25卷 第8期
2008年8月

微电子学与计算机
MICROELECTRONICS & COMPUTER

Vol.25 No.8
August 2008

多民族语言农业信息处理平台的构建

强 静[1,2],张 建[1],李 淼[1]

(1 中国科学院 合肥智能机械研究所,安徽 合肥 230031; 2 中国科学技术大学 信息科学技术学院,安徽 合肥 230027)

摘 要:提出了构建多民族语言农业信息处理平台的方法,重点介绍怎样把机器翻译与农业信息处理平台结合,在原来的农业信息处理平台的基础上,进行了结构性改造,加入了翻译系统,研制了多语言农业信息处理平台.并以汉蒙双语农业技术咨询系统为例,介绍具体的实现与应用.得出了构建汉语/民族语言农业信息处理平台的一般方法.

关键词:农业信息;推理机;翻译模型;机器翻译

中图分类号:H085 **文献标识码**:A **文章编号**:1000-7180(2008)08-0168-04

Construction of Multiracial-Language Agriculture Information Processing Platform

QIANG Jing[1,2], ZHANG Jian[1], LI Miao[1]

(1 Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China;
2 School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China)

Abstract: This paper proposes the the Construction method of Multiracial-Language Agriculture Information Processing Platform, Introduce weightily how to combine machine translation and Agriculture Information Processing Platform. On the base of the Agriculture Information Processing Platform, processes structural reconstruct, adds translation system, develops the Multiracial-language Agriculture Information Processing Platform. Taking Sino-Mongolian reasoning machine for example to illustrate the application and implement of it. It obtained a general method of Constructing Chinese /Minority Languages agricultural information processing platform.

Key words: agriculture knowledge; inference engine; translation model; machine translation

1 引言

随着社会信息化发展,农业信息技术处理平台的应用需求更显迫切.文中在原来"平台"基础上,进行结构性改造,研制了"多语言农业信息处理平台".

在民语即时翻译的技术路线上,考虑到各民族语言的不同特点,采用了以"人工智能"思想为主的基于统计的规则提取方法以及基于逐次筛选法的多引擎汉/民翻译技术.这样的结构设计,可以在民族语言学家提供的民语对应语料库、部分规则以及字库的基础上,随时生成民族语言的"农业专家系统",以便在不同民族地区推广应用.

2 多民族语言农业信息平台体系结构

多民族语言农业信息处理平台的体系结构,分为两部分:一是由农业知识获取系统;二是汉/民双向农业技术咨询系统.文中主要详细介绍涉及民族语言翻译的第二个部分.

如图1所示,汉/民双向农业技术咨询系统由六个模块组成,各个模块及其功能如下:

(1) 显示界面模块,利用网页中表格设置的技术,分别显示汉语和民族语言的推理结果.其中民族

收稿日期:2007-10-23

基金项目:中国科学院知识创新工程重要方向项目(KGCX2-SW-511)

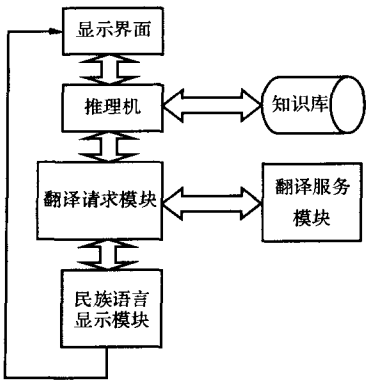

图 1　多语言农业咨询系统的体系结构图

语言的结果按民族语言的显示习惯以弹出对话框的形式出现.

(2) 推理机模块，采用 Agent 技术基于模型知识表示和案例推理的方法，把用户在多种编辑器下写好的专业领域知识(也称知识库)“转变”成计算机应用软件，这个应用软件将模拟专家思维过程，提出专家咨询意见.本模块的主要功能是根据知识库提供的知识进行推理，得出结论.

(3) 翻译请求模块，利用客户机服务器技术，主要是完成翻译的前期的工作和后期的多语言显示可能需要的工作.前期工作主要是对源语言(汉语)进行文本断句，然后进行汉语分词，再进行相应的预处理，把结果发送给服务模块.后期多语言显示的工作是考虑到有的民族语言可能需要进一步的转化才能显示成正确的、通用的民族语言形式.

