文章编号：1003-0077(2012)01-0037-05

面向层次短语翻译的词汇化调序方法研究

肖欣延[1,2]，刘 洋[1]，刘 群[1]，林守勋[1]

(1. 中国科学院 计算技术研究所 智能信息处理重点实验室，北京 100190；
2. 中国科学院 研究生院，北京 100190)

摘 要：词汇化信息在短语调序中有重要的作用。然而层次短语翻译模型调序时并不考虑变量所泛化的短语的词汇化信息，因此该模型调序的歧义性较大。为此该文提出面向层次短语模型的词汇化调序方法。我们定义变量与邻接词语的调序关系，并使用变量所泛化短语片段的边界词信息来指导调序。在大规模语料的汉语到英语翻译评测任务中，我们的方法在 NIST 2003-2005 测试数据上获得了 0.6～1.2 BLEU 值的提高。

关键词：统计机器翻译；层次短语；词汇化调序

中图分类号：TP391 **文献标识码**：A

Lexical Reordering for Hierarchical Phrase-based Translation

XIAO Xinyan[1,2], LIU Yang[1], LIU Qun[1], LIN Shouxun[1]

(1. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. Graduate University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Lexical information plays an important role in the phrase reordering. However, the reordering in the hierarchical phrase-based (HPB) model does not consider the lexical information within the phrases, resulting in the reordering ambiguity. To alleviate this, we propose a lexicalized reordering method for the HPB translation. We distinguish two orientations of a variable comparing to its adjacent words, and use boundary words covered by the variable to guide reordering choices. In the large scale Chinese-English translation evaluation task, the proposed method improves the translation performance ranging from 0.6 to 1.2 BLEU on NIST 2003-2005 test-sets.

Key words: statistical machine translation; hierarchical phrase-based; lexical reordering

1 引言

层次短语模型[1-2]是目前统计机器翻译最好的模型之一。它简洁有效，在统计机器翻译中得到了广泛的应用。这种模型使用同步上下文无关文法(SCFG)形式的规则进行翻译。SCFG 规则将其左步重写为右部，其中右部包含对齐的两个部分：翻译的源语言端及翻译的目标语言端。在本文中，我们将用中文端和英文端来表示这两个部分。SCFG 规则既包含完全词汇化的规则，相当于短语模型[3]中所使用的短语；同时也包括含变量的规则。我们分别称这两种规则为短语规则及泛化规则。如表 1 所示，规则(1)的中文端和英文端都是词语，因此是短语规则；规则(2)是包含一个变量的泛化规则；而规则(3～5)则是包含两个变量的泛化规则。

表 1 SCFG 规则

X→<与 沙龙，with sharon>	(1)
X→<X_1举行，X_1 held>	(2)
X→<X_1举行 X_2，X_1 held X_2>	(3)
X→<X_1举行 X_2，held X_2 X_1>	(4)
X→<布什 X_1举行 X_2，bush held X_2 X_1>	(5)

收稿日期：2011-01-14 **定稿日期**：2011-04-21

基金项目：国家自然科学基金重点资助项目(60736014)；国家自然基金资助项目(60873167)

作者简介：肖欣延(1984—)，男，博士研究生，研究方向为统计机器翻译；刘洋(1979—)，男，副研究员，研究方向为统计机器翻译；刘群(1966—)，男，研究员，研究方向为自然语言处理、机器翻译、信息提取。

层次短语既能捕捉短距离调序，也能捕捉长距离调序。短距离调序主要通过短语规则来实现，而长距离调序则由泛化规则来捕捉。比如规则(4)，变量 X_1 在中文端处于“举行”之前，在英文端它的位置被调整到 X_2 之后。因为变量能够覆盖较长的短语块，使用这样的规则就能实现长距离的调序。

