文章编号：1003-0077（2011）00-0000-00

基于词向量的藏文词性标注方法研究*

郑亚楠[1]，珠杰[1,2]†

（1.西藏大学计算机科学与技术系，西藏 拉萨 850000；）

（2.西南交通大学信息科学与技术学院，四川 成都 610031）

摘要：藏文词性标注是藏文信息处理的基础，在藏文文本分类、自动检索、机器翻译等领域有广泛的应用。本文针对藏文语料匮乏，人工标注费时费力等问题，提出一种基于词向量模型的词性标注方法和相应算法，该方法首先利用词向量的语义近似计算功能，扩展标注词典，其次结合语义近似计算和标注词典，完成词性标注。实验结果表明，该方法能够快速有效地扩大标注词典规模，并能取得较好的标注结果。

关键词：词向量；藏文；词性标注

中图分类号：TP391　　**文献标识码：**A

Research the Method of Tibetan POS Based on Distributed Representation

Yanan Zheng[1], Jie Zhu[1,2]†

(1.Department of Computer Science, Tibetan University, Tibetan, Lhasa 850000, China;)

(2.School of Information Science and Technology, Southwest Jiaotong University, Sichuan, Chengdu 610031,China.)

Abstract: Tibetan Part of Speech (POS) is the foundation of Tibetan natural language processing and is widely applied in the Tibetan text classification, information retrieval, machine translation and other fields. Lack of Tibetan corpus and difficulty of manual tagging, this paper proposed a method and an algorithm of Tibetan POS based on distributed representation. First, this method can extend the dictionary by semantic approximate calculation of distributed representation, and then complete POS combined with semantic similar calculation and dictionary .Experimental results show that, this method can expand the dictionary quickly and achieve better results.

Key words: Distributed Representation; Tibetan; POS;

1 引言

藏文信息处理起步于上世纪 80 年代，经过三十多年的发展，已取得一些令人瞩目的成绩。但由于缺乏统一标准，词处理技术尚不够成熟，加上藏文语料严重匮乏，其研究一直进展缓慢。藏文词性标注作为藏文信息处理中一项重要的基础性工作，其标注效果直接制约着藏文信息处理技术的发展，并对藏文词法分析、句法分析和语义分析等研究领域有很大影响。虽然藏文信息处理研究在技术上充分利用已有的国内外先进的处理方法，但其基础语料资源相对贫乏，各研究单位公开的语料较少且多为未标注语料，应用价值非常有限。因此，针对

*收稿日期：2016 年 6 月 1 日。定稿日期：2016 年 8 月 5 日。
基金项目：国家自然基金项目(61262058)；国家社会科学基金（15ZDB11）；西藏高校青年教师创新支持计划项目（QC2005_18）

藏文词性人工标注费时又费力的问题，本文提出了一种基于词向量模型的词性标注方法。

深度学习模型训练的词向量具有良好的语义特征，是表示词语特征的常用方式，一般用 Distributed Representation 表示。词向量是一个稠密、低维的实数向量，它的每一维表示词语的一个潜在特征，该特征捕获了有用的句法和语义特征。本文充分利用词语之间的语义相似关系扩充原始标注词典，并结合扩充后的标注词典与词向量近似计算对测试语料进行词性标注。

2 相关工作

词性标注是计算机自动语言分析和理解的一个重要环节，其任务是为文本中的每一个词都标记上一个恰当的语境词类标记符号，即确定每个词的名词、动词、形容词或其他词类属性[1]。汉语、英语等语言的词性标注研究较为成熟，都有开源的标注系统。藏文词性标注起步相对较晚，研究基础相对薄弱，采用的标注方法大多借鉴汉语、英语等国内外较为成熟的方法。

