◁ BIT ▷

藏语NLP

组员：崔铭元

时间：2021/10/25

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

德以明理　学以精工

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目录

CONTENTS

01

前言

ABSTRACT

藏文历史

藏族是中华民族中历史悠久、文化源远流长、人口众多、分布较广的古老民族之一。藏文是藏族群众主要的交流工具。藏语属汉藏语系藏缅语族藏语支。是一种具有1400多年的古老拼音文字。

研究意义

类似汉语NLP,藏文NLP一方面是广泛的应用于藏文的舆情分析、文本分类、智能问答、信息抽取等领域，为现代藏语大环境做贡献。另一方面，历史上用藏文撰写的各类典籍数量庞大，在国内仅次于汉文。

藏文NLP的发展历程

分词研究

藏文是一种拼音文字，由字母组成的音节构词。藏语词语之间没有明显的分隔符来进行区分。

词性标注

藏文信息处理技术中的一项基础性课题。应用于信息抽取、信息检索等；也是藏语语块、句法、语义等分析器。

汉藏机器翻译

汉藏机器翻译能打破汉藏两种语言文字之间的障碍，加速双向的信息传播。对藏区的经济发展和文化交流有着巨大的促进。

2.1 藏文分词发展

藏文分词的研究相对较早，1999年始主要是基于语法规则的分词方法，2009年基于统计的分词方法开始萌芽并成长，近几年内统计与规则相结合的方法备受关注。语法规则方法主要是利用词典，格助词和虚词等规则用匹配算法及进行分词。统计方法是通过训练统计模型，按概率做分词的结果。

2.2 藏文词性标注

藏文词性的标祝研究较晚，起始于2005年。初期的研究由于藏语自身的语法特点和相关语料库的匮乏，进展缓慢。到2010年苏俊峰开拓了以隐马尔科夫模型（HMM）为核心的标注方法，2014年华却才让使用了感知机训练模式进行标注（TiPosTag），2015年李亚超团队用条件随机场和最大熵模型开发了目前最高效的藏文分词标注工具TIP-LAS。

2.3 汉藏文机器翻译

于1997年格桑志玛提出了汉藏机器翻译的构想之后，学界开始了对汉藏机器翻译的探索，初期基本停留于对于动词，句法在翻译上的技术讨论。2011年才让加发表了对汉藏语料库的构建技术研究后，汉藏机翻的发展步入正轨，现在的主流技术分为三种：基于规则的，基于统计的和基于神经网络的汉藏机器翻译。

