汉英机器翻译的英文词形态生成*

黄河燕 陈肇雄

(中国科学院计算所智能机译中心 北京 100080)

摘要 在汉英机器翻译译文生成中,一个主要的问题是如何根据句子的上下文语境获取有关时态、语态、句式和主谓性、数、格等信息,生成具有正确单词形态的译文,如动词的过去式、过去分词、现在式形式;名词的所有格、复数形式;助动词生成以及冠词的生成等. 本文提出一种基于SC文法的汉英机器翻译译文词形态生成算法,该方法通过设计一种生成导向的语言特征描述体系,采用译文生成和源文分析一体化的语言分析技术,使得译文生成能够充分利用源文分析过程中所用到的各种知识,准确地形成句子中各个成分的形态特征,并能有效地解决汉英机译译文生成中助动词和冠词的生成这一难题.

关键词 汉英机器翻译,译文生成.

中图法分类号 TP391.2

由于汉语和英语属于两个完全不同的语系,它们之间的表达习惯存在很大的差异. 汉语是分析型语言,其词(词组)的形态特征不丰富,很少有形态变化. 而英语是屈折型语言,其单词(结构成分)有很丰富的形态特征变化,语言表达方式也与汉语有很大不同. 因此,要把汉语这种分析型语言翻译转换为英语这种形态特征很丰富的语言,译文生成时,除了要解决一般机译译文生成中词序调整、多义区分等问题外,一方面需要根据上下文语境和英语的表达习惯解决诸如动词的时态(过去时、现在时等)、动词的形态(过去式、过去分词、现在分词等)、名词的数(单数、复数)、代词的格(宾格、所有格等)、副词形容词的级(比较级、最高级)等的生成问题;另一方面,在汉语中,谓语的形式一般不受主语性、数、格的影响,而在英语中,主语和谓语之间有性、数、格一致的要求,如复数主语和单数主语要求谓语动词也有相应数的变化形态. 同时,句子不同的时态、语态、句式也要求有相应形式的助动词. 因此,如何根据句子的时态、语态和句式,以及主语、谓语的性、数、格特征等生成相应的助动词形式和动词的形态也是汉英机译译文生成中要解决的一个问题.

此外,汉语中表示单个个体和不定个体时不用冠词进行修饰,而在英语中,表示不定个体和单个个体时要用不定冠词,指称确定的个体时要用定冠词. 因此,如何根据表示个体的名词的意义选择适当的冠词是汉英机译译文生成中要解决的另一个重要问题.

---

* 本文研究得到国家自然科学基金和国家863高科技项目基金资助. 作者黄河燕,女,1963年生,博士,研究员,博士副导师,主要研究领域为自然语言处理与机器翻译,面向对象方法,人工智能. 陈肇雄,1961年生,博士,研究员,博士导师,主要研究领域为机器翻译,大型智能应用系统等.

本文通讯联系人:黄河燕,北京 100080,中国科学院计算所智能机译中心 E-mail: Chenzhx@mimi.cnc.ac.cn

本文1996-11-11收到修改稿

在现有的一些汉英机译系统中,一般是通过译后编辑生成合适的单词形态.这一方面大大增加了译后编辑的难度,另一方面,不能充分利用在翻译转换的分析过程中所使用的上下文信息和语法、语义知识来判定单词所应具有的形态.

本文提出一种基于 SC 文法[1,2]的汉英机译译文词形态生成算法,在我们的算法中,根据汉英两种语言的转换生成特点,通过设置生成导向的特征描述体系,把译文单词的形态生成直接与源文分析和译文转换生成算法结合起来,通过在源文分析阶段[3,4]对一组具有相同源文结构、不同头部特征条件及上下文环境条件以及能表达相应语言形态特征的转换生成结构的SC 规则的测试,确定相应生成结构成分的形态生成特征,选择与该特征相应的转换生成结构,从而可以有效地解决上述问题.

下面,我们首先介绍生成导向的汉英机译特征描述体系,然后给出词形态特征的确定方法,最后给出词形态生成的具体实现算法.

1　生成导向的语言特征描述体系

针对汉英机译译文生成中存在的问题,我们在设计译文生成算法时,既要考虑如何从词典知识和上下文语境中获取有关时态、主谓性数格等信息,同时还要考虑如何在源文分析和译文生成时应用这些信息形成具有合适词形态的译文.

