文章编号 :1007-5321(2008)01-0014-04

级联中文组块识别

秦 颖, 王小捷, 钟义信
(北京邮电大学 信息工程学院, 北京 100876)

摘要:基于统计方法的中文组块研究大多借鉴 CoNLL2000 英文组块的思想,建立了组块表示的 BIO 模型,并将组块识别任务作为一种为词序列标注的多分类问题. 为降低分类复杂度,采取了一种分解识别法,即先识别组块的边界,再进行组块类别判定. 基于条件随机场(CRF)构建了级联组块识别器,实验数据集采用宾州大学中文树库(CTB5.1). 在特征选择上,借鉴了中文分词特征选择的方法. 5 倍交叉验证的实验结果为:组块边界识别的 F_1 值为 95.05%;类型识别的准确率为 99.43%;整体 F_1 值为 93.58%. 该方法提高了系统性能,缩短了学习器的训练时间.

关 键 词:中文组块;边界识别;类别识别;条件随机场
中图分类号:TP391 **文献标识码**:A

Cascade Identification of Chinese Chunks

QIN Ying, WANG Xiao-jie, ZHONG Yi-xin
(School of Information Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Most statistical-based Chinese chunking researches was inspired by English chunking of CoNLL2000. After representing chunks within the scheme of tags for words in a chunk, chunk identification task was cast as word sequence tagging and tackled as multi-classification problems. For sake of decreasing classification complexity, a decomposed chunking approach was proposed: first, chunk boundary identification, and then chunk type identification. The vital problem of Chinese chunking is actually boundary identification. Cascade chunk identifiers were built based on conditional random fields (CRF). The experimental dataset was extracted from Chinese tree bank 5.1 (CTB5.1). As to the features selection, some methods often used in Chinese word segmentation were borrowed to chunking task. On 5 cross validation of dataset, F_1-measure of chunk boundary identification is 95.05%, and the precision of chunk type recognition is 99.43% as well. And the total chunking F_1-meausre reaches 93.58%. Comparing with other relative researches, the performance is improved and the training time of learners is sharply shortened.

Key words: Chinese chunking; boundary identification; type identification; conditional random fields

作为一种浅层句法分析[1],组块识别在机器翻译、搜索引擎、信息抽取等领域受到了普遍重视. 英文组块识别在 CoNLL2000[2] 的规范和评测平台上已有了较成熟的研究成果[3-4],而中文组块的研究仍存在不少亟待解决的问题. 中文组块识别大致可分为基于规则的方法[5]和基于统计的方法[6-8]. 基于统计的方法借鉴了英文组块的思想,将组块识别转化为分类问题. 一般组块的边界和类型识别同时

收稿日期:2007-04-05
基金项目:语言司民文语科库工具建设项目(MZ115-022)
作者简介:秦 颖(1971—),女,博士生,E-mail:qinyingmail@163.com.

进行,出现分类类别过多[6-8]的问题.汉语的词没有一致的定义,同样,对于组块(短语)也存在多种不同的看法.但是汉语语言学研究者基本认为,汉语的句子、短语和词语在内部语法构成上很相似,甚至是一致的.因此,组块(短语)研究应与词的研究相结合,互相借鉴.通常中文的分词和词性标注是 2 项不同的任务,所以本文将组块的边界和类型分开识别.通过编码,把组块边界识别转换为词边界的二分类问题,把类型识别视为边界确定组块的多分类问题.2 个分类器级联实现组块识别整体任务.分类类别的减少对于提高组块识别性能有利,同时也明显降低了训练分类器的时间.

1　中文组块的定义和表示

1.1　中文组块的定义

中文组块与英语 chunk 类似,被定义为一种非重叠(non overlapping)、非递归(non recursive)、全覆盖(non exhaustive)、成分边界不交叉(not cross constituent boundaries)的语言单位[2].1 个组块可由 1 个或多个词语构成.目前一些计算语言学研究者使用组块的经验定义,即根据标记语料定义组块,以避免语言学上的争议.宾州大学中文树库(CTB5.1)是由完整句法解析的句子构成中文语料,其中对短语进行了分层标记.文献[6-7]中,将树库词语层紧邻上一层的短语解析为组块.本文也采用类似的经验定义.

