基于字符级特征的日文依存句法自动分析算法

文益民[1, 4]，赵 海[2, 3]，李 健[4]，黄晗文[4]

(1. 湖南大学 电气与信息工程学院，湖南 长沙，410082；
2. 香港城市大学 中文翻译及语言学系，香港 999077；
3. 苏州大学 计算机科学与技术学院，江苏 苏州，215006；
4. 湖南工业职业技术学院 信息工程系，湖南 长沙，410208)

摘 要：基于字符在词位的特定性位置能起标志性作用，使用反映日语的语言倾向性的特征分析依存句法，给出基于字符级特征的依存句法自动分析算法。该算法使用词首的头 2 个字符、词尾的最后 1 个字符以及词尾的最后 2 个字符这 3 种类型的字符级特征加强分析器的学习。采用第 1 种类型的特征，反映日语的词汇形态特点，采用后 2 种类型的特征，则部分反映了日语表达重心后置的语言特性。在 CoNLL-2009 日语语料库上进行实验以及进行实际评估，结果表明：与基线特征相比，该算法有效地提高分析器的 UAS 指标(即无标记依存关系的正确率)和 LAS 指标(即考虑带标记的依存关系的正确率)，大大提高分析器的竞争力。

关键词：计算机应用；依存句法分析；字符级特征
中图分类号：TP391 文献标识码：A 文章编号：1672−7207(2009)04−1035−05

Japanese dependency parsing based on character-level features

WEN Yi-min[1, 4], ZHAO Hai[2, 3], LI Jian[4], HUANG Han-wen[4]

(1.College of Electrical and Information Engineering, Hunan University, Changsha 410082, China;
2. Department of Chinese Translation and Linguistics, City University of Hongkong, Hongkong 999077, China;
3. School of Computer Science and Technology, Soochow University, Suzhou 215006, China;
4. Department of Information Engineering, Hunan Industry Polytechnic, Changsha 410208, China)

Abstract: Based on the indicative impact of character located at a special position in a word, an algorithm was proposed to make use of character-level features that reflect the characteristics of Japanese to enhance the performance of Japanese dependency parsing. Three character-level features denoted by the first two characters, the last character, and the last two characters inside a word were adopted. The first type of features was used for the morphological purpose, and the latter two demonstrate that the emphasis in Japanese trends to locate at the end of an expression segment in the sentence. The results of experiments and evaluation on the Japanese corpus from CoNLL-2009 shared task show that the accuracy of Japanese dependency parser can be effectively improved by using the proposed features.

Key words: computer application; dependency parsing; character-level feature

收稿日期：2008−09−05；修回日期：2008−11−25
基金项目：国家“ 863 ”计划项目(2007AA04Z244)；国家自然科学基金重点资助项目(60835004)；湖南省博士后科研资助专项计划项目(2008RS4005)；湖南省教育科学 “ 十一五 ” 规划课题(XJK08BXJ001)
通信作者：文益民(1969−)，男，湖南桃江人，博士，副教授，从事机器学习、自然语言理解及图像处理等研究；电话：0731-88539059；E-mail: ymwen2004@yahoo.com.cn

在多语种的句法分析中，对日文的依存句法分析较早[1]。目前，基于数据驱动的依存句法分析分为 2 类：基于图模式(Graph-based)的依存句法分析模型和基于转换(Transition-based)的依存句法分析模型。基于图模式的依存句法分析模型需要为依存树制定 1 个概率函数，而此概率函数通常分解为各条弧概率的累计之和。各类基于图的依存句法分析模型的区别主要体现在依存树概率函数的定义和搜索算法。基于图的依存句法分析模型见文献[2−4]。基于转换的依存句法分析模型通过执行连续的多个动作(Action)或转换(Transition)构建依存树[5]。各类基于转换的依存句法分析模型的区别主要在于动作或转换策略的不同和动作概率所使用的机器学习模型(如支持向量机(SVM)和最大熵)不同[6]。基于转换的依存句法分析模型见文献[7−10]。与其他句法分析方法一样，早期日文的依存句法分析技术是基于概率方法的不确定分析。Duan等[10−11]证明了确定性的句法分析性能更优，同时效率更高。他们提出的确定性方法在很大程度上可以归结为移进规约方法，也就是典型的基于转换的方法。但是，它们与典型的Nivre算法存在差别，如文献[7]中的算法需要 4 种分析操作，而文献[10]中算法只需要 3 种。Iwatate等[12]认为：基于图模式的方法在日文分析中不具有性能优势，这实际上佐证了日文的依存句法分析更适合用基于转换的方法来进行处理的观点。在此，本文作者对日文依存句法分析算法进行改进。考虑使用字符(包括假名、汉字、标点等多种日文书写成分)级的特征来加强分析器学习。该研究不同于已有的那些将重点集中于学习框架修正的工作，而是强调系统地使用反映日语的语言倾向性的特征。

