基于翻译记忆库与基于规则的汉维-维汉机器辅助翻译系统方法与框架研究[?]

吐尔根.依布拉音，艾尔肯.伊米尔，阿布力米提.阿不都热依木

新疆大学信息科学与工程学院，乌鲁木齐.830046

Email: turgun@xju.edu.cn

摘要： 本文描述一种基于翻译记忆库和基于规则相结合的方法，对维汉-汉维辅助翻译系统翻译记忆库的框架的构建，维汉-汉维实例的对齐、组合、检索、译词选择、库的扩充，与基于规则翻译引擎的接口等问题作了探讨。提出了基于翻译记忆库与基于规则的汉维-维汉机器辅助翻译系统初步解决方案。

关键词： 维汉-汉维翻译、基于翻译记忆库、辅助翻译系统

Translation-Memory-Lib Based and Rule Based Chinese Uighur, Uighur Chinese Translation System Aid Machine Research

Turgun.Ibrahim, Erkin Imir, Ablimit.abdureyim

Information Science & Engineering College of Xinjiang University

Urumqi 830046

turgun@xju.edu.cn

ABSTRACT: This article describes a method, which is the combination of translation-memory-lib based method and rule based method, discusses the framework of translation-memory-lib of Uighur Chinese-Chinese Uighur translation aid system, align, combination, search of Uighur Chinese –Chinese Uighur instances, selection of interpret word, extension of library, and interface of rule based translation engine etc. Finally put forward the initial solution of the system.

Keywords: Chinese Uighur-Uighur Chinese translation, translation-memory-lib based, translation aid system

一、引言

维吾尔语属阿勒泰语系突厥语族，属该语族的国内还有维吾尔、哈萨克、柯尔克孜、乌

[?] 本文得到国家科技部 2001 年度基础研究快速反应项目资助（项目批准号：国科基字[2001]51）

孜别克、塔塔尔族等。随着信息社会各类知识信息急剧聚增，而且这些知识和信息主要以英语和汉语为载体，如果我们不尽快解决网上的电子翻译问题，我们将在国民收入步入小康的同时在信息和知识上走向贫困，在知识经济的大潮中沦为第四世界、第五世界。

基于翻译记忆库与基于规则的汉维-维汉机器辅助翻译系统的研究就是在上述背景下提出的，该系统研究将解决维吾尔等少数民族同志使用信息技术时的语言障碍问题，使他们能快速翻译获取信息。对提高少数民族的科技文化水平尽快脱贫有着深远的意义；另一方面对将来对哈（哈萨克文）汉、乌（乌孜别克文）汉、柯（柯尔克孜）汉、土（土耳其文）汉双向辅助机器翻译系统的开发打下坚实的基础。对我们快速翻译获取中亚国家的科技、文化、经济的信息，加强与周边国家的友好往来，使我们国家长治久安有着深远的意义。

二、目前机器翻译的方法

2．1 基于规则的机器翻译方法

基于规则的机器翻译方法的界定及其优缺点大家都非常熟悉，这里不再详述。

自从 Chomsky 提出转换生成语法以来，基于规则的方法成了机器翻译研究的主流。在已有的商品化机器翻译系统中，很少有哪个系统声称自己采用的是纯统计（或语料库）的方法或以统计为主的方法。虽然如此，统计方法的影响也是不可低估的。与传统的规则方法相比，现在的规则方法已产生了很多变化。这些变化主要体现在：

1. 在规则的获取方面，传统的规则方法主要依靠语言学家总结规则，进行调试，而现在则更加重视从语料库中获取规则（如采用错误驱动的学习算法）；

2. 传统的规则方法往往偏重于描述粗粒度、全局化的大范围语言学规则知识，而现在则更加重视描述细粒度、局部的小范围的语言学知识，呈现出“小规则库、大词典”的趋势；

3. 在知识表示方面，为了以更小的粒度、更加准确地对翻译知识进行描述，一般对要对单纯的上下文无关规则做一些改进。改进的方法有以下两种：一种是采用特征结构与合一算法，如 LFG、GPSG 等等，这种方法一般要求具有较好的语言学背景；另外一种是采用词汇化的方法对规则加以细化。后一种方法的做法之一也就是下面我们将要介绍的基于模板的机器翻译方法；

