北京大学学报(自然科学版) 第 52 卷 第 1 期 2016 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 52, No. 1 (Jan. 2016)
doi: 10.13209/j.0479-8023.2016.023

# 基于词间关联度度量的维吾尔文本自动切分方法

吐尔地·托合提† 维尼拉·木沙江 艾斯卡尔·艾木都拉

新疆大学信息科学与工程学院，乌鲁木齐 830046; †E-mail: turdy@xju.edu.cn

**摘要** 提出一种基于词间关联度度量的维吾尔文本自动切分方法。该方法从大规模生语料库中自动获取维吾尔文单词Bi-gram及上下文语境信息，在充分考虑维吾尔文单词间结合规则的前提下，将相邻单词间的互信息、*t*-测试差及双词邻接对熵的线性融合作为组合统计量(dmd)，度量文本中相邻单词之间的关联程度。以dmd度量的弱关联的词间位置作为切分点进行自动切分，得到语义及结构完整的词串，而不仅仅是以空格隔开的单词。在大规模文本语料上进行的测试表明，该方法的切分准确率达到88.21%。

**关键词** 语义串；互信息；*t*-测试差；邻接对熵；单词结合规则

**中图分类号** TP391

# Uyghur Text Automatic Segmentation Method Based on Inter-Word Association Degree Measuring

Turdi Tohti†, Winira Musajan, Askar Hamdulla

School of Information Science and Engineering, Xinjiang University, Urumqi 830046; † E-mail: turdy@xju.edu.cn

**Abstract** This paper puts forward a new idea and related algorithms for Uyghur segmentation. The word based Bi-gram and contextual information are derived from large scale raw corpus automatically, and according to the Uyghur word association rules, the liner combinations of mutual information, difference of *t*-test and dual adjacent entropy are taken as a new measurement to estimate the association strength between two adjacent Uyghur words. The weakly associated inter-word position is taken as a segmentation point and the perfect word strings both on its semantics and structural integrity, not just the words separated by spaces, is obtained. The experimental result on large-scale corpus shows that the proposed algorithm achieves 88.21% segmentation accuracy.

**Key words** semantic string; mutual information; difference of *t*-test; dual adjacent entropy; word association rules

文本切分是自然语言处理中的第一步，也是关键的一步。采取何种方法及切分难易程度，在不同语言环境下有所不同，但最终目的是一样的，即从文本中获取能表达具体、完整语义的语言单元的集合。这些语言单元在很多情况下是突破词语概念界限的语义串[1−2]，是文本中上下文任意多个连续字符(字或词)的稳定组合，是结构稳定不可分割且语义完整的语言单元(如固定搭配、习语、对偶词等具有词汇意义和语法意义的模式串[3−4]、词组或短语[5]、复合词或领域术语[6]以及命名实体等)。

在文本中，句子可以表达完整、连贯、易于理解的语义，而语义串蕴含句子的关键信息。因此，用语义串表示文本，可以有效地刻画文本主题[7]，构造泛化能力更强、更紧凑的文本模型[8−9]，从而可以提高相关算法性能及文本处理效率。因此，如何识别语义串边界并完整获取，已成为文本挖掘领域中的关键问题[10−11]。

中文信息处理领域中，分词是研究历史最悠久的问题，经过多年的研究，中文分词已出现多种较成熟的技术和实用分词工具。但是，“文本海啸”的到来，对中文自动分词研究提出一系列新的课题，尤其是新词边界的正确识别及分词系统对开放环境

国家自然科学基金(61262062, 61163033, 61262063, 61562083)和新疆维吾尔自治区高校科研计划重点项目(XJEDU2012I11)资助
收稿日期: 2015−06−07; 修回日期: 2015−08−18; 网络出版日期: 2015−09−30

的适应性及健壮性的需求日益突出。

作为文本中上下文任意多个连续字符(字或词)的稳定组合，语义串是语义及结构完整的语言单元，其内部结合紧密，不可分割。因此，以相邻汉字之间的结合程度作为切分依据，或将它作为补充手段来消除歧义，在中文分词和新词识别方法中已起到很好的作用。孙茂松等[12]从大规模生语料中获取汉字二元信息，用互信息及 *t*-测试差的线性叠加值来衡量相邻汉字之间的结合能力，并设计了一种无词表及无指导学习的自动分词算法。王思力等[13]用双字耦合度和 *t*-测试差的线性叠加值来消除分词中的交叉歧义，但他们是从熟语料中获取二元模型。费洪晓等[14]分别用 *N*-gram、互信息及 *t*-测试 3 种统计量来判断双字构成词的可能性。王芳等[15]用互信息定量估计相邻两个基本词间的结合可信度，提出一种基于可信度的中文完整词自动识别方法。何赛克等[16]将字串邻接变化数(accessor variety)引入基于条件随机场的中文分词系统，提高了分词系统性能。基于词典的分词方法 mmseg 中，蒋建洪等[17]用互信息来度量并过滤非邻接词，使分词系统性能得到提高。

