面向句法分析的样本选择

孙俊，曹海龙，赵铁军

（哈尔滨工业大学计算机学院，哈尔滨市，150001）

摘　要： 句法分析是自然语言处理的一个基本问题，也是目前急待解决的一个问题。目前大多数的句法分析是基于统计方法的，基于统计的句法分析需要大规模的训练语料，而标注一个大规模语料需要很大的人力。为了减少标注句法树库所需的人力，本文对选择样本进行了研究。本文从句法结构上对句子进行聚类，根据聚类的结果精选出一个小的句子集，这个句子集的规则分布近似于整个句子集的规则分布。标注这个句子集就能在保证句法分析器性能的前提下减少标注所需的人力。实验结果证明，通过选取一半的句子训练出的句法分析器，其性能就能近似于用所有句子训练的句法分析器的性能。

关键词： 句法分析，样本选择，聚类

Sample Selection for Statistical Parsing

Jun Sun, Hailong Cao, Tiejun Zhao

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

Abstract: Parsing is one of the fundamental problems in natural language processing, and the main approach is statistical parsing. Statistical parser relies on using many hand-parsed sentences as training examples. However, the task of labeling so many sentences is a labor-intensive task. We proposed to select these samples to reduce the amount of sentences in the training data by clustering based on syntactic structures, thereby reducing the workload of human to label sentences. Our result showed that the amount of training sentence could be reduced by 50% with an approximate performance of the statistical parser.

key words: Parsing, sample selection, clustering

1　引言

句法分析是自然语言处理中的一个基本问题，也是目前急待解决的一个问题[1]。目前主流的句法分析是基于统计方法的，对基于统计的句法分析来说，提出一个好的算法固然重要，但拥有一个好的句法树库也是必要的。一般认为，一个大规模句法树库能比较好的反映自然语言的特征，从而使句法分析有比较好的性能，但标注一个大规模句法树库需要很大的人力。为了在保证句法分析器性能的情况下减少标注所需要的人力，本文提出了一种对这些句子样本进行精选的方法：在标注语料前，从句法结构上对句子进行聚类，并根据聚类的结果选择一部分句子标注后作为句法分析器的训练样本。这样可以减少标注句子的数量，从而减少标注所需的人力，而在质量上，选择出来的句子集的句法规则分布近似于整个句子集的句法规则分布，这样可以使通过它们训练的句法分析器的性能同用整个句子集训练的句法分析器的性能相近。

作者简介：孙俊，男，湖北省鄂州市，硕士研究生，sunjun15@sohu.com

2 面向句法分析的样本选择相关研究

目前面向句法分析的样本选择方法是基于模型的自主学习方法[2] [3] [4]。该方法先用少量的样本训练出一个初始的模型，然后用该模型分析数据，将模型最不确定的数据标注后加入样本集重新训练模型，然后再重新分析数据，直到模型的性能达到要求或者所有的句子全被标注。在这里，衡量数据的不确定性是用计算信息熵的方法。可以计算一个模型的多个分析结果及其概率的熵，也可以计算多个不同模型分析结果的熵，来确定模型对数据的不确定性。

在上面的方法中，每次模型最不确定的数据之间可能有一定的冗余性。为了解决这个问题，唐民提出了从聚类结果中选择数据的自主学习方法[5]。该方法训练出一个初始的模型后，根据该模型对句子的分析结果将句子聚类，然后在每个类中根据类的大小选择一些最不确定的句子标注后加入到训练集中重新训练模型，然后再用该模型对句子分析，如此循环直到模型的性能达到要求或者所有句子全部被标注。

3 面向句法分析的样本选择算法

目前面向句法分析的样本选择方法基本上都是自主学习的方法，自主学习方法离不开一个初始的模型和用来衡量不确定性的熵计算。这样选择出来的样本集一方面会受到模型算法的影响，另一方面，计算信息熵只能从局部最优化样本选择，而不能从全局考虑整个样本集的规则分布。本文提出了一种独立于模型的面向句法分析的样本选择方法：根据句法结构对句子聚类，使每个类的句子在句法结构上相同或很相似，从每个类中选择一部分句子进行标注作为训练集。这样能使选择的句子集在句法规则分布上近似于原来整个句子集，从而能在保证句法分析器性能的前提下节省需要标注的样本和标注所需的人力。本文设计的面向句法分析的样本选择算法主要分为下面两个步骤：

