

基于 NAA 的词性自动标注模型¹

朱靖波 姚天顺²

东北大学信息学院计算机系 沈阳 110006

【摘要】本文提出了一种基于 NA 假设的词性自动标注方法。该方法采用基于 NA 假设自动从无标注语料库中抽取词性三元组数据，训练词性标注统计模型所需参数，对稀疏数据进行平滑处理。对词典中未登录词的词性进行猜测，根据未登录词的上下文评估各种词性的概率，最终选取最大概率词性作为未登录词词性。两万词次的开放性测试，三个模型的测试结果的准确率分别为 80.2%，93.1%和 85.4%。

关键词：词性标注，语言模型，NA 假设

POS Tagging Model Based on NAA

Zhu Jingbo Yao Tianshun

Dept. of Computer Science and Engineering of Northeastern University

Shenyang, Liaoning 110006

Abstract: This paper presents an approach to POS tagging based on raw corpus. This method adapts unsupervised learning technique to get a large amount of trigram data from a raw corpus. The model assumes that an unknown word has all grammatical categories, and its POS is determined by its context. The open test set consists of twenty thousand words. The precision of result is 80.2%, 93.1% and 85.4%.

Key words: POS tagging, language model, NA Assumption

一、前言

词性兼类现象是英汉机器翻译中典型的歧义问题。词性自动标注方法可以分为基于规则的方法和基于统计的方法两种。早期的自动标注方法多为基于规则的方法。七十年代 Greene 和 Robin (1989) 出于语言学的目的设计了一个用 3300 多条规则，86 个标记的标记集对 Brown 语料库进行标注的系统 Taggit，能够达到 77% 的正确率。国内也有许多学者设计了基于规则的词性标注系统。但是基于规则的方法，由于规则很多且杂，它的

¹ 本文获得国家自然科学基金和国家教委博士点基金资助

² 朱靖波，博士，讲师，jibzhu@ies.cs.neu.edu.cn；姚天顺，教授，博士生导师，tspiao@mail.neu.edu.cn

编写和维护也是一个很大的问题。E.Brill (1992) 提出了一种基于转换 (Transformation-based) 的错误驱动学习机制, 从带标语料库中自动获取转换规则用于词性自动标注, 实验结果表明该方法可以用较小的训练集获得较高的分析准确率。词性标注封闭式实验结果达到了 97-99% 准确率。八十年代末 DeRose (1988) 等用二元语法 (Bi-gram) 统计模型, 一下子把对 Brown 语料库的标注精度提到到 96% 以上。De Marcken (1990) 把这一技术引入句法分析器中用来消除歧义, 实验结果表明取得了相当满意的结果。国内清华大学白栓虎 (1991) 等采用二元语法 (Bi-gram) 统计模型, 同时对词典中未登录的词和一些词条未登录的词性 (所谓词典空缺) 进行预测, 封闭测试结果达到了 97% 以上的精确率。周强 (1993) 采用分词与词性标注相结合的策略, 利用规则方法和统计方法相结合的消歧策略, 测试结果达到了 94-96%。

实际上目前很多基于统计的词性自动标注方法取得了很好的实验结果, 很大程度依赖于词性带标语料库构造的代价。为了减轻人工标注训练语料库的瓶颈问题, 如何从更少标注代价的语料库中获取更多、更深层次知识的获取技术研究越来越有意义。本文提出了一种基于 NAA 的词性自动标注方法, 下文将详细介绍带标数据的自动抽取过程、词性自动标注模型和实验结果。

二、基于 NAA 的带标数据自动抽取

2.1 Nonambiguity-Ambiguity 假设

下面介绍 Nonambiguity-Ambiguity 假设 (Nonambiguity-Ambiguity Assumption, 简称 NAA)。为了论述方便, 定义: Ω 为语言现象集合 (真实语料); Ω_A 为具有歧义的语言现象集合 (消歧语料); Ω_N 为无歧义的语言现象集合 (训练语料)。很明显: $\Omega_A \cup \Omega_N = \Omega$ 。

NA 假设 (NAA): 利用 Ω_N 训练语言模型的参数, 完成对 Ω_A 的消歧处理是可行的。

为了从无标注语料库中自动抽取带词性标注的训练数据, 本文定义语料库中具有如下特征之一的词汇为无歧义语言现象, 反之为具有歧义的语言现象。四个特征描述如下:

- (1) 词典中只具有单一词性的词汇, 如英文代词词汇 “we”;
- (2) 可以通过形态分析确定唯一词性的词汇, 如英文动词词汇 “bought”;
- (3) 可以通过前缀和后续分析确定唯一词性的词汇, 如英文形容词词汇 “well-founded”;
- (4) 非英文词汇, 如数字, 标点符号等。

