

融合事件特征及语义角色标注的藏文信息抽取技术

万福成*, 马宁, 何向真

(西北民族大学 甘肃省民族语言智能处理重点实验室, 甘肃 兰州 730030)

摘要: 本文在藏语短语句法分析的基础上, 融入功能语义信息线索, 采用在藏语短语句法树库的基础上加入语义角色标注的方法, 并融合事件触发词信息, 对藏文信息进行抽取, 并通过实验进行了对比分析。本文事件特征融合与已标注信息对藏文信息抽取有帮助, 本文模型可以很好地应用于藏文信息抽取工作。

关键词: 功能语义; 语义角色标注; 藏文信息抽取

中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-6600(2018)02-0018-06

引用格式: 万福成, 马宁, 何向真. 融合事件特征及语义角色标注的藏文信息抽取技术[J]. 广西师范大学学报(自然科学版), 2018, 36(2): 18-23.

WAN Fucheng, MA Ning, HE Xiangzhen. Tibetan information extraction technology integrated with event feature and semantic role labeling[J]. Journal of Guangxi Normal University(Natural Science Edition), 2018, 36(2): 18-23.

Tibetan Information Extraction Technology Integrated with Event Feature and Semantic Role Labeling

WAN Fucheng*, MA Ning, HE Xiangzhen

(Key Laboratory of National Language Intelligent Processing, Gansu Province,
Northwest Minzu University, Lanzhou Gansu 730030, China)

Abstract: Semantic information integrated with syntactic analysis was used for extracting the Tibetan language information. Using the method of semantic role tagging on the basis of the Tibetan phrase syntactic treebank and the information of event triggering words, the information of Tibetan language is extracted. Through comparative analysis of experiments, It is shown that the fusion of event feature and tagged information help Tibetan language information extraction, and the given model can be well applied to Tibetan information extraction.

Keywords: semantic parsing; semantic role labeling; Tibetan information extraction

信息抽取是从文本中自动地抽取主要信息, 并以结构化的形式呈现出来。信息抽取^[1]作为信息处理领域的关键技术, 在信息检索^[2]、自动问答^[3]、文本挖掘等领域有着广泛的应用。融合句法信息和功能语义信息的藏文信息抽取算法可实际运用到藏文舆情监控、藏文文本关联检测、藏文热点话题识别与跟踪等信息安全系统中, 具有较好的学术价值和广阔的应用前景。对藏文数据库和知识库的自动生成、藏文问答系统、藏文信息检索等领域研究有着积极的意义。

收稿日期: 2017-03-07

基金项目: 国家自然科学基金(61602387); 西北民族大学中央高校基本科研业务费资金项目(31920170155)

通信联系人: 万福成(1985—), 男, 辽宁台安人, 西北民族大学副教授, 博士。E-mail: wanfucheng@126.com

1 相关研究

事件抽取工作要追溯到 20 世纪中叶,Zellig Harris 对科技文献作分析,进行了信息抽取的工作,并建立了相应的索引^[4]。从 2000 年开始至今的 ACE 评测会议的召开,使得信息抽取取得了丰富的成果。

信息抽取的内容涉及各行各业,抽取的方法也囊括了绝大部分的机器学习算法^[5],信息抽取的研究从面向特定领域发展到开放领域,使得抽取到的信息更加全面^[6]。事件抽取需要语料库的支持,袁毓林^[7]研究了事件抽取的语义资源建设;赵妍妍等^[8]提出了基于触发词和二元分类技术相结合的事件类别识别;许荣华等^[9]定义了一种事件融合框架 TEFF,根据各类元事件在主题事件中的作用,将主题事件以层次化的形式表示;于洪志等^[10]采用命名实体的同现频率来计算藏文实体间的关系;Ji Heng 等^[11]研究了跨文档事件提取和跟踪以及其评价标准;丁效等^[12]研究面向音乐领域的事件抽取,并将研究成果应用在哈尔滨工业大学的语言技术平台(LTP)上;金明等^[13]通过规则加词典的方法进行藏文命名实体识别;Alan Ritter 等^[14]研究面向 Twitter 的开放域事件抽取,将事件抽取技术进一步实用化;徐涛等^[15]在李响等^[16]的研究基础上,将复句分析等规则与 MEM 相结合,研究了藏文自动断句模型,都在不同程度上推动了事件抽取工作。目前信息抽取工作可与其他应用结合,出现了很多成果,例如:邹博伟等^[17]对不确定性信息进行抽取;高庆宁等^[18]利用文档模型对网页进行抽取;刘春梅等^[19]针对开源论坛网页进行了抽取。很多研究成果都表明,信息抽取技术呈现多方位、多方向的发展。