然而，由于泛化规则中的变量能够匹配任意的短语片段，因此在翻译过程中引起了较大的歧义。图1显示翻译中文片段“与 沙龙 举行 会谈”时使用的两种调序选择。X_1 泛化了(与 沙龙，with Sharon)的短语对；X2 泛化了(会谈，a talk)的短语对。虽然规则(3)、(4)都能够匹配该中文片段，但翻译效果迥异。规则(3)并不改变词语的中文端和英文端的相对位置关系；而规则(4)则大幅度调整 X_1 所覆盖的短语片段在英文端的位置。从语言学的角度来分析，规则(3)认为 X_1 是名词性的成分，英文翻译应位于动词“held”之前；而规则(4)则将 X_1 作为补语，因此在英文端应该位于动词“held”后面。在当前的上下文中显然使用规则(4)更为合适。然而规则(3)在训练语料中十分常见，导致第一种翻译被选择的可能性更高。通过分析可以发现，调序与变量所覆盖的短语片段是相关的。事实上，短语调序模型上的工作[4-7]已经表明，词汇化的信息对调序十分重要。

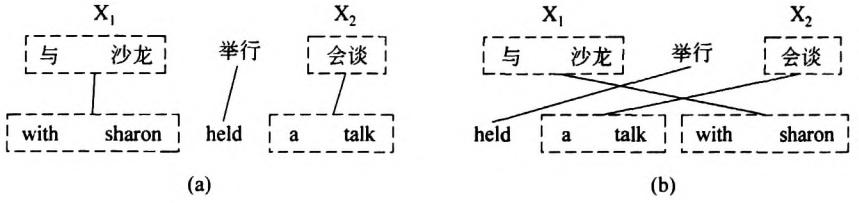

图1 两种调序选择

注：(a) 使用规则(3)进行顺序翻译；(b) 使用规则(4)进行翻译，并将 X_1 的位置调整到最后

为此我们提出面向层次短语模型的词汇化调序方法。我们在规则上定义了四种调序关系，并使用变量所泛化的短语片段的边界词信息来估计调序关系的概率。在判别式模型[8]的框架下，我们训练若干词汇化调序模型并以特征的方式融入到基准系统中，以达到消解调序歧义的目的。在大规模语料的汉语到英语的翻译评测任务中，我们的方法在NIST 2003-2005 测试数据上获得了 0.6～1.2 BLEU[9]值的提高。

文章的组织结构如下：第2节简述词汇化调序模型的相关工作；第3节详细介绍面向层次短语模型的词汇化调序方法。首先定义调序关系；然后介绍词汇化调序模型；第4节描述实验设置和结果；最后是总结和展望。

2 相关工作

调序是机器翻译中非常重要的子问题之一，很多研究者对这个问题进行了深入的研究。常见的词汇化调序模型主要出现在短语模型[4-7]中，包括基于词[5-6]、基于短语[4]、基于层次化短语[7]的调序。特别注意的是，这里所说的基于层次化短语的调序，并非指该模型使用了类似层次短语模型的SCFG规则进行调序，而是该模型解码时通过移近—规约的算法以支持任意长度短语的调序。根据当前短语对与前面短语对之间的位置关系，这些模型使用三种调序方向：单调(Monotone)；交换(Swap)；非连续(Discontinuous)。

如图2(a)所示，基于词的调序模型[5-6]分析位置(s－1，u－1)与(s－1，v＋1)的关系。因为(s－1，v＋1)没有词语对齐，bp 的调序方向定义为非连续。基于短语的调序[4]判断当前短语对与位于(s－1，v＋1)的临近短语对的关系，并将这里的调序方向定义为交换调序。层次化的短语调序模型[7]不需要限制短语的最大长度。这种方法将图2(b)定义为交换调序。

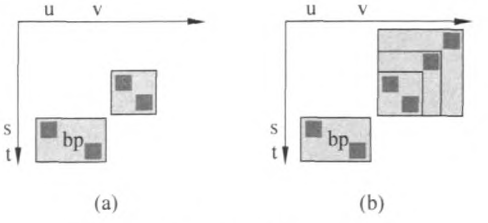

图2 短语模型中的词汇化调序

基于ITG模型的句法也使用词汇化信息[10]。该方法考虑正序和逆序两种调序规则，并将这两种调序转化为分类问题，而分类使用短语的边界词作为特征，并在词语对齐语料上训练最大熵分类器。基于规则类型选择(Rule Pattern Selection)的工

