

副词“都”用法自动识别研究

张静杰 咎红英[†]

郑州大学信息工程学院, 郑州 450001; [†] 通信作者, E-mail: iehyzan@zzu.edu.cn

摘要 首先对副词“都”分别进行基于规则和基于统计两种方法的用法自动识别研究, 并具体分析它们的优点和不足, 然后尝试了规则与统计相结合的方法。3 种方法的准确率分别为 82%, 89.62% 和 98.54%, 实验数据表明, 采用规则和统计相结合的方法使副词“都”用法的自动识别取得较好的效果。

关键词 自然语言处理; 用法规则; 条件随机场; 自动识别; 都

中图分类号 TP391

Automatic Recognition Research on Chinese Adverb DOU's Usages

ZHANG Jingjie, ZAN Hongying[†]

School of Information Engineering, Zhengzhou University, Zhengzhou 450001; [†] Corresponding author, E-mail: iehyzan@zzu.edu.cn

Abstract The authors recognized the adverb DOU's usages with two methods, rule-based and statistics-based, and analyzed their advantages and disadvantages, respectively. And then, The methods of rule-based and statistical-based were combined. The accuracies of these three methods are 82%, 89.62% and 98.54%, the experiments show that combination method of rule-based and statistical-based is effective in automatically recognizing of adverb DOU's usages.

Key words natural language process; usage rule; conditional random fields; automatic identification; DOU

在现代汉语中, 副词的句法功能相对比较简单, 然而, 一些常用副词不仅使用频率很高, 而且用法丰富多样。马真^[1]曾提出每个副词都需要进行综合的分析与研究。也就是说, 对于副词研究, 尤其是一些义项纷繁、用法灵活的常用副词, 需要针对个例, 做仔细的考察和对比, 才有可能做到对汉语句子较为准确的理解。副词“都”是现代汉语中使用频率比较高、意义和用法也比较复杂的副词, 一般认为它具有范围总括、表达语气和表示时间 3 种不同的功能。如下面例句^[2]:

- 无论工作大小, 我们都要把它做好。 (1)
- 真抱歉, 我都忘了你的名字了。 (2)
- 饭都凉了, 快吃吧! (3)

例句(1)中“都”表示范围上的总括, 例句(2)中“都”表达一种语气, 有“甚至”之意, 例句(3)中“都”

表示时间, 有“已经”之意。语境不同, 副词“都”所表达的意思也不同。在咎红英等^[3]目前构建的现代汉语虚词知识库中, 副词“都”共有 3 个义项 11 个用法, 是用法比较复杂的副词之一, 研究副词“都”的用法自动识别对整个副词的用法识别研究以及汉语句子的理解都有着重要的作用。

自然语言处理领域中有两种不同的研究方法: 理性主义方法和经验主义方法, 也就是通常所说的基于规则的方法和基于统计的方法。人们逐渐在研究中发现不能单纯依靠规则或者统计, 应该对两者的优点进行分析, 把理性主义和经验主义结合起来, 进而达到更好的研究结果。因此, 在进行副词“都”的用法自动识别研究中, 首先对基于规则和基于统计两种方法进行分析, 然后提出规则与统计相结合的思想。

国家自然科学基金(60970083)和河南省科技创新人才杰出青年基金(104100510026)资助

收稿日期: 2012-05-28; 修回日期: 2012-08-15; 网络出版时间: 2012-10-26 17:04

网络出版地址: <http://www.cnki.net/kcms/detail/11.2442.N.20121026.1704.002.html>

1 副词“都”的用法描述

在对副词用法进行识别研究时,副词的用法描述分两种:一种是面向人类的用法描述,人可以根据这些描述直观区分各个用法;另外一种就是面向机器的用法描述,在进行研究时便于机器识别各个用法。

根据俞士汶等^[4]提出的“三位一体”构建现代汉语广义虚词知识库的思路,咎红英等^[3]初步构建了虚词用法词典、用法规则库以及用法标注语料库“三位一体”的现代汉语虚词知识库,广义虚词包括副词、介词、语气词、助词、连词及方位词等。根据《现代汉语八百词》、《现代汉语虚词词典》、《现代汉语词典》(第5版)、1998年1月《人民日报》分词与词性标注语料库的统计分布,以及其他语法学家有关论著对虚词进行用法描述(来源在用法描述中分别标记为〈b〉,〈x〉,〈h〉,〈r〉和〈z〉),其中副词“都”共有3个义项11个用法,具体描述见表1。