2.2 带标数据的自动抽取过程

我们采用的源语料是美国 Berlitz 公司的汽车配件真实语料, 未经过任何词性人工标注。利用计算机从无任何标注的生语料中自动抽取词性三元组数据, 自动抽取过程分为

四个步骤:

第一步: 非英文词汇的识别

定义空格和标点符号为不同词汇之间的分隔符号。如果一个词汇中存在一个或多个非英文字母, 称之为非英文词汇。非英文词汇的类型主要分为五种: (1) 时间类型; (2) 数字类型; (3) 标点符号类型; (4) 数学符号类型; (5) 其它非英文词汇类型。

第二步: 英文词汇的词性标注和未登录词识别

这一步处理主要是一个直接查词库过程, 将每个英文词汇在词库中词性集合抽取出来, 作为标注结果。对于一些直接从基本词库和专业词库中无法查询到的词汇, 本文称之为未登录词汇, 词性标注为空 (NULL)。

第三步: 英文词汇的形态分析

形态分析的目的在于确定一些未登录词汇的词性。形态分析过程主要分为两步:

(1) 查询通用缩略语表和不规则动词表。根据不规则动词表, 可以确定一些具有不规则动词变化的未登录词词性, 词性标注为动词(v); 根据通用缩略语表, 可以确定一些以缩略语形式出现的未登录词词性, 词性标注为名词(n)。

(2) 根据前缀和后缀构词规则来确定未登录词词性。系统根据英语词汇前缀和后缀构词规则书写了大约五十多条未登录词词性猜测规则。

第四步: 训练数据的自动抽取

由于语料库中还存在一些没有被准确词性消歧的词汇, 在抽取训练数据时只考虑已经唯一确定词性的英文词汇 (不考虑还没有唯一确定词性的词汇), 同时, 在每个英文句子的开始附加一个特定的词汇: 句子开始词汇 (BEGIN), 词性标注为 B0。每个英文句子的结束符号称之为句子结束符号 (END), 词性标注为 C3, 通常是句号等。

下面给出一个自动抽取训练数据 (词性大类标注) 的例子:

无标注文本	FOLLOW THESE SCHEDULES IF YOU USUALLY OPERATE YOUR VEHICLE UNDER ONE OR MORE OF THE FOLLOWING CONDITIONS .	
第一步 非英文词汇的 识别	Follow these schedules if you usually operate your vehicle under one or more of the following conditions . C3	
第二步 词性标注和未 登录词汇识别	Follow VERB NOUN these PRON schedules NULL if CONJ you PRON usually ADV operate VERB your PRON vehicle NOUN under PREP ADV one NUM or CONJ more ADJ of PREP the ARTI following NULL conditions NULL . C3	
第三步 形态分析	Follow VERB NOUN these PRON schedules NOUNS if CONJ you PRON usually ADV operate VERB your PRON vehicle NOUN under PREP ADV one NUM or CONJ more ADJ of PREP the ARTI following VERBING conditions NOUNS . C3	
第四步 训练数据自动 抽取	词汇三元组 (these, schedules, if) (schedules, if, you) (if, you, usually) (you, usually, operate) (usually, operate, your)	词性三元组 (PRON, NOUNS, CONJ) (NOUNS, CONJ, PRON) (CONJ, PRON, ADV) (PRON, ADV, VERB) (ADV, VERB, PRON)

(operate, your, vehicle)	(VERB, PRON, NOUN)
(one, or, more)	(NUM, CONJ, ADJ)
(or, more, of)	(CONJ, ADJ, PREP)
(more, of, the)	(ADJ, PREP, ARTI)
(of, the, following)	(PREP, ARTI, VERBING)
(the, following, conditions)	(ARTI, VERBING, NOUNS)
(following, conditions, .)	(VERBING, NOUNS, C3)

表 1 自动抽取训练数据的例子

三、词性自动标注模型

3.1 概率模型

给定一个词序列 $W=\{W_0, \dots, W_n\}$, 推测最有可能的词性序列 $C=\{C_0, \dots, C_n\}$ 。定义 $P(C|W)$ 为给定词序列 W 的条件下选择词性序列 C 的条件概率。很明显, 词性自动标注模型主要寻求一个词性序列 C , 使条件概率 $P(C|W)$ 达到最大。即:

$$P(C|W) = \arg \max(P(C|W)) \quad (1)$$

根据贝叶斯公式, 词性序列 C 的条件概率为:

$$P(C|W) = \arg \max\left(\frac{P(C)P(W|C)}{P(W)}\right) \quad (2)$$

其中 $P(C)$ 是词性序列 C 的先验概率, $P(W|C)$ 是词性序列 C 已知情况下词序列 W 发生的条件概率, $P(W)$ 是词序列 W 的非条件概率。也就是说, $P(W)$ 对于所有可能的词性序列来说都是相同的, 因此可以不考虑 $P(W)$ 。计算公式 (2) 只需要考虑 $P(C)$ 和 $P(W|C)$ 两项。同时假设当前词性只与前一个词性和后一个词性相关。这样的话, 我们可以使用下列简化公式来计算词性序列 C 的条件概率:

$$P(C|W) = \prod P(W_i|C_i)P(C_i|C_{i-1}, C_{i+1}) \quad (3)$$

对象的概率计算公式采用最大似然估计公式 MLE:

$$P_{MLE}(y|x) = \frac{R(x, y)}{R(x)} \quad (4)$$

其中 $R(x)$ 表示数据对象 x 在语料库中出现的次数。可以得出给定词序列 W 条件下, 选取词性序列 C 的条件概率为:

$$P(C|W) = \arg \max \prod \frac{R(W_i, C_i)}{R(C_i)} \frac{R(C_{i-1}, C_i, C_{i+1})}{\sum_{c \in \text{词性}} R(C_{i-1}, c, C_{i+1})} \quad (5)$$

其中句子开始符号 $w_0=\text{BEGIN}$ 的词性 $c_0=\text{B0}$ 和句子结束符号 $w_n=\text{END}$ 的词性 $c_n=\text{C3}$

的条件概率不参与计算，只是在计算其它词性的条件概率时作为上下文。

3.2 平滑技术

为了解决稀疏数据问题，后来许多人提出了一些 MLE 的改进方法。基本思想将 MLE 作为最初的评估，然后将数据对的条件概率之和小于 1，留出一部分概率赋给未出现在样本语料中的可能数据对。这个技术我们称之为平滑 (Smoothing) 技术。Jelinek 和 Mercer (1980) 提出了著名的内插法平滑技术。我们采用简化的内插法平滑公式：

$$P_{MLE}(c|o) = \lambda \frac{R(o,c)}{R(o)} + (1-\lambda) \frac{1}{|CT|} \quad (6)$$

其中 $R(o,c)$ 表示 (o,c) 在训练语料中出现的次数， $R(o)$ 表示对象 o 在训练语料中出现的次数。 $|CT|$ 表示局部上下文的数目。本文考虑的局部上下文为对象的前一个词的词性和后一个词的词性，因此可以得出： $|CT|=26$ 种词性 \times 26 种词性 = 676 种局部上下文。在本文的实验中，设置插值参数 λ 为 0.90，取得了很好结果。

3.3 未登录词的词性猜测

系统对于一些直接从基本词库和专业词库中无法查询到的词汇，本文称之为未登录词汇，词性标注为空 (NULL)。未登录词汇的词性确定主要分为两步：

第一步：对具有形态变化的未登录词汇进行形态分析，猜测可能词性：(1) 查询通用缩略语表和不规则动词表。根据不规则动词表，可以确定一些具有不规则动词变化的未登录词词性，词性标注为动词(v)；根据通用缩略语表，可以确定一些以缩略语形式出现的未登录词词性，词性标注为名词(n)。(2) 根据前缀和后缀构词规则来确定未登录词词性。

第二步：对不具有形态变化的未登录词汇，系统假设该词汇的词性后选集合为所有可能的词性。也就是说，系统假设未登录词可能具有各种词性参与公式 (5) 的计算，最终根据计算结果来确定未登录词的词性。

四、实验结果

4.1 自动抽取结果统计

我们采用的源语料是美国 Berlitz 公司的汽车配件真实语料，覆盖了汽车配件的使用和维修内容，没有任何标注，包含大约 4 万个英文句子，带标数据的抽取处理程序在 P200

的微机，处理平均速度大约为 15 句/秒。词性三元组数据的抽取结果统计如下：

源语料库规模（词次）	带标数据规模（词次）	具有歧义的数据规模（词次）
674116	402722	271394

表 2 词性三元组数据的自动抽取结果统计

4.2 自动标注结果

为了测试基于生语料库的词性自动标注效果，我们构造了一个包括两万多词次关于汽车配件的开放性测试集，采用三种计算模型对测试语料进行自动标注，测试结果如下：

编号	计算模型 $P(C W)$	准确率 CP
(1)	$\operatorname{argmax} \prod P(C_i)P(W C_i)$	80.2%
(2)	$\operatorname{argmax} \prod P(C_i C_{i-1}C_{i+1})$	93.1%
(3)	$\operatorname{argmax} \prod P(C_i C_{i-1}C_{i+1})P(W_i C_i)$	85.4%

