

基于规则的名词短语预调序¹

牟小峰 荀恩东

北京语言大学语言信息处理研究所 北京 100083

E-mail: mouxiaofeng@blcu.edu.cn, edxun@blcu.edu.cn

摘要: 短语预调序是提高机器翻译效果的有效手段。本文通过对英汉翻译中名词短语结构的分析, 总结出名词短语调序规则, 通过将这些规则进行短语预调序, 提高了最终的翻译效果。为了缓解专家规则覆盖范围太大的弊端, 本文从大规模双语语料中自动获取实例规则。通过使用专家规则与实例规则进行短语预调序, 进一步提高了英汉翻译的效果。

关键词: 短语预调序, 专家规则, 实例规则

Noun Phrase Pre-reordering based on Rules

Mou Xiaofeng, Xun Endong

Center of Language Information Processing, Beijing Language and Culture University, Beijing, 100083

E-mail: mouxiaofeng@blcu.edu.cn, edxun@blcu.edu.cn

Abstract: Phrase pre-reordering is an effective means to improve machine translation performance. We summarized some reordering rules of noun phrase through analyzing noun phrase structures in English-Chinese translation. And the experiment shows that translation is improved through the use of these rules. In spite of such improvement, expert rules are somewhat coarse. To deal with such defects, lots of reordering examples is collected. And the experiment shows that translation performance is improved, compared with that in previous experiment.

Keywords: Phrase Pre-reordering, Expert rules, Example rules

1 引言

短语预调序即根据目标语言的语序来调整源语言的短语语序, 使源语言的语序与目标语言的语序接近或相同, 经过调整以后, 源语言句子的语序一般会出现一定变化。例如:

- The five ministers also exchange views on international issue of common concern.
- The five ministers also on common concern international issue of exchange views.

给定源语言句子 $f_1^j = f_{1,2}, \dots, f_j, \dots, f_n$, 预调序函数 θ 会根据目标语的语序对 f 中的单词进行重排, 使得 $\theta(f_1^j) = f_1^j = f_{i_1}, \dots, f_{i_n}$ 。

预处理过程并没有直接将调序和解码融合在一起, 而是在解码之前, 根据目标语的语序调整源语言句子的语序。调整后的源语言句子的语序与目标语的语序接近或相同, 然后将调序后的结果送到解码器翻译。

当前主流的短语预调序研究主要在欧洲主要语言之间展开。本文率先对英汉翻译中的预调序进行了探索。

与一般预调序的研究不同, 本文并未完全从基于规则或基于统计进行短语预调序, 而是通过将规则与统计²结合起来进行短语预调序处理。

本文通过分入分析英汉句法结构的差异, 对名词短语中的“NP IN NP”结构的调序素进行了深入分析, 发现了一批指导短语调序的规则。这些规则包括数+量结构、名词短语中心词、代名词等等。实验结果显示, 相对于不进行任何预处理的实验结果, 这些规则在一

¹ 本文得到国家自然科学基金资助(项目号: 60973062)。

一定程度上提高了翻译结果和调序准确率。

由于专家规则比较有限，且覆盖面太宽，因此本文从双语句对中抽取了大量短语调序实例。这些短语调序实例通过双语词对齐和句法分析获得。通过使用调序规则和调序实例，进一步提高了翻译效果和调序准确率。

2 预调序研究综述

已有的预调序研究可以根据句法分析的程度分为基于词性[1]、基于组块[2]和基于深层句法分析[3]三种；也可以根据获取预调序知识的方式分为基于统计的知识获取[4]和基于专家规则[5]两种；还可以根据预调序结果的数目分为产生一个最优结果[5]和产生多个最优结果[6]两种。下面按照第二种分类标准详细介绍预调序研究。

2.1 基于自动方法的预调序

基于自动获取调序知识的预处理通过预先对双语语料进行统计分析，然后自动获得调序规则。前面介绍的许多方法都属于这种思路，只不过其中许多调序策略融入了解码过程之中，而预处理却是在解码前对输入句子的语序进行处理。正因为失去了解码过程中诸多特征的约束，因而在预处理中自动获取调序规则的思路往往需要双语句法分析的支持。有些方法仅仅对双语句子进行浅层分析即可[2]，有些方法需要对句子进行深层结构分析[4, 5]。自动获取调序知识的预处理方法有下列特点：