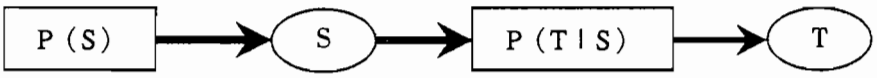
4. 传统的规则方法采用的往往是非此即彼的确定性原则，系统的鲁棒性较差，而现在规则系统中一般都引入各种形式的概率或评分函数，系统的鲁棒性有所提高。这两种方法的区别在于：概率方法一般有比较严格的数学模型做基础，概率值的计算要以对大规模语料库的统计为依据；评分方法的主观性较强，评分规则的确定以及具体规则的分值都是人为的（也不排除统计），人可以根据经验进行调整。不过这两种方法都对提高系统的鲁棒性有较为直接的效果。

2．2 基于统计的统计机器翻译

基于统计的机器翻译方法和基于实例的机器翻译方法都是使用语料库作为翻译知识的来源。这二者的区别在于，在基于统计的机器翻译方法中，知识的表示是统计数据，而不是语料库本身；翻译知识的获取是在翻译之前完成，翻译的过程中不再使用语料库。而在基于实例的机器翻译方法中，双语语料库本身就是翻译知识的一种表现形式（不一定是唯一的），翻译知识的获取在翻译之前没有全部完成，在翻译的过程中还要查询并利用语料库。

基于统计的机器翻译方法源于 Weaver 在 1947 年提出的把翻译看成是一种解码的过程。统计机器翻译的数学模型是由 IBM 公司的 Brown 等人提出的。

统计机器翻译的基本思想是，把机器翻译看成是一个信息传输的过程，用一种信道模型对机器翻译进行解释。假设一段源语言文本 S，经过某一噪声信道后变成目标语言 T，也就是说，假设目标语言文本 T 是由一段源语言文本 S 经过某种奇怪的编码得到的，那么翻译的目标就是要将 T 还原成 S，这也就是就是一个解码的过程。（在这种思想下，我们已知的是目标语言 T，未知的是源语言 S。这与我们一般的说法不同，要注意不要混淆。）



根据 Bayes 公式可推导得到：

$$S ? \max_{S} P(S)P(T \mid S)$$

这个公式被称为统计机器翻译的基本方程式（Fundamental Equation of Statistical Machine Translation）。在这个公式中，$P(S)$是源语言的文本 S 出现的概率，称为语言模型。$P(T|S)$是由源语言文本 S 翻译成目标语言文本 T 的概率，称为翻译模型。语言模型只与源语言相关，与目标语言无关，反映的是一个句子在源语言中出现的可能性，实际上就是该句子在句法语义等方面的合理程度；翻译模型与源语言和目标语言都有关系，反映的是两个句子互为翻译的可能性。

也许有人会问，为什么不直接使用 $P(S|T)$，而要使用 $P(S)P(T|S)$这样一个更加复杂的公式来估计译文的概率呢？其原因在于，如果直接使用 $P(S|T)$来选择合适的 S，那么得到的 S 很可能是不符合译文语法的（ill-formed），而语言模型 $P(S)$就可以保证得到的译文尽可能的符合语法。

统计机器翻译问题被分解为三个问题：

1. 语言模型 Pr(s)的参数估计；
2. 翻译模型 Pr(t|s)的参数估计；
3. 搜索问题：寻找最优的译文；

对于语言模型 Pr(s)，可以采用 n 语法、链语法等语法模型。

对于翻译模型 Pr(t|s)，IBM 公司提出了 5 种不同形式，复杂程度递增的数学模型，这些模型都用到了很复杂数学推导。在模型 1 和 2 中，首先预测源语言句子长度，假设所有长度都具有相同的可能性。然后，对于源语言句子中的每个位置，猜测其与目标语言单词的对应关系，以及该位置上的源语言单词。在模型 3,4,5 中，首先，对于每个目标语言单词，我们选择对应的源语言单词个数，然后再确定这些单词，最后，判断这些源语言单词的具体位置。

注意，在翻译模型中，我们已知的是目标语言句子，要求解源语言句子的概率，这与我们通常所说的翻译顺序刚好相反，因此在理解时注意不要混淆。

这些模型的主要区别在于计算源语言单词和目标语言单词之间的连接（Connection）的概率的方式不同。模型 1 最简单，只考虑词与词之间互译的概率，不考虑词的位置信息，也就是说，与词序无关。好在模型 1 的参数估计具有全局最优的特点，也就是说最后总可以收敛于一个与初始值无关的点。模型 2 到 5 都只能收敛到局部最优，但在 IBM 的实验中，每一种模型的参数估计都依次以上一种模型得到的结果作为初始值，于是我们可以看到最后的结果实际上也是与初始值无关的。

