领域自适应的合成词词性标注研究*

张春荣，赵　琦
（普天信息技术有限公司，北京 100080）

摘　要： 在词性标注研究中，未登录的专业领域合成词给词性标注增加了很大的困难。提出了一种领域自适应的合成词词性标注方法，融合支持向量机（SVM）模型和基于转换学习（TBL）的方法来进行自动词性标注。对专业领域合成词的形态特征进行了详尽的分析，对有关的语法特点和语言现象进行了总结。有效利用这些合成词构词单元的语言学信息，把词类和词内结构信息引入 SVM 特征选择模板和 TBL 转换规则模板中，并采用核心属性渗透方法标注专业领域合成词的词性。实验结果表明，该方案能够有效地提高词性标注的准确率。

关键词： 词性标注；支持向量机；基于转换学习；合成词；领域自适应
中图分类号： TP391　　**文献标志码：** A　　**文章编号：** 1001-3695(2018)05-1350-05
doi:10.3969/j.issn.1001-3695.2018.05.015

Research on domain-adaptive POS tagging for compound words

Zhang Chunrong, Zhao Qi
(*Potevio Information Technology Co., Ltd., Beijing* 100080, *China*)

Abstract: In the part-of-speech(POS) tagging study, it is very difficult to POS tagging for unknown out-of-vocabulary(OOV) compound words in a specific domain. This paper proposed a domain-adaptive POS tagging method for compound words, which integrated support vector machine (SVM) model and transformation-based learning (TBL) method to process automatic part-of-speech tagging. Firstly, it analyzed the morphological features of the compound words in the specific domain in detail, and summarized the related grammatical features and linguistic phenomena. And then, with the linguistic information of these word-formation units, it added the information of lexical and word structure into the SVM feature selection template and TBL transformation rule template. Lastly, it used the head-feature percolation theory to determine the compound word's POS tagging in the specific domain. According to the experiments, the proposed method improves the accuracy of POS tagging.

Key words: part-of-speech(POS) tagging; support vector machine(SVM); transformation-based learning(TBL); compound word; domain-adaptive

0　引言

词性标注是中文信息处理领域内一项很重要的研究内容。在输入文本进行中文分词后，结合该词的上下文环境，词性标注对每个词标记上合适的词性。词性标注不仅是句法分析、语义分析、篇章理解等深层自然语言处理的基础，也是机器翻译、问答系统、信息检索和信息抽取等应用的关键环节[1]。汉语中存在词的兼类现象，而且存在大量专业领域未登录的合成词，这些都给词性标注增加了很大的困难。

国内外自然语言处理研究人员都很重视词性标注，自动词性标注研究也取得了很大的进展，成功设计出很多词性标注模型。目前，归纳起来比较典型的词性标注方法主要有基于规则的方法和基于统计的方法[2]。基于规则的方法是一种传统方法，充分利用现有语言学的成果，通过总结语言学规律而得出许多有用的规则，进而对大多数的语言组合作细致的描述。但是规则库的详尽程度影响词性标注的准确率，如果规则描述过细，词性标注的正确率可提高，但是规则的覆盖面就会大大减小；如果使覆盖面增大，必然要以降低正确率为代价。任何规则库都不可能尽善尽美，所以需要对规则进行人工添加和修改，故急需自动规则学习的方法。另外，当遇到词典未登录的合成词时，则会错误地将未登录词切分成多个语素或多个词，由于系统未能识别这些合成词，所以也无法对其进行词性标注，造成词性标注准确率相应降低。基于统计的方法是通过对大规模语料库进行训练而得到的，在词性标注方面得到了广泛的应用并取得了较好的效果。词性标注是一种序列标注问题，可以通过很多基于统计学习的模型来处理，包括最大熵模型（MEMM）[3,4]、隐马尔可夫模型（HMM）[5,6]、条件随机域模型（CRF）[7,8]、支持向量机模型（SVM）[9]等。许多模型可以进行特征选择和调整，进而可以取得较好的性能。在未登录的合成词词性标注方面，基于统计的词性标注方法可以根据词的上下文信息以及词的构词特点来确定其词性。基于统计的方法是目前应用较多的方法，由于其是通过对大规模语料库进行训练而得到的，所以覆盖面很广，但这种方法只是取大概率事件，并没有考虑小概率的特殊事件，不能很好地解决稀疏问题，必然会降低词性标注的正确率。