基于上述研究，本文提出一种基于词间关联度度量的维吾尔文本自动切分方法。所提方法接近于孙茂松等[12]和王思力等[13]的研究，但又有区别。首先，他们都是用两种基本统计量的线性融合作为组合统计量，度量相邻汉字之间的结合紧密度，目的是提高现有中文分词系统的精度。本文引入邻接对熵(dual adjacent entropy)，并将 *t*-测试差(difference of *t*-test)、互信息(mutual information)及邻接对熵的线性融合作为一个组合统计量 dmd，用以度量相邻维吾尔文单词之间的结合紧密程度。本文目的是从已分好词的单词序列(以空格隔开的词序列)中识别出最终的切分边界，从而获取文本中结构及语义完整的语言单元。除此之外，本文还将维吾尔文不同词性之间的结合规律作为一种规则，融入词间位置判断中，以便提高语义串识别精度。

## 1 维吾尔文分词及存在的问题

维吾尔文是突厥语族中的一个成员，又属于阿尔泰语系，是一种拼音文字，具有黏着语特性。从表面上看，维吾尔文词在文本中以空格与上下文隔开，因此，一直认为维吾尔文中不需要分词。在各种文本处理中也都以空格作为自然分隔符进行分词(简称空格分词)，以词为特征表征文本。

**例 1** 去北京的火车从哪个车站出发？

例 1 由 6 个词组成，经过词干复原处理后再进行空格分词，对应的中文也经 ICTCLAS 分词(省略了功能词)，得到的词序列(维吾尔文书写顺序为从右到左)如图 1 所示。

从分词结果上看，空格分词对以上句子是有效的，切分出来的词都能作为基本的语义单元运用。但对于以下几个新闻标题，这种分词结果几乎是错误的。

**例 2** 科学家研制出禽流感病毒。

**例 3** 首批全国政协委员抵达北京。

**例 4** 奥巴马连任面临就业问题挑战。

例 2~4 采用空格分词的结果见图 1。

根据一个词语的最小语义完整性，例 2 应该被分为5个词语(带下划虚线的串)，但是空格分词把句子分成 8 个词。然而，二词串②，④和⑤都是常用实词，是两个单词的稳定组合，不可分割。例 3 中的二词串①，③，⑥和例 4 中二词串③，④，⑥也都是结构稳定、语义完整的常用实词，不能以空格分开提取。

维吾尔文中能表达一个最基本的、具体而完整语义的语言单元，在很多情况下不仅仅是一个以空格隔开的单词，而是它与上下文若干个词的稳定组合。维吾尔文中能表达一个完整语义或者说在实际语料中能充当一个实词的串，可分为以下两类。

**定义 1** 单词语义串是一个维吾尔文单词，即一个无空格字母串，语义完整且独立运用，可用空格分词切分得到。比如，例 1 中以空格分割的都是单词语义串。

**定义 2** 多词语义串是若干个维吾尔文单词的稳定组合，并且满足如下条件：1)语义完整，在真实语料中充当一个实词，不能以空格分开；2)结构稳定，在大规模语料中具有较高的流通度，是独立运用的语言单元。

随着维吾尔文文本挖掘相关领域研究工作的不断深入及更广范围的开展，空格分词方法开始暴露出其潜在的缺陷和局限性，主要表现如下。

1) 在维吾尔文Web 搜索中，由于空格分词没有考虑切分单元的语义完整性和结构完整性，因此获取的单词难以在文本标引中发挥关键词的作用[18]。另外，空格分词还会导致组合歧义及交叉歧义的产生，并出现低查准率[19–20]。


**图 1　例 1~4 采用空格分词的结果**
Fig. 1　Segmentation results on Example 1–4

2)中、英文文本聚类和分类中，常用词特征来表征文本，聚类和分类效果也比较满意。同样以词特征表征文本并用性能最好的学习算法，维吾尔文文本聚类和分类效果却远不及中文和英文[21]。这是因为，文本中表示关键信息的语义串被空格分词拆分为与其语义完全不符的若干个字母串，因此不仅不能提取更具有表征能力的文本特征，反而提高了特征空间的维度，甚至导致大量冗余、不相关(噪音)甚至类间交叉特征的出现。冗余特征的存在会降低学习算法的效率，不相关特征(噪音特征)的存在会损害学习算法的性能[22]，类间交叉特征的存在会极大地降低聚类和分类准确率[23]。

除搜索、聚类和分类外，空格分词在机器翻译、主题词提取、维吾尔人名(名在前姓在后，以空格隔开)、地名、机构名等命名实体识别以及新词识别等文本处理过程中也会成为一个瓶颈。

## 2　词间关联度度量

本文的主要思路是，从第一个单词开始扫描待处理文本中的词序列，并用一个统计量 $S$ 去观察相邻单词间的结合程度。如果 $S>T$ ($T$ 为阈值)，则保留它们之间“连”的状态，否则插入一个分隔符(维吾尔文单词之间以空格隔开，因此本文以字符“|”作为分隔符)将它们分开，这时它们之间是“断”的状态。例如，一个有 $n$ 个单词的文本的词序列为 $W_1\ W_2\ W_3 \ldots W_{i-1}\ W_i \ldots W_{n-1}\ W_n$，则基于统计量 $S$ 的词间连、断判断如图 2 所示。最后以分隔符“|”进行切分，就得到文本中的所有语义串。其中，统计量 $S$ 是从大规模生语料中学习计算得出。

