基于机器学习方法的哈萨克语词干切分研究

库瓦特拜克·马木提
(伊犁师范大学 电子与信息工程学院,新疆 伊宁 835000)

摘 要:自然语言处理任务中词处理是基础性的工作,其结果直接影响后续任务的效果。词干和构形附加成分是哈萨克语单词的组成成分,其中词干显示单词的主要意义,而构形附加成分中包含着词法和句法信息,因此词干切分是对哈萨克语进行有效处理的基础。文中构建了哈萨克语词干切分语料库,并通过将哈萨克语词干切分看作是序列化标注问题,提出一种有效的哈萨克语词标注方法,并基于最大熵模型和条件随机场模型构建了对比词干切分实验。结果表明基于条件随机场模型的词干切分准确率比现有最好的哈萨克语词干切分系统的准确率有 15% 的提高。该方法对哈萨克语词干切分相较于基于规则的方法有了一定的提升。

关键词:词干切分;统计学习模型;最大熵模型;条件随机场模型
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2020)04-0182-07
doi:10.3969/j.issn.1673-629X.2020.04.035

Research on Kazakh Stemming Based on Machine Learning

Kuwatebaike·MAMUTI
(School of Electronic and Information Engineering, Yili Normal University, Yining 835000, China)

Abstract: Word processing is a basic task in natural language processing, which directly affects the subsequent tasks. Stem and inflectional suffix are the main components of Kazakh words. Stem displays the main significance of the word, and the inflectional suffix contains lots of information of grammar and syntax. As a result, stemming becomes the basis of Kazakh information processing. We build the Kazakh segmentation corpus, and through the Kazakh stemming as serialized label problem, propose an effective Kazakh word labeling method. Based on the maximum entropy model and the conditional random field model, a comparative word-stem segmentation experiment is constructed. It is showed that the stemming accuracy based on conditional random field model is 15% higher than that of the best Kazakh stemming system. Compared with the rule-based method, the proposed method improves the stemming of Kazakh words.

Key words: stemming; statistical learning model; maximum entropy model; conditional random field model

0 引 言

黏着语类型语言包括蒙古语、维吾尔语和哈萨克语等。黏着语类型的语言单词在组成上可以分为:词根、词干、构词附加成分、构形附加成分(附加成分也称为词缀)。一般而言,黏着语的每一个词缀都只表达一种意思或只具有一种语法功能。词根后面附加构词附加成分,形成新的词汇意义从而构成新词;而词干后面附加构形附加成分,形成与词干意义相同,语法含义不同的单词。哈萨克语单词的构造形式是通过将不同的构形附加成分按照一定的规则缀接在词干后来实现的。根据这些规则,构形附加成分是可以层叠的。哈萨克语单词的这种构形方式使哈萨克语单词的形态变化丰富而且复杂。

哈萨克语单词的构形附加成分承载着该单词数、格、体、时等大量语言相关的语法信息。每一个哈萨克语单词与其他语言不同之处在于,其语法意义不仅与单词在句子中的未知有关,也与不同构形附加成分的缀接相关,所以要分析哈萨克语单词的词性属性和语法关系就需要正确切分词干和构形附加成分。但是在现实的语言环境中,哈萨克语单词整体为一个连续的字符串形式,各构形成分之间没有形式上的分隔。首先要从单词中分离出词干和构形成分,才可以利用这些信息。同时词干在缀接构形成分时有些词干会发生相应的变化,需要进行词干的还原处理。构形附加成

收稿日期:2019-04-28 **修回日期**:2019-08-29 **网络出版时间**:2019-12-18
基金项目:新疆自然科学基金(2019D01C337);伊犁师范大学科研项目(2016YSYB09);伊犁师范大学教育教学研究项目(JGZH17151)
作者简介:库瓦特拜克·马木提(1977-),男,硕士研究生,讲师,研究方向为自然语言处理。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20191218.1113.042.html

