面向甲骨文的实例机器翻译技术研究*

袁 冬[1] 熊 晶[2] 刘永革[2]
[1](中国海洋大学计算机科学与技术系 青岛 266100)
[2](安阳师范学院计算机与信息工程学院 安阳 455002)

【摘要】提出基于实例的甲骨文释文机器翻译方案,研究实例库的构建流程、实例句相似度算法和实例检索算法等关键技术,并通过实现一个机器翻译系统,验证所提出方法的有效性。实验结果表明,该方法得到的翻译结果能够满足甲骨文学习者的阅读要求。

【关键词】甲骨文 基于实例的机器翻译 双语平行语料库 句子相似度
【分类号】TP391.2

Research on Example - based Machine Translation for Oracle Bone Inscriptions

Yuan Dong[1] Xiong Jing[2] Liu Yongge[2]
[1](Department of Computer Science and Technology, Ocean University of China, Qingdao 266100, China)
[2](School of Computer and Information Engineering, Anyang Normal University, Anyang 455002, China)

【Abstract】This paper introduces an Example - Based Machine Translation (EBMT) method for Oracle Bone Inscriptions (OBI). The key technologies in the example library building process, similarity algorithm of example sentences and example retrieval algorithm are proposed. And a machine translation system is developed to verify the proposed method. Experimental results show that the target sentences obtained by the machine translation can meet the demands of OBI researchers.

【Keywords】Oracle bone inscriptions EBMT Bilingual parallel corpus Sentence similarity

1 引 言

甲骨文距今已有3 500多年的历史,记载了商代王室的占卜记录,具有极其重要的史料价值[1]、学术价值和文化遗产保护价值[2]。作为我国迄今发现最早的一种成熟文字系统,甲骨文在古代汉语的研究和学习中发挥着重要的作用[3]。

甲骨文研究面临的首要问题是如何利用现代汉语理解和读懂甲骨文语句,国内外甲骨文专家指出将甲骨文用白话文释读很有意义[4]。但是从事甲骨文研究的门槛很高,培养一名甲骨文专家需要一、二十年甚至更长的时间[1],并且专家对甲骨文的辨识和翻译需要长期的学术钻研和经验积累,而这种经验知识仅存储在专家的头脑中,并不能实现知识的有效共享。针对这些问题,本文利用计算机技术和信息技术实现甲骨文白话释读,采用基于实例的机器翻译方法和技术,目的是有效共享和重用甲骨文专家的知识,降低甲骨文研究门槛,为甲骨文的研究和推广、提高甲骨文数字化展示等起到重要的推动作用。

收稿日期:2012 -03 -13
收修改稿日期:2012 -04 -11
*本文系国家自然科学基金项目"基于甲骨文语料库的计算机辅助考释技术研究"(项目编号:60875081)和河南省教育厅科学技术研究重点项目"基于本体的甲骨文知识共享平台构建方法研究"(项目编号:12A520003)的研究成果之一。

2　相关研究概述

机器翻译研究主要有基于规则的机器翻译(Rule - Based Machine Translation, RBMT)、基于实例的机器翻译(Example - Based Machine Translation, EBMT)和统计机器翻译(Statistical Machine Translation, SMT)三种[5]。RBMT 是依赖规则的,其"瓶颈"在于通过人工编写的方式获得大规模的语言规则成本太高,在研究上难以取得更大突破[6],而且甲骨文是迄今为止最早的成系统语言,很多文法规律还处于不确定状态[1],因此深层次的甲骨文规则的获取和维护比较困难;SMT 方法需要大规模的双语平行语料库作为训练各种概率参数的基础[7],但目前收集的甲骨文资料的规模还远远不够,而且一片甲骨上的文字最多百余字,最少的只有一个字,数据稀疏问题严重。由日本机器翻译专家长尾真于 20 世纪 80 年代提出的 EBMT 具有无需编写规则、系统维护容易、产生的译文质量较高、需要的语言知识较少等优点[5],是一个很好的选择。而且,甲骨文学习者都是通过已经存在的翻译实例作为知识源,来进行类比翻译和学习,这与 EBMT 当初的设计思想十分相符。