02

藏语概况

GRAMMAR
BASICS

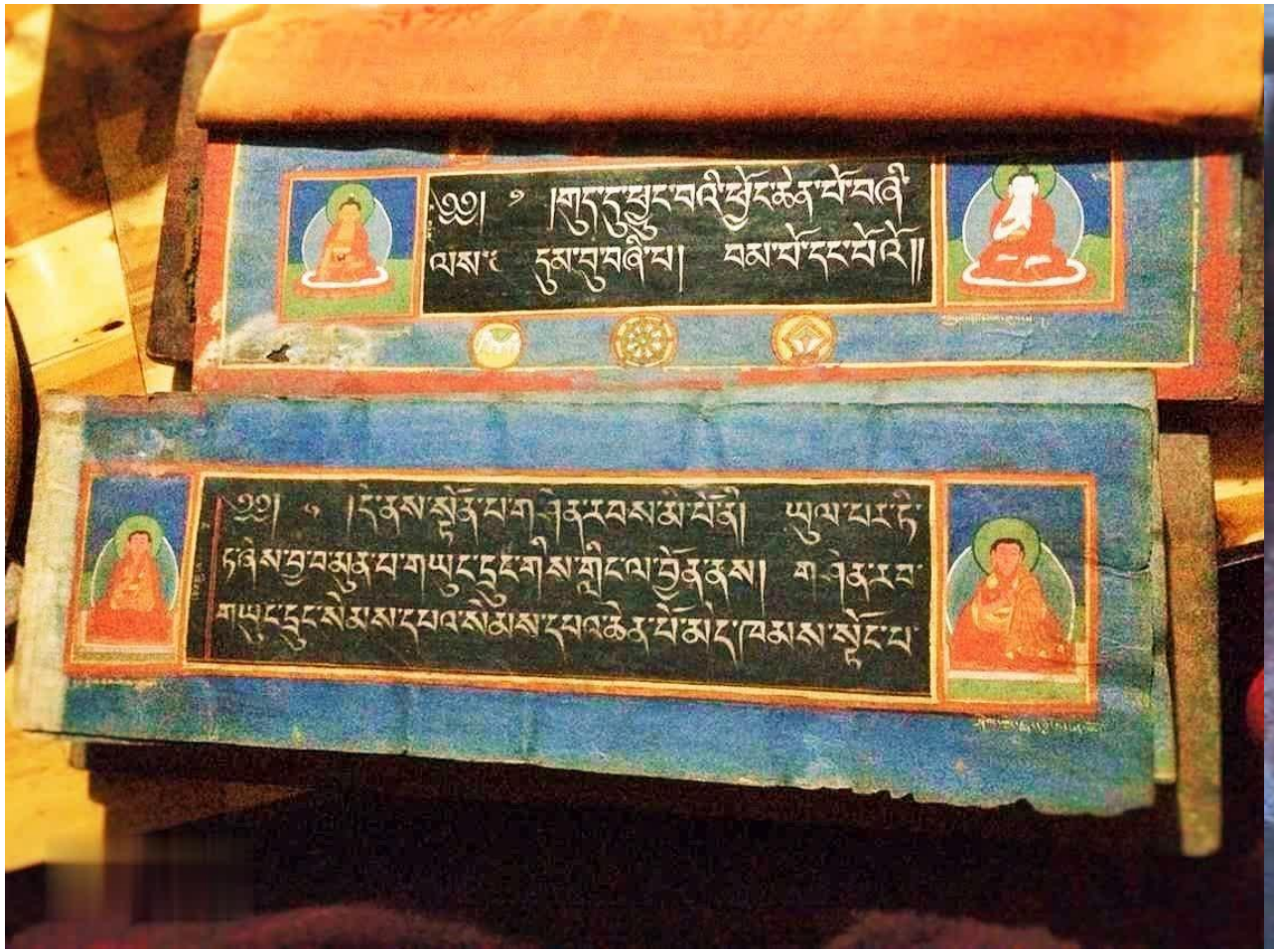
藏文历史

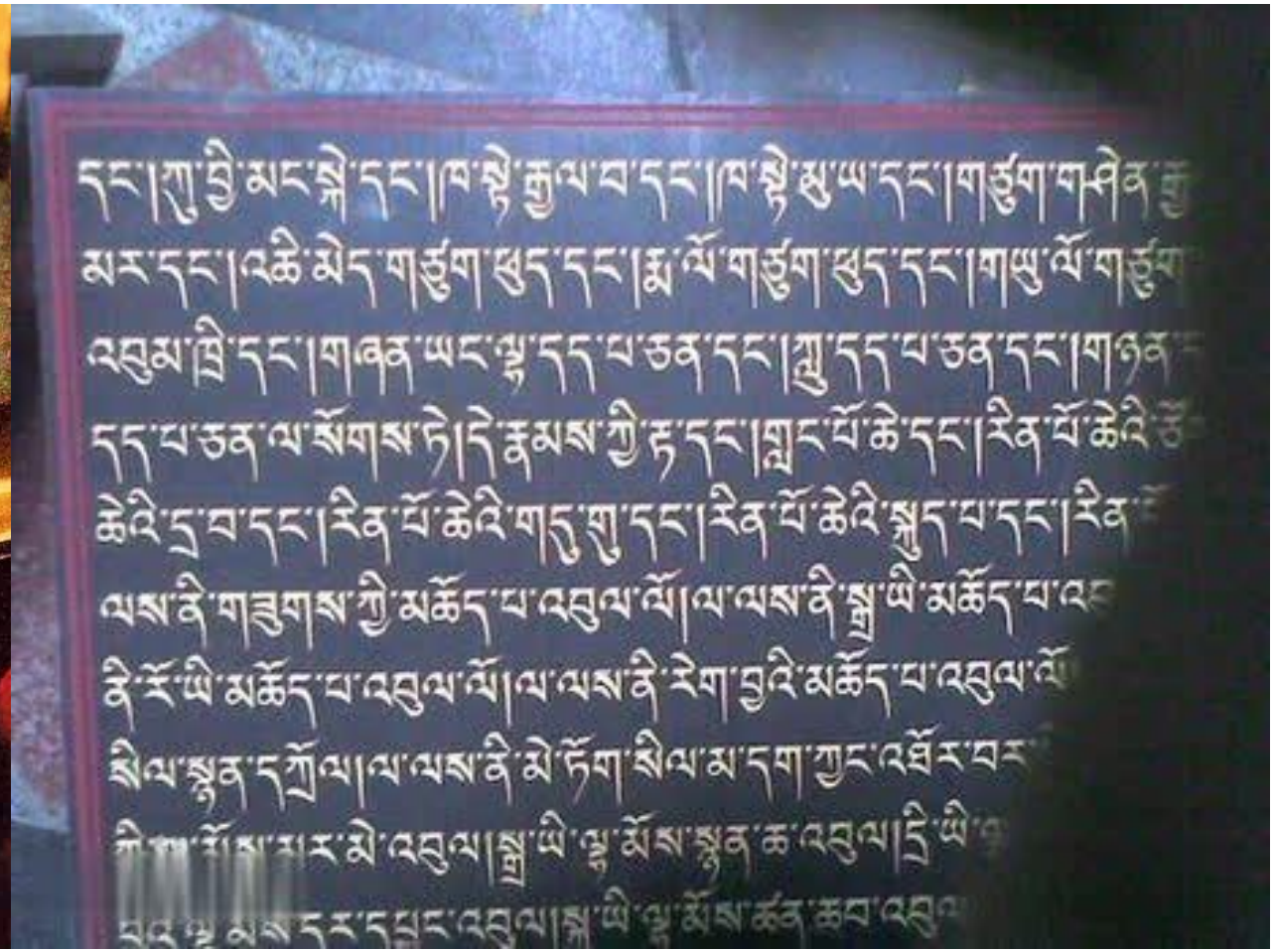
藏文的起源：

1. 由古象雄语言发展而来，说法源自雍仲苯教。象雄王国是吐蕃之前西藏的统治势力，有象雄语写的象雄大藏经流传于世，包罗万象，其汉译工程被中国社科院列为重点课题。

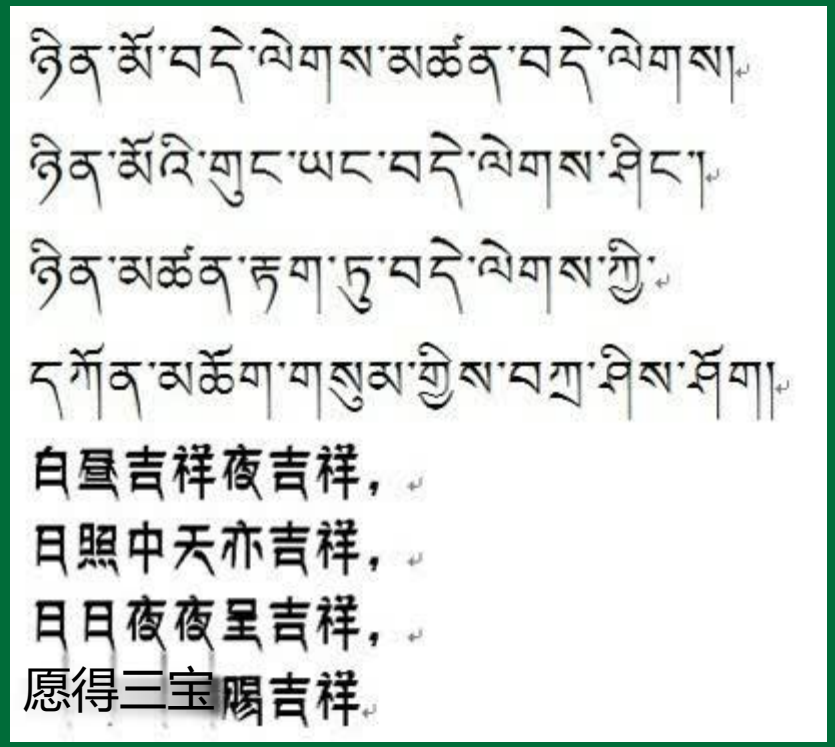
2.源自古印度语，文字诞生于7世纪松赞干布王时期，吞弥·桑布扎去印度学习后创制的。中古藏语语音特点加以改造，去掉不需要的音素字母，增添几个新的音位字母，成功地“创制了一套基本上反映当时藏语语音面貌的拼音文字

藏文历史





藏文历史



藏文的三次厘定：

首次厘定：从8世纪中叶墀松德赞至9世纪初叶墀德松赞时期，是藏文的首次厘定规范时期。这个时期诞生了规范译语的翻译工具辞书《梵藏词典》。

二次厘定：9世纪中叶集藏、印著名译师，专设译场，统一译名，规定译例，校订旧译经典，新译显密经典，进一步对藏文进行规范。

