N-gram 统计模型在机器翻译系统中的应用

张　健　李素建　刘　群
(中国科学院计算技术研究所,北京　100080)
E-mail:zhangjian@ict.ac.cn

摘　要　文章提出了 N-gram 模型在机器翻译系统中的几个应用。模型是在语料库的基础上统计连续几个词的出现概率,以此来筛选翻译过程中的侯选元素,并可以对译文的语序进行纠正。由于此种方法是建立在语料库的基础之上的,从而具有真实可靠和实时等特点。实验表明,这种方法具有良好的性能,且与被处理的语言无关。

关键词　自然语言处理　机器翻译　N-gram　语料库

文章编号　1002-8331-(2002)08-0073-03　**文献标识码**　A　**中图分类号**　TP391.2

Statistical N-gram Method Used in Machine Translation System

Zhang Jian　Li Sujian　Liu Qun
(P.O.Box 2704,Software Department,Beijing 100080)

Abstract: This paper introduces the statistical N-gram method,and used it in the Chinese-English Machine Translation System.Based on the parallel corpora,it performs statistics about consequent words.According to the statistic results,it can eliminate the wrong translations and rearrange the word sequence of the translation results.Experiments show that this method has good performance.

Keywords: Natural Language Processing,Machine Translation,N-gram,Corpus

1　引言

机器翻译(Machine Translation,MT)是利用计算机及其软件系统把一种源语言转换成为目标语言,并使得二者具有相同的意义和作用。传统的机器翻译系统大多数以规则为基础,在早期取得了显著的效果。但是对基于规则的系统来讲,需要将专家的领域知识融入到各种各样的规则中,并且这种方法随着规则库的不断增大而显得低效。例如,随着规则库的不断膨胀,可能其中的某些规则之间会发生直接或间接的冲突,导致整个系统的性能下降。这主要是由于自然语言的复杂性和其中存在着的许多特殊用法所造成的。另外,专家精心设计的规则也不是十全十美的,经常要根据具体的翻译情况来调整。这种调整工作是非常耗时的。

随着 Internet 的迅速发展,可以获得的语料越来越多,机器翻译的研究与大规模的语料库的结合也更加紧密。由于语料库具有信息量大、领域广、实时性强以及真实等特点,它可以作为一种很有价值的机器翻译系统的资源。如何更有效地利用这种资源是当前机器翻译的一个研究方向。

文章主要阐述了如何利用 N-gram 统计模型来提高翻译系统的性能。第二、三部分以一个具体的汉英翻译系统为基础,介绍了翻译系统的基本结构,并以一个具体的实例阐述了其翻译流程。第四部分详细地介绍了 Tri-gram 统计模型及其对机器翻译系统的改进,最后指出了今后的工作方向。

2　翻译系统的基本结构

(1)词法分析:词法分析就是分析输入句子中的各个单词的形态变化,分解出单词和语素,识别和判断单词的各种形态信息。由于汉语的特殊性,如词和词之间无间隔、人名和地名无特殊的标记,无形态的变化,使得对汉语的分析比其他语种要困难一些。对于汉语的词法分析主要有两个任务:分词和词性标注。对于英语来说,词和词之间有空格分开;而汉语的词与词之间难以区分,所以要对汉语进行分析,首先要对汉语句子进行自动分词。汉语的自动分词通常采用基于词典的最大匹配方法,要解决的主要问题就是未定义词问题(如人名、地名和组织机构名称)和切分歧义问题。对于英语来讲,人名和机构名称首字母大写,可以很容易地识别;而汉语中的未定义词没有明显的标记。对于未定义词可以通过分析句子的语法成分和研究其组成规律来正确的标识,也可以通过从对齐的双语语料库提取[3]。由于词法分析是机器翻译的前端处理步骤,此处的很小的错误可能会导致后面非常差的性能,所以我们一般采用非确定的算法,即保留各个可能正确的切分标注结果,尽量的保证正确的切分标注结果在侯选集合中,然后通过统计的方法来对其评估。