目前甲骨文信息处理方面的应用研究较多,并取得了一些成绩:江铭虎等[1]建立的甲骨文字库已收录 3 000 多字,对其中已考释的 1 000 多字用现代汉字、音、意、词性、属性等作出了详尽的标注解释;美国、中国香港和中国台湾等也进行了计算机甲骨文字库方面的研究[1];在计算机辅助甲骨文的缀合、考释、语料标注、甲骨文字编辑、文字库构建等方面也有了不少的研究成果[8-12]。

但是,甲骨文机器翻译方面的研究极为少见,目前最相近的是"汉字叔叔"网站(http://www.chineseetymology.org),但只能实现现代汉字到古代汉字的映射。国内已有针对古籍文字的机器翻译研究,如王爽等[13]设计和实现的基于实例的古文机器翻译系统 EBMTAC,避免了复杂的深层次语法树和语义分析;文献[14]利用统计模型研究了中国古诗在英文翻译时韵律自动选择的问题;郭锐等[15]指出,快速准确地构建大规模古今汉语平行语料库以及检索与输入句子最相似的源句子是基于实例的古今汉语机器翻译必须解决的问题。宋继华等[16]从语料的设计、采集、格式化存储、双语对齐与 XML 标注等方面研究了大型古今汉语平行语料库的构建方法。

甲骨文作为最早具备汉语语法体系的文字,虽有很多特征被延续到后代传世文献,但也有区别于其他古籍的一些特点:一字异形;异字同形;合文普遍,即两个或三个字刻在一起,在行款上只占一个字的位置[17];少数高频字占总字量的高比重和在总字量中占极低比重的低频字占单字总数的极高比重的两端集中特征[18];特有的三宾动词[19];完整的卜辞有前辞、命辞、占辞、验辞 4 个部分,但是大多数卜辞都省略了某些部分,常见的卜辞只保留了前辞和命辞[17]。因此,同其他古籍翻译技术相比较,基于实例的甲骨文的机器翻译还需要结合部分规则、翻译记忆技术和小样本机器学习技术如 SVM 等。

3　实例库的建立

基于实例的机器翻译需要有一定规模的实例库,因此需要对目前已经过甲骨文专家翻译的、在学术界不存在争议的甲骨文语句进行收集和整理。甲骨文原文是没有句读的,且甲骨字很多都是异体字,而甲骨文释文则是经专家考释过的与原文对应的简体或繁体中文(没有考释出来的甲骨字仍然以原始形态出现)。释文已经添加了句读符号,统一了异体字的表示形式,并且对一些残缺的或错刻的甲骨文字进行了补充,因此本文选择甲骨文释文作为机器翻译的源语言,并建立"甲骨文释文 - 现代汉语"的双语平行语料库,对语料库进行处理后,生成服务于机器翻译的实例库。

"甲骨文释文 - 现代汉语"双语平行语料库中的甲骨文释文例句均摘自于真实甲骨文片上的卜辞,其对应的现代汉语翻译均从权威的已公开发表的甲骨文著作中收集。目前的实例库收录了来自著作《甲骨文精粹释译》的 692 片甲骨的共 2 425 条卜辞及其对应的 2 425 个现代汉语翻译句,通过人工方式完成句子对齐和短语对齐,词对齐为自动对齐辅以人工校对,因而实例库质量较高。实例库构建流程如图 1 所示。

3.1　语料库预处理

甲骨文机器翻译的源语言为甲骨文释文,目标语言为现代汉语,两者均以 XML 文件格式存储,分别用 source.xml 和 target.xml 表示。由于甲骨片上的卜辞语句均为短句,且同一甲骨片上可能记录着多次占卜

