第 34 卷 第 11 期 计 算 机 工 程 2008 年 6 月
Vol.34 No.11 Computer Engineering June 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)11—0200—03 文献标识码：A 中图分类号：N945

基于语义的水情电报翻译系统

陈 遥[1]，朱跃龙[2]，冯 钧[2]
(1. 南京信息工程大学计算机与软件学院，南京 210044；2. 河海大学计算机与信息工程学院，南京 210098)

摘 要：介绍针对信息提取的机器翻译模型，分析该模型下基于语法分析的水情电报翻译系统及存在的问题。采用机器翻译中语义分析技术解决问题，利用语义信息提高翻译率。用逻辑语义建立水情电报的逻辑语义模型，结合语义信息实现对几种类型错报的翻译，由此提出基于语义的水情电报翻译模型。对系统的评价与分析结果表明，翻译率即自动化程度得到了提高，报文的不可翻译率从 2%~5%降到 0.96%左右。

关键词：水情电报翻译；逻辑语义；语义分析

Semantics-based Water Information Telegraph Translation System

CHEN Yao[1], ZHU Yue-long[2], FENG Jun[2]
(1. Computer and Software Institute, Nanjing University of Information Science & Technology, Nanjing 210044;
2. Computer and Information Engineering Institute, Hohai University, Nanjing 210098)

【Abstract】 This paper presents Machine Translation(MT) model for extracting information and its application and problems in water information telegraph translation, which is based on rules of water information telegraph. It introduces semantic analysis technique to solve the problems, which is a kind of language processing technique in MT, and provides semantic information to increase the rate of translation. By applying logic semantics, it constructs the logic semantic framework of telegraph to obtain the logic semantics of telegraphs. Semantics-based water information translation model is presented, and its system is evaluated. The result shows the degree of automation is improved, and the telegraphs unable to be translated descend from 2%~5% to about 0.96%.

【Key words】 water information telegraph translation; logic semantics; semantic analysis

1 概述

目前在水利系统中，虽然网络比较普及，但许多水情测站地处偏僻，通过网络传送水情的方式在很长时间内还无法实现，水情电报仍然是一种常用的传达水情信息的方式。水情报汛系统如图 1 所示，各地区的水情测站由水文观测员把现场观测到的本站水情信息(如水位、雨量)按照水文情报预报拍报办法[1]和规定译成电文(水情电报)，通过邮政电报方式拍发到其所属的水情信息接收中心。水情测站的一般类型为水文站、水位站、雨量站、气象站、水库站等，每种测站有各特点，根据不同的水情信息确定该站拍发的报文类型，如水位站拍发河道水情报。在水情信息接收中心，已接收的水情电报经水情电报翻译系统(简称译电系统)翻译，提取出报文中的水情信息，进而处理其他水情信息(如查询、检索)。

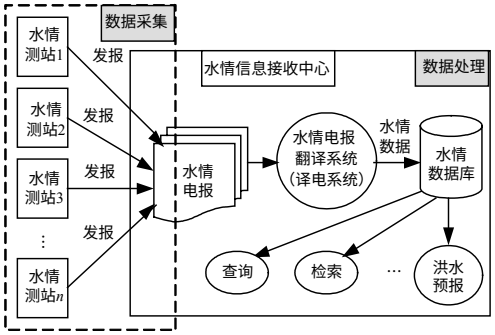

图 1 水情报汛系统

在水情报汛系统中，译电系统是水情数据自动采集与水情信息自动化处理之间的桥梁，因此，译电系统的自动化程度直接影响到水情信息的准确性、时效性。而高翻译率是人们追求的目标。

机器翻译(简称机译)是利用计算机把一种语言转变成另一种语言的过程[2]。机译主要有 4 种基本的翻译需求，其中之一是以信息提取为目的，借助计算机得出的结果传达信息的基本内容[3]，简称为针对信息提取的机器翻译。水情电报翻译是依据水情电报拍报办法，利用计算机将水情电报转变成所需水情数据的过程，因此，可以将它视为一种针对信息提取的机器翻译，将其中按一定规则编码后发送的代码串原文(水情电报)看作一种语言。这样水情电报所包含的水情信息即语义，水情电报规则即语法。水情电报由报头与报体构成，报头为一个字符串，报体由多个 5 位字符串(简称五位码)按水情电报规则组成，报头与报体、各五位码之间都以空格分隔，因此，一条报文可看作一个句子，每个 5 位码可看作某一类单词。

