

# 藏语句子语义组块标注数据集的构建方法研究

旦正吉<sup>1,2,3</sup>, 华却才让<sup>1,2,3</sup>, 完么措<sup>1,2,3</sup>, 白颖<sup>1,2,3</sup>

1. 青海师范大学, 计算机学院, 青海, 西宁, 810008;
2. 青海师范大学, 藏语智能信息处理及应用国家重点实验室, 青海, 西宁, 810008;
3. 青海师范大学, 藏文信息处理教育部重点实验室, 青海, 西宁, 810008  
3338435909@qq.com, peljortsering@qq.com

## 摘要

语义组块对自然语言的语义理解和分析有着重要的作用, 其自动标注技术依赖于良好的语义组块标注训练数据集。目前, 藏语方面未发现语义组块研究方面的分类体系, 考虑到按粗粒度分析语义不利于语义解析和知识抽取等任务, 选择了细粒度语义分析方法, 依据不同藏文句型中语义组块的结构特征, 制定了藏语句子语义组块标注规范 (TSSCTS-13)。在此基础上, 构建了一个实用的藏语句子语义组块标注资源库 (TSSCTL-44302)。截至目前, 共完成了 498619 个语义组块标注, 并在该文提出的藏文音节向量和 BILSTM-CRF 相结合模型上完成了自动识别的实验。综合测试实验结果 F1 值为 95.28%, 精确率为 94.95%, 召回率为 95.62%, 结果表明该文构建的数据集可以应用于藏语语义领域的语义组块识别任务。

**关键词:** 语义组块; 藏语语义组块标注; 语义组块标注库; 藏语语义分析

## A Study on the Construction of a Tibetan Sentence Semantic Block Annotation Dataset

DanZhengji<sup>1,2,3</sup>, HuaQuecairang<sup>1,2,3</sup>, Wanmaicuo<sup>1,2,3</sup>, BaiYing<sup>1,2,3</sup>

1. School of Computer Science and Technology, Qinghai Normal University, Xining 810008, China;
2. The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Qinghai Normal University, Xining 810008, China;
3. Key Laboratory of Tibetan Information Processing of Ministry of Education, Qinghai Normal University, Xining 810008, China  
3338435909@qq.com, peljortsering@qq.com

## Abstract

Semantic chunks play an important role in the semantic understanding and analysis of natural languages, and their automatic annotation technology relies on a good training dataset for semantic chunk annotation. At present, no classification system for semantic chunk research has been found in Tibetan. Considering that coarse Particle size analysis analysis is not conducive to semantic analysis and knowledge extraction

©2023 中国计算语言学大会

根据《Creative Commons Attribution 4.0 International License》许可出版

**基金项目:** 国家自然科学基金 (62166034); 藏语智能信息处理及应用国家重点实验室项目 (2020-ZJ-Y05); 青海省应用基础研究计划项目 (2021-ZJ-727); 青海师范大学创新创业训练项目 (qhnucxy2022028)

and other tasks, fine granularity semantic analysis method has been selected. According to the structural characteristics of semantic chunks in different Tibetan sentence patterns, the Tibetan sentence semantic chunk labeling specification (TSSCTS-13) has been formulated. On this basis, a practical Tibetan sentence semantic chunk annotation resource library (TSSCTL-44302) was constructed. As of now, a total of 498619 semantic block annotations have been completed, and automatic recognition experiments have been completed on the proposed Tibetan syllable vector and BiLSTM-CRF model. The comprehensive test experimental results show that the F1 value is 95.28%, the accuracy rate is 94.95%, and the recall rate is 95.62%. The results indicate that the dataset constructed in this paper can be applied to semantic chunk recognition tasks in the Tibetan semantic field.

**Keywords:** Semantic chunks , Tibetan semantic chunk tagging , Semantic chunk annotation library , Semantic Analysis of Tibetan Language

## 1 引言

自然语言处理 (Natural Language Processing, NLP) 是用计算机处理人类语言的一种技术。语言是人类用来传递信息、交流思想、思考推理、传承文明、身份认同的一个标志, 以往人们认为语言是人与人之间的一种交流工具, 随着信息技术的发展和普及, 人们踏入了人工智能时代, 语言已成为人机交互的工具。研究人工智能的核心是语言处理, 语言处理的难点在于语义理解。语义理解和分析是自然语言处理的重要目标之一, 其致力于获取给定文本所蕴含的语义信息, 并以计算机能理解的某种方式进行展示 (刘亚慧 et al., 2020)。目前, 语义理解和分析的研究主要包括深层语义分析和浅层语义分析, 其中深层语义分析的相关工作主要包括语义依存分析。然而, 深层语义分析存在语义层次涉及范围广, 短期内难以形成具有较强实用性的成果 (张秀龙 et al., 2012)。本文研究的语义组块分析技术是自然语言处理中浅层语义分析和句法分析的代表, 旨在解释自然语言中语法和语义之间的关联。

