中 文 信 息 学 报
第 12 卷 第 3 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 12 No. 3

汉英机器翻译中的冠词处理研究①②

常宝宝　　刘 颖 刘 群

北京大学计算语言学研究所　北京　100871　　中国科学院计算所二室　北京　100080

摘要　对于汉英机器翻译而言，由于汉语中缺乏与英语冠词相应的语言范畴，而且面向人的冠词用法规则很难满足机器翻译处理的需要，冠词的误用严重地影响了最终译文的质量。本文提出一种将基于转换的错误驱动的学习机制用于冠词处理的策略，初步实验显示，这种方法可以有效地提高机器译文中冠词使用的准确率。

关键词　机器翻译　冠词选择　基于转换的错误驱动学习

一、前　言

八十年代以来，世界机器翻译研究取得了长足的进步，研究人员不断提出许多新的翻译方法和翻译策略，包括：基于实例的翻译方法[1]、基于统计的翻译方法[2]以及多引擎机器翻译方法[3]。与此同时，我国机器翻译研究也取得了不少进展，不少系统实现了商品化。总的看来，我国的机译研究主要是外汉和汉外机译研究，尤其是英汉、汉英机译研究。值得注意的是，从译文质量看，英汉系统要远优于汉英系统[4]。就其原因，一般认为，主要是由于汉语的特点造成的，汉语是一种缺少形态的语言，分析难度高于英语，而另一方面，由于英语形态丰富，较难生成，在汉语到英语的翻译过程中，一些汉语中缺乏的语言范畴在英语生成过程中难以处理，主要表现在：汉语中没有明显的数的变化，而英语中有严格的数的变化；汉语中没有确定的区分时态的形式标记，而英语中有严格的时态区分；汉语中没有冠词等概念，而英语中名词短语前常伴随有冠词。总之，汉英系统要走向实用，还需付出艰苦的努力[4]。

冠词在英语中属于限定词，一般位于普通名词之前，它可使得名词含义明确[5]。在汉英机器翻译中其之所以难以处理，一方面是由于汉语中没有相应的语言范畴。另一方面也是由于冠词在英语中的有无常常牵涉到复杂的语义知识、话语知识以及世界知识，用法规则难以形式化。有关英语冠词的使用，许多英语语法论著都有详细的说明，如《朗文英语语法》一书关于冠词用法一章有 30 多页[5]。这些冠词的用法规则无疑都是正确的，但用于机器翻译，它们的可操作性极差，要使机器能够利用这些规则，表示和编码存在极大的困难。人对这些规则的掌握也比较困难。我国学生在英语写作中，冠词是一个容易出错的方面。可见针对人的冠词使用规则显然不适于机器翻译的处理。K. Knight 等曾提出利用决策树的学习方法获取冠词选择规则[6]，但事实上，这种学习方法在该类问题上能力还是有限的。而在实践上，目前汉英机器翻译系统中一般还缺乏有效统一的冠词处理策略，冠词选择问题没有得到足够重视。

① 本文 1997 年 7 月 18 日收到
② 本文工作得到了国家 683 计划的资助（编号：863－306－03－06－2）

本文提出一个将基于转换的错误驱动的学习(Transformation－Based Error－Driven Learning[7])应用于冠词选择规则获取的处理策略,目前已在我们研制的一个汉英机器翻译系统①中开始采用,初步实验显示,该策略可以有效地提高冠词使用的准确率。

二、冠词的统计特征

K. Knight[6] 等曾就冠词的使用做了如下富有启发意义的统计及试验,他们在不考虑零冠词的情况下首先考察了定冠词和不定冠词的统计分布,得出在所统计的文本中定冠词与不定冠词存在如下的分布:

° *a* = 28.2%　　° *an* = 4.6%　　° *the* = 67.2%

其次,他们又以人为主体做了如下的试验。给定一篇英文文本,将其中的冠词替换为空格,在给定不同范围的上下文的情况下由受试者恢复其中的冠词(受试者母语为英语),并统计冠词恢复的准确率。在给定全部上下文的情况下,冠词恢复的准确率可达到 94%～96%。这实际上说明冠词基本上可以由其所处的上下文预测出来,语言在没有冠词的情况下,意义传递不存在特殊障碍,但英语冠词也不是完全可以预测的,大约 5%的冠词不能从完整的上下文中预测出来。这个实验同时也说明了,从香农意义上说,冠词含有较少的信息量。缩小上下文,在仅仅给出中心名词及其前修饰语的情况下,冠词恢复准确率可以达到 79%～80%。在给出中心名词、中心名词后面的两个单词、未知冠词前面的的两个单词的情况下,受试者给出了 83～88%的准确率。

