文章编号：1003-0077(2007)04-0065-08

基于实例的汉蒙机器翻译

侯宏旭[1,2,3]，刘 群[1]，那顺乌日图[2]

(1. 中国科学院 计算技术研究所 智能信息处理重点实验室 北京 100080；
2. 内蒙古大学 内蒙古 呼和浩特 010021；3. 中国科学院 研究生院 北京 100080)

摘 要：本文通过对汉蒙机器翻译方法的研究，给出了一种基于实例的汉蒙机器翻译方法，并加以了实现。本文给出了用于汉蒙EBMT机器翻译的实例搜索以及短语片段划分、匹配、组合的方法。本文给出的方法是基于词语对齐的，利用词语对齐进行词语的匹配，并根据匹配词数和长度计算相似度，选取最好的实例。通过对齐信息，确定片段组合的策略，生成翻译结果。通过对方法的实现和实验，完成了一个基于实例的汉蒙机器翻译系统。

关键词：人工智能；机器翻译；蒙古语；基于实例；词语对齐

中图分类号：TP391 **文献标识码**：A

Example Based Chinese-Mongolian Machine Translation

HOU Hong-xu[1,2,3], LIU Qun[1], Nasun Urt[2]

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China;
2. Inner Mongolia University, Huhhot 010021, Inner Mongolia AR, China;
3. Graduated University of Chinese Academy of sciences, Beijing 100080, China)

Abstract: We have presented an example based Chinese-Mongolian machine translation method, and implemented it. The method is insist of several parts, includes example searching, segment splitting, matching and recombining. The method is based on word alignment. It is using word alignment information for segment matching, and computing the similarity by the number of matching words and length, and selects the best example. Using word alignment information, determined the method of segment recombining, and generates the translation result.

Key words: artificial intelligence; machine translation; mongolian; example based; word alignment

1 引言

蒙古语是我国重要的少数民族语言，也是蒙古国的官方语言，在国际上具有很重要的地位。作为我国蒙古族自治地方的官方语言，使用的范围是非常大的。这样，蒙古语的信息处理就成为了语言信息处理的重要研究方面。其中，汉蒙机器翻译是将汉语资料翻译成蒙古语的重要工具。但是，由于蒙古语自身的原因，以及研究投入不够的问题，导致汉蒙机器翻译的研究进展相对比较缓慢。目前，国内外汉英机器翻译的研究如火如荼，我们正好可以借鉴这些技术，加快汉蒙机器翻译的研究。

目前国内的少数民族语言机器翻译的研究还处在初级阶段，主要的研究还是集中在基于规则的方法上[2]。在蒙古语机器翻译的研究上，国内外的研究和相关文献都比较少，目前的研究还主要集中在研究的基础阶段[1]。尤其是在基于语料库的方法上还没有深入的研究，而且，在由内蒙古大学和中国科学院计算技术研究所合作研究的基于规则的汉蒙机器翻译，是目前我们能够了解到的真正达到部分实用的汉蒙机器翻译技术，但是其方法上能够取得的进展也依然比较有限，无法达到比较好的实用效果[2]。因此，基于语料库的汉蒙机器翻译方

收稿日期：2007-02-01 **定稿日期**：2007-05-10

基金项目：国家自然基金资助项目(60573188)；国家863计划资助项目(2003AA115510)

作者简介：侯宏旭(1972—)，男，硕士，副教授，主要研究方向为自然语言处理；刘群(1966—)，男，博士，研究员，主要研究方向为自然语言处理；那顺乌日图(1959—)，男，博士，教授，主要研究方向为蒙古文信息处理、计算语言学。

万方数据

法就非常值得研究。

国际、国内在机器翻译的研究方面主要集中在RBMT、EBMT和SMT三种基本方法上，其中EBMT和SMT是基于语料库的，RBMT是基于规则的。早年，国内外的机器翻译工作都是集中在RBMT方面的，但是RBMT具有扩充困难等问题，所以目前相关研究较少。EBMT和SMT是基于语料库的，其中EBMT是80年代长尾真提出的一种方法，它具有构造容易，容易生成高质量译文等特点。SMT则是由Brown等提出的方法。在汉蒙机器翻译方面，由于蒙古语和汉语都属于语法比较复杂的语言，所以利用基于语料库的方法可以尽量减少语言知识上面的投入，能够尽快建立比较好的翻译系统。作为今后SMT研究的基础，我们选择了更容易实现和看到效果的EBMT作为研究的切入点。