语义分割是计算机视觉中的基本任务, 其将视觉输入分为不同的语义可解释类别, 并且每一类别在真实世界中都有意义。目前, 国内外的相关学者已经对此做出了不少研究, 取得了很多有意义的成果 (孙广路 et al., 2011; 李业刚 and 黄河燕, 2013a; 丁伟伟 and 常宝宝, 2009; 余正涛 and 樊孝忠, 2005; 魏楚元 et al., 2015; Wang et al., 2015)。但在藏语研究方面, 将语义分割应用于自然语言处理的研究少, 本文针对藏语句子语义分析提出一种语义组块的思路, 即将藏语句子分解成多个语义组块, 然后通过整合这些语义单元来理解句子语义。相近的研究任务有句义分割和组块分析。柔特等人 (2019; 2020) 提出了一种以语义块分析藏文句义的新方法, 处理建立在对句子进行分词和标注的基础上, 对分词结果进行重新组合, 将句子分割为若干个语义块; 诺明花等人 (2011) 因为藏文资源匮乏, 先提取了汉语语块后翻译成藏文, 同时提出藏文词序列相交算法抽取藏文短语; 李琳等人 (2013) 提出了五种句法功能组块及功能组块边界识别策略, 并基于条件随机域模型来识别了功能组块边界问题; 江荻 (2003) 提出藏语组块分析和块内分词的组块自动分词方法, 并进行功能性归并。由此可见, 目前藏语的语义组块并没有统一的分类体系, 也未发现藏语语义组块数据集构建研究方面的相关报道和数据。

作为实现浅层语义分析的基础, 标有语义组块的语料资源至关重要, 其可以极大地促进语义分析相关模型和算法的测试与研究。为了解决上述语义组块资源匮乏的问题, 并更好地满足语义理解分析和知识获取等研究需要, 本文提出了藏语语义组块数据集构建的方法, 并利用该

方法初步构建了一个藏语句子语义组块标注资源库 (Tibetan Sentence Semantic Chunk Tagging Library, TSSCTL-44302), 为藏语句子语义组块分析提供了数据资源。基于对实际语料的考察, 提出了一种藏语句子语义组块标注规范 (Tibetan Sentence Semantic Chunk Type Set, TSSCTS-13), 为制定国家标准的藏语语义组块标注规范奠定了基础。为了更准确地识别语义组块, 进一步将其细分为 53 个训练标注类别。语义组块数据集中标注的语料拥有更细粒度的语义信息, 细粒度的语义信息不仅体现在我们标注的语义组块数据集中拥有更丰富的语义组块类型, 还体现在语义组块数据集标注步骤的多样性。采用本文提出的 TS-BILSTM-CRF 识别方法在构建好的藏语语义组块数据集上进行了实验测试, 验证了本文标注语料的合理性以及构建的藏语语义组块数据集的有效性。目前, 语义组块标注技术被广泛地应用于知识抽取、机器翻译、自动文本摘要、信息检索和自动问答等多种下游任务。

## 2 藏语语义组块的分类

根据 Abney 对组块的定义, 它是一种语法结构, 是符合一定语法功能的非递归短语。每个组块都有一个中心词, 组块内的所有成分都围绕该中心词展开, 任何一种类型的组块内部不包含其他类型的组块 (Abney, 1991)。汉语组块也借鉴了英文的研究方法 (周强 et al., 1999; 赵军 and 黄昌宁, 1999)。藏语研究方面, 语义块是指将一个句子分割为若干个相对独立的语义单元, 长度基于词义之上句义之下, 语义组块是一种语法、语义、语用关联的预处理手段, 各语义块之间是非递归、非嵌套和不重叠的 (柔特 et al., 2019; 柔特 et al., 2020)。语义组块是一个记忆组织的单位, 小到一个词, 大到一个句子, 通常是由两个、三个或更多的词组合而成。语义组块是人类记忆的基本单位, 具有一定的功能或意义, 能够帮助语言学习者在大脑中进行整存整取 (旦正吉 et al., 2022)。

语义组块分析是对非结构文本进行理解的一种重要手段, 对自然语言处理领域中的多项应用具有重要意义。对藏语进行组块分析应该有语义因素的考虑, 这是因为组块可以揭示句子的结构, 而语义能够反映句子的意义, 语义组块正是句子中表达语义的单位。一个语句中可能存在多个语义组块类型, 语义组块类型的确定是语义组块数据集构建过程中最重要也是最复杂的工作之一。需要在语义组块的种类数量方面寻找一个平衡, 即在能够保证获得必要的语义信息的同时, 尽可能地精简语义组块的种类。如果语义组块的种类过少, 会造成无法准确地获取文本的语义信息, 造成后期知识抽取和问答系统等下游任务无法顺利进行。例如: 在 “མཚོ་/B-OBJམཚོ་/I-OBJམཚོ་/I-OBJམཚོ་/I-OBJ/O (头顶帽子)” 和 “རི་/B-LOCརི་/I-LOCམཚོ་/I-LOCམཚོ་/I-LOCམཚོ་/B-EVE/O (堆在山上)” 这两个句子中, 若把 “མཚོ་” 都理解为事物, 这是不合理的, 因为这两个句子在语义上表达的意思不同。而如果语义组块的种类过多, 又会增加标注人员对语义组块标注的难度。

目前, 藏语语义组块分类的研究还没有统一的分类体系。本文借鉴汉语等语言的语义组块分类方法, 结合藏语自身的特征和语言现象并经过实际的藏语语料考察与验证后, 最终藏语语义组块以语义为依据制定了 13 种语义组块类型 (TSSCTS-13), 如表 1 所示。为了在实际语料标注过程中能够准确并容易地区分确定这些主要语义组块类型, 还在表 1 中给出了它们的判断标准以及例句。语义组块标注规范的制定是语料库构建中的首要环节, 也是语料标注合理化的保障。本文提出的藏语语义组块分类体系 (TSSCTS-13) 的目标是能够更加准确的抽取语义, 可扩展细粒度语义组块, 所以相对细致的标注粒度很有必要。