作[11-12]将词汇化信息引入到了层次短语中，文章[11]根据规则的类型（包括变量数目，是否调序）进行分类，分类时使用词汇化的信息。直接应用这种方法到层次短语系统中将导致大量的分类器[11]，文章[12]将规则简化为七种源端类型、17 种目标端类型。因此这种方法受限于规则的种类，难于扩展到规则类型复杂的翻译模型上（如句法模型）。本文所使用方法的不同之处在于，本文直接定义变量的调序方向，因而模型更为简单紧凑，并且不受规则类型数量的影响。

3　面向层次短语模型的词汇化调序方法

3.1　基于层次短语规则的调序关系

由于翻译模型的不同，短语模型的调序方向不能直接用于层次短语翻译中。这里首先举例说明本文使用的调序方向。层次短语的调序都是通过规则完成，我们只需定义规则的调序就能直接获得翻译的调序。本文的调序仅考虑变量与前后词语的位置关系，因此只需分析泛化规则。这主要是由于短语规则本身就是完全词汇化，其调序并没有歧义性。同时本文仅区分单调（Monotone）及交换（Swap）两种调序方向。单调调序是指词语在中文端英文端的相对位置并不改变，而交换调序则表示相对位置发生了变化，如图 3 所示。图 3 是使用规则（5）进行翻译时的调序情况。变量 X_1 与其前面的短语对（布什，Bush）在中文端和英文端的位置关系都是一致，因此是单调调序。而 X_1 相对于短语对（举行，held）则发生变化。虽然 X_1 在中文端位于“举行”之前，然而在英文端 X_1 的位置被调整到了“held”的后面，因此这里的调序是交换调序。同样的 X_2 与“举行”则是单调调序。具体的从图上来看，如果变量的对齐连线与词语的对齐连线并不交叉，则两者的调序关系是单调，否则为交换调序。

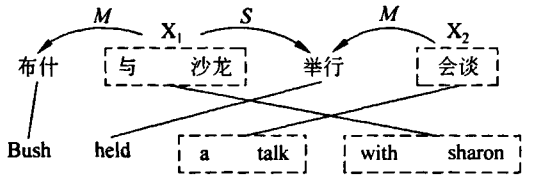

图 3　使用规则（5）进行翻译的调序情况

注：箭头弧线上的字母 M，S 分别是 Monotone，Swap 的首字母缩写

形式化上，给定一个规则及其词语对齐关系 a＝{(i, j)}，根据变量与源端邻接词语的位置关系，我们定义了四种调序关系。需要注意的是，这里的对齐包含变量的对齐。变量必然是一对一的对齐，而中文端的一个词语则可以对应多个英文单词。变量的对齐表示为 (i^x, j^x)；源端词语的对齐为 $a^w=\{(i^w, j^w)\}$；源端邻接单词用 W 表示。我们考虑两种情况，前一个词（－1），后一个词（＋1）；调序方向用 O 表示，共两种情况单调（M）和交换（S）。根据 W、O 的取值情况，我们定义四种调序关系：

- （W＝－1，O＝M）：$i^w<i^x$ *and* $\forall(i^w,j^w)\in a^w: j^w<j^x$；
- （W＝－1，O＝S）：$i^w<i^x$ *and* $\exists(i^w,j^w)\in a^w: j^w>j^x$；
- （W＝＋1，O＝M）：$i^w<i^x$ *and* $\forall(i^w,j^w)\in a^w: j^w>j^x$；
- （W＝＋1，O＝S）：$i^w<i^x$ *and* $\exists(i^w,j^w)\in a^w: j^w<j^x$。

3.2　词汇化调序模型

我们通过变量所泛化短语的边界词估计上述四种调序关系的概率。一个变量在中文端和英文端共有四个边界词分别是：中文端左边界词（flb）；中文端右边界词（frb）；英文端左边界词（elb）；英文端右边界词（rlb）。如图 3 中变量 X_1 四个边界词分别是：flb＝与；frb＝沙龙；elb＝with；erb＝sharon。我们按照如下方式估计变量 X1 与前一个词“布什”的调序的概率：

P(O | W＝－1，flb＝与，frb＝沙龙，elb＝with，erb＝sharon)

类似的，X1 与“举行”、X2 与“举行”调序的概率则为：

P(O | W＝＋1，flb＝与，frb＝沙龙，elb＝with，erb＝sharon)

