一种基于 MIRA 和遗传算法的句法分析模型构造方法*

王 丹[1]，姬东鸿[1]，黄 玮[2]
[1]武汉大学 计算机学院，武汉 430072
[2]天津理工大学 计算机与通信工程学院，天津 300191
[1]E-mail: wanghzc@163.com

摘 要：提出了一种新方法，通过在线学习算法 MIRA（Margin Infused Relaxed Algorithm）和遗传算法 GA（Genetic Algorithm）来构造句法分析模型。首先用 MIRA 用来构造句法分析模型，然后使用 GA 对模型参数进行进一步优化。数字实验表明提出方法构造的句法分析模型具有更好的性能。

关键词：句法分析模型；在线学习算法；遗传算法

Design of Parsing Models by Mean of Margin Infused Relaxed Algorithm and Genetic Algorithm

Wang Dan[1], Ji Donghong[1], Huang Wei[2]
[1]School of Computer, Wuhan University, Wuhan 430072
[2]School of Computer and Communication Engineering, Tianjin University of Technology, Tianjin 300191
[1]E-mail: wanghzc@163.com

Abstract: This study concerns a design procedure of parsing models. We propose an approach to design parsing models with the aid of margin infused relaxed algorithm (MIRA) and genetic algorithm (GA). MIRA is exploited here to construct a parsing model, while DE is employed to optimize the parsing model. Numerical examples are included to evaluate the performance of the proposed model. They are also contrasted with the performance of the parsing models existing in the literature.

Keywords: parsing models; on-line learning algorithm; genetic algorithm

1 引言

近年来，自然语言处理技术已被广泛应用于机器翻译、信息抽取、自动问答、生物医学等多个领域[1-3]。句法分析是自然语言处理的一个关键环节，人们已经提出了多种不同的句法分析模型构造方法。早在 20 世纪 90 年代末，Collins[4] 和 Charniak[5] 就提出迄今为止最好的短语结构分析模型构造方法。但是该方法也存在一些缺陷：不适合大多数的谓语论元信息编码。1959 年由法国语言学家 Tesiniere 在其著作《结构句法基础》一书中提出依存语法的句法分析，该方法能够更有效的学习和分析谓语论元信息编码。2003 年，Yamada 和 Matsumoto[6]提出利用支持向量机训练，在一个转移递减的依存分析中做依存决定。2004 年，Nivre 和 Scholz[7] 提出了基于历史的学习模型。2005 年，McDonald[8] 等提出 MIRA（Margin Infused Relaxed Algorithm）算法构造句法分析模型。此后，其他的学者如 Deyu et al. [9], Lilja et al.[10], Chung-Hsien et al.[11], Ambati [12], and Terry Koo and Michael Collins[13] 等通过不同的方法研究了分析模型。然而，上述所有方法集中关注在如何去构造分析模型，未发现有文献提出采用演化算法对模型参数进一步优化。

* 本文获国家自然科学基金委员会重大研究计划“视听觉信息的认知计算”培育项目“汉语特征结构的资源建设和自动分析研究”（90820005）、基于部分指导的词义学习和词义排岐综合研究（61070082）、复杂网络在词语语义相关性度量中的应用”（61070243）、2008-2009 年度武汉大学人文社科自主创新项目“基于语义的网络舆情智能监测平台研究”、武汉大学 985 二期拓展子项目“基于汉语特征结构的语义描写及其应用”（985yk004）、湖北省教育厅人文社科项目“基于依存语法的语料库标注研究”（2008q275）的资助。

本文提出一种基于遗传算法和MIRA的句法分析模型构造方法。具体构造过程包含两个阶段：模型构造阶段和模型优化阶段。模型构造阶段完成后执行模型优化阶段。模型构造阶段是通过MIRA实现，而模型优化阶段则采用演化算法来完成。一系列对比实验表明，提出方法构造的模型具有更好的性能。文章组织结构如下：第二节提出一种新的两阶段句法分析模型的构造方案；第三节详细阐述模型构造的具体实现步骤；第四节构造了一个模型实例，并和当前一些常用句法模型进行比较；第五节对文章进行了总结。

2 一种句法分析模型构造的新方案

在自然语言处理中，在线学习算法例如MIRA通常被用来生成一个句法分析模型。图1(a)描述了一个传统的句法分析模型生成框架(McDonald[14])。当输入为不同的训练数据，则相应产生不同的句法分析模型。换句话说，如果给定的训练数据是不同的语言，它将会产生基于不同语言的分析模型。所谓的训练或学习实际上是从一个训练集出发，通过在线学习等方法来生成一个句法分析模型。一般来讲，一个句法分析模型包含了参数集和特征集。

