中 文 信 息 学 报
第 20卷 第 5期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 20 No. 5

文章编号: 1003 - 0077(2006) 05 - 0001 - 09

2005统计机器翻译研讨班研究报告

徐 波[1], 史晓东[2], 刘 群[3], 宗成庆[1], 庞 薇[1], 陈振标[1], 杨振东[1],
魏 玮[1], 杜金华[1], 陈毅东[2], 刘 洋[3], 熊德意[3], 侯宏旭[3], 何中军[3]

(1. 中国科学院 自动化研究所, 北京 100080; 2. 厦门大学, 福建 厦门 361005;
3. 中国科学院 计算技术研究所, 北京 100080)

摘要: 2005年 7月 13日 至 15日, 中国科学院自动化研究所、计算技术研究所和厦门大学计算机系联合举办了我国首届统计机器翻译研讨班。本文主要介绍本次研讨班参加单位的测试系统和实验结果, 并给出相应的分析。测试结果表明, 我国的统计机器翻译研究起步虽晚, 但已有快速进展, 参评系统在短期内得到了较好的翻译质量, 与往年参加 863评测的基于规则方法的系统相比性能虽还有差距, 但差距已经不大。从目前国际统计机器翻译研究的现状和发展趋势来看, 随着数据资源规模的不断扩大和计算机性能的迅速提高, 统计机器翻译还有很大的发展空间。在未来几年内, 在基于短语的主流统计翻译方法中融入句法、语义信息, 必将成为机器翻译发展的趋势。

关键词: 人工智能; 机器翻译; 统计机器翻译; 基于短语的翻译模型; 机器翻译评测
中图分类号: TP391 文献标识码: A

Current Statistical Machine Translation Research in China

XU Bo[1], SHI Xiao-dong[2], LIU Qun[3], ZONG Cheng-qing[1],
PANG Wei[1], CHEN Zhen-biao[1], YANG Zhen-dong[1], WEI Wei, DU Jin-hua[1],
CHEN Yi-dong[2], LIU Yang[3], XIONG De-yi[3], HOU Hong-xu[3], HE Zhong-jun[3]

(1. Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China; 2. Xiamen University, Xiamen, Fujian 361005, China; 3. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: Institute of Automation, Institute of Computing Technology of Chinese Academy of Sciences and Department of Computer Science of Xiamen University held the first Statistical Machine Translation Workshop in China together from July 13th to 15th in 2005. This paper describes the tested systems of involved institutions and analyzes the results of their experiments. The test results show that although the research of statistical machine translation started late in China, it develops rapidly. The tested systems got quite good results in a short period. Compared with the rule-based systems reported in the formal "863" evaluation, the performance is somewhat lower, however, the difference is small. According to the state of art and the trend of international statistic machine translation research, we believe that there is still great space for the improvement of statistic machine translation, with larger scale data resources and more powerful hardware. In near future, phrase based method incorporated with syntax and semantic information will become the mainstream of statistical machine translation.

Key words: artificial intelligence; machine translation; statistical machine translation; phrase based translation mod

收稿日期: 2005-12-14 定稿日期: 2006-07-27
基金项目: 国家自然科学基金资助项目(60272041)
作者简介: 徐波(1966—), 男, 研究员, 博士生导师, 研究方向为语音识别, 机器翻译, 中文信息处理等.

1 引言

近十几年来,机器翻译由于巨大的市场需求和广阔的应用前景,受到了越来越多的重视。尤其随着互联网时代的来临,大量信息的涌现,一方面对机器翻译的实用性提出了更加迫切的需求,另一方面也对机器翻译提出了面向海量实际文本翻译的更高的要求。在这个背景下,机器翻译的研究掀起了新的浪潮。

