中文单词聚类的比较研究

王波[1]，王厚峰[2]

（北京大学，计算语言学研究所，100871）

摘　要： 本文对无监督的中文单词聚类方法进行了比较研究，包括最大化平均互信息(MMI)，基于功能词的聚类(FW)，基于高频词的聚类(HFW)以及基于词类的聚类(WC)。采用词性标注准确度以及语义准确度对聚类质量进行评估，最终 MMI 聚类效果最好，其词性标注准确度为 79.1%，语义准确度可达 49%，其它三种方法的词性标注准确度都超过了 50%，语义相似度超出 30%。将上述四种方法生成的词类引入无监督的基于对齐的汉语句法结构自动推导后，准确率、召回率以及 F 值都提高了 1% 左右。

关键字： 单词聚类，句法分析，自然语言处理，无监督学习

A Comparative Study on Chinese Word Clustering

WANG Bo, WANG Hou-Feng

(Institute of Computational Linguistics, Peking University, Beijing, 100871, China)

Abstract: This paper evaluates four unsupervised Chinese word clustering methods. They are maximum mutual information (MMI), function word (FW), high frequent word (HFW), and word cluster (WC) Two evaluation measures, such as part-of-speech (POS) precision and semantic precision, are employed. Testing results show that MMI reaches the best performance: 79.1% on POS precision and 49% on semantic precision, while the other three exceed 50% and 30% respectively. Word clusters generated by the methods mentioned above are introduced into alignment-based Chinese syntactic induction, improve the performance.

Keywords: word clustering, syntactic parsing, natural language processing, unsupervised learning

1　引言

数据稀疏是构建语言模型面临的主要挑战之一，许多低概率事件很少在样本中发生。词类在改善语言模型质量以及提高句法分析的准确性上都可发挥重要作用，本文主要关注采用无监督的方法对中文单词进行聚类，每个词类应尽可能在语法和语义上相似。

使用分布信息识别句法类英语中也有很多工作，Finch 和 Chater [11][12] 以及 Schutze[14][15]通过单词周围共有单词的同现信息抽取一系列特征，然后通过标准的聚类技术以及信息提取技术进行词聚类，对高频词他们获得了较好的结果；Brown et al [1]在大规模数据上采用了一种信息论模型，推导出许多合理的语义类以及句法类。

本文对四种中文单词聚类技术进行详细比较。第一种方法是最大化平均互信息(MMI)，它以 Brown 提出的 n 元词类模型作为基础，在此我们提出一种有效的实现算法。其二是基于功能词的聚类(FW)，通过所有功能词相对

基金资助：国家自然科学基金(No. 60473138)

作者简介：王波(1982-)，男，贵州毕节人，硕士研究生，主要研究领域为自然语言处理；E-mail: wangbo@pku.edu.cn

王厚峰(1965-)，男，博士，教授，主要研究领域为自然语言处理；

内容词的分布刻画内容词的上下文分布。其三是基于高频词的聚类(HFW)，频度最高的 1000 个单词被用来描述单词局部上下文的二元概率分布。第四是基于词类的聚类(WC)，采用单词左右二维类别的概率分布作为其上下文信息。

本文结构如下：第 2 节详细描述四种聚类算法，第 3 节给出聚类结果的详细比较，并给出词类用于基于对齐的无监督句法分析后句法分析效果的提高。最后给出结论。

2 中文单词聚类方法

2.1 最大化平均互信息 (MMI)

交叉熵和困惑度是语言模型的评价标准之一，主要衡量语言模型的不确定性。因此我们需寻找一种划分函数 π，它把单词 w_i 映射到相应类别 c_i，从而降低 n 元语言模型的困惑度。对于二元模型，训练语料为 $L = w_1 w_2 \mathrm{L}\ w_n$，其交叉熵可按(1)进行计算：

$$H(L,\pi) = -\frac{1}{N}\log p(w_1 w_2 \mathrm{L}\ w_N) \approx -\frac{1}{N-1}\sum_{w_1 w_2} C(w_1 w_2)\log p(w_2|w_1). \quad (1)$$

假定当前聚类中的单词的出现仅依赖于前面一个词的类别，(1)可变形为(2)：

$$H(L,\pi) = -\frac{1}{N-1}\sum_{w_1 w_2} C(w_1 w_2)\log(p(c_2|c_1)p(w_2|c_2)) \quad (2)$$