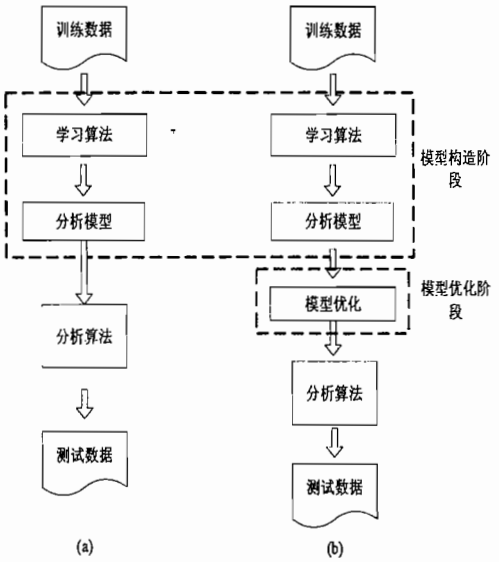


图1 两种类型的框架: (a) 传统建模方案; (b) 本文建模方案

句法分析模型的质量主要依赖于参数设置。因此，可以在模型特征集不变的基础上，考虑模型参数的优化来提高模型性能。图1 (b) 展示了句法分析模型构造的新方法。即首先通过传统方法获得模型，然后使用遗传算法对获得的模型进行进一步优化。

3 分析模型的设计

3.1 模型构造

为阐述MIRA学习算法原理，我们假定: (1) $\tau = \{(x_t, y_t)\}_{t=1}^{T}$ 为一个训练集, 包含一个句子 x_t 和它的依存表征 y_t 对组成，其中 y_t 是通过句子获得的依存树。(2) v 表示一个辅助的权重向量。(3) n 表示迭代的次数，t 表示是第几个句子。(4) $w^{(i)}$ 表示第 i^{th} 训练迭代之后的权重向量。(5) $s(x_t, y_t)$ 记录依存树的得分。(6) $best_k(x_t; w^{(i)})$ 是产生的依存树集中前 k 个最好的依存树。(7) $L(y_t, y')$ 记录产生的不正确依存树根节点个数。其中，y' 表示总的依存树，y_t 表示正确根节点的依存树。(8) w 表

示所有的权重向量的平均值。MIRA 学习算法的伪代码如下：

Training data: $\tau = \{(x_t, y_t)\}_{t=1}^{T}$

1. $w^{(0)} = 0; v = 0; i = 0$
2. for $n:1..N$
3. for $t:1..T$
4. $\min \| w^{(i+1)} - w^{(i)} \|$

   s.t. $s(x_t, y_t) - s(x_t, y') \geq L(y_t, y')$

   $\forall y' \in best_k(x_t; w^{(i)})$
5. $v = v + w^{(i+1)}$
6. $i = i + 1$
7. $w = v/(N * T)$

3.2 模型优化

优化问题是多个科学领域的共同存在的经典问题，遗传算法被认为是解决这些优化问题的一种有效方法[15]。遗传算法包括三个遗传操作：选择，杂交和变异。在句法分析模型中，模型包含参数的设置和特征说明。特征说明表示为一系列语言描述，而参数则表示为数字向量。本文使用遗传算法对模型的参数进行优化(即特征向量 w)，如图 2 所示。句法分析模型优化步骤如下：