表 1: 藏语语义组块种类汇总表

类型	标注范围	标记符号	判断标准—相应例句
人物 (མི་ལྔ།)	姓名、关系、人物介绍、个人、作者 (མིང་དང་འབྲེལ་གཞི་སྒྲིག་སྒྲིག་དང་མི་སྤྱིད་ ཚུལ་པ་པོ་བཅས་ཀྱི་མིང།)	CHA (Character)	发出直接可控行为或思维活动的有意识的主体： a. 表示人的姓名 (如, འཇུ་བྲག་ཤེས་བཟང་པོ་/CHA <sub>1</sub> ) b. 表示人与人之间的关系 (如, བྲག་ཤེས་ཀྱི་བྲག་ཚུང་ཟླ་བ་/CHA <sub>1</sub> ) c. 表示某个角色的基本情况 (如, གཤེས་ཀྱི་འཇུ་མ་པའི་སྤྱོད་མ/CHA <sub>1</sub> ) d. 表示独立的、有自我意识的人类 (如, བྲག་ཤེས་/CHA <sub>1</sub> ) e. 表示创作或其他艺术品的人 (如, ལྷན་པག་ཚུལ་པ་པོ་/CHA <sub>1</sub> )
事物 (བྱ་དངོས།)	语言、颜色、抽象物、技术、人工物、动植物、自然物 (སྐད་བཟང་དང་ལ་དོན་གྱི་མཚན་དངོས་པོ་དང་ ང་ལག་སྐད་མིས་བཟོས་དངོས་པོ་དང་རི་རྒྱུས་ སྐྱེས་ཚུགས་པར་བྱུང་བྱ་དངོས་བཅས་ཀྱི་མིང།)	OBJ (Object)	客观存在于自然界的一切物体或现象： a. 表示不同语言 (如, བོད་སྐད་/OBJ <sub>1</sub> ) b. 表示物件的颜色 (如, དམར་པོ་/OBJ <sub>1</sub> ) c. 表示看不见摸不着 (如, ལེས་ལ་/OBJ <sub>1</sub> ) d. 表示制造一种产品的技能 (如, ལག་ཤེས་/OBJ <sub>1</sub> ) e. 表示人类劳动的产物 (如, མིས་བཟོས་ཚུང་ཟླ་བ་འཇུ་/OBJ <sub>1</sub> ) f. 表示动物和植物 (如, རྩ་བ་/OBJ <sub>1</sub> ) g. 表示存在于自然界中的各种物体 (如, རྣམ་མཁའ་/OBJ <sub>1</sub> )
地点 (ས་གནས།)	国家、省、城市、河流、山脉、建筑、方位、地址、处所、星球 (རྒྱལ་ཁབ་དང་གོང་ཁྲིམ་རྒྱུ་རྒྱུན་དང་རི་རྒྱུད་ འཇུགས་སྐྱུན་དང་ཕྱོགས་མཚམས་ས་གནས་དང་ ང་གནས་ལུ་ལགོ་ལ་བཅས་ཀྱི་མིང།)	LOC (Location)	物体某一时刻所在出或地区： a. 表示在某国家名词 (如, ལྷན་གུ་མི་དམངས་རྒྱུ་མཐུན་རྒྱལ་ཁབ་/LOC <sub>1</sub> ) b. 表示在某省名词 (如, མཚོ་ཚོན་/LOC <sub>1</sub> ) c. 表示某城市名词 (如, རྩ་ལིང་/LOC <sub>1</sub> ) d. 表示某河流名称 (如, མ་ཚུ་/LOC <sub>1</sub> ) e. 表示某座山的名称 (如, ལྷ་ལའི་རྩུགས་པོ་རི་/LOC <sub>1</sub> ) f. 指人工建筑而成的资产 (如, ཁང་བ་/LOC <sub>1</sub> ) g. 表示某事物所在的某方向的词 (如, ཤར་/LOC <sub>1</sub> ) h. 表示所在地或通信地点的词 (如, མོག་ཚེལ་བཟུ་བཞིན་/LOC <sub>1</sub> ) i. 表示某人或某物所在位置的词 (如, སློབ་ཁང་/LOC <sub>1</sub> ) j. 表示天空中发光或反射光的天体 (如, ཟུང་/LOC <sub>1</sub> )
时间 (དུས་ཚོད།)	一般时间、日、周、月、季度、年、时间范围、世纪 (སྤྱིར་བཏང་གི་དུས་ཚོད་དང་ཉིན་གཤམ་འཁོར་ དང་ཟླ་བ་དུས་ཚོད་དང་ལོ་ཚོགས་དུས་ཚོད་ཀྱི་ བྱུང་བའདུག་དང་དུས་རབས་བཅས་ཀྱི་མིང།)	TIM (Time)	事件所发生的时间点或时期： a. 表示时间的词 (如, དུས་ཚོད་ལྔ་/TIM <sub>1</sub> ) b. 表示某一日的词 (如, ལ་སང་/TIM <sub>1</sub> ) c. 表示某一周一周里的某一天 (如, གཤམ་ཉིན་/TIM <sub>1</sub> ) d. 表示某一月或一月里的某一天 (如, ཟླ་བུག་པ་/TIM <sub>1</sub> ) e. 表示某一月的某一个季节 (如, དབྱར་/TIM <sub>1</sub> ) f. 表示某年 (如, ཉིན་རྒྱུད་བཅོམ་པོ་/TIM <sub>1</sub> ) g. 表示开始到结束的一段时间 (如, གནའ་ནས་ད་བར་/TIM <sub>1</sub> ) h. 表示某一世纪 (如, དུས་རབས་ཉི་ལུག་/TIM <sub>1</sub> )
原因 (རྒྱུ་མཚན།)	缘由、因果 (རྒྱུ་རྐྱེན་དང་རྒྱ་འབྲས་ཀྱི་མིང།)	REA (Reason)	造成某种结果或者另一件事情发生的条件： a. 表示某个事件或现象发生的原因集合 (如, རྒྱུ་/REA <sub>1</sub> ) b. 表示某个现象引起另一个现象 (如, རྒྱུ་མཚན་/REA <sub>1</sub> )
组织 (སྐྱོག་འཇུགས།)	机构 (ཁོ་ལས་ཀྱི་མིང།)	ORG (Organization)	实现某种共同目标而协调行动的过程和结果： a. 表示机关、团体或其他工作单位 (如, སློབ་གསོ་རྒྱུ་ཁག་/ORG <sub>1</sub> )
指示 (སྐྱེལ་ཚིག།)	指引、提醒 (ཁྱིད་སྟོན་དང་བློན་སྐྱེལ་ཀྱི་མིང།)	INS (Instructions)	指引、告示、指点、把事物拿出来或指出来使别人知道： a. 表示指点引导 (如, ལྷེ་ཁྱིད་/INS <sub>1</sub> ) b. 表示指点别人或促使别人注意 (如, མ་བྱེད་ཚིག་/INS)
比喻 (དཔེ་འགོད།)	程度、品格、情感、意识 (ཚིག་གཞི་དང་གཤེས་རྒྱུད་ བཅོམ་ལེས་དང་འདུ་ཤེས་བཅས་ཀྱི་མིང།)	ANA (Analogy)	事物之间的相似点, 把某一事物比作另一事物, 还表示喜怒哀乐等情感词： a. 表示某一方面的水平或状况 (如, ཤེན་ཏུ་/ANA <sub>1</sub> ) b. 人的内在道德或伦理方面的修养 (如, ལུག་སྤྱོད་/ANA <sub>1</sub> ) c. 表示喜怒哀乐的词 (如, དགའ་/ANA <sub>1</sub> ) d. 表示一种自我感受 (如, སེམས་འདུ་ཤེས་/ANA <sub>1</sub> )

