口语机器翻译 语音翻译 机器翻译 (4)

计算机科学1998Vol.25№.5

47-50

口语机器翻译研究综述*)

The State of the Art of Spoken Language Translation

王海峰 高 文 李 生

(哈尔滨工业大学计算机系 哈尔滨150001)

TN912.3

TP391.2

摘 要 In this paper, we firstly introduce the state of the art in the research of spoken language translation. Then we discuss some essential differences between spoken language and its written counterpart. Finally, we examine the fundamental requirements of spoken language translation and propose some approaches to solve them.

关键词 Machine translation, Speech-to-speech Translation, Spoken language

一、引言

开发口语语音翻译(Speech-to-Speech Translation)系统是语音识别、机器翻译、自然语言处理、乃至整个人工智能界的终极目标之一,它通过对不同语言的话语的自动互译,使得语言不通的人们之间进行没有语言障碍的自由交流成为可能。它的实现不仅需要准确的非特定人、大词表的连续语音识别,恰当表达话者意图的机器翻译,流畅的目标语语音合成,而且需要各个功能模块作为一个整体的有机结合。

口语语音翻译的应用除了显而易见的科学上和工程上的意义外,还将带来巨大的经济和社会效益。在目前的理论和技术条件下,不受限制的、任意话语的语音翻译系统的实现还只能是一个可望而不可及的美好目标,但即使是一个小型的面向特定领域的语音翻译系统,也会有着非常现实的需求和良好的应用前景。很多公共服务部门,如报警服务、电话号码查询、宾馆客房服务、航空公司机票预订等都可能要面对不懂当地语言的服务对象(如旅游者和移民等),于是面向这些特定领域的口语语音翻译系统就有了用武之地;信息是超越国界的,信息服务系统也必须能够提供多种语言的查询服务,而如果有口语语音翻译系统提供对口语语音查询的自动翻译,数据库的提供商们必将从中受益匪浅;对于两个持不同语言的人,他们之间的交谈就只好借助于人的翻译了,这无疑为自由地交流带来了麻烦,尤其当谈话内容涉及到某些不宜于被第三者知晓时,这种语言的鸿沟尤难以逾越了,而口语语音翻译的实现将会十分理想。

虽然与口语语音翻译相关的研究领域如机器翻译、语音识别等已有很长的发展历史,但直到八十年代末,国际上才出现了有关语音翻译的研究,并先后出现了 SpeechTrans(CMU)、JANUS(CMU)、SL-TRANS(ATR)等系统[1],而北野先生(Hiroaki Kitano)在日本京都大学期间所研制的使用大规模并行计算的、基于实例的语音翻译系统第一次向人们证明了实时(毫秒级)的口语语音翻译是可能实现的,他也因此获得了1993年度 IJCAI 的计算机与思维奖[1]。

鉴于语音翻译重要的科学研究意义和显著的工程实践意义,从90年代初开始各国政府和各种研究机构都对这方面的研究开发作了大量的投入。其中,当数由德国联邦教育、科学、研究与技术委员会(BMBF)支持的 Verbmobil 工程规模最大。Verbmobil 从1993年开始制定了预计为期八年的研发计划,其中从1993年到1996年的第一阶段共吸收了德国、美国和日本的29个来自企业界和高校的成员单位,在这一阶段总共投入来自政府的资金0.649亿德国马克,来自企业界的资金0.31亿德国马克。Verbmobil 第一阶段的目标是建立一个非特定人的、面向会面安排领域的交谈的口语语音翻译系统,现在其原

*) 本课题研究受国家863计划支持,课题合同号863-306-03-01。

型系统已经完成。

1993年成立的国际语音翻译联合会(Consortium for Speech Translation Advanced Research,简称C-STAR)是一个以口语语音机器翻译为基本研究目标的国际合作组织,现在有来自12个国家的共20个成员,包括美国的卡内基-梅隆大学(CMU)、麻省理工学院(MIT),日本的ATR-ITL(Advanced Telecommunications Research Institute International-Interpreting Telecommunications Research Laboratories),德国的卡尔斯鲁厄大学(Karlsurhe Univ)、西门子公司(Siemens)等。我国的哈尔滨工业大学和中科院计算所也于1996年以观察员身份参加入了该组织的活动,开始了口语语音翻译的研究,目前主要承担汉英口语语音翻译系统的研究与开发,与其他成员单位相一致,将翻译领域定为会面安排(Meeting Schedule)[2]。