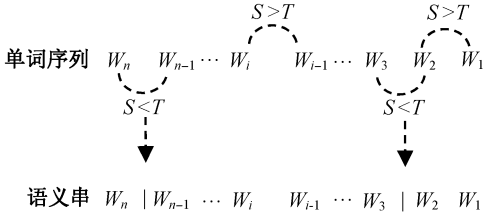


**图 2　基于统计量的词间位置连、断判断**
Fig. 2　Inter-word position judgment based on statistical measurement

本文所用语料都来自网络和正式出版物，包括从互联网收集的人工分类文本(共 20 类)、新疆日报 2008 年 3 和 4 月份全部内容、出版物 8 本(有关文学、社会、法制、经济等)。在实验和算法验证中，我们将大语料分为 3 个语料库，每个语料库都包含以上大语料中一定比例的内容。1) 生语料库 URC (Uyghur Raw Corpus)，共含维吾尔文单词及标点 9443290 个，未经标注；2)开发集 $USC_1$ (Uyghur Segmented Corpus)，共含维吾尔文单词及标点 15708 个，经人工标注；3)测试集 $USC_2$，共含维吾尔文单词及标点 154411 个，经人工标注(以上语料均由新疆大学智能信息处理重点实验室提供)。实验中，除对文本语料进行词干提取处理外，无任何特殊处理和人工干预，算法所需要的所有统计信息直接从生语料中获得。

在实验和算法验证中，我们用单词间的“连”和“断”的判断准确率 $\alpha$ 来调整阈值 $T$ 和其他参数，直到算法给出最好的性能。准确率$\alpha$的定义为

$$\alpha=\frac{\mathrm{PosCount}_{连}(W_{i-1},W_i)+\mathrm{PosCount}_{断}(W_{i-1},W_i)}{\mathrm{PosCount}(W_{i-1},W_i)}, \quad (1)$$

其中，$W_{i-1}$ 和 $W_i$ 是文本中相邻的词对，$\mathrm{PosCount}_{连}(W_{i-1}, W_i)$表示正确判断为“连”的词间位置数，$\mathrm{PosCount}_{断}(W_{i-1}, W_i)$表示正确判断为“断”的词间位置数，$\mathrm{PosCount}(W_{i-1}, W_i)$表示被处理文本中所有的词间位置数。

## 2.1 基本统计量：互信息(mi)

在一个维吾尔文文本以空格隔开的有序词序列中，$A$ 和 $B$ 是相邻的词对，则根据互信息原理，单词 $A$ 和 $B$ 之间的互信息可定义为

$$\mathrm{mi}(A,B)=\log_2\frac{P(A,B)}{P(A)P(B)}, \quad (2)$$

其中，$P(A, B)$为词对 $A$ 和 $B$ 在大规模语料库中出现的概率，$P(A)$为单词 $A$ 出现的概率，$P(B)$为单词 $B$ 出现的概率。假定它们在语料库中出词频分别为 count($A$)，count($B$)和 count($A$, $B$)，$n$ 是语料库中的词频总数，则有

$$\begin{cases}P(A,B)=\dfrac{\mathrm{count}(A,B)}{n},\\ P(A)=\dfrac{\mathrm{count}(A)}{n},\\ P(B)=\dfrac{\mathrm{count}(B)}{n}\end{cases} \quad (3)$$

互信息 mi($A$, $B$)反映相邻词对 $A$ 和 $B$ 之间的关联程度。mi($A$, $B$)越大，表明 $A$ 和 $B$ 之间的关联程度越紧密，如果 mi($A$, $B$)大于给定的一个阈值 $T_{\mathrm{mi}}$，则可以认为 $A$ 和 $B$ 之间是不可分割的。

我们以生语料库 URC 训练维吾尔文单词 Bi-gram 模型，并以 $USC_1$ 为对象考察互信息关于单词间连、断的分布情况。互信息变化范围在–6.75~21.01 之间，当阈值 $T_{\mathrm{mi}}$ 取 4.0 时(根据 URC 统计得到的 mi 均值为 3.63)，$\alpha$ 值最高可达 75.26%。例如，对于例 5 中各位置的判别基本上是正确的(如图 3(a)，“|”为分隔符)。

**例 5** 这种软件可以监测硬盘状态。

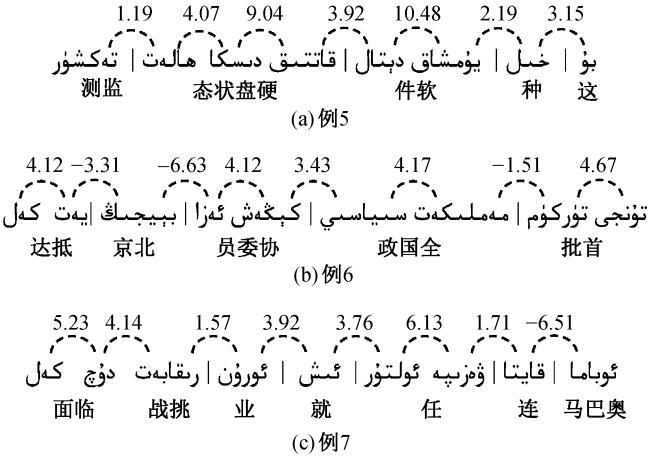


**图 3 对例 5~7 的 mi 切分结果**

Fig. 3 mi segmentation results on Example 5–7

从式(2)可以看出，互信息反应的是相邻词对 $A$ 和 $B$ 之间的静态结合能力，而不考虑它们的上下文，因此仅仅参考互信息这个基本统计量，也会出现错误的连、断判断。比如，对于例 6 和 7 的词间位置判断准确率较低(图 3(b)和(c))。