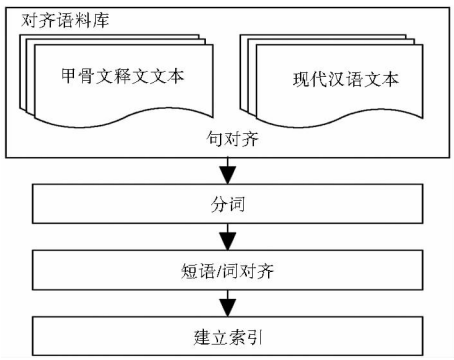

图 1　实例库建立流程

的内容,因此采取下列标记方法:

<OBI> </OBI> 表示 XML 文件中的根元素。

<bone id = "" > </bone > 元素记录某一甲骨片上的卜辞内容,id 表示甲骨片号,这个编号在甲骨文研究中是唯一的。

<u id = "" > </u > 元素位于 <bone > </bone > 元素之间,记录一条卜辞;id 取值为 1,2,3…,表示句子所属的每条卜辞单元编号。

<s id = "" > </s > 元素位于 <u > </u > 标记之间,表示甲骨文释文句子;id 取值为 1,2,3…,表示句子编号。

source. xml 和 target. xml 中均是句子对齐的,因此,两者中的各级 id 值是一一对应的,句子对齐采用人工对齐的方法。一个简单 source. xml 的例子如下:

```
<OBI>
  <bone id = "H12324 正" >
    <u id = "1" >
      <s id = "1" > 丁巳卜,亘,贞:自今至于庚申其雨? </s >
      <s id = "2" > 贞:自今丁巳至于庚申不雨? </s >
    </u >
    <u id = "2" >
      <s id = "1" > 戊午卜,㱿,贞:翌庚申其雨? </s >
      <s id = "2" > 贞:翌庚申不雨? </s >
    </u >
  </bone >
  …
</OBI >
```

其对应的 target. xml 如下:

```
<OBI>
  <bone id = "H12324 正" >
    <u id = "1" >
      <s id = "1" > 丁巳日占卜,贞人亘问卦,贞问:从今天到庚申日会下雨吗? </s >
      <s id = "2" > 贞问:从今天丁巳日到庚申日不会下雨吗? </s >
    </u >
    <u id = "2" >
      <s id = "1" > 戊午日占卜,贞人㱿问卦,贞问:后天庚申日会下雨吗? </s >
      <s id = "2" > 贞问:后天庚申日不会下雨吗? </s >
    </u >
  </bone >
  …
</OBI >
```

3.2　分词及词对齐

甲骨文分词采用基于词典、句法规则和句法分析相结合的办法[8, 20],首先通过甲骨文词典获得初步的分词结果,然后根据甲骨文句法规则和句法分析对初步结果进行再次划分,划分的结果通过句法分析排歧后再进行分词序列优化,得到最终的分词结果。通过未登录词识别规则进行检查,将满足未登录词条件的新词加入词库。笔者曾提出的甲骨文分词流程如图 2 所示:

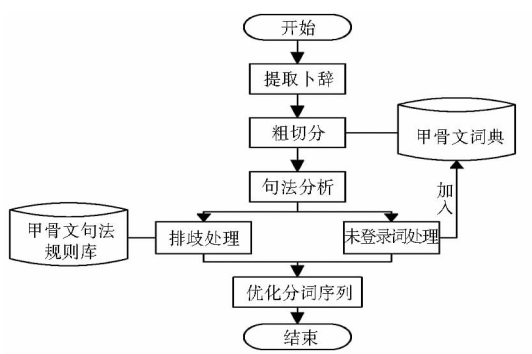

图 2　甲骨文分词流程[20]