2．3 基于实例的机器翻译方法

基于实例的机器翻译思想最早是由著名的日本机器翻译专家长尾真(Nagao, M.)提出的，其基本设想是不通过深层的分析，仅仅通过已有的经验知识，通过类比原理进行翻译。

基于实例的机器翻译方法具有以下一些优点：

- 系统维护容易，系统中知识以翻译实例和语义词典等形式存在，可以很容易的利用增加实例和词汇的方式扩充系统。
- 容易产生高质量的译文，尤其是利用了较大的翻译实例库，或者输入能和实例精确匹配时更是如此。
- 可以避免一些传统的基于规则机器翻译必须进行的深层次语言学分析。
- 同语种相关的知识很少。只要记忆库中存在外形同输入相似的句子，就可以进行匹配。

由于大规模获取语言知识的代价非常大，对于词法、语法和语义的规则的收集概括难以全面，机器翻译系统的性能一直徘徊不前。利用已经存在的双语语料库资源为新的翻译需求提供经验，是目前提高机器翻译系统译文质量的重要途径之一。EBMT 对于相同或相似的文本的翻译有非常显著的效果，随着例句库的规模的增加，其作用也越来越显著。对于实例库中已有的文本，可以直接获得高质量的翻译结果。对与实例库中存在的实例十分相似的文本，可以通过类比推理，并对翻译结果进行少量的修改，构造近似的翻译结果。

要实现一个基于实例的翻译引擎，面临的主要问题是：

- 实例的对齐：实例库表现为对齐的文本。要进行翻译，实例必须至少做到句子一级对齐。从理论上说，对齐的单位越小（如做到子句、短语、句子级对齐），语料库的可重复利用价值就越高，匹配的准确率也越高；不过对齐的单位越小，意味着加工的深度越深，加工的成本越高（尤其是人工的成本），系统的可扩充性也越差；
- 实例的查询：实例库规模很大，因此高效的查询算法也是一个系统实现中要考虑的重要问题；
- 实例的组合：这是基于实例的翻译中的核心问题。其目的是用已有的实例片断组合成被翻译文本的一个覆盖。
- 译词的选择：与所有其他机器翻译系统一样，基于实例的机器翻译也存在译词选择问题，不过由于不作完整的句法分析，与基于规则的方法相比，这里译词选择问题的解决策略可能略有不同。
- 语料库规模：要达到较高的准确率，实例库的规模肯定不能太小。一般认为，例句库的

