维吾尔语多词表达抽取方法研究

麦热哈巴·艾力[1,2],阿孜古丽·夏力甫[3],吐尔根·依布拉音[1,2]
Mairehaba Aili[1,2], Aziguli Xialifu[3], Tuergen Yibulayin[1,2]

1.新疆大学 信息科学与工程学院,乌鲁木齐 830046
2.新疆多语种信息技术重点实验室,乌鲁木齐 830046
3.新疆大学 人文学院,乌鲁木齐 830046
1.School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China
2.Xinjiang Laboratory of Multi-Language Information Technology, Urumqi 830046, China
3.School of Humanity, Xinjiang University, Urumqi 830046, China

Mairehaba Aili, Aziguli Xialifu, Tuergen Yibulayin. Research on extracting methods of multi word expression in Uyghur texts. Computer Engineering and Applications, 2014, 50(8):26-30.

Abstract: Multi word expression is a special language phenomenon, which is combination of words. As a block of meaning, multi word expression appears together more often than by chance. They play more important role in natural language processing applications. In this study, it explores the effect of three more used statistical methods on extracting multi word expression in Uyghur texts. The three methods contain mutual information, log-likelihood and chi-square. According to the characteristics of Uighur, it adds stemmed form of words as features of extraction methods. On the choosing corpus, it considers the coverage and field, and explores its effect on extraction methods.
Key words: collocation; mutual information; log-likelihood; chi-square; Uyghur

摘 要: 多词表达是特殊的语言现象,一般由多个词构成来表示一个意义,语料中常出现在一起。多词表达因是特殊的单元,其抽取在自然语言处理的很多领域有着非常重要的作用。讨论了目前常见的三种统计方法即互信息、对数似然比以及卡方等在维吾尔语多词表达抽取方面的影响。根据维吾尔语的特点,将词干作为一项特征加到抽取方法中。语料的选择上考虑了覆盖面及领域,并探讨了它们对抽取方法的影响。
关键词: 多词表达; 互信息; 对数似然比; 卡方; 维吾尔语
文献标志码: A **中图分类号:** TP391 doi:10.3778/j.issn.1002-8331.1309-0439

1 引言

多词表达自动抽取研究是自然语言处理中较难处理的部分。多词表达的自动抽取对文本分类、信息检索、机器翻译、词典编纂等起到举足轻重的作用。目前为止,多词表达的自动抽取方法主要集中在规则、统计[1-2]和混合[3-4]等方法中。规则方法主要是根据多词表达的结构特征建立规则库,其特点是直观、简单但对语言规则依赖性强,容易出现歧义;统计方法主要利用多词表达中各词的同现概率以及各词之间的紧密度等特征来抽取。常见的有点互信息(Mutual Information)[5]、卡方(Chi-square)、t检验(t hypothesis)以及对数似然比(Log-likelihood)等。混合方法在统计方法得到的多词表达候选上设置一些语言规则,过滤非多词表达降低其噪音,提高正确率。

维吾尔语是典型的粘着性语言,具有丰富的形态变化,维吾尔语中多词表达不仅有复合构词(composition)的形式,还有非复合构词(non-composition)的形式。维吾尔语多词表达自动抽取方法的研究将对维吾尔语信

基金项目: 国家自然科学基金(No.61262061);新疆多语种信息技术重点实验室开放课题。
作者简介: 麦热哈巴·艾力(1973—),女,在读博士,讲师,研究领域为自然语言处理、机器翻译;阿孜古丽·夏力甫(1974—),女,副教授,研究领域为维汉对比研究、计算语言学;吐尔根·依不拉音(1958—),男,通讯作者,教授,博士生导师,主要研究领域为自然语言处理、社会计算。E-mail:marhaba@xju.edu.cn
收稿日期: 2013-09-27 **修回日期:** 2013-11-21 **文章编号:** 1002-8331(2014)08-0026-05
CNKI **网络优先出版:** 2013-12-10, http://www.cnki.net/kcms/doi/10.3778/j.issn.1002-8331.1309-0439.html

息处理起到非常重要的作用。目前为止，国内外很多语言，包括：英语、汉语、德语、法语、俄语、印度语等，都有多词表达自动抽取方面的研究，但对维吾尔语多词表达自动抽取的研究甚少。本文尝试了用现有的多个多词表达抽取方法从维吾尔语文本中抽取了维吾尔语中的多词表达形式、并分析了遇到的问题以及解决的方法。