**例 6**　首批全国政协委员抵达北京。

**例 7**　奥巴马连任面临就业挑战。

## 2.2　基本统计量: *t*-测试差(dts)

Church 等[24]首次引入 *t*-测试来度量一个英文单词 $A$ 与其上下文单词 $x$ 和 $y$ 的结合紧密程度。根据定义，维吾尔文单词序列 $x$ $A$ $y$ 的 *t*-测试值计算公式如下:

$$t_{x,y}(A)=\frac{p(y\mid A)-p(A\mid x)}{\sqrt{\sigma^2(p(y\mid A)+\sigma^2(A\mid x))}},\qquad(4)$$

其中 $p(y|A)$和 $p(A|x)$分别为相邻词对($A$ $y$)和($x$ $A$)的 Bi-gram 概率，$\sigma^2(P(y|A))$和$\sigma^2(P(\mathrm{A}|\mathrm{x}))$分别是二者的方差。由式(4)可以看出，如果 $t_{x,y}(A)>0$，则 $A$ 与后继 $y$ 结合的强度大于与前趋 $x$ 结合的强度，此时 $A$ 应与 $x$ 断而与 $y$ 连；如果 $t_{x,y}(A)<0$，则 $A$ 与前趋 $x$ 结合的强度大于与后继 $y$ 结合的强度，此时 $A$ 应与 $y$ 断而与 $x$ 连；如果 $t_{x,y}(A)=0$，则 $A$ 与其前趋和后继的结合强度相等，无法判断 $A$ 与 $x$ 和 $y$ 的连断关系。

*t*-测试是基于字的统计量，而不是基于字间位置。为了能够在中文分词中直接计算相邻字间连断概率，孙茂松等[12]提出 *t*-测试差的概念。根据定义，对于维吾尔文单词序列 $x$ $A$ $B$ $y$，相邻单词 $A$ 和 $B$ 之间的 *t*-测试差值计算公式如下:

$$\mathrm{dts}(A,B)=t_{x,B}(A)-t_{A,y}(B)。\qquad(5)$$

当 dts($A$, $B$)>$T_{\mathrm{dts}}$($T_{\mathrm{dts}}$ 为阈值)时，相邻词对 $A$ 与 $B$ 之间的位置更倾向于判断为连，否则判断为断。我们仍以 $\mathrm{USC}_1$ 为对象，考察 *t*-测试差关于单词间连、断的分布情况。*t*-测试差变化范围在–264.14~108.41 之间，当阈值 $T_{\mathrm{dts}}$ 取 0.0 时，$\alpha$ 值可达最高为 78.14%。与 mi 相比，切分准确率较高，但对于例 5~7，各位置的判断与 mi 有所不同(图 4)。

## 2.3　基本统计量：邻接对熵(dae)

语义串作为频繁使用的语言单元，在真实文本中具有一定的流通度，能够应用于多种不同的上下文环境。因此，我们可以根据相邻两个单词上下文语言环境的复杂程度来衡量词对的结构稳定性。

对于维吾尔文有序单词序列 $x$ $A$ $B$ $y$ ($x$ 和 $y$ 是任何一个维吾尔文单词)，词对 $A$ 和 $B$ 在文本中每次出现的左邻接元素 $x$ 和右邻接元素 $y$ 构成一个邻接对<$x$, $y$>，那么 $A$ 和 $B$ 的所有邻接对组成邻接对集 $S_{\mathrm{dae}}=\{<x_i, y_i>\}$。$m$ 为集合中所有邻接对个数，$c$ 为集合邻接对种类数(不重复邻接对个数)，$n_i$ 为每个邻接对<$x_i$, $y_i$>的频次，则 $A$ 和 $B$ 的邻接对集合的信息熵(邻接对熵)的计算公式如下:

$$\mathrm{dae}(A,B)=-\sum_{i=1}^{c}\frac{n_i}{m}\log\left(\frac{n_i}{m}\right)。\qquad(6)$$

由式(6)可知，dae($A$, $B$)取值越大，词对 $A$ 和 $B$ 的语言环境变化越灵活多样，其内部结合越紧密；dae($A$, $B$)取值越小，$A$ 和 $B$ 的独立性越弱，很可能是

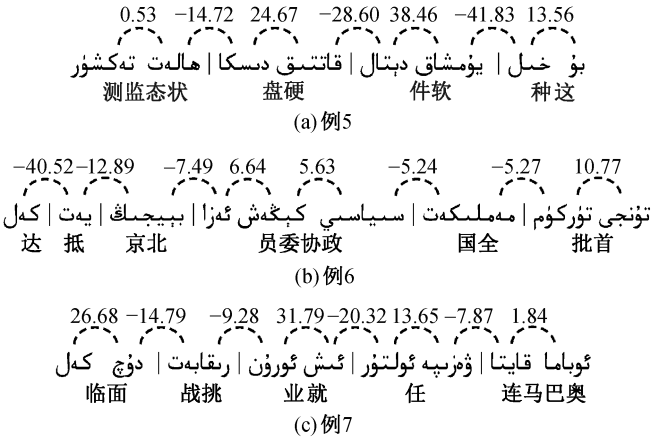


**图 4　对 5~7 的 dts 切分结果**

Fig. 4　dts segmentation results on Example 5–7

一种偶然性组合。因此，当 dae($A$, $B$)>$T_{\text{dae}}$($T_{\text{dae}}$为阈值)时，$A$ 和 $B$ 的单词间位置更倾向于判断为“连”，反之判断为“断”。