实验中共有 12 种类型的组块,主要有:NP-名词组块、VP-动词组块、ADJP-形容词组块、ADVP-副词组块、PP-介词组块、QP 量词组块、DNP-"的"字结构等.其他非组块的词语为第 13 类.

1.2　组块的表示

组块的表示即编码问题,常见的有 BIO 模型.BIO 表示模型在 CoNLL2000 上提出[2],用 B-X 表示 X 类型组块的起始词,I-X 表示 X 类型的组块中非起始词以外的其他词,O 表示不属于任何组块的词.这种表示方法存在 2 个问题.

1) 分类类别过多.由于将组块类别和边界标记合并,若有 m 种组块类型,那么分类器需要能区分 2× m+1 种类别.如果用 SVM 等二分类器完成分类任务,复杂度将会很高[8].即使用多分类算法,如 CRF 等,在一定规模的语料上训练分类器也需要很长时间.

2) 识别结果不一致问题.如 B-X 标记的下一个词可能为 I-Y,X 和 Y 为不同类型的组块.对于不一致问题,一般解决办法是将 I-Y 直接视为 B-Y[2],而这必将导致错误.

为克服上述问题,本文将中文组块表示方法分解为 2 种:一种为组块的边界标记;另一种为组块类型标记.

2　级联组块识别

组块识别的任务是在已经进行了中文分词和词性标记的句子中标记出组块的边界及其类型.在实际应用中,如机器翻译和搜索任务,有时更关注组块边界的正确与否.相应于组块的表示方法,本文将组块识别分解为 2 个级联的子任务:先进行组块边界识别,然后识别组块的类型,其中边界识别为组块识别的中间结果.

组块边界识别同中文分词都可视作对句子的分割问题,不同之处是分词是基于字进行的,而组块是基于词进行的.级联组块识别不仅将复杂问题降阶,同时在组块边界识别上也可更多地借鉴中文分词的技巧和方法.实验结果说明,中文组块识别的关键问题是边界识别问题,类型识别在本文的实验中有很高的准确率,这与组块内部结构和类型良好的对应性有关,如名词组块的尾词一般为名词,一般也是该组块的头词(headword);动词组块的头词一般是首词等.这样,根据组块首词和尾词的词性就能较好地判定组块的类别.

2.1　边界识别和评测

本文采取一种简单的对词的右边界编码的方法标记组块的边界,用"1"表示该词右边界非组块边界,用"0"表示该词右边界同时也是组块边界.如

[外商 NN 1][投资 NN 1][企业 NN 0][在 P 0][改善 VV 0][中国 NR 0][出口 NN 1][商品 NN 1][结构 NN 0][中 LC 0][发挥 VV 1][了 AS 0][显著 JJ 0][作用 NN 0][. P 0]

其中括号内的各项分别为[词 词性 边界标记].该句中的组块有"外商投资企业,在,改善,中国,出口商品结构,中,发挥了,显著,作用,."组块边界识别问题转化为词右边界的二分类问题.分类器的特征选择详见 3.2 节.

组块边界识别评测和分词评测有相通之处,用准确率(P)和召回率(R)及综合指标 F_1 值衡量.

$$P = \frac{\text{边界正确识别的组块数}}{\text{识别的组块总数}} \quad (1)$$

$$R = \frac{\text{边界正确识别的组块数}}{\text{标准组块总数}} \quad (2)$$

$$F_1 = 2P \times R/(P + R) \quad (3)$$

2.2 类别识别和评测

类别分类器的训练数据由解析到的基于BIO模型的组块转化而来,如

外商-投资-企业 NN NN NP

在 P P PP

改善 VV VV VP

中国 NR NR NP

出口-商品-结构 NN NN NP

中 LC LC LCP

发挥-了 VV AS VP

显著 JJ JJ ADJP

作用 NN NN NP

. PU PU O

例中的第1列为组块,中间2列为组块的首词和尾词的词性,最后1列为组块类型标记.