三次厘定：仁青桑布同入藏的印度班智达善护、德护、智护一起，共同修订文字，厘定新译语。

藏语区分布

在中国西藏自治区和青海、四川甘孜藏族自治州、阿坝藏族羌族自治州以及甘肃甘南藏族自治州与云南迪庆藏族自治州5个地区，不丹、印度、尼泊尔、巴基斯坦四个国家的部分说藏语。

藏语主要分为三大方言：卫藏方言、康巴方言、安多方言。通用的文字标注的是古藏语音。

藏语语系

藏语属汉藏语系。虽文字相差较大，但是读音实则很贴近于古汉语

家に帰ってご飯を食べるの？

집에 가서 밥을 먹을까요?

về nhà ăn cơm chưa?

ညစာအတွက်အိမ်ပန်ပါ?

ནང་ལ་ཐོག་ནས་ཁ་ལག་ཟ་གི་ཡིན་བས།

	1	2	3	4	5	6	7	8
日	hi	fu	mi	yo	i	mu	nana	ya
韩	hana	tul	se	ne	taseo	yeoseo	ilkob	yeodeol
越	một	hai	ba	bốn	năm	sáu	bày	tám
缅	tit	ni	thoun	lei	nga	chao	kunni	shit
藏	chik	nyi	sum	shi	nga	druk	dyun	gye