2004 年，江荻[2]最先讨论了藏文词性标注问题。2006 年，才让加[3]等根据藏文词类的功能和性质提出了一种藏文的词性分类及代码。扎西加[4]等以藏文语法理论和汉语、英语词性划分为依据，将藏文词语划分为 26 个基本类和 9 个特殊类。苏俊峰[5]等使用人工标注的语料统计词和词性，并通过训练二元语法的 HMM 模型参数，运用 Viterbi 算法完成了基于统计方法的藏文词性标注。扎西多杰[6]等以 4 万词的语料库作为训练语料，同样采用 HMM 模型对 20 篇文章进行词性标注，其标注正确率达到 84%。华却才让[7]等在分析现有藏文词性标注方法的基础上，提出了感知机训练模型的判别式藏文词性标注方法，并在 573 句人工标注的语料上进行了相关实验，取得了较好的效果。于洪志[8]等研究了融合语言特征的最大熵藏文词性标注模型，并通过实验证明音节特征可以显著提高藏文词性标注的效果。康才畯[9]采用最大熵结合条件随机场模型实现了藏文词性标注，并在小规模语料训练下达到了 87.76%的准确率。综上所述，可以看出在已有的藏文词性标注研究中，均是采用统计模型的方法进行词性标注。由于统计方法需要大规模的语料来提高精度，而藏文公开的语料较少，各研究人员的实验条件和实验语料不统一，使得实验结果相差较大，还达不到可实际应用的程度。

3 词性标注算法

3.1 标注集的确定

由于目前还没有一个统一的藏文词类划分标准，因此，各研究单位和人员所用词类划分的粒度和标记符号并不相同。本文参照前人对标注集的研究，将藏文词语分为一、二、三级类别，其中包括 3 个一级类别，16 个二级类别，70 个三级类别。然后根据藏文文法特点，在该标注集的基础上按照划分粒度不同分别定义了粗切分标注集和细切分标注集，涉及到的相关概念定义如下：

定义 1：将最初统计的各标注类别所包含的词语称为种子。

定义 2：将藏文词语中数量及词性无太大变化的虚词组成的集合称为固定标注集，标注规范及各类词数统计结果如表 3-1 所示，称之为固定标注库。

定义 3：将藏文词类标注集中二级类别和三级类别相结合的标注规范称为粗切分，如表 3-2 所示，称之为粗切分标注库。该表包含了标注规范和种子数量，在标注库扩充算法中将其作为粗切分的种子库。

定义 4：将藏文词类标注集中三级类别的标注规范称为细切分，如表 3-3 所示，称之为细切分标注库。该表包含了标注规范和种子数量，在标注库扩充算法中将其作为细切分的种子库。

表 3-1 固定标注库

类别	标记	标注词数	类别	标记	标注词数	类别	标记	标注词数
人称代词	RP	15	时态助词	UT	8	集饰连词	CS	3
叙述代词	RD	5	语气助词	UN	18	待述连词	CX	3
不定代词	RN	9	祈使助词	UP	6	和聂连词	CH	3
指示代词	RI	15	比较助词	UI	3	提聂连词	CN	6
疑问代词	RQ	18	原因助词	UC	4	总聂连词	CZ	3
程度副词	DX	6	目的助词	UE	4	离合连词	CL	1
否定副词	DN	4	终结助词	UF	11	陈述连词	CC	1
范围副词	DR	9	作格助词	PZ	5	并列连词	CB	8
时频副词	DH	19	属格助词	PS	5	递进连词	CD	7
情态副词	DM	12	位格助词	PW	7	转折连词	CU	2
叹词	E	9	从格助词	PC	2	条件连词	CT	2
藏文符号	TF	31	藏文数字（非常用）	TD	20			

表 3-2 粗切分标注库

类别	标记	种子数	类别	标记	种子数	类别	标记	种子数
普通名词	NN	8	数词	M	26	动词	V	42
人名	NR	5	长量词	QL	8	形容词	A	53
地名	ND	6	状量词	QS	5	状态词	S	57
机构名	NJ	21	重量词	QH	8	区别词	B	11
处所词	NL	5	集量词	QG	7	拟声词	O	12
方位词	NF	18	器量词	QC	6	藏文数字（常用）	TD	10
时间词	NT	23	动量词	QM	3	非藏文字符	TN	
其他专有名词	NZ	12						

表 3-3 细切分标注库

类别	标记	种子数	类别	标记	种子数	类别	标记	种子数
普通名词	NN	8	长量词	QL	8	状态形容词	AS	6
人名	NR	5	状量词	QS	5	颜色形容词	AC	9
地名	ND	6	重量词	QH	8	限定形容词	AX	8
机构名	NJ	21	集量词	QG	7	面积形容词	AR	5