规模一般应达到几百万句对的数量级。

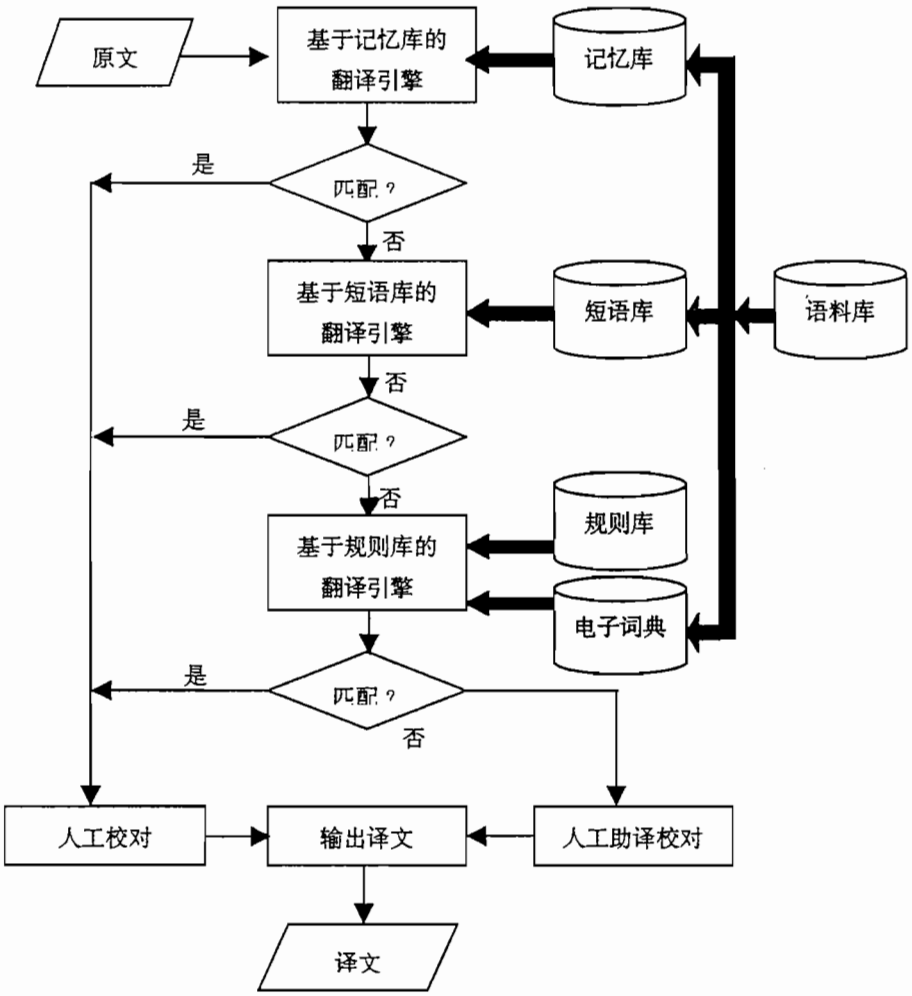
基于实例的翻译具有众多的优点，在具体实现上又是千差万别，很多地方还有相当大的潜力，因此近年来一直是机器翻译的研究的热点之一。但由于语料库规模的限制，基于实例的机器翻译很难达到很高的匹配率，因而，到目前为止还很少有机器翻译系统采用纯粹的基于实例的方法，一般都是把基于实例的机器翻译作为多翻译引擎中的一个，以提高翻译的正确率。

三、本系统采用的基本方法与框架

本方法以计算语言学、语料库语言学、认知科学、翻译学、统计学、信息学、计算机信息处理科学为基础，采用基于翻译记忆库（数据库）和基于规则的翻译系统二者相结合的方法，以新闻领域言语活动为背景，以汉维-维汉双语对齐语料库为依据，提出如何实现完整的人机辅助翻译系统的解决方案。该系统将是基于翻译记忆库（数据库）和基于转换规则的翻译系统二者的结合产物。它可作为翻译人员的辅助工具，其作用是提高翻译速度、规范译文的格式和译法以及降低译文成本等。其原理是对输入句子在翻译记忆库中直接调出在语法和语义结构上完全匹配的句子它的译文；或调出翻译记忆库中与输入句子在语法和语义结构上最相似的句子和它的译文，供翻译人员对译文做简单的修改就变成输入句子的译文；对完全查找不到的句子在通过基于转换规则的翻译系统进行初步翻译，在由翻译人员对译文校验修改变成输入句子的译文；再将输入句子和译文存入翻译记忆库作为后期翻译的备选译文存储的辅助系统。具体方法如下：

- 采用计算语言学、计算词汇学与现代维吾尔语词汇学的方法，以语言学分类为理论基础，继承语言学公认的词性分类方法，又对基本词类进行细分类，研究现代维吾尔语的词法规则，制定《信息处理用现代维吾尔语词性标记集规范》。采用面向对象的程序设计技术，编制出维吾尔语自动词性标注系统，使之正确标注率达到90%，平均标注速度达到150词/秒。
- 在《信息处理用现代维吾尔语词性标记集规范》和维吾尔语自动词性标注系统基础上，从《新疆日报》、《求实》、《参考消息》等新闻媒体中收集筛选出维汉双语对照的真实语料，构建大规模面向新闻领域汉-维双语对齐词性标注语料库。
- 采用现代数据库技术的研究方法，对翻译记忆库的结构进行全面的分析，制定翻译记忆库的存储结构、完整性规则、对等译文的对齐规则及其匹配搜索算法，从上述语料库中抽取对等译文建立基于动态数据结构的句子翻译记忆库和短语实例翻译记忆库。
- 采用计算语言学和计算词典学的研究方法和现代数据库技术，以及项目所涉及的某些理论知识和技术方法，在上述双语平行语料库抽取双语搭配词语对，构建一部每个词条包括句法、语义、搭配、固定词组、构词等方面信息的双语搭配词典。对词典的结构进行全面的分析，建立基于动态数据结构的双语搭配词典，制定词典的存储结构、完整性规则及其存取算法。

