逻辑并列机器翻译中的ＲＳＡ处理方法

On Reservable Structural Ambiguities in Logically Parallel Japanese-Chinese Machine Translation

任福继
Ren Fuji

日本ＣＳＫ技術研究開発本部
MT Dept., CSK Corporation, Japan

摘要

本文中，将描述机器翻译中可保留暧昧关系(RSA)这一新概念及其处理手法．所谓RSA是指既使不解消原入力文中授受构造上的暧昧性也能生成反映原文意思的译文．具体而言，就是研究由日文中的并列助词「と」、连体助词「の」和名词构成的名词句及由用言连体形构成的句子中授受构造上的可保留暧昧性，在此基础上提出了一个新的翻译方法，并建立了一个应用于逻辑并列型日中机器翻译中的RSA处理系统．

Abstract

In this paper a new concept, Reservable Structural Ambiguities (RSA) in machine translation, is presented. The RSA are structural ambiguities of one language which can be translated to another without being resolved. Usually, when translating from one language to another which doesn't belong to the same "language family", structural ambiguites must be resolved. However, Japanese and Chinese don't belong tothe same family, their difference being mostly on sentence structure. However, some parts of their sentences are similar, making RSA possible, and this paper will focus on them. Some RSA patterns will be discussed and a method for generation of Chinese from Japanese sentences with RSA will be given. An experimental system based on this method was constructed and an experiment was carried out. The result was a correct translation rate of about 97.6%, showing that the proposed method is quite effective.

关键词 ：

(1)机器翻译 Machine Translation；
(2)可保留暧昧关系 Reservable Structural Ambiguities；
(3)日本语 Japanese；
(4)中国语 Chinese；

联络地址：〒214 日本国川崎市多摩区菅馬場２－１－１－２０３
TEL & FAX：(081) 44-945-1719

一、引言

一般在进行不同语系间的机器翻译系统中，为了得到正确译文必须解消原语言中授受构造上的暧昧性．目前已被实用化的机器翻译系统，多是在进行构文解析的同时运用动词的框架、名词的意味属性等进行意思解析，解消单词意思或授受间的暧昧性1),2),3)．但是一个由20～40个单词构成的普通句子中，若将多品词与授受构造间的暧昧性综合考虑的话，统语解析的候补个数将有可能上亿10)．因此，为了解消授受构造上的暧昧性就需要大量的处理时间．尤其是解消象”山田の家と寺が火事で焼けた”，”美しい谷間の百合”等文中的授受构造上的暧昧性是极为困难的．特别是后一例文中，也许说者（著者）自己就没有严密考虑过到底要表现什么样的授受关系．

本文中，我们将考虑日中机器翻译中可保留暧昧关系(RSA)．所谓可保留暧昧关系是指既使不解消授受构造上的暧昧性也能生成译文的暧昧关系．

例如，”山田の家と寺が火事で焼けた”文中，具有下述辆种构文构造A，B．

构造A：（山田の家）と寺が火事で焼けた．

构造B：山田の（家と寺）が火事で焼けた．

但构造A，B所对应的中文译文相同：

中国语：由于火灾山田的家和寺庙被烧了．

本文中我们把具有这种性质的暧昧性叫作可保留暧昧关系．也就是说，日语和中文虽分属不同语系，句子整体结构及表现方式不同，但就句子中的某一部分而言，其构文及表现却有极为类似之处，对该处理单元而言，既使保留日文中的授受间暧昧性也能译出符合原文意思的中文．当然，并非指所有的授受构造上的暧昧性均是可保留的，而是指其中的一部分．为此，本文中我们把日语的授受构造上的暧昧性分为可保留暧昧关系和必须解消的暧昧关系两种．并提出了对可保留暧昧关系不进行暧昧性解消，而是用已求出的译文函数直接生成译文的方法．

二、可保留暧昧关系

2.1 日语中授受构造上的暧昧性

若一个文节中的授受对象同时具有二个以上时，我们称之为具有授受构造上的暧昧性．下面，我们将列出日文中具有授受构造上的暧昧性的例子．

（a）由并列助词「と」，连体助词「の」，名词构成的名词句：图1中的例2-1，无论是文法上，还是意思上辆个构文树均成立．也就是说：作为「一太郎」的授受对象，「花子」和「子供」这辆个名词同时成立．

