第47卷第6期 大连理工大学学报 Vol. 47, No. 6
2007年11月 Journal of Dalian University of Technology Nov. 2007

文章编号：1000-8608(2007)06-0904-05

基于支持向量机的日语并列关系解析

周惠巍*， 黄德根， 李 巍

（大连理工大学 计算机科学与工程系，辽宁 大连 116024）

摘要：日语并列关系是日语依存关系的一种. 传统的日语依存关系解析不特殊解析并列关系，识别并列关系较为困难. 为此提出一种日语并列关系解析方法：基于支持向量机(SVMs)识别并列关键字，如果存在并列关键字，基于SVMs解析句中的并列关系，并将并列关系解析结果用于随后的依存关系解析. 用京都大学语料库进行测试，结果表明并列关系解析与依存关系解析均取得较高的正确率.

关键词：日语并列关系解析；支持向量机；并列关键字；依存关系解析

中图分类号：TP391.12 **文献标识码**：A

0 引 言

日语依存关系解析一直被认为是日语解析的一个重要技术. 近年来，大规模依存关系语料库的创建，使人们可以基于统计进行依存关系解析. 基于统计进行依存关系解析可以分为两种方法：一种是基于概率模型[1,2]的方法，这种方法通过计算所有可能依存关系的概率值确定句子的依存关系，因此其效率不高；另一种是工藤拓提出的组块逐步应用算法[3,4]，这种方法通过判断句中每个文节是否依存于其紧后方的文节解析整个句子，该方法简单高效，但使用这种解析算法解析距离较远的两个依存关系文节时，精确率却不是很高. 日语长句中往往存在并列关系，且存在并列关系的两个文节常常距离较远，所以解析并列关系至关重要.

日语长句中的并列结构解析被认为是一个比较困难的问题[5]. 黑桥祯夫提出通过计算句中各文节间的类似度，将并列关键字前后类似度总和最大的文节列作为并列结构范围的方法解析日语长句的并列结构. 该方法需要计算文中所有文节间的类似度，并比较多种可能的并列结构，因此其效率不高.

本文提出基于支持向量机(SVMs)(support vector machines)[6]进行日语并列关系解析方法. 决策树模型[7]与最大熵模型[1]曾应用于日语依存关系解析. 使用这些模型时，为得到较高的精度，必须慎重地选择特征，如果特征选择不当，将产生过学习. 并且为得到有效的组合特征，需要反复实验，所以无法高效地进行特征间的组合学习. 本文基于SVMs进行并列关键字识别、并列关系解析，以及依存关系解析. SVMs是基于分类边界的间隔最大化而提出的分类模型. 核函数的引入不但巧妙地解决了将低维特征向量映射到高维空间所引起的数据灾难问题，而且可以进行特征间的组合学习. SVMs基于所有学习样本求解最优分类超平面，理论与实验证明SVMs具有较强的泛化能力，不易产生过学习. 在自然语言处理领域，支持向量机已应用于文本分类，与传统的学习模型相比，即使增添学习特征也不会产生过学习，而且取得了较高的精度[8,9].

1 支持向量机(SVMs)

SVMs正负两类训练集为$(\boldsymbol{x}_i, y_i), \cdots, (\boldsymbol{x}_l, y_l)$，$\boldsymbol{x}_i \in \boldsymbol{R}^n$，$y_i \in \{+1, -1\}$，$\boldsymbol{x}_i$是数据$i$的$n$次特征向量($\boldsymbol{x}_i = (f_1 \quad f_2 \quad \cdots \quad f_n) \in \boldsymbol{R}^n$)，$y_i$表示数据$i$是正例(1)、负例(−1)的类标. 支持向量机用于分类问题其实就是导出由n次特征向量$\boldsymbol{x}_i \in \boldsymbol{R}^n$求得类标$y_i \in \{\pm 1\}$的识别函数$f: \boldsymbol{R}^n \rightarrow$

收稿日期：2006-03-04； 修回日期：2007-10-10.
基金项目：国家自然科学基金资助项目(60373095；60373096).
作者简介：周惠巍*(1969-)，女，讲师，E-mail：zhou_huiwei@163.com；黄德根(1965-)，男，博士，教授.
?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

± 1.

