭 早日 早已 怎 直 重新 逐步 最初 …
3	1965 年 4 月 本世纪 本周 朝 初三 春节 冬日 二月 逢年过节 会上 今 近期 明清 年中 上次 眼前 一九九三年 一日战后 周末 …
4	按 比 除 根据 经 通过 同 往 于 在 自 …
5	久 长期 沉着 成功 持久 赤 充实 抽象 出色 匆匆 匆忙 粗 错综复杂 大规模 大致 单纯 独资 方便 非常 非同寻常 分明 分散 …

4.2 词聚类的实验结果和分析

利用开发集对于聚类算法进行参数调优, 发现产生 56 类词簇的聚类结果在组块分析实验中得到了最优的性能. 下面以它为例进行分析, 对于训练语料库中的词典词进行自动聚类, 表 2 列出了聚类算法的部分结果. 从结果中可见, 类别 1 中的词体现出数量组块的特性; 2 中体现出副词性组块的特征; 3 中体现出时间组块的特性; 4 中体现出介词性组块的特性; 5 中体现出形容词性组块的特性. 一些兼类词比较难以划分, 但聚类算法可

以将具有相同兼类的词聚为一类. 表 3 列出了采用本文提出的优化聚类算法和不采用优化算法的聚类算法各自所需的聚类时间, 从中可以看出, 优化算法提升了聚类的效率.

表 3　词聚类算法的聚类时间

聚类算法 (产生 56 类词簇)	采用优化算法的聚类算法	未采用优化算法的聚类算法
聚类时间(小时)	8	20

4.3　基于词簇的中文组块分析实验结果

我们进行词聚类的目的是为了给中文组块分析提供更好的预测能力和平滑能力, 词聚类算法也是基于组块分析语料库中的词和组块标记的映射关系的. 因此, 我们通过比较聚类产生的词簇和自动标注的词性对于中文组块分析性能的影响来评价聚类结果的优劣. 我们采用通用的性能指标: 精确率(P)、召回率(R)、和它们的调和平均值(F) 来评价组块分析的性能. 所有的实验都是在开放测试的条件下进行的.

我们对聚类结果进行了不同聚类层的剪枝处理. 例如: 如果我们取剪枝参数为 $n=6$, 那么最多类别数为 2^6, 即 64 个类别. 由于每一层聚类的结果是不均衡的, 所以剪枝后的簇数可能不会达到 2^n. 在实验中, 我们选择了 $n=5\sim10$ 作为剪枝参数. 我们采用 3.2 节和 3.3 节的方法将聚类产生的簇应用到模型中, 并利用开发集和不同的簇得到最优的分析层数为 6. 我们在测试集中进行测试, 表 4 列出了基于不同层数聚类结果的中文组块分析性能.

表 4　不同层数的聚类结果对组块分析性能的影响

聚类的剪枝层数	5	6	7	8	9	10
产生的簇数	31	56	99	165	267	409
中文组块分析结果 F(%)	81.93	82.69	82.59	82.57	82.53	82.41

表 5　基于不同类型特征的组块分析性能比较

模型	P(%)	R(%)	F(%)
1. 基准模型 1	74.91	75.37	75.14
2. 基准模型 2	79.41	81.53	80.46
3. 基准模型 3	82.01	81.07	81.50
4. 基于布朗词簇特征的模型	76.22	76.20	76.21
5. 基于本文词簇特征的模型	83.17	82.22	82.69
6. 基于金本位词性特征的模型	91.36	90.68	91.02

表 5 列出了基于不同类型特征的组块分析性能的比较. 模型 1 仅仅使用了词特征, 而没有使用词类的特征(如: 词性或者词簇); 模型 2 是参考文献[3]中的基于 HMM 的组块分析模型, 使用了词特征和自动标注的词性特征, 词性标注的准确率是 93.5%; 模型 3 使用了词特征和同样标注准确率的词性特征; 模型 4 使用了词特征和基于布朗方法产生的词簇特征; 模型 5 使用了词特征和基于本文方法产生的词簇特征, 其中聚类剪枝层数为 6; 模型 6 使用了词特征和金本位词性特征. 另外, 模型 1, 3, 4, 5, 6 都是基于 MEMM 的模型. 从表 5 中可以看到应用词簇特征的模型 5 的性能优于应用自动词性的模型 2 和模型 3, 分别提升了 3.23% 和 2.19%; 也优于布朗模型产生的词簇结果, 提升了 7.48%; 更优于仅用了词特征的模型 1, 提升了 8.55%. 由于聚类算法是在系统训练时进行的, 聚类所需的时间不会影响系统在测试过程中的分析实时性. 而且系统不需要进行词性标注, 这样也节省了标注需要的时间. 实验结果说明:

(1) 由聚类算法产生的含有句法信息的词簇特征对于中文组块分析问题有很大的帮助.