下一页继续



表 2: 藏语句子类型及语义组块分类

句型	描述	例句
陈述句 (སྨྲ་བའི་རྒྱ་ཚིག་)	说话人用来陈述某一件事或回答某一个问题的句子。(རྒྱ་བའི་རྒྱ་ཚིག་ལ་འགྲེལ་བའི་དོན་གྲངས་ལ་མ་འོངས་པོ་ལ་སྨྲ་བའི་རྒྱ་ཚིག་རྒྱུ་ཡིན།)	མི་ན་/LOCགཅན་གཞན་/OBJམང་པོ་/NUM ཡོད་/EVE <sub>1</sub> /O
疑问句 (འདོད་ལྡན་རྒྱ་ཚིག་)	具有疑问语调的表示提问的句子。(རྒྱ་བའི་རྒྱ་ཚིག་ལ་འགྲེལ་བའི་དོན་རྒྱ་ཚིག་ལ་མ་འོངས་པོ་ལ་འདོད་པའི་རྒྱ་ཚིག་རྒྱུ་ཡིན།)	ར་ཚོ་རྒྱ་བའི་ལས་/OBJམཚོ་བ་/DISམ་མཚོ་ལ་/EVE <sub>1</sub> /O
祈使句 (སྨྲ་བའི་རྒྱ་ཚིག་)	说话人用来表示祝愿、命令、请求、禁止等语气的句子。(མ་འོངས་པོ་ལས་ལ་གཞིགས་པར་སྨྲ་བའི་རྒྱ་ཚིག་ལ་འགྲེལ་བའི་དོན་རྒྱ་ཚིག་ལ་མ་འོངས་པོ་ལ་འདོད་པའི་རྒྱ་ཚིག་རྒྱུ་ཡིན།)	བཤམ་གྱིས་བདེ་ལེགས་/OBJཤོག་/EVE
感叹句 (འདོད་ལྡན་རྒྱ་ཚིག་)	带有浓厚感情的句子。一般处在句子的首位。(ཡིད་ཀྱི་འགྲེལ་བའི་རྒྱ་ཚིག་ལ་འགྲེལ་བའི་དོན་རྒྱ་ཚིག་ལ་མ་འོངས་པོ་ལ་འདོད་པའི་རྒྱ་ཚིག་རྒྱུ་ཡིན།)	མེ་མ་ཉི་/OBJམོ་མོ་/LOCའཕེལ་གྱིས་མཚོགས་/EVE <sub>1</sub> /O

### 3 数据构建与统计

#### 3.1 数据来源与预处理

目前藏语语义方面没有公开的数据集，本文搜集的实验数据为青海民族出版社出版的小学至初中的义务教育课程标准实验教科书《语文》为主的文本语料，详细数据为：二年级至六年级（上下册）共 13211 个句子；七年级至九年级（上下册）共 31091 个句子，具体的相关语料数据统计如表 3 所示。