1 学习模型

采用Nivre方式的移进规约框架作为基本的句法学习框架[7]。基于转换的句法分析方法实际上是一种词对的分类方法，仅限于处理投影型的输入句子。日语就句法统计特性来说，恰恰是一种高度投影性的语言，因此，基于转换的学习模型适合于其结构学习。

在 Nivre 框架的句法分析中，分析器按照一定方向扫描输入的句子，同时，保存已经过分析的部分句子的状态。具体来说，使用 1 个栈来维护已经得到分析的部分句子。在每个状态，分析器检查 2 个词：1 个词位于栈顶(通常用 TOP 表示)，另外 1 个词位于尚未处理的句子的首部(通常用 NEXT 表示)。根据分类器的输出，来决定是否在这个词对之间建立一定的依存关系。若用弧来表示依存关系，则可用 2 种弧来表达 TOP 和 NEXT 之间的关系，左弧代表后者是中心词(上位词)，右弧代表前者是中心词。分析器还需要移进和规约 2 个操作来完成扫描句子的操作。因此，在 1 个无标记的依存分析中，需要 4 类操作。

a. 左弧：增加 1 条从 NEXT 到 TOP 的弧，同时，将 TOP 弹出栈。

b. 右弧：增加 1 条从 TOP 到 NEXT 的弧，同时，将 NEXT 推入栈。

c. 规约：将 TOP 弹出栈。

d. 移进：将 NEXT 推入栈。

在建立左弧和右弧的策略上，选用立即建立的策略，也就是说，若分析动作的分类器判断弧存在，则立即建立弧，而不是滞后等待其他的判别条件成立。在扫描句子的方向上，使用正向扫描，也就是从左至右扫描。

日文的依存关系具有少数的类别标记。左弧有 3 种：D，I 和 P；右弧只有 1 种：D。为了建立 1 个单步系统，1 步获得完整的句法输出，需要扩展建立左右弧的无标记操作到类别相关的弧操作，这样，共有 6 个类别的分类任务需要分析器的分类器来完成。

日语是一个高度投影化的语言，但是，依然有少量的非投影输入句子存在，为了处理这部分特殊情形，使用经过轻微变形处理的伪投影化技术[13−14]。标准的伪投影化技术是将非投影型依存的中心词转移，具体来说，重置这个中心词到原始中心词的中心词，同时，附加额外的依存类别编码，以便在解码时恢复原来的非投影型依存关系。但是，这一编码/恢复操作在另外一种高度投影化的语言——英语上并不能使其性能得到额外提升，而且有时会导致始终无法消解的非投影型依存关系。因此，需要使用 2 个简化策略来解决这个问题：第 1 个是直接重置中心词而不需要任何附加的依存类别编码，也就是在解码时不再恢复原始的非投影型依存；第 2 个是在出现无法消解的非投影型依存时，搜索最近的词，试图将中心词转移到该词，一旦找到，则立即结束搜索。实验证明，采用这种操作总是能消解所遇到的非投影依存关系。

传统的基于转换的依存分析所使用的分类器通常是支持向量机(SVM)或者其他基于边界或者基于记忆

的方法。但是，这些分类器在依存学习中表现为训练时间长和解码低效，甚至比基于图模式的分析器要慢很多。Zhao等[14]使用最大熵作为分类器，并证明将最大熵作为基于转换的依存分析的分类器，使分类器性能更优。因此，本文继续使用这一分类工具。

依存句法分析所依赖的特征涉及多重因素，为了方便表示，定义了一组基本记号，如表 1 所示。

表 **1**　用于表示特征的基本记号

Table 1　Basic marks of features

标记	含义
s	栈顶词
s_{-1},s_{-2},…	栈顶下方第 1、第 2 个词，等
I	未处理部分第 1 个词
i_{+1},i_{+2},…	未处理部分第 2、第 3 个词，等
dprel	依存关系类别标记
h	中心词
lm	最左下位词
ln	左边最近的下位词
rm	最右下位词
rn	右边最近的下位词
form	词形
lemma	词的原形
pos	词性
−	的。例如，s.rm 代表栈顶词的最右下位词
+	串加法，合并 2 个串为 1 个串作为特征

