学校编码：10384　　分类号_____密级___

学号：23020071151284　　UDC____

厦门大学

硕 士 学 位 论 文

藏汉统计机器翻译研究

Tibetan-Chinese Statistical Machine Translation Research

刘 智 文

指导教师姓名：史 晓 东　教授

专　业 名 称：计算机应用技术

论文提交日期：2010 年　5　月

论文答辩时间：2010 年　　月

学位授予日期：2010 年　　月

答辩委员会主席：_____

评　　阅　　人：_____

2010 年　5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“ √ ”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

基于统计的机器翻译方法是目前机器翻译领域主流的研究方法。现存的统计机器翻译模型大体可以分为基于词的模型、基于短语的模型和基于句法的模型三大类。其中基于词的翻译模型提出的时间最早，其数学描述比较严密，但是由于在翻译过程中以词作为基本的翻译单位，难以使用语言的上下文信息，故而在词义消歧、语序调整等方面都存在很大的不足。基于短语的翻译模型以短语为基本翻译单位，可以有效利用语言的上下文信息。基于短语的模型也有缺乏全局信息，远距离调序困难等缺点。基于句法的模型从理论上来讲能够利用语言更深层次的信息，具有更大的潜力，但该类模型目前还没有取得人们预期的突破性进展。短语模型是一种十分健壮的模型，模型本身对训练、解码过程中可能出现的很多错误不是很敏感，因而在处理真实语料时往往能有较好的效果。

本文的主要工作是构建一个基于短语的藏汉统计机器翻译系统，利用短语模型的健壮性和可扩展性，通过在系统的训练和解码过程中引入一些语言学信息和统计信息来提高短语模型系统的性能。由于GIZA++的双语聚类工具mkcls在对藏汉双语词聚类过程还存在一些缺陷，本文利用了一些语言学信息对此予以改进。具体的做法是采用史晓东老师的藏文和汉语分词标注工具进行分词标注，根据词汇的词性标记对藏文和汉语词条进行双语词聚类，从而得到更好的藏文和汉语词类模板，提高翻译系统的性能；我们还使用一部藏汉双语词典，利用词典里的词条对齐可靠性高的特点，将其加入到语料的词对齐训练当中，以提高系统的词对齐正确率进而提高整个系统的翻译性能；还有，利用藏汉词典的词性信息，对语料库中出现的低频词采用藏汉双语词聚类替换的策略获得低频词的聚类翻译模板以缓解数据稀疏问题；另外，由于语料库中真实的藏文文本的在分词后词尾大部分带有音节符而少部分没有音节符，这样就造成了藏文语料中词汇表数量的增加，从而引起数据的更加稀疏。由于有无音节符本身不会改变藏文词汇意思，根据藏文的这个特点，我们在对藏文语料进行预处理时将每个词后所带的音节符删除，以降低词汇表的规模、缓解数据稀疏，最终达到提高系统性能的目的。本文的最后还探讨了语言模型对机器翻译效果的影响。

关键词：对齐模板；双语词聚类；词聚类替换；藏文分词；条件随机场

Abstract

Statistical machine translation method is the mainstream research method in the field of machine translation now. Existing statistical machine translation models can be roughly divided into three categories, including the word-based models, the phrases-based models and the syntax-based models. The first proposed model is the word-based translation model, which has strict mathematical description. But due to the word-based models make the words as the basic unit of the translation process, they can not consider the context of a text, so they inadequacy in word sense disambiguation and word reordering. Phrase-based translation models make the phrases as the basic unit in the translation process, they can effectively use the context information. Phrase-based models are also have shortcomings such as lack of global information and difficult in long-distance reordering. In theory, syntax-based models can use a deeper level of linguistics information and have great potential to improve the translation system. But so far, syntax-based models have not made the expected breakthrough. Phrase-based models are very robust models, those models were not very sensitive to the errors occurred in the training and decoding process, so they always have a good performance in the real corpus.

The main work of this thesis is building a phrase-based Tibetan-Chinese statistical machine translation system. In order to improve the performance of the phrase-based translation system, we introduce a number of linguistic and statistical information into the training and decoding process. As the bilingual clustering tool mkcls used by the GIZA++ has some defects in Tibetan and Chinese bilingual word clustering, this thesis employs some linguistic information to improve it. We use a Tibetan and Chinese word segmentation tool developed by teacher Xiao-Dong Shi to make a POS-based (part of speech based) Tibetan and Chinese bilingual word clustering tool to get a better Tibetan and Chinese bilingual word clustering templates which can improve the performance of the translation system; Meanwhile, as the entries in Tibetan-Chinese dictionary have high reliability, we also add those entries into the

word alignment training process to improve word alignment accuracy of the system and raise the system's Translation performance; We also clustering the low-frequency words appear in the bilingual corpus into their POS so that we can get low-frequency words clustered templates to alleviate the data sparseness problem. In addition, because most of the words in the real Tibetan corpus have a syllable break ending and some words have no syllable break ending, thus cause the number of Tibetan vocabulary of the corpus increased and more sparse data. Since the syllable break ending itself can not change the meaning of a Tibetan word, we remove all the syllable break ending when we preprocess the Tibetan corpus. So that we can reduce the size of the Tibetan vocabulary, ease the data sparseness problem and achieve the purpose to improve the system performance. At the end of this thesis we also discuss how the language models influence the performance of the machine translation system.