甲骨文词典中记录了双语对译信息,目前共收录词条 4 881 个(含异体字和合文),其中单字词 4 687 个,二字词 174 个,三字词 20 个。通过查找甲骨文词典,可得到甲骨文词语对应的现代汉语词汇。由于甲骨文中单字词的数量较多,因此,词对齐的准确率较高。实验表明,该方法的准确率、召回率和 F 值分别达到 97. 49%、97. 61% 和 97. 55%[20]。词对齐效果依赖于甲骨文词典的完善程度,利用词典进行自动词对齐后,一般要辅以人工校对。

3.3　建立索引

建立索引的目的是为实例搜索提供基础。建立索引主要包括按句子排序的索引和按词排序的索引[5]。

按句子排序索引是基于语料库的,按词排序的索

引则基于词表。为便于检索,实例库的最终形式不是文本,否则在检索时效率很低。词表中存储了词的序号,索引时,所有的词将被词序号代替。建立词表后,语料库中的句子也采用序号表示。在生成的实例库索引中,包含了实例的源句子、目标句子和对齐信息[5]。

随着甲骨文研究的不断深入,学者将不断考释出新的甲骨文句子,实例库也因此不断扩充。对实例库中没有收录的例句,若经计算机翻译能得到较为满意的结果,也可以扩充到实例库中。因此,实例库的建立是一个动态完善的过程。

4　甲骨文释文机器翻译

4.1　机器翻译流程

甲骨文释文机器翻译流程主要有以下几个关键步骤:实例检索、实例匹配、片段组合、翻译评价等,如图 3 所示:

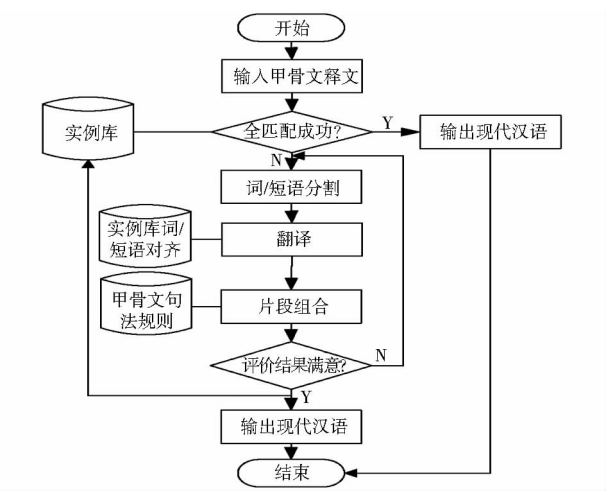

图 3　基于实例的甲骨文释文机器翻译流程

图 3 中的实例匹配有两种情况:全匹配和部分匹配。全匹配是指待翻译句子与实例句相同,此时的翻译过程就是实例检索过程。鉴于甲骨文句子是有限的,在实例库规模足够大的前提下,全匹配成功率可达 100%。部分匹配是指待翻译句子与实例句相似,此时,分别找到与实例句匹配的部分和不匹配的部分,并计算句子相似度。选取相似度最高的实例句,将其对应的现代汉语译文进行词/短语替换,并利用甲骨文句法规则进行组合和调整,得到最终翻译结果。若经过评价,认为翻译结果符合要求,则将其添加到实例库当中。

4.2　实例检索

实例检索是翻译流程中的关键问题,即如何从实例库中检索出与待翻译甲骨文句子相同或相似的实例。实例检索存在两个关键技术[5]:相似度计算和检索算法。

(1)相似度计算

用 S_o 表示输入的甲骨文释文句子,用 S_e 表示实例库中的实例句子。针对甲骨文的特点,考虑到 S_o 来自于某片甲骨而不是用户随意组合的,因此主要从匹配组块和编辑距离两个方面进行句子比较。基于文献[15]和文献[21]的方法,本文采用如下的相似度计算公式:

$$Sim(S_o,S_e)=\alpha\times\frac{2WordCom(S_o,S_e)}{WordNum(S_o)+WordNum(S_e)}+\beta\times\left[1-\frac{EditDist(S_o,S_e)}{len_{max}(S_o,S_e)}\right] \quad (1)$$

