doi:10.6041/j.issn.1000-1298.2015.01.037

面向移动终端的农业知识文语转换系统*

李鑫星[1]　陈英义[1,2]　李道亮[1,2]　傅泽田[1,2]　张领先[1,2]

(1. 中国农业大学信息与电气工程学院, 北京 100083; 2. 农业部农业信息获取技术重点实验室, 北京 100083)

摘要: 设计了一套面向移动终端的农业知识文语转换系统,该系统借鉴语义检索技术中对于关键词的语义处理方式,用语义检索技术处理文本切分过程中所遇到的歧义字段,可将农业文本知识转换为自然流畅的语音,并结合呼叫中心技术、借助移动终端推广农业知识。首先对文语转换中的文本分析流程进行分析,明确歧义字段处理为文语转换的关键点。在明确关键点的基础上,设计出用于进行分词的词典,进而基于词典匹配和统计分析模型对歧义字段进行提取,再基于语义检索对歧义字段进行处理,从而实现歧义字段的切分,最终采用 Cool Edit Pro 2.0 软件实现语音合成和韵律处理功能,开发出面向移动终端的农业知识文语转换系统,可有效解决歧义字段的处理问题。由测试结果可以看出,本文算法的查准率为 94.03%,较最大匹配法和三元语法的查准率分别提高了 5.72 个百分点和 0.97 个百分点;本文算法的查全率为 95.32%,较最大匹配法和三元语法的查全率分别提高了 0.23 个百分点和 1.95 个百分点;本文算法的 $F-1$ 测度为 0.93,较最大匹配法提高了 0.01,与三元语法相同,说明本文算法具有较好的性能。

关键词: 农业知识　移动终端　文语转换　歧义字段

中图分类号: S2　　**文献标识码**: A　　**文章编号**: 1000-1298(2015)01-0266-06

Mobile Terminal-oriented Text to Speech System for Agriculture Knowledge

Li Xinxing[1]　Chen Yingyi[1,2]　Li Daoliang[1,2]　Fu Zetian[1,2]　Zhang Lingxian[1,2]

(1. *College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*
2. *Key Laboratory of Agricultural Information Aquisition Technology*, *Ministry of Agriculture*, *Beijing* 100083, *China*)

Abstract: With the help of semantic processing of keyword in semantic retrieval technology, a mobile terminal-oriented text to speech system was designed for agriculture knowledge. The ambiguous phrase encountered in text segmentation were cutting by using semantic search technology which could converted agricultural text knowledge to smooth and natural voice, and popularized agricultural knowledge combining call center and mobile terminal technology. Firstly, the text analysis flow was analyzed, and the ambiguous phrase which was the key point of text to speech processing was confirmed. Based on this key point, the word segmentation dictionary was designed, and the ambiguous phrase was extracted based on the dictionary matching and statistical analysis. Secondly, the ambiguous phrase was processed based on semantic retrieval and the segmentation of ambiguous phrase was realized. Finally, the functions of speech synthesis and prosodic processing were realized with the software of Cool Edit Pro 2.0, and mobile terminal-oriented text to speech system was developed for agriculture knowledge which could effectively solve the problem of ambiguous phrase processing.

Key words: Agriculture knowledge　Mobile terminal　Text to speech　Ambiguous phrase

收稿日期: 2014-01-08　修回日期: 2014-04-04

* 山东省自主创新专项资助项目(2012CX90204)

作者简介: 李鑫星,讲师,主要从事农业系统与知识工程研究,E-mail: lxxcau@cau.edu.cn

通讯作者: 张领先,副教授,主要从事农业系统与知识工程研究,E-mail: zlx131@163.com

引言

随着信息技术的发展，除了广播、电视等传统方式，互联网已经成为人们获取知识最重要的途径。但对于农民而言，通过互联网获取农业知识必须借助计算机等基础设施，而我国农村地区计算机和互联网的普及率并不高；与计算机相比，固定电话，特别是手机则具备价格低廉、易学易用、灵活方便等特点，其普及率高于计算机，可见，手机在我国农村有广阔的应用前景。