二、口语翻译的特点

一个语音翻译系统需要包含语音识别、机器翻译和语音合成等多种功能模块,其中口语翻译模块的输入是语音识别模块对口语语音进行识别的结果,它除了要面临一般文本翻译的共同困难外,还要处理一些口语翻译所特有的问题,具体来说主要包括以下几个方面[3]。

1. 口语的不连贯性:人在口语交谈中,与说话人有关,经常会出现一些使句子不连贯的成分,如错误的开始、重复、没有结尾和语气词的插入等等。例如:句子"哎,咱们……咱们下午谈一谈";"明天,噢不,后天有个讨论会";"后天上午,嗯…八点吧"。

2. 语法约束相对较弱:一般来讲,即使没有上述的不连贯因素存在,口语中也很少会有严格符合语法约束的结构完整、正确的句子,而大量存在的是系统的语法规则无法处理的现象。这一方面是由于系统的语法知识对语言现象的覆盖程度不足;而更主要的则是口语本身的特性所决定的,例如口语中高度的省略、大量的指代等。

3. 没有明确的句子边界:口语中没有标点符号来标志句子,也基本没有传统意义上的句子。在口语交谈中,一个对话轮次(a dialogue turn,一个说话者在交谈中连续说出的一段话)往往会包含多个意段(utterance,一个对话轮次中表达一个概念的一段话)。例如,"噢 星期三 星期三不行 星期三我有会 星期四怎么样 星期四上午九点可以吧"。

4. 语音识别的错误:在现在的技术条件下,语音识别有一定的错误率是必然的,因此翻译模块就必须对误识别有较强的容忍和适当的处理,使这种错误不至于扩散而严重影响翻译质量。

5. 多种可选输入:语音识别模块所给出的往往不是一个唯一的句子,而是附有可信程度打分的词格形式(Lattice)或N-Best形式,口语翻译模块需要针对这种情况作出分析和选择。

以上是口语翻译相对于书面语翻译较困难的一面。同时,口语也有其相对较容易的一面,那就是口语对话一般结构都不很复杂,而且也不会出现那种动辄数十字甚至数百字的长句。

口语的这些特点决定了口语翻译的研究既要吸取书面语翻译的一些经验和方法,同时更要从其自身特点出发进行有针对性的研究和处理。以下内容将讨论口语翻译研究中要考虑的一些基本问题并介绍其研究中的一些基本方法。

三、口语翻译的基本问题

根据研究者所采用研究策略的不同,口语翻译系统各具特点。一般地,进行口语翻译研究时应考虑如下一些基本问题:

1. 鲁棒性。良了的鲁棒性无疑是口语翻译系统应具备的基本特性。很难期望一个脆弱的系统面对灵活多样甚至是包含着不可预期的错误的口语输入会得出什么好结果。口语翻译系统的鲁棒性应包含如下几个方面:对口语固有的不连贯、语法约束较弱、句子边界不明确等现象的鲁棒性;对一些噪声输入,尤其是语音识别器的错误的鲁棒性;对多种可选输入和歧义现象的鲁棒性。如何提高系统的鲁棒性是口语翻译研究中的首要问题。

2. 各种知识的综合运用。一个语音翻译系统所需要使用的知识是多层次的,大体上,应包含如表1所示的几个层次的知识[4]:

表1 语音翻译中所涉及的多层次知识

层次	内容	作用
1声学层	语言的节奏、韵律和语调、声调	形成音素
2音素层	发出的声音	形成语素
3语素层	构成词的子单元	形成词
4词法层	词	衍生出意义的单元
5语法层	词的结构功能	形成句子
6语义层	上下文无关的意义	衍生出句子的意义
7段落层	句子的结构功能	形成一段对话
8语用层	上下文有关的意义	衍生出句子在上下文中的意义

对于语音翻译中的口语机器翻译模块，将涉及到其中3到8层的知识，鉴于前面提到的口语翻译的特点，照搬针对文本的翻译方法试图将对话逐句逐词地完全翻译显然是不合适的，为了得到有用的翻译结果，一个基本的策略应该是：在一定的对话背景下，作出适当的语义和语用分析，从而恰当地抽取对话的基本语义内容，进而得出准确地传递了话者意图的翻译结果。于是，多种知识的综合使用就成了众多口语分析和口语翻译研究者的共识。如何有效地利用好各种知识是提高翻译能力的关键。

