藏 语 N L P

姓名：马子豪 王冰钰 刘欣阳 郑心茗 杜思楠

目录

CONTENT

01

前言

王冰钰

NLP是什么

自然语言处理，英文Natural Language Processing，简写NLP。

自然语言处理是指机器理解并解释人类写作、说话方式的能力，NLP的目标是让计算机/机器在理解语言上像人类一样智能。

最终目标是弥补人类交流和计算机理解之间的差距。

NLP这个概念本身过于庞大，可以把它分成“自然语言”和“处理”两部分。

1）自然语言是人类发展过程中形成的一种信息交流的方式。

2）处理则必须是计算机处理的，计算机无法像人一样处理文本，它有自己的处理方式。

藏语NLP背景

2005年开始藏文词性的标祝研究。

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。

1999年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。

2011年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。

为什么需要 藏语NLP ？

- 藏语nlp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。
- 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。
- 有可能完成自动语音、自动文本编写这样的任务，比起人工花费的时间更少。
- 通过使用网络抓取和自然语言处理缩短识别研究差距所需的时间， 自然语言处理标记化从最高频率到最低频率对搭配进行排序，提高效率和更完善的资料。

藏语NLP的作用

01

藏语是促进西藏文化和民族文化的主要工具和手段。

02

藏语NLP可以加速双向的信息传播。对藏区的经济发展和对藏族文化了解有着巨大的促进。

03

随着信息化的高速发展，藏语的语言模型成为目前乃至以后的研究趋势。

「02」

藏语相关

马子豪

藏　文　历　史

- 藏语（藏文：བོད་སྐད，藏语拼音：poigê），属汉藏语系藏缅语族藏语支，是以藏族为主的喜马拉雅文化圈使用的主要语言。

- 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。

- 在迄今的1300年里，藏文经历了四次改革。此四次改革分别发生在公元7世纪、公元8世纪、公元8世纪和公元1070年。四次改革先后整理规范了藏文字和语法、统一了用词用语、确立藏文字的书写法。

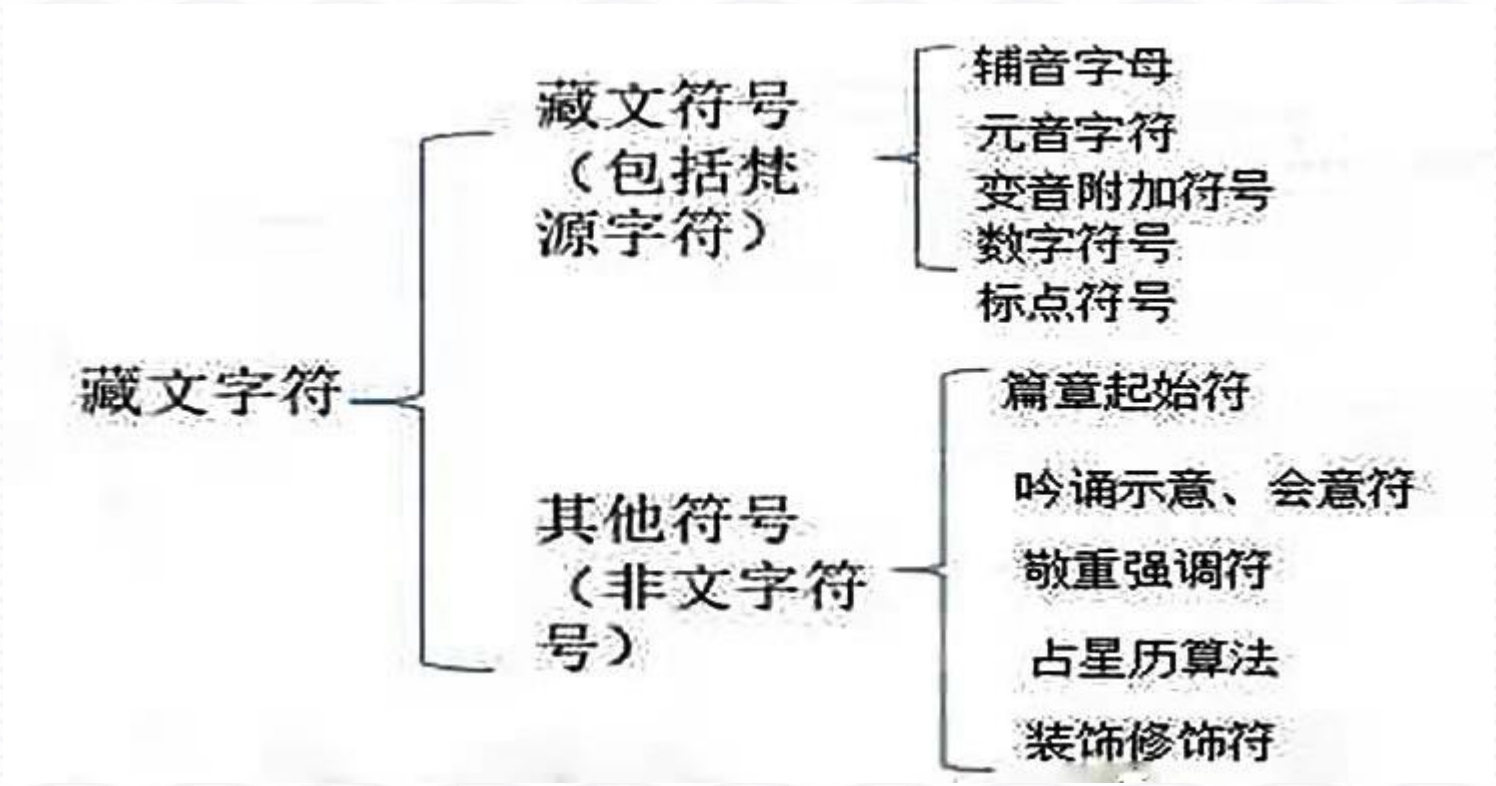
- 西藏和平解放后，为促进藏语的学习、使用和发展，国家和政府历时近20年的研究，1988年西藏正式成立了西藏自治区藏语文工作指导委员会，由党委、政府主要领导人兼任领导。各地市均成立了藏语文工作指导委员会。与此同时，藏文编码国际标准于1997年获国际标准组织的通过，成为中国少数民族文字中第一个具有国际标准的文字。