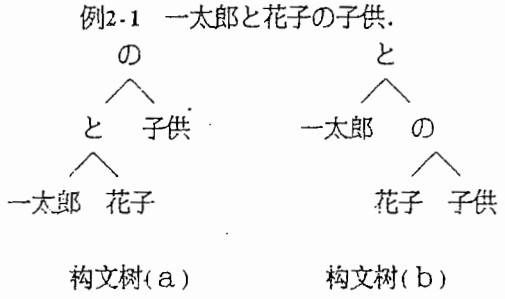

图1 具有授受构造上暧昧性的例子

Fig.1 Examples of structural ambiguities

（b）由形容词构成的句子：例2-2中，作为形容词「美しい」的授受对象，「楊貴妃」和「目」这辆个名词同时成立.

例2-2 美しい 楊貴妃の 目.

（c）与动词具有授受关系的句子：例2-3中，作为「ワープロで」的授受对象，「翻譯した」和「修正する」这辆个动词同时成立.

例2-3 ワープロで 翻訳した 結果を 修正する.

2.2 具有授受构造上暧昧性的日文的中文译文

下面，我们将用前一节中的例文来说明日中辆语言中授受构造上暧昧性的对应关系.

（a）并列助词「と」和连体助词「の」的场合：例2-1是具有授受构造上的暧昧性的日文，对应于构文树（a），构文树（b）的中文构文树虽不同，但所生成的中文译文却一样.

（b）形容词的场合：图1中例2-2的日文与图4中的中文相对应.「美しい」的授受对象无论取①，还是②生成的中文译文均相同.

（c）与动词具有授受关系的场合：例2-3中，若「ワープロで」的授受对象为「翻譯する」则其中文译文为图2的①. 若「ワープロで」的授受对象为「修正する」则其中文译文为图2的②. 也就是说在本例中，因授受对象不同而生成了不同的译文.

总上所述，例文2-1，例文2-2中的2种授受构造均生成同一中文译文，换言之，这种授受关系上的暧昧性既使不进行解消也不会影响译文结果. 然而，例文2-3中因授受对象不同而产生了不同的译文，因此就必须进行暧昧性解消.

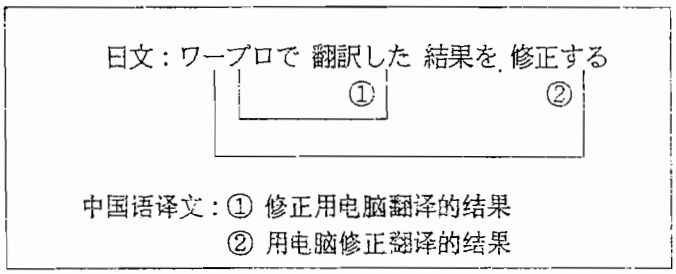

图2 例2-3的中国语译文

Fig.2 The translated Chinese sentence of ex.2-2

2.3 可保留暧昧关系及其特征

下面我们将说明可保留暧昧关系的三个主要特征.

a．可保留暧昧关系必须是被定义在特定的两语言之间.

b．可保留暧昧关系具有双方向性.

c．对可保留暧昧关系，根据已求出的译文关数可直接生成译文。

三、日中机器翻译中可保留暧昧关系的种类

如2．3节所述，日中机器翻译中授受构造上的暧昧性并非全部是可保留的. 因此必须从具有暧昧性的授受构造中抽出可保留暧昧关系.

本章中，我们把可保留暧昧关系的种类分为以下三大类. ①由并列助词「と」，连体助词「の」，名词构成的句子；②由形容词连体形，名词，并列助词「と」，连体助词「の」，名词构成的句子；③由动词连体形，名词，并列助词「と」，连体助词「の」，名词构成的句子，下面将一一进行说明.

