经验知识在机器翻译系统中的应用
Techniques for Machine Translation System Using Empirical Knowledge

任 福 继
Fuji Ren

广岛市立大学
Hiroshima City University

摘 要：本文将介绍融合协调型机器翻译模型中的经验知识的利用策略。利用经验知识进行机器翻译也称为 EKMT (Empirical Knowleage based Machine Translation)。EKMT 是融合协调型机器翻译中的一种翻译生成途径。本文将介绍 KEMT 中的知识描述方法，语义距离计算方式，译文决定机制，协调策略以及部分语境处理手法。实验表明本文提案的策略对实现高质量的口语翻译系统是行之有效的。

Abstract: This paper introduces techniques for Chinese – Japanese Machine Translation system using Empirical Knowledge stored from actual translations (EKMT). In this paper, some methods for 1) describing and updating knowledge, 2) calculating the semantic distance between linguistic expressions, and 3) context processing are resented. Some experiment result are also given in this paper.

1 前 言

如何在自然语言处理特别是在机器翻译系统中有效地利用经验知识去提高翻译系统的质量以及系统的鲁棒性能是近年极其热门的课题。当然，如何从大规模真实语料中有效地抽取知识也是极其重要的研究领域。本文将介绍融合协调型机器翻译模型中的经验知识的利用策略。利用经验知识进行机器翻译也称为 EKMT(Empirical Knowleage based Machine Translation)。EKMT 是融合协调型机器翻译中的一种翻译生成途径。

EKMT 以转换模块为中心协调其他模块来共同完成翻译作业。EKMT 的翻译机制就是最大限度地利用经验知识，因此能生成高质量的译文而且有望以此为基盘开发出高效的会话翻译及自动电话翻译系统。

EKMT 在理论上虽与 EBMT(Exampled – Base MT：基于实例) 和 SBMT(Statistics – Based MT：基于统计)相似，但在 EKMT 中，利用经验知识的层次和范畴则与后者不同。前者从全体句子到任何语言片断均将输入原文与各种语言单位(句、节、短语等)的经验知识相匹配，根

据类似度来选取译文结构，消除歧义，因此 EKMT 将 EBMT 及 SBMT 的优点扩展到了翻译的全过程。

本文将介绍 KEMT 中的知识描述方法，语义距离计算方式，译文决定机制 C 协调策略以及部分语境处理手法。实验表明，本文提案的策略对实现高质量的口语翻译系统是行之有效的。

2 转换知识描述

EKMT 中的转换知识是将原语言表现与目的语言表现的对应关系以语义单位形式表示出来的经验知识。但原语言与目的语言并不是一对一的，所以，我们在描述转换知识时，采用了以关键要素作为指针。其数据结构如下：

原语言表现(键) → 目的语言表现 1 (E11,E12,...)
⋮
→ 目的语言表现 n (En1,En2,...)

在这个知识描述结构中，目的语言表现有 n 种，实例为 Eij 时，对应的译文为第 i 号目的语言表现。Eij 是语句的组合。在 EKMT 的转换处理中，首先求出输入语句与实例的语义距离，然后选择语义距离最近的实例 Eij，以此来选择目的语言表现 i。可以看出，这种方法是非常高效的。

3 语义距离计算

本系统是根据分类辞典概念阶层中的语义概念间位置关系来计算输入语句的词 i 和例句的词 e 之间的语义距离 d(i,e)。d 的值在 0～1 之间。越接近 0 表明 i 与 e 的语义越类似。

例如：d(白菜,胡萝卜) = 0.00
d(白菜,金属) = 1.00
d(会议,停留) = 0.05
d(北京,东京) = 0.00
d(会议,停留) = 0.05

对于输入语句 I 和例句 E 之间的语义距离，可用下式求出：

$$d(I,E) = d((i_1,\ldots i_n),(e_1,\ldots e_m))$$
$$= \sum_{k=1}^{n} d(i_k,e_k)\cdot w_k$$

式中，i_k 与 e_k 分别表示输入语句 I 和例句 E 的第 i 个单词，W_k 表示第 k 号单词的加权值。目前，W_k 取其平均值 1/n。这里需要指出的是：随着大规模语料库处理技术的进展，W_k 有望求出其最佳数值。

4 译文决定机制

在 EKMT 中，转换知识分为若干个类别，如表 1 所示。

表 1 转换知识分类

语言单位	例	
复句	AのでB	→ 因为A′，所以B′
单句	Aおねがいします	→ 请A′
短句	AのB	→ A′的B′
格	AをB	→ B′A′
全文	どうも有难うございました	→ 谢谢

由表 1 可以看出，CTM 是在翻译的各个层次中反复利用经验知识去寻求最佳匹配模式，从而提高了译文精度。下面用一个简单的例子来说明译文的决定机制。

例：I = you eat apple。

(1)转换知识

A eat B → A′吃 B′((He, $1, Vegetables), (tiger, $1, man)....)
A′腐蚀 B′((Acid, $1, metal).(...), ...)
:
:

(2)语义距离

$$d(I,E11) = d((you,apple),(He,Vegetables))$$
$$= d(you,He) * w1 + d(apple,vegetables) * w2$$
$$= 0.00$$
$$(w1 = w2 = 0.5)$$
$$d(I,E12) = d((you,apple),(tiger,man))$$

$$= 0.5 \times 0.5 + 0.25 \times o.5$$
$$= 0.375$$
$$d(I,E21) = d((you,apple),(Acid,metal))$$
$$= d(you,Acid) * w1 + d(apple,metal) * w1$$
$$= 1$$