相对于汉英机器翻译来说，汉蒙机器翻译的特点和难点主要集中在以下几个方面：

蒙古语的语序[7]。蒙古语具有谓语在句子的末尾的特点，所以在以短语为单位进行翻译的时候，存在长距离的调序。而相对来说汉英机器翻译中这样的调序并不多。这样，就给汉蒙机器翻译的调序带来相对更大的困难。

蒙古语的词形变化[8]。蒙古语的词形化非常复杂，动词、名词都存在时态、数、格等多种变化，这些变化是通过添加词缀的方法实现的。这些变化在使用统计的机器翻译方法时就会造成比较多的词形错误。相对来说汉英机器翻译中主要考虑的只是单复数等比较简单的变化，而且相对数量比较少。

很多蒙古语词，尤其是动词，存在着纷繁的变化形式。这些变化构成了复杂的词干—多词缀形式。例如，一个动词词干，后面可以添加格、数、人称、时态等多种词缀，例如，在一个20万词的语料中，动词词干ILA可以衍生出来的词有ILABA、ILAGAD、ILAGDABA等10多种形式，而且这还不是全部。因此单独以词(包含词干词缀的词)为单位来进行汉蒙机器翻译是远远不够的。

和RBMT相比，EBMT具有易于构造的优点。由于蒙古语的特殊复杂性，编写蒙古语的翻译规则是非常复杂的事情，难于得到好的效果。通过先前基于规则的汉蒙机器翻译的尝试，我们发现，虽然基于规则的方法也取得了比较好的结果，但是，一方面规则的维护非常复杂，进一步提高翻译系统的效果需要的工作量非常庞大，另一方面在语料不断的积累中，这些语料很难应用到系统中，无法从语料规模的扩大中获益。

和SMT相比，对于相似度比较高的句子，EBMT具有更好的效果。由于EBMT是基于实例的，因此，如果能够选择到比较好的实例，那么经过简单的替换就可以生成非常好的翻译结果。但是，如果不能找到好的实例，那么翻译效果就会变差。因此，结合EBMT和SMT是比较好的思路。这里，我们给出的EBMT实现的汉蒙机器翻译就可以为今后的工作打下基础。

本文第2节将给出这个基于实例的汉蒙机器翻译系统的总体架构，第3节给出构建实例库的方法，第4节给出实例的匹配和搜索的方法，第5节给出片段的匹配和组合的关键方法，第6节给出候选翻译结果的评价方法，第7节给出系统的实现及初步实验的结果。

2 总体架构

我们知道，EBMT具有以下的主要优点[3~5]：

- 不需要编写规则
- 系统维护容易
- 容易产生高质量的译文
- 需要的相关语言知识少

通过以往的尝试，我们发现，由于汉语和蒙古语分属不同的语系，语言的差别相对比较大，编写规则相对来说比较困难，调试起来工作量比较大。因此，选择基于语料库的机器翻译方法就是比较合适的。

通过分析论证，考虑到我们以前研究的汉蒙双语对齐的技术基础，我们采用了基于对齐的EBMT系统。

在系统的架构中包含几个主要的处理步骤：

(1) 分词和对齐

将待翻译的句子切分成以词为单位的片段。在本系统中，汉语的分词采用的是中国科学院计算技术研究所研发的ICTCLAS汉语分词系统。在最终系统中，蒙古语按空格分词，不做特殊处理。将双语语料库中的汉蒙句对进行词对齐。这里是利用汉蒙双语词典及共现概率为基础的方法进行词对齐。经过对齐后，将双语语料库转换为实例库，为机器翻译提供实例。

