中国民族语言研究与应用

第一辑

龙从军　燕海雄 主编

中国社会科学出版社

中国民族语言研究与应用

第一辑

龙从军　燕海雄　主编

中国社会科学出版社

图书在版编目（CIP）数据

中国民族语言研究与应用. 第1辑 / 龙从军，燕海雄主编.
—北京：中国社会科学出版社，2016.6
ISBN 978-7-5161-7252-0

Ⅰ.①中… Ⅱ.①龙… ②燕… Ⅲ.①民族语-研究-中国 Ⅳ.①H2

中国版本图书馆CIP数据核字（2015）第283392号

出 版 人 赵剑英
责任编辑 任 明
责任校对 朱妍洁
责任印制 李寡寡

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲158号
邮 编 100720
网 址 http://www.csspw.cn
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印刷装订 北京市兴怀印刷厂
版 次 2016年6月第1版
印 次 2016年6月第1次印刷

开 本 710×1000 1/16
印 张 14.5
插 页 2
字 数 318千字
定 价 68.00元

凡购买中国社会科学出版社图书，如有质量问题请与本社营销中心联系调换
电话：010-84083683

编 委 会

心智层次与深度学习

代 序

当代社会，信息技术是如此快速地渗入我们的工作与生活，从孤立的个人电脑到全球网络数字移动系统，人类身处的世界已再度翻篇。忆往昔，数千年前，庄子的混沌天地万物一体，无数据可言，唯思绪可鹏程万里："风起北方，一西一东，有上彷徨，孰噓吸是？孰居无事而披拂是？敢问何故？"察现今，拜大数据之福，世界万物关联的曙光初现，从孔仲尼到今天的你已没有那么远的时间距离，喜马拉雅山离你地理上亦不再遥不可及。工程师笑了，我有数据可计算了；哲学家笑了，我有事物可辩证了；语言学家也笑了，我有词语可聊了。大数据的精髓并非数据之多之大，实为数据之多之广，无论哪个领域、无论何种范畴，凡世界凡人类之数据之集合方可称为大数据，大数据为万物的联系本质提供了比较和计算的可能。

数据的生命在于解读，解读的方法是语言表达。可是，大数据的建立同时又带来数据芜杂难以萃取的困难，这就是大数据陷阱。好在一物降一物，人总是有办法。据我所知，在大数据概念形成过程中，人们已经发展出一种叫作深度学习的算法，这是一种机器学习的模型，具有 AI 性质。它的认知模型来源于心理学领域，即人脑具有深度心智结构，因此可以让机器模拟人脑神经网络，构建逐层深度学习模型。究竟何为心智深度？又如何建立深度学习模型？且举一个藏语案例。$[[skad]_N[yag\text{-}po]_{ADJ}]_N$ 结构上是典型修饰型名词短语$[N+ADJ]_N$，意为"好声音"，这是该结构的表层意思；可是在隐喻作用下该结构词汇化为形容词，充当其他名词的修饰语（还可做谓语，例略），即$[[bu\text{-}mo]_N[skad\text{-}yag\text{-}po]_{ADJ}]_N$，意为"[嗓子好的]姑娘"，这就是所谓深层结构。也就是说，字符串 skad yag po 即可理解为名词短语，也可理解为形容词，在真实文本中如何识别和处理呢？对于这个不难也不易的问题，我乐于建议当代年轻学者不妨利用大数据尝试解决，创造这个问题的深度学习模型。

经过 20 余年艰辛努力，我所自然语言处理从理论到实践、从基础到应用，以藏语文为探索对象，紧随中文信息处理主流研究领域的发展，越过了数次观念的和技术的门槛，成长为中国中文信息处理领域一支重要研究力量。我相信，这股力量仍将保持活力，迈向新的高峰。

这本论文集由龙从军博士、燕海雄博士负责主编，他们联系国内藏语文自然语言处理和基础研究领域多位青年学子共同编撰，又得到部分长辈学者赐稿，我为他们的成长高兴。从军博士请我写一篇短文代作序言，我也欣然领命。目前，研究所机构发生变动，藏语文研究领域重组到应用语言学研究室，该室主任王锋教授提议组成新的编委会，继续编撰出版这个系列集子，是为学科发展之长远之策。

