基于藏语字性标注的词性预测研究 *

龙从军 1,2，刘汇丹 1，诺明花 1，吴健 1
（1.中国科学院软件研究所，北京，100190
2.中国社会科学院民族学与人类学研究所，北京，100081）

摘要：本文选取了藏语文中小学教材的部分语料，构建了带有藏语字性标记、词边界标记和词性标记的语料库，通过比较不同的分词、标注方法，证明分词、词性标注一体化效果比分步进行的效果好，准确率、召回率和 F 值分别提高了 0.067、0.073 和 0.07。但词级标注模型难以解决词边界划分的一致性和未登录词的问题。基于此，作者提出可以利用字性和字构词的规律预测合成词的词性，既可以融入语言学知识又可以减少由未登录词导致的标注错误，实验结果证明，作为词性标注的后处理模块，基于字性标注的词性预测准确率提高到了 0.916，这个结果已经比分词标注一体化结果好，说明字性标注对纠正词性错误标注有明显的效果。

关键词：藏语 语字标注 分词 词性标注
中图分类号：TP391 **文献标识码**：A

Research on POS prediction of Tibetan Based on Tagging of Syllable

Congjun Long^{12} Huidan Liu^{1} Nuo Minghua^{1} Jian Wu^{1}
(1 Institute of software Chinese academy of Sciences, Beijing, 100190
2 Institute of ethnology and Anthropology Chinese Academy of Social Sciences, Beijing, 100081)

Abstract: Authors of this paper construct the corpus with syllable markers, word boundary markers and part of speech markers; its texts have been selected from Tibetan textbooks of Primary and middle school. And then the authors compare several POS tagging methods, the results prove that train data with the multi-level annotation can enhance the effects of POS tagging. There is also a strong relation between the part of speech of words and the part of speech of Tibetan syllables. So as for, authors use the POS of Tibetan syllables to predict POS of words. The results of experiments show that POS of syllables can correct some tagging errors caused in POS tagging.

Keyword: Tibetan Language； Tagging of Tibetan Syllables； Word Segmentation；POS

1. 藏语词性标注的现状和问题

词性标注研究指为给定句子中的每个词确定一个合适的词性的过程。词性标注研究是自然语言处理基础研究内容之一，在语音识别、信息检索等很多领域发挥着重要的作用。

藏语词性标注研究已经取得了一些成果，文献[1]采用隐马尔科夫模型，实现分词和词性标注一体化，最终词性标注的 F 值达到 79.494%；文献[2]采用了融合语言特征的最大熵词性标注模型，标注准确率达到 90.94％；文献[3]提出了利用感知机训练模型的判别式词性标注方法，经测试，准确率达 98.26％；文献[4]采用了最大熵和条件随机场相结合的标注方法，

---

*收稿日期： 定稿日期：
基金项目：本文研究得到国家自然科学基金资助（61202219，61303165，61132009），中国科学院信息化专项经费资助（XXH12504-1-10），中国社科院创新工程项目资助。

最终在开放测试中，标注准确率达到 89.12%。这些研究无疑对藏语文本词性自动标注做出了重要的贡献，但是同样也存在较多的问题，一是各家的词性标注规范不一致，二是词性标注的训练、测试语料不一致，三是都没有公开各自的标注系统，因此难以对各家的系统进行客观评价。这些研究都采用了统计模型进行词性标注，但可供统计训练的藏语标注文本数量不多，过多的未登录词也影响了标注准确率的提高。

本文作者提出基于藏语字性标注的合成词词性预测策略，主要思路是可通过标注藏语字性，根据字构词的规律，预测词的词性。藏字字性可以作为特征加融入统计模型中，也可以加入后处理模块对未登录词或者标注错误校正。文章第二部分比较了几种标注方法，说明多特征融合可以提高标注准确率，但对未登录词作用不大；第三部分讨论藏语字性和词性的关系；第四部分描述了基于字性的词性预测实验及结果。

2. 基于词的词性标注

在进行基于词的词性标注研究中，我们分别训练了几个不同的模型，独立分词模型，独立标注模型和分词标注一体化模型。训练分词、标注和分词标注一体化模型时，都采用了条件随机场工具包①，训练语料选自语素标注库（见 3.1 节介绍），按照 1:4 的比例，随机抽取 3987 句作为测试语料，其余 15952 句作为训练语料。