表 3: 实验数据规模统计

数据源	上册 (句子)	下册 (句子)	数据规模 (音节)
二年级	298	334	5586
三年级	344	371	6352
四年级	1188	2147	28976
五年级	1688	1639	22833
六年级	2293	2909	49677
七年级	5780	5315	106613
八年级	4756	5237	94378
九年级	5612	4391	120708

深度学习模型训练需要一定大小的实验数据，训练需要进行预处理，实验数据预处理的流程如图 1 所示。搜集到的生语料需要进行语料清洗，具体内容包括处理非藏文字符、断句和音节切分。由于直接从语料中删除非藏文字符，包括阿拉伯数字、英文、汉字等，会对句子的可读性和正确性产生影响，存在句子上下文不连贯或不正确的现象，因此本文将设置正则表达式自动去除生语料中存在的非藏文符号。由于 Bi-LSTM 语言模型是按构成语句的词序列或音节序列处理数据，为了提高模型训练的效率需要对数据进行批处理，由于只有断句才方便对数据进行批处理，因此对数据进行断句是有必要进行的。本文把人工收集整理语料中出现的长文本进行切割，以藏文单垂符“།”来进行断句。由于目前的藏文分词技术还未能满足客户的需求，导致准确率较低，会对模型的训练产生一定的影响。因此，本文以藏文音节作为模型的输入研究藏语

语义组块分类方法，采用藏文音节分隔符“.”来进行音节切分。为了保证音节切分质量，对音节切分结果进行了人工校对。

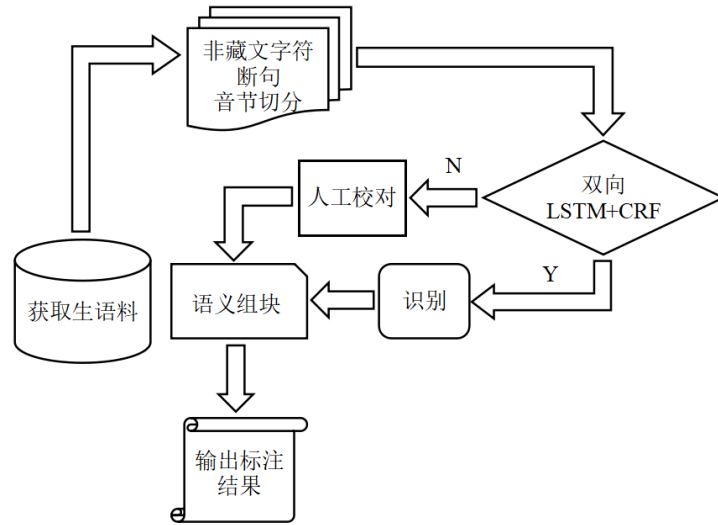


图 1: 实验数据预处理流程图

语义组块的自动标注主要包括三个环节，首先语料进行句子级别的音节切分，然后把这些预处理好的文本送入到系统的自动标注模块中，此模块用已训练的双向长短期记忆网络和条件随机场相结合的解码器层对文本进行自动标注。最后为了确保数据质量，再次对自动标注的语义组块语料进行人工校对。

### 3.2 统计与分析

基于第 1 节的藏语句子语义组块数据集构建方法，截至撰写本论文时，一共完成了 44302 句藏语语义组块标注，共包含 498619 个语义组块，各种类型的语义组块在标注数据集中的分布如图 2 可以得到。初步形成了一个细粒度的句子语义组块标注数据集 (TSSCTL-44302)。在这个数据集中，事物语义组块和事件语义组块所占比例最高，分别为 31.2% 和 29.5%。这是因为本文语料是小学和初中藏语文课文部分，小学课文更偏向一些说明文，初中的课文更偏向一些记叙文，并且课本里短短的一篇课文中也存在事件和事物语义组块。

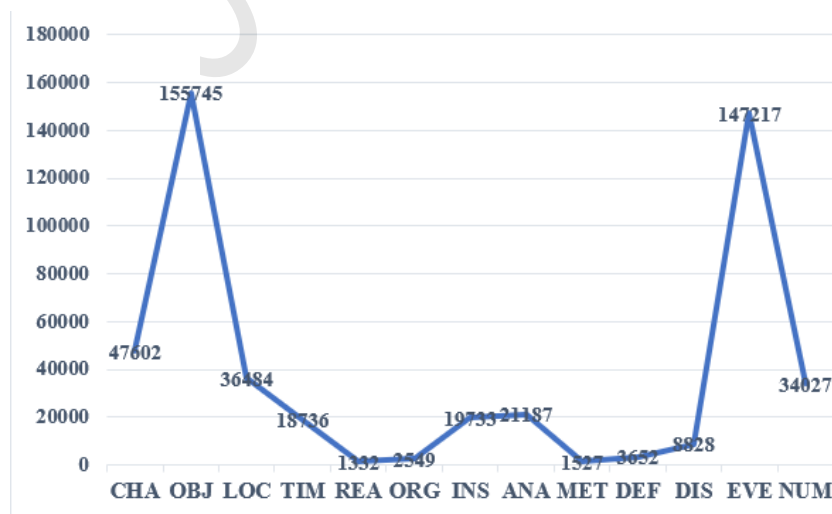


图 2: 语义组块在标注数据集中的分布

## 4 自动识别方法与结果

### 4.1 基于 TS-BiLSTM-CRF 的藏语语义组块分类方法

目前, 由于未发现藏语句子语义组块数据集研究方面的相关报道和数据, 因此, 没有适合的基准 (Baselines) 对比本文标注数据集的有效性。为了测试标注语料的合理性以及构建的藏语语义组块数据集的有效性, 用本文的语义组块数据集进行了语义组块自动识别的实验。采用 TS-LSTM-CRF 和 TS-BiLSTM-CRF 进行对比实验。具体实验过程如下:

采用 BIO 标注法, 其中, B 表示语义组块开始部分; I 表示语义组块非开始部分, 可以出现一个或多个; O 表示非语义组块。例如: “ཁོས་/B-CHAགང་/B-LOCལང་/I-LOCལང་/B-OBJལྟ་/I-OBJལྟ་/B-EVEལྟ་/I-EVEལྟ་/I-EVEལྟ་/I-EVEལྟ་/O”。

目前序列块识别或者标注算法上表现最好的是 Bi-LSTM 算法 (李业刚 and 黄河燕, 2013b; 李丽双 and 郭元凯, 2018), 在预测当前标签时, BiLSTM 善于处理长距离的上下文信息, 但无法处理标签间的依赖信息。CRF 相比其他概率图模型能够利用更加丰富的标签分布信息, 能通过邻近标签的关系获得一个最优的预测序列, 并弥补 Bi-LSTM 的缺点。本实验将 TS-BiLSTM 网络与 CRF 网络相结合, 形成一个 TS-BiLSTM-CRF 网络, 其结构如图 3 所示。首先将数据文本进行句子序列的音节向量化, 通过向量化学习到结果标注的映射, 送入到编码器的 Bi-LSTM 模型。Bi-LSTM 从前向和后向两个方向上学习上下文特征, 然后将双向 LSTM 输出结果作为 CRF 的输入, 最终由 CRF 预测全局最优标签序列。

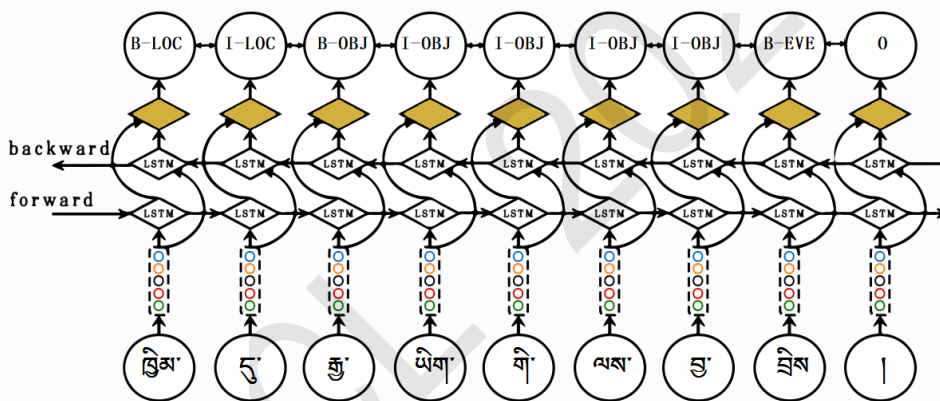


图 3: TS-BiLSTM-CRF 网络

### 4.2 实验结果与分析

将构建的藏语句子语义组块标注数据集 (TSSCTL-44302) 随机打乱, 并按照 7:2:1 的比例划分为训练集、验证集和测试集, 最后得到拥有 31012 条语句的训练集, 8860 条语句的验证集以及 4430 条语句的测试集。本文采用精确率 P (Precision)、召回率 R (Recall) 和 F1 (F-Score) 值来衡量最终的实验效果。

实验表明, 由于双向 LSTM 能够获取上下文有效信息特征, 再加上 CRF 能够充分考虑标注序列的顺序性, 得到全局最优标注序列。相比于 TS-LSTM-CRF 模型方法, 基于 TS-BiLSTM-CRF 模型的藏语语义组块识别的 P、R 和 F1 值三项指标分别提升了 2.47 个百分点、2.75 个百分点、2.61 个百分点。但这并不能说明本文语义组块数据集的语义组块识别任务容易, 出现图 4 中的实验效果是因为本文语义组块数据集在初步构建时所选用语料的长度较短, 并对语料中出现的长文本进行了切割, 而 TS-BiLSTM-CRF 模型更容易捕捉短句中词与词之间的上下文

依赖关系。事实上，本文构建的语义组块数据集有挑战性，这是因为它包含更详细的语义组块类型和更细腻的语义信息，这增加了语义组块自动识别的难度。具体识别效果如图 4 所示。

