多策略汉语句子时间分析与英译

关晓薇
GUAN Xiao-wei

大连理工大学 计算机科学与工程系,辽宁 大连 116023
Department of Computer Science and Engineering,Dalian University of Technology,Dalian,Liaoning 116023,China
E-mail:angels_gxw@yahoo.com.cn

GUAN Xiao-wei.Temporal analysis and English translation of Chinese sentences based on multiple strategies.Computer Engineering and Applications,2008,44(5):22-24.

Abstract: This paper introduces a new Chinese temporal information classification and the concept of temporal pattern.The temporal information of Chinese sentences is formalized based on temporal pattern,and the bases of lexical-information temporal pattern and grammatical-information temporal pattern are constructed.An new temporal analysis and English translation algorithm of Chinese sentences based on multiple strategies is presented,which consists of temporal analysis algorithm of Chinese simple sentence,conjunction-marked sentence,analogous subjunctive mood sentence and context rules.The experimental results show this method has a good effect on resolving temporal analysis and English translation of Chinese sentences in MT.
Key words: temporal patterns;multiple strategies;temporal analysis algorithm;constant;parameter

摘 要:提出了汉语时间信息的新分类和时间模式的概念,基于时间模式对汉语句子的时间信息进行形式化,构建汉语句子的词汇信息和语法信息时间模式库;提出多策略汉语句子时间分析和英译方法,将汉语单句时间分析算法、汉语关联词语标记句时间分析算法、类虚拟语气句时间分析算法和篇章信息识别规则相结合。实验表明该方法能有效解决汉英机器翻译中汉语句子时间分析和英译问题。
关键词:时间模式;多策略;时间分析算法;实量;变量
文章编号:1002-8331(2008)05-0022-03 **文献标识码:**A **中图分类号:**TP391

1 引言

英语借助动词时态变化表达时间信息,而汉语借助词汇和语法手段来表达。在汉英机器翻译中,我们只能以所表达时间信息为桥梁建立两种语言之间时间表达上的转换。

在机器翻译领域,马红妹等[1]提出基于汉语上下文语境模型处理汉语时间引用现象,该方法对处理不易获得写作时间的汉语篇章具有局限性。程节华等[2]提出汉语句子时态和体态抽取分析算法,归纳了汉英时态转换规则,但该方法仅在单句中进行分析,未考虑复合句和上下文等语境信息,规则过于简单。Li W.J 等[3]结合机器学习方法和语言学知识,提出分析汉语多分句句子中分句时间关系的计算模型。Wong S.M 等[4]建立汉语篇章时间引用模型,并用图像方法表示结果。林达真等[5]通过计算句子中每个词对时态确定所作贡献值,判断句子时态类别,该方法引进了对解决时态无关的特征词分析,未考虑上下文信息,对时态词出现较少或未出现的句子处理结果不理想。

根据汉英机器翻译的特点,论文在第二部分提出了汉语句子时间信息的新分类和时间模式概念,并对汉语句子时间信息进行形式化;第三部分介绍汉语句子词汇信息和语法信息时间模式库构建,并提出了多策略汉语句子时间分析和英译方法,将汉语单句时间分析算法、汉语关联词语标记句时间分析算法、类虚拟语气句时间分析算法和篇章信息识别规则相结合。

2 汉语句子时间信息形式化

2.1 汉语句子时间信息的新分类

西方理论语言学、形式语义学的许多研究将自然语言对时间信息表达分为 3 个方面:时相、时制和时体。一般认为汉语中存在 9 种时制类型和 6 种时体类型。语言学上这种对时间信息传统分类主要是一种逻辑上和理论上的分类,存在类别交叉、类别不宜分辨、分类复杂化等问题,不完全适用于汉英机器翻译中时间翻译的需要。而英语中存在 16 种时态类型,分类比较清晰,且受广泛的认可。由于汉英句子可以互译,即给定一个具有某一时间特征的汉语句子,一定可以将其直译或意译成(或通过其他翻译手段)一个具有某一时态的英语句子。