上述试验没有考虑零冠词的使用,然而,从汉英机器翻译译文冠词处理的角度来看,必须处理零冠词。为了进行我们的工作,我们在考虑零冠词的前提下,以北京大学计算语言学研究所汉英机器翻译系统测试集中部分双语语料的英语部分为试验对象,重复了以上的试验。我们采用 1525 个英语句子作为实验材料(总计 16939 个单词),下面是该例句集的一个片段:

0156 The discovery that magnetism can produce electric current is extremely important in the field of electricity.

0157 Such a practice must be put an end to.

0158 It happened during the Long March.

0159 The Yellow River is the second largest river in China.

0160 The meeting will be held over until next week.

在(1)保留全部上下文;(2)给定前修饰语、中心名词;(3)给定冠词前的两个单词、前修饰语、中心名词、中心名词后的两个单词三种情况下抽去其中的冠词,设计测试问卷,并分别由两位通过CET－6 考试的同学按照(2)、(3)、(1)的顺序进行冠词填充,下面是第(3)种情况的测试问卷的一个片段:

0156－1 ____ discovery that magnetism

0156－2 discovery that ____ magnetism can produce

0156－3 can produce ____ electric current is extremely

…

0157－1 Such ____ practice must be

① 中国科学院计算所二室和北京大学计算语言学研究所目前正在联合研制的一个汉英翻译系统

然后对两位受试人员的得分求平均数，得到如下结果：

给定的上下文	准确率
给定全部上下文	94%
给定前修饰语、中心名词	77%
给定冠词的两个单词、前修饰语、中心名词、中心名词后的两个单词	86%

我们也对冠词的分布做了重新统计，例句集中共出现冠词 2324 次，其中零冠词 858 次，定冠词(the)1083 次，不定冠词(a)329 次，不定冠词(an)54 次，计算结果如下：

·零冠词 = 36.9%　·*a* = 14.15%　·*an* = 2.35%　·*the* = 46.6%

从试验结果来看，除了加入零冠词后，冠词分布发生改变之外，冠词恢复的准确率没有发生显著变化。

以上试验说明，冠词选择的正确率有一定的上限，在我们的实现中，我们限定上下文为冠词前面的两个单词、前修饰语、中心名词、中心名词后面的两个单词，我们的冠词处理程序准确率 p 的范围应为：46.6% <= p <= 86%。

三、冠词选择规则的自动学习

基于转换的错误驱动的学习算法是 E. Brill 于 1992 年提出的一个有效的学习算法。[7] 该算法已经在词类标注领域获得了较好的应用。该算法获取的规则以带有一定激发环境(triggering Environment)的转换式(transformation)的形式存在，直观易懂，即不需要花费大量的存储空间，一定程度上又可避免数据稀疏问题。

结合以上分析，我们将该算法用于获取冠词选择规则，我们使用的上下文范围为，冠词位置前的两个单词、中心名词及其前修饰语以及中心名词后面的两个单词，同时在设计中采用了词汇化匹配条件，以保证一些对冠词选择具有较强限制作用的词汇能出现在获取的规则中。

3.1　学习过程

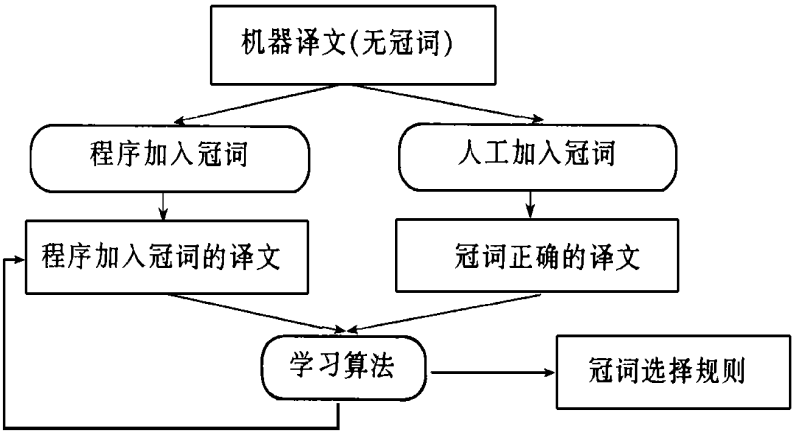
学习过程始于机器产生的译文，调整系统的译文输出，使得输出中不仅含有译文单词序列，同时保留分析、生成过程中产生的能用于冠词选择的语法标记(各标记含义参看文后附录)，例如，

(汉语)澳大利亚是世界上最大的岛。

(英语)(^)Australia(NPROP:NUCNT) is (UBE) (^) biggest(ASUP) island(NCONT:NSNGL) in(P) (^)world(NCONT:NSNGL).(PUNCT)

