

文章编号:1001-9081(2010)08-2038-03

基于条件随机场的蒙古语词性标注方法

应玉龙^{1,2}, 李 森¹, 乌达巴拉¹, 朱 海^{1,2}

(1. 中国科学院 合肥智能机械研究所, 合肥 230031; 2. 中国科学技术大学 自动化系, 合肥 230027)

(yingyul@mail.ustc.edu.cn; mli@iim.ac.cn)

摘 要: 为了保留蒙古语词缀中大量的语法、语义信息和缩小蒙古语词典的规模, 蒙古语词性标注需要对词干和词缀都进行词性标注。针对这一问题提出了一种基于条件随机场(CRF)的蒙古语词性标注方法。该方法利用 CRF 模型能够添加任意特征的特点, 充分使用蒙文上下文信息, 针对词素之间的相互影响添加了新的统计特征, 并在 3.8 万句的蒙古语词性标注语料上进行了封闭测试, 该方法的标注准确率达到了 96.65%, 优于使用隐马尔可夫模型(HMM)的词性标注模型。

关键词: 词干; 词缀; 条件随机场; 词性标注; 词素

中图分类号: TP391.2 **文献标志码:** A

Mongolian part-of-speech tagging approach based on conditional random fields

YING Yu-long^{1,2}, LI Miao¹, Wudabala¹, ZHU Hai^{1,2}

(1. Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei Anhui 230031, China;

2. Department of Automation, University of Science and Technology of China, Hefei Anhui 230027, China)

Abstract: It is necessary to tag both stem and affix in the Mongolian part of speech tagging, in order to save lots of syntax and semantic information of affix and to reduce the size of Mongolian dictionary. This paper presented a new approach of Mongolian part of speech tagging based on CRF. To take advantage of the ability of using arbitrary features as input in CRF, the system exploited not only the contexts of words, but also new statistical features adopted for mutual influence between the morphemes. The system was tested in the 38000 part-of-speech dataset provided by Inner Mongolia University. The closed test results show that POS tagging accuracy of the testing set reaches 96.65%, outperforming the HMM-based model.

Key words: stem; affix; Conditional Random Field(CRF); part-of-speech tagging; morpheme

0 引言

词性标注^[1]是词法分析的一个重要部分,是自然语言信息处理领域的基础性研究课题之一。它的作用就是通过采用适当的方法,根据上下文语境关系,消除句子中词的兼类,使得无论一个词兼有几种词性,在特定的场合下只保留其中最合适的一种。其标注效果将直接影响到后续的句法分析、语义分析、信息检索、机器翻译等诸多领域的研究,因此,一直引起人们的高度关注。

蒙古语的词性标注是蒙古语信息处理工作的基础。蒙古语属于黏着语,其构词、构形都通过在词干后面不断缀接不同的词尾来实现^[2],其词法形态变化非常丰富。蒙古语构形的附加成分承载着大量的语法信息,如果把蒙古语词只作为一个整体来处理,就损失了大量的语法、语义信息。而且由于收录词的形态变化而派生的新词,让蒙古语词典的规模变得非常庞大。例如,在词典中,如果要列出“IR_E”的所有形式,那可能要列出几百种变化。因此,在传统的语言学词典中只收录一个条目,即“IREHU”这个原始现在形动词形式。在蒙古语词性标注过程中,不但要对蒙古语的词干进行词性标注,还要对各个附加成分进行标注,主要采用规则方法^[3]和基于隐马尔可夫模型(Hidden Markov Model, HMM)的方法。

在采用规则方法处理时,由于蒙古语词典中收录到的表面词形有限,因此需要规模庞大的规则库。而规则库的构造需要考虑两个基本问题:规则对语言现象的覆盖率和规则处理的正确率。对于一条规则,这两种性能往往显示反比关系。而且规则方法带有很大的主观性,难以保证规则的一致性,适应性较差,处理歧义长句、未登录词、不规范句子的能力非常弱,词性标注准确率不高。目前基于规则的蒙古语切分标注准确率仅有 72.6%^[4]。

在基于 HMM 的词性标注方法中,假设当前词的词性只与其前面 n 个词的信息有关,而与其后面的词无关。该假设在蒙古语词性标注任务中并不符合上下文关系的实际情况。例如,“HELE”是兼类词,它既可以是名词也可以是动词,在“CITEGUN-DU HELE BICIG JIGAJV BAYN_A UU?”中为名词;“CI TEGUN-DU HELE!”中为动词。而这两句中“HELE”的词性是无法通过其前面的词“TEGUN-DU”来判断的。而且 HMM 不能包含有效的远距离特征。因此,处理歧义长句、未登录词的标注能力仍然比较薄弱。