藏 语 词 性

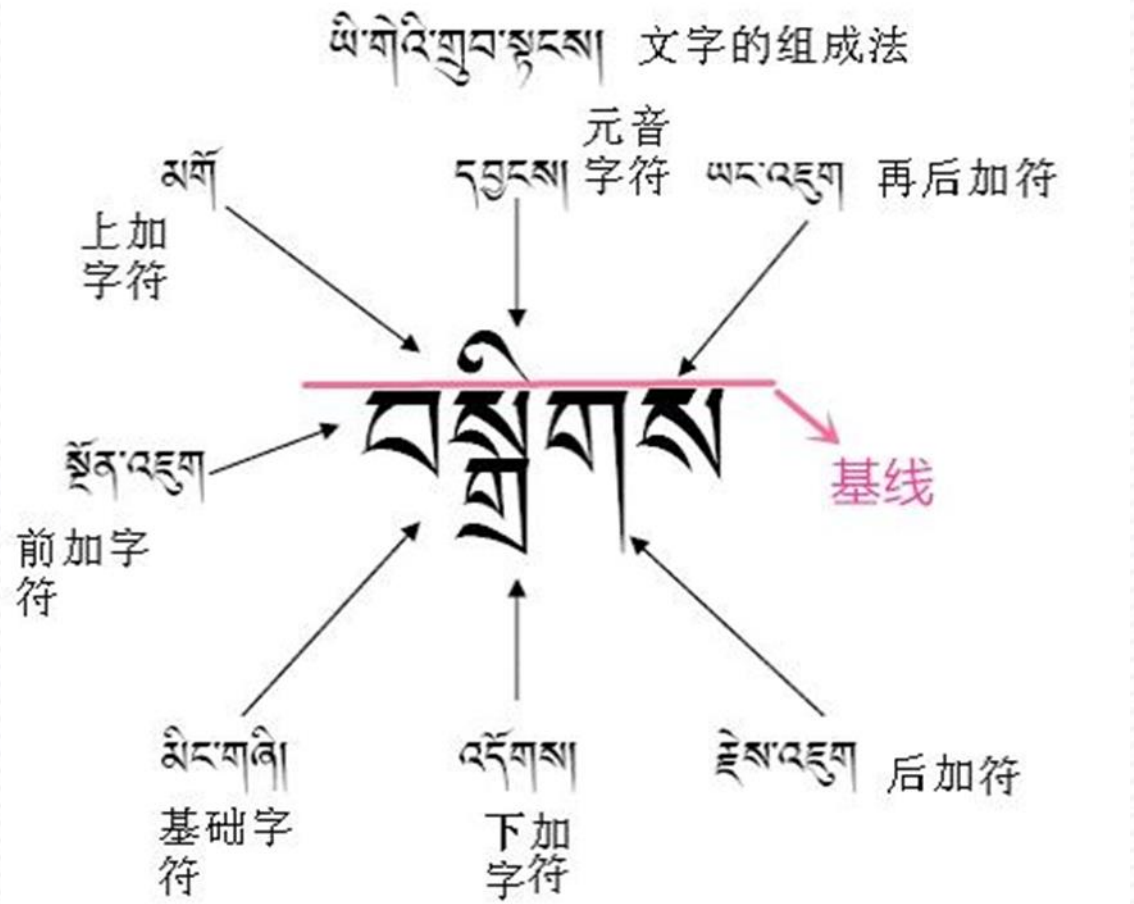
- 藏语词型态相当丰富，给藏语nlp的研究带来很大困难。
- 其动词分四个时态，而且时态的系统呈现很多例外，意即藏文具有屈折变化。

- 吃
  - 现在时 ཟ་ (za)
  - 过去时 ཟོས་ (zos)
  - 未来时 བཟའ་ (bza')
  - 命令时 ཟོས་ (zos)
- 读
  - 现在时 སྐོག་ (klog)
  - 过去时 བསྐྱགས་ (bklags)
  - 未来时 བསྐྱག་ (bklag)
  - 命令时 སྐོགས་ (klogs)

藏文是拼音文字，有着自己固定的语法，它们在发音、构句、书写等方面与汉字截然不同。

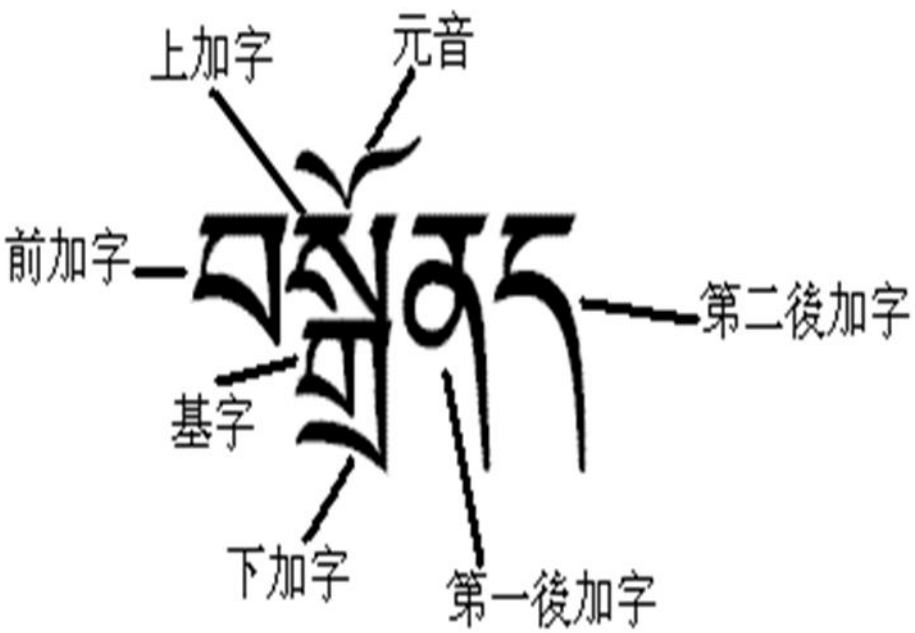


下图中标识出的红色水平线被称为**基线**，书写字母时要在基线处对齐，基线下方长短不一，基线上方是元音字符。在做藏文识别时，基线非常重要，一旦将基线识别准确，便可判断出位于基线上方的肯定是元音符号。