(2)语法分析:基于规则的语法分析主要是以 Chomsky 的句法理论为基础。按照 Chomsky 的理论,形式文法分为分别为 0 型文法、1 型文法、2 型文法和 3 型文法。其中 2 型文法又叫上下文无关文法(Context Free Grammar),因其形式简单且对语言具有较强的解释能力,在计算语言学语法分析等领域得到广泛应用。事实上,自然语言是上下文有关文法,通常采用在规则中加入一些上下文的限制条件来扩充上下文无关文法,从而达到上下文有关文法的功能。语法分析主要完成两个主要任务:确定输入的结构和语法结构的规范化。

(3)语义分析:语义分析就是要在自然语言和一种无歧义

作者简介:张健,男,硕士研究生,导师为白硕研究员,当前研究领域为自然语言处理和机器翻译。李素建,女,博士研究生,当前研究领域为自然语言处理和计算语言学。刘群,男,副研究员,当前研究领域为机器翻译。

内部语义表达之间建立一种映射关系,因此建立一种无歧义的语义表达系统就显得十分重要。传统的语义分析采用语义网络和逻辑表示两种知识表示方式。逻辑表示最大的优点是容易用于推理,而语义网络则较为直观,用节点表示短语和词,节点之间的弧表示语义关系。该文所讨论的机器翻译系统是采用基于合一算法的语义分析,语义分析嵌入在语法分析过程中,通过合一算法来描述节点之间的语义关系,同时也排除了很多符合规则但不符合语义的分支。语法分析和语义分析两部分对应于图 1 中的结构分析。

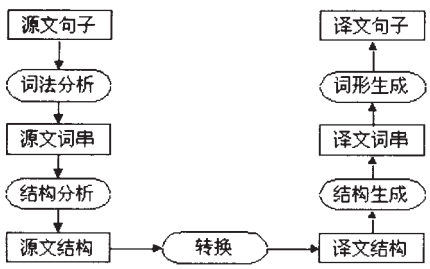

图 1

(4)转换语法树:利用规则库中的转换规则来将源语言的语法树转换成目标语言的语法树。

(5)转换阶段采用自顶向下与自底向上相结合的局部子树变换算法。在转换的过程中,可能会出现源语言中的某些语言现象在目标语言中不存在的情况(例如下面的实例)。这时应根据目标语言的习惯来处理。

(6)结构生成:根据目标语言的生成规则和特征将目标语言的语法树转换成为目标语言的词串。这时得到的词串已经具有了译文的雏形。结构生成阶段采用自底向上的局部子树变换算法和自顶向下的全局子树位移算法。结构生成根据转换得到的树结构完成英语词串的生成,这一阶段的工作主要根据造句规则来完成英语句子结构的重构和生成。

(7)词形生成:根据从语义分析中得到的有关句子中的时态、语态、单复数等等信息,修正目标语言的词串,从而得到译文。词形生成阶段也采用同结构生成同样的方法,这一阶段的工作主要根据词形生成规则来完成单词形态的生成。

3 翻译流程

下面以一个具体的实例来说明翻译流程:

(1)输入句子:统计学在机器翻译中得到了广泛的应用。

(2)词法分析:

分词后结果:统计学 在 机器翻译 中 得到 了 广泛 的 应用。

词性标注:统计学/n 在/p 机器翻译/n 中/f 得到/v 了/u 广泛/a 的/u 应用/v 。/(n)

其中:n-名词 p-介词 f-方位词 v-动词 u-助词 a-形容词

(3)结构分析:句法分析是利用规则库中的有关源语言的规则进行分析,并生成源文树。生成的语法树见图 3。图 2 是该例子中使用的一组简单的规则。

s	→	*np*	*vp*
np	→	*n*	
np	→	*ap*	*vp*
vp	→	*pp*	*vp*
vp	→	*vp*	*np*
vp	→	*vp*	*u*
vp	→	*v*	
ap	→	*a*	
ap	→	*ap*	*u*
pp	→	*p*	*sp*
sp	→	*np*	*f*