根据表 1 中的记号，得出使用的部分基线特征集，如表 2 所示。

表 **2**　部分基线特征集

Table 2　Set of partial baseline features

标记	含义
i.existVerb	未处理词所在短句是否存在动词
x.isComma	x是否是逗号，其中x代表i或者i_{+1}
x.isMD	x的词性是否是MD，其中x代表i或者i_{+1}
i.posSeqOfChildren	未处理词所有下位词的词性合并构成串
i.posSeqOfChildrenRel	未处理词所有下位词的词性合并构成串，未处理词自身同时按照顺序包含进入其中。
i_{-1}.form+i.form	未处理词的词形相关的特征
i.form+i_{+1}.form	
i_{+1}.lemma	未处理词的词的原形相关的特征
i_{+1}.lemma +i_{+2}. lemma	
i.pos+i_{+1}.pos	未处理词的词性相关的特征
i_{+1}.pos +i_{+2}.pos	—
i_{+1}.pos +i_{+2}.pos+ i_{+3}.pos	—
i_{+3}.pos	词性
x.isComma	x代表s,s_{-1}或者s_{-2}

2　基于字符和粗词性的特征

引入字符特征的目的在于日语书写方式和汉语具有一定的相似性。无论是假名还是汉字，字符都在词位的特定位置发挥了标志性作用。特别是，日语中汉字的使用较广泛，单一的汉字具有独特的含义。因此，字符特征有可能揭示部分有益的启发信息用于句法分析。

考虑 3 种类型的字符级特征：第 1 个是词首的头 2 个字符，表示为 firstTwoChar；第 2 个是词尾的最后 1 个字符，表示为 lastChar；第 3 个是词尾的最后 2 个字符，表示为 lastTwoChar。注意到前 1 个字符单独不作为特征字符，这说明日语是一种重心后置的语言。字符级特征如表 3 所示。

除了考虑字符级特征外，还考虑了字符级特征和粗词性特征的结合。粗词性来自原始词性的简化。定义 2 种日语的粗词性：原始词性的第 1 个字母，记为 cpos1，原始词性的头 2 个字母，记为 cpos2。例如，若原始词性为 NNP，则 cpos1=N，而 cpos2=NN。经证明，粗词性有助于光滑常规的词性特征。

3　评　估

为了系统评估所提出的算法性能，使用最新的 CoNLL-2009 国际评测的日语语料作为评估语料，验证基于字符的特征集的有效性[14]。CoNLL-2009 国际评测的日语语料来自京都大学文本语料库(Kyoto University Text Corpus 4.0)，其人工标注可以从网上免费下载[15]，但是，其文本本身需要得到《每日新闻》的授权。

表 3 字符级特征

Table 3 Character-level features

标记	含义
i_{+1}.form.lastChar+ i_{+2}.form.lastChar + i_{+3}.form.lastChar	未处理词词形的最后 1 个字相关特征
i_{+2}.form.lastChar	—
i_{+2}.form.lastTwoChar	未处理词词形的最后 2 个字相关特征
i.lnVerb.form.lastChar	未处理词左边最近下位动词词形的最后 1 个字
i.rmVerb.form.lastChar	未处理词最右下位动词词形的最后 1 个字
s.form.firstTwoChar+ s.form.firstTwoChar$_{+1}$	栈顶词的词形头 2 个字相关的特征
s.form.firstTwoChar$_{+1}$+ s.form.firstTwoChar$_{+2}$ + s.form.firstTwoChar$_{+3}$	—
s.form.lastChar $_{-1}$+ s.form.lastChar	栈顶词的词形最后 1 个字相关的特征
s.form.lastChar + i.form.lastChar	—
s.form.lastChar + i_{+1}.form.lastChar	—
s.form.lastTwoChar + i.form.lastTwoChar	—
s.form.lastChar + s.h.form.lastChar	—

CoNLL-2009 国际评测提供的文本已经转化为基于词的依存形式，因此，可以直接使用这一语料。该日语语料的附加句法训练集(包含 33 257 句)被用作训练集，再将用于语义学习的训练集(4 393 句，同时包括句法信息)的前半部分作为开发集，后半部分作为测试集。不用标准的开发集的原因是该开发集规模太小，仅有 250 句，会使评估结果的统计显著性不强。

评估度量使用依存句法分析领域通行的 UAS(即无标记依存关系的正确率)和 LAS(即考虑带标记的依存关系的正确率)，实验结果如表 4 所示。可以看到，尽管使用基线特征已经获得较高的性能，加入字符特征，UAS 和 LAS 指标正确率提升近 1%，但同时加入字特征和粗词性特征时，系统性能提升超过 1%。

表 4 实验结果比较

Table 4 Experimental results

	正确率/%	
	无标记依存关系(UAS)	考虑带标记的依存关系(LAS)
基线特征	92.10	91.33
基线特征+字符特征	93.05	92.27
基线特征+粗词性特征	92.54	91.69
基线特征+字符特征+粗词性特征	93.21	92.40

表 5 所示为在 CoNLL-2009 上真实测试集的评测结果。在 CoNLL-2009 的评测中，没有使用粗词性特征，仅使用了字符特征就获得了接近最优的结果。同时，CoNLL-2009 日语句法的测试集较小，约 400 句。因此，表 4 中结果的统计显著性更优。