江　获

2014 年 12 月于北京

目　录

计算语言学研究

基 础 研 究

名 著 翻 译

计算语言学研究

现代藏语的机器处理及发展之路

——组块识别透视语言自动理解的方法

江　荻

[提要] 本文对现代藏语机器自动处理方面所涉及的藏文编码标准、句法分析、计算研究各个方面做了简略的回顾性概述，认为藏语计算研究在基础研究不足、经验累积不够、研究成果零散的情况下，应该积极调整思路，结合藏语自身句法特征，发掘新的研究方法。文章指出，在藏语处理尚未经历成熟的自动分词和语料标注条件下，可以直接尝试一种藏语组块分析和分词同步进行的方法，即利用藏语句法形式标记直接识别组块和进行块内分词的策略，同时还提出抽取句法语义信息为下一步实现句法关系的识别铺垫基础。

[关键词] 藏语；究现状；形式标记；句法组块；句法关系；语言计算

1　藏语研究现状

挑这个话题应该是有感而发。虽然不同的语言在机器处理和自动识别上可以有其相同的一面，都一定程度上遵循着普遍的算法规则和语言规律的共性，可论述的范围和焦点以及具体的操作方法却可能千差万别。

就自然语言处理领域而言，藏语差不多还是块处女地，真正作为自然语言处理核心内容的研究和论述还很少，涉及的范围也有限。不过，与这个领域相关的前期研究还是需要论及的。首个相关话题是藏文编码标准或者藏文平台建设，这种百米跑道起点的问题，大概汉语、英语现在不必多谈，藏语处理中还是很重要的内容。更关键的话题是，在藏语的计算处理中，已有的经验是什么，开展了哪些句法分析和算法研究，并解决了哪些形式识别和基础建设问题。

1.1 藏文编码标准及平台建设

藏文平台建设涉及两个主要内容，一是标准，国际标准化组织的ISO/IEC 标准和中国国家标准（GB），二是实际的藏文应用系统。藏文国际标准是 1995 年依据中国提出的方案在斯德哥尔摩会议通过的［ISO/IEC 10646-1（1993/P.DAM.6-ISO/IEC JTC1/SC2/ WG2 N 2627:1995）］，中国国家标准是1998年正式发布的《信息技术 信息交换用藏文编码字符集 基本集》（GB16959—1997），与此同时，国家标准《信息技术 藏文编码字符集（基本集）24*48 点阵字形 第一部分：白体》（GB/T 16960.1—1997）也正式发布。

可是，建立在ISO10646-1 的基本平面 00 组 00 平面的藏文《基本集》（即 UCS 的基本多文种平面，机内码 0F00-0FBF，占用 192 个码位）只提供了 168 个编码字符。就藏文的二维构造方式（字母在纵向和横向两个维度上构字）来说，这样小规模的字符根本就没有实用的可能性。这就导致了利用 B 平面的设想，提出建立信息交换用藏文编码字符集辅助集的办法，把辅助集放在 UCS 的拼音文字辅助平面。2000 年以后，ISO/IEC 10646-1:2000 为藏文编码追加了部分空间，从 0F00 到 0FCF，但仍不能完全解决空间需求问题。不过，近年关于 ISO 10646/Unicode 的持续发展为藏文编码带来了全面实现的可能。

另外，近二十年来，藏文的计算机实用技术一直在不断发展，国内外开发了多种藏文字库和应用软件，最有名的是北大方正的藏文排版系统，包括键盘输入和打印输出。最近，西北民族学院研制的同元藏文系统实现了 Internet 网浏览功能和其他辅助功能。值得指出的是，虽然国外的系统多采用小字符集，但像方正这类的藏文实用系统都是另外设计了中字符集（600—700 字符），这似乎说明已有的标准与社会实际需求存在着差距。

把藏文平台建设作为藏语计算处理的家当是历史的现实，只有有了标准或者有了平台，藏语的计算处理才谈得上可能。举例来说，藏语语料或机器词典的排序、检索以及统计都需要依据一定的序性规则[13]，而这样的规则无法从内码编码中获取。因此，一方面要依据藏语的编码结构和字的构造特征，另一方面要设计恰当的算法予以解决。这些研究就需要依靠编码标准或者在应用平台上来实现[4]、[12]、[13]。

1.2 藏语语法研究

尽管藏语文的计算机处理已有二十年的历程，但绝大部分力量集中在编码和平台建设方面。真正作为自然语言处理核心内容的藏语自然语言处理或者计算语言学研究似乎只有一些零散的表述和对浅层形式的认识。

就现有的研究状况来看，最基本的困难是，缺乏一套现代意义的完整藏语语法体系，甚至缺少至少可以“照猫画虎”的操作起点。有着1000多年历史的传统藏语文法研究不能说没有成就，学龄的藏族蒙童都是从这里起步。但就自然语言机器处理而言，这样的体系犹如月光普照大地：轮廓似乎可见，什么也无法看清。传统藏文文法主要包括两项内容，一是虚词用法，主要包括各类词格、语气词和连词等的用法描述，二是字性语音分析，主要围绕动词语音形态论述动词的用法。这两项内容基本维系了吐蕃时期藏文创始人吞米桑布扎编撰的《三十颂》和《字性组织法》[26]格局，虽然历代学者不断纾难释疑，基本内容不断调整和补充扩展，但传统体例没有变化。