在上个世纪 90年代以前,规则方法在机器翻译中占据统治地位。1990年由 IBM 公司的 P. Brown等人提出的基于信源信道思想的统计翻译模型[1],由于其数学推导严密、模型一致性好、可以自动学习、鲁棒性强等优点,越来越受到人们的重视。但是,最早的 IBM 统计翻译系统是基于词的翻译模型,只考虑了词与词之间的线性关系,没有考虑句子的结构及上下文信息,在两种语言的语序相差比较大时效果不好。随着研究者不断的努力,出现了基于短语的翻译模型[2,3],基本思想是从大规模语料库中抽取大量对齐短语片断,利用这些短语片断来匹配组合要翻译的句子。由于短语的限制使得词的翻译选择更为准确,而且有助于一些常用语及成语的翻译,并且相对确定了翻译的语序,使得翻译的结果更加符合目标语言的特征,因此,基于短语的统计机器翻译模型在近几年的统计机器翻译研究中占据了主导地位。在这种形势下国内的各个研究机构也对统计机器翻译产生了浓厚的兴趣,并将它作为一个重要的方向进行研究。为了尽快跟踪国际统计机器翻译研究的最新进展和趋势,进一步推动国内统计机器翻译研究的快速发展,中国科学院自动化研究所与计算技术研究所和厦门大学计算机系于 2005年 7月 13日至 15日在厦门大学联合举办了国内首届统计机器翻译研讨班。本文主要介绍本次研讨班参加单位的测试系统和实验结果,并给出相应的分析。

本文其他部分按如下结构组织:第二部分介绍本次研讨班的组织情况;第三部分分别介绍参加本次研讨班的几个统计翻译系统;第四部分给出测试结果和相应的分析;第五部分是本文的结束语。

2 研讨会准备

参加本次研讨班的中国科学院自动化研究所(CASIA)、计算技术研究所(ICT)和厦门大学计算机系(XMU)是近几年来国内从事统计机器翻译研究的三个单位。

自动化研究所从 1998年开始就进行了多方面的机器翻译的研究,其间完成了几个比较重要的实验系统,包括:IF(Interlingual exchange Format)[4,5];Word-based SMT[6];EBMT[7,8];Phrase-based SMT[9,10]。由于自动化研究所在语音识别技术上的丰富经验,因此,自动化研究所的统计机器翻译系统的解码部分融入了语音识别技术中解码的经验,并且在限定领域的应用中能与语音识别技术相结合,使其适用的范围更广。此次研讨会 CASIA 参加测试的是基于短语的汉英翻译系统(CASIA-PBT)。

中国科学院计算技术研究所长期从事自然语言处理与机器翻译研究,在此领域有着丰富的经验,计算技术研究所在汉语自动分词和标注、句法分析等很多方面都有着很好的基础,开发了一些成功的系统[11~13]。计算技术研究所在上世纪 90年代开始进行汉英机器翻译的研究[14~16],近年来开始转向统计机器翻译[17~20],是国内从事统计机器翻译研究较早的单位之一。这次研讨会他们在同一个翻译框架和短语词典的基础上提出了三个不同的系统:基于短语的翻译系统(ICT-PBT)[21],基于对齐模板和最大熵的翻译系统(AT)[22,23]、基于双语转换语

法 BTG(SBTG)的翻译系统[24]。

厦门大学计算机系在机器翻译领域有着非常资深的经历和丰富的经验,他们已经做了二十多年的基于规则的翻译方法研究,并且取得了很好的成绩。2004年开始他们进行基于统计的机器翻译方法研究。此次参加研讨会的也是两个基于短语的翻译系统(一个基于单调解码 XMU-MONO,一个基于对齐模板 XMU-AT)。

本次研讨会确定的翻译方向为汉语到英语的翻译。为了能对各个系统更好地做一个比较,这次研讨会统一了训练集和测试集。训练集有:自动化所提供的旅游领域的 5万句(记做 CASIA-5w),计算技术研究所提供的 863训练集 5万句(记做 863-5w),计算技术研究所提供的电影字幕 15万句(记做 ICT-15w)和厦门大学提供的电影字幕 20万句(XMU-20w)。测试集有:自动化研究所提供的 1000句(CASIA-1000),计算技术研究所提供的 863评测 2003年的测试集 350句(863-03)和 2004年的测试集 400句(863-04),以及厦门大学提供的 1500句(XMU-1500)。此次的评测全部是基于通用领域的,评测方法统一采用了计算所提供的评分工具。