2.1 独立分词模型

分词时首先需要处理文本中的黏写形式，如 ངས 切分为 ང/ས，ཁོ་ཚོར 切分为 ཁོ་ཚོ/ར 等。藏文黏写形式切分可以采用多种方法，文献[5][6][7][8]分别做了阐述。本文在对黏写形式切分时，采用了把疑似黏写形式的音节全部切开，然后再根据上下文对非黏写形式进行合并，如，འདི་ག་ཚགས་པར་ཁང་དུ་ལྟ་ས་ནས་གནས་ཚུལ་འགྲོར་གསལ།/中 པར、ནས、གནས、འགྲོར 几个音节为疑似黏写形式，音节切分结果为：འདི་/ག་/ཚགས་/པ/ར་/ཁང་/དུ་/ལྟ་/ས་/ནས་/གན/ས་/ཚུལ་/འགྲོ/ར་/གསལ་/།/，然后采用四词位标注法对切分后的音节进行标注，其结果为：འདི་/B ག་/E ཚགས་/B པ/M ར་/M ཁང་/E དུ་/I ལྟ་/B ས་/E ནས་/I གན/B ས་/M ཚུལ་/E འགྲོ/B ར་/E གསལ་/I །/I。最后进行训练获得分词切分模型。表格 1 中数据为利用独立分词模型切分测试结果。

表格 1：独立分词实验结果②

计量单位	训练语料	测试语料	P	R	F
句（个）	15952	3987	94.0	94.0	94.0
词（个）	191996	48073			
音节（个）	208437	52355			
大小（KB）	2735	437			

2.2 独立标注模型

在独立分词的基础上进行单独标注实验时，为了比较分词结果对标注的影响，我们进行了两轮实验：分词后直接标注和对分词结果校正后再进行了标注。两个实验的结果如表格 2 所示。

表格 2：独立标注实验结果

计量单位	训练语料	测试语料	实验 1（未校对）			实验 2（校对）		
句（个）	15952	3987	P	R	F	P	R	F

---

①本文中使用的 CRF 工具包是 CRF++ 0.58 版，下载地址：http://taku910.github.io/crfpp/.

②本测试结果三项评测指标数据相同，纯属偶然，测试语料词有 48073 个，受测试的词有 48099 个。

词（个）	191996	48073						
音节（个）	208437	52355	0.832	0.830	0.831	0.876	0.875	0.876
大小（KB）	2735	437						

从表格 2 可以看出，分词的准确率对标注的效果影响明显，在分词未校正的情况下，标注准确率为 0.832、召回率为 0.830、F 值为 0.831。当对分词结果进行校对之后，各项测试指标分别提高到了 0.876，0.875 和 0.876，每项指标分别提高了 0.044、0.045、0.045，这说明分词的准确率影响标注的准确率。

2.3 分词标注一体化模型

分词标注一体化是在分词的同时进行词性标注。在训练模型时，把词边界标记和词性标签组合形成新的标注标签，如：ལྷ་ས 这个词的分词标签为 B(词始)、E（词尾），词性标注标签为 ns（地名），组合后标注标签为 ལྷ་/B_nsས་/E_ns。例如：ལྷ་ས་གྲོང་ཁྱེར་གྱི་དང་གི་མང་ཚོགས་ལམ་ཕྱོགས་ཀྱི་སློབ་གསོ་ལག་ལེན་བྱེད་སྒོ་སྤེལ་ཚུལ་སྐོར་ 的标注结果为：ལྷ་/B_nsས་/E_nsགྲོང་/B_ngཁྱེ/M_ngར་/E_ngགྱི་/I_kgདང་/I_niགི་/I_kgམང་/B_ngཚོགས་/E_ngལམ་/B_ngཕྱོགས་/E_ngཀྱི་/I_kgསློབ་/B_ngགསོ་/E_ngལག་/B_ngལེན་/E_ngབྱེད་/B_ngསྒོ་/E_ngསྤེལ་/B_ngཚུལ་/E_ng སྐོ/B_ngར་/E_ng。然后进行合并得到最终的分词和词性标注结果，如 ལྷ་ས་/ns གྲོང་ཁྱེར་/ng གྱི་/kg དང་/ni གི་/kg མང་ཚོགས་/ng ལམ་ཕྱོགས་/ng ཀྱི་/kg སློབ་གསོ་/ng ལག་ལེན་/ng བྱེད་སྒོ་/ng སྤེལ་ཚུལ་/ng སྐོར་/ng。