(4) 翻译服务模块，利用客户机服务器技术，主要完成解码的工作，由解码器构成.解码器就是把客户端发送来的预处理的结果进行解码，解码成所需的目标语言.

(5) 民族语言显示模块，根据各个民族语言的特点并以网页的形式显示其推理结果.

(6) 知识库模块，主要是存储相关领域的知识.

3　多民族语言农业信息平台中的翻译关键技术

本系统中嵌入的翻译技术是采用基于短语的统计机器翻译技术.首先构建农业领域 3 万汉语词条以及 10 万句对汉/民语言双语对齐语料库、汉/民语言 10 万词条双语词典.从这些语料库中自动学习、构建出语言模型与翻译模型[1].关键技术如下：

(1) 汉语分词工具，提出了基于全词哈希的词典结构的方法，在设计词典结构的时候，将每个词映射到一个整数上.这样在加载词典和进行 N 元概率计算的时候，大大降低了 N 元词典所需的空间和计算 N 元概率的时候字符串匹配所需的时间.N 元概率的计算，区别于以往的最大似然估计和加 1 平滑算法，借鉴了统计语言模型中使用的 Good-Turing 概率计算方法和 Katz 平滑算法；

(2) 词语对齐工具，利用对数线形模型的方法，在 IBM3 模型的基础上，通过加入双语词典模型和词性标记模型，改进了原有对齐精度，通过定义启发式搜索路径，减少了搜索空间；

(3) 词语评分工具，利用最大释然法，在双语双向提取后的词语对齐的基础上，进行词语的翻译概率的计算.计算公式：

$$\phi(\bar{s} \mid \bar{t}) = \frac{\mathrm{count}(\bar{s},\bar{t})}{\sum_{t}\mathrm{count}(\bar{s},\bar{t})} \tag{1}$$

(4)短语抽取工具，参考 F. J. Och 的启发式的短语抽取方法，首先进行双向词语对齐，通过将对齐结果进行扩展获得多对多的词语映射.抽取双语句子中词语连贯对齐的所有短语对；F. J. Och 对双语短语 BP 的集合可以定义为

$$BP(s^J,t^I,A) = \{(s_j^{j+m},t_i^{i+n}) : \forall(i',j') \in \boldsymbol{A} : j \leqslant j+m \leftrightarrow i \leqslant i' \leqslant i+n\} \tag{2}$$

式中，s^J 为源句子，t^I 为目标句子，m 为源句子长度，n 为目标句子长度，$\boldsymbol{A}$ 为对齐矩阵.

(5) 短语评分，参考 P. Koehn, F. J. Och 的短语评分方法，计算双向的短语翻译概率和检验短语抽取效果的双向词典化概率.在词语评分和短语抽取的结果的基础上进行短语和词典评分；

(6) 解码器，在信道模型中将翻译过程表示为

$$\hat{e}^I = \arg\max_{e^I}\{\Pr(e^I \mid f^J)\} \tag{3}$$

$$= \arg\max_{e^I}\{\Pr(e^I)\Pr(f^J \mid e^I)\} \tag{4}$$

式中，$\Pr(e^I)$ 为目标语言模型，反映目标语言句子的质量；$\Pr(f^J \mid e^I)$ 为翻译模型，体现源语言句子到目标语言句子的互翻译可能性；argmax 是搜索最大概率 e^I 的算子，这个搜索过程在统计机器翻译中又称为解码过程.

本系统解码方法利用动态编程的 beam search 算法[2]，从代表未翻译的初始状态开始进行扩展，每次扩展翻译源句子中的一个短语生成新的状态.在模型因子的驱动下对扩展的结果进行动态的剪枝，直至源句子翻译完成扩展结束.其特点是加入了一些附加模型如：词语惩罚模型、扭曲模型，提高了翻译的正确率[3-4].

4 汉语民族语言机器翻译流程

农业咨询系统知识推理过程中的汉/民机器即时翻译,采用改进的基于短语的统计机器翻译方法.如图 2 所示,翻译流程主要可分为以下 6 个步骤.