(3)译文表现

由于 d(I, E11) = 0 为语义距离最小的项，所以选取第 i 行作为日的语言表现，即 A 吃 B，所以译为：

你吃苹果

5 协调机制

由于 KEMT 在各个层次反复利用经验知识，所以根据需要，系统自动地与形态解析(单词切分)、构文解析、上下文处理、语义解析等模块交换信息。其概念如图 1 所示。各协调模块是整个系统各路径共用的。

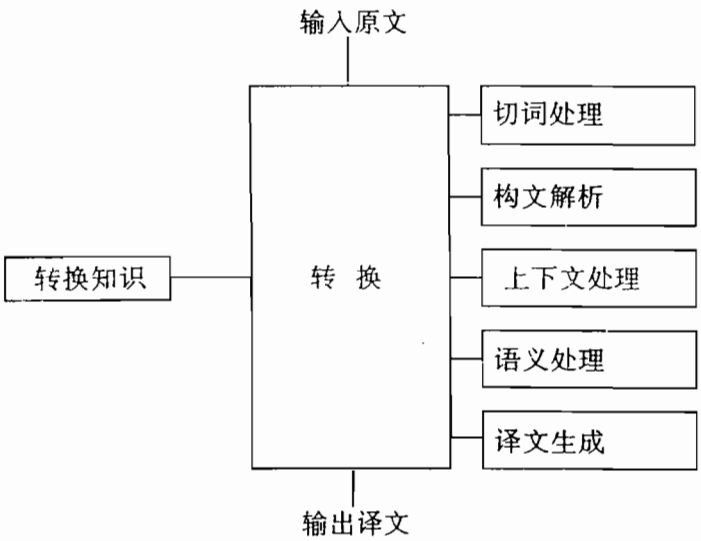

图 1 CTM 中的协调概念

5.1 单词切分

单词切分后，各单词不仅仅带有属性、词性，而且还赋有在语义分类中的编码。对有些长句，例如「おめにかかる」，「もうしわけございません」等则以语义形式作为一个单词处理。

5.2 鲁棒性

系统鲁棒性也可称为系统的容错能力，这里的错并不是指无原则的输入差错，而是人们在日常中几乎默认了的语法错误。例如，在对话系统中，日文常常出现"私山田です"，省略了格助词"は"。CTM 中，根据解析知识首先将原文进行补完，再利用经验知识来加以处理。对该例的解析知识为：

代名词 固有名词 → 代名词は 固有名词

5.3 语境处理

有些句子，单凭该句子的信息很难得到满意的译文，因此有必要引入上下文关系，即语境处理。但目前无论从理论上还是从实践上对语境处理而言都存在着极大的困难。在 CTM 中仅限于国际会议咨询领域导入了语境概念。主要是根据上下文来解决代指和省略问题，以及部分歧义处理。例如：对于日文的"はい"根据上下文关系可以很好地生成：

是的（前文是疑问句）
一定（前文是命令句）
你好（前文是问候句）
明白了（前文是针对询问句的反问句）

6 考　察

本文提出了研制新世代机器翻译应该着重于经验知识的有效利用这一思想，并且主张采用协调融合方式来吸取各种翻译模式的特长。同时，介绍了基于经验知识的机器翻译策略。对有关国际会议咨询语料 270 句进行了封闭实验，结果表明：对这一领域若不采用经验知识的话，其正译率仅为 50%，加入经验知识后其正译率上升为 85%。如果继续增加转换知识的话，正译率还会有所提高。实验表明，该策略是有效的，并且可以应用于会话翻译，电话自动翻译系统中去。

今后需要研究的是如何开发自动求出各单词间的语义距离的算法，以及系统如何从巨大语料库中自动获取经验知识的途径。

本研究是广岛市立大学特定研究课题及日本文部省研究课题，得到了文部省奖励研究（课题号：07780344，研究代表：任福继）及广岛市立大学特定研究费的支助，在此一并致谢。

参 考 文 献

[1] 黒桥祯夫,长尾真:长い日本语文における并列构造の推定,情报处理学会论文志 1991,Vol.33.No.8,pp.1022－1031.

[2] Brown,P.F.,Cocke,J.,Della Pirtra,S.A.Della Pietra,V.J.,Jelinek,F.,A Statistical Approach to Machine Translation,Computational Linguistics,1990,Vol.16,No.2,pp.79－85.

[3] Sumita,E.,Oi,K.,Furuse,O.,Iida,H.,Higuchi,T.,Example－Based Machine Translation on Massively Parallel Processors,Proc.of IJCAI－93,1993,pp.1283－1288.

[4] 范莉馨,任福继,宫永喜一,栃内香次:中日机械翻译における离合词の处理,情报处理学会论文志 A1994,Vol.35,No.8,pp.1702－1713.

[6] 范莉馨,任福继,宫永喜一,栃内香次:关连语をもちいた文のぶんかいにもとづく中日机械翻译,情报处理学会论文志 A1994,Vol.35,No.12,pp.2712－2724.

[7] 任福继:次世代自然言语における超并列处理,广岛市立大学研究学报 HCU－IS－95－023,pp.1－19(1995)