(2) 实例搜索

从实例库中所有最接近的实例。这一步的主要内容包含相似度的计算和搜索两个部分。

(3) 片段匹配、分割和组合

万方数据

在待翻译句子和实例中查找匹配和不匹配的片段。根据匹配和不匹配的片段确定翻译结果片段。将翻译结果片段组合成翻译结果。

(4) 评价

从候选的翻译结果中选择最佳的翻译结果。

3 实例库生成

获得双语语料库后，需要对语料库进行处理，生成实例库。

3.1 原始语料库

采用拉丁转写的原始语料库的存储格式是以 xml 格式。原始语料库最初是由源语言文本文件和目标语言文本文件组成的，它们都是每行一个句子，在自动对齐时生成 xml 格式的语料库。

为了蒙古语的处理方便，蒙古语的存储方式为拉丁转写方式。拉丁转写是采用英文字母和数字 0 作为蒙古语的表示方法。由于采用了 ASCII 表示蒙古语，在存储和处理时就非常方便了。在具体显示的时候，再转换为蒙古文的显现格式。

3.2 分词

将待翻译的句子切分成以词为单位的片段。在本系统中，汉语的分词采用的是中国科学院计算技术研究所研发的 ICTCLAS 汉语分词系统。在最终系统中，蒙古语按空格分词，不做特殊处理。事实上，单从词的基础上进行蒙古语机器翻译的研究还是不够的，我们需要在词根、词干、后缀的层次上才能得到蒙古语更深入的研究，这也是我们将来的研究目标之一。

其中，汉语分词结果中包含每个词的词性，词性的标记集是 ICTCLAS 的词性标记集。蒙古语没有标注词性。

3.3 词对齐

本系统的词语对齐采用了大规模汉蒙词典。双语词典含有大量的词语互译信息，用双语词典进行词语对齐往往准确率很高。由于规模的限制，双语词典的词汇覆盖面往往不够，因此用双语词典进行词语对齐有召回率不高的缺点。在我们的方法中，利用双语词典计算的词语相似度、位置等信息进行词语对齐，并通过对齐窗口得到了多对多的词语对应。通过这样的方法得到的较好的词语对齐结果。

经过自动对齐以后的结果可能存在一些错误，可以通过人工对齐工具进行校对。

3.4 建立索引

建立索引是语料库处理的关键步骤[3]。通过索引的建立为实例库的搜索提供基础。实际上，建立的索引包括两个，一个是以句子排序的索引，另一个是按词排序的索引。

(1) 生成索引

生成索引即生成语料库的索引。索引的内容是以句对排序的。

(2) 生成词表

最终的实例库不是以文本形式存储的。如果以文本形式存储，那么查找需要的时间代价是比较高的，因此，在建立索引的时候，所有的词将被词的序号所代替。这个序号存储在词表中。

倒查表中存储词表中所有的词在哪些双语语料的句子中出现。即纪录某个词出现的所有句子的序号。

(3) 生成倒排索引

建立了词表以后，语料库中的句子将被转换成以序号表示的形式。这样在查找的时候速度会提高很多。

在实例库索引中，包含实例的源句子、目标句子和对齐信息。

3.5 双语词典

对于不能从实例库中获得的翻译，可以利用双语词典进行翻译。也就是说，如果一个待翻译的句子无法找到接近的实例的话，将会退化成基于词典的翻译。

双语词典包含源语言词、源语言词性、目标语言词、目标语言词性(蒙古语的词性是忽略的)以及一个致信度。

4 实例搜索

EBMT 的关键点之一就是从庞大的实例库中搜索到所有相似的实例。这里主要包含两个指标：正确性和完备性。也就是说，一方面要尽量搜得所有相似的实例，另一方面要更准确的评价相似程度。因此，在实例搜索中就存在两个主要的内容：相似度计算和搜索算法。