图 2

(4)转换生成树:利用规则库中的转换规则把生成的源文树转换成为译文树。转换后的语法树见图 4。

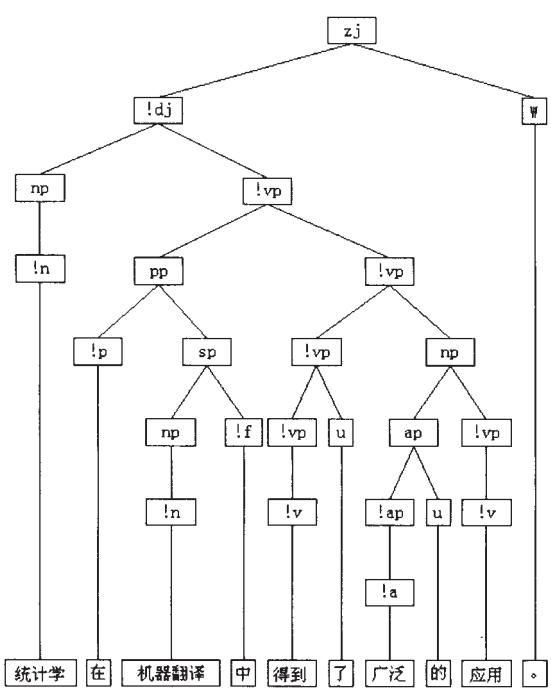

图 3

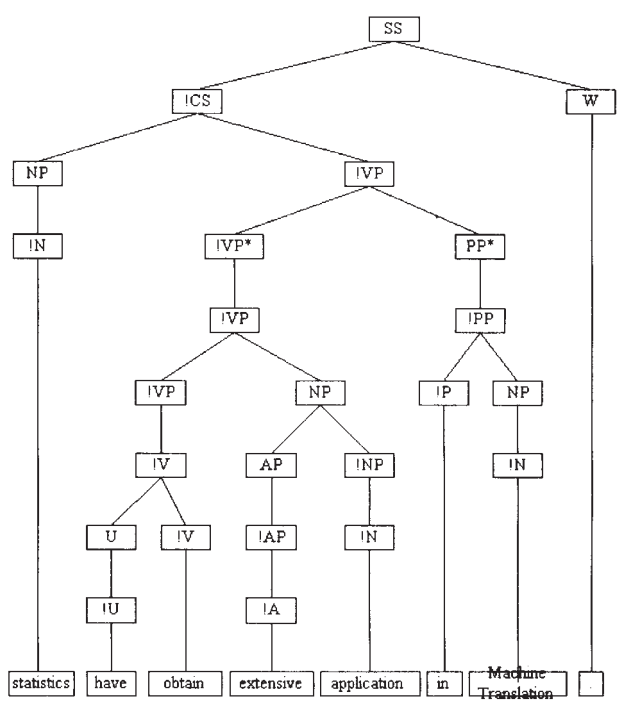

图 4

图中的"!"表示中心词,"*"是由于某些源语言成分在目标语言中无对应导致的冗余。

(5)生成目标语言词串:由译文结构树转化成词串。完成这一部分的功能要使用规则库中的造句规则。

(6)词形生成:根据构词法、语气、时态、人称、单复数和形容词(副词)的比较级和最高级来形成最后的输出。在本例子中即为将转换树中的 have 和 obtain 分别转化为 has 和 obtained。

4 改进

4.1 语料库的建设

随着 Internet 的迅速发展,越来越多的新闻机构、公司和企业将他们的信息发布在互连网上。由于这部分信息是经常被更新的,大家可以从中获得实时的、各个领域的文本资源。依靠人工建立语料库是一项庞大的系统工程,并且存在主观因素的影响。目前一般采用的方法是自动地从各种资源中获取语料库,通过建立起来的语料库,可以对大规模真实语料进行调查并总结自然语言的各种事实和语法规律。但是,仅仅这些资源是不够的,如何组织和索引这些语料库也是要建立大规模的语料库的一个重要的因素。语料库按照语言可以分成源语言语料库、目标语言语料库和双语语料库。相对而言,双语语料库使用价值比较大,但是也比较难获得。一般的获得的双语语料库是篇章对齐的,可以将其自动的进行句子一级的对齐。