在选择保留那些特征时，对其中一些对冠词选择影响较大的词类保留了较细的分类标记，而对其余一些对冠词选择影响较小的词类仅保留了较粗的分类标记。

学习开始时，首先将其中的所有应该出现冠词的位置(^)填充为冠词 the，按照上节的统计结果，这样可以获得 46.6%的准确率，然后，人工校对得到正确文本，提交程序学习。学习过程如下图示：



3.2 候选转换规则激发环境的确定

在基于转换的错误驱动的学习算法中,首先要确定可能的转换空间。同样,冠词选择规则学习程序首先要确定可能的冠词选择规则的激发环境,然后再交由学习算法选择可导致最小错误率的选择规则。这些可能的激发环境可以由语料归纳出来。例如,依据以下的译文输出(*表示在激发环境中不考虑该词或相关标记):

…is in (') next room
UBE P POSI NCONT PUNCT
* * CIN NSNGL *
* *

可以获得 3×3×4×4×3=432 个候选冠词选择规则的激发环境,如:

P_{-2} P_{-1} M_1 H P_1
UBE P POSI NCONT PUNCT
* * next NCONT *
* * next NSNGL *
……

3.3 学习算法

根据 3.1 节的学习过程,由于每一个冠词位置最多有四种可能:零冠词 ø、定冠词 the、不定冠词 a、不定冠词 an,这样学习算法可以用伪码表示如下:

1. 初始冠词选择(一律选择定冠词 the),计算错误率 E_1,令 $E_2=1$;
2. 将指针定位到第一个中心名词处,向前寻找冠词位置;
3. 循环,当 $|E_1-E_2|>t$ 时,反复执行
   循环,变量 FromArticle 顺次取[the, ø, an, a],执行
   循环,变量 ToArticle 顺次取[the, ø, an, a],执行
   循环,在训练语料中,依次定位中心名词,并向前寻找冠词位置;
   若Correct-Article()=ToArticle 且 Current-Article()=FromArticle
   Num-good-transformation((FromArticle, ToArticle))++;
   否则,

若 Correct-Article()=From-Article 且 Current-Ariticle()=ToArticle

Num-bad-transformation((FromArticle, ToArticle))++;

在所有所学到的转换式中，寻找转换式 T，使得：

$\max_T$(Num-good-transformation(T)−Num-bad-transformation(T))

把获取的转换式存储在有序表中

利用获取的规则重新对译文进行冠词选择，$E_2=E_1$，计算错误率 E_1

上述算法中，对每一个带有激发环境的转换式(FromArticle, ToArticle)，都要扫视整个训练文本，对所有可能的转换式都要计算错改为对以及对改为错的次数，两者之差最大的是一个好的转换式。每次循环结束，可以寻找一条最好的冠词选择规则，然后利用该转换式对译文文本重新进行冠词选择，如此往复，直到错误率变化值小于某一阈值 t。

3.4 评价

为了验证学习算法的有效性，我们将前文提到的 1525 句双语料的汉语部分交由机器翻译系统翻译，对结果作适当编辑后(修改译文中的非冠词错误)，然后分别由人加入正确冠词和机器全部加入冠词 the，交由学习程序按 3.1 节的过程反复学习，最后截取获得的有序表中的前 2500 条规则进行分析和试验。

在我们对所学到的规则进行分析后发现，许多规则有一定直观意义，为了便于说明，我们对规则进行了如下改写：

IF=L1=UBE, HN=NMASS, HNINITIAL=C) THEN(the, a)

IF(P1=ASUPE)THEN(a, the)

IF("next"=P*i*)THEN (a, the)

这些规则分别和下述的例子对应：

It is a wonderful tea.

The largest island in China is Taiwan.

Miss Wang is in the next room.

我们又从双语例句集中另外选取 100 个汉语句子(不包含在训练使用的 1525 句中)交由翻译系统翻译，对结果适当编辑后(修改译文中的非冠词错误)，然后顺次应用转换表中的规则，并用下面的公式进行评价：

$$P = \frac{c}{t}$$

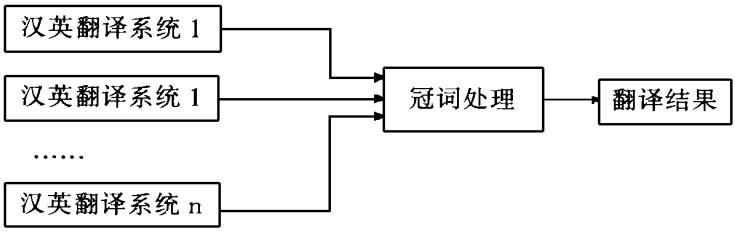
其中，t 表示译文中应该出现的冠词总数，c 来表示由程序准确插入的冠词数量。

最后得到 115/152=76%的冠词选择准确率。

四、进一步讨论

从以上的讨论可以看出，本文所提出的冠词处理策略并不能保障足够高的冠词准确率，该策略依然在理论上存在诸多的不足之处，同时在学习过程中由于考虑了词汇因素，仍然存在数据稀疏问题。但是，我们认为该策略也存在下面的一些优点。