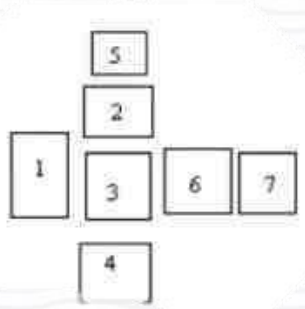


如图所示，藏文的字，左右宽最多为四个字丁，上下最多为四层这种非线性的立体结构，为藏文信息处理中造成了不小的麻烦。
在书写时，整体规则是从左到右，从上到下。
上图中文字的书写顺序为：

1. 前加字符
2. 上加字符
3. 中间的基础字符（基字）
4. 下加字符
5. 元音字符
6. 后加字符
7. 再后加字符



- 书写顺序



藏文中共有5个前加字符、
3个上加字符、
4个下加字符
和10个后加字符。

藏文有30个表示辅音的字母和4个表示元音的符号。按照传统藏文拼读法，辅音和元音拼合时，是先读辅音字母的名称，再读元音符号的名称，然后读拼出的音节音。即：辅音 + 元音 → 音节音。

如有后加字，还要在基字和元音拼出的音节上再和后加字拼读。即：基字 + 元音 → 音节音 + 后加字 → 音节音。

句子结构

S（主）+O（宾）+V（谓）

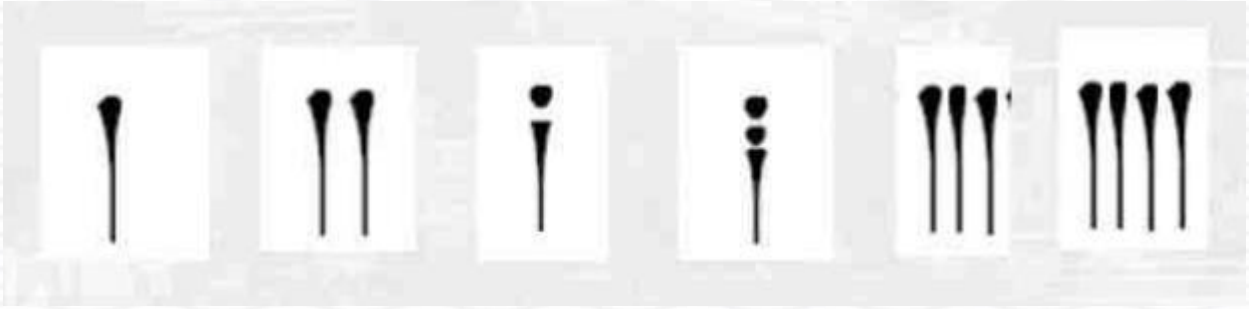
我-格　藏文　　学习　时体

主-施　宾　　谓语

音节点是音节之间重要分隔标记。

垂符是藏语的词，短语，句子，段落的结束符，
相当于逗号，句号，分号等。

音节黏写也是机器翻译中的难点。



དེ་རིང་ཚེས་པ་གཙོད་རེད་དམ།

今天是几号了?

དེ་རིང་ཚེས་པ་གཙོད་རེད་དམ།

今天是几号了?

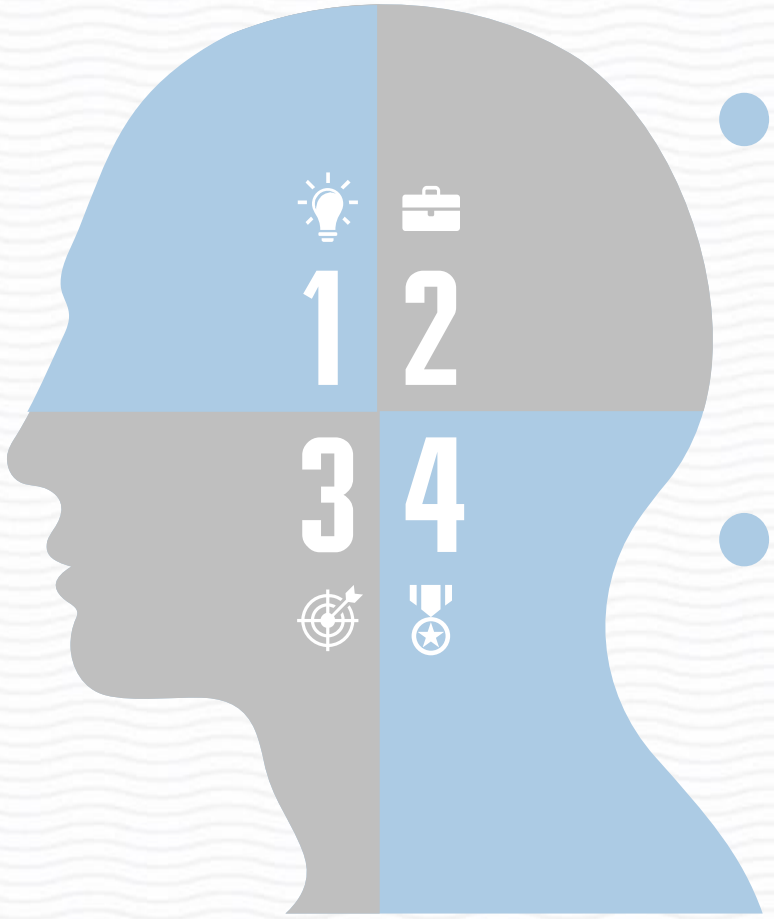
03

藏文NLP理论技术

刘欣阳

Sentencepiece 工具及其算法

Transformer 模型



相关优化器

相关评价指标

Sentencepiece 工具及其算法

2018 年 8 月 Google开源了 Sentencepiece 子词划分工具，与 subword-nm[33] 和 wordpiece[34]等子词划分工具无法指定词表的大小不同，Sentencepiece 可以指定词表的大小，如常用的：8k、16k 和 32k。

N 元语言模型（N-gram）是自然语言处理中的一个重要的统计语言模型，该模型基于一定的语料，假设一个句子有 m 个单词，单词wi的出现概率依赖从第一个单词$w1$到它之前一个单词$wi-1$组成的序列，基于该假设计算句中单词的联合概率$P(w1, w2, \cdots, wm)$，其计算公式如下：

$$P(w_1, w_2, \cdots, w_m) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\cdots P(w_m|w_1, \cdots, w_{m-1})$$

为了解决计算难度过大的问题，需要引入马尔科夫假设，即：

$$P(w_1, w_2, \cdots, w_m) \approx \prod_{i=1}^{m} P(w_i|w_{i-n+1}, \cdots, w_{i-1})$$