3.1 「NとNのN」及「NのNとN」型

关于由并列助词「と」，连体助词「の」，名词构成的句子中授受构造上的暧昧性已有众多研究，并已总结出其种类分类及出现频度的报告4),12)．我们将之归纳分为以下３种类型．

类型１：NとNのN［のN］*

类型２：［Nの］*NのNとN

类型３：［Nの］*NのNとNのN［のN］*

其中，N为名词，「と」为并列助词，「の」为连体助词（下同）．［x］* 为x的反复．

3.2 「ANのN」及「ANとN」型

「美しい谷間の百合」中常々被作为授受构造上的暧昧性的代表例被举出．这种暧昧性在日中机器翻译中为可保留暧昧性．这种与形容词（包括形容动词）相关连的可保留暧昧关系型可分为以下２种类型．

类型４：ANとN

类型５：ANのN［のN］*

其中，A为形容词连体形．

3.3 「VNのN」及「VNとN」型

与动词相关连的可保留暧昧关系型具有以下２种类型．

类型６：VNとN

类型７：VNのN［のN］*

其中，V为动词连体形．

据此，我们总结出了可保留暧昧关系的类型（本文略）．

四、可保留暧昧关系的处理方法

本章我们将说明正在开发中的日中机器翻译系统中可保留暧昧关系的处理概要．

如前所述，一般在机器翻译中，若有授受构造上的暧昧性，就必须对它进行解消．但是，如果它是可保留暧昧关系，不解消其暧昧性而直接使用译文函数就可直接生成译文．所谓译文关数是指对应于可保留暧昧关系中日语各种类型的中文种类．例如：

例文4-1：ラジオと車のバッテリー

该例文的类型为

「N1とN2のN3」 ... ...<1>

其中，「と」为は并列助词，「の」为连体助词．因此为可保留暧昧关系．

例文4-1的中文译文为「收音机和汽车的蓄电池」，用一个形式化式子写出的话为，

「M1和M2的M3」 ... ...(1)

(1)是日语类型<1>所对应的译文函数．Mi是日文Ni所对应的中文名词（i=1,2,3）．

在含有用言的句子中，必须要考虑到日文与中文的词序不同．

例如：对应于日文「ご飯を食べる」的中文为「吃（食べる）饭（ご飯）」．

例文4-2：骨を食べる犬と猫．

在例文4-2中，将宾语（骨を）除开，则该文的类型为：

「VNとN」... ... ... <2>

在此，「V」代表动词「食べる」．与动词「V」相对应的中文动词「D」为「吃」．如果不调整宾语的位置，既：

「D的M和M」 ... ...(2)

则其中文译文为： “骨头吃的狗和猫” 为错误译文.

因此，在分析具有动词，形容词的可保留暧昧关系时，应把与之相关连的文节一并作为对象进行处理 即要把动词和所带的宾语等成分作为一个整体来考虑.

例如，例4-2中「D」的位置应移至宾语的前面，变成「吃骨头（骨を食べる）」。

若宾语也具有复数文节，例如「（鱼の骨を）食べる犬と猫」，则需要在进行宾语解析的基础上，生成「吃鱼的骨头」。有关用言语顺变换处理问题将在别的机会予以说明.

以下，我们将用例文4-4来说明其概要.

例文4-4：制御部が信号を記憶部と演算部の科学演算レジスタへ送出する。

首先，对入力的日文进行形态素解析，生成代码要素14,15). 然后，应用RSA检出器检出含有可保留暧昧关系的文节. 本例中为”記憶部と演算部の科学演算レジスタ”被检出. 之后，用RSA翻译器对检出的含有可保留暧昧关系的文节进行翻译处理. 与此同时，系统把这一结果作为一个要素，变换成宾语码列后生成译文. 详见文献［14,15,16］. ”RSA处理”表示用RSA检出器和RSA翻译器进行的可保留暧昧关系的处理，”通常的翻译”表示采用文献［15］中所提家族模型方式的翻译处理.

下面，我们简单绍介一下RSA检出器和RSA翻译器.

所谓RSA检出器就是把可保留暧昧关系类型与代码要素列进行匹配，检出一致的部分。之后，把这一部分作为一个要素使日文要素列被缩短。

所谓RSA翻译器就是把被检出的可保留暧昧关系部分用译文函数来生成译文。例文4-4中，被检出的部分为”記憶部と演算部の科学演算レジスタ”，用该译文函数就生成了中文译文”存储器和计算器的科学计算寄存器”。

五、实验及考察

目前已将第4章中所述的可保留暧昧关系的处理流程作为子程序放入正开发中的逻辑并列日中机器翻译实验系统中9),15)，并就确认该方式的有效性及正确性等进行了实验.