P(O | W＝－1，flb＝会谈，frb＝会谈，elb＝a，erb＝talk)

给定变量的边界词以及变量与邻接词语的调序方向。我们估计参数 P(O | W，flb，frb，elb，erb)。这组参数同时考虑中文端和目标端的边界词语。到目前为止，我们仅说明了规则上的调序情况。完整的词汇化模型是定义在推导上的。所谓的推导就是生成翻译结果的整个规则序列。这里的调序模型区分邻接词语的位置，因此共有两个调序模型。具体定义如下。

- 与左边邻接词语的调序模型：

$$P_{-1}=\prod_{r\in D}\prod_{x\in r}P(O\mid W=-1,\mathrm{flb},\mathrm{frb},\mathrm{elb},\mathrm{erb})$$

• 与右边邻接词语的调序模型：

$$P_{+1}=\prod_{r\in D}\prod_{x\in r}P(O\mid W=+1,\mathrm{flb},\mathrm{frb},\mathrm{elb},\mathrm{erb})$$

其中D表示推导，r表示规则，x表示r中的变量。

我们还估计其他两组类似的参数P(O|W, flb, frb)；P(O|W, elb, erb)，这两组参数仅依赖与中文端的边界词或者英文端的边界词。这两组参数可以看作对P(O|W, flb, frb, elb, erb)的回退。因为这三组参数十分相似，在下文中我们仅说明第一组参数。

概率估计。我们使用词语对齐的语料来估计参数。当一个规则被抽取的时候，同时记录各种调序实例，以此来估计参数。如图3所示，当规则(5)被抽取的时候，记录如下的实例：

• O=M, W=−1, flb=与, frb=沙龙, elb=with, erb=sharon

• O=S, W=+1, flb=与, frb=沙龙, elb=with, erb=sharon

• O=S, W=−1, flb=会谈, frb=会谈, elb=a, erb=talk

显然这些实例能够直接在规则抽取过程中同时被抽取。这里不再赘述。当抽取完实例后，就可以直接用这些实例来估计上述的条件概率。我们使用加0.1的平滑来估计参数。

解码时使用参数。在对数线性模型的框架下，我们将词汇化调序概率的对数值以特征的方式加入到基准系统中。参考Moses系统[6]，我们也把每一种调序情况当作一个特征加入到系统中，即单调调序与交换调序各自作为一个特征，这里举例说明与左边邻接词语的调序模型中交换调序特征分数的公式：

$$f^{s}_{+1}=\sum_{r\in D}\sum_{x\in r}\log P(O=S\mid W=+1,\mathrm{flb},\mathrm{frb},\mathrm{elb},\mathrm{erb})$$

考虑到有四种调序关系，每种关系有三组估计的方式，因此我们共添加了3×4=12个特征到基准系统中。在抽取规则的过程中，保留规则的词语对齐信息，以便估计这些概率。当一个规则有多种对齐时，我们选择最常见的一个。解码过程中可能出现未发生事件。当出现这种情况时，我们认为两种调序方向是等概率的，概率值都为0.5。特别注意的是，解码过程中还将使用到粘贴规则($X\rightarrow <X_1\ X_2, X_1\ X_2>$)。这个规则确定性的使用单调调序，因此并不估计这个规则中变量的调序。

4 实验

4.1 数据

实验在汉语—英语方向上的翻译进行。所使用的语料如下。

• 双语语料。约155万平行句对。这些句对来自LDC语料的部分子集，包括：LDC2002E18；LDC2003E07；LDC2003E14；LDC2004T07 Hansards部分；LDC2004T08以及LDC2005T06。双语语料用于抽取规则和训练词汇化调序模型。我们首先使用GIZAC++[13]工具获得汉英、英汉两个方向的词语对齐，然后使用grow-diag-final-and[3]的启发式方法获得多对多的词语对齐。规则抽取[2]及词汇化模型训练在多对多的词语对齐数据上进行。

• 单语语料。包含GIGAWORD语料的新华部分，包含约2.38亿的英语单词。我们使用SRILM[14]工具训练四元的语言模型，使用Kneser-Ney平滑估计参数。