3. 增量式翻译。增量工作方式是指边接收输入边进行处理，而不必待接收到完整的输入后才开始工作的方式。增量方式是符合人的认知过程的。对人的认知过程的研究表明，人对句子的语义理解是在接收到一个完整的句子甚至子句之前就开始了，而不会等到一个完整的句子说完后才开始理解的过程。同理，口语翻译系统所包含的多个模块的工作也应是增量式的，只要前一模块产生部分输出，后一模块就应针对这部分输入开始工作[3]。值得指出的是，对综合运用多种知识的口语翻译系统来讲，分别运用不同知识的各模块间的作用是相互的，不但前导模块的工作会对后续模块产生影响，而且后续模块的工作也会反过来影响前面的模块，这就产生了两种可能的方法：后续模块给前导模块一个反馈信息；或者，每一模块保留所有可能的选择并作为下一模块的输入，而当后续模块工作时，对前导模块产生的整个可能空间进行操作，并剔除其中不合理的部分，从而缩小搜索空间。前一种方法由于在系统中引入了较多的反馈和回溯，影响了效率，也使增量工作方式无法实现，所以众多研究者都倾向于采取后一种方式。增量式翻译是口语翻译系统的基本特征之一。

四、口语翻译的方法

机器翻译方法通常可分为基于规则的方法（包括基于转换的方法、基于知识的方法等）和基于语料库的方法（包括基于统计的方法、基于实例的方法以及联结主义的方法等）[5]，随着机器翻译研究的不断深入，混合方法的使用已受到越来越多的研究者的重视[6]，而对口语翻译来讲，综合利用各种方法的混合策略更成为其基本的研究策略。针对口语翻译的特有问题和各种语言本身的固有特点，各国研究者采取了各自不同的策略进行分析和翻译。这些方法大体可归纳为如下三种：

1. 以基于规则的方法为主体的方法，这类方法通常是对原有面向文本翻译的方法加以改造，使之适合口语翻译的需要。典型的例子是 CMU 的 GLR* 算法，众所周知，LR 文法是上下文无关文法的一个子集，它在编译系统中发挥了重要作用，而 Tomita 提出了广义 LR 算法（Generalized LR，GLR），GLR 通过图结构栈的引入可以分析一个入口对应多个动作的非 LR 的上下文无关语言，它在书面语的分析中取得了重要成功。面向书面语的 GLR 算法，以及几乎所有自然语言分析算法都是用来分析"纯净"的符合语法的输入的，它们拒绝接收任何不符合语法（哪怕是一点点）的输入，而这类算法对很少遵循规范语法的口语输入显然是无能为力的。为了适合于口语分析，Alon Lavie 通过对 GLR 的扩展引入了 GLR* 算法[7]，GLR* 通过忽略无法分析的词和片断而找出原始输入的可分析的最大子集。算法的扩展在提高了分析能力的同时也增大了分析的不确定性，为了及时地剪枝以提高效率，Lavie 又引入了概率消歧和属性约束的策略。其它的如德国汉堡大学的 Jan W. Amtrup 等人使用的基于表的方法[8]，SRI International 的 SLT（Spoken Language Translator）采用的是以 LR 算法为基础的规则与统计相结合的方法[9]，CMU 的 L. J. Mayfield 等人使用的则是以 RTN 网络为基础的基于概念的方法，等等。这些方法的共同特点是一方面在原有基础上引入新的机制以加强系统的鲁棒性（如统计信息的引入）；另一方面，其规则系统也更侧重于语义规则的使用，以适应口语翻译的需要。

2. 以基于实例为主的方法。基于实例的机器翻译方法最早由日本京都大学的长尾真教授在1984年提出。这种方法的基本思想是比照相似句子的翻译实例来进行翻译，它的最主要特点就是把翻译实例作为它的主要翻译知识源。基于实例的机器翻译方法有很多优点，对于口语翻译来讲，有一点是尤为重要的，那就是根据实例，而不局限于严格的规则约束，这种灵活性正适合口语语言灵活多变的特点。ATR 采用了基于实例的方法与基于规则的方法的混合策略[10]。在基于实例的翻译中，关键在于合适的实例的选择，而实例的选择又涉及到单词到句子各级相似度的计算，对于口语翻译来讲，一个合适的相似度计算方法应综合利用词法、句法、语义、上下文及语用等多种信息。

