维普资讯 http://www.cqvip.com

计算机科学 1996 Vol. 23 No. 2

语料库和机器翻译*)

王.挺 陈火旺 史晓东
(国防科技大学计算机系 长沙410073)

摘 要 In this paper, at first we discuss the difficulties faced by the Machine Translation(MT) researchers and demonstrate the reasons of the rapid growth of research on corpus. Then the processing technologies of corpus are illustrated. We also discuss in detail how the corpus technology support the MT research. At last, the authors present a corpus-based MT research model (CBMT).

关键词 Corpus, Machine translation.

一、前 言

50年代,机器翻译被视为一个纯粹的"解码"问题,基于经验的、概率统计的方法在机器翻译的研究中非常盛行.但是由于当时计算机性能和资源(如计算速度,存储容量,机器可读的文本(语料))十分有限,对于这类方法的支持远远不够,因此这类方法很快就被人们冷落了.随着乔姆斯基文法体系的建立,为机器翻译的研究提供了一个坚实的理论基础,基于规则的第二代机器翻译系统得到了迅速发展,能力(Competence)模型得到了人们的极大关注.却忽视了对大量真实语料的调查研究,于是当这类系统用于处理大规模的真实语料时,人们发现基于规则的机器翻译系统存在着许多严重的缺陷,主要表现在:

1. 为了描述复杂的自然语言,文法规则的规模越来越大,人们在设计规则的过程中不可避免地会带有主观随意性,因此文法的维护和一致性的保证越来越困难.人们日益感到文法的产生应当基于大量的真实的语料,必须建立一种基于语料的支持文法生成和维护的工具;

2. 基于规则的系统处理歧义的能力差.规则所表示的语言知识是普遍的、理性的和大粒度的,因而在分析自然语言的句子时,常常会生成许多分析树,尽管许多情巧的理论为消除歧义提供了支持,但这些理论的实际运用却差强人意,造成这种困境的一个主要原因是这类系统忽视了语言中特殊的、经验性的和小粒度的知识,而这些知识对于消除歧义起重要作用;

3. 基于规则的系统表示和处理的词法、句法知识的广度和深度十分有限.由于缺乏对大量语料的调查、加上在系统设计和实现上的一些技术原因,这类系统的文法规则的数量是有限的,对于规则描述以外的语句系统无法处理,因而其灵活性较差.

正是由于基于规则的机器翻译系统遇到了这些困难,人们认识到对大量语料的研究是必不可少的.尤其是近二十年来计算机技术的迅速发展,计算机的速度、容量和机器可读的大规模文本都为语料库的研究提供了强有力的支持,基于概率统计的、经验性的机器翻译方法再度唤起人们的关注.基于概率统计的方法有利于表示语言中特殊的、小粒度的知识,善于处理语言的不确定性.因此,将语料库技术、概率统计方法结合到基于规则的机器翻译系统中成为当前研究的一大趋势.

二、语料库及其加工技术

语料库是大量的、能代表某一领域的语言现象的真实语言材料的集合.人们建立语料库,是期望从中获得对真实语言现象和规律的认识.最原始的语料,如果不经过任何加工,对于机器翻译的研究来说意义不大.只有当语料被加工之后,蕴含在语料中的语言知识被标识出来,人们才能在大量的语言现象

*)得到863计划863-306-03-06课题的支持.王挺 博士生,主要研究方向为计算机软件、机器翻译、语料库语言学等.陈火旺 教授、博士生导师,主要研究方向为计算机软件、人工智能等.史晓东 博士,主要研究方向为计算机软件、机器翻译等