传统文法之外，近当代还有不少现代藏语语法论述，包括历代西方学者出版的一些藏语语法专著和国内学者的语法研究。例如匈牙利学者乔玛（Csoma de Körös）1834年出版的 *A Grammar of the Tibetan Language*，其后英国学者 H. August Jäschke 和印度学者 Sarat Chandra Das 等人也分别于19世纪后期和 20 世纪初叶出版了藏语语法专著，最新的论著有 Philip Denwood 1999年的 *Tibetan*。汉族学者在20世纪50年代中期开始陆续展开藏语语法研究，如张琨、金鹏等人。其中最有特色的研究是胡坦等人主编的《拉萨口语读本》，既是教科书又可以称得上拉萨口语语法论述。其他还有一些单篇论文对藏语语法现象做了较深入研究，涉及范围包括名词、动词、形容词以及各类虚词和一些语法范畴，如胡坦《拉萨藏语几种动词句式的分析》（《民族语文》1984年第1期），张济川《藏语的使动、时式、自主范畴》（《民族语文》1989年第2期），周季文、谢后芳《藏文阅读入门》（云南民族出版社，1998年），等等。但总体上说，大多数的论著或者系统性不强，或者只是一个描写的粗框架，语法的分类专项研究基本没有。

词汇研究方面也不乐观，零零星星只发表过少量构词和形容词重叠以及四音格一类的讨论。反倒是词典方面建树较大，除 Jäschke、Das 等国外学者早期的词典外，1949年国内格西曲扎木刻出版了第一本按现代藏文序编纂的单语词典，1957年民族出版社铅印出版了该词典的藏汉对照本。最有影响的词典是张怡荪主编的《藏汉大词典》（民族出版社，1985年），是迄今最大最全的藏文词典，而且部分词条标注了词性。

以上只是大致勾勒了藏语计算处理方面所涉及的词汇、语法面貌，在这份家当中，比较直接可用于机器处理的包括传统文法的部分内容，如传统文法提供了一部分非常精细的关于字、词书面变体形式的续连规则（传统称为字性添接法），而动词时（现在时/未来时/过去时）、式（陈述式/命令式）、态（使动/自动）形式差别的详尽区分也可直接形成算法处理用的动词

词形表。至于传统词格等虚词描述与现代语法观念差异甚巨，很难直接采用。现代藏语词典或双语词典也都需要进行整理，剔除不合适的长词语（如相当数量佛教用语），添加词性标注，等等。至于语法体系方面，虽然词法和句法上确有不少真知灼见，则还需要费很大力气从已有成果中进行扒梳，重构框架和填充细节，或设法另创体系。所以说，藏语语法描写的先天不足将给现在和未来的自然语言处理加载沉重负担。

1.3 藏语计算研究

也许就是在这种条件和环境下，藏语自然语言处理一直处于缓慢发展阶段，各种研究呈现零散、局部状态。究竟这方面我们取得了哪些进展呢？不妨简略叙述（不含编码处理）。

分词标准　像汉藏语这类书面形式无词界的语言一般都存在分词问题，就汉语来说，一般公认需要建立分词的标准。藏语的分词研究刚刚起步之时，已经注意到规范性和标准性预设问题，因此有学者初步提出一个藏语分词标准讨论稿[18]。这项研究对藏语书面语包括标点符号在内设计了36条分词基本规则。例如，讨论稿提出词格标记作为分词单位，这就引出开音节韵尾词根后附黏着型词格标记的识别和切分问题。如“我+属格”，“我+施格”，其中标记ཉ和ས分别只是属格和施格的5种变体形式（属格标记འི、ཡི、ཀྱི、གྱི、ཏི/施格标记གིས、ས、ཀྱིས、ཀྱིས、ཏིས）中的一种。这条规则导致技术处理上要建立相关的续连规则和词根归一化算法，不仅要切分出黏着形式，还必须识别格标记与一般的构词辅音韵尾同形形式。有点像英语，bring的-ing并非现在分词后缀，corpus的-s并非复数后缀[2]。