2　相关工作

Church和Hanks首次引入了互信息方法到多词表达的抽取研究[5]，Pecina[6]在对德语的Adj-N和PP-Verb多词表达进行排列时使用55种相关方法，最后得出采用多个抽取方法的混合使用比单个方法得到更好的结果。Chang[7]测试出在众多多词表达抽取方法中互信息(MI)和对数似然比(LLR)的效果好于其他方法。以上方法中研究者主要采用多词单元同现概率作为评价标准。然而仅仅使用同现概率对多词表达的抽取远远不够，因为这种方法抽取的多词表达候选中还包含着大量的非多词表达。于是，研究者尝试了将语言特性与统计信息结合的方法。Piao[8-9]等提出多词表达同现相关性模型，此模型中利用上下文词汇的统计搭配信息抽取多词表达。Caseli[10]在识别英语动词时态结构以及德语Adj-Noun(形容词-名词)结构时引入了词性标注等语言信息后得到了远比单独使用统计方法好得多的结果。

3　维吾尔语形态和多词表达特点

维吾尔语属于阿尔泰语系突厥语族，是典型的粘着性语言。维吾尔语的词干缀接不同的词尾，生成丰富而复杂的形态，表示多种语法意义。例如：kitabi(他(们)/她(们)的书)，kitabim(我的书)，kitabinglardiki(你们书里的)，都是在词干kitab(书)的后面接不同的词尾后形成的。维吾尔语中，词尾类型和数目非常多，同一个词干可以缀接多个词尾，且可以是多层缀接。由于维吾尔语词尾也带有语义信息，有时出现一个维吾尔词对应一段甚至是一句汉语(或其他)的情况。比如：ölchem-leshtürelmemsiler?(你们不能使它标准化吗?)就是在词干ölchem(标准)后缀接+lesh，+tür，+el，+me，+m，+siler等词缀后形成的。

另外，维吾尔语词干在缀接不同的词尾时因需遵循语音和谐规律，从而出现一些音变现象，包括：弱化、脱落及曾音。因为篇幅有限，本文不详细阐述，可以参考文献[11-13]。

维吾尔语中多词表达也非常多见，其形式多样，总结起来可以包括以下几种：

(1)习语

习语主要特点是其包含的单词不能被替换，几乎每个单词都可以出现形态变化，语义往往跟各词的实际意思有所区别。比如：

közge kir_(打扰)

原词意义：köz+ge(n.，望眼睛)　kir(v.，进)

ichini tingsha_(默默无声)

原词意义：ich+i+ni(n.，里面)　tingsha(v.，听)

bashni ye_(该死的)

原词意义：bash+ni(n.，头)　ye_(v.，吃)

一般，习语中各部分可以出现多种不同的形态，例如：以上第一个例子可以为közumge kiriwaldi (打扰我了)，közunglargha kiriwalmay_(不打扰你们…)；第二个例子也有beshini yeydighan(该死的(指第二个人))，beshimni yemsen?(你要让我死吗?)等。

(2)谚语

维吾尔语谚语具有结构固定、包含的词数较多等特点。其中各词不能被替换，组成词一般很少有形态变化，而形态变化主要出现在最后一词上，含义远比各组成词更深、更广。例如：

aghzida külke chaxchax，qoynida palta pichaq(口蜜腹剑)

原词意义：嘴上讲着笑话，怀里揣着刀剑

oshke janggal chüsheydu，tohu danggal(chüsheydu)(各有所思)

原词意义：山羊做着草原的梦，公鸡做着饲料的(梦)

(3)对偶词

对偶词是维吾尔语中最常见的一种结构，一般有两个词构成，大部分情况两个词之间有个连接符“-”，但也有不带连接符的情况。对偶词有多种组织形式：从组成词都具有语义到组成词都不具有语义的形式；灵活的形态变化：从组成词都没有形态变化到组成词都有形态变化等。汉语中的重叠词在维吾尔语中也算是对偶词的一种，本文就不详细论述了。对偶词的形式可以形式化为：$w_1\ w_1, w_1\ w_1+af, w_1+af\ w_1, w_1+af_1\ w_1+af_2, w_1\ w_2, w_1+af_1\ w_2, w_1+af_1\ w_2+af_2, w_1\ w_2+af_2$ 等。其中 w 表示词干，af 表示词尾，不同小标表示不同的词或词尾。如：

chong-kichik(大大小小)