维普资讯 http://www.cqvip.com

中总结出规律并用之于机器翻译的研究。目前对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等。

2.1 词法标注

词法标注是对语料中的单词标以词类。词类的标注主要有两种方法：一是基于规则的标注方法，如BROWN语料库的标注系统TAGGIT，其准确率为77%；另一种是基于概率的方法，如LOB语料库的标注系统CLAWS，其准确率为96-97%。后者比前者更准确，但是后者通常需要有大量已经标注好的语料作为训练语料，从中获取标注系统的概率参数，来保证标注的准确率[1]。近年来，人们为了克服这个困难，将隐含马尔科夫模型(HMM)引入到标注研究中，把词类标注问题视为：已知一词列(即观察序列)，如何求出其正确的词类序列(即隐藏状态序列)。这样就可以把标注系统的参数作为一个HMM的参数用未标注的语料进行训练。这种方法无需消耗大量的人力来标注训练语料，但训练过程中所需的计算较多，准确率也有所下降[2]。

2.2 句法标注

所谓句法标注，就是将语料中的句法单位标注出来，直观地说，就是标出各个句子的语法分析树。这一过程与机器翻译中的语法分析是一样的。句法标注通常有三种方法。

2.2.1 基于规则的标注方法。使用语法规则及分析程序分析语料，以获取语法分析树。这种方法的好处在于使得标注的结果符合统一的文法，具有较好的一致性。但是这种方法具有基于规则的系统的通病：一是消歧的能力差，分析的结果经常是多棵分析树，必须用手工的方法对这些分析树进行判断和挑选，耗费大量的人力，在标注大规模的语料库时，这个问题更加突出；二是该方法不能处理文法覆盖范围以外的语句，缺乏灵活性。

2.2.2 基于概率统计的标注方法。使用句法单位间的互现概率对语料进行句法标注[2]。其长处是对语料的分析不依赖于文法规则的设计，仅根据句法单位之间的相互信息统计来分析语料。该方法在处理较短的句子时可以获得较好的结果(在分析15词以下的句子时，其准确率可达98%)。但在处理较长的句子，尤其是带连词的复合句时，其准确率有所下降(在分析30词以下的句子时，其准确率为94-95%)。另外，这种方法要用已经标注好的语料训练概率参数(Macus 用BROWN语料库进行训练和测试)。

2.2.3 概率统计与规则相结合的标注方法。将概率参数引入到规则系统中，即使用基于规则的分析方法生成可能的语法分析树，用概率传播来计算各个分析树的可靠性，作为选择最佳树的标准。一般来说，我们可以给每一条规则赋予一个概率参数，如对于上下文无关文法，其规则形为：

S→NPVP(p)

其中，p 是赋予该产生式的概率。Brisoe 和 Caroll 提出了一种更精细的方法，采用广义LR分析方法，将概率参数赋予LR分析表的每一个动作项，而不是文法规则。这种方法能揭示出自然语言中小粒度的相关性[2]。

2.3 语义特征标注

对语料进行语义特征标注的前提是选择一套语义特征。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，对于语义特征的选择与组织还存在较多争议。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有限的。对于机器翻译来说，精心选择一套语义特征系统，并对语料库标以语义特征，那么我们至少可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，动词与名词、形容词、副词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，这些知识都来源于系统的开发者，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。

2.4 双语的对应(Bilingual Aligning)

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，建立对应的难度越大。Gale 和 Church 根据源文中的较长句子对应的译文也较长的事实，提出了一个非常简单的使用统计方法的模型来建立双语(英、法)之间句子与句子之间的对应关系。这一模型根据句子中字符的个数，使用动态规划(Dynamic Programming)方法寻找句子之间的最佳对应关系，其准确率达到96%。这种模型的准确率会由于不同的双语而有所不同[3]。另外，在建立句子内部单位之间的对应的研究方面，Brown 提出了一系列基于概率统计的模型来建立单词之间的对应[4]。