其中$P(wi|wi-n+1, \cdots, wi-1)$采用极大似然估计进行计算

当 n=1 时为 unigram，n=2 时为 bigram，n=3 时为 trigram。
Sentencepiece 中所使用的就是 unigram，其计算公式如下所示，其中 M 表示训练语料中的总的字符数。

$$P(w_1, w_2, \cdots, w_m) \approx \prod_{i=1}^{m} P(w_i)$$

$$P(w_i) = \frac{C(w_i)}{M}$$

具体流程：

算法 1　一元语言模型分词算法

输入：训练语料

1：　定义生成词表大小

2：　根据训练语料设置 seed 词表，一般为语料中的所有字符和最高的子字符串

3：　repeat

4：　　修改词表，使用最大期望算法（Expectation-maximization algorithm，EM）优化 P(x)

5：　　计算每个子词x_i的$loss_i$，$loss_i$表示x_i被移除词表后，似然函数$\mathcal{L}$减小的可能性，$\mathcal{L}$计算公式如下，其中 S 表示对输入句子 X 的候选分割

$$\mathcal{L} = \sum_{s=1}^{|D|} \log P(X^{(s)})$$

6：　　根据$loss_i$排序，保留前列部分子词如 80%

7：　until 达到预先设定的词表大小

Transform模型

Transformer模型的提出是为了解决循环神经网络（Recurrent Neural Network，RNN）无法并行计算的问题以及卷积神经网络（Convolutional Neural Networks，CNN）提取长距离依赖信息弱的问题。该模型结构开启了自然语言处理的新篇章，近年来基于 Transformer 的预训练模型层出不穷，本文所使用的预训练模型、文本分类模型、摘要提取模型其核心均为 Transformer 模型。Transformer 模型采用Encoder-Decoder 模型框架，并在 Encoder 部分和 Decoder 部分都使用了自注意力机制，该机制赋予了 Transformer 模型更好的特征提取能力以及并行计算能力。

相关优化器

Adam 优化器算法

输入：训练数据 X；学习率η，迭代次数 T，用于数值稳定的参数δ，矩估计的指数衰减率$\beta_1 = 0.9$，$\beta_2 = 0.999$

1：初始化一阶和二阶变量$s = 0$，$r = 0$

2：初始化时间步$t = 0$

3：当$t < T$时：

4：　获取一个批次大小的数据x_t

5：　计算梯度$g = \nabla L(w_{t-1}^l, x_t)$

6：　更新时间步$t = t + 1$

7：　更新有偏一阶矩估计：$s = \beta_1 s + (1 - \beta_1) g$

8：　更新有偏一阶矩估计：$r = \beta_2 r + (1 - \beta_2) g \odot g$

9：　修正一阶矩的偏差：$\hat{s} = \frac{s}{1-\beta_1^t}$

10：　修正二阶矩的偏差：$\hat{r} = \frac{r}{1-\beta_2^t}$

11：　计算更新：$\triangle w = -\eta \frac{\hat{s}}{\sqrt{\hat{r}}+\delta}$

12：　应用更新：$w = w + \triangle w$

13：结束循环。

输出：更新后的权重矩阵 w

AdamW 是对 Adam 的一个修正，在随机梯度下降中 L2 正则化等效于权重衰减，但在自适应学习率优化算法中并不成立，AdamW 将权重衰减和损失函数解耦，提升了 Adam 的泛化性能。

相关优化器

LAMB 优化器算法

输入：训练数据 X；迭代次数 T；隐藏层层数 L；学习率η

默认超参数设置：$\lambda = 0.01$，$\beta_1 = 0.9$，$\beta_2 = 0.999$，$\epsilon = 1$

初始化：t=0，权重w

当$t < T$时：

  获取一个批次大小的数据x_t

  $l = 0$

  当$l < L$时：

    $g_t^l = \nabla L(w_{t-1}^l, x_t)$

    $m_t^l = \beta_1 m_{t-1}^l + (1-\beta_1) g_t^l$

    $v_t^l = \beta_2 v_{t-1}^l + (1-\beta_2) g_t^l \odot g_t^l$

    $\hat{m}_t^l = m_t^l / (1-\beta_1^t)$

    $\hat{v}_t^l = v_t^l / (1-\beta_2^t)$

    $r_1 = \|w_{t-1}^l\|_2$，$r_2 = \left\| \frac{\hat{m}_t^l}{\sqrt{\hat{v}_t^l + \epsilon}} + \lambda w_{t-1}^l \right\|_2$

    $r = r_1 / r_2$

    $\eta^l = r \times \eta$

    $w_t^l = w_{t-1}^l - \eta^l \times \left(\frac{\hat{m}_t^l}{\sqrt{\hat{v}_t^l + \epsilon}} + \lambda w_{t-1}^l \right)$

  停止循环

停止循环

输出：训练的模型权重w

Google Brain 使用 LAMB 优化器，在 1024 块 TPU 上使用65536/32768
混合批次大小，只需要迭代 8599 步就可以训练完成，训练时间仅有 76.2 分钟，最终在 SQuAD-v1 任务中超过了使用 AdamW 优化器的 BERT 预训练模型。

文本分类评价指标

二分类混淆矩阵

相关评价指标

混淆矩阵		预测值	
		正类	负类
实际值	正类	TP	FN
	负类	FP	TN

精确率：表示 TP 和 TN 占全部样本的比率，是对分类结果一个最为直观的评价指标，但在样本不平衡的情况下并不能作为一个很好的评价指标

$$acc = \frac{TP + TN}{TP + TN + FP + FN}$$

查准率（Precision），也称准确率，表示的是实际为正类的样本在预测为正类的样本中的占比，代表着模型对正样本的准确程度

$$P = \frac{TP}{TP + FP}$$

查全率（Recall），也称召回率，表示预测为正类在实际为正类的样本中的占比，代表着模型对于分类结果的完备性。

$$R = \frac{TP}{TP + FN}$$

F1 分数又被称为平衡 F 分数（ balanced F score），它是$F\beta$当β = 1的时候的一般形式，$F\beta$于 1979 来被 Rijsbergen 提出，主要为了平衡查全率和查准率来更好的评价模型性能，β是需要调整的参数，用于调节查全率和查准率的比重。