- 1) 需要对句子进行句法结构分析。句法分析是自动获取调序知识的前提。句法分析的程度存在差异，可能是浅层分析[2]，也可能是深层分析[4]；可能是短语结构分析[6]，也可能是依存分析[7]；可能仅仅对单语进行句法分析[6]，也可能进行双语句法结构分析[4]。
- 2) 必须进行单词对齐。除了对单语或双语进行句法分析以外，必须进行单词自动对齐或者短语对齐，这样才能捕捉双语句法结构的差异，进而学习到两种语言间的调序规则。
- 3) 获取的规则一般是语言学规则。

尽管基于自动方法获取预调序知识具有多种优点，但该方法存在如下不足：

- 1) 句法分析的基础——句法理论——不是为统计机器翻译而产生。
- 2) 句法分析结构和短语翻译模型之间存在矛盾。
- 3) 大规模真实文本的环境下句法分析的准确率有待提高。

2.2 基于专家规则的预调序

与基于统计方法的预调序思路不同，基于专家规则的预调序方法通过专家对语言现象进行归纳和总结获取预调序规则。基于专家规则的预调序有以下优点：

- 1) 专家规则往往比较灵活，有针对性。文献[3]利用人工总结德英翻译差异规则进行源语言句法分析树的预调整，这些调序位置主要体现在主句和从句动词、从句主语、动词小品词结构、不定式、否定式等等。
- 2) 专家规则往往比较抽象，泛化能力强。文献[5]针对汉英翻译中 VP³、NP 和 LCP 三种短语的调序规则进行了简单总结。通过对短语结构的分析，提炼得到一些短语调序规则，并利用这些简单的短语调序规则来处理复杂地短语调序问题。

尽管具有上述优点，基于专家规则的预调序方法也存在一些不足：

- 1) 专家规则往往比较细碎，覆盖面不够。
- 2) 专家规则的泛化能力差，往往不具有推广能力，随着语言的不同而不同。许多基于统计的调序方法都是通用的方法，如词汇化的调序思路、基于固定移动距离的调序方法等等。这些方法具有很强的推广能力，不受具体语言的影响。而专家规则往往与具体语言相联系，随着语言和翻译方向的不同而不同，例如英汉的短语调序规则和汉英的短语调序规则就不一样。
- 3) 一般只能得到一个预调序结果，缺乏错误恢复机制。家规则一般是确定性的规则，没有附带概率信息，因此难以产生 n-best 结果。因而一旦作出调序决

³ 本文的短语标记全部采用 Penn Treebank 的标记体系。

定, 则很难有其它手段弥补调序错误。

3 短语调序范围

在“Penn Treebank”中, 主要短语类型不超过十种, 但短语结构的数量非常多。在这些多种多样的英语短语结构中, 有些结构的语序与汉语译文的语序相同, 这些短语在翻译中不需要调序, 例如大部分英语“主谓宾”结构和汉语译文的语序相同, 除非宾语表示时间。大多数短语结构出现频率极低, 少数短语结构的出现频率很高, 例如“NP of NP”。考虑到英汉短语结构调序的差异和短语结构本身的出现可能性大小的不同, 本文仅仅关注简单名词短语。

所谓简单名词短语, 即短语内不嵌套复杂短语结构, 例如一般简单句、定语从句、名词从句和复杂动词短语, 其内部成分为名词短语、介词和简单动词短语。

它不同于基本名词短语, 基本名词短语是不嵌套包含短语的名词短语。英语基本名词短语在翻译为汉语后一般不会出现语序调整的现象, 少数情况除外, 例如“the western Indian state(印度西部省份)”。

高频的简单名词短语结构主要包括如下几种:

- 1) NP IN NP, 如 NP of NP、NP in NP、NP from NP、NP at NP 等等。
- 2) NP VBN IN NP, 如 NP VBN by NP、NP VBN from NP。
- 3) NP VBX⁴ NP, 如 NP VBN NP、NP VBG NP。
- 4) NP VBX PP, 如 NP VBN PP、NP VBG PP。