• 评测语料。使用NIST2002年的评测语料(NIST02)作为开发集。2003～2005年的评测语料(NIST03-05)作为测试集。

我们使用最小错误率训练[15]方法来优化线性模型的参数。采用的评测指标是大小写不敏感BLEU-4。所使用的解码器是层次短语解码器的C++重实现版本。该解码器采用CKY方式进行解码，并使用Cube-Pruning的方法进行减值减少搜索空间。实验所使用的栈为100。

4.2 实验结果

表2是实验结果。从最后一行的涨幅中，可以清楚的看到词汇化调序特征能够稳定的提高翻译的效果，提高幅度从0.6到1.2个点，平均涨幅约0.9个点，实验表明本文的方法是有效的。这也说明变量泛化的子短语的词汇化信息有助于层次短语进行调序。虽然层次短语规则本身已经带有一定的词汇化信息，但是由于没有考虑子短语的信息，因此歧义比较大。我们的词汇化调序方法考虑了这一部分信息，因此一定程度上帮助系统在翻译过程中选择正确的调序方向。

表 2　测试数据上的实验结果

	NIST 03	NIST 04	NIST 05
基准系统	34.49	35.28	32.96
+词汇化调序特征	35.06	36.17	34.19
涨幅	+0.57	+0.89	+1.23

注：第二行是基准系统的结果。第三行是加入 12 个词汇化调序特征的效果。最后一行是改进的系统相对于基准系统的涨幅。

表 3　两个系统在 NIST05 测试集合上的结果

长度限制	基准系统	+调序	涨幅
8	32.66	33.64	+0.98
10	32.96	34.19	+1.23

表 4　不同参数组合对翻译结果的影响

	NIST 03	NIST 04	NIST 05
基准系统	34.49	35.28	32.96
+all	34.52	35.50	33.40
+all+src	34.73	35.79	33.72
+all+trg	34.51	35.96	33.83
+all+src+trg	35.06	36.17	34.19

注：all 表示参数 P(O|W,flb,frb,elb,erb)；src 表示参数 P(O|W, flb,frb)；trg 表示参数 P(O|W, elb,erb)。

为了平衡翻译速度和质量间的关系，层次短语模型对于规则(仅指从语料中抽取的规则)所能覆盖的中文端长度进行了限制，传统的设置为 10。对于超过长度限制的短语都采用顺序翻译。表 3 比较了不同长度限制下，基准系统及加入词汇化调序模型的系统在 NIST05 测试集上的结果。使用更大的长度限制，潜在的调序空间也就更大。使用更大的长度限制，两个系统的 BLEU 值都有所增加，同时涨幅也就更大。这说明词汇化调序特征在更大的调序空间有更好的效果，也进一步验证了词汇化信息对调序的重要性。

在文中我们使用了三种类型的参数 P(O|W, flb,frb,elb,erb)、P(O|W, flb,frb)、P(O|W, elb, erb)，分别记为 all、src、trg。src、trg 参数分别使用源端和目标端对 all 进行回退。一个有趣的问题是，哪种回退更为有效。为此我们比较了不同参数组合的翻译效果。如表 4 所示，仅使用 all 参数能够提高翻译效果(第 2 行)，但涨幅有限。当继续增加 src 或者 trg 参数都能进一步提高效果(第 3～4 行)，所以这两种参数都是有用的。特别注意到在 NIST03 上，trg 未能进一步提高效果，这可能与 MERT 权重优化有关。此外，trg 的效果通常比 src 好些，这说明目标端的信息比源端的信息更有用。最后，同时使用三种参数(第 5 行)效果最佳，这说明目标端与源端的信息都有用，它们的作用并不重叠。

5　总结与展望

本文提出了一种面向层次短语的词汇化调序方法，使用词汇化的信息帮助层次短语的调序，提高了翻译的质量。本文的方法并不需要额外的数据集，简单有效。将来我们希望将这些词汇化的信息也用到其他模型中，比如基于语言学句法的模型[16-17]。我们也希望使用其他的训练方法比如最大熵[10-11]来训练词汇化调序模型。最后希望将 ITG 逆序调序规则加入到层次短语翻译中，将通过词汇化模型来指导该规则的使用。

参考文献

[1] David Chiang. A hierarchical phrase-based model for statistical machine translation [C]//Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. 2005: 263-270.