三、语料库对机器翻译的支持

基于规则的系统本身所固有的缺陷难以克服、

维普资讯 http://www.cqvip.com

人们日益认识到将语料库技术结合到机器翻译中来，对于克服这些困难是大有帮助的。语料库可以从多方面、以多种方式为机器翻译的研究提供支持，主要有以下三大类：

1. 基于双语语料库的、单纯使用统计方法的机器翻译。这种方法完全抛弃了规则，仅使用概率统计来进行语言之间的翻译。根据贝叶斯理论：

$$Pr(T|S)=\frac{Pr(T)Pr(S|T)}{Pr(S)}$$

其中，S、T 表示源语和目标语的句子，Pr(S)、Pr(T)分别表示 S、T 在各自语言中出现的概率，Pr(S|T)被解释为在 T 出现的情况下，其对应的源语言句子为 S 的概率。那么，机器翻译可视为：给定源语句子 S，通过计算来寻求一个目标语句子 T 使得 Pr(T|S)为最大的过程。即：

$$\mathbf{T}=\max_{T} Pr(T|S)=\max_{T}\frac{Pr(T)Pr(S|T)}{Pr(S)}$$

由于 S 是给定的，上式可化为：

$$\mathbf{T}=\max_{T} Pr(T)Pr(S|T)$$

Brown 在文[3]中提出了两个模型：1)用于计算 Pr(T)的语言模型，其参数的值来源于目标语的语料库；2)用于计算 Pr(S|T)的翻译模型，其参数的值来源于源语-目标语的双语语料库。翻译模型要求在双语之间建立单词一级的对应，用这些语料训练模型的参数。虽然这种方法的译文准确率仅为48%，但是若在其译文的基础上进行译后编辑生成正确的译文，则比手工翻译减少60%的工作量。

2. 基于语料的机助翻译。由于现有的全自动机器翻译系统的能力和人们的实际要求相距甚远，机助翻译作为一种折中的方案，力求将现有的机器翻译成果用到实际系统中，尽量发挥计算机的作用，最大限度地支持人工翻译。语料库也为机助翻译提供了有力支持。加拿大的 CITI 为政府的翻译局开发了机助翻译系统 *Translation Workstation*，该系统为英法双语料库建立了句子一级的对应。用户在翻译的过程遇到疑难的单词、词组，可以通过对应关系，找出语料库中包含这些单词和词组的句子及其译文，供用户参考[6]。

3. 语料库对基于规则的机器翻译系统的支持。从上面的分析我们可以看到，单纯的基于统计方法的机器翻译系统和单纯的基于规则的机器翻译系统都有长处和不足：前者能较好地处理语言中的不确定的、小粒度的知识，灵活性好，但不便于反映语言中确实存在的确定的语法规律和语义知识；后者能较好地处理语言中的确定的、大粒度的知识，句法、语义分析方法较丰富，但不能表示语言中的不确定的、小粒度的知识，灵活性差。因此将两种技术结合起来应当是有益的。语料库研究的兴起不应当视为基于规则系统的没落，应当用语料库技术来支持基于规则的系统的研究和开发，在基于规则的系统中结合语料知识和统计方法，以克服现有机器翻译系统中存在的一些问题。我们认为语料库技术可以在下面四个方面支持基于规则的机器翻译系统的研究和开发：

1. 词典信息。机器翻译系统中的词典包含了机器翻译过程中所需要的词法、语法和语义信息，这些信息对于消除歧义非常重要。词典的词汇的选取应当基于对大量语料的统计，这对于科学地保证面向子语言领域的机器翻译系统的词汇量尤其重要。此外词典中的词的固定搭配、语义特征、单词之间的约束条件等知识均可以通过对语料进行加工（如词法标注、句法标注、语义特征标注等），经过统计和提炼而来。尤其是对于一词多义、一词多类等情况，我们可以用从语料中统计出各种词类、词义的出现概率以及相应的上下文的约束条件，为机器翻译中的消歧提供依据。词典的知识只有来源于大量真实语料才是真正可靠和全面的。