在分词标注一体化模型训练中，由于分词和标注组合标签比较多，训练的时间比较长（10 天左右），表格 3 列示了本实验的测试结果。

表格 3：分词标注一体化测试结果

计量单位		训练语料	测试语料	分词标注一体化		
句（个）		15952	3987	P	R	F
词（个）		191996	48073			
音节（个）		208437	52355	0.899	0.903	0.901
大小（KB）		2735	437			

正如我们所料，分词标注一体化模型的标注结果与独立分词、独立标注的结果相比，各项测试指标分别提高了 0.067、0.073 和 0.07；与校对分词后的标注结果相比，各项测试指标分别提高了 0.023、0.028、0.025。详细情况如图 1 所示。

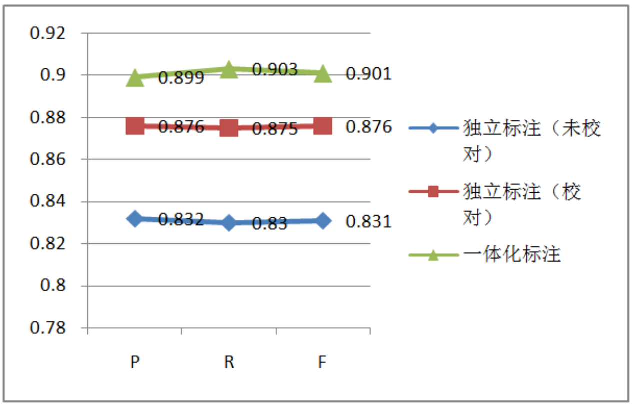

图 1. 标注结果比较图

这说明，在分词和标注一体化时，分词和标注之间相互影响，相辅相成，既可以避免一部分分词的错误，也可以避免部分标注错误，分词和标注实现了两者之间的优化组合。为了进一步考察分词标注一体化中分词的准确性，我们对分词标注一体化测试结果中的分词结果进行测试，发现一体化的分词结果与独立分词结果相比，准确率、召回率和 F 值分别提高到 0.943、0.948、0.945，与单独分词结果相比，各项测试指标分别提高 0.003、0.008、0.005。

3. 字性与词性的关系

在自然语言处理研究中，大多数语言模型采用词作为处理的最小单元，词是最小的能够表达独立意义的语言单位，因此基于词的语言模型具有较好的性能。但是对于汉藏语言来说，词的界限并不十分清晰，因词边界划分不一致导致了很多研究成果之间难以比较；另一方面，未登陆词也难以彻底解决。在一些研究中，字符（Letter）、子词（Subword）概念被提出，并证明了子词层级要比字符层级的语言模型好[9]。这说明在语言处理中，最小的语法单位可能是最佳的统计单位。对照藏语来看，一个非黏写的音节可能是统计语言模型可利用的最好单位。藏语的字大部分有意义（包括词汇意义和语法意义），如果能够充分利用这些有意义的字，可能会改善词性标注的性能。

3.1 字性标注

文献[10]界定了在文本信息处理中藏语字的定义。它不是指传统的前加字、上加字、基字、下加字、后加字和再后加字，也不是指文本中以分音点隔开的音节字，而是指“非黏写的音节字”，非黏写的音节字是指对黏写形式切分后的所有音节字。如ངས་“ngas”（我+格/吗）、ཚོར་“tshor”（复数标记+格/感受），它们既可能是一个音节，也可能是两个音节。如句子***ངས***ཁོང་***ཚོར***དཔེ་ཆ་གསུམ་སྟེར་པ་ཡིན། “ngas khong tshor dpe cha gsum ster pa yin”（我给他们三本书）中加黑斜体音节实际上都是两个音节黏连而成。