类别识别即为组块标记类别,用准确率评测,与词性标注的评测方法相同.

$$P = \frac{\text{正确标记的组块类型数}}{\text{组块总数}} \quad (4)$$

组块整体评测也包括准确率、召回率和 F_1 值,但是识别正确的组块指边界和类型都正确的识别结果.

3 条件随机场与特征选择

3.1 条件随机场

本文运用条件随机场(CRF,conditional random fields)实现组块识别的2个子任务.CRF是一种能保证损失函数(loss function)收敛到全局最优的算法,在用于命名实体识别、词性标注、组块识别等实验中,CRF显示了良好的性能,克服了标记偏问题(label bias)[9].实验中使用了下载的CRF算法包(http://chasen.org/~taku/software/CRF++).

3.2 特征选择

CRF用于组块边界识别的特征有上下文的词(W_i)和词性(P_i),i表示该词和当前词的位置关系.有unigram、bigram、trigram 3类语法特征,即

$$w_{-2}, w_{-1}, w_0, w_1, w_2, w_{-1}/w_0, w_0/w_1$$

$$p_{-2}, p_{-1}, p_0, p_1, p_2, p_{-2}/p_{-1}, p_{-1}/p_0,$$

$$p_0/p_1, p_1/p_2$$

$$p_{-2}/p_{-1}/p_0, p_{-1}/p_0/p_1, p_0/p_1/p_2$$

另外,还抽取训练集中的高频组块,构成组块表(相当于分词词表),共8 135个组块,占训练集中所有组块的3%.利用前向最大匹配法(FMM)判定是否是已知高频组块,将其作为一项特征加入到模型中.

边界识别后的结果作为类别识别的输入.类别识别的特征来自上下文的组块(C_i)和组块中词的词性(P_i)两部分信息,具体为 C_{-1},C_1 和

$$P_{0f}, P_{01}, P_{-1f}, P_{-11}, P_{1f}, P_{11}$$

P_{if}表示位置i组块的第1个词的词性,P_{il}表示位置i组块的最后一个词的词性,若该组块只有1个词,则两值相同.

通常BIO组块识别方法的特征仅为上下文的词和词性[6-7].这里,组块边界识别和类别识别采用了不同的特征,并加入了已识别组块和是否为高频组块等特征,使特征更加丰富.

4 实 验

4.1 实验数据集

CTB5.1包括890篇文章,以新闻语料为主,有12种类型的组块,其数目及平均长度分布见表1.以下实验结果均为数据集的5倍交叉验证结果.

表1 组块识别性能对比

类型	数目	平均长度	BIO	级联	CRF[6]	SVM[6]
ADJP	3 686	1.026	85.86	87.56	84.55	84.45
ADVP	12 996	1.002	85.96	88.07	82.74	83.12
CLP	80	1.063	27.63	30.84	0	5.26
DNP	12 376	1.000	99.80	99.83	99.64	99.65
DP	6 013	1.365	99.30	99.65	99.40	99.70
DVP	638	1.006	99.62	99.62	99.82	96.77
LCP	7 772	1.001	99.81	99.80	99.85	99.85
LST	183	1.000	66.84	68.81	68.25	68.75
NP	146 294	1.369	88.56	92.12	89.79	90.54
PP	17 664	1.002	99.59	99.71	99.66	99.67
QP	17 690	1.686	98.91	98.09	96.53	96.73
VP	77 326	1.503	90.76	93.04	88.50	89.74
总计	302 718	1.355	91.41	93.58	90.74	91.46

4.2 识别性能及分析

组块边界识别的特征选择如3.2节所述,识别性能为 P = 95.31%,R = 94.74%,F_1 = 95.05%.组块类别识别器的准确率为99.43%.整体识别的

F_1 值为 93.58% . 下面仅分析边界识别的主要错误.

1) 并列结构的组块识别错误率最高. 并列词"和""与""、"等在并列成分简单时包含在组块内,并列成分复杂时,位于组块外. 边界识别要正确标记并列结构组块,除上下文外,还需要更多的信息.

2) 带有修饰成分的组块识别也有较多错误. 当名词作修饰成分时. 如"[今年 NT] [经济 NN] [发展 NN]",分类器不能很好地判定为 2 个名词组块,往往识别为 1 个组块.