2 藏语功能语义分析

2.1 藏语词性标注及句法规律

藏语跟汉语的语序有所不同,汉语采用的语序是主谓宾格式,藏语采用的是主宾谓格式,谓语后置是藏语最显著的特点。

为了构建藏语短语句法树库,本文将藏语的实词和虚词分别进行细化标记,如表 1 所示。

表 1 藏语短语树库词性标记集

Tab.1 Tibetan phrase syntax tree labelling

普通名词—NN	人名—NR	时间名词—NT	机构名称—NO	地名—NS	辞藻词—NE
佛教梵语—NF	数词—CD	量词—M	副词—AD	连词—CC	方位词—LC
感叹词—IJ	人称代词—RR	指示代词—RZ	疑问代词—RY	不定代词—RB	终结词—E
拟声词—ON	成语词—I	习惯语—L	缩略词—J	格助词—P	动词—VV

2.2 融合语义信息的藏语短语树库举例

按照表 1 的标注,借鉴宾夕法尼亚大学的汉语树库(宾州大学汉语树库,CTB)关于短语层次的标注规范,本文构建了面向藏文事件的藏语短语句法树库。短语句法树如图 1 及图 2 所示。

在目前藏语句法树库的构建中,依存句法相对比较多,但是依存句法体现更多的是词和词的依赖关系,本文着眼于藏语短语句法,以图 1 和图 2 的样例为标准,对藏语的实词、虚词分别进行人工标记。将经过人工标记的 2 000 句语料库作为训练语料库,对测试语料库进行解码,然后辅助以人工校对,采用 bootstrapping 的方法,逐步扩展藏语短语句法树库。