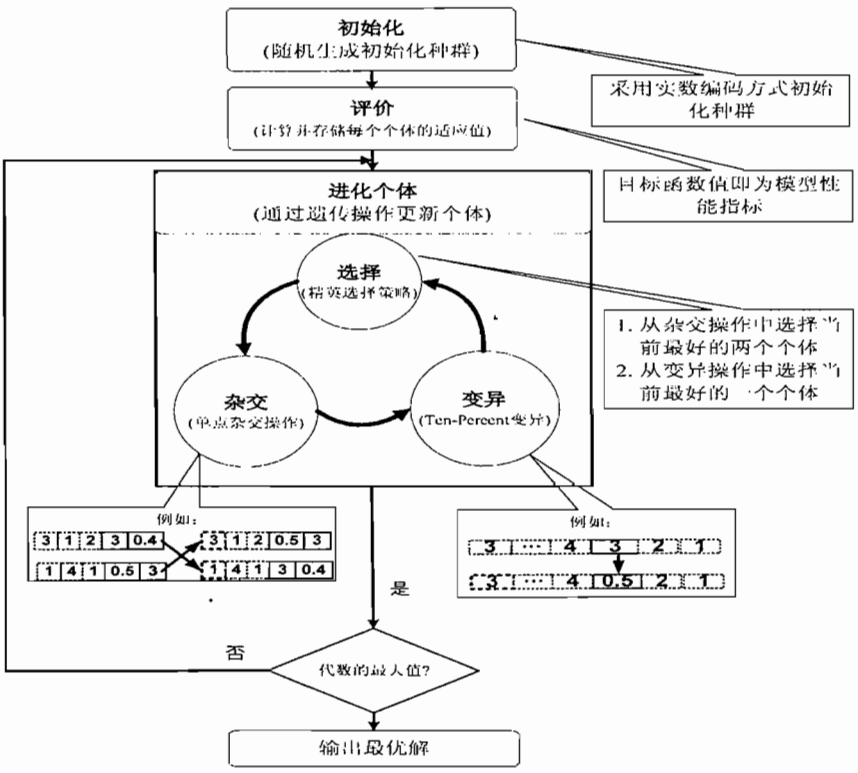


图 2　对于分析模型优化的 GA 流程图

步骤 1. *初始化*。种群的初始化是基于权重向量 w，它的获得是通过 MIRA 运行后的结果。例如，假设权重向量 $w = (w_1, w_2, ..., w_k, ...)$，那么一个个体 $w^{new} = (w_1^{new}, w_2^{new}, ..., w_k^{new}, ...)$ 的获得是通过如下的方式：每一个个体包含的单个向量 w_k^{new} 是从 $[w_k - 0.1w_k, w_k + 0.1w_k]$ 之间的范围随机生成

的值，即

$$w_1^{new} \in [w_1 - 0.1w_1, w_1 + 0.1w_1],$$
$$w_2^{new} \in [w_2 - 0.1w_2, w_2 + 0.1w_2],$$
$$\cdots\cdots,$$
$$w_k^{new} \in [w_k - 0.1w_k, w_k + 0.1w_k],$$
$$\cdots\cdots$$

步骤 2. *评估*。计算并存储每个个体的适应值。

步骤 3. *选择*。根据每个个体的适应值，将所有的个体按从大到小的顺序进行排序。

步骤 4. *杂交*。执行单点的杂交操作。为获得快速收敛速度，进行杂交的个体采用当前两个适应值最好的个体。

步骤 5. *变异*。执行变异的操作。类似于单点杂交，进行变异的个体是基于当前适应值最好的个体。

步骤 6. 重复步骤 3 到步骤 5，直到满足结束条件。

步骤 7. 输出最优的个体。

值得注意的是，这里我们使用了一些点的杂交取代了传统的单点杂交，这是因为一个个体是一个关联的比较大的数字(维数的大小依赖于 w)。图 3 描述了 GA 进行参数优化的个体数据结构。这里采用 *Accuracy* 作为 GA 优化句法分析模型的目标函数。*Accuracy* 是指在依存树中能够被正确识别词语所在语法树节点的父母个数。

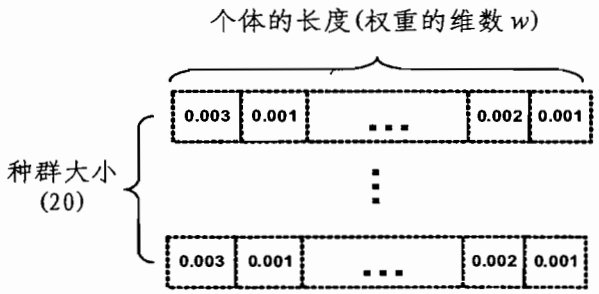


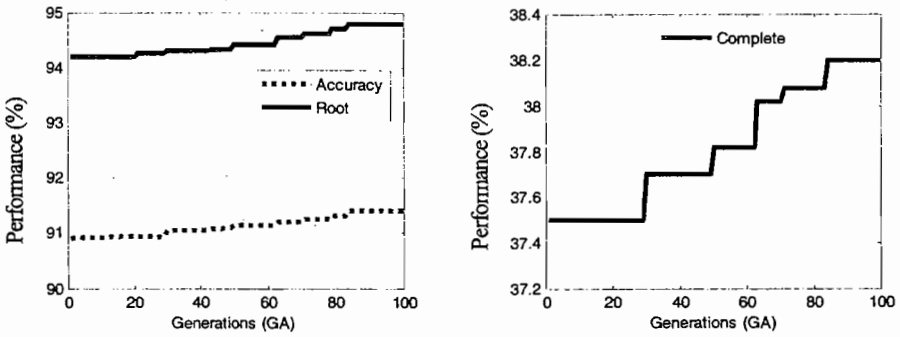
图 3　句法分析模型个体的数据结构

4　数值实验

句法分析模型的构造实验在国际标准数据集 English Penn Treebank[16]上进行. 实验数据包括 3 个部分：02-21 部分为训练集，22 部分为开发集，23 部分为测试集。模型构造阶段借助于 MIRA 来实现：实验中，POS 标注作为输入；依存结构由 Yamada 和 Matsumoto [6]的提炼规则产生，共包含了 6,998,447 特征；开发集和测试集的标注来源于 Ratnaparkhi [17]。在模型优化阶段，GA 被用来估算句法分析模型的参数。GA 的参数设置如下：迭代次数 100 代，种群大小 20，杂交率和变异概率设置分别设为 0.6 和 0.2。