在上述短语结构中, 第2、3和4种短语结构的调序一般比较简单, 只要把后面两个子成分往前移。这3种短语结构调序的关键在于后面几个子成分的语序判定, 它们的关系判定属于动词短语类型。第一种短语结构在简单名词短语中所占的比例最大, 是简单名词短语调序的主要部分, 本文仅仅处理该类型短语的短语预调序问题。

“NP₁ IN NP₂”结构的调序方式主要包括:

- a) NP₁ IN NP₂
- b) NP₂ IN NP₁
- c) 其它方式

在“其它方式”中, 有两种调序方式的出现频率稍高。第一种是把 NP₁ 分成 NP_{1x} 和 NP_{1y} 两部分⁵, 把 NP_{1y} 与 NP_{2s} 的顺序交换。第二种是把 NP₂ 分成两部分 NP_{2x} 和 NP_{2y}, 交换 NP₁ 和 NP_{2x} 的顺序。

4 专家规则

“NP₁ IN NP₂”结构中的主要短语结构规则如下:

- NP₁ 为“数+量结构”, 调序方式为“NP₁ IN NP₂”。英语的数量结构比较多, 例如: “a piece of”, “a bag of”, “a batch of”, “a beam of”, “a block of” 等等, 除了这些明显的数+量结构以外, 还存在一些不明显的数量结构, 例如: “piles of”, “a mount of”, “a lot of” 等等。数量结构的识别并不容易。大量 NP₁ 中心词既可能是量词, 也可能为名词, 如 bag, block。
- NP₁ 表示数量, NP₂ 也以数量开始, 调序方式为“NP₂ IN NP₁”。前后 NP 均表示数量往往意味着 NP₁ 是 NP₂ 数量的一部分, 例如:
 - ◆ eight of the nine charges: 九项指控中的八项
 - ◆ five of the seven persons: 七个人中的五个
- NP₁ 表示数量, NP₂ 不表示数量, 调序方式为“NP₁ IN NP₂”。NP₁ 在翻译后往往直接限定 NP₂, 表示一种数量上的限定关系, 例如:
 - ◆ sixty percent of students: 百分之六十的学生
 - ◆ over 500 of its members: 五百名成员

⁴ VBX

⁵ 究竟从 NP 的何处切分, 这与具体短语相关。在本文的测试语料中, 大部分从第一个词后面分开, 少部分从第二个词后分开, 未见其它的分割情况。

- NP₁ 的中心词为指示代词和表数量的不定代词，调序方式为“NP₁ IN NP₂”。指示代词和表数量的不定代词数量有限，可以列举。指示代词包括 these、those 等等，例如：
 - ◆ these of students: 这些 学生
 - ◆ those of teachers: 那些 老师
 表数量的不定代词包括 all, any, another, both, each, every, either, every, few, little, many, much, no, none, neither, one, other, some 等等。
- NP₁ 的首词或者是代词所有格形式，或者表示数量，或者表示地点，则拆开 NP₁，判断 NP₁ 其它词与 NP₂ 的调序关系。这条规则是针对“其它调序方式”而言，例如：
 - ◆ Switzerland system of direct democracy: 瑞士 直接 民主 体系
 - ◆ their knowledge of credit scores: 他们 对 信用 分 的认识
 - ◆ several areas of the investigation: 几个 调查 领域
 在上述例子中，NP₁ 首词往往表示地名、代词所有格和数量，其它类型的情况比较少见。前三种类型相对而言比较容易识别，具有明显的特征或者可枚举。
- NP₂ 的首词是代词所有格形式，在这种情况下，拆开 NP₂，把代词所有格提前，再估计前后其它成分的调序方式，例如：
 - ◆ two of his friends: 他的 两位 朋友
 - ◆ three of her cats: 她的 三只 小猫
- NP₁ 的中心词为专名，且 NP₂ 的中心词也为专名，调序方式为“NP₁ IN NP₂”。在这种情况下，不进行短语调序，例如：
 - ◆ the Times of London: 伦敦时报
 - ◆ the Agricultural Bank of China: 中国农业银行
 不对专名的语序进行调整与短语翻译模型有关。第一，在短语翻译模型中专名往往以短语的形式出现。在这种情况下不需要调整语序。第二，在短语翻译模型下专名调整语序后的译文常常不正确，因为缺少上下文的约束和限制。
- NP₁ 的中心词表示与地点相关的词语，且 NP₂ 的中心词也表示地点，调序方式为“NP₁ IN NP₂”。与地点相关的词语不是指具体地名，而是指行政区域名的后缀，例如“省、州、市、地区”等等。在这种类型的名词短语中，前后子成分往往同指，表示同一地点，或者 NP₁ 限制说明 NP₂，例如：
 - ◆ the western Indian state of Gujarat: 印度 西部 省份 古杰拉特
 - ◆ the capital of Colombo: 首都 科伦
 - ◆ the southern region of Achaia: 南部地区 亚该亚