表1的用法描述属于面向人类的用法描述,在进行自然语言处理研究过程中很难直接使用,因此需要将人类能够识别的副词“都”的用法描述转化成机器可以识别的描述。咎红英等^[5-6]对面向人类用法

的研究结果进行了BNF规则的形式化描述,其中符号标记采用现代汉语虚词知识库中规则库的标记标准。然后根据副词的用法描述,结合虚词知识库中的例句,对用法用形式化的规则进行了描述,形成了面向机器识别的副词用法规则。与表1相对应的副词“都”的用法规则描述如下。

\$都

@〈d_dou1_1〉→N^N→v|a

@〈d_dou1_1a〉→M^M→谁|哪里|什么|怎么|哪儿|哪|〈ry〉|〈ryw〉

@〈d_dou1_1b〉→M^M→(不论|不管|无论|虽然|尽管|凡是|只要)*{,}

@〈d_dou1_1c〉→NE^N→谁|哪里|什么|怎么|哪儿|哪|〈ry〉|〈ryw〉^E→?

@〈d_dou1_1d〉→FR^F→~^R→是

@〈d_dou1_2〉→N^N→v

@〈d_dou1_2a〉→M^M→连|甚至

@〈d_dou1_2b〉→A~A(不|没|没有|未|〈df〉)^A→v

@〈d_dou1_2b〉→A~(不|没|没有|未|〈df〉)*A^A→v

@〈d_dou1_2c〉→MN^M→~q^N→不|没|没有|未|〈df〉

@〈d_dou1_2d〉→N^N→[,]* (不|没|〈df〉)

@〈d_dou1_3〉→E^E→了

表1 副词“都”用法描述
Table 1 Usages description of adverb DOU

ID	释义	用法
d_dou_1	表示总括全部。〈b〉	除问话以外,所总括的对象必须放在“都”前。也可以说“全都”,总括的意思更明显。〈b〉
d_dou_1a	表示总括全部。〈b〉	所总括的对象可以用表示任指的疑问指代词。〈b〉
d_dou_1b	表示总括全部。〈b〉	所总括的对象前可以用连词“不论、不管、无论、凡是、只要”。〈b〉〈z〉
d_dou_1c	表示总括全部。〈b〉	问话时总括的对象(疑问代词)放在“都”后。〈b〉
d_dou_1d	表示总括全部。与“是”搭配。说明原因,有责备的意思。〈b〉	与“是”搭配。〈b〉
d_dou_2	甚至。“都”轻读。〈b〉	修饰动词或动词短语。〈z〉
d_dou_2a	甚至。“都”轻读。与“连”字同用,有强调语气的作用。〈b〉	与“连”字同用。〈b〉
d_dou_2b	甚至。“都”轻读。〈b〉	“都”前后用同一个动词(前肯定,后否定)。〈b〉A~(不 没 没有 未 df)A 或 A~A(不 没 没有 未 df)。〈z〉
d_dou_2c	甚至。“都”轻读。〈b〉	一+量词+...~+动词(否定式)。〈b〉
d_dou_2d	甚至。“都”轻读。〈b〉	用于表示让步的小句,引出表示主要意思的小句。〈b〉
d_dou_3	已经。〈b〉	句末常用“了”。〈b〉

2 副词“都”的用法自动识别研究

为了适应自然语言处理的应用需求，根据对副词“都”用法的形式化描述以及对真实文本语料中副词“都”不同用法的语境考察，对副词“都”的用法进行了自动识别研究。

2.1 基于规则的副词“都”用法自动识别

基于规则的理性主义方法通常明白易懂，表达清晰，描述明确，很多语言事实都可以用语言模型的结构和组成成分直接地、明显地表示出来，而且这些语言学规则的形式描述能力和形式生成能力都很强，在自然语言处理中有很好的应用价值^[7]。