分的识别及词干还原过程就是哈萨克语的词干切分。哈萨克语词干切分属于词法分析的基础性工作,对哈萨克语的信息检索、句法分析、机器翻译等具有重要作用。

基于机器学习的方法在哈萨克语词干切分的研究中还没有得到应用。文中首先手工标注了 100 万词汇的哈萨克语文本语料,为开展机器学习方法的研究准备了较为充分的词干切分语料;其次在哈萨克语词干切分任务中应用了最大熵模型和条件随机场模型,为哈萨克语信息处理提供了可行的机器学习方法;再次设计并实现了两种机器学习方法的对比性实验,取得了较好的实验结果,哈萨克语词干切分的准确率在条件随机场模型中达到了 85% 以上,相对于传统的基于规则的方法,取得了一定提升,为进一步利用统计学习方法研究哈萨克语信息处理技术奠定了基础。

1　哈萨克语词干切分工作研究现状

目前哈萨克语的词干切分研究工作还处于起步阶段,尤其是在统计学习领域如何将哈萨克语词干切分很好地利用到各个不同的 NLP 任务当中依然是一个值得研究的领域。当前词干切分工作主要有基于词典和规则相结合的方法[1-2]。通过在词干词典的基础上应用哈萨克语词干切分语言学规则实现了哈萨克语词干切分的方法,存在的主要问题是词干切分的准确率不高,在 70% 左右,还不能很好地满足实用性的要求。下面将对包括蒙古语、维吾尔语、哈萨克语等在内的黏着语类型的语言所采用的三种词干切分方法逐一说明。

1.1　基于词干词典和词法规则的方法

基于词干词典和词法规则的方法,在所有的黏着语类型的语言中都进行了许多尝试,2004 年古丽拉·阿东别克老师在维吾尔语词干切分研究中提出了基于规则的方法,实现了维吾尔语的词切分算法[3],利用维吾尔语中语音的同化和和谐规律实现切分。该方法存在的难点是需要收集比较完整的维吾尔语词干词典;需要根据该种语言的语言学规律设置条件规则库,同时语言中又存在规则无法完全覆盖到的特例和不规则变化。2008 年米热古丽·艾力提对维吾尔语词干切分中存在的元音弱化现象进行了讨论,提出元音弱化还原算法有助于提升词干切分的正确率[4]。阿孜古丽·夏力甫则进一步探讨了动词构形附加成分规则,在复杂特征理论的基础上进一步提升了维吾尔语动词还原效果[5]。

热娜·艾尔肯提出利用规则和词典相结合的混合处理方法进行形态还原[6],利用从左到右的分析和 Lovin 算法实现对词干的提取,平均正确率为 77.4%。早克热·卡德尔提出维吾尔语词干提取中使用名词构形词缀分析 DFA 的构造过程[7],利用构形词缀的规律性,使用有限状态自动机从右到左进行描述,最后对自动机进行方向翻转和转换来确定该自动机的操作。史建国提出将词典和规则相结合的方法对斯拉夫蒙古文进行切分[8],通过预处理部分蒙古文词,然后基于词典切分高频和部分不符合规则的词。最后对剩余的词,用切分规则生成多个候选的词切分方案,然后在这些方案中选出最优方案。通过两种方法的有机结合,发挥各自的优点,得到了性能较好的斯拉夫蒙古文词切分系统。

2008 年达吾勒·阿布都哈依尔老师在哈萨克语词干切分任务中提出利用有限状态机(FSM)和前后向切分相结合的方法[1],先对待切分单词使用有限状态机进行分析。如果成功则将输出作为切分结果,否则使用联合的改进方法进行切分。相对于最大匹配法,从正确率和切分速度两方面提高了词干切分的效果。

2011 年达吾勒·阿布都哈依尔老师又提出了利用词干词典和构形附加成分构词规则的哈萨克语词干切分方法[2],构建了 6.2 万词条的词干词典和 436 个构形附加成分构成的规则库;采用全切分算法和词法分析相结合的方式进行词干切分。该方法首先对待切分单词利用词干词典信息抽取出所有可能的词干;随后对对应某一种词干分离后的词的其余部分进行基于规则的分析,利用还原规则得到各种成分,再将其与规则库中的构形附加成分进行匹配,从而确定是否为正确的切分,并将该切分结果作为派生词放入派生词表;最后根据词干最长、概率最高和整词输出作为词干切分的最终结果输出。

基于词典和规则的方法存在以下问题:(1)词干部分存在多种切分时,选择词干最长的切分形式,可能会存在错误,例如:“领导”这个单词باششلىق ,应切分为باششى/لىق;(2)针对词干边界与词缀边界相交的切分情况,简单地采用词频统计的方法来解决;(3)词缀匹配成功,词干部分没有匹配成功时,词缀部分采用词频统计的方法来解决;(4)没有匹配的词干和词缀时,将整词作为词干处理。