例如，三词语义串($A$ $B$ $C$)在语料库中共出现 5 次，其语言环境分别为 $x$ $A$ $B$ $C$ $y$，$y$ $A$ $B$ $C$ $x$,$z$ $A$ $B$ $C$ $x$，$w$ $A$ $B$ $C$ $y$，$v$ $A$ $B$ $C$ $x$，那么相邻词对 $A$ 和 $B$ 的邻接对集合 $S_{\text{dae}}$={<$x$, $c$>, <$y$, $c$>, <$z$, $c$>, <$w$, $c$>, <$v$, $c$>}，此时 $m$=5, $c$=5，因此 $A$ 和 $B$ 的邻接对熵为

$$\text{dae}(A,B)=-\frac{1}{5}\log\frac{1}{5}-\frac{1}{5}\log\frac{1}{5}-\frac{1}{5}\log\frac{1}{5}-\frac{1}{5}\log\frac{1}{5}-\frac{1}{5}\log\frac{1}{5}=0.699。$$

我们仍以 $\text{USC}_1$ 为对象，考察邻接对熵关于单词间“连”、“断”的分布情况。dae 变化范围在 0.06~1.37 之间，当阈值 $T_{\text{dae}}$ 取 0.60 时，$\alpha$ 值可达最高为 73.23%。

与 mi 和 dts 相比，dae 的分词准确率稍低，但它对新词词间位置的连、断判断更准确。例如，对于一个新出现的语义串 $A$ $B$，因为 $A$ 和 $B$ 是两个独立语言单位，在真实文本中会频繁使用，他们结合构成的特定新词 $A$ $B$ 的词频远远小于 $A$ 和 $B$ 的词频，会出现 count($A$)和 count($B$)极大而 count($A$ $B$)极小的情况。在这种情况下，mi 和 dts 几乎都会做出错误的判断，但 dae 中词频不是决定性因素，而是更多地考虑两个词上下文语言环境的变化多样性，因此能够做出正确的判断。

例如，新词“قۇش زۇكام”(禽流感)在生语料库 URC 中共出现 17 次，单词“قۇش”(禽)出现 2378 次，单词“زۇكام”(流感)出现 4927 次。因此，(قۇش, زۇكام)的互信息值为3.78, dts取值不均匀(–3.17~–0.48)，如果以 mi 或 dts 来判断词间位置的连断，是要断开的。但邻接对熵取值为 0.96，用 dae 判别词间位置是连接的。例如，对于例 8 中“زۇكام قۇش”的词间位置判断，mi 和 dts 都是错误的，只有 dae 的判断是正确的(图 5)。

**例 8** 科学家研制出禽流感病毒。

不论互信息、$t$-测试差或邻接对商，都是将词在语言环境中某一方面的信息特征作为计算依据，因此必然存在一定的局限性。中文分词中已有成功的案例，将基本统计量加以组合从而各取所长。我们分别用互信息、$t$-测试差和邻接对熵对 $\text{USC}_1$ 进行实验，发现将它们结合互补有较大的可行性。

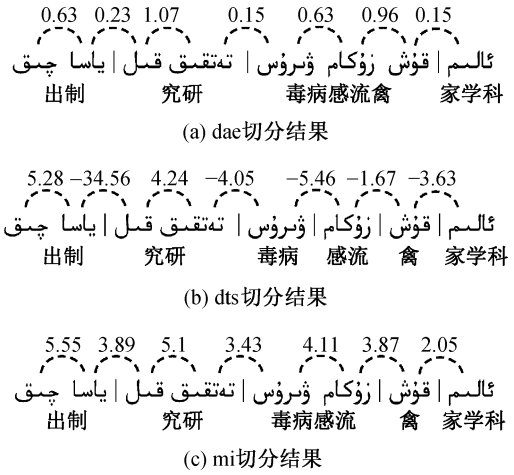


**图 5 对例 8 的 dae, dts 和 mi 切分结果**

Fig. 5 dae, dts and mi segmentation results on Example 8

## 2.4 组合统计量: dmd

我们单独用基本统计量对 $\text{USC}_1$ 进行词间位置判断，其中 $t$-测试差的 $\alpha$ 值最高(78.14%)，其次为互信息(75.26%)，最后是邻接对熵(73.23%)。因此，我们以 dts 为主，将 3 个基本统计量进行线性叠加，融合成一个组合统计量 dmd，并完全根据 dmd 来判断词间位置。由于以上基本统计量取值范围相差较大，$t$-测试差变化范围为–264.14~108.41，互信息变化范围为 –6.75~21.01，邻接对熵变化范围为 0.0~3.97，因此，线性迭加前先进行归一化处理，如式(7)~(9)所示。

$$\text{dts}^*(A,B)=\frac{\text{dts}(A,B)-\mu_{\text{dts}}}{\sigma_{\text{dts}}},\tag{7}$$

$$\text{mi}^*(A,B)=\frac{\text{mi}(A,B)-\mu_{\text{mi}}}{\sigma_{\text{mi}}},\tag{8}$$

$$\text{dae}^*(A,B)=\frac{\text{dae}(A,B)-\mu_{\text{dae}}}{\sigma_{\text{dae}}},\tag{9}$$