图4刻画了GA迭代过程中，句法分析模型性能的优化过程。不难发现，随着迭代次数的增加，模型精度也不断提高。

表 1 总结了本文提出方法和一些常见依存分析方法的实验对比结果。不难发现，采用本文方法论构造的句法分析模型具有更好地性能。其中，*Root* 表示正确的根节点个数；*Complete* 是表示分析的依存树中，能够被完整分析出的依存树句子个数。



(a) accuracy 和 root　　(b) complete

图 4　GA 连续迭代中句法分析模型的性能

表 1　英语的句法分析模型性能对比

	Accuracy	Root	Complete
Y&M2003[7]	90.03	91.6	38.4
N&S2004[8]	87.3	84.3	30.4
Avg. Perceptron	90.6	94.0	36.5
MIRA[9]	90.9	94.2	37.5
Our Model	91.4	94.8	38.2

5　结论

本文提出一种基于 MIRA 和 GA 的句法分析模型构造方法。实验对比分析表明，和一些常见经典模型对比，本文的方法构造的句法分析模型具有更好的性能。

进一步研究思路如下：1.采用其他的方法（例如短语结构）代替 MIRA 进行模型构造；2.采用其他的演化优化技术来进行句法模型的优化。

参 考 文 献

[1] Lee-Feng Chien, Keh-Jiann Chen, and Lin-Shan Lee, "A Best-First Language Processing Model Integrating the Unification Grammar and Markov Language Model for Speech Recognition Applications", IEEE Trans. on Speech and Audio Processing, vol. 1, no. 2, pp.221-240, 1993.

[2] Sampo Pyysalo, Filip Ginter, Tapio Pahikkala, and Jorma Boberg, "Evaluation of Two Dependency Parsers on Biomedical Corpus Targeted at Protein-Protein interactions", International Journal of Medical Informatics, vol. 75, pp. 430-442, 2006.

[3] Gergely Korodi and Loan Tabus, "Compression of Annotated Nucleotide Sequences", IEEE/ACM Trans. on Computational Biology and Bioinformatics, vol. 4, no. 3, pp.447-457. 2007.

[4] Michael Collins, Head-Driven Statistical Models for Natural Language Parsing. Ph.D. thesis, University of Pennsylvania, 1999.

[5] Eugene Charniak, "A maximum-entropy-inspired parser", In Proc. NAACL, pp.132-139, 2000.

[6] Hiroyasu Yamada and Yuji Matsumoto, Statistical dependency analysis with support vector machines. In Proc. IWPT, 2003.

[7] Joakim Nivre and Mario Scholz, "Deterministic dependency parsing of english text", In Proc. COLING, 2004.

[8] Ryan McDonald, Koby Crammer, and Fernando Pereira, "Online large-margin training of dependency parsers", In Proceedings of ACL, 2005.

[9] Deyu Zhou and Yulan He, "Discriminative Training of the Hidden Vector State Model for Semantic Parsing", IEEE Trans. on Knowledge and Data Engineering, vol. 2, no. 1, pp.66-77, 2009.

[10] Lilja Øvrelid, Jonas Kuhn, and Kathrin Spreyer, "Improving data-driven dependency parsing using large-scale LFG grammars", In Proceedings of ACL-IJCNLP, pp.37-40, 2009.

[11] Chung-Hsien Wu, Chao-Hong Liu, Harris M., and Liang-Chih Yu, "Sentence Correction Incorporating Relative Position and Parse Template Language Models", IEEE Trans. on Audio, Speech, and Language Processing, vol. 18, no. 6, pp.1170-1181, 2010.

[12] Bharat R. Ambati, "Importance of Linguistic Constraints in Statistical Dependency Parsing", In Proceedings of the ACL, pp.103-108, 2010.

[13] Terry Koo and Michael Collins, "Efficient Third-order Dependency Parsers", In Proceedings of the ACL, pp.1-11, 2010.

[14] Ryan McDonald, "Discriminative Learning and Spanning Tree Algorithms for Dependency Parsing", Ph.D. thesis, University of Pennsylvania, 2006.

[15] David M. Golderg, "Genetic Algorithm in Search", Optimization & Machine Learning. Addison Wesley, 1989.

[16] Mitchell P. Marcus, Beatrice Santorini, and Mary A. Marcinkiewicz, "Building a Large Annotated Corpus of English: the Penn Treebank", Computational Linguistics, vol. 19, no. 2, pp.313-330, 1993.

[17] Adwait Ratnaparkhi, "A maximum entropy model for part-of-speech tagging", In Proceedings of EMNLP, pp.133-142, 1996.