为了简化问题分析的难度,本文利用从英语的 16 种时态向汉语时间表达的逆向映射关系,从大量汉英双语语料中总结英语 16 种时态所分别对应的汉语时间表达形式,包括:词汇形

基金项目:国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2006AA012140)。
作者简介:关晓薇(1979-),博士生,主要研究领域为:自然语言处理,机器翻译。
收稿日期:2007-07-20 **修回日期:**2007-11-19

式,语法形式,篇章形式等。再反过来根据总结得到的各种汉英时间表达的对应形式来帮助处理汉英机器翻译中的时间转换问题。这样,不直接从汉语时间表达向英语时态表达映射,不必过多地涉及汉语中比较复杂的时相、时制和时体的分类,使问题简化,并且不易忽略对一些无时间特征词的汉语句子以及汉语特殊句式表达的研究。

2.1.1 词汇信息

(1)时间短语包括表时数量结构、时间名词、表时连词、表时方位词、表时介词,历史事件名、节日名以及事件。

将时间短语分为三类:简单时间短语(可译成一个单词或固定词组),复合时间短语(由两个或两个以上简单时间短语复合而成),复杂时间短语(由表时间的连词、介词、方位词与简单时间短语或复合时间短语组合而成)。每一类又可分为能确定时间和不能确定时间两类。

(2)时间副词和时间助词:汉语动词常借助助词和副词来表达时间信息,而且助词和副词的句法位置也是决定时间信息的一个重要因素。因此论文专门针对汉语中主要的助词和副词(虚词),重点研究这些虚词单独或复合出现在汉语句子中所体现出来的时态规律。

(3)否定词和情态助动词。

(4)谓语动词情态:对于一部分汉语句子中未出现时间信息词的句子,根据主谓语动词的情态特点,可以确定时态。

2.1.2 语法信息

(1)特定句式,包括连动句,紧缩条件句,强调句,可译成英语表语从句、主语从句、定语从句、宾语从句的汉语句子,可译成英语时间状语从句的汉语单句,形容词谓语句,固定表达句。

(2)关联词语标记句。

(3)类虚拟语气句(可转换成英语虚拟语气句的汉语句子)

2.1.3 篇章信息

包括文章体裁信息、上下文语境信息和句子主题信息等。每种体裁都具有各自特有的时间基调、时间发展线索以及语言表达形式。主要研究汉语各种文体的时间组织规律,上下文语境和该句的主题对该句时间表达的影响等。

2.2 相关定义

定义 1(变量类型) 包括具有某一语义特征的类型或语法类型。用 N,A,V,S 分别表示类型为名词、形容词、动词和句子。

定义 2(实量) 指模式中固定词语部分。在具体自然语言中可用具体词语表示。

定义 3(变量) 指模式中可以被相同变量类型词语替代的词语部分。在具体自然语言中用变量类型符号表示。

定义 4(时间模式,TP) 指针对句子或短语的时态规律,将句子或短语提取成能够反映其时态特征的形式化表达形式。由实量和变量按照一定的顺序相互插入而成,并且能够进行变量代入的翻译模板。例如:时间模式"$N_{人}$正在 VN",其中的"$N_{人}$、V、N"是三个变量,"正在"是实量。

定义 5 (模式类型,PT) 指整个时间模式所属的类型。例如:"Nt"代表时间短语,"S"代表句子等。

2.3 时间模式的多种形式

定义 6(基本时间模式,BTP) 指最基本存在的且不可再分解的最小时间模式(包括简单时间短语和固定表达句子)。

定义 7(非基本时间模式,NBTP) 指除基本时间模式以外的其它时间模式。包括:(1)由实量和基本时态模式相互插入而成的时间模式,例如"从今天起"的 TP 为"从 Nt 起";(2)由实量和变量相互插入而成的句子时间模式, 例如 "他已经走了"的 TP 为"$N_{人}$已经 V 了"。