3 测试系统简介

根据上述介绍,参加此次研讨会的共有六个系统:三个 PBT系统,两个 AT系统和一个 SBTG系统。这六个系统都利用了短语信息。其中三个 PBT系统是完全基于短语的翻译,没有加入其他的语法、语义知识,后两个系统也都不同程度地应用了短语信息,并且在短语的基础上加入了更多的知识。AT系统利用最大熵模型加入了模板等其他知识,SBTG系统采用了句法分析,基于自底向上的 CKY方法解码。下面我们就分别对这六个系统的特点做简要介绍。

3.1 三个基于短语的翻译系统

3.1.1 ICT-PBT系统

ICT-PBT系统与 Zens[21]的工作相仿,Zens提出:

$$Pr(f_1^J \mid e_1^I) = \sum_B Pr(f_1^J, B \mid e_1^I) = \sum_B Pr(B \mid e_1^I) \cdot Pr(f_1^J \mid B, e_1^I)$$
$$= \alpha(e_1^I) \cdot \sum_B Pr(\bar{f}_1^K \mid \bar{e}_1^K)$$

B 指某种划分,$\alpha(e_1^I)$ 指当前划分的概率,对应某个输入句子有多种划分,认为所有的划分发生概率相同。这个是最典型最简单的基于短语的翻译模型。

ICT-PBT系统在这个基础上对提取短语的方法进行了改进,并通过试验证明了短语的规模对结果影响很大。短语提取是在给出对位信息的双语训练集上进行提取。提取有两个限制条件:1. 短语由顺序连续的词组成;2. 根据对位信息提取。Zens的短语提取方法如图 1所示,ICT-PBT将它的 I J互换提出了反向的提取方法,并将两种方法结合提出了双向提取的方法。这样短语提取数量增加了 2~3倍,并且在 NIST2002年汉英机器翻译测试集上把 BLEU值从 0.1663提高到 0.1748。ICT-PBT系统没有对短语进行重排序。

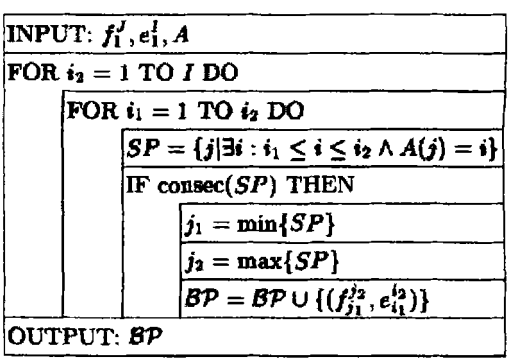

图 1 Zens的短语提取算法

3.1.2 XMU-PBT系统

XMU-MONO系统也是基于短语的翻译系统,参加了去年863评测,今年对它进行了一些改变。它利用了GIZA++[25]提取出来的对位信息。用Och[26]的方法双向提取词组。同ICT-PBT系统相同,都利用了CMU-Cam语言模型工具在BNC语料库上训练2-gram语言模型,应用动态规划的方法解码[27]。

XMU-MONO系统有几个特点:(1)允许单词的翻译为空(如汉语的助词),对翻译为空的单词进行与其翻译概率相关的惩罚。(2)采用fertility模型[29],但是,所不同的是,为了在一对一的单词翻译上增加简单的词组模型(因为只有两个以上的单词才称为词组),词组翻译模型不允许"放弃"翻译成"give up",而fertility模型却可以。(3)对汉语的数词、人名、地名、时间词做了简单的翻译处理。(4)对于个体量词,强制翻译为空。(5)解码算法:标记每个汉语单词的10个翻译;标记所有短语(最多10个),形成一个格;自左至右,采用动态规划寻找最短路径,不做词序调整。

3.1.3 CASIA-PBT系统

CASIA-PBT系统利用了GIZA++[25]提取出来的对位信息。用Och[26]的方法双向提取翻译词组。在小规模语料上训练了tri-gram语言模型,应用基于词组的堆栈搜索的方法解码。

