基于规则和统计的日语分词和词性标注的研究[1]

姜尚仆 陈群秀

清华大学信息科学与技术国家实验室计算机科学与人工智能研究部

清华大学计算机系 北京 100084

E-mail: jsp07@mails.tsinghua.edu.cn

摘要：和中文类似，日语的词法分析需要首先进行分词。基于词的方法是日语分词的主流方法。同时，对中文的研究结果表明，词性标注对分词结果的正确性有帮助，这点在日语中也得到了证实。我们提出了一种基于规则和统计的日语分词和词性标注方法，使用基于单一感知器的联合分词和词性标注算法进行训练和解码，并加入了词语的邻接属性特征。实验结果表明，这种方法无论是分词准确率还是分词加词性标注的准确率都比原有的基于字和词的混合 HMM 算法更高。我们已将这种方法应用到我们的日汉机器翻译系统中。

关键词：日汉机器翻译系统，日语分词，日语词性标注，联合分词

Study on Japanese Word Segmentation and POS Tagging Based on Rules and Statistics

JIANG Shangpu, CHEN Qunxiu

Computer Science and Artificial Intelligence Division, National Laboratory for information Science and Technology, Tsinghua University

Department of Computer Science and Technology, Tsinghua University, Beijing, 100084

E-mail: jsp07@mails.tsinghua.edu.cn

Abstract: Like that of Chinese, Japanese morphological analysis starts with word segmentation. Word-based approach is the mainstream on Japanese word segmentation. Meanwhile, according to the study on Chinese, POS tagging results are helpful to the correctness of word segmentation. This conclusion is also substantiated on Japanese. We propose a Japanese word segmentation and POS tagging approach based on rules and statistics, which uses a single perceptron based joint word segmentation and POS tagging algorithm for training and decoding, and is added with the features of adjacency attribute. The experiment shows that the new approach is better performed than the hybrid character and word based HMM algorithm. We have already applied this approach into our Japanese-Chinese machine translation system.

Keywords: Japanese-Chinese machine translation system, Japanese word segmentation, Japanese POS tagging, joint word segmentation

1 前言

规则和统计相结合的研究方法是当前计算语言学界主流的研究方法，是今后发展的倾向。

[1] 本文相关研究得到国家 863 计划重点项目(项目号:2006AA010109)资助。

我们对基于规则和统计的日语分词和词性标注进行了研究，提出了一个具有高分词准确率的基于规则和统计的日语分词和词性标注算法。

日语分词和词性标注是以日语为源语言的机器翻译系统的重要组成部分，是机器翻译系统的第一个模块，它的正确率很大程度上影响了机器翻译结果的好坏。此外，日语分词和词性标注也广泛应用于日语的各种自然语言处理的任务中。因此，日语分词和词性标注算法的研究有着重要的意义。和中文类似，日语的词语之间没有明显的分隔符，因此日语词法分析也包括了分词和词性标注两个部分。

对于词性标注，近年来提出了很多算法，包括 HMM[5][9]、ME[6][7]和 CRFs[8]等。中文分词通常被认为是特殊的一类序列标注，因而采用字标注的方法，通过对每个字标注 B/I[3]或者 B/M/E/S[4]来实现词语切分。然而，在日语分词中，这种方法并不能取得很好的效果[1]，这是由于日语词语相对较长，而字标注的窗口较小，往往不能获取足够的上下文特征。而通常来说，基于词典的日语分词算法，即使是最大匹配，也能获得 80%以上的正确率。此外，词典能给我们提供词性、邻接关系、词形变换等很多先验知识，这些都是字符特征无法获得的。对于未登录(OOV)词，我们也可以通过抽取词语中的字符特征来进行识别[15]。因此，一些基于词的分词算法成为了日语分词的主流算法[1][10][11][12]。

另一方面，传统的分词和词性标注方法将两个步骤串行执行，带来了误差累积的问题。近年来，很多研究都在尝试将两者合二为一[2][13][14]。实验证明，联合的方法无论是分词正确性还是词性标注正确性都有了一定提高。

本文提出了一种基于规则和统计的日语分词和词性标注方法，利用词典中词条的特征，并且采用了联合分词和词性标注的方法。我们采用了基于单一感知器的算法[9][12][13]对这些特征进行训练和解码。我们原有的日语分词系统采用的是基于字和词特征的混合 HMM 算法，实验结果表明：新的方法可以获得比原有方法更好的效果。我们已经将此方法应用到我们的日汉机器翻译系统中。

2 基于规则的日语分词和词性标注研究

我们的日语分词和词性标注首先是基于规则(邻接属性表)的。基于规则的方法优点是事先总结归纳好的邻接属性可以覆盖绝大部分的语言事实，准确性高并且计算比较简单，速度快。