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

	汉语	藏语
语法结构	主-谓-宾：我是学生	主-宾-谓： ང (我) ཤེས་རབ་ (学生) ཡིན་ (是)
量词	丰富	贫瘠：除了少数 རྩ(根)、ཆ(双)、ཚན་ད(束把)等，其余皆用数词连接名词
定语	前置：美丽的花园	后置：花园美丽的
形容词	使用副词表示程度： 好，更好，最好	需要变型： ཡག་པོ (好) ཤིན་ཡག་ (较好) ཡག་རྩེས་ (最好)
动词时态	无时态：用“将”，“着”，“了”，“过”“等副词表示动作状态	需要变位： ངས་ཡག་བཟའ་གི་ཡོད་ (我在吃饭。) ངས་ཡག་བཟའ་གི་ཡིན་ (我将要吃饭。)

藏文字

现代藏文共有30个辅音字母和4个元音字母，同时使用5个反写字母和5个并写字母，一个长元音字符等。

藏文字形结构以辅音字母为核心，其余字母以此为基础前后和上写拼写。

藏文20个辅音字母和5个反写字母均可作基字

藏语词法

格助词

指处于格位范畴的虚词。通过一定的语法形式，附着于名词，代词，名词性短语的后面。格助词分为八类：主格、业格、具格、为格、从格、属格、与格和呼格。

不自由虚词

指根据添接法是否受前一音节后加字的的限制。

不自由虚词包括四种集饰连词，持续连词，离合连词，终结连词。

自由虚词

自由虚词不受前一音节后加字的影响，共有六种用法：

陈述词，连词，指代词，疑问词，否定词，指人后缀。

动词

藏语的语法更接近屈折语分为自主动词和不自主动词、及物不及物，使动与自动。同时，针对不同的时态：现在时，过去时，将来时 以及 命令式动词会进行变位。

藏语是典型的动居句尾型语言，其语序常态是“施事（S）+受事（O）+动作（V）”

NP+PP+VP句式

藏语的语法主要是通过虚词语法手段来表现的，依靠虚词可以进行句法分析，了解语句含义。
名词性短语（NP）+格助词（PP）+动词性短语（VP）

NP+VP句式

实词和虚词是构成藏语的两个因素。实词是必须的，没有实词无法构成句子，但是多个实词的组合不一定有虚词

03

实现方法

METHODS

语料库

语料库极少，目前大部分研究人员还在用爬虫对中国西藏网，藏语版人人网进行数据收集，在人工生成语料库和词典。
目前现成的语料库普遍来自全国机器翻译研讨会（CWMT）
2014年大型藏文基础语料库的建设已经完成。

分析工具

TIP-LAS是上届CWMT汉藏分词标注比赛的第一名，速度和精准度都不错。
此外还有青海师范学院研发的分词工具，也有一部分科研工作者在使用。

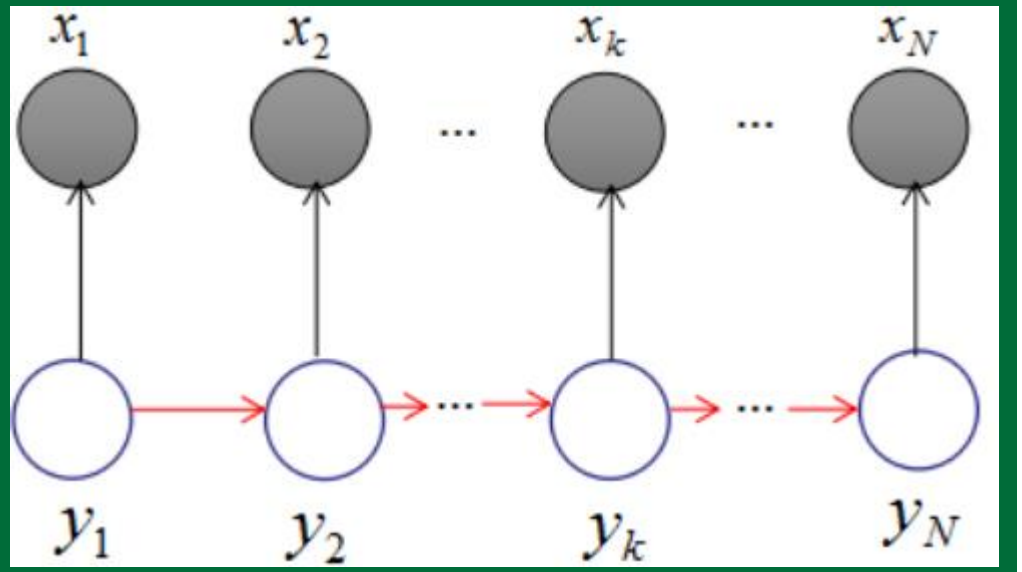
“

ＴＩＰ－ＬＡＳ集成藏文分词、词性标注功能，该系统由Ｃ＋＋实现，提供跨Linux，Windows平台功能，分为藏文分词系统，词性标注系统两大模块。藏文分词系统基于条件随机场模型，实现了基于音节标注的藏文分词方法，藏文词性标注系统基于最大熵模型，并融合了音节特征。该系统的准确度和速度已经基本满足实际应用要求。