呼叫中心可通过固定电话、手机等移动终端，以音频形式为用户提供服务，且其技术已经相对成熟，在电信、金融、旅游、医疗卫生等领域已经得到广泛应用，因此借助呼叫中心技术，通过移动终端为农民提供农业知识，可为其提供有效的知识获取途径。早期的呼叫中心系统多采用人工方式实现语音输出，即事先录制编辑大量的语音文件，存入语音数据库中，需耗费大量的人力物力财力，成为制约呼叫中心发展的瓶颈。文语转换作为一种综合了语言学和声学处理、可将文字形式的信息转换成自然语音的计算机语音合成技术，成为推动呼叫中心发展的关键技术。

现有用于提供农业知识的呼叫中心，如“12316”三农服务热线等，不得不采用成本较高的人工座席或专家在线咨询等方式提供服务，究其原因在于现有文语转换方法并不适用于农业知识的特点，因为农业知识具有较多的专有词汇，如病虫害名称、农药化肥名称等，另外在农业知识中一词多义和一义多词现象十分普遍，同一种作物可能有多个别名，如将现有在工业等领域应用的文语转换技术应用于农业知识，常常产生大量歧义字段，本文也将着重针对这些歧义字段提出相应解决方案，设计出适合农业知识特点的文语转换方法和系统。

另外声调在很大程度上决定了语音的自然度，在一段语流当中，声调与每个字的发音密切相关，因为字的发音和语气受到相邻字的影响，所以欲实现文语转换，需先进行文本分析，结合上下文语境，确定出每个字应该发什么音[1]。当把每个字串联成词句时，必须考虑停顿、语气等韵律相关参数，因此，文语转换过程应包含文本分析、语音合成、韵律处理 3 部分。而文本分析是核心，直接决定着断句和语音效果，歧义字段也都在此产生。本文以棉花知识为例阐述系统的设计方法，系统开发过程中采用现有成熟技术以实现语音合成和韵律处理功能。

1　系统设计

1.1　文本分析流程设计

文语转换与简单的语音合成最大的区别在于它具有某种程度的篇章理解能力，这种篇章理解能力就由文本分析模块实现。它能够对文本进行语言学分析，生成一种适合于语音学的内部表示，并根据文本的上下文关系在一定程度上对文本进行浅层的分析理解，从而确定文本中的字、词、短语、句子等[1]，因此文本分析是文语转换技术中最核心的部分。

目前限制文语转换发展的最大瓶颈是在文本分析过程中遇到的歧义字段，可分为交集型歧义字段和组合型歧义字段[2]。交集型歧义：在字段 AJB 中，AJ 属于 W，JB 属于 W，则称 AJB 是交集型歧义字段，其中 A、J、B 为字串，W 为词典。例如：“把风车”分为“把/风车”和“把风/车”。组合型歧义：在字段 AB 中，AB 属于 W，A 属于 W，B 属于 W，W 为词典，称 AB 为组合型歧义字段，例如：“把手”可以切分成“把手”和“把/手”[3]。据统计，汉语真实文本中，歧义切分现象出现的概率约为 1/110，即平均 110 个汉字中出现一次歧义切分，其中交集型歧义切分现象占 86%[4]，由于歧义字段造成的文本切分多样性给自动分词带来极大的困难，目前大多数文语转换技术，都不得不把歧义字段的每一个字都切分为一个词，并在每个字之间插入停顿间隔标记，而导致合成的语音像“蹦豆”似的一字一断、机械性极强，与人类自然流畅的发音相距甚远，因此要想设计出好的文本切分方法，必须解决好歧义字段的处理问题。

传统分词方法大都借助于词典，根据字符串匹配的原理，把待处理语句流中的字序列和词典库中的词语序列逐个进行比较匹配，从而把词语逐步从文本中分离出来，是从文本到词的映射过程，如图 1 所示。这些分词方法简洁、易于实现，在工程上得到了广泛的应用，但是该方法切分精度不高，对于切分歧义无法有效地克服[5]。比如对于文本“其中第一方面包括”，应该切分为“其中/第一/方面/包括”，但如果按照传统的切分方法，这句文本很可能被切分为“其中/第一/方/面包/括”，因此“面包”也就成为本句文本的歧义字段。

如上文所分析的，文本切分过程实际上是文本到词的映射过程，受此启发，本文联想到了农业知识检索领域中基于语义的文本搜索问题，如图 2 所示。信息检索方法是使用一组具有代表性的关键词来描述数据库中的每一篇文档，而关键词是文档中的一些简单的单词，通过它们可以与数据库中的文档相

