文章编号：1003-0077(2012)02-0018-05

基于主动学习的中文依存句法分析

车万翔，张梅山，刘 挺

（哈尔滨工业大学 计算机学院 社会计算与信息检索研究中心，黑龙江 哈尔滨 150001）

摘 要：目前依存句法分析仍主要采用有指导的机器学习方法，即需要大规模高质量的树库作为训练语料，而现阶段中文依存树库资源相对较少，树库标注又是一件费时费力的工作。面对大量未标注语料，该文将主动学习应用到中文依存句法分析，优先选择句法模型预测不准的实例交由人工标注。该文提出并比较了多种衡量依存句法模型预测可信度的准则。实验表明，一方面，与随机选择标注实例相比，当使用相同数目训练实例时，主动学习使中文依存分析性能最高提升 0.8%；另一方面，主动学习使依存分析达到相同准确率时只需标注更少量实例，人工标注量最多可减少 30%。

关键词：主动学习；依存句法；不确定性度量；委员会投票

中图分类号：TP391 **文献标识码**：A

Active Learning for Chinese Dependency Parsing

CHE Wanxiang, ZHANG Meishan, LIU Ting

(Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: It is necessary to have a large annotated Treebank to build a statistical dependency parser. Acquisition of such a Treebank is time consuming, tedious and expensive. This paper presents a method to reduce this demand via active learning, which selects the most uncertain samples for annotation instead of the whole training corpus. Experiments are carried out on the HIT-CIR-CDT, our results show that the parsing accuracy rises about 0.8 percent by active learning when using the same amount of training samples. In other words, for about the same parsing accuracy, we only need to annotate 70% of the samples as compared to the usual random selection method.

Key words: active learning; dependency parsing; uncertainty-based sampling; query-by-committee

1 引言

在统计学习的模型训练过程中，按照对训练实例的处理方式，可将学习过程分为两类：主动学习和被动学习。被动学习是随机地选择训练实例，被动地接受这些样本信息。主动学习与被动学习不同，它是迭代地从未标注语料中优先选择最富含有效信息的实例（即当前模型预测最不准的）交由人工标注，然后加入训练集重新训练。由于优先选择的是最具训练效用的样本，所以减少了那些对提高学习器精度帮助不大的冗余样本的标注，因而学习器只需更少的样本便能获得相同精度[1-2]。

目前最广泛使用的主动学习方法有基于不确定性度量（Uncertainty-based Sampling）和基于委员会投票（Query-by-committee）两种[1]。

基于不确定性度量的样本选择根据学习器对未标注样本的分类置信度来进行。样本分类置信度越

收稿日期：2011-09-20 **定稿日期**：2011-12-21

基金项目：国家自然科学基金重点项目（61133012）；国家自然科学基金资助项目（60803093）；国家 863 重大项目（2011AA01A207）；核高基重大专项（2011ZX01042-001-001）；哈尔滨工业大学科研创新基金（HIT. NSRIF. 2009069）；中央高校基本科研业务费专项资金（HIT. KLOF. 2010064）

作者简介：车万翔（1980—），男，讲师，主要研究方向为自然语言处理；张梅山（1983—），男，博士研究生，主要研究方向为自然语言处理；刘挺（1972—），男，教授，主要研究方向为自然语言处理，信息检索。

低，说明学习器尚不能很好区分此样本，即学习器缺乏此样本含有的信息。此时将该样本进行人工标注并加入训练集会对学习器精度的提升有很大帮助。对于分类置信度高的样本，不再人工标注，从而免除了在冗余样本上耗费人力。这类学习算法的重点是构造一种合理有效的不确定性度量机制，以此来指导样本选择。

基于委员会投票的样本选择需要构建一组分类器，这些分类器可以是用不同的训练算法得到(SVM、MaxEnt 等)，也可以是用同种训练算法对样本从不同的特征角度训练得到(Multi-view active learning[3])。基于委员会投票的方法优先选择各分类器投票结果最不一致的样本进行人工标注。投票熵(Vote Entropy，Dagan and Engelson，1995)和相对熵(KL divergence to the mean，Pereira et al.，1993)是两种最常用的度量投票结果差异的方法。熵值越高，说明投票差异越大，该样本越应该加入到训练集[4]。