3 融合语义特征的藏文信息抽取模型

3.1 触发词识别

信息抽取首先要识别触发词即句子中核心动词,藏语动词一般处于一个句子的末尾,是整个句子的核

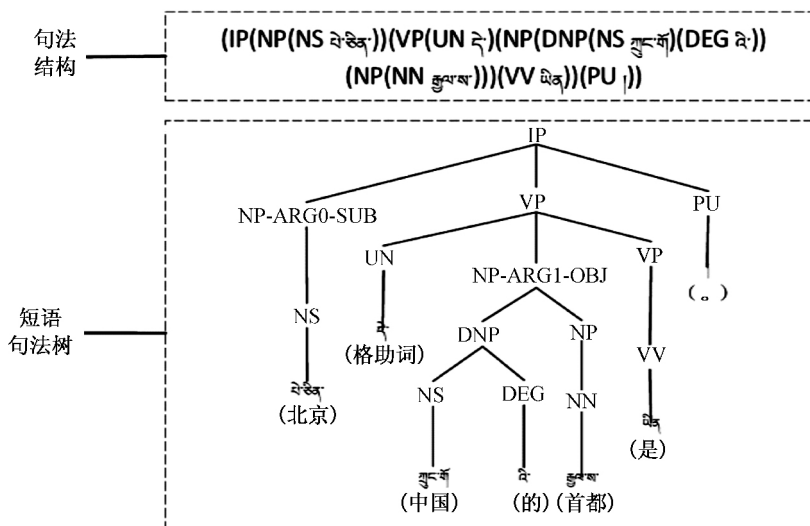


Fig.1 Tibetan Event phrase syntax tree sturcture example one

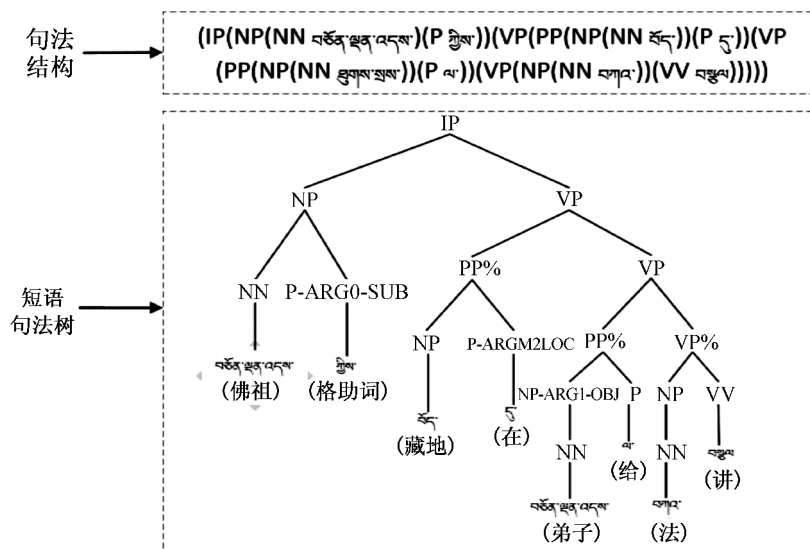


Fig.2 Tibetan Event phrase syntax tree sturcture example two

心,从位置上来看,藏语动词的识别相对于其他成分要容易,但是很多动词并没有出现在训练语料中,很多属于未登陆词。例如,“དགེ་ལྡན་ལ་བརྟུན་ཅི།”(尊重老师),假设“བརྟུན་ཅི།”(尊重)没有在训练语料中作为事件触发词出现,就不易识别出这是一个事件。但是,“བྱུང་བཤུགས་ཅི།”(尊敬)和“བརྟུན་ཅི།”(尊重)词义相近,基于此,本文基于藏语动词词典,并基于词典触发词进行词语语义相似度计算,自动扩充事件触发词,尽可能多地覆盖各种类型事件的触发词。藏语词语语义相似度,采用词语向量空间模型,利用藏语词汇的上下文语境进行统计描述,词语的上下文可以为词语提供丰富的语言学信息。首先,选取一组特征词,然后根据这组词在语料中的出现频率以及 TF-IDF 值,计算这一组特征词与每一个词的相关性,得到一个特征词向量,最后,利用向量之间的相似度作为两个词的相似度。

获取种子触发词与其信息的类别组成一个二元组(trigger, type)。如:(主席, Person/Respect),构建“藏语种子触发词—信息类别”二元对照表,二元对照表两列分别为事件的触发词和信息类别。

表 2 “种子触发词—信息类别”对照表
Tab.2 Trigger words and infomaton category comparisons

种子触发词	信息类别
དུས་ཁྱོད་ (倒闭)	Business/Declare—Bankruptcy
འགོག་གྲོལ་ (抗争)	Conflict/Attack
སྐྱོད་ (逃离)	Movement/Transport
ངམ་སྐྱོད་ (示威)	Conflict/Demonstrate
⋮	⋮

表 2 中第一列表示的意义是种子触发词,第二列表示的意义是每一个触发词的信息类别,每一个触发词对应着唯一的一个信息类别,并且该表的触发词涵盖了所有的信息类别。将信息类别标注为不同的符号,采用最大似然概率算法获取识别模型参数,则状态转移矩阵的参数获得为:

$$Ta_{ij} = \frac{TC_{i,j}}{\sum_{k=1}^N TC_{i,k}}, 1 \leq i, j \leq N,$$

其中 $TC_{i,j}$ 是所有符号对应状态 TS_i 到状态 TS_j 的次数,在语料库中代表种子触发词到信息类别的频率。

$$Tb_j(V_K) = \frac{TE_j(V_K)}{\sum_{i=1}^M TE_j(V_i)}, 1 \leq j \leq N, 1 \leq i \leq M,$$

其中 $TE_j(V_K)$ 是 V_K 在状态 TS_i 中的发射频率, $Tb_j(V_K)$ 是发射概率之和的比重。

$$TP_i = \frac{\text{Init}(i)}{\sum_{j=1}^N \text{Init}(j)}, 1 \leq i \leq N,$$

其中 $\text{Init}(i)$ 是状态 i 开始的总个数。

3.2 语义角色标注

C. J. Fillmore 的框架语义学认为,命题中的格包括:①施事格—Agentive;②工具格—Instrumental;③承受格—Dative;④使成格—Factitive;⑤方位格—Locative;⑥客体格—Objective;⑦受益格—Benefactive;⑧源点格—Source;⑨终点格—Goal;⑩伴随格—Comitative。在藏语句法规律的基础上,参照框架语义学关于格框架的分类,本文以施事、受事、时间、地点、方式等 5 种藏语功能语义学线索为研究内容。