1. 根据句子的词性序列计算句子间的距离，然后利用 k-近邻算法对句子进行聚类[6]
2. 从每个类中取出一定的句子加入到需要标注的训练集。

3.1 计算句子距离并聚类

在这里计算句子间的距离，是为了量化的表示两个句子间句法树的结构的相似程度，并根据句子间的距离聚类，从而使句法结构相同或很相似的句子能够被聚成一类。这里的句子经过分词和词性标注，但没有标注其句法树。因为是独立于模型的，也不能通过句法分析得到其句法树，所以这里只能根据句子的分词和词性标注结果来近似的表明句法结构。

如果两个句子的句法结构相同，那么它们的词性序列一定相同。这里假定，如果两个句子有一个越长的相同词性序列，那么这两个句子的结构越相似，它们的距离也就越小。很显然。计算距离公式如下：

$$D(A,B)=1-\frac{2sameness_{AB}(sameness_{AB}+1)}{num_A(num_A+1)+num_B(num_B+1)} \qquad (1)$$

$sameness_{AB}$ 表示句子 A 和句子 B 的最长的相同词性序列包含的词的个数，num_A 和 num_B 分别表示句子 A 和句子 B 所包含的词的个数。在这里，词性序列是连续的，比如一个句子的词性序列 A B C D 包含的词性序列有：A, B , C, D, AB, BC, CD, ABC, BCD, ABCD，但 AC 不是词性序列，因为它不是连续的。

本文用 k-近邻算法进行聚类，具体步骤如下：

1) 随机选择 n(类的个数)个句子作为初始的中心句子，计算每个句子同这些中心句子的距离，并将其归入最近句子的一类。
2) 对每个类，重新计算其中心句子：距离该类所有句子距离和最小的句子是中心句子。
3) 根据重新得到的中心句子，重新聚类。
4) 如果聚类结果收敛（不一定是绝对的收敛），结束聚类，否则继续步骤 2 和步骤 3。

3.2 根据聚类结果选择句子

在数据不缺稀的时候，通过上面的聚类方法，可以将句法结构相同或很相似的句子聚为同一类，每个类中选择一个中心句子，就能让重新选择的句子集近似于原句子集的规则分布。比如：有两个句法结构 A 和 B，句法结构 A 的类有 3 个句子，句法结构 B 的类有 5 个句子，从 A 类中选择出句子 a，从 B 类中选择出句子 b，让 a 在选择的句子集中出现 3 次，b 出现 5 次，这样选择的 2 个句子的规则分布就能近似于原来的 8 个句子的分布。因为句法分析需要一个很庞大的训练集，现在就是要从一个庞大的未标注的数据集中精选出一个子集，所以数据不缺稀的条件可以满足。

4 实验结果

本文首先用 Kullback Leibler distance（简称 KL-distance）[7]算法观察通过聚类选择出来的句子集同整个句子集的规则分布相似程度，并用随机选择的句子集同整个句子集的规则分布相似程度作为基线进行比较。然后利用基于概率上下文无关文法（PCFG）的句法分析器的性能，即句法分析器的精确率和召回率，来评价选择的句子集。用 1000 个句子作为测试集对句法分析器的性能进行测试。整个供选择的句子集有 12000 个句子，用随机选择句子作为基线进行比较。

4.1 句法规则分布

KL-distance 可以用来表示两个分布的相似程度，两个分布越相似，它们的 KL-distance 越小，反之则越大。两个相同的分布的 KL-distance 为 0。公式(2)是 KL-distance 的计算公式：

$$D_{KLD}(p_1, p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} \partial x + \int p_2(x) \log \frac{p_2(x)}{p_1(x)} \partial x \qquad (2)$$

公式(2)中的 p_1 和 p_2 分别是要比较的两个分布。

表 1 是随机选择的句子集同整个句子集的 KL-distance。第一列的句子数指随机选择的句子集所包含的句子数。由于句法规则很多，在这里只选择三个最具有代表性的规则：NP、VP 和 BNP 来观察，其结果如下：

表 1 随机选择的句子集同整个句子集的 KL-distance

Tab.1 The KL-distance between the subset of the sentences selected randomly and the whole sentences set

句子数	NP	VP	BNP
1000	0.8976	0.397744	0.609431
2000	0.541229	0.232342	0.420626
3000	0.285265	0.133844	0.200517
4000	0.190564	0.0815555	0.125763
5000	0.161916	0.0614129	0.104768
6000	0.156694	0.0526498	0.10389
7000	0.129382	0.0409546	0.0925536
8000	0.0721394	0.0223373	0.0337209
9000	0.0730227	0.0191392	0.0352666
10000	0.0721796	0.0164246	0.0388475
11000	0.0244759	0.00763719	0.0104059
12000	0	0	0