与对语言的机译系统相比，针对信息提取的机译系统主要特点是：句子间相互独立，即句子之间没有联系[3]。根据这个特点，文献[3]提出了一个针对信息提取的机器翻译模型，该模型支持基于语法与语义规则进行信息提取的机器翻译。但由于语义规则的总结比语法规则的总结更为复杂，目

基金项目：南京信息工程大学科研基金资助项目(y617)
作者简介：陈 遥(1977－)，女，讲师、硕士，主研方向：计算机应用技术；朱跃龙，教授、博士；冯 钧，副教授、博士
收稿日期：2007-07-30 **E-mail：**chenyao0077@163.com

前基于该模型的水情电报翻译系统只能根据水情拍报办法(即语法规则)来翻译水情电报，是基于语法分析的水情电报翻译系统。据统计，它通常能翻译约 95%的报文，这些报文都是按电报规则拍发的，另外 4%~5%是报汛人员在拍报过程中因各种原因而拍出的一些不符合拍报办法的报文。目前的做法是将它们确定为错报，经人工更改后再翻译。这种半自动化的译电方式影响了信息的时效性，在汛期尤为突出。本文对语义分析在水情电报翻译中的应用进行了初步研究。

2 系统设计

在机器翻译系统中，语言技术(包括存储于计算机系统内的语言知识库以及语言知识的归纳、表示与运用)居于核心地位[4]。类似地，水情电报翻译模型由机器词典和语言规则库支持，语言规则库占有重要的地位。

2.1 针对信息提取的机器翻译模型

文献[3]提出的机器翻译模型如图 2 所示，其采用了基于规则的方法，由翻译模块(规则分析解释器、规则库和词典库)、规则库维护模块和词典维护模块组成。

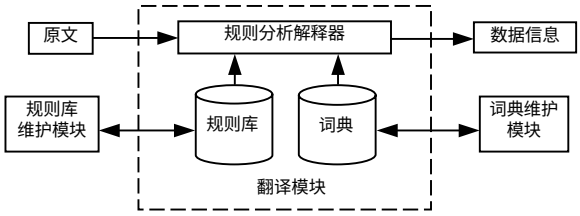

图 2 针对信息提取的机器翻译模型结构[3]

(1)规则分析解释器是系统的核心，按照一定的算法(如下推自动机)，通过规则库和词典的调用完成对句子句法和语义的分析，以提取数据信息。在分析句子的过程中，解释器将每个单词看作一个处理元素。

(2)因为采用了基于规则的方法，这里句子分析可以包括语法分析和语义分析，所以规则库可以包括语法规则与语义规则，为句子的翻译提供动态信息支持。

(3)词典为句子翻译提供静态信息支持，但与常规词典不同，单词本身还须分析和提取所含的数据信息，因此，在单词中还有很多操作函数。

(4)规则库维护模块可以对变化的规则进行修改。

(5)词典维护模块可以对词典库中各个单词的属性信息进行修改。

2.2 基于语义的水情电报翻译模型的算法思想

应用图 2 的机器翻译模型，在基于语法分析的水情电报翻译模型中为语言规则库增加语义规则(采用语义分析技术推出的规则，包括各类报文中水情信息组的语义规则)。

对基于语法分析的水情电报翻译模型中的语言解释器进行改建，在其对水情电报进行语法分析的同时，根据各水情组的语义规则推出报文各语法成分(五位码组)的语义并记录(可用链表等)，当报文中某水情信息组无法识别时，结合此报文的逻辑语义结构，通过已记录的报文语义链及其他一些语义信息(如某些特征水情信息组的语义信息)进行语义分析，即利用语义分析技术归纳推出相应的语义规则，对报文的水情信息组进行识别与调整，实现重新翻译。

2.3 基于语义的水情电报翻译系统总体设计

根据 2.2 节，基于语义的水情电报翻译系统的总体设计见图 3。该系统的解释器仍采用下推自动机，但下推栈中压入的是相应报文逻辑语义链中的语义项。规则库中则增加了各类报文的最大逻辑语义链及归纳出的语义规则，其译电类库也定义了相应的符号类。