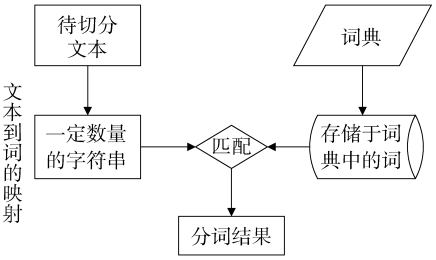

图 1 传统的文本切分流程

Fig. 1 Flow of traditional text segmentation

联系,使用关键词来索引文档内容[6]。而关键词与歧义字段十分相似,都是文本中所包含的词,文本搜索是从词到文本的映射过程,同样与文本切分从文本到词的映射有着千丝万缕的联系,因此本文借助农业知识检索中的文本搜索问题来研究文本切分问题,特别是通过对关键词的研究,解决歧义字段的处理问题:首先构建用于进行分词的词典,进而基于词典匹配和统计分析模型对歧义字段进行提取,最终基于语义检索对歧义字段进行处理,从而最终实现歧义字段的切分。

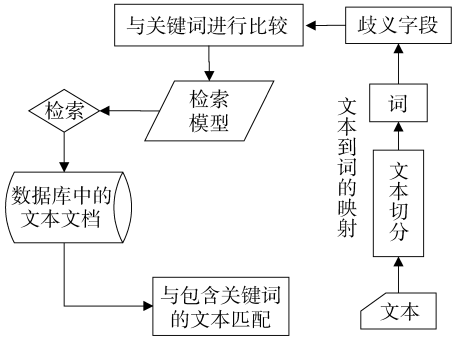

图 2 信息检索流程

Fig. 2 Flow of information search

1.2 分词词典模块设计

分词词典是文本分析的基础,由于自动分词所需的全部匹配信息都要依赖于分词词典,因此词典中词条的覆盖率直接影响到分词的精度,词典结构的好坏也影响到分词的速度。本文在构建词典时借鉴《现代汉语常用词表(草案)》[7]中所提供的词汇和词频信息。该草案收集词语同国家语言文字工作委员会"现代汉语通用语料库"核心语料库、厦门大学的新词语语料库、《现代汉语规范词典》、《现代汉语词典》、《新华词典》等所收词语进行了对比,并查验了该词在人民网《人民日报》报系网页以及 Google 简体中文网页、百度等常用网页上的使用情况,共收录常用词语 56 008 个。本文通过词汇和词频高低进行整理后,最终挑选出 2 万条左右的常用词汇,并用 MySQL 数据库建立和储存通用词典的内容。

考虑到汉字的编码体系特点,本文设计的词典具有三级索引的首字 Hash 表结构,利用 Hash 查找可以高效地检索到词条的位置。并且词典预留了词汇扩充功能,用于存储在后续处理中多次被标记为歧义字段但又按完整词切分的词汇,从而增加系统的自学习功能。

(1)首字索引表。汉字在计算机中以内码的形式存在,将内码计算后转为区位码,汉字的区位码就是 GB2312 码中的汉字部分,它可以唯一确定某一汉字或字符,也就是说汉字与区位码是一一映射的关系。比如"棉"字的区位码是 3562,"花"字的区位码是 2708 等,像这样定义的汉字共有 6 763 个。本文采用首字偏移的方法来直接定位首字在字典索引表中的位置,这一过程不进行任何匹配,理想情况下一次存取便能得到所查词条,提高了分词速度。