Key Words: alignment template; bilingual word clustering; word clustering replacement; Tibetan word segmentation；conditional random field

目　录

Contents

第一章 绪论

机器翻译（Machine Translation，简称 MT）是利用计算机把源语言翻译成目标语言。从 20 世纪 40 年代计算机刚刚被发明出来不久，就有人想到用计算机进行不同语言间的机器翻译。经过几十年的发展，机器翻译在理论和实践方面都取得了很大的进展，世界各地的学者和研究人员提出了不少机器翻译的理论和数学模型，也根据这些模型和理论开发出了相应的机器翻译系统。

机器翻译领域所使用的方法一般可以分为两类：基于规则的（Rule-Based）方法和基于语料库（Corpus-Based）的方法。上个世纪 90 年代之前，居于主导地位的方法是基于规则的方法，其主要思想是由人工构造出可供机器翻译使用的句法语义规则库，然后利用规则信息在两种语言间匹配来实现机器翻译。这种方法也称为理性主义（Rationalist）方法，这种方法一般不需要像基于统计的方法那样在大规模的空间进行搜索，构造出来的系统运行效率很高。缺点是规则库的构造和维护过程需要耗费大量的语言学专家的劳动，同时也离不开具有较强计算机功底的编程人员的劳动，而既懂得大量语言学知识又懂的计算机编程知识的专家少之又少，而且专家们所构造的规则库既难以完备地包容语言中存在的大量不符合语法规则的现象，也不能适应语言日新月异的变化。

基于语料库的方法也称为经验主义（Empiricist）方法，主要是通过对大规模的双语平行语料库进行机器学习，根据不同的数学模型构造出翻译模型和语言模型来实现机器翻译的方法。基于语料库的方法的相对于基于规则的方法优势是构建模型时不需要太多的人力物力参与，研究人员本身对两种语言的理解也不一定要达到语言学家的水平，相对而言研究的门槛没有那么高。这样就能允许更多有兴趣的学者和研究人员投入其中深入研究。基于语料库的方法又可分为基于统计的（Statistical-Based）方法和基于实例的（Example-Based）方法，目前主流的是基于统计的方法。

1.1 统计机器翻译的发展

统计机器翻译（Statistical Machine Translation，简称 SMT）的思想可

以追溯到20 世纪50 年代。由于当时语料规模不足、计算机的计算资源十分有限等原因，该研究并没有获得成功。进入20 世纪80 年代中后期，基于统计的方法在机器翻译中的应用被重新提起。到了90 年代初，IBM 的Brown 等人提出了基于信源信道（Source Channel）思想的统计机器翻译模型[1, 2]，并且在实验中获得了初步的成功，引起了相关领域的学者和研究人员的广泛兴趣。但是，由于其他人无法获取IBM 的源代码，要想进行统计机器翻译的研究，首先需要重复Brown 等人的实验，然后才能对它进行改进，这就需要进行大量的编码工作。于是，在1999 年夏天，很多相关的研究人员汇聚到约翰霍普金斯大学（JHU）的夏季研讨班上，大家通过共同合作重复了Brown 等人的统计机器翻译实验，并开发了一个公开源代码的统计机器翻译工具包——EGYPT[3]。这个夏季研讨班结束以后，这些研究人员回到各自的研究机构，继续开展相关的研究工作，并提出了各种改进的模型和方法，使得统计机器翻译的研究出现了新的高潮。

EGYPT 是个免费的工具包，其源代码可以自由下载,这为相关的研究工作提供了一个很好的基础和平台。EGYPT 中参数训练模块GIZA 的主要开发者是Franz J. Och 博士，他后来还在此基础上改进并发布了增强版的GIZA++[4]。GIZA++现在已经成为统计机器翻译领域最常用的工具之一，其主要用途已经不仅限于用来估计IBM 参数，而更多的是为了得到双语平行语料库的词语对齐。

1999 约翰霍普金斯大学（JHU）的夏季研讨班结束后，随着越来越多的研究人员投入到统计机器翻译的研究，各种新的模型不断涌现，统计机器翻译从理论到技术都取得了长足的进展。其中Och 博士在基于短语的机器翻译模型理论和技术方面都做出了巨大的贡献：他提出了对数线性模型[5]使得机器翻译的模型框架中可以加入更多的特征，在容纳了经典的信源信道模型的前提下，使系统的可扩展性大大提高;在翻译模型方面,他提出了对齐模板模型[6,7]改进IBM 的基于词的翻译模型，减少了数据稀疏现象；他还引入了基于最小错误率（Minimum Error Rate，简称MER）的区别性训练方法[8]应用于参数训练过程。同一时期，还有其他的研究人员也不断地提出各种统计机器翻译模型，如基于短语的模型，基于层次短语的模型，基于浅层句法结构的模型和基于句法结构的模型等等。这些方面的进展使统计机器翻译系统的翻译效果有了很大的提高，使基于统计的机器翻译系统显现出了极大的竞争力。在近几年NIST 等国际评测

