基于最大间隔马尔可夫网模型的汉语分词方法

李月伦　常宝宝

北京大学计算语言学研究所 北京大学计算语言学教育部重点实验室 北京 100871

E-mail: lyldtc.student@sina.com, chbb@pku.edu.cn

摘　要: 分词是汉语自然语言处理研究中非常重要的一个环节，在早先的研究中，最大熵模型和条件随机场(CRF)模型已经广泛运用到汉语自动分词的工作中。最大间隔马尔可夫网（Max Margin Markov Networks，简称M3N）模型是近年来由B.Taskar等[1]人提出的一种新型结构学习模型。本文尝试将该模型用于汉语分词建模，通过一组实验证明基于最大间隔马尔科夫网模型的汉语分词方法可以取得较高的分词精度，是一种有效的汉语分词方法。

关键词: 最大间隔马尔可夫网模型 汉语分词 机器学习

Maximum Margin Markov Networks-Based Chinese Word Segmentation Method

LI Yuelun　CHANG Baobao

Institute of Computational Linguistics, Peking University, Beijing,

Key Laboratory of Computational Linguistics (Peking University), Ministry of Education, CHINA , 100871

E-mail: lyldtc.student@sina.com, chbb@pku.edu.cn

Abstract: Chinese Word Segmentation(CWS) is a crucial step in the study of Chinese Natural Language Processing. In previous researches, the Maximum Entropy model and Conditional Random Field(CRF) model have been widely used in the study of CWS. In recent years, B.Taskar[1] introduced a new method of structural study: Max Margin Markov Networks(M3N as abbreviation). In this paper, we will apply the M3N model to CWS. By doing some experiments, we can verify that the Maximum Margin Markov Networks-based Chinese Word Segmentation Method, which can achieve fairly high segmentation precision, is a very useful segmentation approach.

Keywords: Maximum Margin Markov Networks, Chinese Word Segmentation, Machine learning

1　引言

在汉语自然语言处理中，首要做的工作就是自动分词。长期以来，研究人员一直把未登录词和分词歧义并列为影响分词精度的两大因素[2]。汉语分词问题有很多不同的解决方法，根据这些方法是否使用了词典可将它们分为基于词典的分词方法与基于字标注的分词方法。

基于词典的分词方法需要一个提前编制好的词典，将文本中的字符串与词典中的词条进行比较，如果匹配成功，则划分出一个词。基于词典的分词方法需要引入专门的未登录词识别模块。

基于字标注的分词方法的典型代表是Xue[3]提出的分词方法。该方法对语料库中每个汉字进行标注，设置四种标志：LL，RR，MM和LR，分别表示词的左边界，右边界，中间部分以及单字词。该方法不需要预先编制的词典，但需要大量的被标注好的训练数据。

目前常用于分词的机器学习模型有最大熵模型[3]和条件随[1]机场模型[4, 5]。最大间隔马尔可夫网模型（Max Margin Markov Networks，简称M3N模型）是由B.Taskar等[1]提出的一种新型结构学习的方法，本文将M3N模型用于解决汉语分词问题，考察这一方法在分词问题中的可用性。

本文工作得到国家自然科学基金项目(60303003)；国家社会科学基金项目(06BYY048)的支持。

本文的组织方式如下：在第二节中，对 M3N 模型进行简单介绍。第三节是特征的提取方法。在第四节中，详细阐述每组实验的做法，报告结果并做一些相应的分析。

2 最大间隔马尔可夫网模型

与 CRF 模型类似，M3N 模型的基础也是无向图(马尔可夫网)模型，但该方法把马尔可夫网与最大间隔原则联系起来，是对马尔可夫网进行最大间隔训练的一种非概率化模型。

传统无向图模型可以有效地表示出序列标注间的相互关系，但与最大间隔模型相比，推广能力不佳，不支持核函数，因而无法用于高维特征空间。M3N 模型将马尔可夫网的学习过程视作寻找最大间隔决策函数的最优化问题，因而将最大间隔模型与无向图模型的优点结合在一起，在改善推广能力的同时也考虑到了序列标注间的相互联系。