SVMs 是在 n 次空间构造一个最优分类超平面

$$(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}) + \boldsymbol{b} = \boldsymbol{0} \tag{1}$$

为增强分类器的泛化性，提高分类精度，最优化以下目标函数：

$$\begin{cases} \min L(\boldsymbol{w}) = \|\boldsymbol{w}\|^2 \\ \text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + \boldsymbol{b}) \geqslant 1 \ (i = 1,2,\cdots,l) \end{cases} \tag{2}$$

引入 Lagrange 乘数 $\alpha_i \geqslant 0 (i = 1,\cdots,l)$，通过求解最优化问题得到最终的识别函数 f 为

$$f(\boldsymbol{x}) = \mathrm{sgn}\Big(\sum_{i=1,\cdots,l;\, x_i \in S_{\mathrm{vs}}} \alpha_i y_i (\boldsymbol{x}_i \cdot \boldsymbol{x}) + \boldsymbol{b}\Big) \tag{3}$$

其中 S_{vs} 为支持向量.

核函数 $K(\boldsymbol{x},\boldsymbol{y}) = \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{y})$，巧妙地解决了将低维空间向量映射到高维空间所引起的数据灾难问题. 引入核函数后的识别函数为

$$f(\boldsymbol{x}) = \mathrm{sgn}\Big(\sum_{i=1,\cdots,l;\, x_i \in S_{\mathrm{vs}}} \alpha_i y_i K(\boldsymbol{x}_i \cdot \boldsymbol{x}) + \boldsymbol{b}\Big) \tag{4}$$

研究者提出了很多种核函数，经测试本文采用以下核函数：

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = (\boldsymbol{x}_i \cdot \boldsymbol{x}_j + 1)^d \tag{5}$$

d 次多项式核函数可以将 d 个以内的特征组合起来进行学习. 在本文中取 d 等于 3.

2　基于 SVMs 的日语并列关系解析

2.1　日语并列关系

日语并列关系是依存关系的一种，如下句所示：

私がはしみにしたのは、美味しい料理と、ハワイの美しい景色だ.

(I am interested in delicious food and Hawaii's beautiful scenery.)

以上例句中两个标着重点的文节为存在并列关系的文节，其中表明句中存在并列关系的文节为并列关键字(句中第一个标着重点的文节). 并列关系解析是解析句中文节间的并列关系，应首先识别句中是否存在并列关键字.

本文基于 SVMs 识别并列关键字，简单地把训练语料中是并列关键字的文节作为正例，把不是并列关键字的文节作为负例.

2.2　日语并列关系解析

如果句中存在并列关键字，解析此并列关键字与其右方的哪个文节存在并列关系. 本文基于 SVMs 进行并列关系解析. 选取训练语料中具有并列关系的文节对作为正例，将并列关键字与其右方不具有并列关系的各个文节对作为负例. 因为在日语中，文节倾向依存于距离其较近的文节，将距离并列关键字最近的依据下式类标为正的文节作为并列关键字的依存对象.

$$Para(ij) = \mathrm{sgn}\Big(\sum_{m,n;\, f_{mn} \in S_{\mathrm{vs}}} \alpha_{mn} y_{mn} k(\boldsymbol{f}_{mn}, \boldsymbol{f}_{ij}) + \boldsymbol{b}\Big) \tag{6}$$

式(6)中文节 i 为并列关键字，文节 j 为句中文节 i 后方的所有文节. 文节 m、n 为支持向量文节对. $\boldsymbol{f}_{mn} = (f_1 \ \cdots \ f_n) \in \boldsymbol{R}^n$ 为文节 m 和文节 n 的特征值，$y_{mn} \in \{$具有并列关系，不具有并列关系$\}$.

3　日语依存关系解析

3.1　日语依存关系解析模型

日语依存关系解析是解析句中各文节间的依存关系. 句子 S 表示为带有特征的文节列 $B = \{b_1, b_2, \cdots, b_m\}$，句子 S 的依存关系为有序集合 $D = \{Dep(1), Dep(2), \cdots, Dep(n-1)\}$. 其中，$Dep(i)$ 为文节 $b_i (i = 1,\cdots,n-1)$ 所依存的对象文节号.