条件随机场模型(Conditional Random Fields, CRF)是非常优秀的序列标注模型,它克服了传统的 HMM 的独立性假设及最大熵模型的标记偏置等缺陷。因此针对蒙古语的语言特点,本文提出了一种基于 CRF 的蒙文词性标注方法,结合

收稿日期:2010-02-03。 基金项目:中国科学院知识创新工程项目(KGCX2-SW-511)。

作者简介: 应玉龙(1984-),男,安徽界首人,硕士研究生,主要研究方向:自然语言处理、统计机器翻译; 李森(1955-),女,教授,博士生导师,主要研究方向:人工智能、农业知识工程; 乌达巴拉(1981-),女,助理研究员,主要研究方向:自然语言处理; 朱海(1984-),男,硕士研究生,主要研究方向:统计机器翻译。

蒙古语词素之间的影响关系选择合适的特征模板,这样不仅可以利用更多的上下文信息,而且可以缓解统计模型面临的数据稀疏问题。

1 CRF 模型及其训练方法

Lafferty 等人首先将 CRF 模型引入到自然语言处理的序列标记学习任务中^[5-6],其核心思想是利用无向图理论使序列标注的结果达到在整个序列上全局最优。

1.1 CRF 模型定义

CRF 是一个无向图上概率分布的学习框架。常用的一类 CRF 是线性链 CRF,即模型中各个节点之间构成线性结构。一个线性的 CRF 对应于一个有限状态机,非常适用于拉丁蒙文词素序列的词性标注学习。给定数据序列随机变量 X , CRF 定义了标注结果序列随机变量 Y 的条件概率分布 $P(Y|X)$,它通过训练学习来使得条件概率 $P(Y|X)$ 最大。记 $X = (x_1 x_2 \cdots x_n)$ 为待标记的序列(即:拉丁蒙文的词素), $Y = (y_1 y_2 \cdots y_n)$ 为对应的标记结果序列(即:词素的词性标记),并且 X 和 Y 的长度相同。例如: X 可以表示一个拉丁蒙文句子 $X = (\text{HOYAR}, \text{GVRBA}, +, \text{N}, \text{MINUIt}, \text{BURI}, +, \text{BEL}, \text{VVGCV}, +, \text{JV}, \text{BOL}, +, 0, +, \text{N_A}, \dots)$, Y 则表示该句子中每个词素的词性序列 $Y = (\text{M}, \text{M}, \text{Fc}, \text{Qc}, \text{Ve}, \text{Fn}, \text{Ve}, \text{Fn}, \text{Vz}, \text{Zv}, \text{Fs}, \text{Wp})$ 。线性链 CRF 对一个给定序列 X 的标注,其概率定义为:

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_j \lambda_j F_j(Y, X)\right)$$

$$Z(X) = \sum_Y \exp\left(\sum_j \lambda_j F_j(Y, X)\right)$$

$$F_j(Y, X) = \sum_i f_j(y_{i-1}, y_i, X, i)$$

其中: Y 是串的标注序列, X 是待标记的字符, f_j 是特征函数, λ_j 是对应的特征函数的权值,而 i 是位置标记, $Z(X)$ 是归一化因子,使得上式成为概率分布。

1.2 CRF 模型训练

建立 CRF 模型的主要任务就是从样本数据中估计得到特征权重 λ ,参数估计使用最大似然估计。在训练集 $T = \{X^k, Y^k\}$ 中,最大似然参数估计就是假设 $P(Y|X, \lambda)$ 为 λ 的函数,使 $P(Y|X, \lambda)$ 的对数值最大的 λ 为估计值,其似然值为:

$$L_\lambda = \sum_T \log P(Y^k | X^k, \lambda) = \sum_T \log \frac{1}{Z(X^k)} \exp\left(\sum_j \lambda_j F_j(Y^k, X^k)\right) = \sum_T \left(\sum_j \lambda_j F_j(Y^k, X^k) - \log Z(X^k)\right)$$

$$\text{最大值: } \lambda^* = \operatorname{argmax}_\lambda \sum_T \log P(Y^k | X^k, \lambda)$$

由于 L_λ 为凸函数,导数为零的点即为最值点。故对 λ 求导,则偏导数公式为:

$$\frac{\partial L_\lambda}{\partial \lambda_j} = \sum_T \left(\sum_i F_j(Y^k, X^k) - E_{P(Y|X^k)}[F_j(Y, X^k)]\right)$$

可简写为 $\frac{\partial L_\lambda}{\partial \lambda_j} = O_j - E_j = 0$ 。其中: O_j 为 λ_j 在训练集 T 中出现的频率, $E_j = \sum_T E_{P(Y|X^k)}[F_j(Y, X^k)]$ 是 λ_j 在条件随机场模型分布中的特征期望, E_j 使用前向—后向算法(Forward-Backward)求解。