◆ 采用计算语言学、形式语法相结合的方法，对现代维吾尔语词法和句法结构进行全面的分析，建立形式化的结构；在基于统计的方法的基础上，揭示出传统语言学定性研究未发现的规律，在这些规律的基础上制定出表示各类语言现象的规则。制定《信息处理用现代维吾尔语词法、句法规则规范》，构建多个模式规则集（维语生成规则集、双语转换规则集）；以这些规则为基础编制维吾尔语句法分析器和句子生成器。



多引擎辅助翻译系统基本框架图

◆ 采用先进的数据库搜索技术，深入分析基于句子实例翻译记忆库和基于短语实例翻译记忆库中双语实例的对齐结构。设计和优化双语实例的精确匹配算法、模糊匹配算法和转换算法。构造基于翻译记忆库汉-维双向翻译引擎。

◆ 在本系统中我们将引进中科院计算技术研究所和北大计算语言学研究所联合开发的“通用机器翻译开发平台”，将在此平台上研制的“汉-英机器翻译系统”应用于基于规则的汉-维翻译引擎，开发时对其描述语言和实现算法、转换生成作适当的修改以至适应汉-

维翻译引擎。着重研究维吾尔语生成技术，即对维吾尔语词的派生和各种语法形态的生成问题进行系统地分析，研制符合维吾尔语语法规则的句子生成器，构成基于规则的汉-维翻译引擎。在此平台上研制的“英-汉机器翻译系统”应用于基于规则的维-汉翻译引擎，开发时对其描述语言和实现算法、转换生成作适当的修改以至适应维-汉翻译引擎。着重研究维吾尔语分析技术，即对维吾尔语词法和各种语法形态的系统地研究，研制符合维吾尔语语法规则的句法分析器，构成基于规则的汉-维翻译引擎。

四、结束语

通过对目前较为流行的机器翻译方法的比较研究，本文提出了一种采用基于翻译记忆库和基于规则相结合的方法实现汉维-维汉机器辅助翻译系统初步方案。并以此方案为主线开始进行大规模面向新闻领域汉-维双语对齐词性标注语料库的翻译资源建设工作。当然为了实现汉维-维汉机器辅助翻译系统，要考虑和解决的问题还很多，如怎样从双语对齐语料库抽取对等译文、维吾尔语句法分析器转换器和句子生成器研制时分析算法、转换算法和生成算法的描述、译文的对齐规则及其匹配搜索算法的描述、几种翻译引擎的冲突等一系列问题有待于进一步探讨。这些方法虽然比较零碎，不成体系，但也是我们经过一段时间探索的结果，希望能引起同行们的兴趣。

参考文献

[1]刘群，俞士汶.机器翻译技术综述及面向新闻领域的汉英机器翻译系统。综合考试报告

[2]柏晓静，常宝宝，詹卫东，吴拥华.构建大规模的汉英双语平行语料库.机器翻译研究进展，2002，10

[3]常宝宝，詹卫东，柏晓静，吴拥华.服务于汉英机器翻译的双语对齐语料库和短语库的建设.CJNLP2002 2002，10

[4] 程适良.现代维吾尔语语法，新疆人民出版社.1996

[5] 周殿生.适应于机器翻译需要的汉维语句型结构对比.新疆大学学报 1996

[6]董振东.中国机器翻译的世纪回顾.中国计算机世界.2000，1

[7]刘倬.机器翻译的发展和突破.机器翻译研究进展，2002，10

[8]刘群，张彤.汉英机器翻译扩充词典的建造。机器翻译研究进展.2002，10

[9]冯志伟.计算语言学基础。商务印书馆，2001

[10]黄昌宁，李涓子。语料库语言学.商务印书馆，2002

[11]柏晓静，常宝宝，詹卫东，吴拥华.构建大规模的汉英双语平行语料库.机器翻译研究进展，2002，10

[12]吐尔根.依布拉音，艾尔肯.伊米尔.英—维 Internet 在线翻译系统开发的初探.机器翻译研究进展，2002，10

[13] 刘群，俞士汶.汉英机器翻译的难点分析，中文信息处理国际会议论文集.清华大学出版社，1998