表 3 三种计算模型的测试正确率

类型	词数 N	标注正确数 N_c	标注准确率 CP
非英文词汇	5192	5192	100%
单词性英文词汇(不包括形态变化的单词)	8060	8060	100%
多词性英文词汇(不包括形态变化的单词)	4552	3610	79.3%
未登录英文词汇(包括形态变化的单词)	5675	4996	88%
总共	23479	21858	93.1%

表 4 计算模型 (2) 的详细测试统计

4.3 讨论

从实验结果我们发现：

1、从表 3 中可以发现一个问题，计算模型 (2) 的测试准确率最好，比计算模型 (1) 的准确率高 12.9%，比计算模型 (3) 的准确率高 7.7%。计算模型 (2) 比计算模型 (1) 效果好，这与一些研究人员的研究结果一样，但是不同的是在于计算模型 (2) 的效果比计算模型 (3) 好。不同的主要原因在于计算模型 (3) 中条件概率 $P(W_i|C_i)$ 存在严重的数据稀疏问题，因为基于 NAA 的带标数据的自动获取过程中不考虑未消歧的数据，导致带标数据缺乏大量的具有多词性的词汇。由此导致计算模型 (3) 处理表 4 中第二项和第四项时，效果没有计算模型 (2) 好。

2、从实验结果我们可以发现，跟一些关于词性自动标注的研究论文相比，准确率比它们低，主要有两个原因：(1) 系统测试的语料是美国 Berlitz 汽车配件开放性真实语料，

由于存在大量未登录词（大约占测试语料的 20%）；（2）统计模型的构造是基于没有任何词法、语法语义标注语料库。跟基于人工标注的训练语料的统计模型相比，准确率相对下降，这是符合实际情况的。

3、由于带标数据（词汇三元组和词性三元组）在没有人工干预的前提下自动获取而来，这样真正可以实现利用大规模真实语料（甚至可以达到上亿词次）进行训练统计模型参数，这一点对于传统采用人工标注是所不能的（特别对中文而言）。

五、结束语

本文提出的基于 NAA 的词性自动标注方法的主要特色在于，基于 NAA 从无标注语料中自动抽取大量的词性三元组，避免了人工加标的过程，真正可以实现利用大规模真实语料（甚至可以达到上亿词次）进行训练统计模型参数（语料库加工平均速度大约为 20 万词次/小时）。实际上该方法也适用于中文词性自动标注。但是从实际应用中发现还存在一些不足之处，为了提高准确率，下一步系统将在三个方面进行改进：（1）扩充基本词典的词汇量，同时增加专业词汇量；（2）增加一些基于固定搭配的规则进行消歧处理；（3）加强未登录词汇的识别功能，这一点对于处理真实文本特别重要。

备注：目前刚开发完成的基于 NAA 的英文分词与词性自动标注系统（ENAAPOS Ver1.5.1）和中文分词与词性标注系统（CNAAPCS Ver 1.5）及其源代码、相关知识源和开发文档，可以通过访问东北大学计算科学研究所主页（[Http://ics.cs.neu.edu.cn](http://ics.cs.neu.edu.cn)）的“研究成果”自由下载，希望对大家有所帮助。

参考文献

- [1] Gulider,Linda, Van., Automated part-of-speech Tagging: A brief overview, Handout for LING361 Georgetown University,1995.
- [2] Eric Brill. A simple rule-based part of speech tagger, Proceedings of the Third Conference on Applied Natural Language Processing,1992.
- [3] DeRose S J, Grammatical Category Disambiguation by Statistical Optimization, Computational Linguistics, V14, No1, p23-44, 1988.
- [4] De Marcken G G, Parsing the LOB Corpus, 28th Annual Meeting of the ACL Proceeding of the Conference, 6-9, June, 1990.
- [5] 白栓虎, 夏莹, 黄昌宁, “汉语语料库词性标注方法研究”, 机器翻译研究进展, 陈肇雄主编, 电子工业出版社, p401-412, 1991年12月.
- [6] 周强, 俞七汶, “一种切词和词性标注相融合的汉语语料库多级加工方法”, 第二届全国计算语言学联合学术会议, 1993年6月.
- [7] Cutting,D.et al.(1992), A practical part-of-speech tagger, In Proc. 3rd Conference on Applied Natural Language Processing,ACL,Trento,Italy,1992.
- [8] Jelinek, Frederick and Robert L. Mercer. Interpolated Estimation of Markov Source Parameters from Sparse Data. In Proceedings of the Workshop on Pattern Recognition in Practice. Amsterdam. 204-216. May,1980.