在词性标注研究中，未登录的合成词给词性标注增加了很多困难，特别是在专业领域，存在大量的领域合成词，因此对这种合成词的词性标注具有很重要的现实意义。鉴于以上两种方法在合成词词性标注方面的缺陷，本文提出了一种领域自适应的合成词词性标注方法。该方法融合了支持向量机词性标注模型和基于转换学习（TBL）的方法来进行自动词性标注。

收稿日期： 2016-12-30；**修回日期：** 2017-02-24　　**基金项目：** 国家重点研发计划资助项目(2016YFB0101100)

作者简介： 张春荣(1975-)，女，河北唐山人，工程师，硕士，主要研究方向为自然语言处理、人工智能(zhangchunrong@potevio.com)；赵琦(1989-)，男，工程师，硕士，主要研究方向为大数据.

合成词,从某种意义上说,可以看做是词组,是由一些基本的构词单元组成的。本文提出的方法能有效利用这些构词单元的语言学信息,把词类和词内结构信息引入规则模板和转换规则中,使所学的规则更有效,也更具代表性。

1　相关工作

1.1　支持向量机模型

支持向量机(support vector machine, SVM)是 Cortes 和 Vapnik 于 1995 年提出的,它在解决小样本、非线性及高维模式识别中表现出许多特有的优势,并能够推广应用到函数拟合等其他机器学习问题中。SVM 是一种很有效的分类方法,其主要思想是最大边缘和核函数原则。假设原始输入空间 $X \subseteq R_n$(其中 n 为输入空间的维数),定义训练集:

$$M = \{(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l)\} \tag{1}$$

其中:$x_i \in X$;$y_i \in \{-1, 1\}$ 是 x_i 的标记。若 x_i 属于正类,则 $y_i = 1$;若 x_i 属于负类,则 $y_i = -1$。l 为样本的个数。SVM 即寻找能够将训练数据划分为两类的最优超平面,该超平面可以通过求下面的凸二次规划方程的解得到。

$$\max \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i=j=1}^{l} \sum_{i=1}^{l} a_i y_i a_j y_j k(x_i, x_j)$$

$$\text{s.t.} \sum_{i=1}^{l} y_i a_i = 0 \quad 0 \leqslant a_i \leqslant c; i = 1, 2, \cdots, l \tag{2}$$

其中:$k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ 为 Kernel 函数,其满足 Mercer 条件,$\phi(x)$ 为原始输入空间到高维特征空间的非线性映射;a_i 为与每个样本对应的 Lagrange 乘子;$c > 0$,是自定义的惩罚系数。给定一个测试实例 x,它的类别由下面的决策函数决定:

$$f(x) = \text{sgn}\left[\sum_{x_i \in s_v} a_i y_i k(x_i, x) + b\right] \tag{3}$$

其中:s_v 为支持向量;b 是分类阈值,可用任一支持向量或通过两类中任一对支持向量取中值求得。

针对词性标注任务来讲,利用 SVM 模型进行训练和解码。SVMTool[10] 是建立在支持向量机原理上的序列标注工具,它使用待标注语言的特征。特征提取的好坏直接影响词性标注的准确率,特征收集得越好、越贴切,则准确率会越高。王丽杰等人[9] 将 SVMTool 应用于中文词性标注,在原特征基础上加入了中文词语的特征部首和重叠特征,有效地提高了未登录词的标注准确率,进一步提高了 SVMTool 的总标注准确率。

1.2　基于转换学习模型

1992 年,基于转换学习(transformation-based learning, TBL)的概念被提出来[11]。接着,Brill[12] 提出基于转换学习模型,它是一种从语料库中自动获取规则的机器学习算法。TBL 模型基于转换的标注有两个重要组成部分,即错误校正转换的详细说明和学习算法。一个转换(transformation)包含激活环境和改写规则两部分。激活环境指的是上下文信息,当上下文某些位置的标注满足一定条件时,应用改写规则。改写规则的形式是 $t^1 \rightarrow t^2$,表示用 t^2 替换 t^1。一个转换规则 T^1 的例子如表 1 所示。