国外学者已经将主动学习应用到诸多自然语言处理相关的任务中，比如信息抽取(Thompson et al.，1999)、文本分类(McCallum and Nigam，1998)和基于短语结构的句法分析(Thompson et al.，1999；Hwa，2000)[5-6]等。在国内，清华大学覃刚力、北京理工大学宋鑫颖等将主动学习应用到文本分类上[7-8]；中国科技大学冯冲、上海交通大学陈霄分别用最大熵模型和支持向量机模型将主动学习应用到组织机构名识别中，并取得了一定效果[9]。就作者所掌握的文献，目前还没有将主动学习和中文依存句法的训练过程相结合的研究。在应用最大熵或者支持向量机模型进行预测的自然语言处理任务中，前者可以得到每个样本属于某一类别的概率，后者可以得到每个样本到分类超平面的距离。这些预测任务的置信度比较容易获得，比如基于 SVM 的文本分类中距离分类超平面最近的样本置信度就比较低等。基于短语结构的句法分析可以根据每个产生式的概率计算最终生成的短语结构树的概率，并利用此概率值进行各种可信度计算；而依存句法通过 Online 算法训练权值，最终求一棵权值最大的生成树，很难得到生成树的概率，原有的基于短语结构的可信度度量方法也就不能直接应用到依存分析上。因此，本文尝试将主动学习应用到依存分析上，并尝试了多种衡量依存句法模型预测可信度的准则。

本文内容组织为，第二部分介绍依存句法分析相关概念和基于图的依存分析算法；第三部分介绍主动学习的算法流程，其中详细讨论了如何衡量依存句法模型的预测可信度；第四部分是实验；第五部分给出结论和下一步工作。

2　中文依存句法分析

主动学习需要事先在小数据集上训练一个依存句法分析器，用来对未知样本进行可信度预测。本文采用基于图的依存分析算法来训练依存分析器，以下简要介绍基于图的依存句法分析。

2.1　基于图的中文依存句法分析

McDonald 首先提出将依存分析问题归结为在一个有向图中寻找最大生成树(Maximum Spanning Tree)的问题。边权使用 Online Learning 算法学习获得，解码使用 Eisner 算法[11]。

Eisner 算法以 span 为解码的基本单位，span 表示输入句子的一个片段对应的子树。与组块不同的是，span 中的核心词只能位于片段首或尾，即 span 只包括了这个词左边或者右边的子孙节点。另外，除核心词外的另外一个片段首或尾词的修饰成分可以是不完整的，即 span 可以不包括这个词左边的子孙节点或者右边的子孙节点。对于其他词，span 包括它们所有的子孙节点。span 的这种特性使得解码算法独立地确定一个词左边的修饰成分和右边的修饰成分，从而降低算法的复杂度[10]。

基于图的依存分析算法是目前性能最高的依存分析方法之一。

3　基于主动学习的中文依存句法分析

本文将主动学习应用到基于图的依存句法训练过程中，具体的算法流程如下。

L：人工标注后的实例(句法依存树库)
U：未标注的实例(已经过分词和词性标注的句子)
C：当前已标注实例训练得到的模型(基于图的依存分析器训练)
Φ：衡量实例可信度的函数
Batch-Size：每轮主动学习挑选实例的个数
初始化：
　从 U 中取出小部分未标注实例，人工标注，加入 L；
C ← Train(L)；
重复：
　U1 ←用 C 预测所有 U；

N ← 根据函数 Φ，选择不可信度最大的 Batch-Siz 个实例；

U ← U − N；

人工标注 N 后，加入 L；

C ← Train(L)；

直到：

以下三种停止准则中至少满足其一。

3.1 可信度度量

可信度度量是主动学习选择某一训练实例交由句法分析器训练的依据。由于依存句法中如何衡量预测结果的可信度并没有现成方法可用，本文利用依存分析对每一句子输出的 K-Best 预测结果来衡量可信度。该方法的思想：一个句子的 K-Best 结果越相似，说明其越容易混淆，相应地该句的预测结果可信度应越低。本文采用了基于不确定性和基于委员会投票两大类方法。