定义 8(汉语时间模式,CTP) 指将汉语句子或短语提取成时间模式,且模式中实量用汉语词语表示。

定义 9(英语对照时间模式,ETP) 指将汉语句子或短语的译文提取成时间模式,且模式中实量用英语单词表示。

2.4 汉语句子时间信息形式化

2.4.1 词汇信息形式化

(1)时间短语形式化

根据 BTP,将简单时间短语作为一个实量,PT 可为 Nt(时间名词或短语)、Nf(节日)、Nhe(历史事件)或 Ne(事件)。

将复合时间短语表示成两个或两个以上 BTP,例如"昨天早上"是由两个 BTP"昨天"和"早上"组成。

将复杂时间短语用 NBTP 表示,即将表时间的连词、介词、方位词作为实量,将该词前后搭配的简单时间短语或复合时间短语作为变量。例如"三个月前"表示成"Nt 前",其中变量 Nt 可以被其他变量类型相同词语所替换。

(2)时间副词和时间助词形式化

将在句子中单独出现的时间副词和时间助词提取成实量,其他成分提取成变量。例如"他正在看书。"的 CTP="$N_{人}$正在 VN"。

将在句子中复合出现的时间副词和时间助词提取成实量,其他成分提取成变量。例如"他正玩着呢。"的 CTP="$N_{人}$正 V 着呢"。

(3)动词情态形式化

将动词提取成实量,其他成分提取成变量。例如"他获得奖学金。"的 CTP="$N_{人}$获得 N"。

2.4.2 语法信息形式化

(1)特定句式形式化

将汉语句子用 NBTP 表示。汉语原文和英语译文对照提取。例如"他是上星期买的这本书。"的 CTP="$N_{人}$是 TV 的 N",ETP="It was Nt that $N_{人}$ $V_{一般过去时}$ N"。

(2)关联词语标记句形式化

用 NBTP 表示,即将句子中能够体现时态的关联词语提取成实量,其他部分作为变量。例如"他一到工厂,就开始下雨了。"的 CTP="$N_{人}$一 V_1N_1,N_2就 V_2N_3了",其 ETP="No sooner/Hardly/Scarcely had $N_{人}$ $V_{1ed}N_1$,than/when/when $N_2V_{2一般过去时}N_3$"。

(3)类虚拟语气句形式化

用 NBTP 表示,即将句子中能够体现虚拟语气的词语提取成实量,其他部分作为变量。例如"如果他听到事情经过,他就会采取其它做法。"的 CTP="如果 S_1,$N_{人}$就会 VN" 其前一分句时态判断结果="S_1=过去完成时",后一分句的 ETP="$N_{人}$ would have V_{ed}N."

3 多策略汉语句子时间分析与英译

3.1 构建时间模式库

包括(1)时间短语时间模式库 TPTPB;(2)单独时间副词和时间助词时间模式库 SATPB;(3)复合时间副词和时间助词时态模式库 CATPB;(4)动词情态时间模式库 VMTPB;(5)特定句式时间模式库 TSTPB;(6)关联词语标记句时间模式库 CSTPB;

(7)类虚拟语气句时间模式库 SSTPB。

各模式库共同存放数据:时间短语或句子 CTP、PT、ETP 或时间判断结果、汉语原文、英语译文。

另外(1)在 TPTPB 加入“能否确定时间”,能确定时间则直接给出时间判断结果,不能确定时间则给出 ETP;(2)若句子为复句,在 SSTPB 加入“复句前一句 ETP 或时间判断结果、复句后一句 ETP 或时间判断结果”。