原词意义：chong(大)　kichik(小)

aldi-arqisi(arqa+si)(周围)

原词意义：aldi(前)　arqa(后)

bir birlep(一个一个)

原词意义：bir(一)

qoghun-tawuz(瓜果)

原词意义：qoghun(甜瓜)　tawuz(西瓜)

(4)复合词

维吾尔语中复合词也比较常见，它通过两个或两个以上的词构成表示一个语义。复合词主要有名词、动词、形容词和数词构成，可形式化描述为：$n_1\ n_2, a\ n, m\ n, n\ v, a\ v, n\ a, a\ a, v\ v$ 等，其中 n 是名词，a 是形容词，m 是数词，v 是动词。复合词的形态往往出现在后面一词

上,但 v v 结构时两个词都有词形变化的情况较常见。如:

aq kongül(心底善良)

原词意义:aq(白色) kongül(心)

tereqqiy qil_(发展,v)

原词意义:tereqqiy(发展,n) qil(做)

Körüp baq_(看看,v)

原词意义:kör+üp(看) baq(助动词)

(5)其他

维吾尔语命名实体也是由多词表达构成,包括人名、地名和机构名、时间词、数词等。这些结构组织较灵活,重复性较少,一般采用规则的方式识别优越于统计方法。由于本文尝试通过统计的方式识别多词表达,以上实体名识别暂未考虑在本次讨论中。

4 多词表达抽取方法介绍

多词表达抽取方法最常见的有共现频率、点互信息、假设检验、t检验、卡方和对数似然比等,本文根据国内外众多研究者目前为止对这些方法的研究与总结,从中选择点互信息、卡方和对数似然比作为维吾尔语多词表达的抽取方法。

(1)点互信息(Mutual Information,MI)

点互信息(Church & Hanks 1990)主要是测量两个变量之间的相互依赖性,认为两个词的共现频率越高,则之间的紧密性越高。如果以 w_1, w_2 依次表示两个词,则互信息即可表示为:

$$MI(w_1, w_2)=\mathrm{lb}\frac{P(w_1, w_2)}{P(w_1)\cdot P(w_2)} \tag{1}$$

如果 w_1, w_2 相互独立,则两个词同时出现的概率即可等于每个词单独出现的概率 $(P(w_1, w_2)=P(w_1)\cdot P(w_2))$,则式(1)的结果为0,表示 w_1, w_2 不能构成多词表达。可以看出,两个词的点互信息值越高,则构成多词表达的可能性越高。

(2)卡方 (χ^2) 检验(Chi-square,CHI)

χ^2 检验又叫做皮尔逊卡方,通过对比表中观测频度和期望频度,以验证是否独立。当他们之间的差别很大时,可以否定观测值互为独立的假设。χ^2 检验理论上适合任何大小的表,但对于 2×2 其公式可简化为式(2)。

$$\chi^2=\frac{N(a\cdot d-b\cdot c)^2}{(a+b)\cdot(a+c)\cdot(b+d)\cdot(c+d)} \tag{2}$$

其中,N 为语料规模。

表1 联列表

	w_1	$\neg w_1$
w_2	$w_1w_2=a$	$\neg w_1w_2=b$
$\neg w_2$	$w_1\neg w_2=c$	$\neg w_1\neg w_2=d$

(3)对数似然比(Log Likelihood Ratio,LLR)

似然比[14]的值表示一个假设的可能性比其他假设大多少。多词表达抽取中可设置两个可选的假设,即

假设1:$P(w_2|w_1)=p=P(w_2|\neg w_1)$

假设2:$P(w_2|w_1)=p_1\neq p_2=P(w_2|\neg w_1)$

假设1表示二元组 w_1, w_2 的出现互为独立,即 w_2 的出现与 w_1 的出现没关系,称为独立性假设;假设2表示二元组 w_1, w_2 是一个整体,称为非独立性假设。实用最大似然估计法计算 p, p_1, p_2:

$$p=\frac{c}{N}, p_1=\frac{c_{12}}{c_1}, p_2=\frac{c_2-c_{12}}{N-c_1}$$

其中,c_1, c_2, c_{12} 分别代表 w_1, w_2 和 w_1w_2 的出现次数;N 是语料总规模。在二项式分布的假设下,对数似然比 λ 的计算公式为:

$$\begin{aligned}\mathrm{lb}\,\lambda=\mathrm{lb}\frac{L(H1)}{L(H2)}=\mathrm{lb}\frac{b(c_{12};c_1,p)b(c_2-c_{12};N-c_1,p)}{b(c_{12};c_1,p)b(c_2-c_{12};N-c_2,p)}=\\ \mathrm{lb}\,L(c_{12};c_1,p)+\mathrm{lb}(c_2-c_{12};N-c_1,p)-\\ \mathrm{lb}\,L(c_{12};c_1,p)-\mathrm{lb}\,L(c_2-c_{12};N-c_2,p)\end{aligned} \tag{3}$$

其中,$L(k;n,p)=x^k(1-x)^{n-k}$。Mood(1974)指出 $-2\,\mathrm{lb}\,\lambda$ 渐近 χ^2 分布,则若 $-2\,\mathrm{lb}\,\lambda$ 的值小于 χ^2 值,则接受两个观测值互为独立的假设;否则,接受被观测对象为一个多词表达的假设,其值越高,被观测对象成为多词单元可能性更高。

同时,LLR对稀疏数据也很有效,所以同现次数并不高的多词表达也可被这个方法检测出来。

以上介绍的公式都是通过统计同现率来计算两个词之间的相关程度,文献[15]对其进行了详细的说明。这些方法都有各自的优缺点,比如,LLR更适合于数据稀疏情况,而互信息对出现频率相差太大的情况不具有可比性。本文采用单个方法抽取多词表达的同时比较和分析了几种方法的不同结合对抽取的影响。

5 维吾尔语多词表达抽取过程

维吾尔语多词表达中词数少则2个,多则达到10个(如:谚语)等不同,但词数越多,语料中出现的频率就越低、数据稀疏较突出,而词数为2的情况最为常见。针对这个问题,本文主要以识别词数为2的多词表达为主,其余的情况暂时忽略。

(1)预处理

首先对语料做一些必要的预处理,比如:过滤非文本信息,不同编码的维吾尔文本使其转换成Unicode标准编码。像其他语言类似,维吾尔语中也有一些使用频率较高的词,如连词、代词、虚词等,根据维吾尔语词的使用频率以及其意义构造了停用词表,由将近300多个符号及词构成。根据多词表达一般不包括标点符号、数字、连词等特点,将语料按停用词来分句。

(2)维吾尔语多词表达的抽取

本文尝试了以上介绍的三种多词表达抽取方法即:互信息、对数似然比以及卡方,以及它们的不同组合,每

种方法都设置阈值，只有大于阈值的候选多词表达才被看成是“合理的”多词表达并被接受，而阈值是根据多词实验后被确定，本文使用的阈值设置为0.75。

(3)维吾尔语形态的引入

由于维吾尔语形态变化的特点，统计时不难出现同一词的不同形态，如，güle qaq(干果)一词和di'agnoz qoy(诊断)为例，语料中统计了它们不同形态出现的次数，如表2。

表2 同一个词不同形态的统计

word	w_1	w_2	w_1w_2
güle qaq*lar*	82	7	3
güle qaq*tin*	82	6	2
güle qaq*qa*	82	8	2
güle qaq	82	105	34
güle qaq*ni*	82	45	24
güle qaq*ning*	82	21	6
di'agnoz qoy*ushtiki*	236	10	7
di'agnoz qoy*ushta*	236	26	16
di'agnoz qoy*ghili*	236	29	14
di'agnoz qoy*ulup*	236	61	27
di'agnoz qoy*ush*	236	245	62
di'agnoz qoy*ushqa*	236	55	11