其中,$WordCom(S_o,S_e)$ 表示 S_o、S_e 两者中相匹配的词语数量;$WordNum(S_o)$ 和 $WordNum(S_e)$ 分别表示 S_o、S_e 两者中的词语个数;$EditDist(S_o,S_e)$ 为 S_o、S_e 之间的编辑距离,指仅通过插入、删除、替换操作,把一个字符串变成另一个字符串所需要的最小操作数目[15];$len_{max}(S_o,S_e)$ 为 S_o、S_e 两者中长度的最大值;α 和 β 为权重参数,且 $\alpha+\beta=1$。在选择 α 和 β 时,参考文献[15, 22, 23]的方法,基于人工对齐 的 2 425 句对,采用遗传算法进行确定。具体参数为:染色体编码采用二进制,编码长度为 20;初始群体大小设定为 50;算子选择过程采用赌轮盘选择方法;交叉概率取 0.7;变异概率取 0.001;算法终止条件为最优个体在连续 10 代没有改进或平均适应度在连续 10 代基本没有改进时停止。

(2)检索算法

检索算法主要用于实例部分匹配的情况。为考虑实例的检索效率,采用词的倒排索引进行搜索。即针对待翻译句子中出现的词,查找所有出现这些词的实例句子,然后只计算这些句子的相似度[5]。

但是,甲骨文中有些词在卜辞语句中频繁出现,检索这些词将对应着大量的实例句子。为避免高频词查找返回过多结果,又保证尽可能不遗漏潜在的相似例句,因此引入词的信息熵[15]:

$$H(ch)=\lg\left(\frac{M}{m}\right) \quad (2)$$

其中,ch 表示一个词,M 表示语料库中的甲骨文

释文句总数(目前收集的卜辞为来自 72 112 片甲骨上的共 129 519 条句子),m 表示释文中出现 ch 的句子数。引入信息熵的概念后,可以计算释文中各词的信息熵,高频词有较低的信息熵。设定其最小阈值 D,信息熵低于 D 的词将不再参与检索。出现频率最高的前 10 个甲骨文词语及其信息熵如表 1 所示:

表 1　前 10 个频率最高的甲骨文词语及其信息熵

甲骨文词语	出现次数	语料库卜辞句总数	信息熵
贞	53 877	129 519	0.381
卜	46 827	129 519	0.442
□(表示残字)	31 731	129 519	0.611
王	27 788	129 519	0.668
亡	24 853	129 519	0.717
一	18 396	129 519	0.848
其	16 701	129 519	0.890
二	16 688	129 519	0.890
于	12 982	129 519	0.999
三	12 511	129 519	1.099

实例句检索算法描述如下:

①将输入的待翻译句子 S_0 进行分词,剔除信息熵小于阈值 D 的词,得到词集合 W;

②对每个词 $w_i \in W$,通过词的倒排索引检索出所有包含 w_i 的实例句,得到句子集合 S_i;

③求 S_i 的并集得到句子集合 S;

④对每个句子 $s_i \in S$,利用公式(1)求出 $Sim(S_0, s_i)$ 并按降序排列;

⑤取 $Sim(S_0, s_i)$ 值最大的句子 s_i 作为目标句。

4.3　翻译结果评价

目前,还没有针对甲骨文的机器翻译自动评价机制,对经过部分匹配得到的甲骨文释文机器翻译结果,需要进行人工评价。人工评价关注“忠实度”和“可理解度”两个方面,前者考察的是译文忠实原文表达意图的程度,后者则考察存在残缺字、未释字等的甲骨文语句经系统翻译后能达到通读全句的程度。同目前最常用的两种基于 n 元匹配的自动评测方法 BLEU[24] 和 NIST[25] 所采用的只对译文在字面字形上的相似性度量[26] 相比较而言,本文的评价方法并没有提供多个参考译文用于给系统译文进行打分,而是采用对照甲骨文专著或咨询甲骨文专家的方式,因此对甲骨文专家的依赖性较高,但对系统译文翻译质量考察的目标要求则相对较低。若评价结果满意,可以将翻译结果及其对应的源语言句子作为新的实例句对添加到实例库中。