处所词	NL	5	器量词	QC	6	性质状态词	SQ	29
方位词	NF	18	动量词	QM	3	体态状态词	SP	12
时间词	NT	23	自动词	VT	7	声态状态词	SO	8
其他专有名词	NZ	12	他动词	VI	6	动态状态词	SD	8
基数词	MC	12	助动词	VU	17	区别词	B	11
序数词	MO	5	存在动词	VE	9	拟声词	O	12
倍数词	MI	4	判断词	VS	3	藏文数词（常用）	TD	
分数词	MF	5	性质形容词	AQ	25	非藏文字符	TN	

3.2 标注库扩充算法

Mikolov[10]通过三个词向量的计算，例如 $X = vector("king") - vector(man) + vector(woman)$ 可以预测出 *queen* 的结果。本文提出的标注库扩充算法利用词向量的语义近似计算功能对种子库中的词语进行近似计算，进而得到扩充后的标注库。算法的具体过程如图 3-1 所示。

在标注库扩充算法中，初始状态下，含有已标记词性的词库称为种子库。在种子库中，每个词性只将少数的典型词作为种子，称之为目标词。算法执行过程中，遍历种子库中所有的目标词，并通过词向量对每一个目标词进行语义相似计算。按照相似度计算值的大小降序排列，取出前 n 个相似词，作为扩充词的候选词。遍历所有候选词，若该候选词已存在于种子库和固定标注库中，则不添加到种子库，否则就将该候选词添加到种子库中，并以目标词的词性来标注该候选词。通过反复迭代，使得种子库中所含已标记词性的词数量不断地增加，直至迭代结束，可得到扩充后的标注库。

设虚词和词性集合为：$X = \{w_i^1 : p_i^1\}$，其中 w_i^1 表示一个藏文虚词，p_i^1 表示藏文虚词 w_i^1 对应的词性，该集合即固定标注库；设标注库集合：$Y = \{w_i^2 : p_i^2\}$，其中 w_i^2 表示除藏文虚词 w_i^1 之外的其他词语，p_i^2 表示词 w_i^2 对应的词性，该集合即种子库；设词向量集合：$V = \{v_1, v_2, ...v_k\}^T$，其中 v_i 表示词 w_i 对应的词向量，k 为词的个数。

算法 1：标注库扩充算法

输入：X，Y，V

输出：扩充之后的 Y

```
(1) for(j=1;j<=m;j++)
(2)        read  w_j  in  Y
(3)              A=similar( w_j , V );
(4)              取 A 集合中的前 n 个词语存入 B 集合；
(5)              if(each element in B not belong to Y )
(6)                  then 用 w_j 的词性标记这该词语，并添加至 Y ；
(7) end for
```

图 3-1 标注库扩充算法

3.3 词性标注算法

词性标注算法是通过固定标注库、标注库和语义近似计算相结合的一种标注方法。该算

法中，首先输入已分好词的句子，然后遍历句子中所有词，判断该词是否存在于固定标注库和标注库中。若存在，则直接标记该词语；否则，先将其作为目标词进行语义相似计算，再确定其词性。根据语义近似计算的结果降序排列，取出前 n 个词，从计算值最高的词开始，逐个与标注库进行比对，一旦找到一个词与标注库中的词相匹配，就用该词的词性来标注目标词。最后，既不在固定标注库和标注库中，也不能通过词向量的语义近似计算来标注词性的词，就用 NULL 来标记。

设句子集合：$S=\{s_1,s_2,....,s_n\}$，其中 $s_i=\{w_1/w_2/...../w_m\}$ 是词序列组成的一个句子，n 为句子的个数。X 表示固定标注库，Y 表示标注库，V 表示词向量，与上节表示方式相同。具体标注算法如图 3-2 所示。

算法 2：词性标注算法

输入：X，Y，V，S
输出：词性标记句子
(1) for(i=1;i<=n;i++) //读取每个句子
(2) for(j=1;j<=m;j++) //检查句子中每个词的词性，并完成标记
(3) read w_j in s_i；
(4) if($w_j \in X$)
(5) w_j 对应的词性作标记;
(6) else
(7) if($w_j \in Y$)
(8) w_j 对应的词性作标记;
(9) else
(10) *A=similar(* w_j *,*V *)*;
(11) 取 *A* 集合中的前 n 个词语存入 *B* 集合;
(12) if(a element in *B* belong to Y)
(13) Y 中该词的词性标注 w_j 的词性;
(14) else 用 NULL 标注 w_j 的词性;
(15) end for
(16) end for