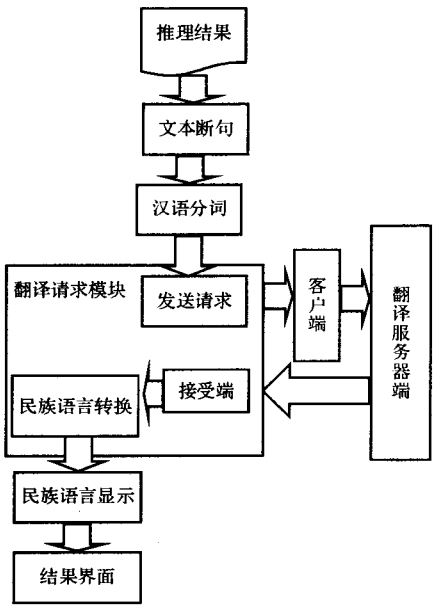

图 2 汉民翻译过程的体系结构图

(1) 首先把咨询工作中所用的汉语知识进行文本断句.

(2) 把进行文本断句的结果进行汉语分词.

(3) 分词后的结果交给发送请求模块,然后发送请求模块对分词后的结果进行打包.建立客户端,通过客户端向翻译服务模块发送要翻译的文件.

(4) 翻译服务器端响应翻译请求,把应答的结果返回给接收端.应答的结果即为翻译后的拉丁转写形式的民族语言.

(5) 接收端接收到翻译后的结果进行民族语言的编码转换.即把民族语言拉丁转写形式转化成民族语言传统的书写形式.

(6) 民族语言转换后以网页的形式显示在推理界面上,即为推理后的最终咨询结果.

5 多民族语言农业信息平台中基本问题

在多民族语言农业信息处理平台中,由于各种语言的文字形状、语法等都不同,在显示时可能遇到一些问题.另外训练模型和翻译通信过程也可能存在一些问题.具体叙述如下.

5.1 控件中的民族语言显示问题

民族语言的语系不同,文字的写法和形态也都有各自的特点.例如,蒙文属于阿尔泰语系,在形态学方面以词根或词干为基础,后接附加成分派生新词和进行词形变化.其显示是从上到下,竖排.彝文则属于藏缅语系,独体字多合体字少,没有音类和意类的区分.传统书写方式是从右向左横排.

根据各个民族语言的特点采用不同的方法解决民族语言的显示问题.例如蒙文的显示,首先要在机器上安装蒙文字库,在网页或控件中的显示,都采用从上到下,竖排,并且按从左到右的顺序显示.并且还可以根据自己的喜好设置蒙文字体及其大小.使显示的效果更好的符合民族语言显示习惯.

5.2 语言模型与翻译模型的训练问题

语言模型的训练,语言模型用于评价译文的忠实度和流利度.解码过程中使用了 SRILM 语言模型训练工具[5]训练的 N-gram 的语言模型.其中采用 N-gram 统计信息的平滑处理.加入了 backoff 值,求出 N-gram 统计信息量,N 一般取 3,即统计出了语料库中的所有词的一元、二元和三元词频;同时统计出了所有词的发射频率;对所有统计出的统计信息作了平滑处理;将最终的词的相应概率信息以 ARPA 格式写入言模型文件中.

翻译模型的训练,根据第二节中描述的方法构建短语翻译模型,使用 GIZA + + 进行源语言与目标语言的双向词语对齐,利用启发式的规则在双向对齐求交集和并集的基础上进行对齐扩展,然后根据扩展得到的词语对齐进行短语抽取构建短语翻译模型.

对抽取的短语使用极大似然法计算短语翻译对的双向翻译概率,并在双向词语对齐词典翻译概率基础上计算抽取短语的词典概率.

5.3 翻译通信问题

推理过程需要翻译的中文文字信息,通过文本断句、中文分词后利用翻译请求端向翻译服务端发送翻译请求.翻译服务端不断监听翻译请求端的的请求,并将翻译请求提交给服务端的解码器进行翻译,最后将翻译的结果返回.

由于翻译过程在翻译请求端和翻译服务段的两个不同进程中进行,这样避免了对推理机系统资源的抢占开销,也提高了对翻译模块的独立性.

民族语言从翻译客请求端获取翻译译文还不能作为我们的翻译结果直接显示.要进行相应的转换以便显示.