5　实验及分析

最终的双语实例句对存储在关系数据库中,若是全匹配方式,则直接从数据库中检索得到翻译结果。如输入待翻译的句子“贞:帝弗其及今四夕令雨?”,得到的结果如图 4 所示:

图 4　全实例匹配翻译结果

若为部分匹配方式,则利用检索算法得到目标句(公式(1)中 $\alpha = 0.44, \beta = 0.56$),借助甲骨文双语词典对其进行词替换后,经过调整生成最终译文。如输入待翻译句子“丙子卜,韦,贞:我受年?”,由于实例库没有收录此句,通过检索算法从实例库中检索出最相似的例句为“□□ 卜 韦 贞 我 受年”(下划线部分为匹配部分),此实例句对应的现代汉语翻译为“某日占卜,贞人韦问卦,贞问:我商王朝会丰收吗?”,将不匹配的词进行替换并调整,得到最终翻译结果为“丙子日占卜,贞人韦问卦,贞问:我商王朝会丰收吗?”。目标例句的选择依据如图 5 所示:

图 5　实例句相似度计算结果

本文随机选取 30 句甲骨文释文进行实验,结果为:常规的甲骨文句子通过全实例匹配或部分实例匹配,其平均正确率为 84.5%(其中全实例匹配的正确率为 100%),已能满足信息工作者的研究需求。但是,对于存在省刻、错刻、缺刻以及残辞等现象的甲骨

文句子,翻译结果并不理想,平均正确率仅为 30.1%(全实例匹配的正确率仍为 100%)。而且,对有歧义的甲骨文句子,此方法无法较好地完成词义消歧。这些将作为特殊句于下一步进行专门研究。

6　结　语

本文针对目前甲骨文信息化处理研究中存在的问题,提出基于实例的甲骨文释文机器翻译技术研究方案,目的是充分共享和重用现有的甲骨文专家知识,减轻甲骨文专家的工作负担,降低甲骨文研究门槛。本文详述了实例库建立的关键技术、实例检索及相似度计算方法。但目前的实例库规模较小,且对齐方法及翻译评价均是人工操作的,在今后的研究工作中,将考虑实例的自动对齐技术,不断扩充实例库,并在此基础上研究面向甲骨文释文机器翻译的自动评价机制,进一步减少对甲骨文专家的依赖。同时,研究特殊甲骨文语句的翻译方法。

参考文献:

[1] 江铭虎. 自然语言处理[M]. 北京:高等教育出版社,2006. (Jiang Minghu. Natural Language Processing[M]. Beijing: Higher Education Press, 2006.)

[2] 顾绍通. 甲骨文数字化处理研究述评[J]. *西华大学学报:自然科学版*,2010, 29(5): 38 - 42. (Gu Shaotong. Review on Digitization Processing of Jiaguwen[J]. *Journal of Xihua University: Natural Science Edition*, 2010, 29(5): 38 - 42.)

[3] 陈光田. 古文字与古代汉语学习的关系研究[J]. *新乡学院学报:社会科学版*,2010,24(4):125 - 127. (Chen Guangtian. Discussion on the Function of Ancient Words in Learning Ancient Chinese[J]. *Journal of Xinxiang Teachers College: Social Sciences Edition*, 2010,24(4): 125 - 127.)

[4] 王宇信, 杨升南, 聂玉海. 甲骨文精粹释译[M]. 昆明: 云南人民出版社, 2004. (Wang Yuxin, Yang Shengnan, Nie Yuhai. Oracle Bone Inscriptions Pithiness Explanation[M]. Kunming: Yunnan People's Publishing House, 2004.)