表 1　转换规则的一个例子 T^1

改写规则	将一个词的词性从动词(v)改为名词(n)
激活环境	该词左边第一个紧邻词的词性是量词(q),第二个词的词性是数词(m)
$S0$	他/r 做/v 了/u 一/m 个/q 报告/v
运用 T^1	
$S1$	他/r 做/v 了/u 一/m 个/q 报告/n

基于转换的标注学习算法选择了最佳的转换,并且确定了它们的应用次序。其工作的主要步骤是:首先用最常见的标记标注每个词。在训练迭代的过程中,本文选择最可能减少错误率的转换,通过标注过的语料库中被错误标注的词语数目来衡量错误率。当没有能够降低超过预先制定阈值 ε 大小的错误率的转换时将停止。这是一个转换最优序列的贪婪搜索过程。

2　研究内容

2.1　领域合成词的形态特征

2.1.1　领域合成词

本文的词性标注方法是基于对轨道交通工程招中标领域知识的研究。在工程招中标这样的专业领域,含有大量的领域合成词,因此对这种合成词的词性标注具有很重要的意义。工程招中标领域的词汇类别涵盖工程施工词汇、工程项目类别、轨道标志词汇、地名/站名、公司机构名以及招投标词汇等。这些合成词的识别(关于合成词识别可以使用词典或者基于统计的方法[13,14]等,本文不作重点陈述)和词性标注,是信息检索和信息抽取等应用的关键环节。本文研究把这些工程招中标领域的知识很好地融入到特征选择和训练模型中。为便于本节对合成词词性标注工作的叙述,先给出一个示例文本,利用哈尔滨工业大学的语言技术平台 LTP 平台[15] 分词及词性标注结果如表 2 所示。表 3 示例文本识别出来的合成词有五个,分别是轨道交通、18 号线、沪南公路站、工程施工、中标公示。表 3 示例文本中的合成词见 0(注:本文使用的词性标记集是国家“863”词性标注集)。

表 2　合成词词性标注示例文本

待标注示例文本	分词及词性标注结果
“上海市轨道交通 18 号线工程沪南公路站 Φ800 污水管拆除、秀龙桥拔桩、清障及回填等工程施工项目中标公示”	上海市/ns 轨道/n 交通/n 18/m 号/q 线/n 工程/n 沪/j 南/j 公路/n 站/n Φ/v800/m 污水管/n 拆除/v 、/wp 秀龙桥/n 拔桩/v 、/wp 清障/n 及/c 回/v 填/v 等/u 工程/n 施工/v 项目/n 中标/v 公示/v

表 3　示例文本中的合成词

带标注的各原子词	合成词	词类特征
轨道/n 交通/n	轨道交通	工程词汇
18/m 号/q 线/n	18 号线	轨道标志词汇
沪/j 南/j 公路/n 站/n	沪南公路站	地名/站名
工程/n 施工/v	工程施工	工程项目类别
中标/v 公示/v	中标公示	招投标词汇

2.1.2　领域合成词形态特征分析

工程招中标领域的领域合成词、词汇类别涵盖工程词汇、轨道标志词汇、地名/站名、工程项目类别以及招投标词汇等。研究提取这些领域合成词词语的特征并将其付诸于实验中,观察其对未登录词标注准确率的影响,对可以提高未登录词标注准确率的特征进行模型训练及优化组合,进一步提高词标注准确率。本文主要分析前后缀以及合成词构词两个方面的特征。