本文作者在前期研究中，构建了中小学藏文语文教材语料库，语料带有多种标记，藏字字性标记、分词切分标记、词性标记。语料格式为：<ནམ་/nམཁའ་/n>ng <སྔོ་/a>a <ཞིང་/c>c <དྭངས་/a>a <པ་/h>h <ལ/c>c <།/xp>xp，其中“<>”是词的分界标记，“/”是藏字分界标记，“/”右边的标注符号是藏字字性标记，“>”右边的标注符号是词性标记。语料库共有 19939 句（按照藏文单、双垂符作为分句标准，切分结果中有些不是完整意义的句子），总词数 240280，音节数 261412。语料库加工过程中，分词和词性标注遵循了文献[11][12]中各项原则，字性分类和标注遵循了文献[10]中的各项原则。

在藏字字性标注过程中，对人名、地名、音译名的藏字统一标注为 k，根据不同的专有名词类别，给 k 赋予区分标记，区分标记为词性标注符号的二级符号，构成人名的藏字标注为 kh，如<ཚེ་/khརིང་/kh>nh，构成地名的藏字标注为 kq（由于 ks，已经做为其他标注符号，为了区分，这里采用 kq），如<ཐྲུ་/kqརན་/kqསི/kq>ns，构成其他专有名词的藏字标注为 kz，如<ཡུམ་/kzབུ་/kzབླ་/kzསྒང་/kz>nz 等。

3.2 合成词词性特点

藏语的合成词由藏字构成，合成词根据不同的构造形式，可以分词复合型合成词，派生型合成词和重叠型合成词，复合型合成词指构成合成词的藏字有词汇意义，如བློ་བཟང(blo bzang, 意识 $_{[n]}$+好 $_{[a]}$)“智者”，构成该合成词的两个藏字分别是名词性的和形容词性的；派生型合成词指构成合成词的部分藏字有词汇意义，部分只有语法意义，表示实在意义的为词根，表示语法意义的为词缀。词缀包括前缀和后缀，后缀在藏语构词中占重要地位。如ཆེ་བ(che ba, 大 $_{[a]}$+ba $_{[s]}$)“大的”，构成该派生合成词的两个语素 ཆེ་(che)为词根，(བ)ba 为后缀。重叠型合成词指以重叠词根或者词缀构成的合成词，藏语中也包括通过语音屈折变化构成的表状态的形容词，如 ཀྱག་གེ་ཀྱོག་གེ(kyag ge kyog ge)“弯弯曲曲”，ཀྱག་ཀྱོག་(kyag kyog)“弯曲”，དགའ་དགའ་སྤྲོ་སྤྲོ་(dgav dgav spro spro)“高高兴兴”等。

合成词的词性可以根据构成合成词的藏字的字性推断。与名词相关的构造方式有：n+n->n，如：མི་གྲངས“人数”、ལས་དོན“事务”；n+v->n，如：རྒྱུན་མཐུད（持续）、ཁྱད་ལྡན（特色）、རྒྱུན་རིང་（持久）、ཤུགས་སྣོན（压力）、མེ་ཟོན་（灭火）、སྙུན་བྲལ（健康）、རང་རྟོགས་“自觉”、རང་འགུལ་“自动”；n+a->n，如：ཡུན་རིང་“长期”、སྟོབས་ཆེ“大力”、རླབས་ཆེན“伟大”、ཁྱོན་ཡོངས་“全面”；n+nf->n，如：རྩ་བ“根”、རྣམ་པ་“形象”、ལག་པ“手”；n+m->n，如：ཕྱོགས་བཞི“四方”、མཐའ་གཅིག“一端”；v+n->n，如：འགོག་རྐྱེན“故障”、ཐོན་ལས“产业”、འཐབ་ཇུས“战略”、གྲུབ་ཆ་“成分”、སྒྲིག་སྲོལ“体制”等。

与动词性藏字相关的构造方式有: v+v->n，如：ལྟ་སྐུལ“监督”、འཆར་འགོད“规划”、འདེགས་སྐྱོར“支持”、འདོན་སྤེལ“发挥”、གཟིགས་སྐྱོང“优惠”；v+vf->n，如：དགོས་པ“需求”、རྩོམ་པ་པོ“作者”。