4.2 N-gram 模型及应用

4.2.1 N-gram 模型是 N 阶 Markov 过程

一列随机变量 $S_1,S_2,\cdots,Sm$ 中,如果其中任何一个随机变量 S_i 发生的概率只与其前面的 n 个变量 $S_{i-1},S_{i-2},\cdots,S_{i-n}$ 有关,即

$$P(S_i|S_{i-n}S_{i-n+1}\cdots S_{i-1}S_{i-1})=P(S_i|S_1S_2\cdots S_{i-2}S_{i-1})$$

则称其为 n 阶 Markov 过程。N-gram 模型是把所有的连续可重叠的 n 个词作为一个单元(见表 1),并假设其为一个 N 阶 Markov 过程。对于一个由 m 个词 $w_1w_2w_3\cdots w_{m-1}w_m$ 组成的语句 S,这里定义

$$P(S)=P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\cdots P(w_m|w_1w_2w_3\cdots w_m)$$

作为一个语句的出现概率。由前面的假设可得:

$$P(S)=\prod_{i=1}^{m}P(w_i|w_1w_2\cdots w_{i-1})$$

$$\approx P(w_1w_2\cdots w_{N-1})\prod_{i=N}^{m}P(w_i|w_{i-N+1}w_{i-N+2}\cdots w_{i-1})$$

式中 $P(w_i|w_{i-N+1}w_{i-N+2}\cdots w_{i-1})$ 的是训练集合中 N 词串 $w_{i-N+1}w_{i-N+2}\cdots w_{i-1}w_i$ 占所有的以 $w_{i-N+1}w_{i-N+2}\cdots w_{i-1}$ 打头的 N 词串的比率。

表 1

语句	统计学在机器翻译中得到了广泛的应用
分词后的语句	统计学 在 机器翻译 中 得到 了 广泛 的 应用
Uni-gram	统计学) 在) 机器翻译) 中) 得到) 了) 广泛) 的) 应用)
Bi-gram	统计学 在) 在 机器翻译) 机器翻译 中) 中 得到) 得到 了) 了 广泛) 广泛 的) 的 应用)
Tri-gram	统计学 在 机器翻译) 在 机器翻译 中) 机器翻译 中 得到) 中 得到 了) 得到 了 广泛) 了 广泛 的) 广泛 的 应用)

在应用中一般采用 Tri-gram,因为对 $N>3$ 的 N-gram 模型需要非常庞大的训练集,而且需要很强的计算能力和存储能力。事实上,取连续的 3 个词作为一个单元对于大多数的应用已经够用了。当某一 Tri-gram 的在训练集合中不存在的时候($P=0$),这里可以用 Bi-gram 和 Uni-gram 来代替。

由于汉语中一级和二级汉字共 6 千多个,英语的常用词汇近万个,统计所有的 Tri-gram 是没有必要的。可以对其设置一个阈值,将语料库中出现频率大于此阈值的单元作为可用单元。定义 thresholdI 作为 I-gram 的阈值,即只统计大于等于阈值的 I-gram。

$$Uni\text{-}gram\text{-}Set=\{w_i|p(w_i)>threshold_1\}$$

$$Bi\text{-}gram\text{-}Set=\{w_{i-1}w_i|p(w_i|w_{i-1})>threshold_2\}$$

$$Tri\text{-}gram\text{-}Set=\{w_{i-2}w_{i-1}w_i|p(w_i|w_{i-2}w_{i-1})>threshold_3\}$$

4.2.2 N-gram-Set 的计算

(1)计算 1-gram-Set

For each word w in the Training-Set

If P(w)> threshold1

Then add w to 1-gram-Set

(2)计算 N-gram-Set

由前面可知,

$$P(w_i|w_{i-N+1}w_{i-N+2}\cdots w_{i-1})=\frac{C(w_{i-N+1}w_{i-N+2}\cdots w_{i-1}w_i)}{C(w_i|w_{i-N+1}w_{i-N+2}\cdots w_{i-2}w_{i-1})}$$