3. 以神经网络为主的方法。神经网络方法由于其较强的学习能力和良好的鲁棒性而受到了众多口语分析研究者的重视。但由于单纯的神经网络分析

几乎不包含什么语言学知识，其分析结果不利于进一步的利用和处理，尤其是不利于口语翻译中目标语的生成。因此，将符号主义与联结主义相结合的方法就受到了格外的青睐。Karlsruhe 大学的 Finn Dag Buo 等人将神经网络用于口语特征结构的分析，在进行组块分析时使用了多个含有一个隐藏层的前向网络，并采用 BP 算法和 PCL 算法进行网络参数的学习[11]。汉堡大学的 SCREEN 系统[12]等也使用了神经网络的方法。在将神经网络用于口语翻译时，需着重考虑两个问题，一是网络的拓扑结构，另一个是网络的学习算法。

结束语 口语翻译的研究极具理论和实践意义，对于我国计算语言学的研究者们来讲，汉外口语翻译的研究尤其是一个大有可为的领域。目前，我们已实现了一个面向会面安排领域的汉英口语语音翻译的原形系统[6]，取得了满意的效果。我们将继续在这方面进行深入研究。

参 考 文 献

[1] Hiroaki Kitano, Speech-to-Speech Translation: A Massively Parallel Menory-Based Approach, Kluwer Academuc Publishers, Boston, 1994

[2] 王海峰、高文、李生，面向受限领域的汉英口语翻译，《语言工程》，第三届中国计算机智能接口与智能应用学术会议论文集，清华大学出版社，1997.8

[3] Alex Waibel, Interactive Translation of Conversational Speech, IEEE Computer, 29(7), 1996

[4] Bill Z. Manaris, Brian M. Slator, Interactive Natural Language Processing: Building on Success, IEEE Computer, 29(7)1996

[5] John. Hutchins, Latest Developments in MT Technology: Beginning a New Era in MT Research, MT Summit IV, Kobe, Japan, 1993.7

[6] 赵铁军、李生、高文，机器翻译研究的现状与发展方向，计算机科学，23(3)1996

（下转第55页）

（上接第76页）

(4)重新计算各子类的中心点：$C_i=\frac{1}{n_i}\sum_{x_j\in i}X_j$；$n_i$ 为子类 i 中的样本总数，X_j 为所有属于子类 i 的样本。

(5)计算各子类的偏差值 b_i，以子类中所有样本到中心点的距离的平均值来度量，即：

$$b_i^2=\frac{1}{n_i}\sum_{x_j\in i}\|X_j-C_i\|^2$$

2.2 有监督学习过程

这一步通过学习得到输出层与隐层之间的权矢量，使网络的输出值与样本的理想输出之间的误差最小。定义误差函数 $E=\frac{1}{2}\|Y_{des}-Y\|^2$，$Y_{des}$ 为样本理想输出矢量，Y 为网络实际输出矢量，由式(3)求得。采用梯度下降法求解，并使用动量因子 α 提高学习速度，在迭代过程中，学习速率 η 和动量因子 α 自动调整。权植学习的主要步骤如下：

(1)以小的随机数初始化权值矩阵；

(2)由公式(3)计算网络输出值 Y；

(3)计算绝对误差

$$e=Y_{des}-Y;$$

(4)修正权植：

$$\Delta W_i(k+1)=\eta\cdot e\cdot R_i(X)+\alpha\cdot\Delta W_i(k),$$

$$W_i(k+1)=W_i(k)+\Delta W_i(K+1);$$

(5)计算系统误差，判断终止条件。若条件成立，退出循环；否则转(2)继续迭代。

小结 通过模拟实验，我们发现：采用 RBFNN 学习时，聚类粒度对规则总数和学习精度有影响，但测试精度基本稳定，而且精度较高。另外，RBFNN 学习速度快，一般只需几十次或几百次迭代就能得到稳定的输出。除聚类粒度外，与其它学习参数的初始值基本无关，因此学习精度容易控制。

参 考 文 献

[1] 刘有才、刘增良，模糊专家系统的原理与设计，北京：北京航空航天大学出版社，1995

[2] Nikola K. Kasabov, Learning fuzzy rules and approximate reasoning in fuzzy neural networks and hybrid systems, Fuzzy Sets and Systems, V82, 1996

[3] Y. J. Chen, Rule combination in a fuzzy neural network, Fuzzy Sets and Systems, Same to [2]

[4] 赵群、保铮，RBFNN 的分类机理，通信学报，No. 2, 1996

[5] 李凡，模糊专家系统，武汉：华中理工大学出版社，1994