该方法不依赖于具体的翻译系统，具有较高的独立性，可以作为一个部件在不经过大的调整的前提下供多个不同的翻译系统使用，有效地提高代码的重用性，具体可图示如下：



冠词的处理看似简单,但其处理却要牵涉复杂的语义知识、世界知识以及话语知识,而这些领域的形式化及相应的处理技术还相当有限,该方法为在不引入语义知识、世界知识以及话语知识的前提下处理冠词选择提供了一种思路。

同时,冠词选择规则的自动学习机制避免了手工编制规则所带来的不一致性。

五、结束语

本文针对汉英机器翻译,提出了一种冠词处理方法。其主要思想首先利用基于转换的错误驱动的学习算法,利用人工正确编辑过的文本,在错误率最小的原则下,自动学习冠词选择规则,然后在将习得的规则用于汉英翻译的冠词选择。该方法可以在不引入话语知识、世界知识的前提下有效地提高冠词使用的准确率,同时该方法较少依赖具体的机器翻译系统,具有较好的独立性,在软件实现方面易于模块化。

值得指出的是,由于英语冠词的准确使用牵涉的复杂语言知识,该方法所能获得的准确率还是有限的。而且目前,在我们的实验中,我们仅仅考虑名词短语前的冠词情况,并未考虑冠词使用的其他情况。我们正在进行进一步的实验并期望对这种方法进行改进。

参 考 文 献

[1] M. Nagao, A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, in: A. Elithorn et al. eds., Artificial and Human Intelligence, NATO Pubilcation

[2] P. F. Brown et al., A Statistical Approach to Machine Translation, Computational Linguistics, Volume 16, Number 2, 1990

[3] R. D. Brown et al., Applying Statistical English Language Model to Symbolic Machine Translation, in Proceedings of TMI 95, 1995

[4] 段慧明, 俞士汶, 关于 1995 年度机器翻译评测的总结报告,《计算机世界报》评测版, 1996 年 3 月 25 日

[5] L. G. 亚历山大,《朗文英语语法》, 外语教学与研究出版社, 1991, pp. 105—136

[6] K. Knight et al., Automated Postediting of Documents, in proceedings of AAAI94, 1994

[7] E. Brill, Transformation— Based Error— Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging, Computational Linguistics, Volume 21, Number 4, 1995

附录: 文中一些语法标记的含义

1. ACOMP: 形容词比较级形式	2. ASUPE: 形容词最高级形式	3. POSI: 形容词原级形式
4. C: 连词	5. D: 副词	6. M: 数词
7. NCONT: 可数名词	8. NUCNT: 不可数名词	9. NSNGL: 单数名词
10. NPLUR: 复数名词	11. NMASS: 整体名词	12. NCOLL: 集合名词

13. NPROP: 专有名词
14. NTITL: 称呼名词
15. R: 代词
16. P: 介词
17. UDO: 助动词 DO
18. UBE: 助动词 BE
19. UHAVE: 助动词 HAVE
20. UMOD: 情态动词
21. V: 动词
22. PUNCT: 标点

Research on Article Selection in Chinese-English Machine Translation

Chang Baobao
Institute of Computational Linguistics,
Peking University

Liu Ying Liu Qun
Institute of Computational Technology,
Chinese Academy of Sciences

Abstract Because Chinese language has no corresponding category with English articles and the usage rules of articles oriented to human are difficult to operationalize for machine translation, there are many cases in using articles incorrectly in Chinese-English machine translation system, which degrade the quality of the output translation severely. In this paper, we proposal a strategy for article selection which based the Transformation-Based Error-Driven Learning Algorithm, an initial experimental result shows the strategy can improve the accuracy in using articles.

Keywords Machine Translation Article Selection Transformation-Based Error-Driven Learning

---

（上接第 20 页）

Discuss the dictionary in the system of Japanese—Chinese Machine Translation

Yong Dianshu Hu Haiwen Chen Jiajun Wang Qixiang
Department of Computer scinces, Nanjing University
State Key Laboratory for Novel Software Technology
(Jiangsu Nanjing 210008)

Abstract In this paper, we discussed several problems, such as: homonym, potysemant. compatible type word, how to process idioms. We make a lot of researches in solving these problems which great effect the quality of generation languages in the system of Japanese—Chinese Machine Translation.

Keywords Machine Translation Dictionary Homonym Polysemant Idiom