For each N-gram $w_1w_2\cdots w_{n-1}w_n$

If $C(w_1w_2\cdots w_{n-1}w_n)>threshold_N\times C(w_1w_2\cdots w_{n-2}w_{n-1})$

Then add N-gram $w_1w_2\cdots w_{n-1}w_n$ to N-gram-Set

关于如何更高效的计算 N-gram,请参考[3]。

4.2.3 筛选元素

由于目前大多数的机器翻译系统都是采用非确定的算法,既各个步骤的输出结果不唯一,目的是为了保证正确的结果在侯选集合中。一般的对侯选元素的评价方法是根据翻译系统和规则对各个结果赋予一个权值,根据权值来进行选择。如果这种权值能够准确的表明翻译结果的优劣的话,可以直接选择权值最大的作为输出的结果。事实上,这种权值只能反映一个大致的标准。判断一个翻译的结果的最可靠的方法是将此结果和真实的语料进行比较,从而确定其可靠的程度。根据前面的公式计算句子(短语)的概率,能从一定的程度上反映侯选集合中各个元素可能在真实的语境中出现的概率,从而可以增加翻译结果的可靠程度。

$$Candidate\text{-}Set=\{S_1,S_2\cdots S_{m-1},S_m\}$$

$$Score(S_i)=P(S_i)$$

4.2.4 纠正译文的语序

同时,由于源语言和目标语言的语序的差异,有时单纯通过转换规则会产生错误的语序。如果只是用对词的词性进行统计的 N-gram 模型(即此模型只是考虑词的词性,只要组成两个 N-gram 的各个词的是词性分别相同即认为两个 N-gram 相同)可以对译文的顺序进行矫正(见表 2)。

表 2

未经过纠正的顺序	纠正以后的顺序
名词+形容词+形容词	形容词+形容词+名词
名词+形容词+连词+形容词	形容词+连词+形容词+名词
not+情态动词	情态动词+not

5 结论

文章介绍了一个汉英机器翻译系统的组成,并以一个具体的实例介绍了其翻译流程,随后提出了 N-gram 模型和基于此模型的改进措施。由于这种模型建立在语料库的基础之上,它

(下转 78 页)

述边录入的特点，然后采用分词连写的思想[3]自己编制了一个分词程序，能够按作者的意图准确分词），其次确定它所用的表达句式为哪一类方式，然后分析其句式中各部分（词或词组）是否具有正确的逻辑关系，并计算句式中各核心成分、强化修饰成分、否定修饰成分、转义修饰成分等的权值，最后对照正确论述语句的权值完成判读操作。

5 基本算法描述

国家863中文自动摘要OA系统的研制者之一王永成教授在介绍他们的经验时曾指出，在计算机上研制与开发高级算法的捷径是“仿人，选突破口，先易后难，稳步迈进，坚持不懈”[7]。

为此，又结合该课题的情况，仔细分析阅卷老师在批改试卷时的具体过程。发现学生在答题时都有希望自己能够答对而得分的心理，同时列举题、简述题、简答题等该类问题本身的答题要点比较简明、清晰，各要点间的相互关系也较为明确，学生在作答时所用的语言一般均属简单论述文字的范围，因此老师在阅卷时一般都是边看答卷边从中扫描是否有该道题目所要考查的知识点的出现，并且各要点在表述上的相互关系是否合乎题目要求，出现了该知识点且相互关系无误则认为回答正确。

在清楚了老师阅卷这一过程后，将依照王永成教授提出的思想，用计算机对该过程进行模拟，模仿老师的阅卷过程，采用前述基于关系的带权匹配技术，为有限领域内简单论述的自动识别与判断确定如下基本算法（见图1）。