TIP-LAS 基于条件随机场的藏文分词方法

条件随机场（Conditional Random Field，CRF）是Lafferty等提出的一种统计的序列标记模型。

在TIP-LAS分词系统中，把藏文分词看成是序列标记问题。在序列标记问题中生成一个基于无向图G=（V，E）的一阶线性链式CRF



TIP-LAS 基于条件随机场的藏文分词方法

V 是随机变量 Y 的集合Y = {Y i | 1 ≤ i ≤ n}，对于输入一个句子的 n 个需要标记单元，E = {(Y i - 1，Y i) | 1 ≤ i ≤ n} 是 n - 1 个边构成的线性链。对于每个句子 x，其对应的标记序列y的条件概率为：

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,k} \lambda'_k f'_k(y_i, x) \right)$$

其中两个求和式子分别为对于每个句子x，对于每个边和每个节点定义两个非负因子，其中f_k是一个二值特征函数，K 和 K'是定义在每个边和相应节点的特征数量。Z(x)是归一化函数。给定训练集D，训练模型的参数是用来最大化条件似然值。当给定了要标记的序列 x，其对应的标记序列 y 由参数

Argmax P(y | x)输出。

TIP-LAS 基于条件随机场的藏文分词方法

表 2　音节标记示例

音节数	藏语词汇	标记示例
1	ང་(我,nga)	ང་/S
2	སློབ་མ(学生,slob ma)	སློབ་/B མ/E
3	གསར་འགོད་པ(记者,gsar vgod pa)	གསར་/B འགོད་/M པ/E
4	རྒྱུན་ལས་ཀྲུའུ་ཞི(常务主席,rgyun las kruvu zhi)	རྒྱུན་/B ལས་/M ཀྲུའུ་/M ཞི/E

基于字标注的分词方法中，需要对每一个字在词中的位置信息进行标注，选用“B M E S”标记集，根据每个藏文音节在词中出现的位置，给予不同的标签，B 代表词的左边界，E 代表词的右边界，M 代表词的中间部分，S 代表单音节词，超过三音节的词中间部分都标记为 M。在分词中，把输入的原始藏文文本切分成音节序列，音节序列包含藏文音节，英文，汉语标点符号等，采用 C R F 模型对音节进行位置标注，根据标注结果还原出分词结果

TIP-LAS 基于最大熵的藏文词性标注方法

最大熵原理的基本思想是，首先利用给定的训练样本，选择一个与训练样本一致的概率分布，它必须要满足所有已知的事实。在没有更多的约束和假设的情况下，对于那些不确定的部分，则会赋予均匀的概率分布。熵是用来表示随机变量的不确定性，不确定性越大，熵越大，分布越均匀

$$P^* = \arg\max_{p \in C} H(P)$$

$$P^*(y \mid x) = \frac{1}{Z(x)} exp\left(\sum_i \lambda_i f_i(x,y)\right)$$

$$Z(x) = \sum_y exp\left(\sum_i \lambda_i f_i(x,y)\right)$$

最大熵模型需要需要构建特征做为判断依据

上下文特征

藏语中一个词的词性很大程度上由其上下文的环境决定，因此当前词的前后 n 个词可以作为判断当前词词性的依据。

特征	说明
$C_n(n=-2,-1,0,1,2)$	当前词的前后第 n 个词
$C_nC_{n+1}(n=-2,-1,0,1)$	连续的两个词
$C_{-1}C_1$	当前词前、后的两个词

词内部特征

藏文动词的现在、将来、过去时和命令式是通过词缀及附加词缀来表现的。将当前词的词首音节、词尾音节，前、后词，前驱词的词尾音节、后继词的词首音节等特征结合在一起作为特征。

特征	说明
$w_0(\text{prefix}(w))$，$w_0(\text{suffix}(w))$	当前词的首、尾音节
$w_{-1}(\text{suffix}(w))$	前驱词的词尾音节
$w_1(\text{prefix}(w_1))$	后继词的词首音节