相似度计算的方法很多，例如利用功能词的相似度计算方法，利用编辑距离的相似度计算方法

万方数据

等[3~6]。在设计算法的时候，我们既要考虑到计算的可靠性，又要注意计算的速度[10]。因此，我们采用了下面的方法。

4.1 相似度计算

在本系统中，采用了片段分割与组合的方法。即，将实例划分成几个片段，分别得到这些片段的翻译，然后组合成最终的句子。这样，相似度的主要评价指标就是如何使得匹配上的片段最多、最长。因此，设计的相似度计算公式如下：

$$m = \sum w(pos(i)) \times match(i) \times w2(i)$$

这里，$pos(i)$是相应词的词性，$w2$是行程长度。

在评价相似度的三个元素中，首先是词性。举例来说，一个明显的结论是匹配上的动词要比匹配上的名词更重要。

例如，两个实例"我 有 一 支 笔"和"我 拿走 了 一 支 铅笔"对于待翻译的句子"我 有 一支 铅笔"来说，显然动词匹配的第一个实例更符合要求。因此在相似度计算的第一个参数就是词性的权重。通过实验，我们给出了各个词性的权重。以下是几个典型的词性权重：

表1 几个典型词性的权重

V动词	P介词	W标点	N名词	M数词
2.0	1.5	1.2	0.5	0.6

第二个既是匹配值。

$$match(i) = \begin{cases} 0 & \text{if } w(i) = e(i) \\ 1 & \text{if } w(i) \neq e(i) \end{cases}$$

其中$w(i)$是待翻译句子中的词，$e(i)$是实例句子中的词。

第三个是行程长度。考虑到相邻匹配的越多，那么相似程度越高，所以在计算相似度时，还要考虑连续匹配的长度。

举例来说，实例"我 有 一 支 笔"和"我 还 有 一 支 蓝色 的 铅笔"对于待翻译的句子"我 有 一 支 铅笔"来说，虽然后一个实例的匹配词更多，但是前一个实例的匹配片段要长，实际上，还是前一个实例更合适。

4.2 实例搜索

由于实例空间相对比较大，所以搜索时要考虑搜索的效率。

通常的搜索方法就是顺序搜索，这样将计算所有实例的相似度。这样的方法最准确，但是效率也最低。

我们采用的方法是，利用词的倒排索引进行搜索。其基本方法是，根据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。

这样做有两个问题，一个是，即使是用这样的方法，出现某些词，例如"的"，"一"等的句子非常多，还是不能有效地减少搜索量。另一个是，可能漏掉匹配比较好的句子，例如实例"我 讨厌 狗"和待翻译句子"他 喜欢 书"，可以采用一一替换的方式进行翻译，可能是比较好的例子。

对于第二个问题，我们认为，这样的匹配实例无法确定其真正的结构相似性，即使获得了相应的实例也很难得到较好的翻译结果。所以，在实际处理中，这种情况不考虑。也就是说，只考虑有多个词匹配的情况。

对于第一个问题，常见词由于在很多实例中出现，对于评价句子的匹配程度的贡献是非常小的，因此在匹配的时候，这些词都被过滤掉了。

目前，被过滤掉的词性包含 cc、ude、w、ns、nr。出现次数超过一定阈值的也被过滤掉。

5 片段的匹配与组合

片段的匹配与组合是构成翻译结果的关键步骤。首先，将匹配的实例进行拆分，拆分成匹配和不能匹配的片段，然后，将实例的翻译结果分割成与上述片段对应的翻译片段，最后，将这些片段组合成翻译结果。

5.1 片段匹配

片段匹配就是匹配实例和待翻译句子，形成一段一段的匹配和不匹配片段。匹配片段中的词有两种情况，一种是词本身完全相同的，称为完全匹配的词，另一种是词虽然不同，但是，词性相同的，称为词性匹配的词。一般来说，词虽然不同，但是词性相同的词可以通过查词典的方法直接得到翻译结果。引入词性匹配的概念，可以更好地利用实例。

经过这样的匹配，相连的匹配词就构成匹配段，相连的不匹配的词就构成了不匹配的段。

5.2 片段分割

片段分割部分是整个翻译过程的核心阶段。如何

万方数据

准确的确定片段的翻译是这部分要解决的主要问题。

本系统采用了基于词语对齐的 EBMT 方法。片段分割的核心就是对齐信息。

图 1 是一个汉蒙机器翻译的例子。

S：　要/v　提防/v　小偷/n

S_1：　要/v　提防/v　那/r　个/q　人/n

(1)　　(2)