用“谓词—论元角色”的结构形式,对句中的每个谓词(动词性谓词或名词性谓词等)标注出句子中谓词的相应语义角色成分,主要论元角色:施事者 AGR0—SUB、受事者 ARG1—OBJ;辅助论元角色:时间 ARGM—TEMP、地点 ARGM—LOC、方式 AGRM—MNR,将语义信息融入到句法结构树中,如图 1 及图 2 的短语层次表所示,进行训练,句法分析器采用 Berkeley Parser。

4 实验及评测

4.1 实验设置

语料库选取:藏文文本分类语料库,包括历史、文学、宗教、修辞、哲学、政治 6 大类的 1 200 篇文本,其中 1 000 篇作为训练文本,200 篇作为测试文本。

结果评测方案:本文以抽取的藏文信息结果,跟标准答案库对比分析,采用准确率(P)、召回率(R)和 F 值进行综合评测。

$$P = \frac{\text{正确抽取的属性个数}}{\text{实际抽取到的属性个数}} \times 100\%, R = \frac{\text{正确抽取的属性个数}}{\text{应抽取到的属性个数}} \times 100\%, F = \frac{2 \times P \times R}{P + R} \times 100\%.$$

4.2 实验方法及结果

本文实验分别以词性标注方法、触发词识别方法、词性标注融合触发词方法、语义角色标注融合触发词方法进行对比。

词性标注方法:仅采用藏语词汇的词性标注方式,例如动词་ཐོག་ཐོག་标注为 VV,名词标注为 NN;

触发词识别方法:以模型识别的结果作为信息抽取模型的输入,得到信息抽取结果;

词性标注融合触发词方法:前两种方法的结合,并与单独的方法进行对比;

语义角色标注融合触发词方法:即本文的方法,触发词作为事件特征,语义角色信息作为藏语本体特征,如图 1、2 所示,统一标注于树库中。

实验结果如表 3 所示。

表 3 藏语信息抽取结果分析

Tab.3 Result analysis

%

方法	<i>P</i>	<i>R</i>	<i>F</i>
词性标注	60.52	57.63	59.04
触发词识别方法	65.34	53.26	58.68
词性标注+触发词	68.74	62.34	65.38
触发词+语义角色标注方法	76.78	75.32	76.04

4.3 结果分析

由表 3 可以看出,仅采用词性标注的方法和仅采用触发词识别的方法整体 *F* 值都比较低。可见,仅使用词典信息以及词性标注方法对于藏文信息抽取工作贡献率不高,原因在于无论词典还是词性标注方法,更多关注于客观知识,没有将藏语语言学信息融入到模型中。而前两种方法的融合 *F* 值并没有大幅度的提升,这也验证了先验知识或是人工知识对于模型效率的增加并没有起到很好的作用。

语义角色标注信息融合触发词,也就是本文的方法,对于藏语抽取工作,在整体的准确率和召回率上都有所提升,由此可见,融合藏语句法以及语义的语言学信息,对于藏语信息处理任务有所帮助。此外,信息抽取也跟句法分析的准确率有很大关系。基于语义角色标注的藏文信息抽取模型要依赖于藏语短语句法分析模型的准确率,尤其是长句子的句法分析的准确率。

5 结语

融合事件特征及语义信息的藏文信息抽取模型,在事件特征的基础上融合了功能语义学的信息,即藏文语义角色标注技术,从而为藏文信息的抽取提供支持,本文同时可以为研究藏语信息抽取工作提供一个基线系统,下一步的研究方向将是对触发词的语义相似度进行计算,扩充触发词集合,从而更好地为信息抽取服务。

参 考 文 献:

- [1] 梁晗,陈群秀,吴平博.基于事件框架的信息抽取系统[J].中文信息学报,2006,20(2):40-46.DOI:10.3969/j.issn.1003-0077.2006.02.006.
- [2] DANIEL N, RADEV D, ALLISON T. Sub-event based multi-document Summarization[C]//Proceedings of the HLT-NAACL 03 on Text Summarization Workshop; Volume 5. Stroudsburg, PA: Association for Computational Linguistics, 2003:9-16. DOI:10.3115/1119467.1119469.
- [3] 姜吉发.开放领域汉语知识问答方法[J].计算机工程,2005,31(11):1-3. DOI:10.3969/j.issn.1000-3428.2005.11.