在无向图中，我们在每条边上定义基函数（特征函数）$f(\mathrm{x}, y_i, y_j)$，其中$(i,j)\in E$，并且假定所有边同质，即每条边都具有相同的基函数。如果每条边都具有 n 个基函数，那么定义特征为 $f_k(\mathrm{x},\mathrm{y})=\sum_{(i,j)\in E} f_k(\mathrm{x}, y_i, y_j)$。其中，$k\in[1,n]$，即每个特征是每条边上相应基函数的和。

M3N 模型将序列标注问题视为如下的决策问题，其中，**x** 为观察序列，**y** 为标注序列。

$$h_{\mathbf{w}}(\mathrm{x})=\arg\max_{\mathrm{y}}\sum_{i=1}^{n} w_i f_i(\mathrm{x},\mathrm{y})=\arg\max_{\mathrm{y}} \mathrm{w}^{\mathrm{T}}\mathrm{f}(\mathrm{x},\mathrm{y})$$

M3N 模型的参数为权值向量 **w**，需要通过间隔最大化的原则求解这些参数。在 M3N 模型中，函数间隔定义为$\Delta\,\mathrm{f}_{\mathrm{x}}(\mathrm{y})=\mathrm{f}(\mathrm{x},\mathrm{t}(\mathrm{x}))-\mathrm{f}(\mathrm{x},\mathrm{y})$，其中，$\mathrm{t}(\mathrm{x})$为 x 的正确标注序列。

M3N 模型扩展了传统的$0-1$损失函数，使用逐标记损失函数$\Delta\,\mathrm{t}_{\mathrm{x}}(\mathrm{y})=\sum_{i=1}^{l}\Delta\,\mathrm{t}_{\mathrm{x}}(y_i)$，其中$\Delta\,\mathrm{t}_{\mathrm{x}}(y_i)\equiv \mathrm{I}(y_i\neq(\mathrm{t}(\mathrm{x}))_i)$，即$\Delta\,\mathrm{t}_{\mathrm{x}}(y_i)$是$y_i$的$0-1$损失。

逐标记损失函数的引入，使得$\mathrm{t}(\mathrm{x})$与 y 之间的间隔依赖于序列 y 的损失函数值，更好地照顾到序列标记的结构特性。按照最大间隔原则，M3N 模型参数训练所对应的原始优化问题和对偶优化问题分别如下：

原始优化问题

$$\min\ \frac{1}{2}\|\mathrm{w}\|^2+C\sum_{\mathrm{x}}\xi_{\mathrm{x}};$$

$$\text{s.t. } \mathrm{w}^{\mathrm{T}}\Delta\,\mathrm{f}_{\mathrm{x}}(\mathrm{y})\geqslant\Delta\,\mathrm{t}_{\mathrm{x}}(\mathrm{y})-\xi_{\mathrm{x}},\ \forall \mathrm{x},\mathrm{y}.$$

对偶优化问题

$$\max\ \sum_{\mathrm{x},\mathrm{y}}\alpha_{\mathrm{x}}(\mathrm{y})\,\Delta\,\mathrm{t}_{\mathrm{x}}(\mathrm{y})-\frac{1}{2}\left\|\sum_{\mathrm{x},\mathrm{y}}\alpha_{\mathrm{x}}(\mathrm{y})\,\Delta\,\mathrm{f}_{\mathrm{x}}(\mathrm{y})\right\|^2;$$

$$\text{s.t. } \sum_{\mathrm{y}}\alpha_{\mathrm{x}}(\mathrm{y})=C,\ \forall \mathrm{x};\ \alpha_{\mathrm{x}}(\mathrm{y})\geqslant 0,\ \forall \mathrm{x},\mathrm{y}.$$

3 文字标注及特征设置

3.1 用不同的标记集对文字进行标注

为了测试标记集对切分性能的性能，我们在实验中采用了下面三种标记集：

（1）S:单字词，B:词的左边界，M:词的中间部分，E:词的右边界

（2）B:词的左边界，I:词的后续部分

（3）I:词的前续部分，E:词的右边界

3.2 特征设置

相关工作表明，使用当前字前后各两个字的特征是比较理想的[2]。因此，特征模板设置如下：

基本特征：(a)C_{-2} (b)C_{-1} (c)C_0 (d)C_1 (e)C_2 (f)$C_{-1}C_0$ (g)C_0C_1 (h)$C_{-2}C_{-1}$ (i)C_1C_2 (j)$C_{-1}C_1$