3.2　日语依存关系解析算法

目前，工藤拓的组块逐步应用算法成为解析精确率与效率较高的一种日语依存关系解析算法. 组块逐步应用算法最早用于英语依存关系解析，工藤拓将其应用于日语依存关系解析. 其解析过程如下：

(1) 对于待分析句子的所有文节，因为依存关系没有确定，所以全部赋予 O 标识.

(2) 除了最后的文节，计算所有赋予 O 标识的文节与其紧后方文节是否存在依存关系. 如存在依存关系，将这个文节的标识改为 D.

(3) 在所有紧跟在标识为 O 的文节后的文节中，如果其标识为 D，依据依存关系不交叉原理将其删除.

(4) 当只剩下句末一个文节时，结束；否则，返回至(2).

组块逐步应用算法巧妙地利用了日语中大多数文节依存于其紧后方文节的特点，通过判断句中每个文节是否依存于其紧后方文节解析整个句子. 但是，当句子较长时，距离较远的依存关系的解析正确率却不是很高. 较长的日语句子中往往

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

存在并列关系，存在并列关系的两个文节常常距离较远，所以首先解析并列关系，然后基于并列关系解析句子中剩余文节的依存关系．基于并列关系的日语依存关系解析过程如下：

（1）判断句中是否存在并列关键字．如果不存在并列关键字转至上述组块逐步应用算法解析整个句子．

（2）确定与此并列关键字存在并列关系的文节，赋予并列关键字 D 标识（表明此文节的依存对象已经确定），与并列关键字存在并列关系的文节 C 标识，其余文节 O 标识．

（3）转至上述组块逐步应用算法解析并列关键字与 C 文节之间的文节，C 文节紧前方的文节无条件依存于 C 文节．

（4）对于并列关键字紧前方的文节，确定其与并列关键字还是与 C 文节存在依存关系．确定后，将并列关键字紧前方的文节标识改为 D．如果此文节与 C 文节存在依存关系，根据依存公理删除并列关键字．

（5）转至上述组块逐步应用算法解析句中剩余文节．

解析例如图1（标下线的文节为存在并列关系的文节）．解析结果如图2．

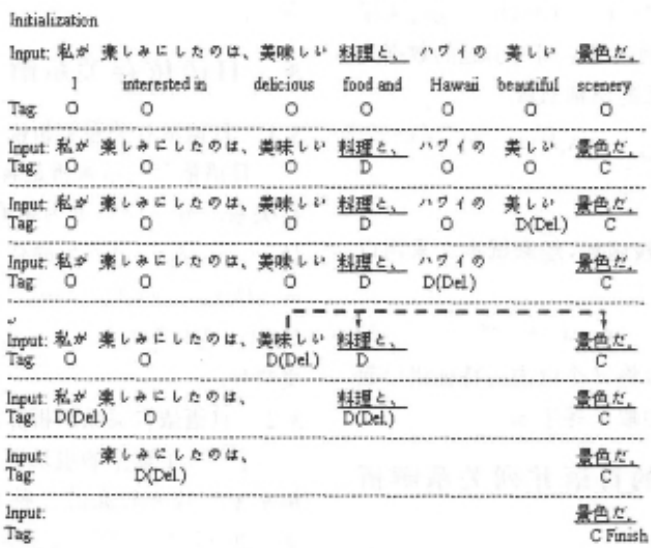

图1　日语依存关系解析例

Fig. 1　Example of Japanese dependency parsing process

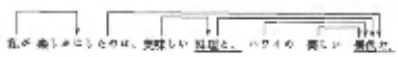

图2　依存关系解析结果

Fig. 2　Parsing result with parallel model

对于并列关键字紧前方的文节 q，依据以下公式确定其依存对象文节：

$$Dep_Seg(q) = \arg\max_{p}\Big(\sum_{m,n;f_{mn}\in S_{vs}} \alpha_{mn} y_{mn} k(\boldsymbol{f}_{mn}, \boldsymbol{f}_{qp}) + \boldsymbol{b}\Big) \tag{7}$$