与形容词性藏字相关的构造方式有：a+v->n，如：དམ་འཛིན“抓紧”、གསར་གཏོད“创新”、གསར་འབྱེད་“开拓”、མཉམ་གནས“共处”；a+a->n，如：ཞི་མཐུན“和谐”、བརྟན་ལྷིང“稳定”、དམ་ཟབ་“密切”、མཛེས་སྡུག“优美”；a+af->a，如：གསར་པ“新的”、བཟང་པོ་“好的”；a+n->n，如：མཉམ་ཤུགས་“合力”、ཆུང་གྲས“小辈”等。

4. 词性预测实验及结果

4.1 实验设计

我们原本设想，联合利用字性、分词标记和词性标记训练一个模型，以此考查标注效果，但由于训练时间过长而中断。因此采用了另一种方案，利用藏字字性标记和词边界标记两个特征，训练了一个能同时给出藏字字性标注和分词标记的模型，然后利用藏字构词的规则来对基于词的标注模型的错误例子进行校正。整个过程如图 2 所示。

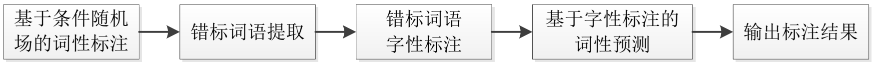

图 2. 实验流程

4.2 词性预测结果及分析

如果采用分词、标注一体化模型，错误标注结果中区分不开是分词还是标注导致的错误，因此我们采用了分词校正后独立标注模型进行实验，然后提取标注错误例子，对错误例子进行字性标注和利用字构词的规则对复合词或结构进行预测。

从评测结果中提取出了约 5900 个错误例子，通过分析发现标注错误包括：在语料中，存在同一个词的相同用法却标注不一致现象，一些特殊符号未给出正确标注，这种错误占比约 20%，这种问题可以通过进一步调节语料，提高训练和测试语料的一致性，对特殊符号进行统一处理等方法来解决。在其余错误标注中，两个藏字及以上的复合词或者短语标注错误和单字词标注错误各占约 40%。利用藏字字性和构词规则，有 1888 个标注错误得到修正，标注准确率提高了约 0.04，这个标注结果已经高于分词标注一体化的效果。几种标注结果如图 3 所示。

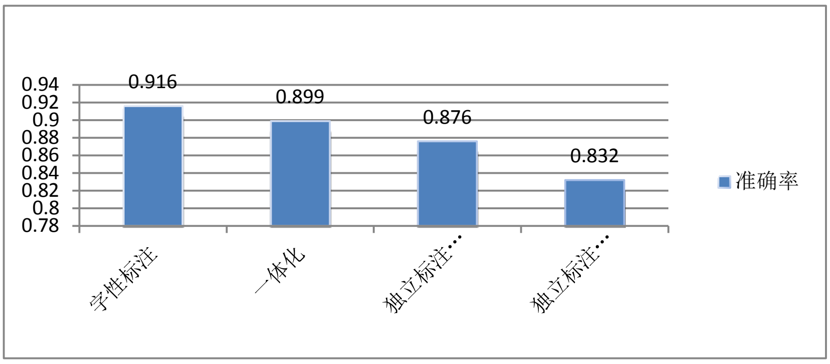

图 3. 几种标注实验结果对比

表格 4 列出了部分标注错误能够通过规则预测得到正确的标注结果。

在表格 4 中，一些标注错误利用简单的规则就可以校正，例如 མགྱོགས 形容词性藏字，པོ་为后缀，a+af 的形式一定不是一个动词，也不大可能为名词；同样 བཞི་ 是数词性藏字，加后缀后一般是表序数的数词，极少为名词的情况。