(2)可化简为(3)：

$$\begin{aligned} H(L,\pi) &\approx -\left[\sum_{w_1 w_2}\frac{C(w_1 w_2)}{N-1}\left[\log p(w_2|c_2)+\log p(c_2)\right]+\sum_{w_1 w_2}\frac{C(w_1 w_2)}{N-1}\left[\log p(c_2|c_1)-\log p(c_2)\right]\right] \quad (3)\\ &= -\left[\sum_{w_2}\frac{\sum_{w_1} C(w_1 w_2)}{N-1}\log p(w_2|c_2)p(c_2)+\sum_{c_1 c_2}\frac{C(c_1 c_2)}{N-1}\log\frac{p(c_2|c_1)}{p(c_2)}\right] = H(w)-I(c_1,c_2) \end{aligned}$$

(3)表明可通过合适的划分函数 π 可获取最小交叉熵，这等价于使互信息最大化，因此可选择使平均互信息最大的聚类结果作为最优聚类结果。词类可通过一种聚类算法获取，对二元模型，我们提出了一种基于交换的算法，见图 1。假定所有单词被划分到 G 类，频度最高的 G-1 个单词作为前 G-1 个类，每类一个单词，其余单词组成类别 G。聚类终止条件是预定义的迭代次数，当没有单词在类别间交换时，聚类也终止。

```
Repeat
For every word w in corpus
    Get the current cluster cur in which w is
    For every cluster c in corpus
        Calculate the increase of mutual information if word w is moved w from cur to c
    Determine the target cluster tar to which if word w is moved from cur the mutual
    information can increase fastest.
    Move word w from cur to tar
Until no words are moved or the iteration time is above some pre-specified threshold.
```

图 1 最大化平均互信息的实现算法

2.2 基于功能词的聚类 (FW)

功能词含有丰富的句法知识，因此可通过所有功能词相对内容词的分布信息刻画内容词的上下文分布，一些典型的功能词列于表 1 中。

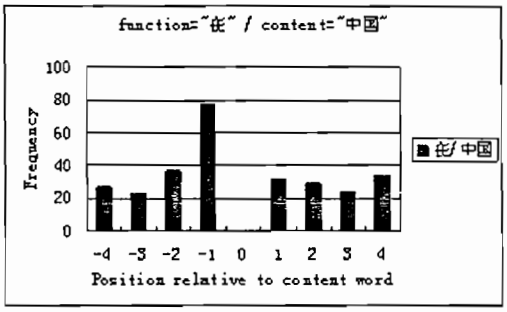
表 1 功能词举例

方位词(orientation word)：左 右 东 南 东部 西部 南部 北部 中部…
助词(auxiliary word)：的 了 等 地 着 …

介词(preposition)：在 对 为 与 从 自 自从 以 把 向 …
连词(conjunction)：关于 为了 按照 依照 因为 除了 …

某功能词相对特定内容词的分布信息可通过它在该内容词的某窗口中出现的频度信息计算，这里窗口被定义为内容词两边计算功能词频度时需考虑的距离(以单词数目度量)。对 98 年 1 月的人民日报语料，在相对于中国的频率信息见图 2(a)。

内容词窗口中每个位置可视为一维，把功能词每维频度连接起来可形成一个向量，功能词在相对内容词中国的分布信息的向量表示见图 2(b)。每个内容词的分布可通过一个向量表示，这个向量是所有功能词相对内容词的分布向量的连接。单词向量表示完成后，每个分布向量都应加以规范化，最后采用 KMEANS 算法进行单词聚类。



(a) "在"相对"中国"的频度信息

Relative Position	Frequency
-4	27
-3	23
-2	37
-1	77
0	0
1	32
2	29
3	24
4	34

27
23
37
77
0
32
29
24
34

(b) "在"相对"中国"的频度向量

图 2 频度向量构造举例

2.3 基于高频词的聚类 (HFW)