1)前后缀特征　汉语中有些词缀相对固定,并与它们之前或之后的字组成词语。将合成词中存在构词能力比较强,且在词 AB 中处于 A 位置的字或词称为前缀。所述的构词能力比较强是指与其他字词构成未登录词的概率超过 90% 的前缀,将所述的前缀字归类为前缀字集,如“上”“前 * ”“最高 * ”“依次 * * ”“中和 * * ”等。类似且在词 AB 中处于 B 位置的字或词称为后缀。工程招中标领域常用的有表示线路、地名、站名等的尾词,如“ * * 线”“ * * * 路”“ * * 站”“ * * 街”等。表 4 是合成词语词前后缀示例。这些词缀对词性标注较为有益,如凡后缀有“路”“站”“线”等字的合成词,大多

可以标为名词。

表4 合成词语词前后缀示例

词缀	例子
前缀	“第*”“前*”“最高*”“最大*”等
后缀	“*线”“*站”“*高速”“*公路”“*市”“*标”等

2)合成词构词模式 合成词,从某种意义上可以看做是词组,是由一些基本的构词单元组成的。合成词的构词模式主要是研究构词单元的词性搭配规则。有效利用这些构词单元的语言学信息,把词类和词内结构信息引入规则模板和转换规则中,使所学的规则更有效,也更具代表性。通过对工程招投标领域合成词构词的研究,常见的构词模式有 v + v(服务招标)、n + v(交通运输)、a + n(智能交通)、v + n(招标公告)、n + n(轨道交通)、m + q + n(九号线)、n + nd + n(新华东路)等。表5是示例文本中的合成词构词模式。

表5 示例文本中的合成词构词模式

合成词	构词模式
轨道交通	n + n
18号线	m + q + n
沪南公路站	j + j + n + n
工程施工	n + v
中标公示	v + v

一些通用的构词模式如一种非常普遍的词语模式是叠字组合。这些叠字组合很多情况下都应被当做一个词。另外,偏旁部首是汉字表意的基础,在汉字形义关系中有着重要的作用。所以知道了一个汉字的偏旁部首,就可以初步猜测它的词性,然后结合它所在的词以及该词所在的上下文环境作进一步判断。

2.1.3 合成词的词性标注

对于合成词的词性标注,有的文献仅仅取最前面一个词的词性来确定合成词的词性,但有时合成词词性构成复杂,最前词难以反映整个合成词的词性。本文参考的方法理论是:合成词核心成分的语法属性渗透到整个合成词上,从而影响合成词的语法属性。关于核心成分的确认,杨梅[16]对北京大学中文系和北京大学计算语言研究所《现代汉语语法信息词典》(电子版)中除成语和惯用语之外的40 778个双音节、多音节词进行了分析和统计,对其中的39 454个词(排除单纯词、专名、音译词和结构关系不明的词等共1 324条)逐一进行构词方式、构成成分的素性和核心的确认等分析,统计结果表明90%以上合成词都符合核心属性同化规则。

合成词根据其结构方式主要有复合式、附加式和重叠式三种构词类型。复合式合成词按构词方式可分为定中式、状中式、主谓式、联合式、连谓式、重叠式、述宾式、连补式、附加式、指量式、复量式、名量式。表6给出了表2示例文本中合成词的构词类型、核心成分及词性标注结果。

表6 示例文本中的合成词核心成分及词性标注结果

合成词	构词类型	核心成分	标注词性
轨道交通	定中式	交通	n
18号线	附加式	线	n
沪南公路站	附加式	站	n
工程施工	主谓式	施工	v
中标公示	联合式	中标	v

2.2 基于SVM和TBL的词性标注方法

2.2.1 算法流程

本文提出的词性标注方法融合了SVM和TBL模型,其系统总体结构如图1所示。在使用SVM标注模型的基础上,采用TBL作为有效的词性标注增益手段。系统主要分为模型训练和标注应用两大部分。模型训练部分由SVM训练而得到SVM模型,以及由TBL学习而建立TBL规则库两个模块组成。词性标注应用部分则先由SVM标注器对语料进行词性标注,然后由TBL标注器根据学习所得到的规则列表对标注结果进行有效的增益加工,从而得到最终的标注输出。通过定义相应的SVM训练和TBL学习的接口来达到训练的核心算法和用户应用分离的目标。