04

科技前沿

CUTTING-EDGE TECHNOLOGY

藏汉神经机器学习结合注意力机制

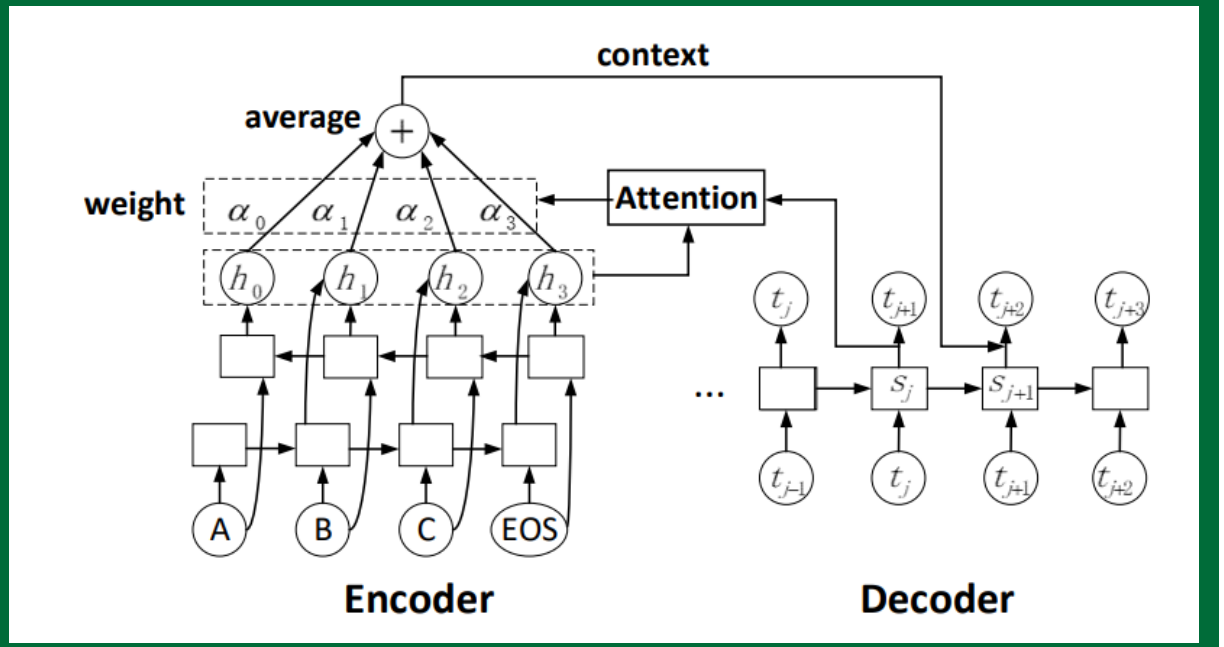
使用短路径单元的线性插值改进神经机器翻译

寻找更好的藏语神经机器翻译子词

基于句法树的藏文最大名词短语识别

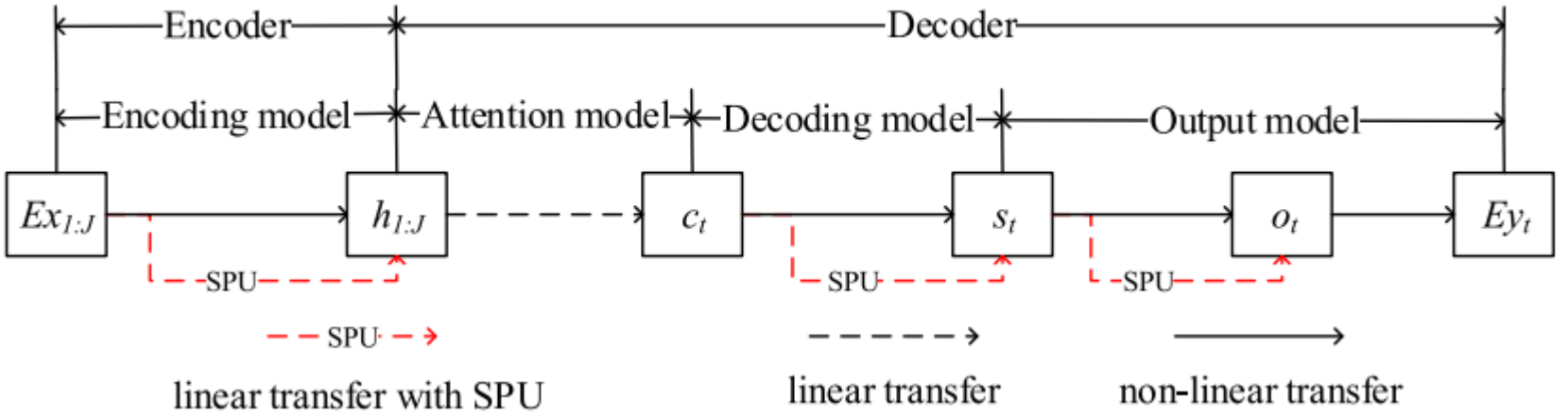
藏汉神经机器翻译结合注意机制

将注意力机制和神经机器翻译相融合，实现了基于注意力机制的藏汉神经机器翻译系统。 通过实验对比分析，验证了注意力机制可以有效提高机器翻译的效果



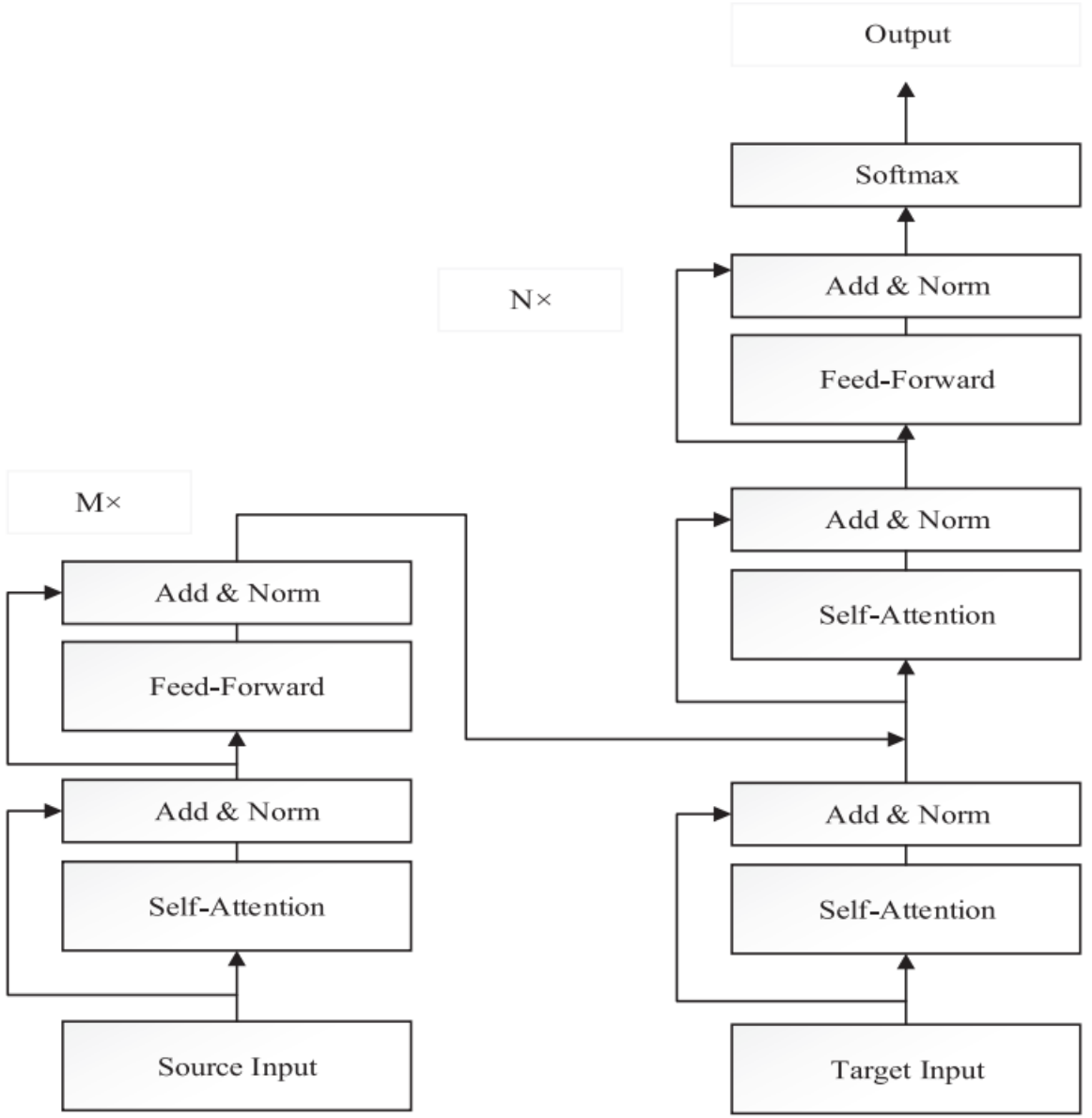
使用短路径单元(SPU)的线性插值改进神经机器翻译

提出了一种方法来加强源词与其相应翻译的关联。 这是通过在编码模型、解码模型和输出模型中引入短路径来实现的。 使用这些 SPU，可以在 DNN 中选择性地逐层线性传输信息。 对藏汉翻译任务的实证研究表明，与常规神经机器翻译系统相比，所提出的方法实现了更好的翻译性能和对齐质量



寻找更好的藏语神经机器翻译子词

针对藏语词结构，该方法提出两种藏语子词分词方法，即基于音节的分词和基于字符的分词。此外，我们研究了不同的子词分割方法具有 Transformer 架构的低资源藏汉神经机器翻译器。从实验结果可以看出，使用子词作为翻译的基本单位可以显着提高藏汉神经翻译的翻译质量。此外，在相同条件下，藏到汉带有子词的神经机器翻译实现了更好的翻译性能。