为了反映汉英两种语言在结构表达形式上的差异,方便汉英译文生成,我们设计了一种生成导向的汉英机译特征描述体系.在对语言的词汇和结构成分进行语法分类的基础上,进一步根据结构成分在汉英转换生成时的特性进行再分类.同时还根据每种词类可能具有的形态,定义相应的词形态特征.词形态生成特征用于标记每个结构成分所对应的译文单词所应具有的形态.根据英语的特点,我们对不同的词类设置相应的形态特征如下:

名词(NP)	SIG	表示名词单数形式	PLUR	表示名词复数形态
	POS	表示名词所有格形态(加'或's)		
代词(R)	SIG	表示代词单数形式	PLUR	表示代词复数形式
	POS	表示代词所有格形式	SUB	表示代词主格形式
	OBJ	表示代词宾格形式	SEL	表示反身代词形式
	WPOS	表示代词的物主代词形式		
动词(VP)	PAST	表示过去式形式	VEN	表示过去分词形式
	PRES	表示现在时第三人称单数形式	ING	表示现在分词形式
数词(Q)	NUMO	表示数词的序数词形式	SIG	表示数词单数形式
	PLUR	表示数词复数形态		
副词/形容词(D/A)	COM	表示形容词/副词的比较级形式		
	SUP	表示形容词/副词的最高级形式		
其它	TAN	表示形容词或名词所对应英文单词是以元音发音开头,当要在该形容词或名词前面加不定冠词时是加"an",而不是"a"		
	MOSY	表示形容词或副词所对应英文单词的比较级、最高级变形是不规则的,或是加 er,est 形式		

2　词形态特征的确定

要形成适当的英文词形态,首先要解决的是如何确定单词和结构成分所应具有的形态特征.在不同的情况下,形态特征的确定可以采用不同的方式.在我们的系统中,根据结构成分的不同情况,译文生成形态特征的确定有如下几种方式.

2.1　字典定义方式

译文生成特征的字典定义方式是在具有形态特征的单词字典定义中直接给出其相应形态的译文,并在其特征说明中给出对应译文所应具有的形态特征.这种方法适用于英文单词形态特征能够在字典定义时静态确定的情况下使用.例如,"我们"的字典定义如下:

$我们

R(PLUR,SUB) 〈上下文环境条件1〉 "we"

R(PLUR,OBJ) 〈上下文环境条件2〉 "us"

其中给出了中文词"我们"的两种不同格的译文表示"we"和"us",这二种不同译文表示所要求的上下文环境条件以及它们的相应形态特征SUB(主格)、OBJ(宾格)和PLUR(复数).

当源文分析通过上下文环境条件的测试[5]选择了其中的一个词条时,相应地也就确定了"我们"这个词的相应结构成分所具有的形态特征和其相应译文词所应具有的形态.

2.2 上下文相关语境的测试

英语词形态特征的第2种确定方式是在源文分析过程中,根据所归约的结构成分所处的上下文环境条件的测试来确定.例如规则:

VP(…) NP(…)→S(L,(1,n),昨天)|VP(PAST,…) ! VP(PAST) ! NP.

这条规则表示,当源文分析过程利用这条规则对结构成分VP NP进行归约时,如果VP NP的特征属性与当前状态的匹配成功,并且在VP NP结构成分的左边存在表示时间的词"昨天",则当源文分析过程利用这条规则进行了成功的归约时,则归约后形成的新的状态将具有PAST的特征.同时,译文转换生成模式中相应动词结构也标志上相应的形态特征PAST,表示该句子的时态为过去时,其相应动词应为过去式形式.在译文生成时,将调用特殊的处理过程生成该动词的过去式形式.这样,在源文分析过程中就利用对上下文环境条件的测试给出了结构成分VP应具有的形态特征PAST(过去式形式).

2.3 从被归约成分的特征集中导出、继承

有些结构成分的形态特征只能在句子分析过程中从其进行归约的更小的结构成分中推导出来.这时就需要根据英文的语法规律和表达习惯,如主谓性数格一致等的要求,从被归约成分的特征中导出或继承新的归约状态所应具有的形态特征.因此,译文形态特征的第3种确定方式是在分析过程中从被归约的结构成分的特征中继承或导出.例如,当"我们"这个词作为主语与一个动词成分归约时,根据英语主谓语性数格一致的要求,可以推出与其归约的动词成分的译文所应具有的形态特征和相应句子的助动词.例如规则:

AP(…) NP(…,PLUR,…)→|NP(…,PLUR,...),! AP ! NP(PLUR).(继承)

AP(…) NP(…,SIG,…)→|NP(…,SIG,...), ! AP ! NP.