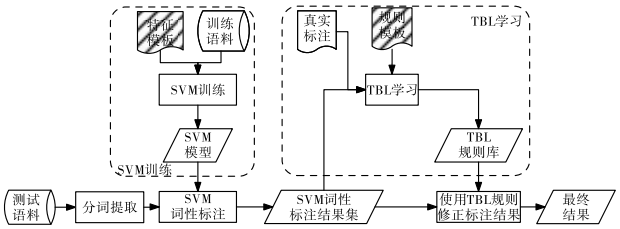

图1 基于SVM和TBL的词性标注方法

算法领域自适应的关键工作有以下两点:a)SVM特征模板的选取;b)TBL规则模板的定义。基于SVM的词性标注算法中,特征模板的好坏直接影响词性标注的准确率。而应用TBL方法自动获取规则列表时,TBL规则库中转换规则的选择空间受限于学习过程中TBL规则模板的定义。

2.2.2 SVM的特征模板

在利用SVMTool训练模型时,提取训练语料的信息,定义丰富的特征集,然后应用SVM原理来训练分类器。在训练模型时不仅生成了已登录词的标注模型,还会根据给定的比率选出部分词作为未登录词,应用这些词生成未登录词的词性标注模型,用于标注未登录词,由此可提高未登录词的标注准确率。

SVMTool应用SVM原理主要是利用待标注语言的特征,特征收集得越好、越贴切,则准确率会越高。词性标注方面,本文使用词而不是字作为基本的标注单元。特征模板如表7所示,包括当前词的模式和前后词缀特征则等。

表7 词性标注模型使用的特征模板

类别	特征
一元词 word-unigram	w[-2],w[-1],w[0],w[1],w[2]
二元词 word-bigram	w[-2]w[-1],w[-1]w[0],w[0]w[1],w[1]w[2]
三元词 word-trigram	w[-1]w[0]w[1]
最前/后字 last-first-character	ch[0,0]ch[0,n],ch[-1,n]ch[0,0],ch[0,-1]ch[1,0]
长度 length	length
前缀 prefix	ch[0,0],ch[0,0:1],ch[0,0:2]
后缀 suffix	ch[0,n-2:n],ch[0,n-1:n],ch[0,n]

在汉语中,汉字中同一偏旁部首的字,一般具有近义或同义关系。这种特征可以帮助分析一些复音词语的构成。同时,利用偏旁的性质,也可以帮助自动标注过程更准确地判定词性。依照偏旁部首的性质,可以在一定程度上判定是词性的大类。例如,“足”部的合成词,一般都是动词,而具有“木”偏旁部首的,都是名词,甚至可以分出凡“木”部的都是树名、木材或木制器具。另外一种非常普遍的词语模式是叠字组合,这些叠字组合很多情况下都应被当做一个词。此外两组叠字一般也是词,如果结合上述的偏旁部首来看,同偏旁的AABB叠字,则皆是词。而且这类词语通常具有形容词属性,因此这类词语模式特征可以对词性标注带来帮助。本文参考王丽杰等人[9]

的研究成果并加入待标注语言的特征,中文词性标注的新特征组合如表 8 所示。

表 8　中文词性标注的新特征组合

新特征组合	实现时表示
部首前/后缀化特征	bsa(1),bsa(2),bsa(3),bsz(1),bsz(2)
重叠特征	DOU

各特征含义为:a)bsa 表示部首的前缀组合,bsz 表示部首的后缀组合,bsa(i)/bsz(i)中的 i 表示距离词开始(bsa)或词结束(bsz)的偏移量;b)DOU 表示词重叠特征。

2.2.3　TBL 规则模板的定义

应用基于转换的错误驱动学习方法来自动获取规则库时,主要包括以下三点:a)对输入文本进行合成词的初始词性真实标注;b)转换规则模板的定义;c)TBL 学习获得规则。其中,规则模板定义了要寻找的候选规则搜索空间,将会影响最终学习得到的规则库的质量。考虑合成词的内部结构是由基本构词单元按照特定的语法规则构成的,本文称这些结构信息为词内结构。例如,合成词“18 号线”的词内结构就是由基本构成单元数字、量词和名词后缀所组成的。