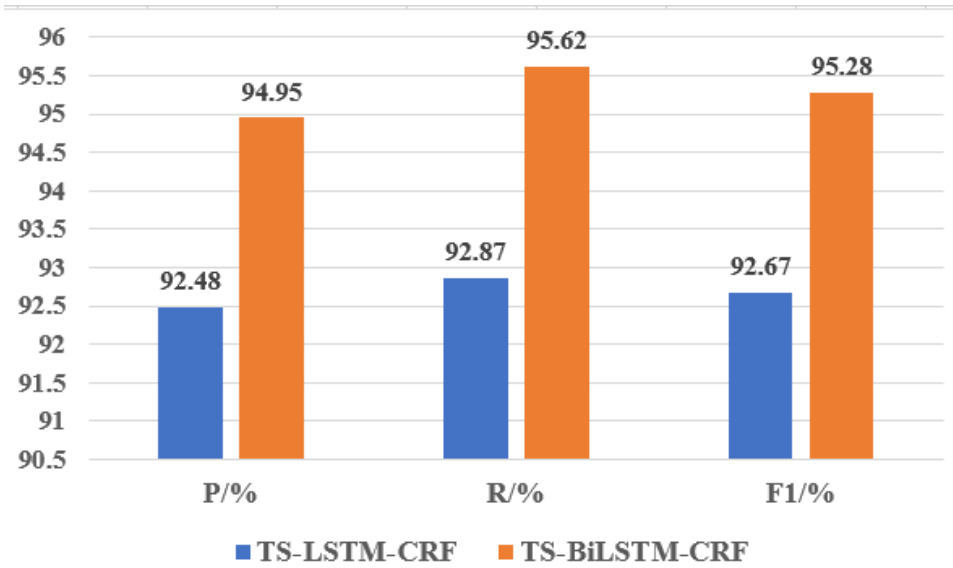


图 4: TS-LSTM-CRF 与 TS-BiLSTM-CRF 评测指标

图 5 为 TS-BiLSTM-CRF 模型对于各个语义组块类型的 Pre、Recall、F1 值，可看出模型对原因语义组块 (REA)、方法语义组块 (MET) 的识别效果较好，由于区别语义组块 (DIS) 和数字语义组块 (NUM) 在数据集中出现较少，导致模型对于 DIS、NUM 的识别效果略差，但总体效果较好。造成不同语义组块类型间精确率不尽相同的主要原因有：(1) 人工标注的语义组块分类不可避免地存在误差；(2) 藏语语义组块类别的结构特征有关；(3) 部分语义组块类别的训练量不足，数据覆盖率小，对整体识别精度影响不大。在后续的研究实验中，将获得更多的藏语语义组块数据，从而提高识别的准确性和有效性。