训练语料中的噪声是边界识别错误的 1 个来源,另一方面,的确存在有歧义和争议的组块边界.

4.3　对比工作

在同样算法和语料的基础上对比了组块级联识别和 BIO 模型识别的性能. BIO 模型的特征只有上下文特征,上下文窗口的设置同级联方法. 首先在模型的训练时间上,级联方法仅为 BIO 方法的 16.5% . 设备条件为 Intel(R) Xeon™ CPU 3.00 GHz, 2.00 GB 内存. 在算法参数设置相同的条件下,级联方法两类分类器的训练时间总和为 4 959.17 s,而 BIO 模型的分类器训练时间为 30 083.25 s. 表 1 反映了 2 种组块识别的性能. 级联组块识别在总体指标和大多数组块的识别上都优于 BIO 模型的结果.

文献[6]在 CTB 语料上基于组块 BIO 模型对比了 SVM、CRF、TBL 和 MBL 等各种算法,其中基于 SVM 的性能最优. 从表 1 可见,在 12 种类型的组块中,级联方法在 8 种组块的识别指标上胜出. 需要说明的是,文献数据是在 CTB4 上的实验结果,而 CTB4 语料规模为 838 篇,比 CTB5.1 略少. 本实验中 BIO 方法与文献[6]使用了类似的模型特征,但是由于训练语料增大,性能只提高了 0.67% ,但级联方式的性能却提高了 2.84% ,这说明性能提高的主要原因是方法的改进.

5　结束语

本文将中文组块识别分解为 2 步实现,即组块的边界识别和类型识别,分别用 CRF 构建了 2 个级联的分类器实现识别任务,与目前通用的组块识别方法相比,大大减少了学习器的训练时间,提高了大多数组块的识别性能和总体性能. 实验表明,组块识别的关键是边界识别. 在下一步的工作中,将借鉴更多中文分词的特征用于边界识别任务中,进一步提升组块边界的识别性能. 组块边界的识别性能决定了组块识别的整体性能.

参考文献:

[1] Abney S P. Parsing by chunks[C] // Steven P, Abney, Carol Tenny. Principle-Based Parsing. MA: [s.n.], 1991: 257-278.

[2] Erik F, Tjong Kim Sang, Sabine Buchholz. Introduction to the CoNLL-2000 shared task: chunking [C] // CoNLL-2000 and LLL-2000. Lisbon: [s.n.], 2000: 127-132.

[3] Sha Fei, Fernando C N, Pereira. Shallow parsing with conditional random fields[C] // Edmonton Alberta. Human Language TechnologyNAACL. CA: [s.n.], 2003: 213-220.

[4] Sun Guanglu, Huang Changning, Wang Xiaolong, et al. Chinese chunking based on maximum entropy Markov models [J]. Computational Linguistics and Chinese Language Processing, 2006, 11(2): 115-136.

[5] 詹卫东. 面向中文信息处理的现代汉语短语结构规则研究[D]. 北京:北京大学, 1999.
Zhan Weidong. A study of constructing rules of phrases in contemporary Chinese for Chinese information processing [D]. Beijing: Peking University, 1999.

[6] Chen Wenliang, Zhang Yujie, Isahara Hitoshi. An empirical study of Chinese chunking [C] // Coling-ACL2006 (Poster Session). Sydney: [s.n.], 2006: 97-104.

[7] Tan Yongmei, Yao Tianshun, Chen Qing, et al. Applying conditional random fields to Chinese shallow parsing [C] // Proceedings of CICLing 2005. Mexico City: Springer, 2005: 67-176.

[8] 李珩, 朱靖波, 姚天顺. 基于 SVM 的中文组块分析[J]. 中文信息学报, 2004, 18(2): 1-7.
Li Heng, Zhu Jingbo, Yao Tianshun. SVM based Chinese text chunking [J]. Journal of Chinese Information Processing, 2004, 18(2): 1-7.

[9] Lafferty John, McCallum Andrew, Fernando Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C] // International Conference on Machine Learning (ICML01). San Francisco: Morgan Kaufmann, 2001: 282-289.