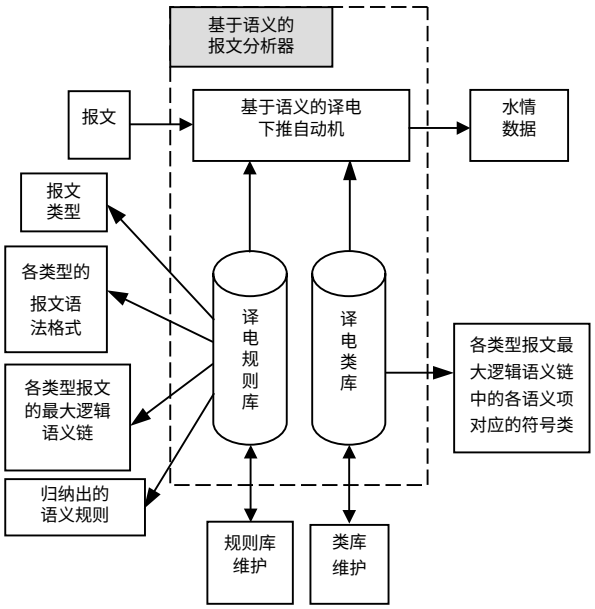

图 3 基于语义的水情电报翻译系统结构

将水情电报看作是一种基于拍报规则的语言L_h，则组成这一语言的有限字母表δ_h＝{{报头组}, {五位码组}, , (,)}，每条报文是该语言的一条句子，其中，{五位码组}={站号组, 时间组, …}。为此译电系统建立基于语义的水情电报翻译下推自动机，它的电报翻译规则库包含了图 3 中的语义规则，以实现对报文语义分析的支持。

定义 基于语义的水情电报翻译下推自动机是 1 个七元组：$A_p=(S, \delta, L, \delta, K, X_0, F)$，其中，

(1)$S=\{S_0\}$，控制器中的状态始终不变。

(2)δ={电报报体中所有的五位码字符串}。

(3)L={电报中所有的逻辑语义项符号，分为终结符和非终结符(语义项中的复合组)}。

(4)δ 表示转换函数的集合，具体转换函数如下(其中，a 为五位码字符串)：

1)$\delta(S_0, a,$ 非终结符$)=(S_0,$ 构成该非终结符语义项的非终结符和终结符语义项)

当下推栈的栈顶为非终结符时，该非终结符出栈，同时将构成该非终结符的非终结符和终结符入栈，读头不移动。例如，$\delta(S_0, a,$ 雨量组$)=(S_0,$ 时段雨量组)。

2)$\delta(S_0, a,$ 终结符$)=(S_0, \varepsilon)$

当下推栈的栈顶为终结符，且该终结符出栈时，判断读头所指五位码 a 是否符合该终结符的语义条件，如果符合，按此终结符的含义进行翻译，读头右移，同时将此终结符(语义项)记录到此报文的语义链中；如果不符合，则读头不动。这里的语义条件是根据其 5 位码的特征及其周围五位码的特征与语义进行判断的。

3)下推栈开始时的转换函数如下(其中，*station* 为代表站号的五位码)：

$\delta(S_0,$ “*station*”, 报体$)=(S_0,$ 由*station*确定该报文的逻辑语义链)，读头不动；

$\delta(S_0,$ “*station*”, 站号组$)=(S_0,$ 时间组)，读头右移。

(5)$K=\{S_0\}$。

(6)F={电报中所有逻辑语义链中的语义项终结符}。

3 关键技术

综合比较各种语义分析方法如格语法、逻辑语义学及优选语义学，根据水情电报的特点，采用逻辑语义水情电报进行语义分析，并用一阶谓词逻辑描述水情电报语义分析。

3.1 逻辑语义

逻辑语义[6]是一种语义关系。在特定的交际环境下，某一语言片断(通常指句子)的各个基本单元之间必然存在某种逻辑关系，这些逻辑关系称为逻辑语义，其集合称为逻辑语义结构。

一条报文由一些包含水情信息的五位码(水情信息组)构成，传达一定的水情信息(语义)。而且一条报文的各个水情信息组之间必然存在着某种逻辑关系，这些逻辑关系则称为报文的逻辑语义，其集合称为报文的逻辑语义结构。例如有报文：55994 29080 00069 00559，则此条报文逻辑语义为

55994：站号组，即施事，用于确定报文类型；

29080：时间组，即时间，用于观测时间；

00069 00559：雨量水情组，即状态；

00069：时段雨量组，末位为 7、8 或 9(天气状况的标志)；

00559：日雨量组，末位为 7、8 或 9(天气状况的标志)。

该报文的逻辑语义结构为：站号组(雨量报)-时间组-雨量组，即：施事-时间-状态。

在此报文中，站号组 55994 决定了该报文类型——雨量报，相应地确定了其后水情信息组的语义为：时间组和 2 组降水量组，同时由这三者之间的逻辑关系确定了后 2 组降水量组分别为时段雨量组和日雨量组。

其他报文与之类似。在报文的逻辑语义结构中，“站号组”是核心，站号组决定了报文的类型，报文类型决定了站号组后面各水情信息组的语义，即逻辑义项，从而确定了报文的逻辑语义模型。报文的最大逻辑语义链是与一种报文类型(站号组)相关的所有逻辑语义项的顺序排列。