定义入口地址的计算公式为

$$O = (C_1 - 16)94 + (C_2 - 1) \tag{1}$$

式中 O——汉字在 Hash 表中的位置

C_1、C_2——汉字的区码和位码

(2)词长索引表。因各个词的长度不同,分词的切分算法采用改进的正向最大匹配法。对于相同首字的词,采用按词长递减顺序进行存储。

(3)词汇正文。用词语链表储存相同首字词长的词语及词频等相关信息,词典的存储结构如图 3 所示。

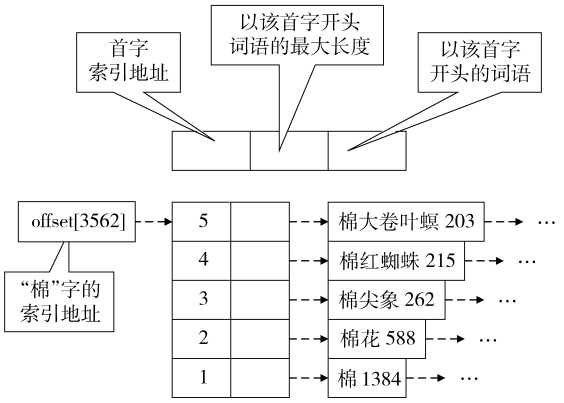

图 3 词典存储结构图

Fig. 3 Storage structure of dictionary

1.3 歧义字段提取模块设计

在分词词典的基础上,首先采用改进的正向最大匹配算法进行分词。

设待切分的字符串 $Str = S_1S_2\cdots S_n$。取 S_1,通过 Hash 表搜索,找到该字的索引项,获取相关数据。若 $L_{max} = 1$,则没有以 S_1 为首字的多字词,即 S_1 为单字词,将 S_1 切分出来。令 $Str = S_2S_3\cdots S_n$,继续下一次切分。若 $L_{max} > 1$,则取 $k = L_{max}$,取 $Str_1 = S_1S_2\cdots S_k$,在词典中搜索 Str_1。若搜索成功,则获得切分词 $Stmp_1 = S_1S_2\cdots S_k$,将其切分出来。若搜索失败,则

取字符串 $Str_1' = Str_1 - S_k = S_1S_2\cdots S_{k-1}$ 进行搜索，若搜索成功，则直接切分成 $Stmp_2$；若失败则取字符串 $Str_1'' = Str_1' - S_{k-1} = S_1S_2\cdots S_{k-2}$ 进行搜索，如此循环直到搜索成功为止。再取其余字符串 $Str_2 = S_{k+1}S_{k+2}\cdots S_n$ 继续按照此规则切分，直到整个字符串都切分完毕。

通过以上的分词，对于与词典匹配成功的文本已完成切分，但对于未匹配部分，还不能作为歧义字段直接提取出，因为歧义字段被提取后，待切分文本应当被分割为在语义层面上完全独立的两部分，即歧义字段和非歧义字段，歧义字段将作为关键词输入语义检索接口，而语义检索模型具备语义处理能力，因此所有可能产生歧义的元素都应并入歧义字段；而非歧义字段则不做任何处理直接存入切分结果库。因此，在非歧义字段中，不应再含有任何可能产生歧义的元素，否则造成的歧义将无法再进行处理。为能更准确地提取歧义字段，本文采用统计分析模型对于歧义字段进行提取，即通过统计分析方法计算分词过程中，与词典不匹配文本前后分界处相邻两字还有无构成词汇的可能（产生歧义的可能），将不会与其相邻部分构成任何歧义的文本提取出，作为最终的歧义字段。

本文采用的统计分析方法为互信息概率统计算法，该算法的本质体现在字与字结合的紧密程度上，而紧密程度可通过互信息表达，对于有序汉字串 XY 中 X、Y 之间的互信息定义公式为

$$I(X,Y) = \mathrm{lb}\frac{P(X,Y)}{P(X)P(Y)} \tag{2}$$

式中　$P(X,Y)$——X、Y 相邻出现的概率

$P(X)$、$P(Y)$——X、Y 单独出现的概率

字符串 X、Y、XY 出现的次数分别定义为 $f(X)$、$f(Y)$、$f(XY)$。N 表示整个语段的词语总数，则有

$$P(X) = \frac{f(X)}{N} \tag{3}$$

$$P(Y) = \frac{f(Y)}{N} \tag{4}$$

$$P(X,Y) = \frac{f(XY)}{N} \tag{5}$$

将式(3)～(5)代入到式(2)中得

$$I(X,Y) = \mathrm{lb}\frac{Nf(XY)}{f(X)f(Y)} \tag{6}$$

若与词典不匹配文本前后分界处相邻两字的互信息 $I(X,Y)\neq 0$，则将分界处的字并入，并继续计算新分界处的互信息，直到 $I(X,Y)=0$ 为止，此时的文本与相邻部分不会再产生歧义，也即成为最终的歧义字段。