1.2　基于监督学习的统计方法

在统计自然语言处理理论的基础上,哈萨克语还没有基于统计学习方法的词干切分方面的研究,汉语的分词与黏着语类型语言的词干切分有一定的相似性,同时汉语的分词技术相对较为成熟,研究的也较为深入,因此基于统计方法的汉语自动分词技术对哈萨克语的词干切分在研究中有借鉴意义。第一篇基于字标注的汉语分词是 Xue 根据汉字在词语中出现的位

置将汉字分为 4 类[9],然后利用最大熵模型标记的方法进行切分;Tseng 基于字标注方法采用条件随机场模型[10];2014 年 Liu 等提出了利用条件随机场模型分词系统在拥有自然分词边界的网络文本中使用,从而提高了领域适应性[11]。Zeng X 提出了一种基于图的标记扩展技术[12],构建了一个最近邻相似图覆盖所有已标注的 3-gram 和扩展句法信息的未标记数据即标记分布。派生的标记分布被视为隐含的证明去正则化线性条件随机场在未标记数据,最终获得一个基于字符的联合模型。

而同属于黏着语类型的蒙古语和维吾尔语提出了基于统计学习方法的相关研究。2009 年 Aisha B 提出利用特征模板和手工标记的基于统计的词干提取算法[13]。首先以特征模板为基础使用手工切分的词库和最大熵方法学习一个字符转移模型,用该模型来切分维吾尔语单词,随后利用语言知识使用条件随机场将切分结果映射为词干、词缀。该方法需要较大的手工切分词库,人工成本较高。

2011 年薛化建基于词缀库及维吾尔语构词结构,提出了规则与统计相结合的词干切分方法[14]。该方法对单词进行规则切分,采用 MAP(最大后验概率)切分评价模型对基于规则的切分结果进行赋分,选择最高分数的切分结果作为该单词的切分结果。实验结果表明,使用该方法进行维吾尔语词切分具有更高的准确率。2015 年赛迪亚古丽·艾尼瓦尔利用维吾尔语构词规则、词性特征和上下文信息[15],提出基于 n-gram 模型的词干提取方法,实验准确率达到 96.60%。2009 年候宏旭老师和刘群老师在蒙古语词干切分中提出基于 SKIP-N 语言模型方法[16]。模型对单词的上下文信息及词性信息进行考虑,解决切分规则中的二义性。首先给出单词所有可能的切分候选集合,该集合由蒙古语词切分规则获得;然后利用 SKIP-N 语言模型对候选集合中的切分进行赋分,选取打分最高的切分为结果。

2010 年赵伟提出了基于条件随机场模型的蒙古语词干切分系统[17],该方法将蒙古语词干切分问题描述为序列标注问题,利用多维度特征,使词干切分的正确率达到了较高的水平。

2011 年姜文斌老师提出了蒙古语有向图形态分析器的判别式词干词缀切分方法[18],以图状结构刻画句中词干和词缀之间的概率关系,从而借助上下文信息为每个单词确定最佳的切分标注候选。与之前词干表与附加成分表结合的枚举方法相比,提出判别式分类的切分方法,对 OOV(未登录词)的词干切分具有很好的泛化能力。以 20 万词规模的三级标注人工语料库为训练数据,采用判别式词干词缀切分的有向图形态分析器,对于含有未登录词干的情形,词级切分标注正确率提高了 7 个百分点。2011 年李文提出基于短语的统计机器翻译形态蒙文切分模型和最小上下文构成代价模型分别对词表词和未登录词进行形态切分[19]。前者选取了短语机器翻译系统中三个常用的模型,包括短语翻译模型、词汇化翻译模型和语言模型,最小上下文构成代价模型考虑了一元词素上下文环境和词缀 N-gram 上下文环境。实验结果显示基于短语统计机器翻译形态切分模型对词表词切分,最小上下文构成代价模型对未登录词处理后,总体的切分准确率达到 96.94%。