图 3-2 词性标注算法

4 实验及数据分析

4.1 实验语料

实验中使用的词向量，是以 2009 年、2010 年和 2014 年西藏日报的文本内容作为语料，经过断句、分词和特殊标点符号的处理之后，利用 word2vec 训练得到。按照 word2vec 工具提供的 skip-gram 模型，在窗口大小 5、迭代次数 100、学习参数 0.025 的条件下，在 50 维度下完成训练。

本文采用的测试语料是分词后由人工标注的 500 条句子，并按两种方案完成实验。第一种方案中采用粗切分种子库和固定标注库相结合进行词性标注；第二种方案中采用细切分种子库和固定标注库相结合进行词性标注。其中固定标注库共包含 35 个词性，粗切分种子库

共包含 22 个词性，细切分种子库共包含 36 个词性。

4.2 不同实验方案下的结果对比

本次实验采用了三个评测指标，分别为召回率、精确度和 F1 值。

4.2.1 实验 1

该实验是粗切分种子库和固定标注库相结合的一种词性标注方法。

（1）固定语义近似词数 n=2，通过调整迭代次数来完成词性标注。算法 1 的实验参数如表 4-1 所示。

表 4-1 方案 1 实验参数设置

训练语料	2009 年、2010 年和 2014 年三年西藏日报的文本内容
测试语料	人工标注的 500 句藏文句子
X（虚词库）	固定标注集，共 35 个词性，507 个词
Y（标注库）	粗切分标注库，共 22 个词性，694 个词
V（词向量）	word2vec 输出的二进制文件

实验中迭代次数 t 分别设为 5，10，15，20；算法 2 的词性标注结果如表 4-2 所示。

表 4-2 不同迭代次数下词性标注结果（粗分集+固定标注集）

迭代次数	精确度（%）	召回率（%）	F1 值（%）
5	58	46	51.31
10	60	48	53.11
15	59	47	52.32
20	58	47	51.92

从实验结果可以看出，随着迭代次数的增加，词性标注的精确度和召回率均呈现出先增加后减小的趋势，在迭代次数在 10 的情况下，F1 值得到了最好的标注结果。

（2）固定迭代次数 $t=10$，通过调整语义近似词数 n 来完成词性标注，实验中 n 分别设为 1，2，3，4；算法 2 的词性标注结果如表 4-3 所示。

表 4-3 不同近似词组数下词性标注结果（粗分集+固定标注集）

词数	精确度（%）	召回率（%）	F1 值（%）
1	67	44	53.12
2	60	48	53.11
3	54	48	50.38
4	56	50	52.83

从实验结果可以看出，随着近似词组数的增加，词性标注效果精确度逐渐下降，召回率逐渐上升，在词数为 1 的时候，F1 值取得了最好的效果。

4.2.2 实验 2

该实验是细切分和固定标注集结合的一种词性标注方法。

（1）固定语义近似词数 n=2 的，通过调整迭代次数来完成词性标注。算法 1 的实验参数如表 4-4 所示。

表 4-4 方案 2 实验参数设置

训练语料	2009 年、2010 年和 2014 年三年西藏日报的文本内容
测试语料	人工标注的 500 句藏文句子
X（虚词库）	固定标注集，共 35 个词性，507 个词
Y（标注库）	细切分标注库，共 36 个词性，684 个词
V（词向量）	word2vec 输出的二进制文件

实验中迭代次数 t 分别设为 5，10，15，20；算法 2 的词性标注结果如表 4-5 所示。

表 4-5 不同迭代次数下词性标注结果（细分集+固定标注集）

迭代次数	精确度（%）	召回率（%）	F1 值（%）
5	55	42	47.63
10	53	43	47.40
15	42	34	37.58
20	53	43	47.48