表 5 CoNLL-2009 日语依存句法的评测结果比较

Table 5 Evaluation of Japanese dependency parsing on CoNLL-2009

实际评测排序	正确率/%	
	无标记依存关系(UAS)	考虑带标记的依存关系(LAS)
CoNLL-2009 最好成绩	—	92.57
CoNLL-2009 次好成绩	—	92.34
CoNLL-2009 第 3 成绩(提出的算法)	—	92.32
CoNLL-2009 第 4 成绩	—	92.21
CoNLL-2009 第 5 成绩	—	91.71

4 结 论

a. 基于字符在词位的特定性位置的标志性作用，使用反映日语的语言倾向性的特征分析依存句法，使用词首的头 2 个字符、词尾的最后 1 个字符和词尾的最后 2 个字符这 3 种类型的字符级特征来加强分析器的学习，给出了基于字符级特征的依存句法自动分析算法。

b. 在 CoNLL-2009 日语语料库上的实验以及 CoNLL-2009 上的真实测试集的评估结果表明，加入字符级特征使 UAS 指标和 LAS 正确率提升近 1%，同

时，加入字特征和粗词性特征时系统性能提升超过1‰。因此，与基线特征相比，该算法有效地提高了分析器性能。

参考文献：

[1] Safir K. The syntax of (in)dependence[M]. Cambridge: MIT Press, 2004.

[2] McDonald R, Pereira F, Ribarov K, et al. Non-projective dependency parsing using spanning tree algorithm[C]// Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP). Vancouver: ACL, 2005: 523−530.

[3] Nakagawa T. Multilingual dependency parsing using global features[C]//Proceedings of the Joint Meeting of the Conference on Empirical Methods on Natural Language Processing and the Conference on Natural Language Learning(EMNLP/CoNLL). Prague: A CL, 2007: 952−956.

[4] Carreras X. Experiments with a high-order projective dependency parser[C]//Proceedings of the Joint Meeting of the Conference on Empirical Methods on Natural Language Processing and the Conference on Natural Language Learning(EMNLP/CoNLL). Prague: ACL, 2007: 957−961.

[5] 段湘煜, 赵　军, 徐　波. 基于动作建模的中文依存句法分析[J]. 中文信息学报, 2007, 21(5): 25−30.
DUAN Xiang-yu, ZHAO Jun, XU Bo. Chinese dependency parsing based on action modeling[J]. Journal of Chinese Information Processing, 2007, 21(5): 25−30.

[6] Yamada H, Matsumoto Y. Statistical dependency analysis with support vector machines[C]//Proceedings of the 8th International Workshop on Parsing Technologies. Nancy: ATOLL, 2003: 195−206.

[7] Nivre J. An efficient algorithm for projective dependency parsing[C]//Proceedings of the 8th International Workshop on Parsing Technologies. Nancy: ATOLL, 2003: 149−160.

[8] Titov I, Henderson J. A latent variable model for generative dependency parsing[C]//Proceedings of the Tenth International Conference on Parsing Technologies. Prague: ACL, 2007: 144−145.

[9] Yamada H, Matsumoto Y. Statistical dependency analysis with support vector machines[C]//Proceedings of the 8th International Workshop on Parsing Technologies. Nancy: ATOLL, 2003: 195−206.

[10] Duan X, Zhao J, Xu B. Probabilistic parsing action models for multi-lingual dependency parsing[C]//Proceedings of The Joint Meeting of the Conference on Empirical Methods on Natural Language Processing and the Conference on Natural Language Learning(EMNLP/CoNLL). Prague: ACL, 2007: 940−946.

[11] Kudo T, Matsumoto Y. Japanese dependency analysis using cascaded chunking[J]. Transactions of Information Processing Society of Japan, 2002, 43(6): 1834−1842.

[12] Iwatate M, Asahara M, Matsumoto Y, Japanese dependency parsing using a tournament model[C]//Proceedings of the 22th International Conference on Computational Linguistics. Manchester:ACL, 2008: 361−368.

[13] Nivre J, Jens N. Pseudoprojective dependency parsing[C]// Proceedings of the 43th Annual Meeting on Association for Computational Linguistics. AnnArbor: ACL, 2005: 99−106.

[14] Zhao H, Kit C. Parsing syntactic and semantic dependencies with two single-stage maximum entropy models[C]//Proceedings of the Twelfth Conference on Computational Natural Language Learning. Manchester: ACL, 2008: 203−207.

[15] Kawahara D, Kurohashi S, Hasida K. Construction of a Japanese relevance-tagged corpus[C]//Proceedings of the 3rd International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria: European Language Resources Association, 2002: 2008−2013.