通过聚类选择的 6000 个句子同整个句子集的 KL-distance 为：BNP = 0.0103952，NP = 0.0497208， VP =

0.0337655，4000 个句子的结果：BNP = 0.0165644，NP = 0.0923995，VP= 0.0631438，同表 1 比较可以看出，通过聚类选择的 4000 个句子的 BNP 规则分布比随机选择的 10000 个句子的 BNP 规则分布更接近于整个句子集的 BNP 规则分布，同样，其 NP 和 VP 的规则分布也超过了随机选择 7000 和 4000 个句子的分布；而通过聚类选择的 6000 个句子的 BNP，NP 和 VP 的分布也分别好于随机选择 11000，10000 和 7000 个句子的分布。由此可以看出，通过聚类选择的句子其规则分布比随机选择的句子的规则分布更相似于整个句子集的规则分布。

4.2 随机选择句子

表 2 是用随机选择方法从 12000 个句子中选择一部分句子标注后作为训练集训练句法分析器。句子数是指随机选择句子的个数，将这些句子标注后训练句法分析器，然后用 1000 个句子对该句法分析器进行测试。后面两列分别是相应的句子集训练出的句法分析器的精确率和召回率。本文用随机选择的句子作为基线。

表 2 随机选择句子训练的句法分析器性能

Tab.2 Effect of the parsers trained on the corpus selected randomly

句子数	句法分析器精确率	句法分析器召回率
1000	74.537541%	70.112590%
2000	74.563506%	74.410659%
3000	73.917422%	75.282051%
4000	73.326706%	75.683995%
5000	74.098578%	76.607387%
6000	73.950413%	76.504788%
7000	74.165841%	76.778386%
8000	74.143610%	76.983584%
9000	74.735799%	77.393981%
10000	75.355137%	78.009576%
11000	75.016545%	77.530780%
12000	75.157337%	77.599179%

4.3 通过聚类选择句子

表 3 是通过聚类的方法从相同的 12000 个句子里面选择一部分句子标注后作为训练集训练句法分析器。句子数是指通过聚类选择的句子集所含句子的数量，后面两列是相应的句法分析器的精确率和召回率。

表 3 通过聚类选择句子训练的句法分析器性能

Tab.3 Effect of the parsers trained on the corpus selected by clustering

句子数	句法分析器精确率	句法分析器召回率
1000	71.974522%	69.419795%
2000	72.052402%	73.333333%
3000	72.546419%	74.829001%
4000	73.307035%	76.213260%
5000	75.400534%	77.257182%
6000	75.418060%	77.120383%

由以上实验可以看出，通过聚类选择的句子集其句法规则分布比通过随机选择的句子集的句法规则分布更接

近于整个句子集的句法规则分布。根据表 2 和表 3 可以看出，通过聚类选择的句子数超过 4000 的时候，用其训练的句法分析器的性能会很大地超过用随机方法选择标注句子训练的句法分析器的性能。并且通过聚类选择 6000 个句子，用其训练的句法分析器的性能就已经近似于用整个句子集训练的句法分析器的性能。

5 结论

基于统计的句法分析需要大规模的句法树库作为训练样本，而标注一个大规模句法树库需要很大的人力。为了在保证句法分析器性能的情况下减少标注所需要的人力，本文提出了一种面向句法分析的独立于模型的样本选择方法：在标注语料前，从句法结构上对句子进行聚类，并根据聚类的结果选择一部分句子进行标注作为句法分析器的训练样本。这样减少标注所需的人力，而在质量上，选择出来的句子集的句法规则分布近似于整个句子集的句法规则分布，这样可以使通过它们训练的句法分析器的性能同用整个句子集训练的句法分析器的性能相似。本文通过实验证明了通过选取一半的句子训练出的句法分析器，其性能就接近于用所有句子训练的句法分析器。

参考文献：

[1] 赵铁军. 机器翻译原理. 哈尔滨工业大学出版社, 2000:156~202

[2] Hwa, R. Sample selection for statistical grammar induction. Proceedings of the 2000 Joint SIGDA Conference on EMNLP and VLC, 2000, pp. 45-52.

[3] Steedman, M, Osborne, M, Sarkar, A, et al. Bootstrapping statistical parsers from small datasets. The Proceedings of the Annual Meeting of the European Chapter of the ACL, 2003b, pp. 120-127.

[4] Anderson, B, & Moore, A. Active Learning for Hidden Markov Models: Objective Functions and Algorithms. Appearing in Proceedings of the 22nd International Conference on Machine Learning, 2005.

[5] Tang, M, Luo, X, Q, & Roukos, S. Active learning for statistical natural language parsing. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2001, pp. 120-127.

[6] Richard, D, Peter, H, & David, S. Pattern Classification. Second edition, 2003, pp. 143-153.

[7] Minh, D, & Martin, V. Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance [J]. IEEE transaction on image processing, 2002, 11(2): 146-158.