[5] 侯宏旭, 刘群, 那顺乌日图. 基于实例的汉蒙机器翻译[J]. *中文信息学报*, 2007, 21(4): 65 - 72. (Hou Hongxu, Liu Qun, Nasun Urt. Example Based Chinese - Mongolian Machine Translation[J]. *Journal of Chinese Information Processing*, 2007, 21(4): 65 - 72.)

[6] 刘群. 汉英机器翻译若干关键技术研究[M]. 北京: 清华大学出版社, 2008. (Liu Qun. Research on Some Key Techniques in Chinese - English Machine Translation [M]. Beijing: Tsinghua University Press, 2008.)

[7] 姜迎春,雪艳. 词语对齐与机器翻译问题研究——以汉蒙机器翻译为例[J]. *民族翻译*, 2010(1): 91 - 95. (Jiang Yingchun, Xue Yan. Word Alignment and Machine Translation Research - Taking Chinese - Mongolian as Example[J]. *Minority Translators Journal*, 2010(1): 91 - 95.)

[8] Cai H Y, Jiang M H, Deng B X, et al. Method Combining Rule - based and Corpus - based Approaches for Oracle - bone Inscription Information Processing[C]. In: *Proceedings of the* 2006 *International Conference on Intelligent Computing* (*ICIC*'06). Berlin, Heidelberg: Springer - Verlag, 2006: 736 - 741.

[9] 马小虎,杨亦鸣,黄文帆,等. 甲骨文轮廓字形生成技术研究与通用甲骨文字库的建设[J]. *语言文字应用*,2004(3): 105 - 111. (Ma Xiaohu, Yang Yiming, Huang Wenfan, et al. Research on the Technology of the Automatic Generation "Jiaguwen" Outline Font and Building of Universal Jiaguwen Font[J]. *Applied Linguistics*, 2004(3): 105 - 111.)

[10] 顾绍通,马小虎,杨亦鸣. 基于字形拓扑结构的甲骨文输入编码研究[J]. *中文信息学报*,2008,22(4): 123 - 128. (Gu Shaotong, Ma Xiaohu, Yang Yiming. Topological Frame Based Input Method Coding of Jiaguwen[J]. *Journal of Chinese Information Processing*, 2008,22(4): 123 - 128.)

[11] Wang A M, Wang J P, Ge Y Q, et al. Research on Intelligent Oracle Bone Fragments Rejoining Technology [J]. *Applied Mechanics and Materials*, 2011, 50 - 51: 594 - 598.

[12] Gao F, Liu Y G, Xiong J. Ontology - based Semantic Annotation for Oracle Bone Inscriptions[C]. In: *Proceedings of the* 2*nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce* (*AIMSEC*). 2011: 5087 - 5090.

[13] 王爽, 熊德兰, 王晓霞. 基于实例的古文机器翻译设计与实现[J]. *许昌学院学报*, 2009, 28(5): 88 - 91. (Wang Shuang, Xiong Delan, Wang Xiaoxia. The Research and Implemention of Example - based Machine Translation of Ancient Chinese[J]. *Journal of Xuchang University*, 2009, 28(5): 88 - 91.)

[14] Fang M, Jiang X, Zhao Q, Jiang Y. Automatic Choosing of English Rhymes in Translation of Chinese Ancient Poems[C]. In: *Proceedings of the* 2*nd International Symposium on Knowledge Acquisition and Modeling*(*KAM*'09). 2009: 434 - 437.

[15] 郭锐, 宋继华, 廖敏. 基于自动句对齐的相似古文句子检索[J]. *中文信息学报*, 2008, 22(2): 87 - 91. (Guo Rui, Song Jihua, Liao Min. Ancient Sentence Search Based on Sentence Auto - Alignment in Parallel Corpus of Ancient and Modern Chinese[J]. *Journal of Chinese Information Processing*, 2008, 22(2): 87 - 91.)