1.4　歧义字段处理模块设计

文语转换中歧义字段的产生，究其原因也是没有结合上下文的语境从语义的层面审视待切分的词，文本检索中对于关键词的语义处理方法可为歧义字段的处理提供借鉴。在文本检索领域，为使关键词语义能够被计算机更好的理解，通常采用基于本体的语义检索方法，即将词典本体引入关键词检索过程中，通过本体对关键词的语义进行解析，然后不是使用关键词本身，而是用其语义概念与待检索文本进行匹配。这也表明，即使文字完全相同，也只有表达出与关键词相同语义的文本才会被检索出，如图 4 所示。

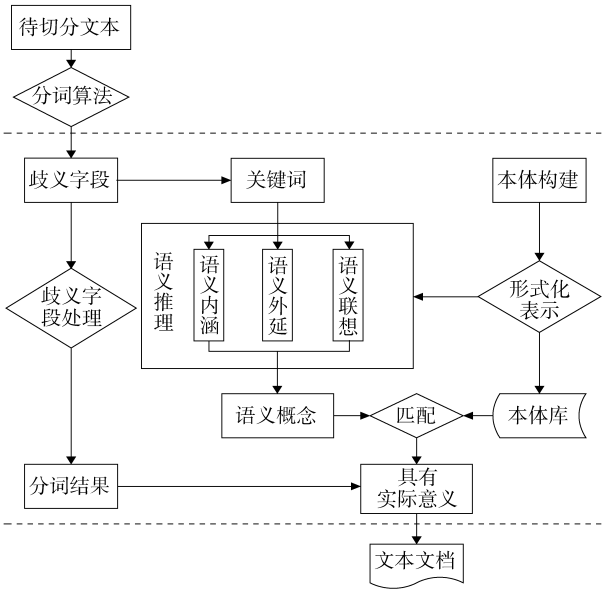

图 4　基于语义检索的文本检索流程

Fig. 4　Semantic retrieval based on text retrieval

图 4 中语义检索中的关键词和有实际意义的语义概念，分别与文本切分中的歧义字段和切分词汇具有极大的相似性，而两条虚线之间的部分，是语义检索的核心模块，其功能恰恰可用于进行歧义字段处理。因此，本研究拟将语义检索引入文语转换领域[6]，将歧义字段作为关键词输入语义检索接口，对其进行语义推理和语义提取，而语义的提取可通过计算歧义字段与知识库中语义节点间的路径距离（语义相似度）来实现。设 d 表示歧义字段与语义节点的路径距离，语义相似度的计算公式为

$$S(n_1,n_2) = \frac{\gamma}{d+\gamma} \tag{7}$$

式中　n_1、n_2——歧义字段和语义节点

d——n_1、n_2 的路径距离

γ——可变系数

即使歧义字段与知识库中的某些语义节点相关，具体的相关关系也要分 3 种情况考虑：①直接相

关,包括实例与属性、实例与子类、具备相同属性的类之间的相关关系。②包含相关,子类与父类的继承相关。③传递相关,经过直接或包含相关的传递过程产生的相关关系。显然,这 3 种相关关系影响相似度的权重不同,本文通过计算语义相似度,可区分不同相关关系的影响度,最终提取出与歧义字段密切相关的语义节点,该节点的知识内容也即成为歧义字段的语义概念,计算式为

$$S(n_1,n_2)=\sum_{i=1}^{3}\beta_i\prod_{j=1}^{i}S_j(n_1,n_2) \tag{8}$$

式中 β_1、β_2、β_3——在计算语义相似度的计算过程中,$S_1(n_1,n_2)$、$S_2(n_1,n_2)$与$S_3(n_1,n_2)$3 种不同相关关系所占的权重

在语义检索中,当通过语义推理获得关键词的语义概念后,就应用语义概念检索存储于文本库中的文本。但本文研究的是歧义字段切分问题,并不需要检索文本,而只要获知歧义字段的语义概念具有实际意义,就会将歧义字段作为一个完整词进行切分,并同时对此歧义字段标注 1 次,若同一歧义字段累计被标注 3 次,则将此歧义字段加入分词词典,供后续分词使用;否则将其切分为若干个单字词。