6 实验结果

在上述研究基础上,以汉/蒙为例,构建了汉蒙

农业信息处理平台,在此平台中开展了汉/蒙农业咨询即时翻译系统的实验.针对蒙文的特点,该平台前期语料库的预处理中,增加了蒙文语料库、拉丁蒙文词性标注工具,汉蒙双语词语对齐工具.拉丁蒙文语料库词性标注采用改进的基于转换的错误率驱动算法实现.平台界面由三大模块构成,分别是推理路径显示窗、交互视图和最终推理结果视图.

最终的推理结果包括两部分,一是汉语的推理结果,以网页的形式显示;二是蒙文的推理结果,采用对话框的形式,对话框中的蒙文也是以网页的形式显示.推理结束时汉语和蒙文的推理结果同时显示.对话框中显示蒙文的推理结果,对话框下面覆盖的是的汉语的推理结果,移动对话框即可以看到.此外,两种语言的推理结果分别有保存和打印的功能.汉蒙推理图、推理结果图分别如图3、图4所示.

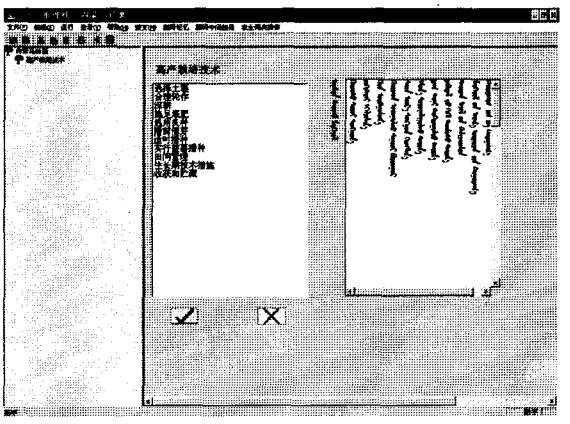

图3　汉蒙推理图

7　结束语

多民族语言农业信息处理平台中的这个翻译系统不仅可以用于汉蒙翻译中,同样还可以应用在其他双语的翻译中.在此基础上也可以配合其他民族的农业语料库,以及相应的字库和显示方法,形成其他民族的农业信息咨询即时翻译系统.目前以同样的方法移植汉/维、汉/彝等农业信息咨询即时翻译系统.

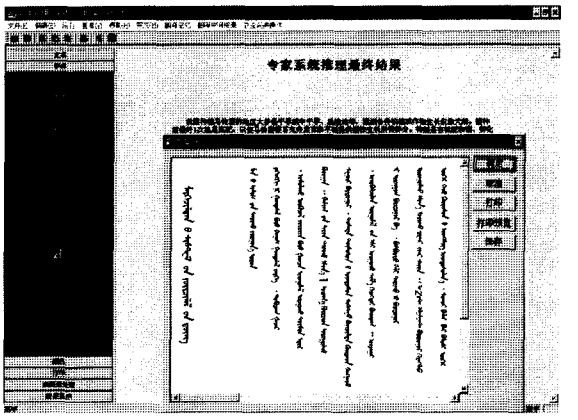

图4　推理结果图

参考文献:

[1] Koehn P, Och F J, Marcu D. Statistical phrase-based translation[C]// Proceedings of the Human Language Technology Conference. Beijing, 2003.

[2] Koehn P. Pharaoh: a beam search decoder for phrase-based statistical machine translation models[C]// Proceedings of the Association of Machine Translation in the Americas (AMTA-2004). USA: University of California, 2004.

[3] 朱文忠.基于VC的串行通信实现方法探析[J].微电子学与计算机,2007,24(1):159-161.

[4] 周光明,张喜生,贾郭军.分布式农业专家系统的设计[J].微电子学与计算机,2006,23(4):156-159.

[5] Stolcke A. SRILM-an extensible language modeling toolkit [C]// Proc. Intl. Conf. Spoken Language Processing. USA: Menlo park, 2002.

作者简介:

强　静　女,(1982-),硕士研究生.研究方向为自然语言处理与机器翻译.

张　建　男,(1954-),副研究员.研究方向为人工智能及其应用.

李　淼　女,(1955-),研究员,博士生导师.研究方向为人工智能与农业知识工程.