CASIA-PBT系统在解码上有一定的优势,这里我们就主要介绍一下它的解码算法。由于单纯的一个词也可看作一个短语,所以基于词的翻译模型包含到了短语的翻译模型中。解码的过程可分为两步:首先从训练语料中提取短语,生成一个大的短语表,然后根据输入文本通过查表得出具体需要的短语对,有关的源语言短语信息、目标语言短语信息、短语翻译概率信息在解码真正开始之前存储起来,然后开始解码搜索。搜索过程从生成初始模型开始,初始模型或者是一个空的状态,或者对应于一个出现频率高且繁衍度为0的目标词(F-zerowords)的状态,这相当于在句子开头加了一个空单词NULL。从初始状态开始继续扩展,生成新的状态,如果新的状态对应一个F-zerowords则扩展到对应于一个未翻译的源语言短语及相应翻译候选[25]的状态,否则既可按前述方法扩展,又可扩展到一个对应于F-zerowords的状态。所有的状态根据已经翻译的输入文本中词的数目存于不同的堆栈中,一边扩展一边进行状态的裁减和合并,以满足空间和搜索速度的需要。直到输入文本中的所有词都被翻译了为止。

生成新状态的概率得分是当前短语翻译模型概率、语言模型概率、扭曲概率及长度模型概率的乘积。即:

$$p(e \mid c) = \lambda_t p_T(c \mid e) \times \lambda_t p_L(e) \times \lambda_d p_D(e, c) \times \lambda_w p_w(e)$$

$p_T(c|e)$是翻译模型,$p_L(e)$是语言模型,采用tri-gram语言模型,$P_w(e)$为句长模型,$p_D(e, c)$为扭曲模型,λ_t、λ_t、λ_d、λ_w分别为相应模型的参数,这里的几个λ都是用经验值得到的。

回溯时从最后的几个堆栈中找出的概率最大得分即为最优路径的最后一个状态。此时的概率得分需要考虑未来得分,计算方法如文献[25]中的方法。这样便可以回溯得出最后的翻译结果。

CASIA-PBT对于解码方法也进行了一些实验和改进。首先加入繁衍度为0的词,解决了很多中文词中冠词、介词、助词等虚词的翻译问题,在翻译模型、语言模型、扭曲模型的共同驱动下,在搜索时被补充到了最后输出结果中,实验也证明了加入这些词使得翻译结果更加合理。另外在回溯时不是仅在最后堆栈中找最优路径,而是按一定比例选取了几个堆栈中的候选最优路径。这是因为一般情况下汉语翻译成英语时,表达方式的不同使得文本中的某些词不需要翻译,因此,我们认为没必要强制翻译必须进行到最后一步。实验也证明了这样结果的

质量会有一定的提高。

3.1.4 各家 PBT系统特点的比较

PBT的系统在本次评测中取得了最好的成绩，并且是各家的主流系统，其中，ICT-PBT参加了 2005 NIST评测，XMU-MONO参加了去年 863评测，这次的系统改变了短语提取方式。

纵向来看，我们发现这三个系统有一个共同点，就是都利用了 Giza++提取出来的对位信息。除此之外，自动化所还研究了其他几种短语抽取方法：利用 IBM 模型 4直接提取，利用 ISA 方法提取[30]、HMM 对位信息提取[31]、Giza++工具提取等，而 Giza++提取的短语较多且质量较高。

同时它们也有很多不同点：

（1）利用对位信息进行短语抽取的方式不同，造成了短语的规模差别较大。在同样的 5万句上进行抽取，自动化研究所用的 Och双向抽取的方法抽出 24万对短语，计算技术研究所用 Zens单向抽取的方法抽出 35万对，厦门大学用 Och双向抽取的方法抽出 78万对。

（2）短语概率的计算方法各不相同。自动化所利用 IBM 训练的词对词的翻译概率，计算技术研究所利用抽取短语对的同现概率和单词的同现概率，厦门大学将抽取出来的短语和单词放在一起计算同现概率。其中 ICT及 XMU 对短语有很强的倾向性，自动化研究所的概率计算比较依赖词对词的翻译概率。

（3）搜索算法不同。自动化研究所用基于词组的堆栈搜索，这种方法可以进行跳序，但是速度相对较慢（8分钟 /1500句）。计算技术研究所用 Beam Search搜索算法，厦门大学用动态规划（DP）算法。这两种方法都是单调搜索，没有考虑词语的插入、删除和重新排序，速度很快（4秒 /1500句）。厦门大学的系统还采用了一些特殊的启发式策略（见 3.1.2）。