当源文分析过程利用其中的一条规则进行了成功的归约时,生成的新的归约成分NP就从其头部成分的属性特征中继承了NP的形态特征PLUR(复数)或SIG(单数).源文分析过程通过对这两条规则的测试选择,相应地也选择了新的归约成分NP所具有的形态特征PLUR和SIG.当头部成分NP带有复数特征时,其相应成分的译文结构也加上复数的生成特征.而对于名词的单数形式则不必对其相应译文作进一步的处理,直接取其译文即可.

又例如规则:

VP(…) XU(…,过,…)→〈上下文环境条件3〉|VP(…,PAST,…),! VP(PAST)(导出)

NP(TIM,FUT,…) VP(STNA,…)→S(R,(1,1),.)|CS(…),it ! AUX ! VP(ING) ! NP.

当源文分析过程利用它们进行了成功的归约后,归约后的结构成分 VP 从其被归约成分的特征中导出它所应具有的形态特征分别是过去式形式(PAST)和现在分词形式(ING). 其中第 1 条规则说明了当一个动词成分和一个虚词归约时,如果对应的虚词是"过",并且满足〈上下文环境条件 3〉,则相应的归约状态中应包括特征 PAST,相应动词的译文形态应该是过去式形式;第 2 条规则说明了当一个表示将来(FUT)时间(TIM)特征的名词和一个描述自然现象(STNA)的动词进行归约,并且其左边没有其它原文成分,右边句末为句号(即为陈述句)时,它们形成了一个子句结构. 它的译文生成结构根据英语的表达习惯,加入了表示自然现象的形式主语"it"和表示将来正要发生的时态特征的助动词"will be"(由! AUX 生成),及作为谓语的动词结构 VP 的相应译文! VP 的进行时特征 VNG,以生成该动词的进行时态,并把作为实际主语的结构成分 NP 的相应译文! NP 的顺序作了调整,放在最后,以符合英文的表达习惯.

3　词形态生成算法

如上所述,基于 SC 文法的译文转换生成与源文的分析过程是同时进行的. 因此,一旦源文分析过程通过对 SC 规则结构生成条件的测试进行了成功的结构分析,相应的也就唯一地选择了译文的转换生成结构. 当源文分析产生了成功的内部结构分析树,则树中每个节点在译文生成中所要使用的转换规则即已确定,即由转换规则中的转换体来说明的译文生成结构也就确定了,从而使译文生成过程能够充分利用归约过程中所使用的各种知识形成准确的单词形态.

一旦规则中的转换体部分确定了合适的结构成分的形态特征,则译文生成时,只要把具有形态特征的相应部分译文的形作为参数,调用独立的单词形态生成模块便可得到其相应的形态. 词形态生成模块分为两个部分,一是助动词的生成;另一个是一般词形态的生成处理. 下面分别给出这两个过程的实现算法.

3.1　助动词生成

助动词的生成是根据当前句子归约状态的特征生成与该句子的时态、语态、句式相应的助动词形式. 它的具体实现过程为

(1) 从当前句子归约状态中取与译文生成相关的特征;

(2) 判断所要生成的助动词的类型,如果是系动词,则转(3);如果是一般实义动词的助动词,则转(4);如果是情态动词,则根据其相应的情态动词标记和其时态特征分别加不同的情态动词,如:can/could, may/might, will/would, shall/should等,然后,转(5);

(3) 根据所取得的句子时态特征,为系动词生成不同的助动词:

将来时:will　　完成时:have/has(PLUR/SIG)

过去时:were/was(PLUR/SIG)　　现在时/进行时:am/are/is(第一人称单数/PLUR/SIG)

然后转(5);

(4) 根据所取得的时态特征和句子的句式结构等特征,生成相应的助动词:

将来时:will　　将来过去时:would　　过去进行时:were/was

过去时被动态:were/was　　过去时的否定句、疑问句:did　　一般完成时:have/has

一般进行时:am/are/is　　一般现在时被动态:am/are/is　　过去完成时:had

现在时的否定句、疑问句:do/does

然后转(5);

(5) 判断句子是否有否定特征,如果有,则在助动词后加上否定词"not".