表2中可以看出，由于形态的不同，同一个多词表达会出现多个不同的形式，使得其同现词数下降，从而影响同现概率。对于以上问题，可采用将原词用词干来替代的方法解决，但通过维吾尔语多词表达的性能特点的介绍可得知：多词表达中各组成词的形态有时是非常必要的，若用词干替代，则避免不了很多噪音的掺入。如：[kozumge kirwelishidin] xoymu zeriktim.(真受不了他这样烦我)这一句中，方括号中的内容为多词表达(习语)，其中词尾-ge是必要的词尾；[kozumdin kirip] ketken nersini chiqiralmaywatimen(我无法将进到眼睛的东西取出来)这一句中，方括号中的内容不是多词表达，但用词干来代替，则两个方括号中内容变为同一种形式。可以看出直接用词干代替原词会出现将很多根本不是多词表达的内容因为丢失词尾而被统计成“正确”的多词表达的情况。

针对以上问题，本文采用折中的方法即：第一步，用词级别统计出多词表达候选；第二步，对候选列表进行词干提取后，再次聚合并统计。

6 实验及分析

6.1 语料

实验中准备了3个不同规模、不同领域的语料，目的是为了测试语料规模及领域对不同抽取方法的影响。语料规模、领域特点等信息参考表3。

6.2 评价标准

多词表达自动抽取常用的评价标准为精确率和召回率。为计算精确率和召回率需要手工标注语料中多词表达，而手工标注语料中所有多词表达本身就费时费力又很艰难。为了自动进行评价并减少人工标注时的偏见性和不一致性，文献[16]等采用在较小规模语料中计算精确率和召回率的方法。本文采用了类似的方法，具体为：

表3 语料信息

语料	领域	句子数	词数	标记数(token)
语料1(A)	新闻	7万	1 423 068	1 644 150
语料2(B)	文学	4万	458 940	597 220
语料3(C)	日常用语	1万	236 825	271 579

步骤1 准备一定规模的句子(本文选了500条句子)。

步骤2 请语言学家手工标注这些句子中出现的多词表达并作为标准答案保存。

步骤3 将这500条句子参合到以上三个规模的语料中。

步骤4 通过某种统计方法抽取语料中候选多词表达。

步骤5 计算$S1$包括的句子中被自动识别出的正确的多词表达数量。

精确率P(Precision)、召回率R(Recall)及F值分别用以下方式来计算：

$$P=\frac{\text{正确识别的多词表达个数}}{\text{标准答案中多词表达个数}}$$

$$R=\frac{\text{正确识别的多词表达个数}}{\text{被识别多词表达的个数}},\ F=\frac{P+R}{2}$$

6.3 标准答案的准备

为了评价多词表达“合理性”，请三个语言学家，以词与词之间的紧密性、实用性和结构固定性作为标注，从语料中抽取的500个句子进行手工标注。第一次标注后，比较三个答案，抽取了两个以上标注者认为“合理”的多词表达并作为正确答案；对有争议的部分(某一个标注者认为合理，而其他标注者没有选择)，再次让标注者讨论其合理性，得到两个票数的多词表达再被收集；对于一个标注者标为正确，而其余两个标注者认为错误的多词表达，作为错误的搭配，没被使用。

6.4 实验结果

三种方法以及它们不同的组合在三种规模的语料上识别多词表达的结果为表4所示。分析表中数据可得到以下几点信息：

单独使用三种方法时，每一种语料中，LLR和MI的效果远比卡方的效果好，其中LLR的F值为最高。使用两种方法的混合时，发现LLR+MI的F值最高。CHI与其余两种方法的结合普遍不高，甚至MI和CHI的结合比单独使用更低。这主要是CHI单独使用时的低识别率带来的影响，这个影响也体现在三种方法的混合使用中。

语料规模对各种方法的影响有一个共同点：虽然语

表4 实验结果 (%)

方法	语料A			语料B			语料C		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
LLR	48.62	52.71	50.67	54.43	55.91	55.17	43.21	46.82	45.02
MI	45.66	53.21	49.43	50.72	53.47	52.09	41.18	43.56	42.37
CHI	40.13	41.83	40.98	43.24	43.32	43.28	39.69	40.02	39.86
LLR+MI	50.23	53.76	52.01	58.36	60.02	59.19	54.55	57.78	56.17
LLR+CHI	40.24	42.41	41.33	46.84	48.26	47.55	42.70	44.52	43.61
MI+CHI	37.71	39.01	38.36	40.62	41.22	40.92	38.04	39.65	38.85
LLR+MI+CHI	39.42	40.98	40.20	42.30	43.55	42.93	41.48	43.09	42.29