图 1　片段划分

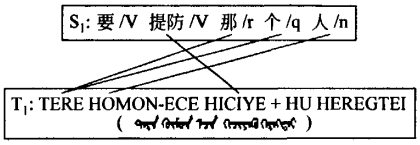


图 2　对齐

S 是待翻译的句子，S_1 是一个实例。根据前面的片段匹配规则，形成了两个片段(1)和(2)，其中(1)是匹配片断，(2)是不匹配的片段。

图 2 是实例 S_1 及其翻译 T_1，连接两者的线就是词语对齐信息。对于不匹配的片段“那个人”，这里存在不同的翻译片段的可能，其中“TERE HOMON”是其必然的翻译，因为它们之间存在连线。可后面的蒙古语附加成分“－ECE”是它的翻译片段的组成部分吗？同样“提防”的必然翻译是“HICIYE”，那么附加成分“＋HU”和情态动词“HEREGTEI”是不是“提防”的翻译呢？可以断定的一点是，“TERE HOMON”不是“提防”这个词的翻译，因为“TERE HOMON”和“那个人”是对齐的。这样我们就可以得到一个最小片段和最大片段。至于哪种分割更合适，交给评价阶段进行处理。

5.3　片段组合

获得片段的翻译后，下一个步骤就是将片段组合成一个翻译结果。由于在上一步片段分割的时候保留了片段的位置信息。所以只需要将片段的翻译置入相对的位置就可以了。

以下是一个翻译片段组合的例子：

S：要/v　提防/v　小偷/n

S1：要/v　提防/v　那/r　个/q　人/n

T1：TERE HOMON －ECE HICIYE ＋HU HEREGTEI

T：HVLAGAYICI － ECE HICIYE ＋ HU HEREGTEI

6　生成结果的评价

从上面的例子可以看出来，生成的翻译结果包含了不应该加入的词。其原因是我们仅仅是把所有的最大翻译片段简单地合在了一起，而不管它是不是真正的翻译的一部分。

解决的方法是，对于未对齐的部分，生成若干保留或者不保留的两个结果，这样就形成了一个翻译的候选集。然后评价哪个结果更适合作为翻译的结果。

下面，我们用语言模型来评价。

我们采用了 SRILM(SRI 口语技术与研究实验室(SRI Speech Technology and Research Laboratory)发布的开源的语言模型工具包)。

在训练中选用了两个参数：-unk 保留〈unk〉和-kndiscount 采用 modified Kneser-Ney 平滑算法。

以下是一个汉蒙机器翻译的例子：

(-90.41) T_1：HVLAGAYICI －ECE HICIYE ＋HU HEREGTEI HEREGTEI

(-85.758) T_2：HVLAGAYICI －ECE HICIYE ＋HU HEREGTEI

在两个候选结果中，T_1 和 T_2 分别利用语言模型进行打分。其中 T_2 的得分要好于 T_1，因此 T_2 更适合于作为翻译结果。

当然，从语法上来说，上面的两个句子都不是最好的结果。事实上，T_2 中有明显的语法错误，即“HVLAGAYICI”是一个阳性词，而其后的附加成分“－ECE”却是阴性的。

7　实验

7.1　蒙古语的表示

实例库是由原始的汉语蒙古语平行语料转换得来的，并存储为易于表示和操作的形式，这就涉及到蒙古语的表示问题。

蒙古语文本的表示是近几年来研究比较多的问题[9]。由于蒙古语是竖写的文字，书写时从上到下，从左到右书写。蒙古语的词并非是用空格分开的，或者说用空格分开的串未必就是一个词。蒙古语的词虽然是由字母顺序组成的，但是蒙古语的字母在词首、词中、词尾会有不同的形式变化，这也给蒙古