$$F_\beta = \frac{(\beta^2 + 1) * P * R}{\beta^2 * P + R}$$

$$F_1 = \frac{2 * P * R}{P + R}$$

宏平均是首先计算每个类别的查全率和查准率，然后对查全率和查准率取算术平均得到宏平均查全率和宏平均查准率，最后根据宏平均查全率和宏平均查准率计算宏平均 F1，其计算公式如下所示，在不平衡样本数据集中，宏平均可以更好地评价模型对于小样本的分类能力。

$$MacroR = \frac{1}{p}\sum_{i=1}^{p} R_i$$

$$MacroP = \frac{1}{p}\sum_{i=1}^{p} P_i$$

$$MacroF_1 = \frac{2 * MacroP * MacroR}{MacroP + MacroR}$$

加权平均是根据每个类别在样本总的占比，对每个类别的查全率，查准率和进行加权，如公式 2-23 所示，其中 C 表示样本的总数，ci表示第 i 类类别的样本总数，加权平均是对模型整体性能的评价指标，是一种通用的指标。

$$WeightedR = \sum_{i=1}^{p} \frac{c_i}{C} R_i$$

$$WeightedP = \sum_{i=1}^{p} \frac{c_i}{C} P_i$$

$$WeightedF_1 = \sum_{i=1}^{p} \frac{c_i}{C} F1_i$$

自动摘要评价指标

相关
评价指标

Rouge(Recall-Oriented Understudy for Gisting Evaluation)，是 2004 年被提出的用于评估自动文摘以及机器翻译的工具集，根据基本单元选定方法的不同，也有不同的 Rouge 计算方法。

ROUGE-N

$$= \frac{\sum_{S \in \{Ref\,remmaries;\ gram_n \in S} \sum \sum_{gram_n} \mathrm{Count}_{match}(\mathrm{gram}_n)}{\sum_{S \in \{Referencestummariesgram_n \in S} \mathrm{Count}(\mathrm{gram}_n)}$$

Rouge-L 使用的是最长公共子序列作为基本单位，其中 L 就是 LCS（longest common subsequence）的首字母，Rouge-L 可以自动匹配到最长子序列，不需要像 Rouge-N 预先设置 N-gram 的长度，其查全率，查准率和$F\beta$分数的计算公式如下所示，m 和 n 分别表示参考摘要和生成摘要的长度。

$$R_{lcs} = \frac{LCS(X,Y)}{m}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n}$$

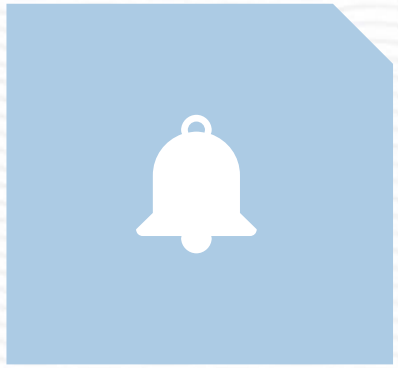
$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$$

「04」

技术难点与研究热点

郑心茗

技　术　难　点



内容的有效界定

日常生活中句子间的词汇通常是不会孤立存在的，需要将话语中的所有词语进行相互关联才能够表达出相应的含义,一旦形成特定的句子，词语间就会形成相应的界定关系。

消歧和模糊性

词语和句子在不同情况下的运用往往具备多个含义,很容易产生模糊的概念或者是不同的想法。

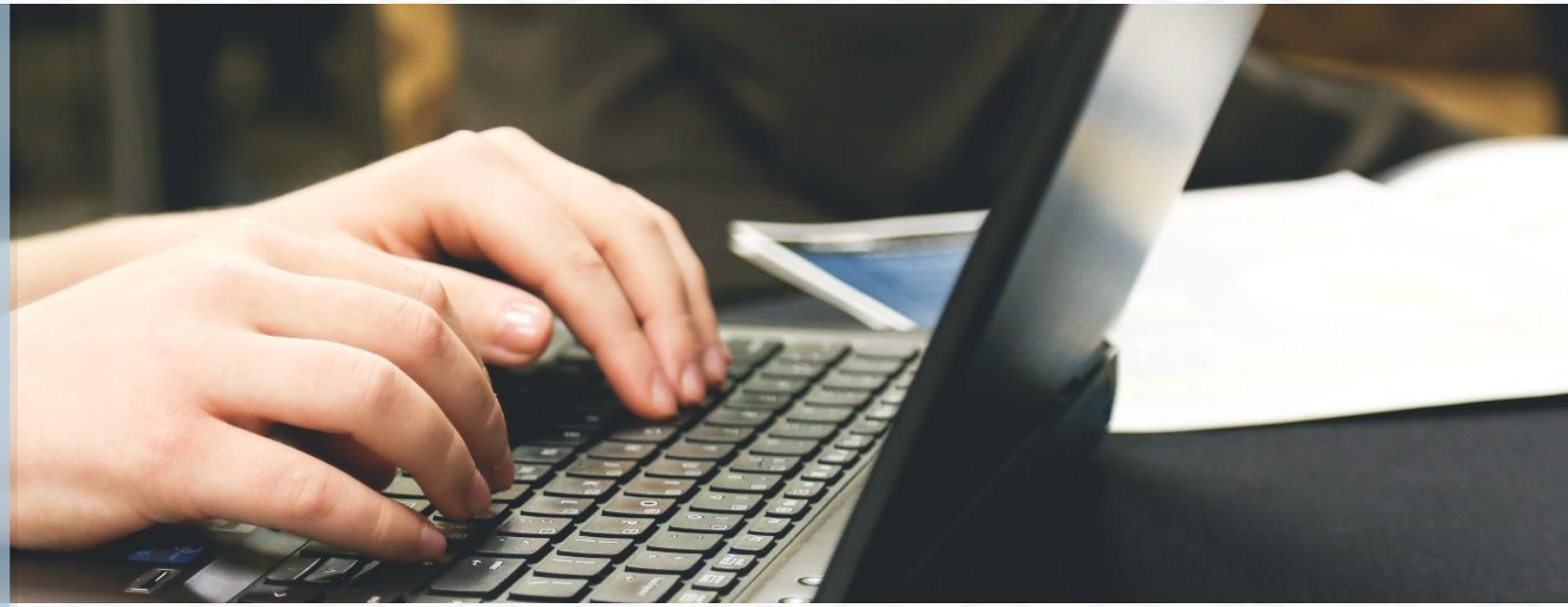
语言行为与计划

句子常常并不只是字面上的意思。

有瑕疵的或不规范的输入

例如语音处理时遇到外国口音或地方口音,或者在文本的处理中处理拼写,语法或者光学字符识别(OCR)的错误。

藏语
研究现状



藏语是一种少数民族语言，对藏语语言模型的研究目前还处于最基础的，最初级阶段，依然使用的是N-CRAM语言模型，是基于N元文法模型的研究，而对神经网络方面的研究是少之又少。