表格 4：藏字字性校正合成词标注错误示例

预测结果（正确）	测试结果（错误）	预测规则	预测结果（正确）	测试结果（错误）	预测规则
གཡོ་སྒྱུ་ཅན་/ng	གཡོ་སྒྱུ་ཅན་/a	v+v+vf=ng	ཕྱིར་འཁྱུར་/iv	ཕྱིར་འཁྱུར་/ng	n+kx+v=iv
བཞི་པོ/m	བཞི་པོ/ng	m+mf=m	ལྷག་པར་/id	ལྷག་པར་/ng	v+vf+uf=id
རྟག་ཏུ་/d	རྟག་ཏུ་/ng	a+uf=d	ལྷིང་རེ་རེ་/ia	ལྷིང་རེ་རེ་/ng	v+k+k=ia
མགྱོགས་པོ་/a	མགྱོགས་པོ་/ng	a+af=a	གསེར་ཉ/ng	གསེར་ཉ/a	a+n=n
མ་ཐག་ཏུ་/d	མ་ཐག་ཏུ་/ng	dn+v+uf=d	འཕྲལ་དུ་/d	འཕྲལ་དུ་/ng	n+kl=id
མཆོངས་རྒྱུག་རྒྱུག་/iv	མཆོངས་རྒྱུག་རྒྱུག་/ng	v+v+v=iv	སྐམ་པོ་/a	སྐམ་པོ་/ng	a+af=a
མགྱོགས་པོ་/a	མགྱོགས་པོ་/vt	a+af=a	ཧྲོ་སོབ་སོབ་/ia	ཧྲོ་སོབ་སོབ་/ng	a+k+k=ia
ཁ་ཡ་མ་བྱས/iv	ཁ་ཡ་མ་བྱས/a	k+k+dn+v=iv	དྲུག་པོ/m	དྲུག་པོ/ng	m+mf=m

5. 结论

字的概念在汉藏语研究中有着独特的地位，以字（基本上叫语素）为单位进行研究是语言学家长期关注的对象，但是近些年，在文本信息处理、语音识别、语音合成研究中，字的概念（Sub-Word,Sub-Syllable）也得到广泛关注。本文比较多种标注方法，尽管复合特征能够提高标注准确率，但是未登录词等问题不能根本解决。为此，我们利用藏字字性，通过字构词的规律预测合成词或短语的标注问题，经过测试标注准确率提高到 0.916。尽管语料规模有限，加工精度有待提高，但这个研究策略值得进一步探究。

6. 参考文献

---

[1] 史晓东,卢亚军.央金藏文分词系统[J].中文信息学报,2011,25(4):54-56.DOI:10.3969/j.issn.1003-0077.2011.04.011.

[2] 于洪志,李亚超,汪昆等.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.DOI:10.3969/j.issn.1003-0077.2013.05.023.

[3] 华却才让,刘群,赵海兴等.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. DOI:10.3969/j.issn.1003-0077.2014.02.008.

[4] 康才畯.藏语分词与词性标注研究[D].上海师范大学,2014.

---

[5] 康才畯,龙从军,江荻.基于词位的藏文黏写形式的切分[J].计算机工程与应用,2014,(11):218-222.

[6] 才智杰.藏文自动分词系统中紧缩词的识别[J].中文信息学报,2009,23(1):35-37.

[7] 巴桑杰布,羊毛卓玛,欧珠等.藏文分词系统中紧缩格识别和藏字复原的算法研究[J].西藏科技,2012,(2):73-75,79.

[8] 李亚超,加羊吉,宗成庆等.基于条件随机场的藏语自动分词方法研究与实现[J].中文信息学报,2013,27(4):52-58.

[9]Tomáš Mikolov, IlyaSutskever,Hai-Son Leetc.Subword Language Modelingwith Neural Networks, www.fit.vutbr.cz/~imikolov/rnnlm/char.pdf.

[10] 龙从军,刘汇丹,吴健.藏语字性标注研究, 第十五届中国少数民族语言文字信息处理学术研讨会,延边,2015 年 8 月 11-13 日.

[11] 赵小兵,孙媛,龙从军等.藏文拉丁转写、分词和词性分类规范-信息处理用现代藏语分词规范（草案）,商务印书馆, 2015 年 6 月.

[12] 赵小兵,孙媛,龙从军等.藏文拉丁转写、分词和词性分类规范-信息处理用现代藏语词性标注规范（草案）,商务印书馆，2015 年 6 月.

作者简介：

龙从军（1978——），男，博士，主要研究领域：藏语语法、藏语信息处理。Email:longcj@cass.org.cn.

刘汇丹（1982——）男，博士，高级工程师，主要研究领域：藏语信息处理。

诺明花（1981——），女，博士，主要研究领域：藏语信息处理。