2016 年 Manaal Faruqui 等提出基于图模型的半监督学习方法[20],利用词之间的句法和语义关系,从小的种子词汇集自动构建广泛覆盖的词典,这个词典提供了形态标签和依存句法分析功能。这种半监督学习方法是不依赖于语言的,在作为黏着语类型的芬兰语和匈牙利语的实验中,芬兰语的 F_1 值为 71.9%,匈牙利语的 F_1 值为 79.7%。

有监督的统计学习方法具有以下优点:(1)基于坚实的数学理论,提出了有效的消歧方法;(2)充分利用语料库知识,提供更多基于统计的实例化模型;(3)基于训练语料,可以学习到有效的语言学规律;(4)具有一致性、健壮性好的特点。能够处理 OOV(未登录词)以及不规则词形变化等问题。其中基于最大熵和基于条件随机场的方法将词干切分看作是序列化标注问题,能够加入更多语言本身所具有的特征,体现不同构形成分之间的不同,有利于词干切分正确率的提升。

1.3 基于无监督学习的统计方法

2002 年 Mathias Creutz, Krista Lagus 提出了基于无监督的方法构建词干切分模型[21],首先利用最小描述长度方法(minimum description length, MDL)获得词干切分模型,然后利用极大似然方法(maximum likelihood, ML)优化词干切分模型对目标语言的切分,得到基于统计获得的类似于词干和附加成分的子词。并基于此开发了基于数据驱动的 Morfessor 开源工具。Morfessor 的 MDL 切分同时很好地处理了切分歧义和 OOV 切分问题。

基于无监督学习的统计方法的不足之处是由于黏着语具有形态丰富,词缀数量大和词缀有层叠现象,导致无监督学习方法切分精度较低,无法满足实际需要。

1.4 小 结

通过以上基于词典和词干切分规则的方法、有监督的统计方法和无监督统计方法这三种词干切分方法的比较,可以看出每种方法都有各自的特点。第一种方法对人工的依赖较大,同时由于词干切分存在歧义和兼类现象,所以基于切分规则的方法很难正确的切

分。无监督的统计方法具有语言无关性,不需要标注语料等优点,但因为黏着语具有形态丰富,词缀数量大和词缀有层叠现象,导致无监督方法切分精度较低,无法满足实际需要。

因此为了减少对人工因素的依赖,利用已有的标注语料,同时结合蒙古和维吾尔文基于统计的词干切分方法分析,文中提出了一种哈萨克语词干切分的基于统计学习的方法。

2　基于统计学习的哈萨克语词干切分

2.1　问题描述

词干切分的问题可形式化描述为序列标注问题。基于统计学习的哈萨克语词干切分方法,将每个单词作为字符串序列进行按字符标注,从而得到标注序列,这一标注序列对应该单词的一个词干切分。为方便统计学习方法处理,将哈萨克语转换为标准化哈萨克语拉丁字符表示。

例如:哈萨克语单词“merekedeg1”(节日中的),“mereke”为名词词性的词干,“deg1”是一个构形附加成分,则单词“merekedeg1”的一个词干切分所对应的标注序列就是“SBSISISISISEBIIE”,其中标记“SB”表示词干的首字母标识,“SI”表示词干的除首尾以外的其他字母标识,“SE”表示词干的尾字母标识,“B”表示构形附加成分的首字母标识,“E”表示构形附加成分的尾字母标识,标记“I”表示构形附加成分中除首尾以外的其他字母的标识。

文中对哈萨克语已标注好的语料,分别用最大熵模型和条件随机场模型对哈萨克语单词中每个字符进行标注。设 m 个字符组成的输入单词用 $W = c_1 c_2 \cdots c_m$ 表示,目标是输出一个相应的标识序列,用 $T = t_1 t_2 \cdots t_m$ 表示,则求解该单词所有可能的标识序列中最大概率值的序列值。

2.2　最大熵模型

最大熵模型(maximum entropy,ME)建立在最大熵理论基础之上,在序列标注问题中,设训练集样本用 (x,y) 表示,其中 x 表示单词字符序列信息的上下文,y 表示字符序列标注结果,根据已知的样本集合构建一个在已知上下文条件下,能够准确预测未知标注结果 y 的概率统计模型 $p(y \mid x)$。这一模型获得的概率分布应与训练集语料的经验分布相符。最大熵原理说明,在满足已知约束的情况下,x,y 的正确分布信息熵最大。按照这一原理构建的模型即为最大熵模型,形式化为:

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left[\sum_{i=1}^{k} \lambda_i f_i(x,y)\right] \tag{1}$$

其中,$Z(x) = \sum_{y} \exp\left[\sum_{i=1}^{k} \lambda_i f_i(x,y)\right]$ 为归一化因子,保证对所有可能的字符序列上下文 x,满足 $\sum_{y} p(y \mid x) = 1$,从而保证 $p(y \mid x)$ 为概率值;$f_i(x,y)$ 为特征函数;λ_i 为特征参数,是反映每个特征对于模型重要程度的权重;k 为特征函数的数目。已知约束条件通过特征函数来描述,一般情况下特征函数 $f(x,y) \rightarrow \{0,1\}$ 是一个二值函数,形式为:

$$f(x,y) = \begin{cases} 1 & \text{如果}(x,y)\text{满足某种约束} \\ 0 & \text{否则} \end{cases} \tag{2}$$

通过式(1)可知,对 $p(y \mid x)$ 概率的求解是通过对特征参数 λ_i 的求解来实现,一般采用迭代算法 GIS 和 IIS。

2.3　条件随机场模型

条件随机场模型(conditional random fields,CRFs)是常用于序列标注任务的概率模型。在中文分词、命名实体识别、词性标注等任务中取得了不错的效果。与隐马尔可夫(HMMs)模型相比,它不需要严格独立性假设,并可以很好地表示交叉特征和长距离依赖关系,还很好地解决了最大熵隐马尔可夫(MEMMs)模型标注偏置问题。对于序列标注任务常用的是链式 CRFs 模型,对于输入序列 x,对应标注序列的 y 的条件概率为:

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left[\sum_{c \in C} \sum_{k} \lambda_k f_k(c, y_c, x)\right] \tag{3}$$

$$Z(x) = \sum_{y} \exp\left[\sum_{c \in C} \sum_{k} \lambda_k f_k(c, y_c, x)\right] \tag{4}$$

其中,$f_k(c, y_c, x)$ 是一个布尔型的特征函数,$Z(x)$ 是一个归一化因子。

运用维特比算法,在给定一个输入序列 x 的条件下,可求解出观测序列最大化条件概率的标注序列:

$$\hat{y} = \arg\max_{y} p(y \mid x) \tag{5}$$

2.4　数据预处理

哈萨克语中的某些单词在构形过程中存在有形变现象,即词干或构形附加成分缀接其他构形附加成分时会发生其中字符的变化。如“qep”是“干”这一单词的词干原形,“ip”是一个构形附加成分,当“qep”词干后缀接构形附加成分“ip”时,“qep”会发生形变变成“qew”。所以“qep”和“ip”组成词的形式就是“qewip”。训练语料库中“qewip”对应的切分是“qew”和“ip”,而不是“qep”和“ip”。而在统计学习方法中,输入序列 x 与标注序列 y 一一对应。在训练语料中,如果词干和构形附加成分都是原型形式,则由于单词中存在的形变,对单词进行切分时就无法识别已经形变的词干与构形附加成分,从而无法获得正确的切分结果。

因此为了正确切分,需要将训练语料中处于原型的词干和构形附加成分对应转换为变形形式。同样在

切分后,需要将变形形式的词干和构形附加成分还原为原型形式。文中构建了 50 多条变形和还原规则,对数据进行互为逆的操作处理。

2.5 哈萨克语词的标注方法

选择一种合适的标注方式有助于序列标注任务的研究。结合其他文献的标注方式和哈萨克语构词特点,文中提出了一种标注方法。对单词的词干部分和构形附加成分部分采用不同的前中后标记模式。这样可以使统计学习方法在训练过程中针对词干和构形附加成分学习到有针对性的信息,从而提高词干切分性能。

文中使用的标注集“SBSISEBIE”如表 1 所示。实验结果表明,在特征函数不变的条件下,区分词干和构形附加成分的标注集比不区分的标注集在切分准确率上有显著提升。

表 1 “SBSISEBIE”标记集

标记符号	标记含义
SB	词干部分的首字母
SI	词干部分除首和尾字母以外的其他字母
SE	词干部分的尾字母
B	构形附加成分部分的首字母
I	构形附加成分部分除首和尾字母以外的中间字母
E	构形附加成分部分的尾字母