2.1 日语词语特征

和中文类似，日语的词语间没有明显的分隔符，然而，日语又具有一些有别于中文的特点，了解这些特点，对于进行较好的日语词法分析有着重要的意义。除了上文中提到的，日语词语的平均长度要长于中文词语以外，日语词语还具有如下一些特征：

1) 日语依靠助词或者助动词的粘着来表示每个词语在句中的成分，因此助词和助动词的正确识别对词法分析的正确性非常重要。

日语中助词(particle)和助动词(auxiliary verb)可以统称为附属词，从语法功能上和中文的助词比较接近，但是数量要远远多于中文中的助词。更重要的是，附属词多由平假名组成。日语中有三种字符类型：平假名(hiragana)、片假名(katakana)和汉字。汉字常用于实词，而且汉字数量众多，比较不容易产生切分和词性标注的歧义。片假名一般用于外来词汇，出现较少且分界明显。

而平假名一共只有 50 多个字符，且广泛存在于各种词性中，尤其是在附属词中数量繁多且词语长度较短，词语边界的划分更加困难。因此，在日语的词法分析中，附属词通常会词汇化(lexicalized)，即词语本身作为和词性类似特征来使用[1][11]。

2) 日语的动词、形容词、形容动词和助动词有活用形。

对于属于这些词性的词，其原始形态被成为基本形。而根据这些词在句子中的不同成分和作用，又有连体形、连用形、未然形、终止形、假定形、命令形、推量形等不同的活用形。

词语的活用会影响到邻接关系。例如，连体形后面通常会连接体言。这种邻接关系对可以确定一些分词或者词性标注的结果，因此，将这种邻接关系加入特征模板是很有必要的。至于如何将邻接关系加入特征模板，可详见 2.2 节。

3) 日语句子成分多数没有严格的次序，可以灵活放置，有些成分则经常可以省略。

除了谓语一般位于句子的末尾，日语句子中的其他成分没有严格的顺序。这个特征更多的应用于句法分析中，但也从另一个方面说明了助词和助动词对于词法分析的重要性。

2.2 邻接属性

词性标注算法通常使用 n-gram 模型来表示连续 n 个词语词性之间的相关性。然而，仅仅使用词性的 n-gram 模型表示能力有限，往往不能描述复杂的语法性质。ME 和 CRFs 成功的解决了这个问题，通过引入各种复杂的、可重叠的特征模板，实现了性能的提高。例如，在中文词性标注中，字符的特征被广泛应用[3][4]。

同样，在日语词法分析中，仅仅依靠词性的 n-gram 模型是不够的。对于两个相邻的词语，一些细化的词类别，例如动词、形容词、形容动词和助动词的活用形类型，人名、地名等命名实体等都可以作为邻接关系的特征来使用。

我们的日语分词系统最早是基于规则即基于邻接表的。邻接表是事先根据语言学规律归纳总结出来的一套用来表示日语相邻词语之间可能的邻接组合的规则。对于每个词语，都指定一个左邻接属性和一个右邻接属性。对于任意两个相邻的词语，后一个的左邻接属性和前一个的右邻接属性的组合决定了这两个词语之间的匹配程度，邻接表就是用来表示这种匹配程度的。在我们的邻接表中，左邻接属性和右邻接属性的种类分别为 102 种和 99 种。例如，在我们的词典中有这样的词条：

五　8　6
分　11　66

表示“五”(五)的右邻接属性和左邻接属性分别为 8(代表“JRN8 数詞”)和 6(代表“JLN6 数詞”)，“分”(分钟)的右邻接属性和左邻接属性分别为 11(代表“JRNB 单位”)和 66(代表“JSF9 後助数詞”)。如果“五”的右邻接属性 8 和“分”的左邻接属性 66 的组合在邻接表中存在，则“五分”就成为一个可能的词语搭配。

而在基于统计的算法中，可以通过训练每种邻接属性的组合的参数来实现软匹配规则。由于邻接属性通过人工分析了各种可能会影响相邻词语搭配的特征，因此能实现较好的分词和词性标注结果，同时又不会造成过拟合。

2.3 词典构成

我们使用的词典由名词词典、形容词词典等 18 部分类词典组成的大规模的词典，共有词条

72.7 万。每个词条除了词语本身以外，还记录了词语的词性、左邻接属性和右邻接属性。对于动词、形容词、形容动词和助动词这些有活用形的词语，我们根据一个动词基本形词典，通过活用形变化规则，生成其所有活用形的词条。例如下面的动词词条：