[16] 宋继华, 胡佳佳, 孟蓬生, 等. 古今汉语平行语料库的语料构

建[J]. *现代教育技术*,2008,18(1):92-99.(Song Jihua, Hu Jiajia, Meng Pengsheng, et al. The Construction of Corpora in a Classic-Cotemporary Chinese Parallel Corpus[J]. *Modern Educational Technology*, 2008,18(1):92-99.)

[17] 刘一曼. 甲骨文字的特点及主要内容[J]. *档案管理*, 2000(1):40-41.(Liu Yiman. The Characteristics and Main Content of Oracle Bone Inscriptions[J]. *Archives Management*, 2000(1):40-41.)

[18] 刘志基. 简论甲骨文字频的两端集中现象[J]. *语言研究*, 2010,30(4):114-122.(Liu Zhiji. On the Two Concentration Features of the Character Frequency of Oracle Bone Inscriptions[J]. *Studies in Language and Linguistics*, 2010,30(4):114-122.)

[19] 郑继娥. 甲骨文祭祀卜辞语言研究[M]. 成都:巴蜀书社, 2007.(Zheng Ji'e. Oracle Bone Sacrifice Inscriptions Language Research[M]. Chengdu: Bashu Publishing House, 2007.)

[20] Xiong J, Gao F, Liu Y. Word Segmentation Method for Oracle Bone Inscriptions Based on Dictionary and Syntactic Rules[C]. In: *Proceedings of the* 3*rd International Conference on Computer Design and Applications* (*ICCDA*). IEEE, 2011: 592-595.

[21] 吕学强, 任飞亮, 黄志丹, 等. 句子相似模型和最相似句子查找算法[J]. *东北大学学报:自然科学版*, 2003, 24(6):531-534.(Lv Xueqiang, Ren Feiliang, Huang Zhidan, et al. Sentence Similarity Model and the Most Similar Sentence Search Algorithm[J]. *Journal of Northeastern University: Natural Science*, 2003, 24(6):531-534.)

[22] 贾兆红, 陈华平. 基于改进遗传算法的权重发现技术[J]. *计算机工程*,2007,33(5):156-157.(Jia Zhaohong, Chen Huaping. Weights Finding Based on Improved Genetic Algorithms[J]. *Computer Engineering*, 2007,33(5):156-157.)

[23] 张刚, 杨海成, 经小川, 等. 基于遗传算法的技术成熟困难度计算方法[J]. *北京理工大学学报*, 2011, 31(4):472-476.(Zhang Gang, Yang Haicheng, Jing Xiaochuan, et al. Computation of Advancement Degree of Difficulty Based on Genetic Algorithm[J]. *Transactions of Beijing Institute of Technology*, 2011, 31(4):472-476.)

[24] Papineni K, Roukos S, Ward T, et al. BLEU: A Method for Automatic Evaluation of Machine Translation[C]. In: *Proceedings of the* 40*th Annual Meeting of the Association for Computational Linguistics*(*ACL*). 2002: 311-318.

[25] Doddington G. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics[C]. In: *Proceedings of the* 2*nd International Conference on Human Language Technology Research*. San Francisco: Morgan Kaufmann Publishers Inc., 2002: 138-145.

[26] 黄瑾, 刘洋, 刘群. 机器翻译评测介绍[EB/OL]. [2012-02-05]. http://lib.ict.ac.cn/libraryict/ITL/data/2007/5/机器翻译评测介绍.pdf.(Huang Jin, Liu Yang, Liu Qun. Introduction to Machine Translation Evaluation[EB/OL]. [2012-02-05]. http://lib.ict.ac.cn/libraryict/ITL/data/2007/5/%BB%FA%C6%F7%B7%AD%D2%EB%C6%C0%B2%E2%BD%E9%C9%DC.pdf.)

(作者 E-mail:yuandong1222@gmail.com)