001.

- [4] COHEN W W, McCALLUM A. Information extraction from the World Wide Web[R]//KDD 2003 Tutorial. Washington DC: ACM Press, 2003.
- [5] SODERLAND S G. Building a machine learning based text understanding system[C]//Proceedings of IJCAI—2001 workshop on Adaptive Text Extraction and Mining. San Francisco, CA: Morgan Kaufmann Publishers, 2001: 133—154.
- [6] 谭红叶. 中文事件抽取关键技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2008.
- [7] 袁毓林. 信息抽取的语义知识资源研究[J]. 中文信息学报, 2002, 16(5): 8—14. DOI: 10.3969/j.issn.1003—0077. 2002. 05.002.
- [8] 赵妍妍, 秦兵, 车万翔, 等. 中文事件抽取技术研究[J]. 中文信息学报, 2008, 22(1): 3—8. DOI: 10.3969/j.issn. 1003—0077. 2008. 01.001.
- [9] 许荣华, 吴刚, 李培峰, 等. 基于事件框架的主题事件融合研究[J]. 计算机应用研究, 2009, 26(12): 4542—4545. DOI: 10.3969/j.issn.1001—3695. 2009. 12.040.
- [10] YU Hongzhi, JIANG Tao, ZHANG Bing, et al. Relation discovery by named entity recognition from Tibetan websites [C]//2009 1st IEEE Symposium on Web Society. Piscataway, NJ: IEEE Press, 2009: 177—179. DOI: 10.1109/SWS. 2009.5271789.
- [11] JI Heng, GRISHMAN R, CHEN Zheng, et al. Cross—document event extraction and tracking: task, evaluation, techniques and challenges[C]//Proceedings of the International Conference Recent Advances in Natural Language Processing 2009. Stroudsburg, PA: Association for Computational Linguistics, 2009: 162—172.
- [12] 丁效, 宋凡, 秦兵, 等. 音乐领域典型事件抽取方法研究[J]. 中文信息学报, 2011, 25(2): 15—20. DOI: 10.3969/j. issn. 1003—0077. 2011. 02.003.
- [13] 金明, 杨欢欢, 单广荣. 藏语命名实体识别研究[J]. 西北民族大学学报(自然科学版), 2010, 31(3): 49—52. DOI: 10. 14084/j.cnki.cn62—1188/n. 2010. 03. 018.
- [14] RITTER A, MAUSAM, ETZIONI O, et al. Open domain event extraction from Twitter[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge discovery and data mining. New York, ACM Press, 2012: 1104—1112. DOI: 10.1145/2339530.2339704.
- [15] 徐涛, 加羊吉, 于洪志. 统计与规则相结合的藏文句子自动断句方法[J]. 云南大学学报(自然科学版), 2012, 34(6): 653—657, 663.
- [16] 李响, 才藏太, 姜文斌, 等. 最大熵和规则相结合的藏文句子边界识别方法[J]. 中文信息学报, 2011, 25(4): 39—44. DOI: 10.3969/j.issn.1003—0077. 2011. 04. 008.
- [17] 邹博伟, 钱忠, 陈站成, 等. 面向自然语言文本的否定性与不确定性信息抽取[J]. 软件学报, 2016, 27(2): 309—328. DOI: 10.13328/j.cnki.jos.004860.
- [18] 高庆宁, 吴鹏, 张晶晶. 基于文档对象模型与行块分布算法的网页信息抽取[J]. 情报理论与实践, 2016, 39(4): 133—137. DOI: 10.16353/j.cnki.1000—7490. 2016. 04. 025.
- [19] 刘春梅, 郭岩, 俞晓明, 等. 针对开源论坛网页的信息抽取研究[J]. 计算机科学与探索, 2017, 11(1): 114—123. DOI: 10.3778/j.issn.1673—9418. 1510016.

(责任编辑 黄 勇)