例如在前文中提到的“merekedeg1”对应的不区分词干和词缀的标注序列是“BIIIIIBIII”,而如果使用有区分的表 1 标记集,“merekedeg1”这个单词对应的标注序列就是“SBSISISISISEBIIE”,词切分系统从标注形式上就可以区别词干和构形附加成分。

2.6 特征函数的选择

对于统计学习方法最大熵模型和条件随机场模型,特征函数的选择至关重要。特征函数反映训练语料包含的统计规律,而合适的特征函数可以很好地表示这些统计规律。

哈萨克语中构形附加成分表现为若干字符相连的固定形式,从统计的角度观察,这些构形附加成分的固定搭配形式在训练集中出现频率较高。为了提取出这些固定搭配的相邻位置关系信息,构建具有相邻关系的特征函数。例如在特征函数中定义当前字母用 C_0 表示,当前字母的前一个字母用 C_{-1} 表示,当前字母的后一个字母用 C_1 表示。从而构建特征函数 $C_{-1}C_0C_1$ 来表示当前字母与前一个字母和后一个字母的位置关系。例如在单词“merekedeg1”中,选取当前字母为“k”,则特征函数 $C_{-1}C_0C_1$ 提取出特征“eke”。

在哈萨克语的构词规则中,某些构形附加成分与另一部分构形附加成分之间存在依赖关系,即一类构形附加成分的出现会对另一类附加成分的出现起到约束作用,表现为远距离依赖关系。这时设置间隔字符位置关系的特征函数来提取这一类特征。例如:特征函数 $C_{-4}C_{-3}C_3C_4$,表示当前字母左侧和右侧第 3 和第 4 个位置上字符之间的关系特征。

窗口长度表示一个特征函数包含的字符个数,通过实验结果观察,选择适合哈萨克语词干切分的相应窗口大小。表 2 列出了文中用到的部分特征函数的表示。

表 2 特征函数与单词中字母对应关系

特征	例词“$merekedeg_1$”
$C_{-1}C_0C_1$	“eke”
$C_{-2}C_{-1}C_0$	“rek”
$C_0C_1C_2$	“ked”
$C_{-2}C_{-1}C_1C_2$	“reed”
$C_{-4}C_{-3}C_3C_4$	“meeg”

2.7 后处理

文中构建了包含 436 个哈萨克语构形附加成分的词典库,用于监督词干切分系统可能对构形附加成分的错误识别。通过切分结果中的构形附加成分与该词典库中条目进行比对,确定是否正确切分。对切分系统按照 $p(y \mid x)$ 概率大小给出的 n-best 结果,依次重复比对过程,选择 n-best 结果中第一个与词典库对应匹配成功的切分结果为最终输出结果。

3 实 验

实验中的训练语料为 2008 年新疆日报(哈文版),其中包含 10 万个哈萨克语句子,约有 100 万哈萨克语词。同时使用 2009 年新疆日报(哈文版)和人民网(哈文版)的 500 个哈萨克语句子作为测试集,并人工编写了对应的标准切分结果。分别使用张乐博士的 maxent-master 实现最大熵模型和 Taku Kudo 开源工具 CRF++根据需要进行修改实现的条件随机场模型。

3.1 实验步骤和评价指标

文中用不同的标注集对训练集已切分语料进行标注,实验比较了不同标注集对词干切分效果的影响。颗粒度最大的是不区分词干和构形附加成分的 BI 标注集,颗粒度最小的是区分词干和构形附加成分的 SBSISEBIE 标记集。对最大熵模型和条件随机场模型实验对比了颗粒度最小的 SBSISEBIE 标记集,也在该标注集上测试了不同窗口大小对词干切分准确率的影响。采用了在序列标注任务中经常使用的准确率指标,定义如下:

$$prec = \frac{正确切分的单词数}{切分的单词数} \quad (6)$$

其中,切分的单元为词干或构形附加成分。

3.2　实验结果对比和分析

在统一窗口大小为 4 的情况下,给出不同标注集的开放测试实验结果,如表 3 所示。

表 3　不同标记集在开放测试的实验结果对比

标记集	准确率/%
BI	82.4
SBSIBI	84.2
SBSISEBIE	86.3

通过表 3 可以看出,选择颗粒库越小的标注集,切分结果越准确。区分词干和构形附加成分的标注集比不区分词干和构形附加成分的标注集有 2 个百分点左右的提升。文中在统计学习方法的实验中统一使用颗粒度小的区分词干和构形附加成分的标注集。