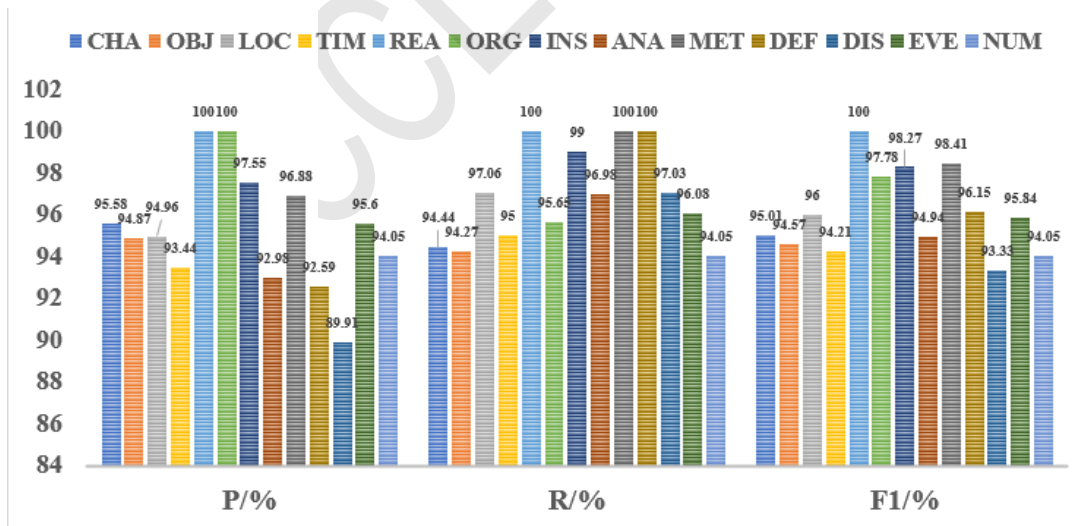


图 5: TS-BiLSTM-CRF 模型对于各个语义组块的 Pre、Recall、F1 值

为了进一步展现实验的识别效果，本文设计了藏语句子级语义组块识别系统，主要功能有音节切分、语义组块识别、语义组块分类等。其可视化系统界面如图 6 所示。

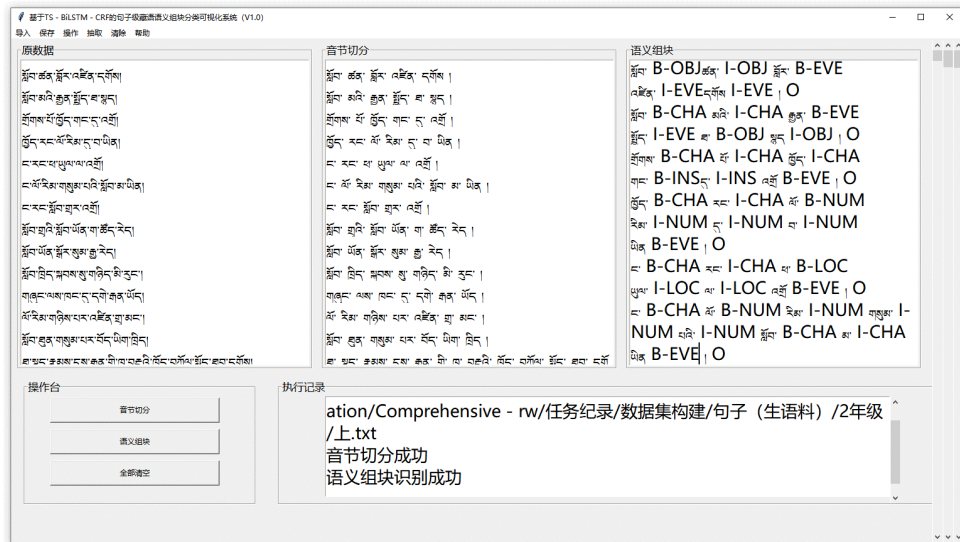


图 6: 藏语句子语义组块分类可视化系统

## 5 结束语

语义组块数据集的构建对自然语言语义分析和理解等研究有着重要的作用。本文对实际的藏语语料进行了详细的考察，并依据不同藏文句型中的语义组块的特征结构，制定了藏语句子语义组块标注规范 (TSSCTS-13)。为制定国家标准的藏语语义组块标注规范奠定了基础。为了更准确地识别语义组块，进一步将其细分为 53 个类别，其中包括更细致的种类和更细腻的语义信息，有助于语义组块准确的分类。在此基础上，研究了藏语语义组块数据集构建方法。选择合适的尺度进行分析是至关重要的，细粒度的语义信息不仅体现在本文语义组块种类的丰富性，还体现在本文语义组块标注步骤的多样性。最后，本文构建了一共拥有 44302 条语句的藏语句子语义组块标注资源库 (TSSCTL-44302)，并将语义组块数据集在 TS-BILSTM-CRF 的模型上进行了自动识别的实验。

目前，本文语义组块数据集的构建还有较长的路需要走，后期的工作重心将集中于扩大语义组块数据集的规模，以及考虑如何更好地对长语料进行细粒度的语义组块标注。此外，利用已有的语义组块数据集，设计相关语义组块的自动识别算法，提高语义组块识别的准确率，并将其运用于语义分析和知识获取等下游任务，也是未来的主要工作。

## 参考文献

Steven Abney. 1991. *Principle-based parsing: Computation and psycholinguistics*. Dordrecht.

Z Wang, T Jiang, B Chang, et al. 2015. Chinese semantic role labeling with bidirectional recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1631.

丁伟伟 and 常宝宝. 2009. 基于语义组块分析的汉语语义角色标注. *中文信息学报*, 23(05):53–61+74.

余正涛 and 樊孝忠. 2005. 基于最大熵模型的汉语问句语义组块分析. *计算机工程*, (17):3–5+8.

刘亚慧, 杨浩苹, 李正华, and 张民. 2020. 一种轻量级的汉语语义角色标注规范. *中文信息学报*, 34(4):10–20.

周强, 孙茂松, and 黄昌宁. 1999. 汉语句子的组块分析体系. *计算机学报*, (11):1158–1165.

孙广路, 郎非, and 薛一波. 2011. 基于条件随机域和语义类的中文组块分析方法. *哈尔滨工业大学学报*, 43(7):135–139.

- 完么才让. 2014. 基于规则的藏语句法分析研究. Ph.D. thesis, 青海民族大学.
- 张秀龙, 李新德, and 戴先中. 2012. 基于组块分析的路径自然语言语义角色标注方法. 东南大学学报 (自然科学版), 42(S1):127-131.
- 拉巴顿珠, 欧珠, and 赵栋材. 2017. 藏文自动分词系统中虚词识别算法研究. 计算机应用与软件, 34(09):299-301+333.
- 旦正吉, 华却才让, 完么措, and 白颖. 2022. 基于藏文音节结合 bilstm-crf 的藏语语义组块分类标注. In 全国少数民族自然语言处理青年论坛.
- 李业刚 and 黄河燕. 2013a. 汉语组块分析研究综述. 中文信息学报, 27(3):1-8.
- 李业刚 and 黄河燕. 2013b. 汉语组块分析研究综述. 中文信息学报, 27(03):1-8.
- 李丽双 and 郭元凯. 2018. 基于 cnn-blstm-crf 模型的生物医学命名实体识别. 中文信息学报, 32(01):116-122.
- 李琳, 龙从军, and 江荻. 2013. 藏语句法功能组块的边界识别. 中文信息学报, 27(06):165-168.
- 柔特, 色差甲, and 才让加. 2019. 藏文句子语义块识别方法. 中文信息学报, 33(06):42-49.
- 柔特, 色差甲, and 才让加. 2020. 藏文句义分割方法. 计算机工程, pages 286-291.
- 江荻. 2003. 现代藏语组块分词的方法与过程. 民族语文, (04):30-39.
- 诺明花, 张立强, 刘汇丹, 吴健, and 丁治明. 2011. 汉藏短语抽取. 中文信息学报, 25(02):105-110+121.
- 赵军 and 黄昌宁. 1999. 基于转换的汉语基本名词短语识别模型. 中文信息学报, (02):2-8+40.
- 高定国, 扎西加, and 赵栋材. 2014. 计算机识别藏语虚词的方法研究. 中文信息学报, 28(01):113-117.
- 魏楚元, 湛强, 樊孝忠, 毛煜, and 张大奎. 2015. 融合事件信息的中文问答系统问题语义表征. 中文信息学报, 29(01):146-154.