3.1.1 基于不确定性(Uncertainty-based Sampling)

该方法依据 K-Best 结果对每个句子预测结果的不确定性做出量化，排序后优先选择不确定性最大的，也即可信度最低的。不确定性度量 Φ 本文采用了以下三种。

a) Uncertainty-1. K-Best 结果中任两个不同结果的分值差的和的倒数，如公式(1)所示：

$$\Phi = \frac{1}{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}(score_i - score_j)} \quad (1)$$

b) Uncertainty-2. K-Best 结果中 1-Best 相对 2-Best 的增长率的倒数，如公式(2)所示：

$$\Phi = \frac{score_2}{score_1 - score_2} \quad (2)$$

c) Uncertainty-3. K-Best 结果的 Sentence Entropy，如公式(3)所示：

$$\Phi = \sum_{i=1}^{k} - p_i \log p_i, \quad \text{其中} \quad p_i = \frac{score_i}{\sum_{j=1}^{k} score_j} \quad (3)$$

前两种度量方法的思想是：若 K-Best 不同结果的分值差异越显著，说明学习器对做出的选择越有信心。第三种度量方法先对 K-Best 每棵树的分值归一化处理，然后通过 K-Best 结果计算熵。之所以要进行归一化处理，是由于基于图的依存分析是判别模型，输出的分值不是概率，也就无法计算熵。

3.1.2 委员会投票(Query-by-committee)

委员会方法需要构造一组不同的学习器来进行投票。基于图的依存句法分析已在 2.2 节介绍过，这里介绍另一种基于转移的依存分析方法。基于转移的方法最早由 J. Nivre 提出[12]，他将依存树的搜索过程建模为一个动作序列，依存分析问题转化为寻找最优动作序列的问题，每次需要根据历史采取动作时由一个分类器给出，例如，支持向量机、最大熵分类器等。

委员会投票优先选择基于图和基于转移两种依存分析方法预测结果最不一致的句子。句法分析结果通常用 LAS 和 UAS 度量，本文得到两种依存分析方法的自动结果后，以基于图的结果为准则，以基于转移的结果做测试结果，计算它们的 LAS 和 UAS 值作为投票的一致性评价指标。本文分别记为 QBC-LAS 和 QBC-UAS。

4 实验

4.1 实验数据

实验数据来自 HIT-CIR-CDT(哈尔滨工业大学信息检索研究中心汉语依存树库)，该树库共包含 59 996 个句子，每个句子均经过人工标注，全部来源于《人民日报》。为了更好检验主动学习的作用，本文只筛选句长在 10～25 个词(不含标点)之间的句子，共约 42 000 句。

4.1 实验结果及分析

主动学习的效果往往通过一条学习曲线(Learning Curve)来评价[1]。本文将主动学习的批处理参数 Batch-Size 设置为 1 000，初始采用 1 000 句训练基于图的依存分析器，以后每次通过主动学习的抽样方法加入 Batch-Size 个句子，算法共迭代 10 次。基于图的依存分析器 K-Best 结果设置为 5。同时为了与主动学习对比，本文用每次随机选择 1 000 句的方法作为基准测试。每次训练出的模型都在 3 000 句测试集上测试，结果如表 1 所示。

从表 1 看出，基于不确定性和委员会投票这两类主动学习中的任一种方法在 10 次主动学习的迭代过程中都取得了比随机抽样更好的性能。一方面，使用相同数目的训练实例，主动学习能使句法分析取得更好的性能：10 次迭代结束时，Uncertainty-1 比 Random 提高了 0.78 个百分点(其中在第 7 次迭代时达到最大，提高 0.8 个百分点)；另一方面，主动学习使句法分析取得同样的性能只需人工标注更