可选特征组一：(k)$C_{-1}C_0T_{-1}$ (l)$C_{-2}C_{-1}C_0T_{-2}T_{-1}$

可选特征组二：(m)T_{-1} (n)T_{-2}

可选特征组三：(o)T_{-2} T_{-1}

其中，C_0表示当前字符，C_{-2}、C_{-1} 、C_1、C_2分别表示当前字符左邻第二、第一，右邻第一、第二个字符，T_{-2}、T_{-1}分别表示当前字符左邻第二、第一个字符的标记。

从特征模板提供的信息来看，第一组特征为基本特征，为每次实验的必选特征。第二组特征～第四组特征均为可选特征，我们将测试这些可选特征对分词性能的影响。

4 实验

我们选择北京大学计算语言所所开发的 1998 年 1 月份《人民日报》的部分语料作为训练语料和测试语料。该语料共有 57 万词左右，实验 4.5 中 1：50 的实验所用的训练语料为全部训练语料。语料中包含切分和词性标注信息，均经过人工校对。该语料中包含新闻题材、文学题材等。

我们一共进行了五组实验，分别验证不同的机器学习模型、不同的标记集、不同的特征模板集、不同的训练迭代次数以及训练不同规模的语料对汉语分词效果的影响。

4.1 实验一：用不同的机器学习模型对语料进行训练

在这个实验中，我们分别用最大熵模型，CRF 模型和 M3N 模型对相同的语料进行训练和测试，比较不同机器学习模型性能在词语切分性能方面的差异。

考虑到训练效率的问题，选择 20 万词左右的语料作为训练语料。为了保证训练语料与测试语料划分的合理性，我们把全部语料中的句子以每 50 句为单位划分，其中第一句做测试语料，第 2 至 21 句做训练语料。训练语料共有 22.5 万词，测试语料 1.1 万词。在测试语料中，共有未登录词 568 个，占测试语料的 5%。使用特征模板 a～n，SBME 标记，实验结果如表 1 所示

	准确率	召回率	F 值	OOV_recall
最大熵模型	0.914618	0.914291	0.914454	0.651408
CRF 模型	0.943211	0.938154	0.940676	0.748239
M3N 模型	0.950420	0.950846	0.950632	0.727113

表 1

可以看出，在分词方面，结构化模型要明显好于非结构化模型，因为结构化模型考虑了句子内部标记之间的相关性。而最大熵模型因为没有考虑到这一点，因而产生了“标记偏见”问题[4]。从CRF和M3N的结果可以看出，M3N得到的F值比CRF高 1%，但在OOV_recall方面却有下降。总的来说，将M3N模型运用到汉语分词中取得了不错的效果，基于M3N模型的性能略优于CRF模型。

4.2 实验二：用不同的标记集对语料进行训练

在该实验中，使用与实验一相同的训练语料、测试语料及特征模板，基于 M3N 模型进行训练和测试。使用三种不同的标记集对训练语料进行标记，如 3.1 节所示。实验结果如表 2 所示：

实验名称	准确率	召回率	F 值	OOV_recall
SBME 标记	0.950420	0.950846	0.950632	0.727113
BI 标记	0.939311	0.947538	0.943406	0.690141
IE 标记	0.935909	0.946113	0.940983	0.640845

表 2

从实验结果可以看出，三个实验的结果依次递减，说明 SBME 标记集使分词结果达到最好。

4.3 实验三：用不同的特征模板集对语料进行训练

在这个实验中，我们使用的语料与实验一相同，使用 SBME 标记，基于 M3N 模型进行训练和测试。在特征模板方面，我们要对 3.2 节所述特征模板集中那些效果相似的模板进行排列组合，与必须包括的模板 a～j 合并构成若干个特征模板集，找出效果最好的特征模板集合，并对其中的模板进行一些分析。具体而言，在每次实验中，我们都必须要得到当前字符左邻第一个字以及第二个字的标注信息，要得到该信息可以对特征模板 k～o 进行如下组合，实验结果如表 3 所示。