其中$\mu_{\text{dts}}$, $\mu_{\text{mi}}$ 和$\mu_{\text{dae}}$ 分别是 dts, mi 和 dae 的均值，实验值依次为–6.51, 3.63 和 0.52。$\sigma_{\text{dts}}$, $\sigma_{\text{mi}}$ 和$\sigma_{\text{dae}}$ 分别是 dts, mi 和 dae 的均方差，实验值依次为 24.29, 3.54 和 0.31。通过下式将它们叠加:

$$\text{dmd}(A,B)=\text{dts}^*(A,B)+\lambda\times\text{mi}^*(A,B)+\gamma\times\text{dae}^*(A,B),\tag{10}$$

其中，$\lambda$和$\gamma$的值经实验测定，发现当$\lambda$=0.35, $\gamma$=0.30 时的分词效果最好。dmd 在 $\text{USC}_1$ 上的变化范围为–11.5~6.9，当阈值 $T_{\text{dmd}}$ 取为 0 时，$\alpha$ 值最高(84.31%)，比单独使用 $t$-测试差、互信息或邻接对熵分别提高 6.17%, 9.05%和 11.08%。

### 2.5　基于规则的词间关联度度量

以组合统计量 dmd 判断词间位置的准确率达到 84.31%，但是与理想的准确率还存在一定的距离。我们从维吾尔文本身的语言特性中寻找有助于词间位置判断的信息，发现以下特性。

**特性 1**　维吾尔文中的助词(ئىدى、ئىكەن等)、连词(ياكى、بىراق等)、副词(بەك、ئاران 等)、量词(نەپەر、دانە 等)、代词(سەن、مەن等)以及感叹词(ئاھ、پاھ等)等功能词，在文本中始终不与其他单词结合成为语义串。本文将这些词称为“独立词”(independent word, IW)。

**特性 2**　维吾尔文单词间的结合主要是在名词(N)、形容词(ADJ)和动词(V)之间发生，并构成语义串。当形容词与名词或与动词结合时，形容词总是作为前驱，而不会出现在后继位置。因此，N+ADJ 或 V+ADJ 关系的相邻单词不可能结合构成一个语义串。

根据特性1和特性2，我们归纳出用于词间关联识别的单词结合规则(word association rule, WAR)，并定义如下。

**定义 3**　单词结合规则(WAR)：对于文本中的相邻词对“*A B*”，如 $A\in${IW}或 $B\in${IW}或 $B\in$ {ADJ}，则判断 *A* 与 *B* 不能结合成为关联模式，要断开。

因此，我们建立两个辅助词表：独立词表和形容词表，并用单词结合规则判断词间位置。这样，既减少了词间位置的 dmd 计算量，又明显提高了准确率。

## 3　基于词间关联度度量的切分算法

确定组合统计量 dmd 和单词结合规则后，基于词间位置判断的维吾尔文语义串识别及切分整体流程如图 6 所示。

对于训练语料，将所有的标点符号都替换为分隔符“|”，并进行词干提取处理，然后计算语料库中所有词对的 dmd 值，构建双词结合度(dmd)词典。对于待处理文本，进行同样的预处理(标点符号的替换以及词干提取)，然后依次提取词间位置(词对)，按以下步骤判断词间的相邻性。

1) 对于当前词对“*A B*”，如 $A\in${IW}或$B\in${IW}或 $B\in$ {ADJ}，则判断 *A* 与 *B* 断开，并插入分隔符“|”来消除 *A* 与 *B* 间的相邻性，否则转步骤 2。

2) 从双词结合度词典中读取词对“*A B*”的 dmd 值，如 dmd(*A*, *B*)>$T_{\mathrm{dmd}}$，则判断 *A* 与 *B* 连接并保留相邻性，否则插入分隔符“|”，消除 *A* 与 *B* 间的相邻性，转步骤 3。