表 1　10 次迭代过程中各种主动学习方法在测试集上的 LAS 和 UAS/%

		1	2	3	4	5	6	7	8	9	10
Random	LAS	70.46	73.36	75.13	76.16	76.78	77.51	78.06	78.42	78.86	79.09
	UAS	74.02	76.70	78.34	79.30	79.87	80.63	81.13	81.44	81.85	82.10
Uncertainty-1	LAS	70.46	73.79	75.68	76.82	77.61	78.11	78.86	79.43	79.67	79.87
	UAS	74.02	77.19	78.93	80.06	80.73	81.07	81.89	82.37	82.60	82.81
Uncertainty-2	LAS	70.46	73.79	75.68	76.82	77.61	78.11	78.86	79.43	79.67	79.87
	UAS	74.02	77.19	78.93	80.06	80.73	81.07	81.89	82.37	82.60	82.81
Uncertainty-3	LAS	70.46	74.10	75.56	76.93	77.60	78.36	78.73	79.14	79.34	79.88
	UAS	74.02	77.43	78.88	80.12	80.74	81.44	81.76	82.12	82.32	82.87
QBC-UAS	LAS	70.46	73.59	75.46	76.46	77.40	78.01	78.41	78.94	79.29	79.57
	UAS	74.02	76.94	78.72	79.67	80.56	81.16	81.47	81.98	82.27	82.63
QBC-LAS	LAS	70.46	73.80	75.18	76.43	77.05	77.65	78.40	78.71	79.40	79.53
	UAS	74.02	77.20	78.45	79.70	80.20	80.78	81.45	81.71	82.38	82.50

少量的训练语料：Random 第 10 次迭代，即使用 10 000 句训练达到 79.09% 的准确率（LAS），而 Uncertainty-1 在第 7 次迭代时，只使用 7 200 句，就达到同样性能，减少了近 30%的人工标注量。

将以上表格绘制成折线图，如图 1 所示：

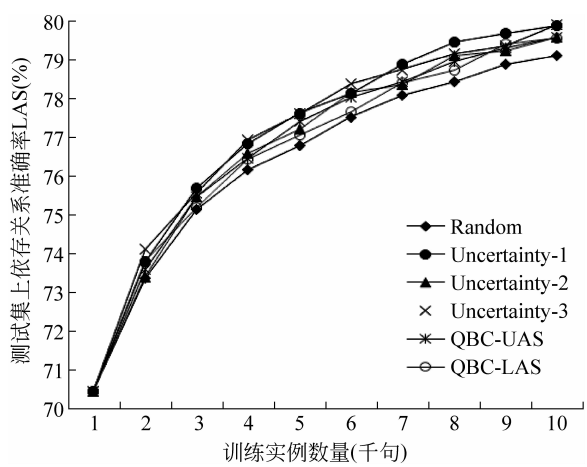

图 1　10 次迭代过程中各种主动学习方法在测试集上的 LAS 递增曲线

图 2 给出了测试集 LAS 达到 79.09%时，基准测试和主动学习各方法所需训练实例数目的对比。

实验表明，在五种可信度度量方法中，Uncertainty-1 的效果最明显，减少了近 30%标注量，其次是 Uncertainty-3，效果最低的是 QBC 的两类方法。分析原因为：与 Uncertainty-2 相比，Uncertainty-1 和 Uncertainty-3 综合考虑了所有 K-Best 结果的差异，而不仅仅是最高的前两个，因此受噪声影响的可能性最小，效果最好；QBC 只减少了 15%标注量，很可能在于基于转移的依存分析性能上与基于图的方法仍存在不小差异，实验使用的基于转移的依存分析器 LAS 为 72.83%（8 000 句训练，1 000 句测试），而同样的训练和测试数据，基于图的 LAS 达到 78%以上。因此，基于转移方法引入的噪声很可能影响投票效果，导致 QBC 实验效果较低。在具体应用时，应优先考虑 Uncertainty-1 的可信度度量方法。

4　结论及下一步工作

本文提出一种基于主动学习的中文依存句法分析方法。该方法从大量未标注语料中优先选择不确定性最大，也即当前句法模型预测可信度最低的句子交由人工标注，随后加入训练集迭代训练，如此往复，直到满足停止准则。其中，主动学习使用了不确定性度量和委员会投票两种最经典的方法。实验证明，一方面，使用相同数目的训练实例，主动学习使 LAS 最高提升 0.8 个百分点；另一方面，主动学习使得只需标注更少量的句子就可达到相同的依存分析准确率，人工标注量最多可减少 30%，从而大大降低了树库标注的人力成本。