あたら　56　9　あたる

表示基本型“あたる”的未然形为“あたら”，其右邻接属性和左邻接属性分别为 56(表示“JEM5 未然 a-nal”)和 9(表示“JLV1 動詞”)。

3 基于统计的日语分词和词性标注研究

由于基于规则的方法灵活性较差并且对语言事实的覆盖面不够全面等固有缺陷，结合基于统计的方法往往能为结果带来较大提升。因此，我们使用了基于统计的感知器算法[9]

3.1 感知器算法

感知器算法是用来对 HMM 进行辨别训练的一种方法。它是 CRFs 的一种替代算法，并且具有和 CRFs 类似的性能。这种方法被广泛应用与词性标注[9]和中文分词[12][13][14]中。

给定一个句子 x，其正确的分词和词性标注结果 $F(x)=\arg\max_{y\in GEN(x)}(score(y))$。其中 GEN(x)表示所有的可能解，$score(y)=\Phi(y)\cdot\overline{\alpha}$，$\Phi(y)\in R^d$ 是 y 的全局特征向量，d 是特征数量，$\overline{\alpha}$ 是特征向量对应的参数。感知器的训练算法如图 1 所示。

输入：训练集(X, Y)

初始化：$\overline{\alpha}=0$

算法：

For t = 1..T, i = 1..N

计算 $z_i=\arg\max_{y=GEN(x_i)}(\Phi(y)\cdot\overline{\alpha})$

如果 $z_i\neq y_i$，则 $\overline{\alpha}=\overline{\alpha}+\Phi(y_i)-\Phi(z_i)$

输出：$\overline{\alpha}$

图 1 感知器训练算法，根据[9]

我们的分词和词性标注方法所选取的特征模板设计如表 1 所示。在基本模板中，我们对未登录词使用了基于字符的特征，对助词、助动词和标点等词语进行了词汇化(见 2.1 节)，还使用了词性的 n-gram 特征。我们还加入了邻接属性(见 2.2 节)的特征，由于邻接属性和词性基本上是多对一的关系，因此我们没有使用邻接属性和词性结合的特征。

表 1 特征模板 t=<p, l, r, w>，t’和 t”是分别向前两个词语的特征

其中 p, l, r, w 分别表示词性、左邻接属性、右邻接属性和词的原形（基本形）。

基本模板	Unigram	<p> 词典词：<p, w> 未登录词：<w 的长度>, <首字符>, <首字符, p> <尾字符>, <尾字符, p>, <字符类型[2]>, <字符类型, p>

[2] 字符类型包括平假名、片假名、汉字、西文字符、数字和标点。

	Bigram	<p', p> w'词汇化：<p', p, w'> w 词汇化：<p', p, w>
	Trigram	<p", p', p>
邻接属性	Unigram	<l>, <r>, <l, r>
	Bigram	<r', l> w'词汇化：<r', l, w'> w 词汇化：<r', l, w>

3.2 解码

对基于感知器的中文分词算法进行解码可以使用集束搜索(beam search)算法[12]。而对与单一感知器联合分词和词性标注，可以使用多重集束搜索(multi-beam search)算法[13]进行解码，从而解决使用集束搜索时由于搜索空间过大导致的准确性下降的问题。集束搜索和多重集束搜索作为一种近似算法，通常能得到较优解，它可以用来处理 viterbi 算法难于处理的复杂特征，而且速度较快。

然而，由于我们的特征状态空间比较简单，使用 viterbi 算法不但可以求得最优解，而且速度也不慢。因此，我们使用 viterbi 算法来进行解码，状态转移方程为：

$$score(p', p, r, lex) = \max_{p'', r', w, lex'} \{score(p'', p', r', lex') + uni(p, w) + bi(p', p) + tri(p'', p', p)\}$$

其中，$score(p', p, r, lex)$ 是当前状态的得分，w 是当前词语，p", p', p 是最后三个词语的词性。当 w 需要词汇化时，lex=w；否则 lex=NULL。Uni, bi 和 tri 分别表示当前位置 unigram, bigram 和 trigram 特征的得分。

4 日语分词和词性标注实验结果和分析

我们使用的训练语料来自北京外国语大学的日汉双语语料库，里面的文章来自日语小说原著和翻译。我们从中选取了 7M 字节左右的日语原文，对它进行了预处理，划分出段落 46,730 段，句子 114,228 句。我们分别使用我们原有的分词系统和一个开源的日语分词系统 MeCab[3]对这些句子进行词法分析，其中有 10,475 句切分结果完全一致，在切分不同的句子中，我们取出部分针对句子不一致的部分进行人工校对，再据此修改统计数据，共整理出 11,000 句句子作为训练语料。