3) 如“*A B*”是最后一个词对，则转步骤 4，否则提取下一个词对并转步骤 1。

4) 结束当前文本词间位置的判断。

对当前文本中所有词间位置判断结束后，以分隔符“|”进行切分，得到文本中所有语义串。算法流程如图 7 所示。

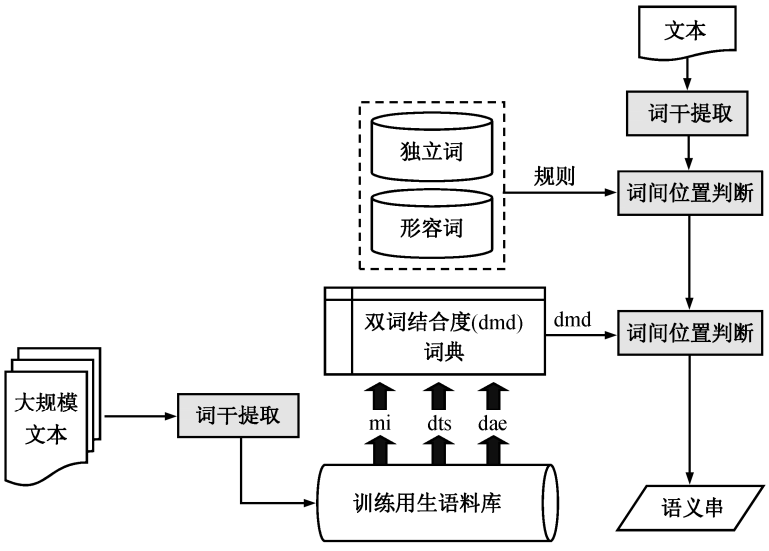


**图 6　维吾尔文语义串识别及切分整体流程**

Fig. 6　Uyghur semantic string recognition and segmentation process

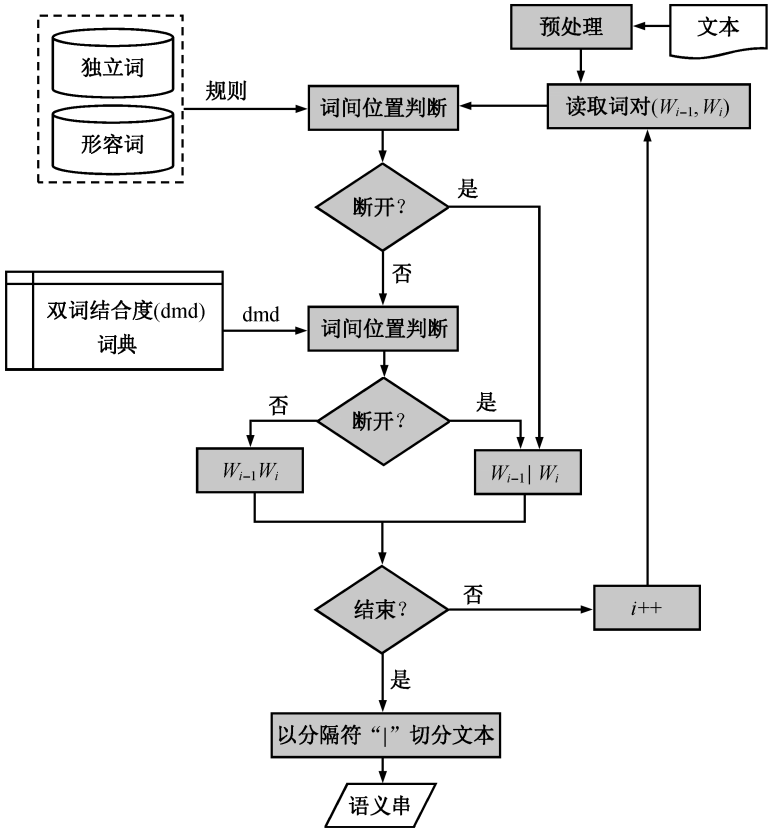


**图 7 维吾尔文语义串提取算法流程**

Fig. 7 Algorithm process of Uyghur semantic string extraction

## 4 实验与分析

我们基于生语料库 URC 得到维吾尔文单词(词干)统计模型，并构建双词结合度词典，以 $USC_1$ 为对象，用不同统计量判断词间位置的准确率并调整阈值，确定式(10)中的$\lambda$和$\gamma$来检验组合统计量 dmd 的有效性以及验证语义串提取算法在开放环境下的健壮性。因此，我们分别在开发集和测试集上进行词间位置判断实验，分析 dts, mi 和 dae 组合前和组合后词间位置正确判断情况。

以分界符“|”替换为所有标点符号后，开发集 $USC_1$ 和测试集 $USC_2$ 共含的维吾尔文单词及需要判断的词间位置如表 1 所示。

使用不同策略情况下的开发集和测试集实验结果如表 2 和 3 所示。

从测试结果看出，算法在测试集中的性能没有下降，表明本文提出的组合统计量 dmd 及各类参数的确定是有效的，尤其是引入语言特性的单词结合

**表 1 $USC_1$ 和 $USC_2$ 中单词及词间位置数**

Table 1 Words and inter-word positions in $USC_1$ and $USC_2$

| 数据集 | 单词个数 | 词间位置数 |
|---|---|---|
| 开发集($USC_1$) | 120948 | 108998 |
| 测试集($USC_2$) | 137757 | 125699 |

**表 2 开发集切分结果**

Table 2 Segmentation result in $USC_1$

| 使用策略 | 判断正确的词间位置数 | 关于词间连断的判断准确率$\alpha$/% |
|---|---|---|
| dts | 85174 | 78.14 |
| mi | 82024 | 75.26 |
| dae | 79819 | 73.23 |
| dmd | 91900 | 84.31 |
| dmd + WAR | 96211 | 88.27 |

**表 3　测试集切分结果**

Table 3　Segmentation result in $USC_2$

| 使用策略 | 判断正确的词间位置数 | 关于词间连断的判断准确率$\alpha$/% |
|---|---|---|
| dts | 97881 | 77.87 |
| mi | 95446 | 75.93 |
| dae | 90936 | 72.34 |
| dmd | 105150 | 83.65 |
| dmd + WAR | 110877 | 88.21 |

规则后，词间位置判断准确率有明显提高。

我们发现，词干切分工具的局限性、维吾尔文中难以避免的拼写错误、词间位置的不规范性以及名词术语的不规范缩写等因素在一定程度上影响词间位置判断准确率。关于词干切分算法的局限性，除算法本身的缺陷外，拼写错误也是一个主要的因素，现有的方法和工具还不能对批量文本进行全自动检错和纠错。对于词间位置和名词术语书写规范化，还没有相关的研究报道。不管是算法上的缺陷，还是原始文本的不规范性，都会影响词间判断准确率。因此，对语料库进行训练或对待处理文本进行处理前，应尽量排除以上负面因素的影响，在较规范的文本语料上可以获得更高的切分准确率。这也是我们将来工作的研究重点。