根据应用的需求,可制定合成词词性标注的 TBL 规则模板,由系统负责生成模板文件以便后续的规则库建立模块调用。在本文实验中定义了两类模板,TBL 的规则模板如表 9 所示。

表 9　TBL 的规则模板

模板类型	规则模板	示例
基于词或词性的组合	如果当前词及上下文范围内的若干词为特定词语词性,则修改词性标记	pos_ -2 pos_ -1 word_0 = >pos
基于词性或词性的组合	如果当前词性及前后的若干词性为特定词类,则修改词性标记	pos_ -1 pos_0 = >pos

表 9 中的规则模板 pos_ -2 pos_ -1 word_0 = >pos 定义了一套基于前第一个、第二个词性和当前词而改变词性标记的规则,而 pos_ -1 pos_0 = >pos 则是基于当前词性以及前一个词性标记而改变词性标记。本文将生成的规则用于合成词中的基本构词单元,然后利用合成词核心成分的语法属性渗透到整个合成词上,从而标记合成词的词性。

3　实验结果与分析

3.1　性能评测指标

在对中文词性标注性能进行评估时,本文采用常用的评测指标——标注准确率(P)。标注准确率表示在对全部词语标注的词性中,正确标注词性的词语所占的百分比。计算方法如下:

$$\text{标注准确率}(P)=\frac{\text{正确标注词性的词语数}}{\text{所有待标注词性的词语数}}\times 100\% \qquad (4)$$

3.2　实验设计及结果分析

为了验证本文方法的有效性,评测其在领域适应性上的合成词词性标注性能。此次实验研究算法在招中标领域的合成词词性标注。实验对抓取的 2016 年 1 月 ~10 月轨道交通领域的工程招标、中标公示文件 6 215 篇进行分析。使用 LTP 平台进行分词、初始标注,其中合成词在所有词中所占的比例约为 15%。为了制作测试集,随机选出 200 篇,利用 2.1 节制定的领域合成词词性标注标准进行少量人工分词标注,再利用本文方法对其他的进行自动标注。本文设计了两组实验,分别从不同的角度对合成词的词性标注性能进行研究。

实验 1　本文领域自适应的词性标注与无领域自适应 SVM 算法词性标注的比较,验证本文领域自适应的词性标注算法对改进合成词词性标注系统性能的重要性。

实验 2　本文算法的性能与训练语料规模的关系,TBL 算法的性能与训练语料的规模有关。手工标注语料的花费很大,但是训练语料越多,学到的规则就越多,统计意义上也越可信,所以这个分析有助于确定适当的语料规模。

3.2.1　与 SVM 算法的比较

这里选取的 SVM 算法是 LTP 平台的词性标注算法。LTP 平台基础语料库是人民日报 1998 年 2 ~6 月(后 10% 数据作为开发集)作为训练数据,1 月作为测试数据。LTP 词性标注准确率达到 98.34%。实验将 SVM 和本文领域自适应的词性标注算法应用于轨道交通的工程招标中标领域,对词性标注准确率进行了比较,包括已知词和合成词,结果如表 10 所示。

表 10　本文算法与 SVM 算法的词性标注准确率比较

方法	已知词准确率/%	合成词标注准确率/%
SVM	98.05	86.43
本文	98.15	90.33

由表 10 中的结果可以看出,当应用于招中标领域时,对于已知词的标注准确率都比较高,SVM 和本文方法相差不大,都超过 98% 的准确率;但是对于合成词的标注 SVM 只有 86.43%,低于基础语料中 LTP 的准确率,这是由于合成词大多是领域专有的未登录词,性能标注降低。而本文方法有所提升,达到 90.33% 的准确率。通过这个评测结果可以看出,本文领域自适应的词性标注结果高于无领域自适应的 SVM 算法,说明领域自适应对于改进词性标注系统性能很重要。

3.2.2　训练语料规模对词性标注结果的影响

本文应用 TBL 算法进行规则学习时,需要对训练语料进行人工真实标注。由于 TBL 算法的性能与训练语料的规模有关,训练语料越多,学到的规则就越多,统计意义上也越可信,当然,手工标注语料的花费也越大,所以需要确定适当的语料规模。本文分别随机选取 50、100 和 200 句进行标注,然后对合成词的标注准确率进行评测,其结果如表 11 所示。