（4）所用的语言模型不同。自动化研究所用的是小规模语料训练的三元语言模型，计算技术研究所用的是 CMU-Cam 语言模型工具在 BNC 语料库上训练二元语言模型。厦门大学用的是 Srilm 语言模型工具在 BNC 语料库上训练二元 gram 语言模型。

3.2 ICT-AT系统

ICT-AT系统是计算技术研究所开发的基于对齐模板和最大熵的翻译系统，该系统首先利用 GIZA++训练句子对齐双语语料库，然后利用 Zens[21]描述的短语抽取算法抽取短语和对齐模板，利用 CMU-Cam 语言模型工具在 BNC语料库上训练 bi-gram 语言模型，最后再利用最大熵模型进行短语划分和翻译[22]，利用 Beam-Search单调搜索进行解码。

ICT-AT的系统有以下几个特点：1）目前大多数基于短语的统计机器翻译系统没有专门的模块做短语划分，一般认为一个句子的各种短语划分是等概率的，ICT-AT利用从语料库中抽取的短语和对齐模板，在词语切分的基础上对句子进行短语划分，将短语划分概率化。2）短语划分及翻译模型均利用最大熵模型。融合了丰富特征信息，并通过系数调整各个特征的比重。3）用 Och的方法进行词语聚类[28]，以此将抽取的双语短语泛化为对齐摸板，将对齐模板的选择和对齐模板的应用作为 2个特征加入到翻译模型中。其中对齐模板的选择指的是从一个中文模板对应的多个英文模板中选择一个合适的英文模板，比如：双语短语（发表声明-> issued a statement; statement issued; …），泛化以后对应的对齐模板是（2 3 -> 14 15 20; 20 14; …），其中阿拉伯数字代表类别。解码时对于短语（2 3）首先决定选择使用哪个英文模板，例如选择模板（14 15 20），从而使用对齐模板（2 3 -> 14 15 20）。其次是应用对齐模板，即对英文模板中的每个词类选择合适的译文单词，比如 14对应的单词有：published announced issued 等，选择 issued作为翻译候选词。利用词类对短语进行泛化拓展了短语的覆盖范围，但由于目

前采取的词语聚类方法正确率比较低，如果全部使用对齐模板效果比较差，因此，ICT-AT在解码时首先查找短语表，只有在短语表中没有匹配时，才去使用对齐模板。

3.3 基于 BTG的机器翻译系统

ICT-SBTG 系统是基于 BTG[24]的方法。基于短语的翻译系统比基于词的方法在选词和常用语的表述上更有优势，但是同样也存在问题，比如短语级上重排序问题。BTG 提供了一种很好的重排序策略，但仍然有值得改进的地方，ICT-SBTG 在重排序时，不仅要考虑目标语言语言模型的概率，还要考虑源语言中心词之间的约束关系，如图 2所示。其基本思路是：(1)源语言句法分析(Collins)和目标语言的生成同时进行；(2)采用自底向上的搜索策略(CKY-Based)，搜索过程中利用抽取出来的短语作为基本翻译单元；(3)在每个节点生成正序和倒序两种结果，其概率分别为 P_mlm * p_m 和 P_ilm * p_i 其中 P_mlm 和 P_ilm 分别是正序和倒序语言模型概率，P_m 和 P_i为基于源语言中心词约束的概率，用宾州树库(PennU Chinese Treebank)训练，并采用回退方法，以在一定程度上缓解数据稀疏问题。但是，我们从后面将要介绍的实验结果可以看出，由于这个方法是在严格意义的句法树上训练的，而分析过程中所采用的短语都不是句法意义上严格的短语，所以目前得到的结果并不比一般的基于短语的方法好。