每个单词的局部上下文可形式化为单词组成的二元组*<PREVIOUS, NEXT>*。因此，单词的上下文分布可表示为在所有单词上的二维概率分布。KL 距离可用来计算两个单词的相似度：

$$D(p_1 \| p_2) = \sum_{i=1}^{V} p_1(i) \log \frac{p_1(i)}{p_2(i)} \tag{4}$$

p_1 表示前一个单词的概率分布，p_2 表示后一个单词的概率分布，V 表示词典。为获得对称的距离度量，可使用散度作为距离的度量：

$$Div(p_1, p_2) = D(p_1 \| p_2) + D(p_2 \| p_1) \tag{5}$$

因此，两个单词的距离可按(6)计算：

$$Dist(w_1, w_2) = Div\left(p_1^{left}, p_2^{left}\right) + Div\left(p_1^{right}, p_2^{right}\right) \tag{6}$$

2.4 基于词类的聚类(WC)

基于高频词的聚类有两大缺陷：首先是数据稀疏导致了估计模型参数的不准确，其次是它假设单词左右两个位置独立。引入词类对刻画单词分布信息可解决此问题，假定上下文分布可表示为二元类别的概率与给定类别下单词的条件概率之乘积：

$$p(\langle w_1, w_2 \rangle) = p(\langle c(w_1), c(w_2) \rangle)\, p(w_1 | c(w_1))\, p(w_2 | c(w_2)) \tag{7}$$

$c(w)$ 表示单词 w 的类别，c_1 是 $c(w_1)$ 的简化。假设 p_1 和 p_2 的条件分布一样，KL 距离可化简为(8)：

$$KL(p_1 \| p_2) = \sum_{c_1, c_2} p_1(\langle c_1, c_2 \rangle) \log \frac{p_1(\langle c_1, c_2 \rangle)}{p_2(\langle c_1, c_2 \rangle)} \tag{8}$$

3 实验结果

3.1 评估标准

词聚类的评估是一项较为困难的任务，本文提出了两种全新的评价标准：词性标注准确率(POS precision)以及语义准确度(semantic precision)。根据词性进行计算某类准确度的方法如下：

对类中的每个词，在已标注语料库中查找其可能出现的所有标注;将类别按标注进行累加选择出现次数最多的词性作为其类别的代表词性;代表词性出现的次数与类中单词总数的比值即为词性标注的准确度.

实验中采用同义词词林评估聚类的语义相似性，同义词词林由树结构表示，具有相同祖先的单词可认为具有相似语义，选择树根的二级孩子结点作为语义代表。语义准确度被定义为出现次数最多的语义代表的出现次数与类别总词数之比。

3.2 实验设置

我们采用的训练语料为 1998 年 1 月的人民日报，根据北京大学词性标注规范进行预标注，单词总数有1093083。对 FW 方法，选择了 233 个功能词，窗口大小为 2，单词距离由 Manhattan 公式计算。对 HFW 方法，频度最高的 1000 个单词作为特征。为公平评测聚类质量，考虑对 120 个类 4096 个单词进行聚类。

3.3 实验结果

表 2 列出了四种聚类结果，词性标注准确率都超过了 50%，语义准确率超过 30%。 MMI 聚类质量最好，其词性标注准确率可达 79.1%，语义准确率可达 50.%。

表 2 各种聚类方法的比较结果

Method	HFW(%)	WC(%)	FW(%)	MMI(%)
POS Precision	53.09	51.09	62.61	**79.09**
Semantic Precision	29.78	32.63	36.68	**49.75**

表 3 列出 MMI 聚类结果中随机抽出的几个类别，从结果看它可识别人民，机构名，部分专有名词，职业以及官衔等。其中包括一些语义相同的类，如词类{财产,成本,抵押,贷款,工资,金额,利率,利润,投资,关税,债权,增幅 资金}都表示金融上的资金。

表 3 MMI 方法聚类结果中随机抽出的几个词类

到 抵 赶赴 进驻 往 于 至 …	这部 这次 这个 这家 这项 …
阿尔及利亚 阿根廷 奥地利 澳大利亚 …	打破 带动 抵制 调动 夺取 遏制 …
局长 秘书长 会长 社长 省长 首相 司长..	除夕 春节 虎年 佳节 年初 年底 农历 ..
规范 和谐 合格 合理 缓慢 混乱 激烈 …	财产 成本 抵押 贷款 工资 金额利率 …
10 月 11 月 12 月 1 月 2 月 3 月 4 月 …	邹家华 朱镕基 叶利钦 希拉克 吴邦国 …
省人大 省政府 书记处 铁道部 外交部…	膨胀 上升 上涨 通胀 下跌 下滑 增产 …
球员 师生 投资者 消费者 志愿者 老百姓	石油 烟草 邮电 油气 医疗 医药 …