表 11　标注规模对词性标注准确率的影响

标注语料规模/句数	合成词占比例/%	合成词标注准确率/%
50	15.70	89.59
100	15.94	89.77
200	14.99	90.33

通过对比在不同规模标注语料上的评测结果可以看出:a)表 11 中所有结果都高于无领域自适应的算法,说明专业领域人工标注语料对词性标注的领域自适应有重要帮助,即使少量的 50 句语料就有明显效果;b)随着标注语料的增大,词性标注的准确率一直有所提升,200 句的标注语料还未达到饱和状态,由此预测,随着经过筛选的标注语料的加入,系统的性能还有提升空间。

4　结束语

词性标注是自然语言处理重要和基础的研究课题之一。迄今为止,词性标注任务已使用了多种技术方法,并取得了不错的成绩。基于规则的方法能准确地描述词性搭配之间的确定现象,但是其编写和维护工作则显得过于繁重,而基于统计的方法适用范围很大程度上又受到训练语料的制约。本文围绕领域自适应的合成词词性标注方法进行探索,针对大规模人

工训练语料难以获取的问题,提出了SVM与TBL相结合进行词性标注的方法,并应用于轨道交通工程招中标领域自适应的任务中,验证本文研究成果的有效性。通过分析招中标领域词汇形态特征,对有关的语法特点和语言现象进行总结,设计SVM的词性标注特征模板和TBL规则模板,并采用核心属性渗透方法标注合成词的词性。在该语料库的构建方面,基于通用基础训练语料,并辅助以小规模的人工标注的领域数据语料,使用领域自适应技术提升合成词的自动标注的准确率。在未来的研究工作中,将针对涉及影响标注结果的特征、规则等因素进行更深入的研究,以期能找到一些更有实用价值的特征信息以及分辨能力更强的转换规则,使本文方法的整体标注性得到进一步的提升,并对领域自适应的方法进行更深入的探索。

参考文献:

[1] 姜维,王晓龙,关毅,等.基于多知识源的中文词法分析系统[J].计算机学报,2007,30(1):137-145.

[2] 梁喜涛,顾磊.中文分词与词性标注研究[J].计算机技术与发展,2015,25(2):175-180.

[3] 赵岩,王晓龙,刘秉权,等.融合聚类触发对特征的最大熵词性标注模型[J].计算机研究与发展,2006,43(2):268-274.

[4] 余昕聪,李红莲,吕学强.最大熵和HMM在中文词性标注中的应用[J].无线互联科技,2014(11):122-124.

[5] 袁里驰.基于改进的隐马尔可夫模型的词性标注方法[J].中南大学学报:自然科学版,2012,43(8):3053-3057.

[6] 姜芳,李国和,岳翔,等.基于粗分和词性标注的中文分词方法[J].计算机工程与应用,2015,51(6):204-207,265.

[7] 洪铭材,张阔,唐杰,等.基于条件随机场(CRFs)的中文词性标注方法[J].计算机科学,2006,33(10):148-151.

[8] 王艺帆,王希杰.基于双层条件随机场的汉语词性标注方法研究[J].安阳师范学院学报,2016(5):87-91.

[9] 王丽杰,车万翔,刘挺.基于SVMTool的中文词性标注[J].中文信息学报,2009,23(4):16-21.

[10] Giménez J, Màrquez L. SVMTool: a general POS tagger generator based on support vector machines[C]//Proc of the 4th International Conference on Language Resources and Evaluation. 2004.

[11] Florian R, Ngai G. Fast transformation-based learning toolkit[EB/OL]. [2008-09-10]. http://nlp.cs.jhu.edu/~rflorian/fnllbl/documentation.html.

[12] Brill E. Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging[J]. Computational Linguistics, 1995, 21(4): 543-565.

[13] 李文坤,张仰森,陈若愚.基于词内部结合度和边界自由度的新词发现[J].计算机应用研究,2015,32(8):2302-2304,2342.