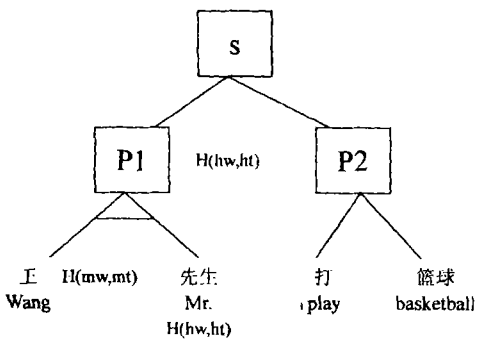

图 2 BTG 句法树

4 实验结果及分析

4.1 863测试集

测试集和训练集都是‘863中文信息处理与智能接口评测’提供的。

表 1 测试集 863-03 训练集 863-5w

结果	CASIA_PBT	ICT_PBT	ICT_AT	ICT_SBTG	XMU_AT	XMU_MONO	03BEST
Nist	6.078		6.6844	6.2962	5.5946	6.4168	
Bleu	0.222	0.2679	0.2696	0.2425	0.1514	0.2833	0.36

注：03BEST是 2003年 863评测的最好结果

表 2 测试集 863-04 训练集 863-5w

结果	CASIA_PBT	ICT_PBT	ICT_AT	ICT_SBTG	XMU_AT	XMU_MONO	04BEST
Nist	4.79		4.5544	4.5842	4.4631	3.8910	
Bleu	0.179	0.1225	0.1287	0.1191	0.1112	0.1105	0.21

注：04BEST是 2004年 863评测的最好结果

从表 1和 2可以看出，PBT系统的结果比另外两个系统要稍好一些，XMU-MONO 在 863-2003测试集上表现非常不错，这一点应该归功于它对于细节的处理比较好。而 CASIA-PBT在 863-2004测试集上的表现非常突出。我们分析了一下语料发现，2004年 863的测试集相对来说问句较多，长句较少，这就使得 CASIA-PBT系统的可以跳转的解码发挥了很大的作用，另外，CASIA-PBT采用了三元语法模型，比另外两个系统采用的二元语法模型有较大的优势。因此，结果比其他几个系统高出很多。

与当年 863评测的最好系统相比，这次参加评测的系统还有一定的差距。但这种差距并

不太大。而且考虑到这些系统都是利用现有的数据资源,在几个月的时间内开发出来的,应该说取得这样的结果是相当可观并令人鼓舞的。

4.2 字幕测试集

测试集是厦门大学提供的电影字幕语料 1500句,训练集是这次研讨会的四个训练集组合而成的四组。

表 3 测试集 XMU-1500;训练集:(1) ICT-15w;(2) XMU-20w;(3) ICT-15w+XMU-20w;(4) CASIA-5w+863 5w+ ICT-15w+XMU-20w

	15w	20w	35w	45w
CASIA_PBT	0.1	0.117	0.116	0.117
ICT_PBT	0.0816	0.1266	0.1274	0.1279
ICT_AT	0.0771	0.1213	0.1292	0.1298
ICT_SBTG	0.1265	0.0927	0.1275	0.1276
XMU_AT	0.0757	0.0775	0.0890	0.0931
XMU_MONO	0.0948	0.1416	0.1445	0.1459

从表 3的实验结果可以看出,在训练集变化的情况下部分系统表现得比较稳定,并且都是随着训练集的增加,评测分值稳步上升。两个加入了语法知识的系统 AT和 SBTG的结果和三个 PBT比较起来没有优势,虽然 SBTG方法对于某些特定的测试集和训练集会有比较好的效果,但是表现得不是很稳定。因此,从这一点考虑,语言知识如何融入统计模型中还需要进一步研究和探索。

5 总结

本次研讨会可以说是对国内统计机器翻译研究的初步检阅。总体来看,虽然国内各家单位起步较晚,但进步很快。各单位都已初步掌握了统计机器翻译的基本方法,搭建了完整的系统。在这次研讨会的现场测试中,取得的结果达到了跟往年 863评测最好结果可比较的水平。

我们也应该看到,在统计机器翻译方面,国内的研究可以说还处在起步阶段,相关的技术仅仅是初步掌握,技术的细节理解还不深刻,一些复杂的技术还没有尝试。这些都需要我们进行进一步的深入探索。

参加本次研讨班的各个统计机器翻译系统都有各自的特点和优势,但它们都有一个共同点,就是都利用了短语信息。通过 CASIA在 863-03、04测试集上所做的实验也可以看出,统计机器翻译正在逐步和规则的机器翻译缩小差距,在比较短的时间里就已经接近了往年的规则翻译的最好成绩,更多的研究单位也已经开始重视并着手于统计机器翻译的研究。随着研究的深入,更多的知识的融入也为人们所重视, ICT在这方面作了很多的工作,但结果提高不大,因此知识的融入方法还需要探讨和研究。可见在未来几年内,在基于短语的主流统计翻译方法中融入句法、语义等信息,必将成为机器翻译发展的趋势。

参 考 文 献:

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Pobert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics[J]., vol. 19, no. 2, 263-311.