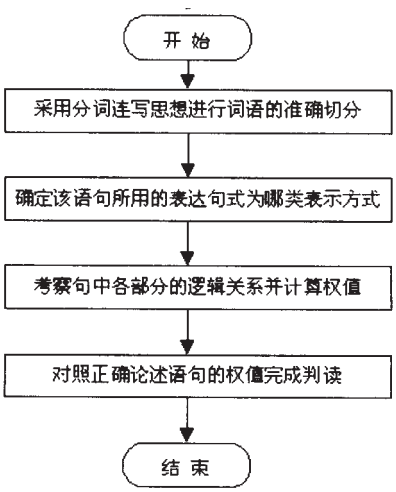

图1

6 整个实现系统的整体结构

在实现上，该系统的整体结构可分为三大部分，即句子中词语的切分、后台数据库管理以及在此基础上实现的简单论述的自动识别和判定。各部分的关系和系统流程如图2所示。

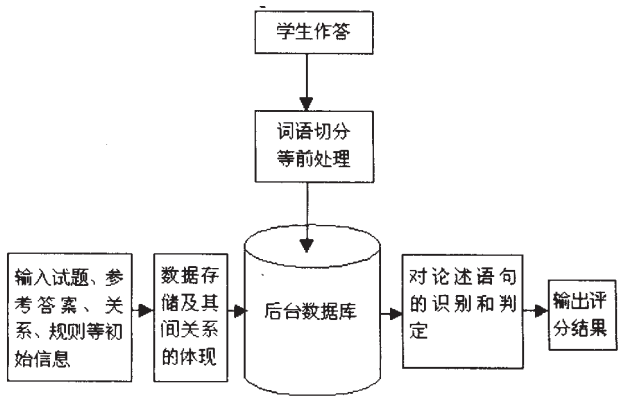

图2

由于前述算法的开放性，基于概算法的实现系统也具有较好的开放性，使该系统本身也可随着其它相关技术的发展而发展。

7 结束语

对于涉及中文信息处理方面的问题一般是人工智能中较为困难的问题，文章借鉴自然语言处理的一些思想，并结合CAI应用问题本身的具体特征，面向具体应用而有别于一般自然语言处理工作中的理论研究，提出了一个可以绕开自然语言语法、文法和语义分析，对CAI中简单论述进行自动识别和判定的方法，可以基本解决CAI在实际应用中遇到的对给定概念及所作论述的正误进行判别的问题，使CAI软件在学习效果的检测上具备相应的能力。由于该方法的开放性构建而成的是一个开放系统，可以通过其它一些技术，以及自然语言处理技术的新发展的引入，使它本身也在不断采用一些新技术新思想的过程中不断向前发展。（收稿日期：2001年7月）

参考文献

1.蒋焕文，王洪.计算机辅助教学CAI综述[J].高等教育研究，1994.1
2.陈郑汉，周经野.自然语言与机器学习[J].计算技术与自动化，2000.9；19（3）
3.陈力为.汉语书面语的分词问题[J].中文信息学报，1996.1
4.刘开瑛，郭炳炎.自然语言处理[M].科学出版社，1991
5.王开铸.自然语言理解[M].哈尔滨工业大学出版社，1996.4
6.郭艳华，周昌乐.自然语言理解研究综述[J].杭州电子工业学院学报，2000.2
7.王永成，许慧敏.OA-1.4版中文自动摘要系统[J].高技术通讯，1998.1

（止接75页）

具有实时、客观、符合语言习惯等优点。但由于该文的N-gram是以词为最基本的单位，所以这种模型仍然不能摆脱分词的困扰。今后笔者将进一步挖掘语料库中的可用信息，使机器翻译的结果更符合人们的表达习惯，具有更好的可理解性。

（收稿日期：2001年3月）

参考文献

1.Ralf Brown，Robert Frederking.Applying Statistical English Language Modeling to Symbolic Machine Translation.Carnegie Mellon University
2.LiuYing，LiuQun，ZhangXiang et al.A Hybrid Approach to Chinese-English Machine Translation[C].IEEE ICIPS'97，1997
3.Wang Xiaolong，Daniel Yeung，Guan Yi.An Algorithm for Constructing Higher Order N-grams of Chinese Words[C].International Conference on Machine Translation & Computer Language Information Processing
4.王斌，刘群，张祥.汉英双语库词汇对齐研究[C].计算语言学文集，清华大学出版社，1999
5.刘颖，常宝宝，刘群等.语言工程——用概率方法处理汉英机器翻译系统中的歧义[M].清华大学出版社，1997