基于语料库的机器翻译研究环境的设计

王挺　陈火旺　史晓东
（国防科技大学计算机系，410073）
李志伟
（空军第一航空学院115#，464000）

摘要：目前基于语料库的机器翻译的研究越来越受到人们的关注。本文首先介绍了一个基于语料库的机器翻译研究模型（CBMT），为了支持这一模型，我们设计了一个基于语料库的机器翻译的研究环境，为机器翻译系统的开发提供一个平台。本文详细介绍了这一环境的设计思想。

关键词：语料库，机器翻译

The Design of Corpus-Based Machine Translation Research Environment

Wang Ting　Cheng Huowang　Shi Xiaodong
（Department of Computer, National University of Defense Technology, 410073）
Li Zhiwei
（P.O. 115, The 1st Aeronautic College of Air Force, 464000）

Abstract: Today, corpus-based machine translation causes people's more and more attention. In this paper, at first, we introduce the Corpus-Based Machine Translation model (CBMT). Then, we describe in detail the design of the Corpus-Based Machine Translation research Environment (CBMTE) which is built to support the CBMT model and provide a platform for the development of machine translation systems.

Keywords: corpus, machine translation

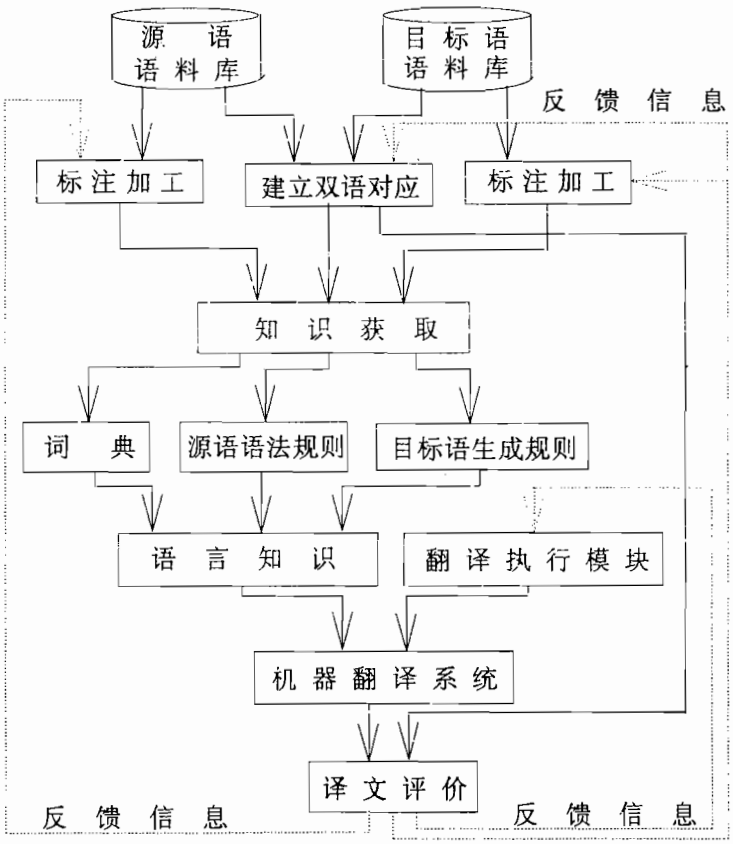
1　前　　言

随着机器翻译研究的深入，人们日益认识到高质量的机器翻译必须基于对大量的真实语料的研究。作为研究对象的语料规模越大、越具有代表性，所得到的语言知识就越准确。然而，当我们面对大规模的语料时，如果没有合适的支撑环境和工具，就很难对其进行深入而系统的研究，难以从语料库中获取准确的知识，所得到的机器翻译系统也就不尽人意了，这与语料库研究的初衷相悖的。为此，我们首先将机器翻译系统模块化为两个部分：一是语言知识模块，包括词典知识、句法知识、目标生成规则等；二是翻译执行模块，该部分使用语言知识进行翻译。并在此基础上提出了基于语料库的机器翻译研究模型

（CBMT）。为了支持该模型，我们设计了基于语料库的机器翻译研究环境（CBMTE），以实现支持CBMT模型的各种工具，从而为机器翻译的研究提供一个多功能的平台。

2　CBMT模型简介

基于规则的方法与语料库技术相结合是当前机器翻译研究的一个热点。传统的基于规则的方法脱离了对真实语言现象的研究，难以处理复杂的自然语言，灵活性较差。为了覆盖各种纷繁复杂的语言现象，机器翻译的研究人员不得不经常添加新的规则，因此文法的维护和一致性保证越来越困难。另外，基于规则的系统忽视了语言中特殊的、经验性的和小粒度的知识，消歧能力较弱。因此，在机器翻译系统中，必须以语料库为基础，将规则方法与统计方法相结合，从语料库中获得规则和统计方法所需要的先验概率。为此我们提出了基于语料库的机器翻译研究模型（CBMT），如下图所示：