[2] Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model[A]. In: Proceedings of the 39th Annual Meeting of the ACL[C], pages 523-530.

[3] Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, Alex Waibel. 2003. The CMU Statistical Machine Translation System[A]. In: proceedings of the Ninth Machine Translation Summit[C]. 402-409.

[4] Xie Guodong, Chengqing Zong and Bo Xu. 2002. Chinese Spoken Language Analyzing Based on Combination of Statistical and Rule Methods[A]. In: Proceedings of the International Conference of Spoken Language Processing (ICSLP'2002) [C]. Sept. 16-20, 2002, Colorado, USA. Pages 613-616.

[5] Wu Hua, Taiyi Huang, Chengqing Zong and Bo Xu. 2000. Chinese Generation in a Spoken Dialogue Translation System[A]. In: Proceedings of COLING, [C]. July 27 - August 4, 2000, Germany. Pages 1141-1145.

[6] Zhou Yu, Chengqing Zong and Bo Xu. 2005. Various Aligned Models In Chinese-to-English Statistical Machine Translation[A]. In: Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)[C]. October 30th-November 1st, 2005, Wuhan, China. Pages 443-448.

[7] Zong Chengqing, Yumi WAKITA, Bo Xu, Kenji Matsui and Zhenbiao Chen. 2000. Japanese to Chinese Spoken Language Translation Based on the Simple Expression[A]. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP)[C]. October 16-20, 2000, Beijing. Pages 418-421.

[8] 胡日勒. 2005. 口语翻译知识自动获取方法研究[D]. 博士学位论文, 中科院自动化研究所.

[9] Pang Wei, Zhendong Yang, Zhenbiao Chen, Wei Wei, Bo Xu and Chengqing Zong. 2005. The CASIA Phrase-based Machine Translation System[A]. In: Proc. IWSLT-05[C], Oct. 24-25, 2005. Pittsburgh, USA. 114-121.

[10] Xu Bo, Zhenbiao Chen, Wei Wei, Wei Pang and Zhendong Yang. 2005. Phrase-based Statistical Machine Translation for MANOS System[A]. In: Proc. MT Summit X[C]. Sept. 12-16, 2005, Phuket, Thailand. i23-i26.

[11] Hua-Ping ZHANG, Qun LIU, Hong-Kui YU, Xue-Qi CHENG, Shou BAI. Chinese Named Entity Recognition Using Role Model[J]. Computational Linguistics and Chinese Language Processing, Vol. 8, No. 2, August 2003, 29-60.

[12] 刘群, 张华平, 俞鸿魁, 程学旗. 基于层次隐马模型的汉语词法分析[J]. 计算机研究与发展, 2004. 8.

[13] Deyi XIONG, Shuanglong LI, Qun LIU, Shouxun LIN, and Yueliang QIAN. Parsing the Penn Chinese Treebank with Semantic Knowledge[A]. The Second International Joint Conference on Natural Language Processing (IJCNLP-05)[C], Jeju Island, Republic of Korea, October 11-13, 2005.

[14] 刘群, 詹卫东, 常宝宝, 刘颖. 一个汉英机器翻译系统的计算模型与语言模型[A]. 第三届全国智能接口与智能应用学术会议[C], 吴泉源, 钱跃良主编, 智能计算机接口与应用进展, 第 253-258页, 电子工业出版社, 1997. 8.

[15] Qun Liu, Shiwen Yu. TransEasy: A Chinese-English Translation System based on hybrid approach[A]. Third Conference of the Association for Machine Translation in the Americas (AMTA-98)[C], Langhorne, PA, USA, Oct. 1998. In: David Farwell et al. Eds., Machine Translation and the Information Soup, Lecture Notes in Artificial Intelligence Vol. 1529, Springer, 514-517, 1998.