3.2 汉语单句时间分析算法

(1)用 TPTPB 匹配单句/分句
If 匹配成功
  If 能确定时间
    输出时间判断结果并转(2)
  Else
    输出 ETP 并转(2)
Else
  转(2)
(2)用 SATPB 匹配该句
If 匹配成功
  输出时间判断结果并转(3)
Else
  转(3)
(3)用 CATPB 匹配该句
If 匹配成功
  输出时间判断结果并转(4)
Else
  转(4)
(4)用 VMTPB 匹配该句
If 匹配成功
  输出时间判断结果
Else
  转(5)
(5)用 TSTPB 匹配该句
If 匹配成功
  输出 ETP 或时间判断结果
Else
  输出翻译失败
(6)人工综合所有的 ETP 或时间判断结果。

3.3 汉语关联词语标记句时间分析算法

(1)按照“,”将汉语关联词语标记句分成几个分句待处理。

(2)用单句时间信息分析算法分别处理每个分句,输出各分句时间分析结果待用并转(3)。

(3)用 CSTPB 匹配整个复合句
If 匹配成功
  输出该 ETP 或时间判断结果待用并转(4)
Else
  转(4)
(4)比较(2)和(3)的结果中每个分句 ETP 或时间判断结果
If 相同
  输出该 ETP 或时间判断结果
Else
  输出(3)的结果

3.4 类虚拟语气句时间分析算法

用 SSTPB 匹配整个句子(单句或复句)
If 匹配成功
  输出 ETP 或时间判断结果
Else
  输出翻译失败

3.5 篇章信息识别规则

(1)文章体裁时间规则:人工判断所给篇章的体裁。总结出新闻等 19 种体裁,共 41 条规则。

(2)句子主题时间规则:提取句子主语,若主语=$N_{无生命物}$或$N_{事件}$,则该句时间判断结果默认为一般时;若该句主语=该句前第 x 句主语,则该句时间判断结果=该句前距离该句最近的第 x 句时间判断结果。

(3)上下文时间规则:对于无法判断时间的句子,若为首句,则令该句时间=下一句时间;若下一句仍是无法判断时间的句子,则令该句时间=下 x 句时间。若不为首句,令该句时间=上一句时间;若上一句仍是无法判断时间的句子,则令该句时间=上 x 句时间;若上 x 句都是无法判断时间的句子,则令该句时间=下 x 句时间。

3.6 汉语句子时间信息分析与英译算法

begin
  while <汉语句子非空>
  begin
    <根据文章体裁时间规则,确定篇章的基准时间>;
    <确定汉语句子为单句或复句>;
    <执行汉语单句时间分析算法>;
    <执行汉语关联词语标记句时间分析算法>;
    <执行类虚拟语气句时间分析算法>;
    <若输出翻译失败,则根据句子主题时间规则进行判断>;
    <若仍无法判断时间,则根据上下文时间规则进行判断>;
  end
    <输出 ETP 或时间判断结果或输出翻译失败>;
end

4 实验分析

4.1 实验设置

本文分别选取“中文自然语言处理开放平台”中的“汉英对照例句集(400 句)”和“双语句对齐语料库(1 500 句中前 500 句)”的双语句子级语料,“《走遍美国》(第 23~24 课)”和“Internet 上双语新闻(共 17 篇)”的双语篇章级语料进行开放测试。人工统计实验结果。

表 1 显示了测试语料的基本信息。

表 1 测试语料

	语料	句对总数
开放测试 1	汉英对照例句集	400
开放测试 2	双语句对齐语料库(1 500 句中前 500 句)	500
开放测试 3	《走遍美国》双语篇章级语料(第 23~24 课)	361
开放测试 4	双语新闻(共 17 篇)	186

表 2 显示了开放测试 1、2 和 3 的结果。表中第一行中的①表示输出的 ETP 或时间判断结果是正确的,且与原英语译文时态相同;②表示输出的 ETP 或时间判断结果虽与原英语译文时态不同,但经人工检查判断为正确;③表示输出的 ETP 或时间判断结果与原英语译文时态不同,经人工检查判断为翻译错误;④表示输出翻译失败。

(下转 45 页)