2. 文法规则。困扰基于规则的机器翻译系统的一个问题是，文法规则的规模往往很大，而且在处理真实而复杂的语料时又必须不断地对文法进行扩充、修改，在设计和修改规则的过程中又常常带有很强的主观随意性，缺乏科学的依据，文法的一致性维护也很困难。但是如果我们以大量的语料为基础，建立一套界面友好的工具，支持人们从语料中抽取文法规则，由计算机辅助人们按照一定的方式来设计文法规则，完成一致性的维护，这个问题的解决就稍微容易一点了。从语料中抽取文法规则是一个交互的过程，在这个过程中一个方便而有效的工具是必要的。

另一方面，在本文的语料库的句法标注一节中，我们看到概率参数可以赋予文法的规则甚至是 LR 分析表中的动作，以表示不确定的、小粒度的知识，而语料库则可以为这些参数提供依据。我们可以通过直接对标注过的语料进行统计获得参数值，或者用语料逐步训练文法中的参数，使其达到最优。

3. 目标语的生成。目标语生成规则的表示和设计是机器翻译中的一个难点。语料库为我们提供了一条值得尝试的途径：通过在双语语料库中建立句子内部的语法单位一级的对应关系，找出源语和目

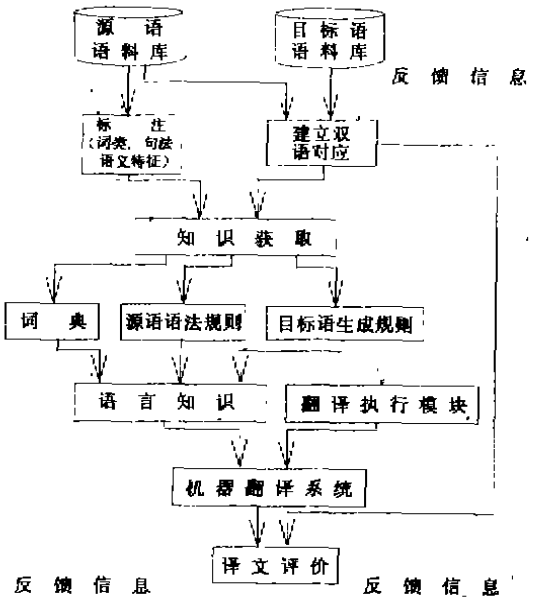
维普资讯 http://www.cqvip.com

标语的对应的语法单位之间的位置变化关系，进而从中抽象出源语到目标语的生成规则。这种从语料中获取生成规则的方法对于提高机器翻译系统的模块性、可伸缩性非常有益。

4. 测试与评价。双语语料库可以为机器翻译系统的测试与评价提供平台。将机器翻译系统的源语句子的分析结果及译文，与语料库中的源语和目标语的对应关系进行比较，可以得到一个大致的评价。在这里源语与目标语之间的对应关系的层次决定了语料库在测试和评价机器翻译系统中所起的作用，对应的层次越深，我们所作的评价就越准确。目前这方面的研究还相当欠缺。

四、一个 CBMT 模型

通过上面的分析，我们可以看到基于规则的方法和统计方法都各有优劣，因此我们设计了一种基于语料库的机器翻译研究模型(CBMT)，用以将统计方法与基于规则的方法结合(图1)



CBMT 模型

图1

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，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、语法分析、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作出评价，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 CBMT 模型中，语料的加工(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。

参 考 文 献

[1]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

[2] Briscoe T. 等，Generalized Probabilistic LR Parsing of Natural Language (Corpora) with Unification-Based Grammars，Computational Linguistics，Vol. 19(1)，1993

[3] Brown P. F. et al.，A Statistical Approach to Machine Translation，Computational Linguistics，Vol. 16(2)，1990

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

[5] Gale W. A. 等，A Program for Aligning Sentences in Bilingual Corpora，Same to [2]

[6] Isabelle P. 等，Machine Translation Today，CITI Report 1990

[7] Magerman D. M. 等，Parsing a Natural Language Using Natural Information Stattistics，Proc. of National Conference on Artificial Intelligence，1990

[8] Merialdo B.，Tagging English Text with a Propabilistic Model，Computational Linguistics，Vol. 20(2)，1994