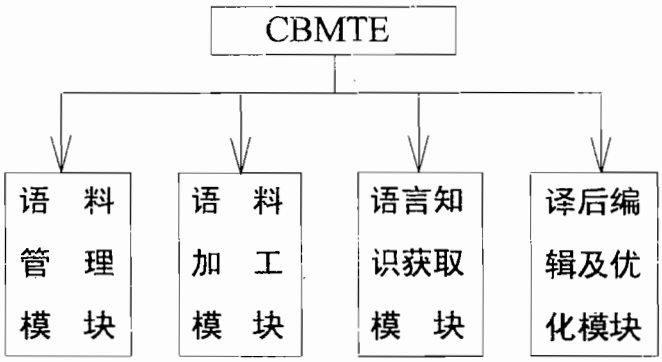
CBMT 模型

在CBMT模型中，我们将机器翻译系统视为两个部分：第一部分是语言知识模块，包括词典、源语言语法规则、目标语生成规则等语言学知识，这些知识包含了相关的统计信

息，它是机器翻译系统的“智能”部分；第二部分是翻译的执行模块，包括词法分析、句法分析、结构调整、目标语生成等过程（也可能包含中间语言处理过程），这些过程使用语言知识进行分析和推理，并与语言知识模块保持相互独立，这一部分可视为机器翻译系统的“机械”部分。基于这种思想，CBMT模型从加工后的语料库中获取语言知识，并与执行模块相结合形成机器翻译系统。由译文评价模块根据双语语料库的对应关系或译后编辑的历史信息作出评价并提出优化意见，并将有关的信息反馈到语料加工模块，沿着知识的流向对各个模块进行优化。在这个模型中，语料的加工（包括各种标注、建立对应关系）是基础，它从根本上决定了所形成的机器翻译系统的质量；知识获取模块是关键，它决定了我们能在多大程度上从语料库中得到帮助，获得机器翻译系统所需要的知识。

3　CBMTE的设计

为了支持CBMT模型，我们设计了一个支持英汉机器翻译研究的环境CBMTE，该环境包括了语料的管理、语料加工、语言知识的获取、译后编辑及反馈等功能，如下图：



各个模块的具体设计如下文所述。

3·1　语料管理模块

关于语料管理，参考文献［1］给出了较全面的论述。针对机器翻译的研究需要，我们的语料管理模块应当具有语料的增加、删除、编辑以及语料的合并与分割等功能。另外，语料管理模块还有一个检索与统计子系统，该子系统对加工后的语料进行词法、句法和语义单位的出现频率的统计、各种搭配的同现概率的计算，并提供查询界面。

3·2　语料加工模块

语料的加工主要有词法标注、句法标注、语义特征标注和双语对应(Aligning)等，对于汉语的语料还存在分词的问题。本文主要介绍前四种加工方法的设计。

3·2·1　词类标注

词类标注就是对语料的单词标以词类。在设计词类标注系统时，首先应当选取一个合适的标注集。如果分类太粗，不利于分析工作，难以准确发现语言中的规律；分类太细，会产生矛盾交叉现象，标注结果的准确率下降，为后续工作带来困难。我们选择词类的标注集的原则是：1. 遵循“分类宜粗不宜细”的原则[3]，分类不宜过细；2. 词类标注集的设计既要体现词的语法功能，又要体现词的形态特征。基于这种思想，我们选择了一套词类标注集（约40多个词类）。

关于词类的标注方法，我们采用一种渐进式的、二元标注方法，具体方法如下：

(1) 设初始二元标注模型为M_0，且M_0为等概率模型；

(2) i=0；

(3) 用M_i标注一部分语料S_i；

(4) 对S_i中不正确的标注进行修改；若S_i和S_{i-1}的标注结果的正确率之差小于阈值δ，说明M_i和M_{i-1}的标注结果提高不明显，转(9)；

(5) 用修改后的S_0，S_1，…，S_i的标注结果训练标注模型的参数，得到标注模型N_{i+1}；

(6) 用隐马尔科夫模型（HMM）的FB迭代法训练N_{i+1}，得到一个局部最优模型M_{i+1}；

(7) i=i+1；

(8) 转(3)；

(9) 用M_i标注所有剩下的语料，并对不正确的标注进行修改；

我们还提供了一个对标注结果进行标后修改的工具，能够方便修改不正确的标注。采用这种标注方法，使我们能够当面对一个从未标注过的领域的语料时，用最少的人力来训练一个标注模型，并标注所有的语料。

3·2·2　句法标注

句法标注就是对语料进行句法分析，标以句法单位或生成语法分析树。我们采用CFG作为语料的句法标注的文法，主要基于以下的考虑：1. CFG 适合表达句法结构，能够用语法分析树来直观地描述句子的结构；2. CFG 的自动分析方法比较成熟，在分析大量语料时，能减少人的工作量，而且标注的结果的一致性易于保证；3. CFG 是许多其他文法的基础，用 CFG 来标注语料可以为其他的文法的研究提供支持，具有较好的通用性。因此我们选择 CFG 来标注语料。