3.4 词类用于汉语句法分析

我们曾采用基于对齐的方法[9][10]对汉语句法结构进行自动分析，获得较为理想的结果，其基本思想是将汉语句库中的句子对两两对齐，以不同的单词片段作为句子成分。借助于词类，句对之间的对齐可被适当调整从而提高句法推导的质量。表 4 给出基于对齐的中文句法结构自动分析结果以及将词类结合起来学习的结果，FW，MMI，HFW，WC 表示采用相应的词类调整对齐，NC 表示未引入词类。最终 NC 的准确度，召回率以及 F 值都低于其它几种引入词类的方法，这表明通过词类可改善对齐结果从而提高句法分析的质量。性能上没有非常明显提高的原因在于所采用的汉语句库其标注标准与人民日报不一样且词类训练仅在人民日报上进行，词类在汉语句库中出现的较为稀疏。

表 4 词类引入后中文句法分析的实验结果

Method	FW(%)	MMI(%)	HFW(%)	WC(%)	NC(%)
Precision	28.18	28.19	**28.39**	28.01	27.58
Recall	46.54	**46.70**	46.62	46.62	46.38
FScore	35.11	35.15	**35.29**	35.00	35.00

4 结论

通过词聚类可构建小型并且有效的语言模型，它可改善 n 元语言模型以及句法结构自动获取的质量。本文我们关注于无监督的中文单词聚类研究，提出了四种不同的聚类方法，最终发现 MMI 方法和 FB 方法可获得句法功能相同，语义相近的词类。句法类的准确度最高可达 79.09%，语义上可达 0.497549。词类被用于基于对齐的句法分析后，句法分析性能提高了 1%。

参考文献：

[1] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifier C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. In Proceedings of the IBM Natural Language ITL, pages 283-298, Paris, France, March.

[2] Andrew Roberts .Automatic Acquisition of Word Classification using Distributional Analysis of Content Words with Respect to Function Words , November 17 2002.

[3] Rile Hu, Chengqing Zong and Bo Xu.Semi-automatic Acquisition of Translation Templates from Monolingual Unannotated Corpora. pages 163-173, IEEE 2003.

[4] Sven Martin , Jorg Liermann , and Hermann Ney. Algorithms For Bigram And Trigram Word Clustering, October 05 1995.

[5] Y. Wang and John Lafferty and A. Waibel .Word Clustering with Parallel Spoken Language Corpora

[6] F. Pereira and N. Tibshy and L. Lillian .Distributional Clustering of English Words ,CL, 1993.

[7] Klein D. The Unsupervised learning of natural language structure. PHD thesis, Stanford University. 2005

[8] Clark A. Unsupervised induction of stochastic context-free grammars using distributional clustering. In: Proc. of CoNLL 2001, July 2001, Toulouse, France.105-112

[9] Van Zaanen, M. and Adriaans, P. Alignment-Based Learning versus EMILE: A comparison. In Proc. of the Belgian-Dutch Conference on Artificial Intelligence (BNAIC); Amsterdam, the Netherlands, 2001, 315–322.

[10] Van Zaanen, M. ABL: Alignment-based learning. In Proceedings of the 18th International Conference on Computational Linguistics (COLING 18),2000, 961–967.

[11] Finch, S., & Chater, N.(1992a). Bootstrapping syntactic categories. In Proceedings of the 14th Annual Meeting of the Cognitive Scienoe Society, pp. 820–825.

[12] Finch, S., & Chater, N.(1992b). Bootstrapping syntactic categories using statistical methods. In

[13] Daelemans, W., & Powers, D. (Eds.), Background and Experiments in Machine Learning of Natural Language, pp. 229–235. Tilburg University: Institute for Language Technology and AI.

[14] Sch¨utze, H. (1993). Part of speech induction from scratch. In Proceedings of the 31st annual meeting of the Association for Computational Linguistics, pp. 251–258.

[15] Sch¨utze, H. (1997). Ambiguity Resolution in Language Learning. CSLI Publications.

[16] Martin Redington, Nick Chater and Steven Finch. Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. Cognitive Science, Vol 22 (4), pp 425-469. Cognitive Science Society. 1998

[17] 俞士汶 等编著，现代汉语语法信息词典详解（第 2 版），清华大学出版社，2003