(2) 不含有句法信息的词簇对中文组块分析的性能帮助不大.

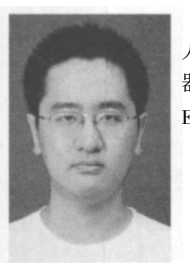
(3) 相对于自动标记的词性特征, 词簇特征更大程度上改善了中文组块分析的性能; 相对于金本位的词性特征, 词簇特征还有很大的改进和提升空间.

5　结论与未来展望

本文以中文组块分析任务为背景, 提出了基于信息熵的词聚类算法, 并将聚类产生的词簇作为特征应用到中文组块分析系统中, 实验表明: 基于聚类算法产生的词簇特征比自动标记的词性特征更好地提升了中文组块分析的性能.

在今后的工作中, 笔者拟在两个方向上继续探索: 通过对于聚类算法的改进, 构建更好的词簇, 并将其应用在其他中文信息处理的任务中; 结合词簇特征与词性特征, 进一步提升中文组块分析的性能.

作者简介:



孙广路　男, 1979 年生, 黑龙江省哈尔滨市人, 博士, 副教授, 主要研究方向为组块分析、机器学习算法等.

E-mail: guanglu.sun@gmail.com;
hati_sun@hit.edu.cn

参考文献:

[1] 徐延勇, 等. 基于最大熵模型的汉语句子分析[J]. 电子学报, 2003, 31(11): 1608-1612.
Xu Y, et al. Chinese sentence parsing based on maximum entropy model[J]. Chinese Journal of Electronics, 2003, 31(11): 1608-1612. (in Chinese)

[2] Chen W, et al. An empirical study of chinese chunking[A]. In Proceedings of the 44th Annual Meeting of ACL[C]. Sydney: ACL Press, 2006. 97-104.

(下转第 2399 页)

作者简介:

戴喜增　男, 1978 年生于辽宁调兵山, 2004 年于电信科学技术研究院信息与通信工程专业毕业, 获工学硕士学位, 现于清华大学电子工程系攻读博士学位. 主要从事 MIMO 雷达、雷达波形设计和 ISAR 信号处理技术方面的研究, 已发表论文数篇.
E-mail: dxz04@mails.tsinghua.edu.cn.

许　稼　男, 1974 年生于, 清华大学电子工程系博士后, 副教授. 研究领域包括雷达及水声领域的目标检测和识别、参数估计、仿真模拟、合成孔径/逆合成孔径成像、混沌非线性理论等. 目前已经在 IEEE Trans. on GRS, IEEE Trans. on AES, IET Proceeding of RSN, 中国科学等国内外刊物及各类学术会议上发表和录用论文六十余篇, 其中被 SCI、EI、ISTP 等检索四十余篇.

彭应宁　男, 1939 年生, 清华大学电子工程系教授, 博导. 原清华大学电子工程系高速信号处理和网络传输研究所所长. 长期从事雷达信号处理领域的研究, 主要研究方向包括检测与估计理论, SAR/ISAR 成像技术, 阵列信号处理, 自适应信号处理. 已发表学术论文近 200 篇, 其中被 SCI、EI 和 ISTP 收录的论文 120 多篇, 出版专著 4 部, 并获十多项国家级和部委级科技进步奖.

王永良　男, 1965 年生于浙江嘉兴, 教授, 博导. 已发表学术论文 200 多篇, 其中被 SCI、EI 和 ISTP 收录的论文 90 多篇, 出版专著 2 部.

(上接第 2453 页)

[3] 李宏乔. 汉语组块分析技术及应用研究[D]. 北京: 北京理工大学计算机系, 2004.
Li H. Research of Chinese Chunking and its Applications[D]. Beijing: Beijing Institute of Technology, Computer Science and Technology, 2004. (in Chinese)

[4] Dagan I, et al. Similarity-based methods for word sense disambiguation[A]. In Proceedings of the 35th Annual Meeting of ACL[C]. Madrid: ACL Press, 1997. 56-63.

[5] Brown P, et al. Class-based n-gram models of natural language [J]. Computational Linguistics, 1992, 16(2): 79-85.

[6] Gao J, et al. The use of clustering techniques for language modeling application to asian languages[J]. International Journal of Computational Linguistics and Chinese Language Processing, 2001, 6(1): 27-60.

[7] Sun G, et al. Chinese chunking based on maximum entropy markov models[J]. International Journal of Computational Linguistics and Chinese Language Processing, 2006, 11(2): 115-136.

[8] McCallum A, et al. Maximum entropy markov models for information extraction and segmentation[A]. In Proceedings of ICML' 2000[C]. Stanford: MIT Press, 2000, 591-598.