直接使用最大似然估计,可能会发生过度学习问题,可通过引入惩罚函数的方法解决这一问题。例如使用惩罚项

$$\sum_j \frac{\lambda_j^2}{2\sigma^2}, \text{则原式变为 } L_{\lambda'} = L_\lambda - \frac{\sum_j \lambda_j^2}{2\sigma^2} + \text{const}, \text{其导数式变为}$$

$$\frac{\partial L_{\lambda'}}{\partial \lambda_j} = \frac{\partial L_\lambda}{\partial \lambda_j} - \frac{\lambda_j}{\sigma^2}。 \text{于是 } \lambda \text{ 的参数估计问题使用有限记忆 BFGS 算法 (Limited memory Broyden-Fletcher-Goldfarb-Shanno, L-BFGS) 算法计算。}$$

1.3 CRF 模型的序列标记

建立统计模型后,求解序列标记就是求得 Y^* 满足 $P(Y|X, \lambda)$ 最大, $Z(X)$ 与 Y 无关,所以 Y^* 可表示如下式,使用 Viterbi(维特比)算法求最优解 Y^* 。

$$Y^* = \operatorname{argmax}_Y P(Y|X) =$$

$$\operatorname{argmax}_Y \frac{1}{Z(X)} \exp\left(\sum_i \sum_j \lambda_j f_j(y_{i-1}, y_i, X, i)\right) =$$

$$\operatorname{argmax}_Y \sum_i \sum_j \lambda_j f_j(y_{i-1}, y_i, X, i)$$

2 基于 CRF 的词素标注模型

使用 CRF 进行蒙文词素(词干和词缀)词性标注的过程就是给定蒙文句子 $X = (x_1 x_2 \cdots x_n)$,通过 Viterbi 算法找出其对应的词性标注序列 $Y = (y_1 y_2 \cdots y_n)$,使得条件概率 $P(Y|X, \lambda)$ 最大。

2.1 系统实现

整个蒙文词素标注模型的识别流程如图 1 所示。

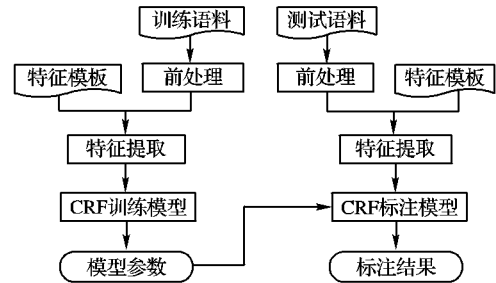


图1 词性标注模型

主要分为两步:

步骤1 模型参数训练,首先对带标注的蒙文语料(如1所示)进行预处理(如2所示),然后把处理结果和特征模板输入模型中,进行参数训练。

1) 语料: YOS0: Ne GORIM-DV: Ne-Fc MONGGO: Ne HEREGLE + HU: Ve + Ft BISI: Su. / Wp;

2) 预处理后: YOS0: Ne GORIM: Ne -DV: Fc MONGGO: Ne HEREGLE: Ve HU: Ft BISI: Su. : Wp。

步骤2 标注测试语料,首先对测试语料进行预处理(如3所示),然后把处理结果输入模型中,进行词性标注(如4所示)。

3) 测试语料: KLARA BOL SONI-BER AJILLA DAG HOMON;

4) 词性标注结果: KLARA: Nt BOL: Sd SONI: T -BER: Fc AJILLA: Ve DAG: Ft HOMON: Ne. : Wp。

2.2 标注集合

在使用 CRF 模型对词素序列进行词性标记时,首先要定义合适的标注集合,标记集确定了词性标记的标准。通用的

标记标准,对自然语言处理非常有必要,有助于不同研究者之间的交流和用户的使用,有助于资源更为合理的使用。我们采用的是较通用的内蒙古大学的词性标注集^[7]。

2.3 特征选择

使用 CRF 进行标注时,为了充分利用训练集的统计信息和蒙文的构词特点,非常重要的一步是选择合适的特征集。原则上是选择的特征越多越好,但是特征过多又会产生冗余信息,反而降低识别精度。本文选择两类特征:原子特征和复合特征。

1) 原子特征。主要考虑词素本身包含的信息、上下文信息对词性标注的影响。词素本身包含的信息非常丰富,而且是最容易得到的,因此是必不可少的一类特征。

表1 原子特征表

原子特征	意义
$C(n)$	位置 n 的词素信息, n 表示相对于当前位置词素的变量,取值为: $-2, -1, 0, 1$ 。 $n = 0$ 表示当前位置, $n = -2$ 表示当前位置之前第二个位置, $n = -1$ 表示当前位置的前一位置, $n = 1$ 表示当前位置的后一位置