3.2 一般词形态生成

一般词形态生成是根据译文生成模式中相应译文结构成分的语法分类及其相应的形态特征,生成不同形态的译文串.形态生成过程的调用形式为

gen_morph(syn_type, morph_flag)

其中 syn_type 为结构成分的语法类(词类)标记;morph_flag 是形态特征标记.其处理算法如下

(1) 判断结构成分的词类,并转向不同的处理;

(2) 如果译文结构成分的词类是名词结构,则根据该结构成分的形态特征:

- 如果是 POS(所有格),则根据该词不同的词尾分别加 ',或加 's;
- 如果是 PLUR(复数),则先查名词复数不规则变形表,如查到则返回相应的不规则变形,否则根据不同的词尾分别加 s,es 或删 y 加 ies;

(3) 如果译文结构成分的词类是代词结构,则根据该结构成分的形态特征查代词变形表,并返回相应形态的代词;

(4) 如果译文结构成分的词类是动词结构,则根据该动词结构成分的形态特征查动词的不规则变形表,如查到则返回具有相应形态的不规则变形,否则根据成分的形态特征:

- 如果是第三人称单数,则根据不同词尾加 s,es 或去 y 加 ies;
- 如果是过去式或过去分词,则根据不同词尾加 d,ed 或删 y 加 ied;
- 如果是进行式,则加 ing;

(5) 如果译文结构成分的词类是数词结构,

- 如果形态特征为序数词形式,则根据不同情况加 th,st,nd 或 rd.;
- 如果形态特征为复数形式,则根据不同情况分别加 s,es 或删 y 加 ies.;

(6) 如果译文结构成分的词类是形容词/副词结构,则根据成分的形态特征,查形容词/副词的不规则变形表,如查到,则返回相应形态的不规则变形,否则根据成分的形态特征:

- 若为比较级形式,则加 er 或 r 或去 y 加 ier;
- 若为最高级形式,则加 st 或 est 或去 y 加 iest.

上述英文单词形态特征的这种生成算法很好地利用了句子分析过程中所用到的各种知识,以形成准确的形态特征.同时在基本不改变 SC 文法的译文结构转换与生成算法[6]的前提下,只需增加一个独立的英文单词形态生成模块便可生成适当的英文单词形态和句子的助动词形式,保证了汉英、英汉两个机器翻译系统中译文生成算法的兼容性.

4 结束语

综上所述,我们给出了一个基于 SC 文法的汉英机译系统译文词形态生成算法,该算法已经在 SUN/SPARC 工作站上用 C 语言实现,并应用于实际的汉英机器翻译系统中.目前,该系统已能够翻译几万句日常方面的句子,在所选的语料中其正确率达到 85%以上.当然,汉英机译是一个高难研究课题.我们希望这里的工作能够为汉英机译的进一步研究和发展提供一些有益的参考.

致谢 本文在算法的设计和实现工作中,中科院计算所机译中心和科智语言信息处理研究所汉英课题组的全体成员在汉语、英语语言规律总结方面作了许多工作,作者向他们表示衷心的感谢.

参考文献

1 陈肇雄,高庆狮. 智能化英汉机译系统 IMT/EC. 中国科学(A 辑),1989,19(2):186～194.

2 陈肇雄. SC 文法功能体系. 计算机学报,1992,15(11):801～808.

3 Huang Heyan, Chen Zhaoxiong. Design and implementation of the Chinese-English machine translation system. The First China-Korea Joint Symposium on MT(CKJSMT'94), 1994. 19～24.

4 陈肇雄. 机器翻译研究进展. 北京:电子工业出版社,1991.

5 Huang Heyan, Chen Zhaoxiong. Forward and feedback context-sensitive processing. Inter. Conf. on Computer Processing of Oriental Language, ICCPOL'97, 1997.

6 黄河燕,陈肇雄. 机器翻译译文生成算法. Inter. Conf. on Chinese Computing 1994, Singapore, 1994. 40～46.

MORPHOLOGIC GENERATION IN CHINESE-ENGLISH MACHINE TRANSLATION

HUANG Heyan　CHEN Zhaoxiong

(*IMT Research Center　Institute of Computing Technology　The Chinese Academy of Sciences　Beijing*　100080)

Abstract　In the development of a Chinese-English machine translation system, one of the key problems is how to generate different morphologies of words(such as, the possessive and plural forms of noun, the past participial form of verb etc.) according to current context situation. In this paper, a SC grammar-based generation algorithm for Chinese-English machine translation is proposed. In which, by designing a generation-driven linguistic feature description scheme and taking a unified language analysis approach of incorporating the generation with parsing, through using various knowledge which is produced and used in parsing, it is possible to generate proper morphologic feature of each component and to efficiently generate corresponding auxiliary verb and articles in sentence.

Key words　Chinese-English machine translation, generation.

Class number　TP391.2