[16] 刘群, 俞士汶. 汉英机器翻译的难点分析[A]. International Conference on Chinese Information Processing, 黄昌宁主编, 1998中文信息处理国际会议论文集[C], 第 507-514页, 清华大学出版社, 1998. 11.

[17] 刘群. 统计机器翻译综述[J]. 中文信息学报, 2003, 17(4): 1-12.

[18] 刘群. 汉英机器翻译若干关键技术研究[D], 北京大学博士论文, 2004. 5.

[19] 刘群. 基于微引擎流水线的机器翻译系统结构[J]. 计算机学报, 2004, 27(5): 1-12.

[20] Yang LIU, Qun LIU, and Shouxun LIN. Log-linear Models for Word Alignment[A]. The 43rd Annual Meeting of Association of Computational Linguistics (ACL-05)[C], Michigan, USA, June 25-30, 2005.

[21] Richard Zens, Franz Josef Och, Hermann Ney. September 2002. Phrase-Based Statistical Machine Translation[A]. In: Proc. German Conference on Artificial Intelligence (KI 2002)[C], Springer Verlag, 18-32.

[22] Franz Josef Och. 2002. Hermann Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation[A]. ACL2002[C], 295 – 302.
[23] Och Franz Josef. Statistical Machine Translation: From Single-Word Models to Alignment Templates[A]. Ph. D. thesis, Computer Science Department, RWTH[C]. Aachen, Germany, October, 2002.
[24] Wu Dekai. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora[A]. Computational Linguistics[C], 377 – 404.
[25] Koehn P., Och F. J., and Marcu D. 2003. Statistical Phrase Based Translation[A]. In: Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics[C]. 127 – 133.
[26] Och F. J., Tillmann C., Ney H. 1999. Improved alignment models for statistical machine translation[A]. Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora[C], University of Maryland, College Park, 20 – 28.
[27] http://svr-www.eng.cam.ac.uk/~prc14/toolkit_documentation.html.
[28] Franz Josef Och. 2002. Hermann Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation[A]. ACL2002[C], 295 – 302.
[29] Brown P. F., Della Pietra S. A., Della Pietra V. J., and Mercer R. L. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation"[A]. Computational Linguistics[C], 263 – 311.
[30] Ying Zhang, Stephan Vogel and Alex Waibel. Integrated Phrase Segmentation and Alignment Model for Statistical Machine Translation[A]. Submitted to Proc. of International Conference on Natural Language Processing and Knowledge Engineering(NLP-KE)[C], 2003.
[31] Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based Word Alignment in Statistical Translation[A]. In: COLING'96: The 16th Int. Conf. On Computational Linguistics[C], pp. 836 – 841, 1996. 19(6): 1 – 6.
[32] 胡日勒, 宗成庆, 徐波. 基于统计学习的机器翻译模板自动获取方法[J]. 中文信息学报, 2005, 19(6): 1 – 6.
[33] 张捷, 陈群秀. 日汉机器翻译系统中的 Agent研究[J]. 中文信息学报, 2003, 17(1): 7 – 12.
[34] 黄河燕, 陈肇雄, 宋继平. 一种人机互动的多策略机器翻译系统 IHSMTS的设计与实现原理[J]. 中文信息学报, 1999, 13(5): 43 – 50.
[35] 张剑, 吴际, 周明. 机器翻译评测的新进展[J]. 中文信息学报, 2003, 17(6): 1 – 8.

重 要 通 知

因宾馆大会场安排问题, 中国中文信息学会第六次全国会员代表大会暨成立二十五周年学术年会将于 2006年 11月 20 – 22日在北京中苑宾馆隆重召开。

这次盛会的重大活动如下:

(一) 2006年 11月 20日召开"中国中文信息学会第六次全国会员代表大会"。

(二) 2006年 11月 20日 – 22日举办一次高规格的"中文信息处理重大成果汇报展"。

(三) 2006年 11月 21日 – 22日举行"中国中文信息学会成立二十五周年学术年会"。

(四) 2006年 11月 20日隆重颁发"钱伟长中文信息处理科学技术奖"。

中国中文信息学会

2006年 8月 20日