（1）l（11_1）　（2）o（11_2）　（3）m，n（12_1）　（4）k，l（12_2）

（6）m，o（12_4）　（5）l，n（12_3）　（7）k，l，o（13_1）

（8）k，l，m，n（14_1）　（9）k，l，m，o（14_2）　（10）k，l，m，n，o（15_1）

实验名称	模板	准确率	召回率	F 值	OOV_recall
11_1	abcdefghijl	0.933127	0.942801	0.937939	0.656690
11_2	abcdefghijo	0.943723	0.948700	0.946205	0.704225
12_1	abcdefghijmn	0.943467	0.948610	0.946031	0.711268
12_2	abcdefghijkl	0.937318	0.947538	0.942400	0.651408
12_3	abcdefghijln	0.944970	0.946912	0.945940	0.723591
12_4	abcdefghijmo	0.937544	0.948436	0.942959	0.688380
13_1	abcdefghijklo	0.948619	0.948789	0.948704	0.732394
14_1	**abcdefghijklmn**	**0.947988**	**0.951291**	**0.949636**	**0.718310**
14_2	abcdefghijklmo	0.946023	0.949236	0.947627	0.718310
15_1	abcdefghijklmno	0.944701	0.951292	0.94794	0.705986

表 3

实验 14_1 得到了最高的 F 值。通过不同实验之间的对比，可大致总结出如下结论：

（1）11_1 与 11_2，12_1 与 12_2 比较均表明：在 OOV_recall 方面，m，o 似比 k，l 有效。

（2）11_1 与 12_2 表明：特征模板 l 易造成 OOV_recall 降低。

（3）11_2 与 12_1 比较表明：特征模板 o 与 m，n 相比，分别对 F 值与 OOV_recall 有贡献。

（4）13_1 与 14_1 比较表明：F 值越高并不意味 OOV_recall 也越高。

（5）整体来说，在特征模板给定合理的情况下，特征模板的数量越多，训练效果越好。

（6）每个特征模板都有存在的必要，缺少一个都会使 F 值下降。

（7）特征模板 o 对 OOV_recall 的贡献比较大。

4.4 实验四：用不同的迭代次数对语料进行训练

在这个实验中，我们使用实验一中的语料，SBME 标记，特征模板 a～n，在 M3N 模型上进行训练和测试，通过进行不同迭代次数的实验检验其对分词结果的影响。实验结果如表 4 所示：

迭代次数	10	15	20	21	22	25	30
准确率	0.947988	0.949317	0.950420	0.950973	0.950063	0.950139	0.949665
召回率	0.951291	0.950846	0.950844	0.951653	0.950402	0.950223	0.949240
F 值	0.949636	0.950080	0.950632	0.951313	0.950232	0.950181	0.949453
OOV_recall	0.718310	0.732394	0.721713	0.732394	0.728873	0.728873	0.730634

表 4

为了更直观地对实验结果进行分析，得到该结果的线状图如图 1 所示：

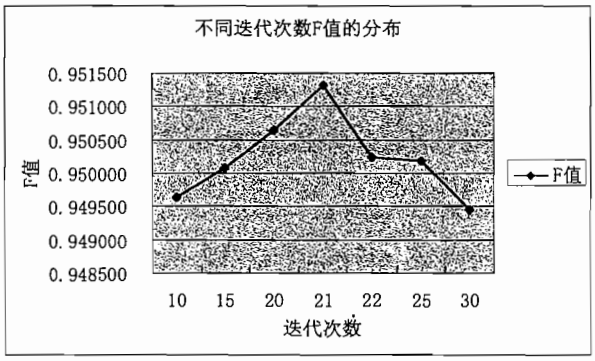

图 1

图 1 表明，F 值不随迭代次数的增加而增加，F 值在 21 次迭代处达到最大值。表 4 表明，OOV_recall 在 15 和 21 次迭代处达到最大值，即 OOV_recall 不随 F 值的增大而增大。此外，随着迭代次数的不断增加，训练时间增加的幅度也比较大，但时间的代价并没有得到效果的提升。

4.5 实验五：用不同规模的语料进行训练

在该实验中，使用 SBME 标记，特征模板 a～n，在 M3N 模型上进行训练和测试。测试语料保持不变（11191 词），每次增加 5%的训练语料（在实验（1：50）中增加了 20%的训练数据）。