藏语研究现状

- 2011年，北方民族大学多拉和才让三智在研究中发现，建立藏语语言模型，重新探索新的藏语语法体系。

- 根据藏语特点，2012年李冠宇和孟猛，提出了一种藏语识别声学模型，并利用高级藏语知识来减少模式匹配的模糊性，在HTK平台上建立了依赖于上下文的连续隐马尔可夫声学模型，实现了西藏拉萨的连续词汇连续语音识别。最后发现，在最优情况下，模型词的错误率大大降低。

- 2014年，李照耀主要研究语言模型在藏文连续识别系统中的应用，结合西藏拉萨的特点，提出了一种新的文本筛选方案。通过比较各种算法的混淆和语音识别系统的识别率，将改进的Kneser-Ney平滑算法最终应用于基于HTK的藏文连续语音识别系统。实验结果表明，Kneser-Ney平滑算法的修改版本在各种平滑算法中具有最少的混淆。

藏语研究现状

- 2015年青海民族大学仁青吉和安见才让在对藏语语言模型的研究中，发现一个可靠的语言模型对比：在自然语言处理领域，例如语音识别，机器翻译和文本校对，起着至关重要的作用。通过在藏语语音识别系统中构建藏语模型来提高识别率，采用了一些算法来比较混淆。
- 2017年中央民族大学周楠主要探讨深度神经网络在藏语拉萨话连续语音识别任务中的应用，研究了深度神经网络的网络结构，预训练和参数设置，训练的深度神经网络的输出层特征用于训练HMM的声学模型。
- 2018年，黄晓辉、李京探索将循环神经网络和连接时序分类算法应用于藏语语音识别声学建模，实现端到端的模型训练。发现循环神经网络模型有更好的识别性能，拥有更高的训练和解码效率。
- 2019年，孙媛、王丽客和郭莉莉等人提出了一种优化词向量的GRU神经网络模型进行藏语实体关系抽取的方法，加入了优化的词向量，在传统的词向量模型中结合藏语音节向量、音节位置向量、词性向量等特征对词向量进一步优化，并且选取了藏语词汇特征和藏语句子特征。
- 以上这些是研究者不懈努力的成果。

探 索 与 展 望

- 许多研究证明神经网络语言模型的性能已超过传统的N-gram模型，但同时神经网络语言模型的构建需要大量的训练语料库。对于藏语语言模型的研究，在训练神经网络模型的数据资源严重匮乏的情况下，如何选择相对于目标任务的合适语言模型，如何找到适合藏语的训练语言模型方法？如何对藏语语言模型的预训练及后续的微调过程进行优化等这些问题将成为研究藏语语言模型的重要途径。

- 目前在还存在一些问题需要探索研究：
- （1）藏语数据集的不足与缺乏。互联网上公开共享的数据集非常缺少。
- （2）藏语语言模型的研究领域单一。语言模型数据集集中在新闻领域，在社交媒体、法律等急需建设。
- （3）评价标准不一致。研究员各抒己见，没有严格的标准，导致评价不一致，加大了研究中的难度。

任　重　道　远

- 藏学研究是非常广泛、影响深远且意义重大的研究领域，国内对于藏语的自然语言处理领域虽然具有较早的前瞻性，但实际起步较晚，且受到方言的影响、语料库资料的缺乏等客观原因，进展较为缓慢。