藏语分词规范讨论稿也有一些明显矛盾的论述，如一方面规定动词后的语尾助词或动词语尾助词组块为分词单位（含时、体、人称、意愿等句法范畴意义），另一方面又规定中嵌于多音动词和形容词的否定词为分词单位，那么中嵌于多音语尾助词组块之间的否定词却没有规定。如果包含否定词的多音节语尾助词组块在算法处理上不宜割裂，那么多音动词或形容词中嵌的否定词是否需要另建一套处理规则？

语料规范　藏语书面语与口语脱节是众所周知的现象，即使书面语也存在巨大的差别，区分韵文体和散文体。长篇史诗般的口传文学《格萨尔传》就是韵文体传记。更棘手的是，相当多的文献往往是各种文体和语体混用，那么藏语自然语言处理的对象究竟怎样定位呢？文献[24]对书面藏语语料做了切实的分析和统计，并从题材、文体、语体、著译四个角度进行观察，其中题材分类包括文学类、政论历史类、专门类（宗教、历算、因明、医药等）；文体分类包括散文体、韵文体、散韵混合体；语体分类包括

文言和口语两大类，文言又分四小类：古体文言、质朴文言、藻饰文言、浅近文言；著译分类主要指直接用藏文创作的或从梵文、汉文翻译的作品。通过 500 万词语料的统计，发现文学类和口语类兼纳了常用和次常用性最高的词以及通用性最高的词，政论类则以通用性词为主，罕用词基本出现在题材分类的专门类中以及语体分类中的古体文言、藻饰文言中。根据这个分析，藏语计算处理的语料对象可以初步确定为现当代书面文学作品和报刊著译作品。

机读词典　研制藏语机读词典也是开展藏语自然语言处理的基本建设，语言中的基本词汇或语素是自然语言处理中最可靠和最有效的知识来源。根据藏语的特点，所建立的藏语机读词典既要有共性也要有个性。首先，确定基本原则，包括通用原则、规范原则、语法原则、稳定原则、构词原则、双语原则，等等[10]。其次，利用语法标注集尽量给出每个词的词法（和部分句法）描述，包括词性标注和语用属性标注（敬语词、佛教词、梵语借词、异体词等）。藏语机读词典不仅可以服务于文本语料分词匹配处理等用途，而且它自身也能提供词语的静态结构面貌和统计数据。例如，在 5 万多词条的词典中可以基本了解双音动词、双音名词以及双音形容词的构造情况和统计比率。全部双音词中，名词占双音词的 86%，动词占 3%，形容词占 6%，三项总共占双音词的 95%。而双音名词中，约 1/6 的名词是通过添加词缀构成的。其他绝大部分名词是词根语素复合构成。双音动词都是由名词语素加动词语素构成的，又分为动宾式和主谓式两种。双音形容词的构词方式主要是附加词缀法、重叠法和复合法。在 1800 个形容词中附加词缀的占 45%，音节重叠的占 16%，复合方式构成的占 39%。不过，名词和形容词的语素构成都比较复杂，还需要进一步展开研究。

计算处理　藏语计算处理方面的尝试大多比较零散和具体，尚未形成持续的技术路线。例如，针对书面语标点的贫乏（书面藏语真正作标点的单垂符可用来分开短语或句子，相当于顿号、逗号和句号，以及其他如问号、感叹号等句终符号），文献[10]提出利用句尾标记依照句式类型识别句界的问题，并对藏语疑问句进行了具体实践[19]。在分词方面，早期的分词方法主要是基于词典的机械分词方法，算法简单。进一步的改进是通过识别虚词标记的分段切分方法，在每个分段内进行词典匹配。同时为了解决歧义切分问题，还采用了任意词和句尾词的人工干预“后校验”辅助手段[10]。其他相关的计算研究还有藏语信息熵分析和语料库技术处理[14-15]。

最近，作为藏语分词体系提出的方案有两类，一是依据藏语句法特征，通过定义藏语组块提出藏语句法理解的组块分析和分词方法[11]，该方案的核心在于：以标记组块的结构关系分析为跳板，直接逼近句法语义理解的

功能成分，而分词只是为了抽取组块句法信息用于揭示组块句法关系的辅助手段。具体操作上，该方案实施步骤是先对文本进行预处理，包括采用续连规则集和各类专项词表对句法标记进行根词归一化的处理，然后进行组块及组块边界识别，并展开组块内部分词、词性标注和信息提取，最后进行组块功能分析与整合以达到对句子的自动理解。有关这方面的论述以及相关认知理论基础下文还会进一步论述。第二种方案称为基于格助词和接续特征的分词方案，该方案“利用字切分特征和字性库先‘认字’，再用标点符号和关联词‘断句’，用格助词‘分块’，再用词典‘认词’”，最终达到分词的目的[6]。由于这个方案以分词为具体实施目标，特别注重格助词（即词格标记）的分块作用，并力图“摆脱词典的束缚”，可以说很有特色。