本文使用歧义字段的语义概念进行检索,但被检索的内容不是大量的文本,而是本体实例。即如果把本体实例视为检索问题中文本的话,那么待检索文本都只是单个的词。如果能够检索到,就认为歧义字段的语义概念具有实际意义。因此需要计算语义概念与本体实例间的词汇关联度,具体方法是提取出可表达词汇潜在关系的基因对,进而计算出基因对之间的关系。任一基因对 k 和 l 的关联度计算式为

$$a[k][l]=\sum_{i=1}^{M}W_i[k]W_i[l] \tag{9}$$

其中 $$W_i[k]=T_i[k]\lg(M/n[k])$$

式中 k——歧义字段的语义概念,是文档中的第 k 个基因项

l——本体实例

M——本体实例的个数

$T_i[k]$——第 k 个基因项在文档 d_i 中出现的频率

$n[k]$——包含第 k 个基因项的本体实例个数,由于文档的内容是单个的词,$T_i[k]$、$n[k]$的取值只能为 0 或 1

当关联度 $a[k][l]$超过设定的阈值时,即可检索到本体实例,也即认为歧义字段能够表达出明确的语义概念,则将其作为完整词切分;反之,将其切分为单子词组合。

2 系统实现

本文设计的农业知识文语转换系统,基于面向对象的思想,使用 C + + 语言设计开发,开发过程中使用上文提出的方法实现文本分析,进而采用 Cool Edit Pro 2.0 实现语音合成和韵律处理功能(图 5),开发软件为 Visual Studio 2005,数据管理使用 MySQL数据库。

图 5 Cool Edit Pro 2.0 界面

Fig. 5 Interface of Cool Edit Pro 2.0

图 6 为系统的文本导入界面,点击界面右边的“导入文本”按钮,选择需要导入到 txt 文档,即可将文档中的内容导入系统中(图 7)。

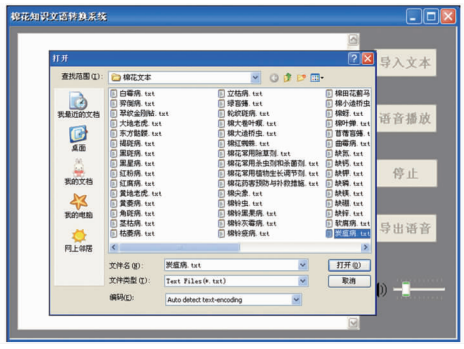

图 6 文本导入界面 1

Fig. 6 Text import interface 1

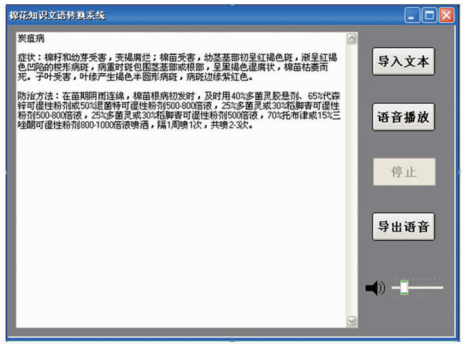

图 7 文本导入界面 2

Fig. 7 Text import interface 2

输入完待转换文本或选择完待转换文本后,点击“语音播放”按钮就可以通过计算机的扬声器或耳机听见合成的语音,并根据个人需要调节音量,且

“语音播放”按钮标题变为“暂停”。当文本转换为语音结束后，若想要导出并储存语音文件，则点击“导出语音”按钮，选择想要保存的目录位置即可（图 8）。

图 8　语音播放与导出界面

Fig. 8　Voice playback and export interface

3　试验结果分析与讨论

卡耐基梅隆大学 2004 年 1 月公布了由 Joy 开发的汉语自动分词评估工具包 EDWS（Edit Distance of the Word Separator）[8]，本文就采用 EDWS 工具包对本文设计的分词方法进行评价。一般来说，查准率和查全率是评价一个分词系统优劣的经典的指标，而 F 测度是综合评价二者的一个重要指标。EDWS 采用查准率和查全率及 $F-1$ 测度评价分词系统的性能，各评价指标分别定义如下：查准率 P[3] 为分词结果中切分正确的总词数占所有输出结果总词数的比例。召回率（查全率）R[3] 为净分词结果中切分正确的总词数占标准文本中的总词数的比例。F 测度[4] 计算式为