表 4　在不同窗口长度的实验结果对比

模型选择	窗口长度	(开放测试)准确率/%	(封闭测试)准确率/%
基于词典和规则		70	
最大熵模型	1	59.1	59.9
	2	65.0	65.5
	3	75.6	76.0
	4	79.8	80.7
条件随机场模型	1	59.7	60.0
	2	67.5	67.6
	3	80.0	80.3
	4	86.0	87.6

表 4 是在使用 SBSISEBIE 标记集的条件下,不同窗口长度的基于词典和规则方法、最大熵方法和条件随机场方法的词干切分实验结果对比。在这里需要说明的是,文献[1-2]的测试环境由于无法获得,因此第一种基于规则的方法和后两种基于统计学习方法的测试环境存在一定的差别,此处的数值比较只能作为参考。从实验结果可以看出,文中的最大熵方法和条件随机场方法在词切分准确率上比基于词典和规则的方法有显著的提升,其中条件随机场模型有了 15% 的性能提升。基于统计学习方法的哈萨克语词干切分方法显示出了很好的性能,在窗口长度从 1 到 4 的对比可以看到字符串的上下文信息对词干切分的影响显著。

特征函数的窗口长度越长,特征集中所包含的上下文信息越多,但同时随着窗口长度的增加数据稀疏问题就会越显著。模型的训练时间开销和生成的模型文件的规模也会随着窗口长度的增加而成倍增加。综合考虑窗口大小和时间空间开销,认为窗口大小为 4 是对哈萨克语统计学习方法词干切分比较适合的选择。

通过对词干切分实验结果的分析,发现对于哈萨克语单词中单个构形附加成分组成的单词切分准确率较高,但对由多个构形附加成分构成的识别准确率较低。这可能是在词中以字符为单位的字符上下文信息较少导致切分系统没有足够的信息做出正确判断造成的。对于这种类型的问题可能的解决方法是加入单词所在句子的上下文信息和该单词的词性信息进行判断。同时根据单词所处句子的上下文信息不同,存在两种或两种以上的切分形式,切分系统给出的都是正确的切分形式,但在当前的句子上下文环境中可能是错误的。解决这类歧义问题的方法就是引入更多以词为单位的上下文信息。这两类错误切分在文中所提到的以字符为单位的模型中无法完全解决。

4　结束语

哈萨克语词干切分问题在统计学习方法中属于序列标注任务的一种,通过分析哈萨克语单词构形上的特点,提出了一种基于统计学习方法的区分词干和构形附加成分的标注方法,其次利用机器学习方法中的最大熵模型和条件随机场模型对转化为序列化标注问题的哈萨克语词干进行切分,实验对比结果表明基于机器学习的方法能够提高哈萨克语词干切分的性能。

文中使用的哈萨克语词干切分方法主要以字符为单位,考虑了单词中字符之间的上下文信息,但没有加入单词所在的句子上下文信息和单词的词性信息。同时在实际语言环境中,哈萨克语单词会根据上下文语境的不同采用不同的切分方法。同时随着神经网络的兴起,在下一步的研究中会尝试使用深度学习神经网络方法和加入以词为单位的句子上下文信息和单词词性信息,来进一步提高哈萨克语单词的词干切分正确率,降低歧义性。同时利用词干切分的结果来影响词性标注的效果,从而进一步在哈萨克语词法分析应用中利用已取得的经验。

参考文献:

[1]　达吾勒·阿布都哈依尔,古丽拉·阿东别克. 哈萨克语词法分析器的研究与实现[J]. 计算机工程与应用,2008,44(19):146-149.

[2]　达吾勒·阿布都哈依尔,海拉提·克孜尔别克. 基于规则的哈萨克语词干提取算法研究[J]. 新疆大学学报:自然科学版,2011,28(2):238-241.

[3]　古丽拉·阿东别克,米吉提·阿布力米提. 维吾尔语词切分方法初探[J]. 中文信息学报,2004,18(6):61-65.