式中 p 为并列关键字以及并列关键字的依存对象文节．在日语中，文节 q 有时依存于并列关键字，有时依存于并列关键字的依存对象文节．基于组块逐步应用算法解析文节 q 的依存对象时，首先解析其与并列关键字的依存关系．具有并列关系的各文节在语法上非常相似，经解析如果文节对 q 与并列关键字的类标为正，则并列关键字被确认为文节 q 的依存对象．而此类依存关系常常是错误的．

4　实验结果及分析

4.1　实验环境

实验采用京都大学的语料库（Version3.0）的一部分，训练语料为 01-01、03-09，测试语料为 01-10．

用于并列关键字识别的特征如表1所示．依存关系解析采用的特征包括静态特征和动态特征，如表2、3所示．在这里，主词是指文节内除了

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

词性是特殊、助词、结尾词，最靠近文末的词．语形是指文节内除了词性是特殊，最靠近文末的词．表形为字典读音．

表 1　用于并列关键字识别的特征

Tab. 1　Features used in parallel key recognition

主词	语形	位于文节最右侧的词或标点
表形，词性，词性细分类，活用，活用形	表形，词性，词性细分类，活用，活用形	表形或标点，词性，词性细分类，活用，活用形

表 2　用于依存关系解析的静态特征

Tab. 2　Static features used in the dependency parsing process

前 / 后文节	文节间
主词表形，主词词性，主词词性细分类，主词活用，主词活用形，语形表形，语形词性，语形词性细分类，语形活用，语形活用形，有无括弧，有无句读，文节的位置（文头、文中、文末）	距离（1，2 ～ 5，6 以上），所有助词，括号有无，句读有无

表 3　用于依存关系解析的动态特征

Tab. 3　Dynamic features used in the dependency parsing process

依存于依存对象或依存源的文节	依存对象所依存的文节
语形词性是助词、副词、连体词、接续词则取词本身；如果语形有活用形，则取活用形；其他的情况取词性和词性细分类	主词的词性和词性细分类

动态特征如图 3 所示．在依存关系解析过程中，已经获得的依存关系可以作为动态特征使用．当判定两个文节的依存关系时，具体考虑 3 种动态特征：(1) 依存于依存对象的文节 A；(2) 依存于依存源的文节 B；(3) 依存对象所依存的文节 C．

图 3　3 种动态特征

Fig. 3　Three types of dynamic features

4.2　实验结果

测试语料中包括含有并列关系的句子与不含有并列关系的句子．并列关系正确率与依存关系正确率（包括并列关系）为衡量解析器性能的重要标准．依据下式分别测试并列关系正确率与依存关系正确率：

$$\text{并列关系正确率} = \frac{\text{正确识别的并列关系个数}}{\text{并列关键字个数}} \tag{8}$$

$$\text{依存关系正确率} = \frac{\text{正确识别的依存关系个数}}{\text{所有的依存关系个数}} \tag{9}$$

本文提出的并列算法与基于 NN-LSVM 的组块逐步应用算法[4] 的解析结果如表 4 所示．基于 NN-LSVM 的组块逐步应用算法是本文提出的在组块逐步应用算法的基础上改进的一种算法．此算法通过删减训练集中对分类不起作用或起反作用的样本提高解析精度，降低解析时间．本文也基于 NN-LSVM 对训练集进行了删减．

并列算法的并列关系正确率较组块逐步应用算法有较大的提高．依存关系正确率也较组块逐步应用算法有所提高，提高的幅度没有并列关系提高的幅度大，其原因是语料中不是每个句子都含有并列关系．

表 4　并列算法与基于 NN-LSVM 的组块逐步应用算法的解析结果

Tab. 4　Result of parallel model vs. cascaded chunking model using NN-LSVM

算法	并列关系正确率 /%	依存关系正确率 /%
并列算法	79. 13	90. 03
基于 NN-LSVM 的组块逐步应用算法	76. 33	89. 86

4.3　与以往方法的比较

本文提出的方法与以往方法（基于 NN-LSVM 的组块逐步应用算法[4]、工藤拓的组块逐步应用算法[3]、内元清贵的最大熵法[1]、内元清贵的考虑后方文节法[2]）的解析正确率如表 5 所示．