$$F=\frac{(\beta^2+1)PR}{\beta^2P+R}$$

在通常情况下，取 $\beta=1$，那么得到 $F-1$ 测度的计算公式为

$$F=\frac{2PR}{P+R}$$

在测评过程中，本文同时选取最大匹配法和三元语法与本文方法做比较，结果如表 1 所示。

表 1　试验结果

Tab. 1　Experiment results

分词方法	P/%	R/%	$F-1$	分词速度/（字·min^{-1}）
本文方法	94.03	95.32	0.93	36 000
最大匹配法	88.31	95.09	0.92	65 000
三元语法	93.06	93.37	0.93	42 000

由测试结果可以看出，本文算法的查准率为 94.03%，较最大匹配法和三元语法的查准率分别提高了 5.72 个百分点和 0.97 个百分点；分词正确率有所提高；本文算法的查全率为 95.32%，较最大匹配法和三元语法的查全率分别提高了 0.23 个百分点和 1.95 个百分点；对于分词查准率和查全率的一个综合测度 F，在参数为 1 的情况下，本文算法的 $F-1$ 测度为 0.93，较最大匹配法提高了 0.01，与三元语法相同，说明本文算法具有较好的性能。虽然在分词速度上，本文算法慢于最大匹配法和三元语法，但以 36 000 字/min 的分词速度而言，仍可以接受。

4　结束语

本文设计了一套面向移动终端的农业知识文语转换系统，该系统将语义检索技术引入文本切分领域中，用于对歧义字段进行处理，进而实现农业知识的文语转换。首先设计出用于进行分词的词典，进而基于词典匹配和统计分析模型对歧义字段进行提取，再基于语义检索对歧义字段进行处理，从而实现歧义字段的切分，最终采用 Cool Edit Pro 2.0 软件实现语音合成和韵律处理功能，设计开发出农业知识文语转换系统，可结合呼叫中心技术、借助移动终端推广农业知识。

参　考　文　献

1　郭锋. 基于 PSOLA 的汉语文语转换技术研究[D]. 南京：南京航空航天大学，2007.

2　闫引堂，周晓强. 交集型歧义字段切分方法研究[J]. 情报学报，2000，19(6)：637－642.
Yan Yintang，Zhou Xiaoqiang. Study of segmentation strategy oil ambiguous phrases of overlap type [J]. Journal of the China Society for Scientific and Technical Information，2000，19(6)：637－642.（in Chinese）

3　张彩琴，袁健. 改进的正向最大匹配分词算法[J]. 计算机工程与设计，2010，31(11)：2595－2633.
Zhang Caiqin，Yuan Jian. Improved forward algorithm for maximum matching word segmentation [J]. Computer Engineering and Design，2010，31(11)：2595－2633.（in Chinese）

4　杨超. 基于最大匹配的书面汉语自动分词研究[D]. 长沙：湖南大学，2004.

5　Niu Zhengyu，Chai Peigi. Segmentation of prosodic phrases for improving the naturalness of synthesized mandrine Chinese speech [C]// Proceedings of ICSLP－2000，2000：39－45.

6　岳峻. 基于本体的蔬菜供应链知识获取系统研究[D]. 北京：中国农业大学，2007.

7　http：// www. moe. edu. cn/publicfiles/business/htmlfiles/moe/s230/201001/75598. html.

（下转第 239 页）

28 Baret F, Guyot G, Begue A, et al. Complementarity of middle-infrared with visible and near-infrared reflectance for monitoring wheat canopies[J]. Remote Sensing of Environment, 1988, 26(3):213 – 225.

29 Tim J M, Bruno A, Danson F M, et al. Candidate high spectral resolution infrared indices for crop cover [J]. Remote Sensing of Environment, 1993, 46(2): 204 – 212.

30 Vaesen K, Gilliams S, Nackaerts K, et al. Ground-measured spectral signatures as indicators of ground cover and leaf area index: the case of paddy rice[J]. Field Crops Research, 2001, 69(1): 13 – 25.

31 Bradley C R. The influence of canopy green vegetation fraction on spectral measurements over native tall grass prairie[J]. Remote Sensing of Environment, 2002, 81(1):129 – 135.