在分析器的设计方面，我们采用基于概率的广义LR分析方法。在分析语料时，先由分析器自动分析语句，有歧义或无法分析时，由人作出选择或增添、修改文法，而这些人机交互和分析的历史信息被记录下来，用来完善和修改文法产生式及其概率参数。这样对某一类语料的标注过程，也是一个生成和完善该类语料的文法的过程。

3·2·3　语义特征标注

目前语义特征标注研究较少，其根本原因在于人们对于语义问题的研究还很不成熟，尤其是对于语义特征集的选择与设计还存在较多的争议。但是我们应当看到，没有经过语义加工的语料所能反映的知识是相当有限的。因此我们选择设计了一套语义特征系统，并

对语料库标以语义特征，以从中获得一些在机器翻译的消歧过程中非常有用的知识：各个单词的语义特征集合，动词与其支配成分之间的语义约束条件等。而在目前的大多数机器翻译系统中，这些知识大都来源于系统的开发者的经验和感性认识，缺乏对真实语料的研究，知识的全面性、系统性和真实性难以保证。

3·2·4 建立源译文的对应 (Aligning)

即在两种语言的语料中建立源文和译文之间的对应关系。这种对应是多层次的，可以是文章与文章之间、段落与段落之间、句子与句子之间、句法单位与句法单位之间、单词与单词之间的对应。随着对应层次的深入，自动建立对应的难度越大。Gale 和Brown 分别给出了一系列基于概率统计的模型来建立句子与句子之间、单词与单词之间的对应[4][5]。作者认为建立句法单位之间的对应对于机器翻译的研究来说意义更为重要，因为只有建立了句法单位的对应，我们才能对比源译文在句法结构上的差异。但是在大规模的语料库中建立这种对应关系，难度很大，还有待进一步的研究。

3·3 语料知识获取模块

从语料中获取语言知识是语料库研究的目的所在。如果我们能够从语料库中获取科学而系统的语言知识，将其填充到机器翻译系统的语言知识模块中，进而与翻译执行模块结合，就能方便地生成机器翻译系统了。这种思想对于设计高质量的专业机器翻译系统尤为重要。理想的情形是，给定某专业的语料（具有足够的代表性），通过对其的加工，从中获取诸如词法、句法及生成规则等知识，并与机器翻译的翻译执行模块相结合，生成专业翻译系统。

3·3·1 词典知识的获取

对于采用基于概率的分析方法来说，词典中必须包含足够的概率信息。我们的词典中下列信息可以来源于语料：多类词的各个词类出现的概率，各个词类与上下文的相关（同现）概率；固定搭配的出现概率；单词的复杂特征集；动词与其支配成分之间的选择约束条件等。这些信息可以通过对经过词类标注和语义特征标注的语料进行统计而得到。另外，部分词典的知识可以从译后编辑的反馈信息中得到。

3·3·2 文法知识的获取

在我们的环境中，文法知识的获取有两种途径：一是从未经用句法标注的语料中获得CFG 的产生式和其他知识，如3·2·2节所述，通过人机的交互作用，在标注语料的过程中同时获取、完善文法；二是对已经标注了的语料，可以直接进行统计和归纳，生成该类语料的 CFG 文法。

3·3·3 译文生成规则的获取

译文生成规则的获取，可以通过两条途径：一是从建立了句法单位之间的对应关系的双语语料库中获得，通过对比源译文之间在句法结构上的差异，抽象出源语结构转化生成目标语结构的规则，但是，正如我们前面所述，由于建立双语的句法单位之间的对应关系还缺乏有效的方法，这条途径难度较大；二是通过译后编辑的反馈信息，对目标语生成规则进行优化、修改。

3·4 译后编辑及反馈

要从用户的译后编辑中获得反馈信息，关键在于获取系统输出的译文与用户编辑之后的译文之间的差异。不加任何限制地允许用户修改译文，不利于我们获得用户修改的真正意图，难以把握修改后的译文的句法结构。我们采取的策略是：为用户提供有限的修改译文的操作手段及相应的工具，跟踪并记录用户修改译文的操作过程，根据这些记录调整译文的句法结构（语法分析树），再根据所作的调整，找出引起调整的原因，逐级优化生成规则、源语的文法乃至词典的知识。

4 结束语

本文简要介绍了 CBMT 模型，并提出了 CBMTE 环境的设计思想。由于不论是语料库还是机器翻译的研究，都是工作量巨大的工程。因此，该环境的设计必然还有许多难以预料的实际困难没有考虑到，只能在实现的过程中不断完善设计思想。

本文得到 863 计划 863-306-03-06-3 课题的支持。

参考文献

[1] 苑春法，黄昌宁等，新一代语料库的建设与管理，计算语言学研究与应用，1993.

[2] 黄昌宁，苑春法，国外语料库述评，机器翻译研究进展，电子工业出版社，1992.

[3] 刘开瑛等，自然语言处理，科学出版社，1991.

[4] Brown, P.F.; et al. The Mathematics of Statistical Machine Translation : Parameter Estimation, Computational Linguistics, Vol.19(2), 1993, 263-311.

[5] Gale, W.A.; and Church, K.W. A Program for Aligning Sentences in Bilingual Corpora, Computational Linguistics, Vol.19(1), 1993, 75-102.