[4]　米热古丽·艾力,米吉提·阿不力米提,艾斯卡尔·艾木都拉. 基于词法分析的维吾尔语元音弱化算法研究[J]. 中

文信息学报,2008,22(4):43-47.
[5] 阿孜古丽·夏力甫.维吾尔语动词附加语素的复杂特征研究[J].中文信息学报,2008,22(3):105-109.
[6] 热娜·艾尔肯,李 晓,艾尼宛尔·托乎提.基于混合方法的维吾尔语词干提取方法研究[J].计算机应用研究,2015,32(1):112-114.
[7] 早克热·卡德尔,艾山·吾买尔,吐尔根·依布拉音,等.维吾尔语名词构形词缀有限状态自动机的构造[J].中文信息学报,2009,23(6):116-121.
[8] 史建国,侯宏旭,飞 龙.基于词典、规则的斯拉夫蒙古文词切分系统的研究[J].中文信息学报,2015,29(1):197-202.
[9] XUE N,CONVERSE S P. Combining classifiers for Chinese word segmentation[C]//Proceedings of the first SIGHAN workshop on Chinese language processing. Taipei:[s. n.],2002:63-70.
[10] TSENG H,CHANG P,ANDREW G,et al. A conditional random field word segmenter for SIGHAN Bakeoff 2005[C]//Proceedings of the fourth SIGHAN workshop on Chinese language processing. Jeju Island,Korea:[s. n.],2005:168-171.
[11] LIU Y,ZHANG Y,CHE W,et al. Domain adaptation for CRF-based Chinese word segmentation using free annotaions[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Doha,Qatar:ACM,2014:864-874.
[12] ZENG X,WONG D F,CHAO L S,et al. Graph-based semi-supervised model for joint Chinese word segmentation and part-of-speech tagging[C]//Proceedings of the 51st annual meeting of the association for computational linguistics (ACL). Sofia,Bulgaria:ACM,2013:770-779.
[13] AISHA B,SUN M. A statistical method for Uyghur tokenization[C]//IEEE international conference on natural language processing and knowledge engineering. Dalian,China:IEEE,2009:1-5.
[14] 薛化建,董兴华,王 磊,等.基于词缀库的非监督维吾尔语词切分方法[J].计算机工程与设计,2011,32(9):3191-3194.
[15] 赛迪亚古丽·艾尼瓦尔,向 露,宗成庆,等.融合多策略的维吾尔语词干提取方法[J].中文信息学报,2015,29(5):204-210.
[16] 侯宏旭,刘 群,那顺乌日图,等.基于统计语言模型的蒙古文词切分[J].模式识别与人工智能,2009,22(1):108-112.
[17] 赵 伟,侯宏旭,从 伟,等.基于条件随机场的蒙古语词切分研究[J].中文信息学报,2010,24(5):31-35.
[18] 姜文斌,吴金星,乌日力嘎,等.蒙古语有向图形态分析器的判别式词干词缀切分[J].中文信息学报,2011,25(4):30-34.
[19] 李 文,李 淼,梁 青,等.基于短语统计机器翻译模型蒙古文形态切分[J].中文信息学报.2011,25(4):122-128.
[20] FARUQUI M,MCDONALD R,SORICUT R. Morpho-syntactic lexicon generation using graph-based semi-supervised learning[C]//Transactions of the association for computational linguistics.[s. l.]:[s. n.],2016:1-16.
[21] CREUTZ M,LAGUS K. Unsupervised discovery of morphemes[C]//Proceedings of workshop on morphological and phonological learning of ACL. Philadelphia,Pennsylvania,USA:ACL,2002:21-30.

CCF 招聘秘书长

CCF 现任秘书长将于 2021 年 1 月 30 日到任,根据 2018 年 10 月 24 日理事会表决通过的《CCF 秘书长遴选聘任条例》的规定,现公开招聘新任秘书长。

CCF 秘书长作为学会的首席执行官,应具备如下条件:

1. 中国公民,具有在计算机领域的从业经历,对社团有深刻的理解并熟悉运作;
2. 有领导力、运营能力和管理经验;
3. 公正、守约、廉洁;
4. 被聘用后能全职从事学会工作,且在 62 岁前至少能任职四年。

欢迎自荐和推荐,联系 ccf@ ccf. org. cn 索要推荐申请表。

截止日期:2020 年 6 月 30 日

登录学会网站(https://www. ccf. org. cn/c/2020-01-17/694649. shtml),查看《CCF 秘书长遴选聘任条例》。

中国计算机学会

2020 年 3 月 1 日