2) 复合特征。在真实文本中,影响分词的因素往往不止一类,因而需要考虑多个因素,才能很好地反映实际情况。对原子特征模板进行适当的组合,得到复合特征模板。

表2 复合特征模板

复合特征	意义
$C(-2)C(0)$	当前词素与前第二个词素的组合
$C(-1)C(0)$	当前词素与前一个词素的组合
$C(0)C(1)$	当前词素与后一个词素的组合

3 实验与分析

3.1 评测标准与语料

本文词性标注实验使用的语料是内蒙古大学提供的 3.8 万句词性标注语料,含有约 52 万词素。选择其中的 3.35 万句作为训练语料,含有约 45 万词素;选择剩余的 0.45 万句作为测试语料,含有约 7 万词素。词性标注测试一般分为封闭测试和开放测试。由于封闭测试只允许从同组的训练语料中获取词性标注知识,可以对词性标注技术本身作出有效的评价,因此,本文在封闭测试条件下进行对比实验。

3.2 实验设计

我们设计了 3 组实验:为了和 CRF 标注模型对比,实验 1 采用 HMM 模型;实验 2 采用 CRF 模型,根据蒙文语言特点,设计的特征模板为: $C(-1), C(0), C(1), C(-1)C(0), C(0)C(1)$; 实验 3 也采用 CRF 模型,特征模板设计为 $C(-2), C(-1), C(0), C(1)C(-2)C(0), C(-1)C(0), C(0)C(1)$ 。实验结果如表 3 所示。

表3 实验结果

分组	采用模型	准确率
实验1	HMM	0.9417
实验2	CRF	0.9665
实验3	CRF	0.9655

通过实验 1 和实验 2、3 的标注结果对比,使用 CRF 的词

性标注准确率要优于使用 HMM 的词性标注模型。这是由于 CRF 最大的优点之一就是它能够加入任意的特征作为输入,相对于 HMM 只能利用中心词素的前 n 个词素作为上下文信息的弱点,CRF 能够同时使用中心词素的前 n 个词素和后 m 个词素作为该词素的上下文信息。这样,中心词素的词性不仅与它前面的词素有关,还与它后面的词素有关,更加符合实际情况。

相对于实验 2,虽然实验 3 使用了更复杂的特征模板,但是前者的词性标注结果要优于后者。对 3.8 万句词性标注语料进行了蒙文词语包含词素个数的频率统计,如表 4 所示。

表4 蒙文语料词长(包含词干和词缀的个数)频率分布

词长	1	2	3	4	≥ 5
频率	0.591	0.335	0.068	0.005	0.001

可以看出由 3 个及 3 个以下词素构成的蒙文词语占到了总词数的 99.46%,三词素窗口已包含了足够的上下文信息。实验 2 的特征模板只考虑了当前词素、及其前一个和后一个词素,为三词素窗口。而实验 3 的特征增加考虑了当前词素的前面第二个词素的特征,为四词素窗口,但是四词素长度的蒙文词语仅占到总词数的 0.5%,而使标注模型增加了大量的特征信息。原则上是选择的特征越多越好,但是特征过多又产生了冗余信息,反而降低了拉丁蒙文词性标注的精度,因此实验 2 的词性标注准确率要优于实验 3。

4 结语

本文提出了基于 CRF 的蒙古语词素的词性标注模型。针对蒙古语的构词、构形都是通过词干后面不断缀接不同词尾来实现的语言特点,并且根据对蒙古语词长频率的统计,设计了适合蒙古语词性标注的特征模板。该方法能够更充分地利用蒙古语的上下文信息,避免了标记偏置问题,得到全局最优的词性标注结果,有效提高了词性标注的性能。

参考文献:

- [1] 宗成庆. 统计自然语言处理[M]. 北京: 清华大学出版社, 2008.
- [2] 那顺乌日图, 雪艳, 叶嘉明. 现代蒙古语语料库加工技术的新进展——新一代蒙古语词语自动切分与标注系统[C]// 第十届全国少数民族语言文字信息处理学术研讨会. 西宁: [s.n], 2005: 122-127.
- [3] 叶嘉明. 基于规则的蒙古语词法分析研究与实现[D]. 北京: 北京大学, 2005.
- [4] 雪艳. 汉蒙词语对齐及相关技术研究[D]. 呼和浩特: 内蒙古大学, 2009.
- [5] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the 18th International Conference on Machine Learning. Washington, DC: IEEE, 2001: 282-289.
- [6] SUTTON C, MCCALLUM A, ROHANIMANESH K. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data[J]. The Journal of Machine Learning Research, 2007, 8: 693-723.
- [7] 图格木勒. 蒙古语语言资源库建设相关技术研究[D]. 呼和浩特: 内蒙古大学, 2007.