下一步工作主要是探索主动学习的其他有效方法，比如从训练语料的多样性出发，可以先聚类再从每个簇抽样；或者优先选取训练语料中最具代表性的实例，因为最具代表性的实例可以覆盖到更多的实例信息，它本身是噪声点的可能性也往往最低，难

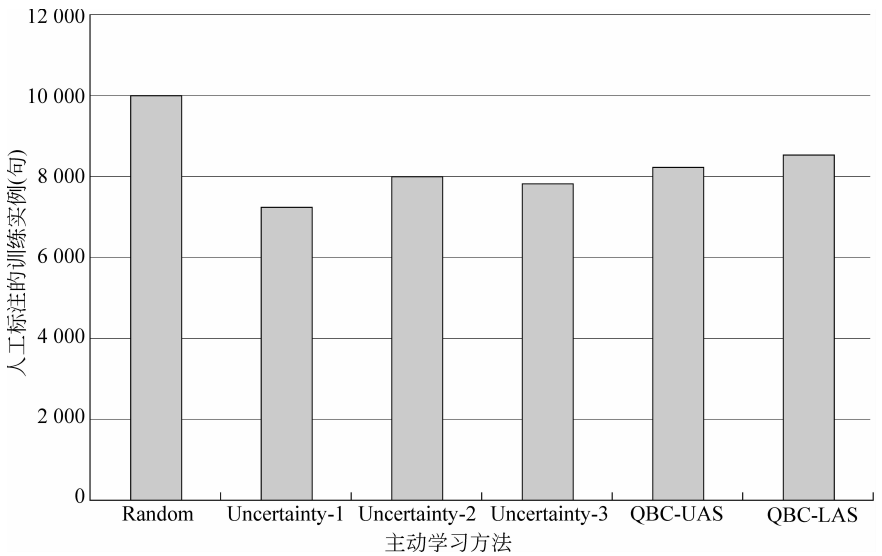

图 2 测试集 LAS 达到 79.09%时,基准测试和主动学习各方法需要的训练实例数目

点在于最具代表性的度量,这可以通过计算相对熵得到。另外,可以尝试将主动学习应用到句法分析的领域移植上,以减少获取新标注语料时的压力。

参考文献

[1] Olsson Fredrik. A literature survey of active machine learning in the context of natural language processing [R]. Swedish Institute of Computer Science. 2009.

[2] Min Tang, Xiaoqiang Luo, Salim Roukos. Active Learning for Statistical Natural Language Parsing [C]//Proceedings of the 40th ACL. 2002:120-127.

[3] Ion Muslea, Steven Minton, Craig A. Knoblock. Active Learning with Multiple Views[J]. Journal of Artificial Intelligence Research. 2006, 27:203-233.

[4] Yoav Freund, H. Sebastian Seung. Selective Sampling Using the Query by Committee Algorithm[J]. Machine Learning. 1997, 28:133-168.

[5] Cynthia A. Thompson, Mary Elaine Califf, Raymond J. Mooney. Active Learning for Natural Language Parsing and Information Extraction[C]//Proceedings of the Sixteenth International Conference on Machine Learning. 1999:406-414.

[6] Rebecca Hwa. Sample Selecting for Statistical Parsing [J]. Computational Linguistics. 2004, 30(3):253-276.

[7] 覃刚力,黄科等. 基于主动学习的文档分类[J]. 计算机科学. 2003, 30(10):45-48.

[8] 宋鑫颖,周志逵. 一种基于 SVM 的主动学习文本分类方法[J]. 计算机科学. 2006, 33(11):288-290.

[9] 陈霄. 基于支持向量机的中文组织机构名识别[D]. 硕士学位论文, 上海交通大学. 2007.

[10] 李正华. 依存句法分析统计模型及树库转化研究[D]. 硕士学位论文, 哈尔滨工业大学. 2008.

[11] R. McDonald, K. Crammer, F. Pereira. Online Large-margin Training of Dependency Parsers [C]//Proceedings of ACL. 2005:91-98.

[12] J. Nivre. Algorithms for Deterministic Incremental Dependency Parsing[J]. Computational Linguistics. 2008, 34(4):513-553.