得到的训练语料相关信息如表 5 所示，实验结果如表 6 所示。

	1:5	1:10	1:15	1:20	1:25	1:30	1:50
训练词数	56682	111079	166626	225405	281117	336897	561737
未登录词数	1200	761	563	568	339	286	71
oov_rate	0.107229	0.068001	0.050308	0.050755	0.030292	0.025556	0.000126

表 5

实验名称	1:5	1:10	1:15	1:20	1:25	1:30	1:50
准确率	0.912812	0.941839	0.952991	0.947988	0.968714	0.972118	0.992763
召回率	0.904736	0.942175	0.951202	0.951291	0.965770	0.972205	0.992940
F 值	0.908756	0.942007	0.952096	0.949636	0.967240	0.972161	0.992851
oov_recall	0.709167	0.734560	0.721137	0.718310	0.728614	0.713287	0.718310

表 6

在实验 1:50 中,F 值达到了 0.992851,这是因为在该实验中,OOV_rate 非常小,只有 0.000126(71 个词)。为了更直观地对实验结果进行分析,得到该结果的线状图如图 2 所示:

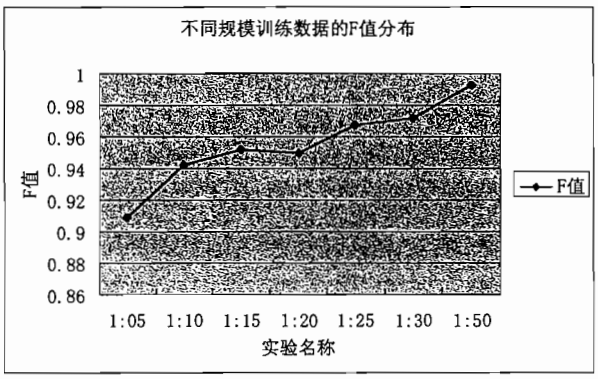

图 2

从图中可以看出,当训练数据集合较小时,F值较低。F值大体上随着训练语料规模的增大而增大,但实验 1:15 得出的结果比实验 1:20 得到的结果略高。从表 6 中可以看出,OOV_recall 在整个实验中变化幅度不大。该实验同时说明,字符标注的方法比较好地解决了未登录词的识别问题,且解决能力在训练集较小时也很有效。这是因为字标注的过程能够平衡地看待词表词和未登录词的识别问题[2]。

5 结束语

M3N 模型是一种将最大间隔原则与无向图模型结合所提出的一种结构学习模型,本文将该模型用于汉语分词任务并进行了相关实验验证,总体来看,该模型的分词性能较好,与基于 CRF 模型的分词系统性能具有可比性,甚至略为领先。不过,与 CRF 模型类似,M3N 模型训练的时间代价也是相当大的,训练算法的优化应是结构化学习模型需要关心的问题。此外,M3N 模型支持核函数,我们也将在未来的工作中考察核函数的引入是否会引起分词性能的改善。

参 考 文 献

[1] Ben Taskar, Carlos Guestrin and Daphne Koller, "Max-Margin Markov Networks", *Proceedings of Neural Information Processing Systems Conference(NIPS 2003)*

[2] 黄昌宁,赵海,"中文分词十年回顾",中文信息学报,第 21 卷,第 3 期,2007 年 5 月

[3] N. Xue, "Chinese Word Segmentation as Character Tagging", *Computational Linguistics and Chinese Language Processing*. 8(1), pp. 29-48, 2003.

[4] 李双龙,刘群,王成耀,"基于条件随机场的汉语分词系统",《微计算机信息》,2006 年第 22 卷第 10-1 期

[5] 迟程英,于长远,战学刚,"基于条件随机场的中文分词方法",《情报杂志》,2008 年第 5 期

[6] Huang Chu-Ren, Yo Ting-Shuo, Petr Simon and Hsieh Shu-Kai, "A Realistic and Robust Model for Chinese Word Segmentation", Proceedings of the 20th Conference on Computational Linguistics and Speech Processing(ROCLING 2008)

[7] 孙茂松,肖明,邹嘉彦,"基于无指导学习策略的无词表条件下的汉语自动分词",计算机学报,第 27 卷,第 6 期,2004 年 6 月