5．1 实验

从有关情报处理的论文，研究报告及有关科学技术论说，人物传记、教科书等抽出了含有可保留暧昧关系的句子进行了实验.

（1）数据情报

- 总文数 =2,919 个
- 总文字数=197,400 个字
- 平均文长=197,400个字／2,919个=68个字／句
- 被抽出的可保留暧昧关系句=806句
- 可保留暧昧关系句的出现率=812 个／2,919个×100%=27.8%

（2）抽出方法

一部分由计算机自动抽出，另一部分由手工作业抽出. 但抽出的均是含有可保留暧昧关系的部分.

（3）实验结果

本实验是与翻译系统分开进行的. 评价指标为译文准确度既正译率. 实验结果为2919句中错误的有67句. 因此正译率为97.7%.

5．2 考察

关于可保留暧昧关系类型，如5．1中所述，被抽出的可保留暧昧关系句为812个，可保留暧昧关系句的

出现率为27.8%（可保留暧昧关数句／总句数＝806／2919）．其中一个句子中含有复数个可保留暧昧关系类型时，作为复数个可保留暧昧关系句来计算．此外，本文中所示的类型不是可保留暧昧关系的全部．例如，出现频度高的类型「N の N の N」是具有授受构造上的暧昧性的类型，但在上记数字中没有包括．若将此记入，则可保留暧昧关系句的出现率约为47%．

当然，目前由本方法所检出的可保留暧昧关系并非完全都能译出。主要是并列助词”と”的判断问题。有的文中的”と”与通常的并列助词”と”的功能不同。作为这种问题的解决可以考虑用追加并列助词的判断规则或利用名词的意思属性等方法。

六、结束语

日中两国语虽分属不同语系13)，但就句中的某一部分而言，例如由并列助词”と”和连体助词”の”构成的句子中两国语的构造却是相同的．本文从机器翻译的观点出发着眼于上述特征，提出了日中机器翻译中可保留暧昧关系及其翻译方式．并进一步用实验确认了本文所提方法的有效性．另外，ＲＳＡ这一概念和处理手法也可以推广到他语种机器翻译系统中。

参考文献：

1）牧野武則：機械翻訳，オーム社，(1989).

2）COLING-90、Computational Linguistics,(1990).

3）向仲：”技術文書の機械翻訳における常識と文脈情報の利用”，情処学論，31,8,pp.1168-1173,（1990）

4）田村直良，田中穂積：”意味解析に基づく並列名詞句の構造解析”，情処研報，NL59-2（1987）

5）平井章博，梶　博行，芦　沢実：”機械翻訳向け前編集のための日本語係り受け構造の曖昧性検出方式”，情処学論，31,10,pp1425-1437（1990）

6）有田英一，福島正俊，進藤静一：”日英機械翻訳システムにおけるプリエディットについて”，情処研報，NL48-7，（1985）

7）長尾　真，田中伸佳：”制限文法に基づく文書作成援助システム”，情処研報，NL44-5，(1984)

8）Tomita,M.：”Sentence Disambiguation by Asking”，Computers and Translation,Vol.1,pp.39-51（1986）

9）任　福継，宮永喜一，栃内香次：”日中常用文型機械翻訳システム”，信学論（Ｄ－Ⅱ），Vol.J74-D-Ⅱ,No.8,PP.1060～1069（1991）.

10）天野真家，平川秀樹：”英日機械翻訳用パーサについて”，情処研報，NL32-1，(1982)

11）東条　敏：”自然言語処理入門”近代科学社（1988）

12）稲垣博人，壁谷喜義，小橋史彦：”意味連結パターンを用いた係り受け解析”，情処研報，NL67-5，（1988）

13）望月十八吉：”中国語と日本語”，光生館，（1974）

14）任　福継，宮永喜一，栃内香次：”コード方式日中機械翻訳の実験システムJCMTの概要”，情処研報，Vol,NL72-7，(1989,5)

15）任　福継，范莉馨，宮永喜一，栃内香次：”家族モデルを用いた文の分解に基づく日中機械翻訳システム”，情報処理学会論文誌Vol.32,No.10,PP.1249～1258（1991）

16）任　福継，宮永喜一，栃内香次：”意味属性による中国語補助語の推定アルゴリズム”，情報処理学会論文誌，Vol.32,No.11,PP.1374～1382（1991）