我们的测试语料来源于网页，共有 9,154 句日语句子，我们将分词和词性标注结果和人工标注的结果进行比较。我们使用的对比系统是我们原有的日语分词和词性标注系统，其使用的算法是基于字和词混合特征的 HMM 算法。该算法利用词典来识别登录词，利用字特征来识别未登录词，并加入了邻接属性的软规则，然后用 HMM 对分词结果和词性同时进行解码。我们分别

[3] http://mecab.sourceforge.net

测试了使用基本特征模板(见表 1)和基本模板加邻接属性两种方法，并和原有系统进行对比，结果如表 2 所示。

表 2 实验结果

	分词			分词+词性标注		
	P	R	F	P	R	F
混合 HMM	97.18%	97.85%	97.51%	93.24%	93.67%	93.45%
基本模板	97.20%	97.05%	97.12%	92.64%	92.47%	92.55%
基本模板+邻接属性	**98.23%**	**98.02%**	**98.12%**	**94.80%**	**94.41%**	**94.60%**

实验结果表明，在只使用基本模板的情况下，我们新的系统略差于原有的系统。可以看出，相对于原有的基于 HMM 的算法，由于缺少了邻接规则的判定，无论是分词还是分词加词性标注的 F 值都有所下降。而加上邻接属性后的算法比原系统和基本模板都有了较大的提高。一方面，相对于原系统，感知器算法由于采用了判别训练，在训练集较小的情况下比生成模型具有更好的效果；另一方面，相对于基本模板，由于加入了邻接属性的特征，相邻词语的搭配将更加符合语法规则。

下面给出一个正确分词和词性标注的例子：

原文：あなたは歩いて五分で行ける。(译文：你走着去五分钟可以到。)

分词词性标注结果：あなた P は XS 歩い V て XC 五 M 分 U で XN 行ける V 。 T

其中词性标注的缩写表示：あなた 代词 は 系助词 歩い 动词 て 接助词 五 数词 分 量词 で 格助词 行ける 动词 。 标点

5 结论和展望

本文提出了一种基于规则和统计的日语分词和词性标注方法，并且使用基于单一感知器的联合分词和词性标注算法进行训练和解码。

我们引入了邻接属性的特征，使算法的正确性得到了较大提高。邻接属性作为一种人工确定的标准，具有很高的效率，我们仅仅使用了一些简单的特征模板，就得到了较好的结果，而且并没有使训练和解码过程更加复杂。而感知器算法作为一种判别训练算法，对小训练集有较好的效果。分词和词性标注联合的算法也对减少误差传递有很好的效果。

当然，更多的特征可能会对结果有进一步的提高，例如更高阶的 n-gram 特征，或者更多的字符特征。我们只在未登录词中使用了字符特征，实际上，字符特征可能对已登录词的识别也有帮助。而多重集束搜索算法的提出也保证了加入更多特征后的解码过程仍然可解。

参考文献

[1] T. Kudo, K. Yamamoto, and Y. Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Proceedings of EMNLP, 2004.

[2] H. Ng and J. Low. Chinese part-of-speech tagging: one-at-a-time or all-at-once? Word-based or character-based? In Proceedings of EMNLP, 2004.

[3] F. Peng, F. Feng, and A. McCallum. Chinese segmentation and new word detection using conditional random fields. In Proceedings of COLING, 2004.

[4] N. Xue and L. Shen. Chinese word segmentation as LMR tagging. In Proceedings of ACL SIGHAN Workshop, 2003.

[5] Lawrence. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recogonition. In Proceedings of IEEE, 1989.

[6] Patnaparkhi and Adwait. A maximum entropy part-of-speech tagger. In Proceedings of the EMNLP, 1996.

[7] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy markov models for information extraction and segmentation. In Proceedings of ICML, 2000.

[8] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML, 2001.

[9] Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In Proceedings of EMNLP, 2002.

[10] T. Nakagawa. Chinese and Japanese word segmentation using word-level and character-level information. In Proceedings of COLING, 2004.

[11] M. Asahara. Corpus-based Japanese morphological analysis. PhD Thesis, NAIST, 2003.

[12] Y. Zhang and S. Clark. Chinese segmentation with a word-based perceptron algorithm. In Proceedings of ACL, 2007.

[13] Y. Zhang and S. Clark. Joint word segmentation and POS tagging using a single perceptron. In Proceedings of ACL, 2008.

[14] W. Jiang, L. Huang, Q. Liu, Y. Lu. A cascaded linear model for joint Chinese word segmentation and part-of-speech tagging. In Proceedings of ACL, 2008.

[15] K. Uchimoto, C. Nobata, A. Yamada, S. Sekine, H. Isahara. Morphological analysis of the spontaneous speech corpus. In Proceedings of COLING, 2002.