- 如今国内对于藏学领域研究大多集中于人文社科方面，对于科技领域尤其是信息技术领域的关注相对而言较少，相关机构也不多，所以，未来藏语研究领域的发展仍旧任重道远！

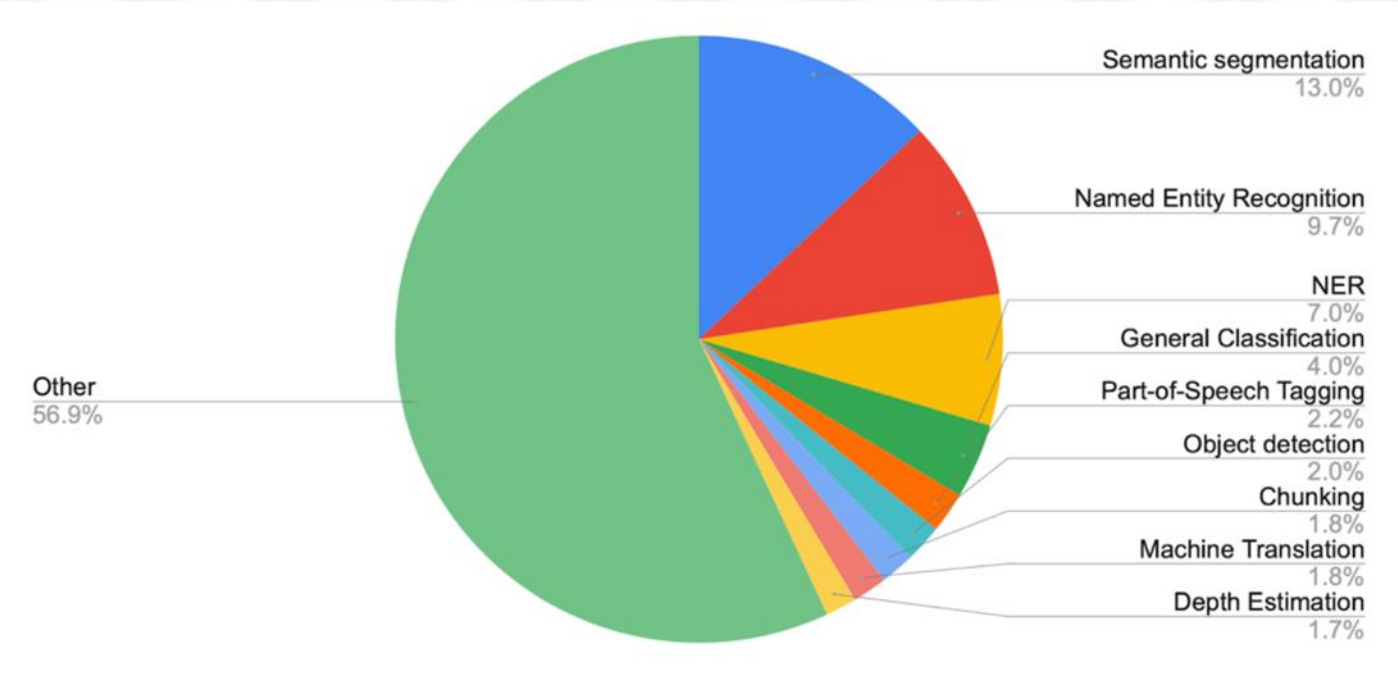
「05」

Demo介绍

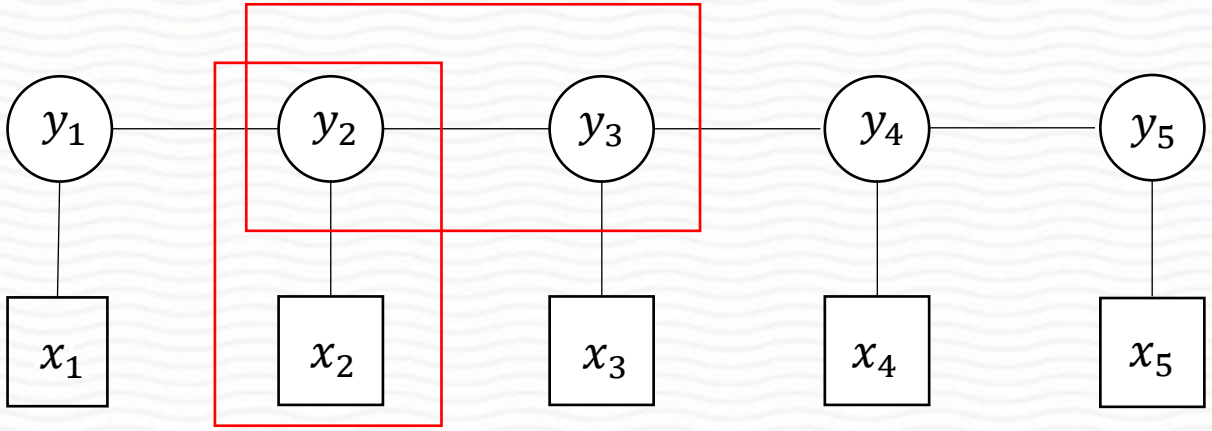
杜思楠

T I P - L A S

- TIP-LAS是一个开源的藏文分词词性标注系统,提供藏文分词、词性标注功能。该系统基于条件随机场模型实现基于音节标注的藏文分词系统,采用最大熵模型,并融合音节特征,实现藏文词性标注系统。



藏文分词-CRF



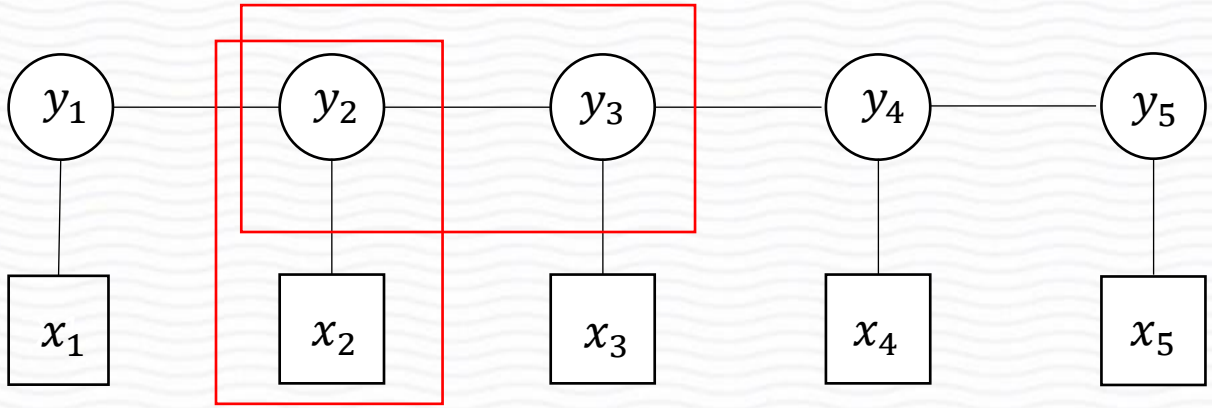
- (Hammersley-Clifford定理) 概率无向图模型的联合概率分布$P(Y)$可以表示为如下形式：

$$P(Y)=\frac{1}{Z}\prod_{C}\psi_C(Y_C) \qquad Z=\sum_{Y}\prod_{C}\psi_C(Y_C)$$

  - 其中，C是无向图的最大团，Y_C是C的节点对应的随机变量，$\psi_C(Y_C)$是C上定义的严格正函数，乘积是在所有最大团上进行的。Z是归一化因子，目的是保证构成一个概率分布。

[1]李亚超,江静,加羊吉,于洪志. TIP-LAS:一个开源的藏文分词词性标注系统[J]. 中文信息学报, 2015, 29(6): 203-207.

藏文分词-CRF



$$P(y|x) = \frac{1}{Z(x)} exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,l} u_l\, g_l(y_i, x)\right)$$

$$Z(X) = \sum_{y} exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,l} u_l\, g_l(y_i, x)\right)$$