万方数据

语的表示造成了比较大的困难。因此，蒙古语的表示问题一直是蒙古语计算机处理研究的重要问题之一。随着蒙古文 Unicode 标准的制定，这一问题在逐渐的得到解决。但是，Unicode 的表示同样存在表示的不便，而且其中存在的转义字符也对机器翻译的处理造成额外的问题。因此，在我们的系统中采用的拉丁转写作为蒙古语文本的表示方式。

拉丁转写是将蒙古语字符利用读音转写成拉丁字母(英文字母)的方法，这种撰写方式具有表示容易的优点。但是其缺点是表示不唯一，从蒙古文到拉丁转写和从拉丁转写到蒙古文的转换都具会产生一定的二义性。这些问题还需要进一步的解决，但是，这些问题对翻译效果的影响是比较有限的。

7.2 语料的规模

进行汉蒙机器翻译的研究还要遇到的一个问题是语料规模的问题。由于蒙古语相关的信息处理发展相对落后，在蒙古语语料的积累和汉蒙平行语料的积累上还做得比较差[12]。目前，我们制作了大约 6 万句对的汉蒙平行语料，基本可以完成一些基础的研究。这些语料还需要进一步扩大。

在这次实验中，我们采用了 12 000 个汉语蒙古语句对的语料库，这些句对是已经完成词语对齐的。

7.3 实验结果

由于汉蒙机器翻译的测试平台相对缺乏，没有汉英机器翻译的国际评测的便利，我们自己设计了一个汉蒙机器翻译的测试平台。

在这里，我们给出了一个具有 100 个日常对话句子的测试集，并由以蒙古语为母语的人翻译成蒙古语。每个句子有 4 个蒙古语参考答案。

```
我 想 参加 一 个 旅游 团 。
    BI NIGE JIGVLCILAL-VN BOLHOM-DU 0R0LCAY_A GEJU B0D0JV BAYIN_A.
    BI NIGE JVGACIL-VN BOLHOM-DU 0R0LCAHV SANAG_A-TAI.
    BI NIGE JIGVLCILAL-VN BOLHOM-DU 0R0LCAY_A GEJU SANAJV BAYIN_A.
    BI NIGEN JVGACIL-VN BOLHOM-DU 0R0LCAY_A GEJU BAYIN_A.
有 这 个 吗 ？
    ENE BAYIN_A VV?
    ENE YAGVM_A-TAI VV?
    ENE YAGVM_A BAYIHV VV?
    ENE YAGVM_A BAYIN_A VV?
...
```

在实验中，我们采用了 12 000 句对的语料作为实例库，实验结果如下：

表 2 翻译系统的实验结果得分

	NIST	BLEU
100 个句子的测试集	3.524 8	0.187 1

评测工具采用的是 NIST 评测工具 mteval-v11b.pl。

NIST 和 BLEU 是目前最常用的机器翻译自动评测的指标，它们都是基于 N-gram 的，它们依赖于翻译结果和参考答案匹配的 N-gram 数目[13]。从得分上可以看出，在 BLEU 得分上我们得到了基本满意的结果，但是 NIST 得分较汉英等机器翻译的结果稍低。以 2005 年的 863 汉英对话机器翻译评测结果为例，我们的系统和汉英系统相比，得分相对比较高，但是 NIST 得分处于比较低的位置，其原因是，由于实例片断的来源不一，而蒙古语的词形变化很大，因此往往由于词形的变化而无法匹配。而 NIST 是匹配 N-gram 的算术平均值，而词匹配的数量相对比较少，所以得分相对比较低。相对来说 BLEU 得分更强的反应了连续词串的匹配程度，从

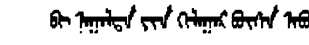
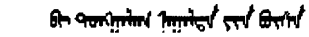
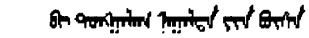
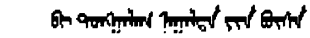
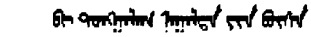
表 3 和汉英机器翻译评测结果的对比

ID	NIST	BLEU
System1	7.139 2	0.250 6
System2	5.921 6	0.181 4
System3	6.209 7	0.174 7
(中略)		
System7	5.522 6	0.145 4
System8	4.227 3	0.071 0