[2] David Chiang. Hierarchical phrase-based translation [J]. Computational Linguistics. 2007, 33(2): 201-228.

[3] Philipp Koehn, Franz Joseph Och, Daniel Marcu. Statistical Phrase-Based Translation [C]//Proceedings of NAACL 2003. 2003.

[4] Christoph Tillman. A unigram orientation model for statistical machine translation [C]//Proceedings of HLT-NAACL 2004: Short Papers. 2004: 101-104.

[5] Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, et al. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation[C]//Proceedings of IWSLT 2005, 2005.

[6] Philipp Koehn, Hieu Hoang, Alexandra Birch, et al. Moses: Open Source Toolkit for Statistical Machine Translation [C]//Proceeding of ACL 2007, demonstration session. 2007.

[7] Michel Galley, Christopher D. Manning. A simple and effective hierarchical phrase reordering model [C]// Proceedings of EMNLP 2008. 2008: 848-856.

[8] Franz Josef Och, Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation [C]//Proceedings of the 40thAnnual Meeting of the Association for Computational Linguistics. 2002: 295-302.

(下转第 50 页)

[11] Hoifung Poon, Pedro Domingos. Joint Inference in Information Extraction[C]//Proceedings of the 22nd National AAAI Conference on Artificial Intelligence. Vancouver, British Columbia, Canada, 2007: 913 - 918.

[12] Hoifung Poon, Pedro Domingos. Joint unsupervised coreference resolution with Markov logic[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii, USA, 2008: 650-659.

[13] Ivan P. Fellegi, Alan B. Sunter. A Theory for Record Linkage[J]. Journal of the American Statistical Association, 1969, 64(328): 1183-1210.

[14] Alvaro Monge, Charles Elkan. The field matching problem: Algorithms and applications[C]//Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. 1996: 267-270.

[15] Steven N. Minton, Claude Nanjo, Craig A, et al. A heterogeneous field matching method for record linkage[C]//Proceedings of the 5th IEEE International Conference on Data Mining. Houston, Texas, USA, 2005: 314-321.

[16] Indrajit Bhattacharya, Lise Getoor. A Latent Dirichlet Model for Unsupervised Entity Resolution [C]//Proceedings of the Sixth SIAM International Conference on Data Mining. Bethesda, MD, USA. 2006: 47-58.

[17] Akiko Aizawa. A fast linkage detection scheme for multi-source information integration[C]//WIRI, Tokyo, Japan, 2005: 30-39.

(上接第 41 页)

[9] Kishore Papineni, Salim Roukos, Todd Ward, et al. Bleu: a method for automatic evaluation of machine translation [C]//Proceedings of ACL 2002. 2002.

[10] Deyi Xiong, Qun Liu and Shouxun Lin. Maximum Entropy Based on Phrase Reordering Model for Statistical Machine Translation [C]//Proceedings of ACL 2006,2006.

[11] Zhongjun He, Qun Liu, Shouxun Lin. Improving statistical machine translation using lexicalized rule selection[C]//Proceedings of EMNLP 2008, 2008.

[12] Zhongjun He, Yao Meng, Hao Yu. Maximum Entropy Based Phrase Reordering for Hierarchical Phrase-based Translation [C]//Proceedings of EMNLP 2010, 2010.

[13] Franz Josef Och, Hermann Ney. A systematic comparison of various statistical alignment models[J]. Computational Linguistics, 2004, 29(1): 19-51.

[14] Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit [C]//Proceedings of the 7th International Conference on Spoken Language Processing. 2002: 901-904.

[15] Franz Joseph Och. Minimum error rate training in statistical machine translation [C]//Proceedings of ACL 2003. 2003.

[16] Yang Liu, Qun Liu, Shouxun Lin. Tree-to-String Alignment Template for Statistical Machine Translation [C]//Proceedings of ACL 2006. 2006.

[17] Michel Galley, Jonathan Graehl, Kevin Knight, et al. Scalable Inference and Training of Context-Rich Syntactic Translation Models [C]//Proceedings of ACL 2006. 2006.