## 5　结语

以空格作为自然分隔符的维吾尔文传统分词方法，会把多词结构的语义串拆分成与其本义完全不符的若干个片段，表现出非常明显的不足和局限性，在维吾尔文文本挖掘领域研究中已成为一大瓶颈。本文提出一种基于词间关联度度量的维吾尔文本自动切分方法，利用单词关联规则和统计结合的方法度量相邻单词之间的关联紧密程度，从而识别出语义串的边界，达到以语义及结构完整的词串为单位进行文本切分的目的，实现了相应的自动切分算法。在大规模测试语料上进行的切分实验表明，该算法表现出较高的准确率和健壮性。本文提出的方法还能应用到哈萨克文、柯尔克孜文等其他语言文本自动切分中。

## 参考文献


[1] 贺敏. 面向互联网的中文有意义串挖掘[D]. 北京: 中国科学院研究生院, 2007: 1–8

[2] 吴庆耀. 无监督的中文语义词抽取技术研究[D]. 深圳: 哈尔滨工业大学深圳研究生院, 2009: 5–10

[3] Chien L F. PAT-Tree-Based keyword extraction for Chinese information retrieval // Proceedings of the 20th annual international ACM SIGIR Conference on Research and Development in Information Retrieval. Philadelphia, PA, 1997: 50–58

[4] Candito M, Constant M. Strategies for contiguous multiword expression analysis and dependency parsing // 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014). Baltimore, MD, 2014: 743–753

[5] Luo R L, Zhang H X, Wu M H. Ambiguity analysis model of word segmentation based on word group. Journal of Applied Sciences, 2013, 13(16): 3153–3160

[6] Masaki M, Masao U. Compound word segmentation using dictionary definitions-extracting and examining of word constituent information. ICIC Express Letters, Part B: Applications, 2012, 3(3): 667–672

[7] Liu X L. Automatic summarization method based on compound word recognition. Journal of Computational Information Systems, 2015, 11(6): 2257–2268

[8] Zheng H T, Kang B Y, Kim H G. Exploiting noun phrases and semantic relationships for text document clustering. Information Sciences, 2009, 179(13): 2249–2262

[9] Sreya D, Narasimha M M. Using discriminative phrases for text categorization // 20th International Conference on Neural Information Processing. Daegu, 2013: 273–280

[10] Rais N H, Abdullah M T, Kadir R A. Multiword phrases indexing for malay-English cross-language information retrieval. Information Technology Journal, 2011, 10(8): 1554–1562

[11] Zhang Y F, Long F, Bin L. Identifying opinion sentences and opinion holders in internet public opinion // Proceedings of the 2012 International

Conference on Industrial Control and Electronics Engineering. Xi'an, 2012: 1668–1671

[12] 孙茂松, 肖明, 邹嘉彦. 基于无指导学习策略的无词表条件下的汉语自动分词. 计算机学报, 2004, 27( 6): 736–742

[13] 王思力, 王斌. 基于双字耦合度的中文分词交叉歧义处理方法. 中文信息学报, 2007, 21(5): 14–17

[14] 费洪晓, 康松林, 朱小娟, 等. 基于词频统计的中文分词的研究. 计算机工程与应用, 2005, 30(7): 67–69

[15] 王芳, 万常选. 基于可信度的中文完整词自动识别. 中文信息学报, 2009, 23(3): 17–23

[16] 何赛克, 王小捷, 董远, 等. 归一化的邻接变化数方法在中文分词中的应用. 中文信息学报, 2010, 24(1): 15–19

[17] 蒋建洪, 赵嵩正, 罗玫. 词典与统计方法结合的中文分词模型研究及应用. 计算机工程与设计, 2012, 33(1): 387–391

[18] Tohti T, Musajan W, Hamdulla A. Efficient term extraction and indexingapproach in small-scale web search of Uyghur Language. Journal of Multimedia, 2013, 8(5): 481–488

[19] Liu J Y, Liu Y. Resolution to combinational ambiguity of Chinese word segmentation // 2009 International Conference on E-learning, E-Business, Enterprise Information Systems, and E-Government. Hong Kong: IEEE, 2009: 141–145

[20] Qiu L K, Hu H L, Wu Y F. Corpus-based method for differentiating genuine and spurious combinational ambiguity. ICIC Express Letters, 2013, 7(4): 1437–1441

[21] 阿力木江·艾沙, 吐尔根·依布拉音, 艾山·吾买尔, 等. 基于机器学习的维吾尔文文本分类研究. 计算机工程与应用, 2012, 48(5): 110–112

[22] 徐峻岭, 周毓明, 陈林, 等. 基于互信息的无监督特征选择. 计算机研究与发展, 2012, 49(2): 372–382

[23] 孟春艳. 用于文本分类和文本聚类的特征抽取方法的研究. 微计算机信息, 2009, 25(3): 149–150

[24] Church K W, Gale W, Hanks P, et al. Using statistics in lexical analysis // Zernik U. Lexical acquisition: exploiting on-line resources to build a lexicon. Hillsdale NJ: Lawrence Erlbaum Associates, 1991: 115–164