万方数据

这一点可以看出这些连续词串正确的比例比较高，验证了 EBMT 词串片断匹配的优势。当然，汉蒙机器翻译和汉英机器翻译的得分直接比较是没有意义的，这样的比较仅为说明汉蒙机器翻译中存在的问题。

通过对翻译结果的分析我们可以看到，对于具有比较接近的实例的句子，在翻译的时候，良好的体现了蒙古语的特点，都能够保持蒙古语句子的正常语序，仅仅是在一些小的词序及形态上存在问题。

待翻译句子：我 想 要 透明 胶带 。
匹配的实例：
　我/r　想/v　要/v　预约/v　。/w
　BI JAHIYALAG_A ABVY_A GEJU B0D0JV BAYIN_A .
翻译结果：BI NAGALTA-YIN GILAGAR BOSE ABVY_A GEJU B0D0JV BAYIN_A .

参考答案：
　BI TVNGGALAG NAGALTA-YIN BUSE ABVY_A GEJU B0D0JV BAYIN_A.

　BI TVNGGALAG NAGALTA-YIN BUSE ABVY_A GEJU B0D0JV BAYIN_A.

　BI TVNGGALAG NAGALTA-YIN BUSE ABHV SANAG_A-TAI.

　BI TVNGGALAG NAGALTA-YIN BUSE ABVY_A GEJU SANAJV BAYIN_A.


虽然 EBMT 在实例比较接近的取得相对比较好的结果，但是在实例匹配并不很好的时候，仍然会出现比较差的结果。通过对翻译结果的分析，我们发现，在这个 EBMT 系统中存在比较严重的片段边界问题。对于未对齐的词是否应该出现在片段中，并没有一个比较好的方法来处理。如果条件比较严格，就会出现丢词现象，反过来，如果条件比较宽松，就会出现大量的冗余词。比较好的解决方法是通过概率化的方法来处理。

8　结论

通过以上实验，我们给出了一个基本的基于词语对齐的汉蒙 EBMT 系统。这是在蒙古语机器翻译方面的一个新的尝试。通过这次尝试，我们把蒙古语机器翻译的方法从规则方法转移到基于语料库的方法，并为将来统计的汉蒙机器翻译方法的研究打下了基础。

通过实验结果的分析我们可以看到，虽然 EBMT 可以部分解决汉蒙机器翻译的问题，但是其中两个重要的内容还有待研究。一是蒙古语词形的变化问题，蒙古语的词形因词性、时态等原因会产生较大的变化，这些形态变化多是通过增加后缀来进行，经过 EBMT 生成的结果在词形上的错误很多，需要进一步处理产生词形正确的结果。如果要利用蒙古语的词形信息，我们需要对蒙古文的词进行切分，并在这个基础上研究蒙古文的语言模型。这些相关的工作还在起步阶段，还需要进一步的研究。二是评价生成结果的方法过于单一，还需要进一步的研究。

参考文献：

[1] HOU Hongxu etc., An EBMT System Based on Word Alignment[A]. In: proceedings of the IWSLT [C]. 2003. 47-49.

[2] 那顺乌日图，刘群，巴达玛敖德斯尔. 汉蒙机器辅助翻译系统[A]. ALTAI HAKPO(JOURNAL OF THE ALTAI SOCIETY OF KOREA)2001, 11.

[3] Satoshi Shirai, Francis Bond and Yamato Takahashi. A Hybrid Rule and Example-based Method for Machine Translation [A]. In: Natural Language Processing Paci c Rim Symposium 97: NLPRS-97 [C]. 49-54, 11.

[4] Lambros Cranias, Harris Papageorgiou, Stelios Piperidis. A Matching Technique in Example-Based Machine Translation [A]. In: Proceedings of the Fifteenth International Conference on Computational Linguistics [C]. Kyoto: 100-104.

万方数据

[5] Ying Zhang, Ralf Brown, Robert E. Frederking. Adapting an Example-Based Translation System to Chinese [A]. In: Proceedings of HLT 2001: First International Conference on Human Language Technology Research [C]. 7-10.