中，基于统计模型的翻译系统显示出了极大的优势，统计机器翻译逐渐成为机器翻译研究领域的主流方向[9]。

1.2 本文研究背景和目的

随着世界化的进程的加速，国际间的交流越来越频繁，人们对可以实用的机器翻译系统或机器辅助翻译系统的需求越来越大。但是，由于机器翻译所处理的自然语言是人类千百年生产生活中积累和传承下来的结果，是十分复杂的。所以，虽然已经历了几十年的发展，但现有的机器翻译系统在翻译效果方面还无法令人满意。不但如此，由于缺乏平行语料库等原因，小语种机器翻译的研究水平还远远不如英汉、英法等目前国际上主流的语言之间的机器翻译。这个问题已经引起了国际社会的关注。随着机器翻译技术的发展，将现有的较成熟的机器翻译方法应用到小语种机器翻译中已经不是很难的事情，而且相对几十年前毫无基础的机器翻译研究而言人力物力的投入也不会很大，于是近年小语种机器翻译的研究也逐渐开展开来。最为明显的变化是，作为国际上权威的机器翻译评测会议，NIST 机器翻译评测会议在 2008 年和 2009 年的评测上就在原有语种的基础上加入了训练语料库规模较小的乌尔都语到英语的机器翻译。同样为了推动世界上繁多的小语种的信息处理的发展，第三届国际联合自然语言处理会议（IJCNLP-2008）专门建立了一个研讨会（workshop）讨论较少被重视的语种（less privileged languages）信息处理的研究。

我国是个多民族的国家，随着经济社会的发展汉族和少数民族之间的交流日益广泛和深入，这就催生了越来越多的少数民族语言和汉语之间的机器翻译需求。我国在少数民族语言文字的规范化、标准化和信息处理工作的研究已经有一段历史。截止到 2009 年，中国少数民族语言文字信息处理学术研讨会已经举办了 12 届；我国第五届全国机器翻译研讨会（CWMT2009）中也将汉蒙日常用语机器翻译列入评测项目，意在探讨汉语和少数民族语言间的机器翻译。但是，由于我国民族众多，各个民族的语言文字发展不平衡，总体而言目前我国少数民族语言信息处理研究工作还处在比较初级的阶段，特别是少数民族语言和汉语之间的机器翻译研究工作开展的还比较少，研究的水平还不够高。就藏汉机器翻译工作而言，相应的基础研究和应用都比较少，如藏文分词标注语

料库、藏汉平行语料库等资源还十分稀缺，这给构建统计机器翻译系统造成了很大的困难。在藏语和汉语之间的机器翻译研究方面，最早的报道见于青海师范大学陈玉忠（德盖才郎）、李延福等人在 1995 年实现一个基于规则的汉藏机器系统[10]。后来该课题组在原有的系统上进行改进，实现了公文和科技两大汉藏翻译系统[11]。基于统计的藏汉机器翻译系统目前还没有见到公开的报道。

基于短语模型的机器翻译系统是当前主流的统计机器翻译系统，短语模型是一种十分健壮的模型，模型本身对训练、翻译过程中可能遇到的很多错误都不是很敏感，因而在处理真实语料时往往能有较好的效果。基于短语的机器翻译系统在目前主流的大语种机器翻译如英汉、汉英机器翻译中已经有很多的研究和应用。但是各个研究机构对面向小语种的机器翻译系统还涉及的比较少，即使有少量研究，由于相应的各种语料的缺乏也使得翻译效果也明显不如大语种之间的机器翻译系统。基于这样的现状，本文的主要思想是在藏汉机器翻译方面引入基于短语的统计机器翻译模型，利用现有的比较先进的机器翻译系统，着重研究在平行语料库规模很小的前提下如何提高机器翻译系统的性能。

1.3 本文的主要工作与结构

由于目前藏文信息处理方面的研究还比较薄弱，基于统计的藏汉机器翻译方面的研究鲜有报道，故而本文从藏文分词、藏文语料预处理等方面入手，建立藏汉统计机器翻译系统并探讨使用各种方法加以改进。本文的总体思想是将目前汉英和英汉统计机器翻译中比较成熟的基于短语的模型应用到藏汉统计机器翻译中，在藏汉双语平行语料库规模很小的前提下，结合藏文本身的特点，利用语言学知识，对翻译系统加以改进。

在藏文分词方面，本文通过分析藏文的特点，藏文字的结构和藏文字的识别入手，通过藏文分字、词位标注和由字构词的思想，描述了一种利用条件随机场模型进行藏文分词的方法。在藏汉机器翻译方面，本文采用基于短语的统计机器翻译框架，利用了藏文分词标注工具和藏汉双语词性标注词典等语言学信息对藏汉双语词聚类进行改进，并针对藏汉语料数据稀疏的问题提出用藏汉词典的词性信息对语料库中出现的低频词进行词聚类替换的方法，通过聚类替换翻译模板来缓解数据稀疏问题，改善低频词及其周边词汇的对齐效果进而改