[14] Song Yan, Xia Fei. Using a goodness measurement for domain adaptation: a case study on Chinese word segmentation[C]//Proc of the 8th International Conference on Language Resources and Evaluation. 2012: 3853-3860.

[15] Che Wanxiang, Li Zhenghua, Liu Ting. LTP: a Chinese language technology platform[C]//Proc of the 23rd International Conference on Computational Linguistics: Demonstrations. Stroudsburg, PA: Association for Computational Linguistics, 2010: 13-16.

[16] 杨梅.现代汉语合成词构词研究[D].南京:南京师范大学,2006.

(上接第1341页)

[12] Palla G, Derényi I, Farkas I, *et al*. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043): 814-818.

[13] Evans T S. Clique graphs and overlapping communities[J]. Journal of Statistical Mechanics Theory & Experiment, 2010, 2010(12): 257-265.

[14] Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure of complex networks[J]. New Journal of Physics, 2009, 11(3): 19-44.

[15] Vahdat A, Becker D. Epidemic routing for partially-connected Ad hoc networks[D]. Durham: Duke University, 2000.

[16] Spyropoulos T, Psounis K, Raghavendra C S. Spray and wait: an efficient routing scheme for intermittently connected mobile networks[C]//Proc of ACM SIGCOMM Workshop on Delay-Tolerant Networking. New York: ACM Press, 2005: 252-259.

[17] Lindgren A, Doria A, Schelén O. Probabilistic routing in intermittently connected networks[J]. ACM SIGMOBILE Mobile Computing & Communications Review, 2004, 7(3): 239-254.

[18] Grossglauser M, Tse D N C. Mobility increases the capacity of Ad hoc wireless networks[J]. IEEE/ACM Trans on Networking, 2002, 10(4): 477-486.

[19] Bettstetter C, Hartenstein H, Pérezcosta X. Stochastic properties of the random waypoint mobility model[J]. Wireless Networks, 2004, 10(5): 555-567.

[20] Hui Pan, Crowcroft J, Yoneki E. Bubble rap: social-based forwarding in delay tolerant networks[J]. IEEE Trans on Mobile Computing, 2011, 10(11): 1576-1589.

[21] Chen Kang, Shen Haiying. SMART: lightweight distributed social map based routing in delay tolerant networks[C]//Proc of the 20th IEEE International Conference on Network Protocols. Piscataway, NJ: IEEE Press, 2012: 1-10.

[22] Eagle N, Pentland A, Lazer D. Inferring social network structure using mobile phone data[J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(36): 15274-15278.

[23] Srinivasa S, Krishnamurthy S. CREST: an opportunistic forwarding protocol based on conditional residual time[C]//Proc of IEEE Communications Society Conference on Sensor, Mesh and Ad hoc Communications and Networks. Piscataway, NJ: IEEE Press, 2009: 342-350.

[24] Tournoux P U, Leguay J, Benbadis F, *et al*. The accordion phenomenon: analysis, characterization, and impact on DTN routing[C]//Proc of IEEE INFOCOM. Piscataway, NJ: IEEE Press, 2009: 1116-1124.

[25] Liu Cong, Wu Jie. Routing in a cyclic mobispace[C]//Proc of the 9th ACM Interational Symposium on Mobile Ad hoc Networking and Computing. New York: ACM Press, 2008: 351-360.

[26] Keränen A, Ott J, Kärkkäinen T. The ONE simulator for DTN protocol evaluation[C]//Proc of the 2nd International Conference on Simulation Tools and Techniques. Brussels, Belgium: ICST, 2009: Article No. 55.

[27] CRAWDAD data set[EB/OL]. http://crawdad.cs.dartmouth.edu.

[28] Hui Pan, Crowcroft J. Predictability of human mobility and its impact on forwarding[C]//Proc of the 3rd International Conference on Communications and Networking. Piscataway, NJ: IEEE Press, 2008: 543-547.

[29] Ekman F, Keränen A, Karvo J, *et al*. Working day movement model[C]//Proc of the 1st ACM SIGMOBILE Workshop on Mobility Models. New York: ACM Press, 2008: 33-40.