[6] Sue J. Ker, and Jason S. Chang. Align more words with high precision for small bilingual corpora [J]. Computational Linguistics and Chinese Language Processing, 1997, 2(2):63-96.

[7] 那顺乌日图. 关于现代蒙古语定格问题[J]. 蒙古语文 1988(1).

[8] 那顺乌日图. 计算机处理现代蒙古语 TAI/TEI 形式的尝试[J]. 民族语文,1991,(3).

[9] 那顺乌日图,确精扎布. 关于蒙古文编码[A]. 中国民族语言学会第六次年会[C]. 1994.

[10] 黄河燕,等. 大规模句子相似度计算方法[J]. 中文信息学报,2006,增刊,47-52.

[11] 敖其尔,王斯日古楞. 英蒙机器翻译系统的设计[J]. 内蒙古大学学报(自然科学版),2003(5).

[12] 赵斯琴,高光来,何敏. 蒙古语语料库的研究与建设[J]. 内蒙古大学学报(自然科学版),2003,(5).

[13] K. Papineni, S. Roukos, T. Ward, and W. Zhu, BLEU: a method for automatic evaluation of machine translation [A]. In: Proc. of the 40th ACL [C]. Philadelphia, USA, 2002. 311-318.

[14] 张孝飞,陈肇雄,黄河燕,胡春玲. 多策略机器翻译系统 IHSMTS 中实例模式泛化匹配算法[J]. 中文信息学报,2005,19(4): 1-9.

[15] 刘洋,刘群,林守勋. 机器翻译评测中的模糊匹配[J]. 中文信息学报,2005,19(3): 45-53.

[16] 徐波,史晓东,刘群,宗成庆,等. 2005 统计机器翻译研讨班研究报告[J]. 中文信息学报,2006,20(5): 1-9.

(上接第 64 页)

参考文献:

[1] Mitra M, Chaudhuri B B. Information Retrieval from Documents: A survey. Information Retrieval [J]. 2000, 2(2): 141-163.

[2] Tan C L, Huang W, Yu Z, et al. Imaged Document Text Retrieval without OCR. IEEE Trans. Pattern Analysis and Machine Intelligence [J]. 2002, 24(6): 838-844.

[3] Doermann D, Sauvola J, Kauniskangas H, et al. The development of a general framework for intelligent document image retrieval [A]. In: The 3rd intl workshop on Document Analysis Systems [C]. Malvern, Pennsylvania, USA, 1996. 605-632.

[4] Doermann D, Li H, Kia O. The detection of duplicate in document image database. Image and Vision Computing [J]. 1998, 16(12):907-920.

[5] Niyogi D, Srihari S. The Use of Document Structure Analysis to Retrieve Information from Documents in Digital Libraries [A]. In: Proc SPIE, Document Recoginition IV[C]. 1997. 3027: 207-218.

[6] Wang C L, Chen T S, et al. Chinese document image retrieval system based on proportion of black pixel area in a character image [A]. In: The 6th International Conference on Advanced Communication Technology (IEEE ICACT) [C]. 2004. 25-29.

[7] 卜飞宇,刘长松,丁晓青. 灰度名片图像快速倾斜检测和校正方法[J]. 中文信息学报,2004,18(1):62-69.

[8] 陈艳,孙羽菲,张玉志. 灰度图像中字符切分方法的研究[J]. 中文信息学报,2004,18(4):44-49.

[9] 丁明跃,彭嘉雄. 基于内点保留的二值图像细化算法[J]. 华中理工大学学报,1994,22(1): 79-83.

[10] Huttenlocher D P, Klanderman G A, William J R. Comparing images using the Hausdorff distance. IEEE Transactions on Pattern Analysis and Intelligence[J]. 1993,15(03): 850-863.

[11] Lu Y, Tan C L, Huang W, et al. An Approach to Word Image Matching Based on Weighted Hausdorff Distance [A]. In: Proc. Sixth Int'l Conf. Document Analysis and Recognition[C]. 2001. 921-925.

万方数据