基于句法树的藏文最长名词短语识别

最长名词短语是指中心词为名词的所有短语，位置可以居于短语的首、中、尾；可以由单个名词、代词、数词等构成。

最长名词短语识别是对一个语言分析的挑战。在这个实验中句法分析方法的结果是弱于序列标注的方法。

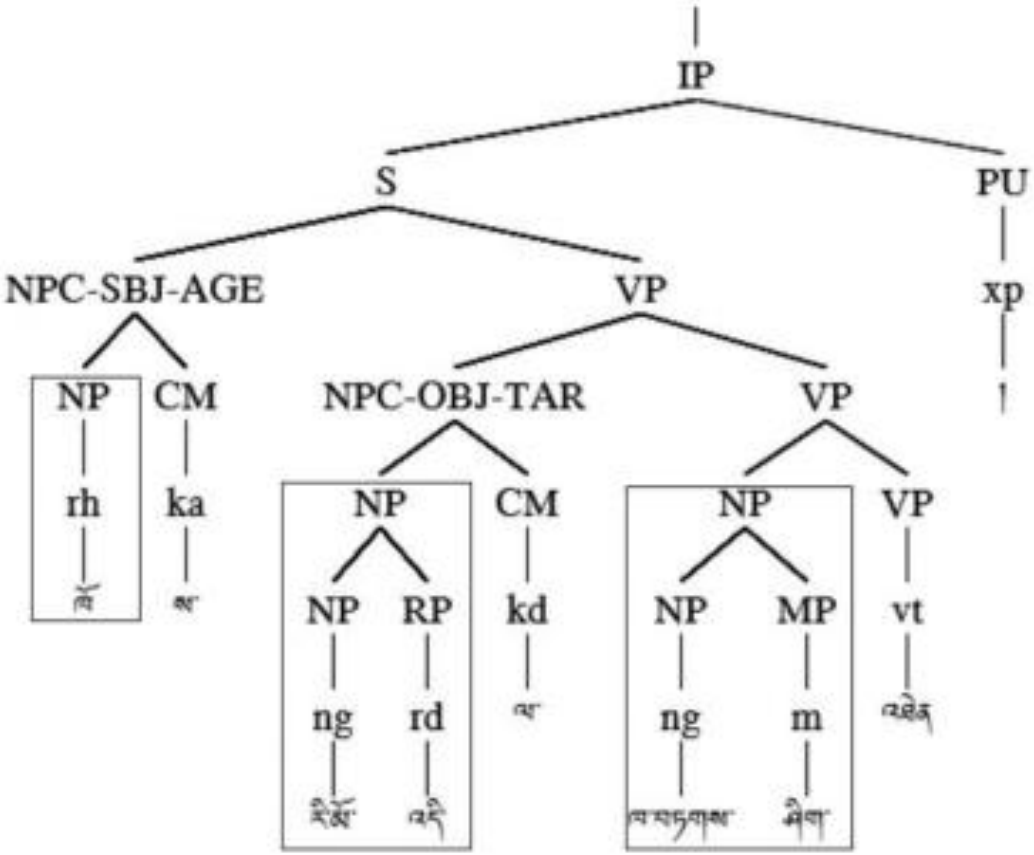


Fig. 2. TMNP in syntax tree.

05

Demo 演示

DEMO

使用TIP-LAS进行分词

由于高质量的藏语语料库的获取有较高难度，本实验直接在TIP-LAS导入训练好的模型，进行分词。

为了更好的分析，将分词出来的结果进行翻译

测试的文本来源是西藏网。

使用TIP-LAS进行分词

藏文文本：ཨང་ཆུན་ཀྱང་གི་ཏ་ཀང་དབུ་བརྒྱེས་ནས་ལོ་ངོ་བརྒྱ་འཁོར་བའི་འཁྲུངས་སྐར་རྗེས་སུ་སྐྱེ་ཚབ་ཐུའུ་ཅི་ཞེ་ཅན་ཡིན་སྐུ་ངོ་མ་བོད་དུ་གཟིགས་ཞིབ་གནང་བར་ཕེབས་ནས་མི་རིགས་ཁག་གི་ལས་ཀྱེད་པ་དང་མང་ཚོགས་ལ་འཚམས་འདྲི་གནང་བ་མ་ཟད། བོད་ཞི་བས་བཅིངས་འགྲེལ་བཏང་ནས་ལོ་ངོ་70འཁོར་བར་རྗེན་འཛིལ་ཞུས་པ་དང་འཛིལ་གསུང་བཤད་གལ་ཆེན་དང་མཇུག་སྡོམ་གལ་ཆེན་ཡང་གནང་ནས་བོད་ཀྱི་ལས་དོན་ལ་ཁས་ལེན་གང་ལེགས་དང་ཆང་ཀ་གསལ་བཀོད་གནང་ཡོད།

例句：ཨང་ཆུན་ཀྱང་གི་ཏ་ཀང་/nh དབུ་བརྒྱེས་/iv ནས་/c ལོ་ངོ་/ng བརྒྱ་/m འཁོར་/vi བ/h འི་/kg འཁྲུངས་སྐར་/ng རྗེས་/ng སུ་/kl སྐྱེ་ཚབ་/ng ཐུའུ་/ng ཅི་/ng ཞེ་ཅན་གིང་/nh སྐུ་/ng ངོ་མ་/a བོད་/ns དུ་/kl གཟིགས་ཞིབ་/ng གནང་/vt བ/h

ཨང་ཆུན་ཀྱང་གི་ཏ་ཀང་/nh 成立/iv ，/c 一/ng 百/m 转动/vi 。/h 的/kg 圣诞/ng 后，/ng 在/kl 总/ng 书记/ng ཞེ་ཅན་གིང་ /nh 身/ng 真/a 西藏/ns 在/kl 审阅/ng 了/vt 。/h

译文：ཨང་ཆུན་ཀྱང་གི་ཏ་ཀང་建国一百周年的生日后总书记ཞེ་ཅན་གིང་亲自到西藏去考察，各民族的干部和群众看望慰问，西藏和平解放60周年七十周年表示祝贺并作重要讲话和重要指导了西藏工作的充分肯定，提出明确要求。

德以明理　学以精工