从实验结果可以看出，随着迭代次数的增加，词性标注的精确度和召回率逐渐下降，，且低于粗分集+固定标注集的结果。这是符合客观规律的，标注集越细，区分难度越大。

（2）固定迭代次数 $t=10$，通过调整语义近似词数 n 来完成词性标注，实验中 n 分别设为 1，2，3，4；算法 2 的词性标注结果如表 4-6 所示。

表 4-6 不同近似词组数下词性标注结果（细分集+固定标注集）

词数	精确度（%）	召回率（%）	F1 值（%）
1	68	40	50.37
2	53	43	47.48
3	49	44	46.37
4	52	47	49.37

从实验结果可以看出，随着 n 的增加，词性标注效果精确度依然呈现出逐渐下降的趋势，但召回率有所上升，整体 F1 值均低于第一种实验方案。

由以上实验可知，精确度最高可达 68%，召回率最高值为 50%。实验整体上随着近似词数逐渐增大，迭代次数逐渐增加，呈现出精确度逐渐下降，召回率逐渐上升的趋势。该实验结果证明了本文提出的方法对标注词典扩展和词性标注是行之有效的。

5 结论与展望

在充分研究现有藏文词性标注方法的基础上，本文提出了一种基于词向量的藏文词性标注方法。该方法首先利用词向量的语义相似计算完成种子库的扩充，然后结合已扩充的标注库和语义相似计算对测试数据进行词性标注。同时，分别以粗分集+固定标注集和细分集+固定标注集进行实验，并将其结果进行了对比分析。

与现有的藏文词性标注方法相比较，该方法不依赖大规模的词典，摆脱了人工标注词典耗时耗力的局限性，较好地解决了未登陆词的词性标注，为研究藏文词性标注提供了一种新视角。但分析其标注结果，该方法还有很大的提升空间，离实际应用还有一定的距离。本文认为造成实验结果偏低的原因主要有以下几点：（1）训练出来的藏文词向量不是最好的，因此直接影响语义近似计算结果；（2）测试数据可能包含一些错误标注；（3）种子库扩充时未考虑兼类词的情况；（4）词向量中未包含的词语，无法获得其向量表示，故不能进行近似计算。针对以上问题如何进行改进是我们今后研究的主要方向。

参考文献

[1]洛桑嘎登,赵小兵.藏文词级处理研究现状及热点方法[J].电脑知识与技术,2015.11:183-185.

[2]Jiang D. Text-annotation Oriented Tibetan-Chinese Dictionary and Its Construction[C].The 4th China-Japan Joint Conference to Promote Cooperation in Natural Language Processing.(CJNLP-04), HongKong,2004,10-15.

[3]才让加,吉太加.藏语语料库中词性分类代码的确定[C].中文信息处理前沿进展-中国中文信息学会二十五周年学术会议论文集.北京:清华大学出版社,2006.

[4]扎西加,珠杰.面向信息处理的藏文分词规范研究[J].中文信息学报,2009.24(3):113-123.

[5]苏俊峰,祁坤钰,本太.基于 HMM 的藏语语料库词性自动标注研究[J].西北民族大学学报:自然科学版.2009,30(1):42-45.

[6]扎西多杰,安见才让.基于 HMM 藏文词性标注的研究与实现[J].计算机光盘软件与应用.2012,12:100-101.

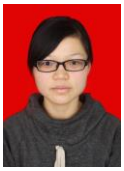
[7]华却才让,刘群,赵海兴.判别式藏语文本词性标注研究[J].中文信息学报.2014,28(2):56-60.

[8]于洪志,李亚超,江昆等.融合音节特征的最大熵藏文词性标注研究[J].中文信息学报:2013,27(5):160-165.

[9]康才畯.藏语分词与词性标注研究[D].上海师范大学博士论文,2014.

[10]T. Mikolov, W.-T. Yih, G. Zweig. Linguistic regularities in continuous space word representations. In NAACL-HLT, 2013: 746-751

作者简介：



郑亚楠（1992-），女（汉族），硕士研究生，主要研究方向为藏文信息处理、数据挖掘，E-mail:zs_zyn@yeah.net；

珠杰(1973-)，男（藏族），副教授，硕士生导师，主要研究方向为藏文信息处理、数据挖掘，†E-mail: rocky_tibet@qq.com。