3.2 一阶逻辑谓词

用自然语言描述水情电报语义知识简单方便，但其致命弱点是存在歧义性，不易于计算机处理。目前每种知识表示方法都适用于表示某一特定领域的知识。一阶谓词演算系统有相当强的形式表述能力，用它可以将极其广泛的一类自然语言语句及所有数学命题形式化。因此，本文采用一阶谓词逻辑定义语义谓词，如：

(1)YLB(Station)：表示站号 5 位码 Station 发出的报文为雨量报；

(2)(ddddd)(x)：表示 5 位码为数字串，并且两边用括号括起来。

并用一阶谓词表示语义规则，如：非降水量组

$$(x) \leftarrow Xdddd(x) \vee (ddddd)(x) \vee dddd1(x) \vee dddd2(x) \vee dddd3(x) \vee dddd4(x) \vee dddd5(x) \vee dddd6(x)$$

其中，“←”表示蕴涵。该规则表示出报文中哪些组为非降水量组。

3.3 语义规则的构建

根据抽象出的水情电报的逻辑语义核及层次结构确定水情电报的几个最大逻辑语义链，并以此为基础建立了一套水情电报的逻辑语义模型。利用逻辑语义模型可以推出各报文的逻辑语义(结构)，再结合报文中其他语义信息可以确定报文的逻辑语义结构，进而识别出一些非规则报文中水情信息组的语义。

对一条报文中同类型的层中水情信息组的语义分析是相同的，因为它们的逻辑义项的构成与排列(即最大逻辑语义链)是一样的。因此对于一条报文，先分析出其中不同的层，然后对不同的层选用其相应的逻辑语义结构，分别进行语义分析，从而完成对报文的分析。

通过对各种类型报文的分析，本文提取出一个通用的逻辑语义框架：

施事-时间-状态{-时间-状态}[-站号间隔标志 99099-施事-时间-状态]

其中所用符号的约定如下：

(1){}中的义项组表示此义项组在此处没有或者有连续几组；

(2)[ ]中的义项组表示此义项组在此处没有或者有一组；

(3)带下划浪线的组表示复合组。

本文通过对这些错报进行研究，组合出 4 种类型的错误。针对这些错误类型，利用提供的逻辑语义模型，结合水情电报规则的语法图，找出报文中尽可能多的语义信息，然后在现有语义信息的指导下，依据其逻辑语义与电报规则，调整并纠正错报，从而实现对这些类型错报的翻译，即归纳推出相应的语义规则，并用一阶逻辑谓词表示。

4 系统评价与分析

为了对本文的翻译系统进行评价与分析，抽取太湖流域 2002 年 8 月份的水情电报进行抽样分析。这个月共收到 4 497 条报文，包含各种主要类型的报文，其中，雨量报有 359 条；河道水情报有 1 800 条；闸坝水情报有 899 条；水库水情报有 1 439 条。一般情况下，太湖流域的水情电报的平均错报率为 2%~5%。经本文的译电系统翻译后，错报率为 4.2%，其中，类型 1 可译的错报占 0.01%；类型 2 可译的错报占 1.68%；类型 3 可译的错报占 1.56%；类型 4 可译的错报占 0.01%。这几种类型的错报可用基于语义的水情电报翻译系统自动翻译，不可翻译率降到了 0.96%。

5 结束语

本文在基于语法分析的水情电报翻译模型的基础上，采用语义分析技术，提出了基于语义的水情电报翻译模型。本模型结合语义信息实现了对某些类型非规则报文的翻译，并降低了错报率。但还有一些非规则报文不能处理，因此，需要结合其他机器翻译技术，进一步提高系统翻译率，以更好地满足水利系统高翻译率的需求。

参考文献

[1] 水利电力部. 水文情报预报拍报办法[Z]. 1964.

[2] 董振东. 机器翻译漫谈[Z]. (2000-07-06). http://tech.sina.com.cn/soft/2000-07-06/480.html.

[3] 朱跃龙, 王 勇, 濮淼清, 等. 面向对象技术的机器翻译系统的实现与设计[J]. 计算机工程, 2001, 27(11): 47-49.

[4] 朱学锋, 俞士汶. 自然语言处理与语言知识库[M]//计算机时代的汉语和汉字研究. 北京: 清华大学出版社, 1996: 107-118.

[5] 冯志伟. 自然语言机器翻译新论[M]. 北京: 语文出版社, 1995.

[6] 董振东. 逻辑语义及其在机译中的应用[J]. 情报学报, 1980, (2): 48-60.