在面向机器的副词“都”的用法描述中，除了目前无法准确描述的读音、短语、小句等特征外，某些规则描述还有形式上或语法上的交叉现象。而在规则用法自动识别过程中，系统按照给出的规则逐条进行匹配，直到找到第一个满足规则描述的用法进行标注输出。如果按照用法编码的自然顺序对汉语句子里副词“都”的用法进行自动识别，可能会出现错误识别的情况，因为系统默认编码在前的用法具有较高的优先级。例如用法〈d_doul_1〉仅仅要求副词“都”后面出现一个动词或者形容词，条件非常宽泛，这也是用法〈d_doul_1d〉、〈d_doul_2b〉等规则的要求之一。因此，为了保证规则识别具有较高的准确率，对副词“都”规则的顺序进行调整，即按照规则之间的互相包含或覆盖关系，将比较独立的规则或易于准确识别的规则赋予较高的优先级，即排在比较靠前的位置。例如用法〈d_doul_2a〉要求句子中副词“都”的前面含有“连”或“甚至”等词，规则易于准确识别。实验时需要进行多次调序，对比每次调序后的准确率，将用法标注准确率最高的规则顺序作为最终的规则调序。

根据管红英等构建的虚词用法词典对北京大学提供的分词和词性标注后的《人民日报》(2000 年 4—6 月出版)语料进行了用法标注，副词“都”各个用法在 3 个月语料中的分布率及词频见表 2。

利用调整顺序后的副词“都”用法规则对未标注用法信息的语料进行用法自动标注，用法标注前后的语料对比如下：

党/n 的/ud 各项/r 工作/vn 都/d 要/vu 坚持/v 、/wu 体现/v 和/c 贯彻/v “/wyz 三/m 个/qe 代表/v ”/wyy 的/ud 要求/n
党/n 的/ud 各项/r 工作/vn 都/d 〈d_doul_1〉 要

表 2 副词“都”的用法分布
Table 2 Usages distribution of adverb DOU

ID	分布率/%	词频
d_doul_1	83.9	5699
d_doul_1a	2.7	180
d_doul_1b	9.00	611
d_doul_1c	0.13	9
d_doul_1d	0.12	8
d_doul_2	0.60	43
d_doul_2a	2.40	160
d_doul_2b	0.06	4
d_doul_2c	0.21	14
d_doul_2d	0.18	12
d_doul_3	0.75	51
合计	100.00	6791

/vu 坚持/v 、/wu 体现/v 和/c 贯彻/v “/wyz 三/m 个/qe 代表/v ”/wyy 的/ud 要求/n

各个用法的识别情况如表 3 所示，在出现的 6791 个副词“都”中，正确标注的有 5555 个，准确率为 82%。整体上分布率差距对各个用法的识别情况影响不大，个别用法的识别准确率很低，如用法〈d_doul_1d〉，因为其规则描述只能体现出表 1 用法〈d_doul_1d〉释义描述中的“与‘是’搭配”，而无法体现带有感情色彩的“责备”之意。如以下例句是规则标注的结果，经分析该句中的副词“都”应标注为用法〈d_doul_1〉。

20000607-10-002-009/m 每次/rz 法制/n 讲座/n 的/ud 内容/n , /wd 都/d 〈d_doul_1d〉 是/vl 根据/p 国

表 3 基于规则的副词“都”的用法自动识别
Table 3 Rule-based method automatic recognition of adverb DOU's usages

ID	准确率/%	召回率/%	F 值/%
d_doul_1	99.61	80.19	88.85
d_doul_1a	74.46	76.11	75.27
d_doul_1b	89.40	99.35	94.11
d_doul_1c	100.00	77.77	87.50
d_doul_1d	3.64	100.00	7.02
d_doul_2	0	0	0
d_doul_2a	85.25	97.50	90.96
d_doul_2b	100.00	50.00	66.67
d_doul_2c	61.11	78.57	68.75
d_doul_2d	1.09	75.00	2.15
d_doul_3	68.57	94.12	79.34
合计	82.00	81.80	81.90