料规模从大到小的排序为A-->B-->C,但是实验数据表明,同一种方法中,语料B的结果大于其他两种语料。对语料以及被测试的语料内容进行比较研究后发现,语料A是新闻领域语料,而被测句子集是日常用语。新闻语料中出现习语等特定多词表达的频率不高,从而导致虽语料规模大,但抽取的多词表达正确率不高的“怪”现象。

不管哪一种方法,不难看出维吾尔语多词表达的抽取的 *F* 值普遍较低,最高达到了59.19%,远远不能满足实际需求。分析原因,目前本文使用的抽取方法主要是以统计方法为主,不加任何规则或语言现象。实际上,多词表达具有很多可利用的语言特征,比如:动词复合词的最常见的出现形式为v1+“P型副动词词尾”+v2等。如果将这些特征信息适当地融合到以上方法中,对提高多词表达抽取正确率具有积极的作用。这也是下一步进行研究的内容。

7 总结与展望

本文尝试了采用目前广泛用于抽取多词表达的三种统计方法即:互信息、对数似然比以及卡方等抽取维吾尔语中多词表达。实验数据表明,光用统计方法抽取多词表达不能满足实际需要,而维吾尔语中多词表达有一些可利用的特征,通过某种方法融合到统计方法中可以提高抽取率的正确性,这也是下一步研究的计划。同时,不仅语料规模对抽取结果有影响,语料的领域对结果的影响也不可忽略。为了提高抽取率,语料规模以及覆盖率必须是首先考虑的因素。

参考文献:

[1] 梁铭.基于英汉平行语料库术语词典的自动抽取[J].电脑知识与技术,2009,5(19):5081-5083.

[2] Metin S K,Karaoğlan B.Collocation extraction in Turkish texts using statistical methods[M]//Advances in natural language processing.Berlin:Springer,2010:238-249.

[3] 肖健,徐建,徐晓兰,等.英中可比语料库中多词表达自动提取与对齐[J].计算机工程与应用,2010,46(31):130-134.

[4] Xu R,Lu Q.A multi-stage chinese collocation extraction system[C]//Advances in Machine Learning and Cybernetics,2006:18-21.

[5] Church K W,Hanks P.Word association norms,mutual information,and lexicography[J].Computational Linguistics,1990,16(1):22-29.

[6] Pecina P.Lexical association measures[D].[S.l.]:Charles University,2008.

[7] Chang Baobao,Danielsson P,Teubert W.Extraction of translation unit from Chinese-English parallel corpora[C]//Proceedings of the 1st SIGHAN Workshop on Chinese Language Processing,2002:1-5.

[8] Piao S S,Rayson P,Archer D,et al.Comparing and combining a semantic tagger and a statistical tool for MWE extraction[J].Computer Speech and Language,2005,19(4):378-398.

[9] Piao S S,Sun G,Rayson P,et al.Automatic extraction of Chinese multiword expressions with a statistical tool[C]//Workshop on Multi-Word-Expressions in a Multilingual Context,2006:17-24.

[10] Caseli H M,Ramisch C,Nunes M G V,et al.Alignment-based extraction of multiword expressions[J].Language Resources and Evaluation,2009,44(1/2):59-77.

[11] 麦热哈巴·艾力,姜文斌,王志洋,等.维吾尔语词法分析的有限图模型[J].软件学报,2012,23(12):3115-3129.

[12] 麦热哈巴·艾力,姜文斌,吐尔根.依布拉音维吾尔语词法中音变现象的自动还原模型[J].中文信息学报,2012,26(1):91-96.

[13] 麦热哈巴·艾力,王志洋,吐尔根.依布拉音一种提高维汉词语对齐的方法研究[J].小型微型计算机系统,2012,33(11):2551-2555.

[14] Dunning T.Accurate methods for the statistics of surprise and coincidence[J].Computational Linguistics,1993,19(1):65-74.

[15] Evert S,Krenn B.Methods for the qualitative evaluation of lexical association measures[C]//Proceedings of the 39th Annual Meeting on Association for Computational Linguistics,2001:188-195.

[16] Wang Z,Chen Y.MWUs extraction based on continuous measurement of inter-word association with frequency adjustment[C]//2nd International Conference on Computer Research and Development,2010:647-651.