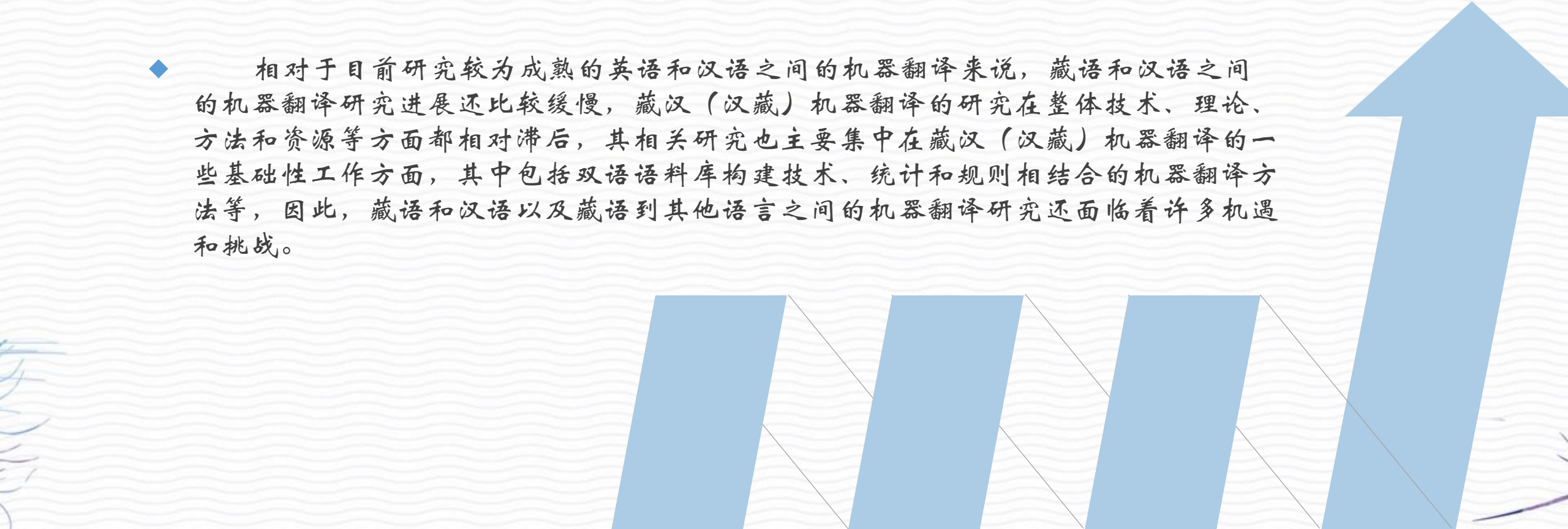
其中，f_k和g_l是二值的特征函数，λ_k和u_l是对应的权重。$Z(X)$是归一化因子，求和是在所有可能的输出序列上进行的。

- 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，在系统中选用“BMES”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。

机器翻译(Tibetan to Chinese)

- 由于世界上大部分语言之间的语言数据对照资源（如对照词典、平行语料等）非常匮乏，这对机器翻译理论、方法和技术的研究及应用带来了挑战，就目前机器翻译研究的发展趋势来看，这种语言资源稀缺条件下的机器翻译是目前重要的研究课题之一。

- 相对于目前研究较为成熟的英语和汉语之间的机器翻译来说，藏语和汉语之间的机器翻译研究进展还比较缓慢，藏汉（汉藏）机器翻译的研究在整体技术、理论、方法和资源等方面都相对滞后，其相关研究也主要集中在藏汉（汉藏）机器翻译的一些基础性工作方面，其中包括双语语料库构建技术、统计和规则相结合的机器翻译方法等，因此，藏语和汉语以及藏语到其他语言之间的机器翻译研究还面临着许多机遇和挑战。



数据集

- 第十七届全国翻译大赛(CCMT 2021)藏汉翻译评测任务提供的数据集，提供了多组训练语料，包括双语平行语料和源语言单语语料。
- 对于双语语料，我们将其整合为一个文件用于模型的训练;对于单语语料，我们先将文档中的句子进行提取，用于模型的预训练和反向翻译。

	藏语-汉语
训练集	157,959
验证集	1,000
测试集	1,000
单语语料	5,435,135

数据预处理

1. 由于平行语料的文件格式不一致，对所有文件进行平行语料的提取，生成格式一致的文件，在这个过程中，对汉语和藏语的句子进行处理，如全角半角转换，非法字符删除等；
2. 汉语分词工具采用 jieba 分词，藏语分词采用按音节划分；
3. 采用 Moses 中的 normalize-punctuation.perl、tokenizer.perl 脚本，对两种语言的数据进行标点符号标准化和进一步切分；
4. 对以词为单位分割的语料，采用 subword-nmt 进行训练 BPE 并应用于语料，BPE 操作符规模设置为藏语 10k、汉语 20k；
5. 采用 Moses 中的 clean-corpus-n.perl 脚本，过滤句子过长或双语句子长度比过大或过小的句对；
6. 从中文单语数据中根据句子长度选取大约三百万条数据，利用反向翻译对语料进行扩展。

过滤后的训练集	155,082
扩展后的训练集	3,895,088
验证集	1,287
测试集	1,000

[1] Sennrich R. Improving Neural Machine Translation Models with Monolingual Data[J]. Computer Science, 2015.

[2] Sennrich R. Neural Machine Translation of Rare Words with Subword Units[J]. Computer Science, 2015.

方法

1. 汉语-藏语翻译模型训练。利用已有的藏汉平行语料训练一个汉语-藏语的神经机器翻译模型，直到该模型收敛。用训练好的汉语-藏语翻译模型将筛选后得到的汉语语料做翻译，得到大量的伪语料，用于之后的藏语-汉语翻译模型的训练；
2. 在伪语料中的藏语句子开始部分添加标记“<BT>”，表示该条数据是由翻译模型生成的，也可以认为该条数据的分布与真实的数据分布具有差异性，方便模型从混合的数据中学到真正有用的信息和内容；
3. 使用真实的平行语料再进行一次微调；
4. 译文后处理。主要包括了全角半角转换，重复字符的删除等。

[1] Caswell I,Chelba C,Grangier D. Tagged Back-Translation[J]. 2019.

实验设置信息

操作系统：CentOS Linux release 7.9.2009 (Core)
深度学习框架：Pytorch 1.7.0
机器翻译框架：fairseq 0.10.2
CPU：AMD EPYC 7742 64-Core Processor
内存：200 GB
GPU：GeForce RTX 3090
显存：24GB

Emb_dim	1024
FFN_dim	4096
编码器层数	6
解码器层数	6
Dropout	0.3
Label_smoothing	0.1
Optimizer	Adam

[1] Vaswani A,Shazeer N,Parmar N, et al. Attention Is All You Need[J]. arXiv, 2017.

实验结果

beam_size=8

编号	设置	测试集
1	基线模型	45.29
2	+Transformer 大参数版本	45.67(+0.38)
3	+带标签的反向翻译	45.88(+0.21)
4	+双语语料微调	46.25(+0.96)
5	+后处理	46.37(+1.08)

THANK YOU