内外/s 形势/n 的/ud 发展/vn 、/wu 针对/p 我国/rz 改革/v 和/c 发展/v 的/ud 重要/a 问题/n 确定/v 的/ud 。/wj

基于规则的用法自动识别方法存在一些缺点:规则比较脆弱、鲁棒性差;在生成规则时往往需要语言学家等各种专家的配合,进行知识密集的研究,研究工作的强度很大;规则不能通过机器学习的方法自动地获得,也无法使用计算机自动地进行泛化,这将阻碍自动识别准确率的提高。

2.2 基于统计的副词“都”用法自动识别

条件随机场(conditional random fields, CRF) 是一个在给定输入节点条件下计算输出节点条件概率的无向图模型,它考察给定输入序列对应的标注序列的条件概率,训练目标是使得条件概率最大化。CRF 的概念自被 Lafferty 等^[8]提出以来,在 NLP 和图像处理等领域中得到了广泛的应用。在中文分词^[9]、词性标注^[10]、语义角色标注以及命名实体识别^[11]等自然语言处理领域取得了一定的成功。

本文在基于 CRF 的副词“都”用法标注实验中,利用 CRF++(CRF++: Yet Another Toolkit (<http://www.chasen.org/~taku/software/CRF++>)) 作为自动标注工具。实验时选取 3 个特征模板,第一个是“都”前 m 个词语和后 n 个词语的词形特征,第二个是“都”前 m 个词语和后 n 个词语的词性特征,第三个是“都”前 m 个词语和后 n 个词语的词形和词性的复合特征,其中 m 和 n 控制在 2~10。选择 2000 年 4—6 月《人民日报》语料进行交叉实验,其中两个月的语料作为训练语料,另外一个月的语料作为测试语料进行测试。

经过实验,选取第三个特征模版,并且左窗口为 7 右窗口为 3 时实验的结果最佳。采用该实验方法将 3 个月的识别情况汇总后得到各个用法的识别情况,如表 4 所示。在出现的 6791 个副词“都”中,正确标注的有 6086 个,准确率达到 89.62%,明显高于基于规则的用法自动识别结果。与基于规则的方法相比,用法的分布率对各个用法的识别有很大的影响,分布率较高的用法的识别准确率也较高。

对于副词“都”来说,虽然基于统计的方法在总体上优于基于规则的方法,但是分析基于规则与基于统计的实验结果,可以看出对于某些用法来说,规则的结果又优于基于统计的结果,例如用法〈d_doul_1c〉等。因为这些用法的规则相对独立,而且易于准确识别,因此在基于规则的方法中识别

表 4 基于统计的副词“都”的用法自动识别
Table 4 Statistics-based method automatic recognition of adverb DOU's usages

ID	准确率/%	召回率/%	F 值/%
d_doul_1	89.84	99.28	94.32
d_doul_1a	78.81	51.67	62.42
d_doul_1b	86.92	36.99	51.90
d_doul_1c	0	0	0
d_doul_1d	0	0	0
d_doul_2	50.00	2.33	4.45
d_doul_2a	100.00	65.63	79.25
d_doul_2b	0	0	0
d_doul_2c	0	0	0
d_doul_2d	0	0	0
d_doul_3	42.86	5.88	10.34
合计	89.62	89.62	89.62

准确率较高,但是又因这些用法在语料中的分布太少,无法进行统计学习,以致在基于统计的方法中不能准确的识别。因此,为了取得更好的用法自动识别结果,我们考虑综合分析规则和统计的优点,采用规则与统计相结合的方法对副词“都”进行用法的自动识别研究。

2.3 规则与统计相结合的副词“都”用法自动识别

分析基于规则和基于统计两种方法的实验结果,发现针对每一个用法进行评价时,我们一般从分布率、准确率、召回率等参数进行讨论,因此可以考虑这些参数在规则和统计中的相互联系,将其作为两者相结合的信息。例如,分布率比较低的用法在规则中的识别效果优于在统计中的识别,所以可以考虑将这些用法的规则信息加入到统计标注过程中;或者在采用统计方法进行用法自动识别时,先识别出那些分布率较低的用法,然后识别分布率较高而且统计效果较好的用法。总之,在进行规则与统计相结合时,可以有很多不同的结合方式。