32 李存军,赵春江,刘良云,等. 红外光谱指数反演大田冬小麦覆盖度及敏感性分析[J]. 农业工程学报, 2004,20(5): 159 – 164.
Li Cunjun, Zhao Chunjiang, Liu Liangyun, et al. Retrieval winter wheat ground cover by short-wave infrared spectral indices in field and sensitivity analysis[J]. Transactions of the CSAE, 2004,20(5): 159 – 164. (in Chinese)

33 Baret F, Jacquemoud S, Hanocq J F. The soil line concept in remote sensing [J]. Remote Sensing Reviews, 1993, 7(1): 65 – 82.

34 池宏康. 遥感数据的裸沙土壤线校正方法 [J]. 地理学报,1999,54(5): 454 – 461.
Chi Hongkang. A method for correcting remote sensing data by bare-sand soil line [J]. Acta Geographica Sinica, 1999, 54(5): 454 – 461. (in Chinese)

35 Robertson A N, Gitelson A A, Peng Y, et al. Green leaf area index estimation in maize and soybean: combining vegetation indices to achieve maximal sensitivity[J]. Agronomy Journal, 2012, 104(5):1336 – 1347.

36 Viña, A, Gitelson A A. New developments in the remote estimation of the fraction of absorbed photosynthetically active radiation in crops [J]. Geophysical Research Letters, 2005, 32(17): L17403.

37 Gitelson A A. Remote estimation of crop fractional vegetation cover: the use of noise equivalent as an indicator of performance of vegetation indices [J]. International Journal of Remote Sensing,2013,34(17):6054 – 6066.

38 Viña A, Gitelson A A, Robertson A N, et al. Comparison of different vegetation indices for the remote assessment of green leaf area index of crops [J]. Remote Sensing of Environment, 2011, 115(12): 3468 – 3478.

39 翟清云,张娟娟,熊淑萍,等. 基于不同土壤质地的小麦叶片氮含量高光谱差异及监测模型构建[J]. 中国农业科学, 2013, 46(13): 2655 – 2667.
Zhai Qingyun, Zhang Juanjuan,Xiong Shuping,et al. Research on hyperspectral differences and monitoring model of leaf nitrogen content in wheat based on different soil textures[J]. China Agriculture Science, 2013, 46(13): 2655 – 2667. (in Chinese)

(上接第 271 页)

8 http://projectileAs.Cs.emu.edu/research,publie/toolstsegmentationteval/index.html.

9 Schubert Foo, Hui Li. Chinese word segmentation and its effect on information retrieval [J]. Information Processing & Management, 2004,40(1):161 – 190.

10 刘异,黄魏,高兵,等. 基于词条组合的中文文本分词方法[J]. 科学技术与工程,2010,10(1):85 – 89.
Liu Yi, Huang Wei, Gao Bing, et al. Word combination based Chinese word segmentation methodology[J]. Science Technology and Engineering,2010,10(1): 85 – 89. (in Chinese)

11 杜璞.基于领域语料库的中文自动分词系统的研究[D].西安:西安科技大学,2007.
Du Pu. The research of automatic Chinese word segmentation system based on domain corpus[D]. Xi'an: Xi'an University of Science and Technology,2007. (in Chinese)

12 李宏波.综合字典和统计分析的中文分词系统的研究与实现[D].武汉:武汉理工大学,2010.
Li Hongbo. The research and implementation of the system for Chinese word segmentation base on dictionary and statistic [D]. Wuhan: Wuhan University of Technology, 2010. (in Chinese)

13 王瑞雷,栾静,潘晓花,等.一种改进的中文分词正向最大匹配算法[J].计算机应用与软件,2011,28(3):195 – 197.
Wang Ruilei, Luan Jing, Pan Xiaohua,et al. An improved forward maximum matching algorithm for Chinese word segmentation [J]. Computer Applications and Software, 2011,28(3):195 – 197. (in Chinese)

14 Richard Tzong-Han Tsai. Chinese text segmentation: a hybrid approach using transductive learning and statistical association measures [J]. Expert Systems with Applications,2010,37(5):3553 – 3560.

15 Fu Guohong, Kit Chunyu, Webster J J. Chinese word segmentation as morpheme-based lexical chunking [J]. Information Sciences, 2008,178(9): 2282 – 2296.