以上论述算是对藏语自然语言处理领域现有的家底做了一番粗略的清理，总结其中的经验和教训对于选择将来的路子很有裨益。比较明显的一个倾向是，在现有的两种分词方案中都涉及藏语句法结构上的“自然块”，其中组块分词方案采用组块分析和块内分词合一方法既吸收了近年自然语言处理新思路，又结合了藏语句法特征，应该是一种有前景的方法。

2 分词与机器理解

机器能思考吗？这是新近出版的《剑桥五重奏》一书的副标题。该书机智风趣地把图灵、霍尔丹、薛定谔、维特根斯坦、C. P. 斯诺“纠集”在一起讨论思维与机器、意义与机器、语言与机器问题。把这个命题放在自然语言理解领域，似乎可以换上一个问法：机器理解能够向人的理解学习吗？

这段话的意思是要引起这样的思考：为什么要分词？人对话语或句子的理解是以词为基础的，词是最小的能独立运用的意义单位。显而易见，机器要理解自然语言，也要从词的理解开始。最典型的范式是查词典，人理解需要查词典，机器理解也需要查词典。这里的比对、这里的逻辑似乎并没有什么深奥费解的地方。

我们犯错误了吗？很有可能。人们理解话语并不经常查词典，他们有许许多多其他办法理解语句。机器则不一样，任何时候都需要查词典。人具有抽象能力，他们并不完全直接依赖句中的词来理解。从语言理解的过程看，人对语句的理解是一种抽取深层命题结构和逻辑推理加工过程，然后构建关系和意义。具体过程是词识别、结构组合、关系整合。所谓构建句子意义，实际上就是解释句子中功能成分的各种关系，而功能成分的形成则是对词语关系的加工（还要包括经验、语境、语用等因素）。换句话说，

无论语句由多少词语构成，人们总是把其中代表命题和推理的核心词抽象出来，形成有限的表达关系，如陈述、指称、修饰、补足等。用句法关系来说，也就是主谓关系、述宾关系、修饰关系和补足关系等。如果我们把这样的关系类型告诉机器，机器设法实现对这些有限关系的识别，这似乎就可以称得上是一种浅层机器理解。而我们即使把“苍蝇”和“蝴蝶”的各种细微句法语义属性（如 wordnet 描述）都告诉机器，它们所蕴含的人的理解意义仍然不会为机器所理解，前者不会带来厌恶，后者也不会带来愉悦。

按照这种观点，机器理解可以向人的理解学习。但是，它要学习它所能学的，也就是说自然语言处理专家要知道机器能够学习什么。当然，这并不是说机器不能够学习词语知识，俞士汶和董振东分别把大量汉语词语的句法和语义属性知识教给了机器，但所有这些知识都是属性描述和词与词线性关系的描述，它可以帮助机器区别不同词语，指明词语间的关联关系，但不能创造新的内涵、直觉和新的关系，不能包含未指明的经验和属性，况且我们所能给予它的知识本身也是不完备的。看来，机器理解与人的理解仍不可同日而语。

总结以上关于机器理解和理解内容的讨论，我们看到机器的理解实际上是关于不同词串能否组合以及组合类型的问题。乔姆斯基那句有名的范例 Colorless green ideas sleep furiously 机器同样可以“正确无误”地理解，无非是主语与谓语或者主题与述题的关系。至于意义，那是人给机器所教知识完备与否问题，或者人现在还没有办法获得完备的关于世界和语言的知识并将其教给机器。

我们认为，正是基于这样的认识，人们开始对于词在机器理解中的作用开始发生变化，同时开始调整分词技术在自然语言处理中的地位。词是理解的基础，是构成语言下位成分以及关系的要素，而完整句子的理解是由上位成分及相互关系构成的。既然机器并不能像人那样真正理解这些基础要素本身，不必与之为难，它能告诉我们各种关系即可。从语素到词，到短语，再到句子，各种关系逐层获取。对机器来说，句子层面的理解可以说就是各种短语结构，或者功能成分的相互关系以及关系类型，即一般所指的句法树。

刘源等学者指出：“在汉语信息处理中，词的处理是基础，短语的处理是中心。”[17] 这个观念意味着词的处理要转移到以短语的处理为重点上来。后来，周强、孙茂松、黄昌宁正式提出了汉语句子的组块分析体系[25]，并开始构建大规模的汉语语块库[26]。

组块分析是依据语言的句法特征建立起来的，开始逼近句法关系的理