本文在进行规则与统计相结合实验时,基于统计的方法采用条件随机场模型,将准确率作为规则与统计相结合的结合信息。相结合的具体思想是对包含副词“都”的句子分别进行基于规则和基于统计的方法自动识别,结果分以下 3 种情况: 1) 两者都没有识别结果,识别失败; 2) 两者只识别出一个结果,即两者之一没有识别结果或者两者识别结果相同,则标注该用法; 3) 两者识别结果不同,对比两

者各自在原模型中识别准确率的大小，标注准确率较高的用法。

对 2000 年 4—6 月《人民日报》语料进行上述的规则与统计相结合的实验，结果如表 5 所示。在出现的 6791 个副词“都”中，正确标注的有 6692 个，准确率达到 98.54%，比规则方法高 16.54%，比 CRF 方法高 8.92%。从用法识别角度分析，除了分布率较高的用法〈d_dou1_1〉具有较高的识别准确率外，分布率较低的用法的识别准确率与基于统计的方法相比也有很大的提高，如用法〈d_dou1_3〉等。

表 5 规则与统计相结合的副词“都”用法自动识别
Tabel 5 Combination method of rule-based and statistical-based automatic recognition of adverb DOU's usages

ID	准确率/%	召回率/%	F 值/%
d_dou1_1	99.58	99.65	99.62
d_dou1_1a	97.77	97.22	97.49
d_dou1_1b	96.45	97.87	97.15
d_dou1_1c	77.78	77.78	77.78
d_dou1_1d	87.50	87.50	87.50
d_dou1_2	7.41	4.65	5.71
d_dou1_2a	99.38	99.38	99.38
d_dou1_2b	100.00	50.00	66.67
d_dou1_2c	73.33	78.57	75.86
d_dou1_2d	44.44	66.67	53.33
d_dou1_3	88.00	86.27	87.13
合计	98.54	98.54	98.54

3 结语

副词“都”是现代汉语中使用频率比较高、意义和用法也比较复杂的副词。本文对副词“都”的用法分别进行了基于规则、基于统计以及规则和统计相结合的自动识别研究。从实验结果可以看出，基于规则和基于统计的两种方法各有优点和不足，而规

则与统计相结合的方法将理性主义和经验主义很好的结合在一起，其结果明显优于两者。

在下一步工作中，我们将考虑使用其他的规则和统计相结合方法进一步提高用法的自动识别的准确率。同时，考虑在用法自动识别的前提下，将用法识别应用到句法分析等其他自然语言处理领域。

参考文献

- [1] 马真. 现代汉语虚词研究方法论. 北京: 商务印书馆, 2004
- [2] 吕叔湘. 现代汉语八百词. 北京: 商务印书馆, 1980
- [3] 笱红英, 朱学锋. 面向自然语言处理的汉语虚词研究与广义虚词知识库构建. 当代语言学, 2009, 11(2): 124-135
- [4] 俞士汶, 朱学锋, 刘云. 现代汉语广义虚词知识库的建设. 汉语语言与计算学报, 2003, 13(1): 89-98
- [5] 笱红英, 张坤丽, 柴玉梅, 等. 现代汉语虚词知识库的研究. 中文信息学报, 2007, 21(5): 107-111
- [6] Hao Liping, Zan Hongying, Zhang Kunli, et al. Research on Chinese adverb usage for machine recognition // Proceedings of the 7th International Conference on Chinese Computing (ICCC2007). 2007: 122-125
- [7] 宗成庆. 统计自然语言处理. 北京: 清华大学出版社, 2008
- [8] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data // Proceedings of the 18th ICML-01. 2001: 282-289
- [9] 黄德根, 焦世斗, 周惠巍. 基于子词的双层 CRFs 中文分词. 计算机研究与发展, 2010, 47(5): 962 - 968
- [10] 应玉龙, 李森, 乌达巴拉, 等. 基于条件随机场的蒙古语词性标注方法. 计算机应用, 2010, 30(8): 2038-2040
- [11] 史海峰. 基于 CRF 的中文命名实体识别研究[D]. 苏州: 苏州大学, 2010