因为并列关系语料只是依存关系语料中的一少部分，所以本文算法取得的结果优势并不明显．其他方法没有单独解析并列关系，而把并列关系当做普通的依存关系进行解析，所以也具有一定的解析并列关系的能力．本文提出的方法是在工藤拓的组块逐步应用算法的基础上改进的一种方法，试验结果表明无论是依存关系解析正确率还是句子解析正确率都较原方法有所提高，并高于考虑后方文节法与最大熵法．

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表5 与以往方法的比较

Tab.5 Comparison with the related work

算法	训练语料(天数)	依存关系正确率/%	句子正确率/%
并列算法	京都大学语料(8)	90.03	49.48
基于NN-LSVM的组块逐步应用算法[4]	京都大学语料(8)	89.86	49.14
组块逐步应用算法[3]	京都大学语料(8)	89.29	47.53
最大熵法[1]	京都大学语料(8)	87.14	40.60
考虑后方文节法[2]	京都大学语料(8)	87.93	43.58

5 结 论

本文提出一种基于SVMs解析日语并列关系的方法.该模型基于单独的并列关系解析器解析并列关系,有效地提高了并列关系解析正确率.将并列关系用于随后的依存关系解析,使得依存关系解析正确率也得到了提高.使用京都大学语料库进行训练并测试,结果表明本方法有效地提高了并列关系解析以及依存关系解析的正确率.

参考文献:

[1] 内元清貴,丸 根,井佐原均.最大エントロピー法に基づくモデルを用いた日本語係り受け解析[J].情報処理学会論文誌,1999,40(9):3397-3407

[2] 内元清貴,村田真樹,丸 根,等.後方文節を考慮した係り受けモデル[J].自然言語処理,2000,7(5):33-17

[3] 工藤拓,松本裕治.チャンキングの段階適用による係り受け解析[J].情報処理学会論文誌,2002,43(6):1834-1842

[4] 周惠巍,李 巍,黄德根.基于NN-LSVM的日语依存关系解析[C]//孙茂松,陈群秀.自然语言理解与大规模内容计算.北京:清华大学出版社,2005:295-301

[5] 黒橋禎夫,長尾真.長い日本語文における並列構造の推定[J].情報処理学会論文誌,1992,33(8):1022-1031

[6] CORTES C, VAPNIK V. Support vector networks [J]. **Machine Learning**, 1995,**20**(3):273-297

[7] 春野雅彦,白井諭,大山芳史.決定木を用いた日本語係り受け解析[J].情報処理学会論文誌,1998,**39**(12):3177-3186

[8] JOACHIMS T. Text categorization with support vector machines: learning with many relevant features [C] // NEDELLEC C, ROUVEIROL C, ed. **Proceeding of ECML-98, 10th European Conference on Machine Learning**. Heidelberg: Springer-Verlag, 1998:137-142

[9] JOACHIMS T. Transductive inference for text classification using support vector machines [C] // BRATKO I, DZEROSKI S, ed. **Proceedings of ICML 99, 16th International Conference on Machine Learning**. San Francisco: Morgan Kaufmann Publishers, 1999: 200-209

Japanese parallel relation analysis based on support vector machines

ZHOU Hui-wei*, HUANG De-gen, LI Wei

(Dept. of Comput. Sci. and Eng., Dalian Univ. of Technol., Dalian 116024, China)

Abstract: Japanese parallel relation is one kind of Japanese dependency relation. It is difficult to recognize the parallel relation since conventional Japanese dependency parsers do not parse parallel relation in particular. A method of Japanese parallel relation analysis is proposed. The proposed model identifies parallel key based on support vector machines (SVMs). If there is a parallel key in a sentence, the parallel relation is parsed based on SVMs. The parallel relation is also used in the dependency analysis later. Experiments using the Kyoto University Corpus show that the system can obtain higher precision in Japanese parallel analysis and dependency analysis.

Key words: Japanese parallel relation analysis; support vector machines; parallel key; dependency analysis

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net