5 实例规则

此处的调序实例与机器学习中的基于实例的学习方法并不完全一样。实例规则是具体的词例知识，而基于实例的方法是一种机器学习模型。二者使用的方式也有一定区别。调序实例在使用中往往采用词例匹配的方式寻找调序实例，并将完全匹配的调序实例的调序方式作为测试短语的调序方式；而基于实例的方法往往根据相似性判断标准选择多个实例，并从中估计目标决策函数，并用此目标决策函数判断测试短语的调序方式。在一定程度上可以把调序实例的方法当作特殊的基于实例的方法。这种特殊性即“相似性”定义为相同实例，“目标决策函数”定义为以相同实例的短语调序作为输入短语的调序方式。

调序实例知识主要通过人工标注和自动获取得到。

人工标注的实例知识主要通过从带标树库中获取。带标树库通过人工标注得到，带有大量单词翻译、短语翻译和结构对齐等信息。由于英语短语结构和汉语译文并存在同一棵句法树上，因此短语结构的调序信息非常容易获取。从双语树库中获取的调序实例知识的特点在于：第一，调序信息的准确性比较高，歧义问题不明显；第二，调序信息的规模有限。由于树库中的短语调序信息的标注代价比较高，因而不能完全依赖双语树库来获取所有短语调序信息。

尽管无法利用人工标注获得大量树库标注信息，但可以利用自动标注工具自动构造双

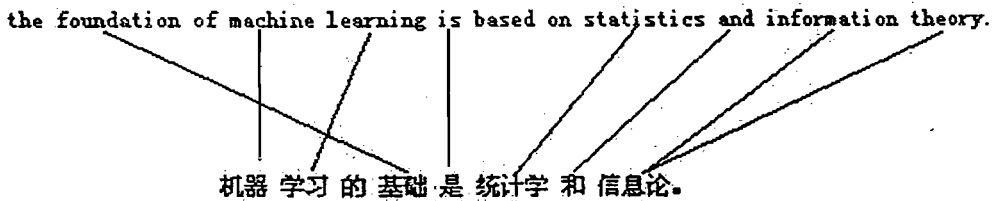


图 1

语树库。自动构造的双语树库在准确性上不如人工标注语料，但在数据的规模上要远远超过后者。

从双语语料中抽取 320 万英汉双语句对，并利用 GIZA++ 进行双向词对齐和英语句法分析。对双向对齐结果取交集以保证对齐准确率。调序实例通过英汉短语中对齐的词的相对位置变化得到，例如：给定如图 1 所示的双语句子对齐信息，根据英语句子的句法树可知，“the foundation of machine learning”是 NP₁ of NP₂ 结构。NP₁ 与 NP₂ 在英语句子中是顺序出现，而二者对齐的译文之间的语序为逆序。根据名词短语中心词的判断方法可知，“foundation”是 NP₁ 的中心词。在该短语中，调序实例为“foundation”，它的调序方式为“逆序”。

6 实验

1) 实验准备

“NP IN NP”按照介词的不同可以分为多种不同的短语结构，例如：“NP of NP”、“NP at NP”、“NP in NP”、“NP from NP”等等。在这些短语结构中，本文把“NP of NP”从“NP IN NP”中分离出来，自称一类。其它“NP IN NP”结构组成一类。进行这种区分的主要原因在于短语调序方式出现频率的差异很不一样。在“NP of NP”中，主要短语调序方式的出现比例非常接近，而在其它“NP IN NP”中，主要短语调序方式的出现比例很不均匀。

测试语料采用“NIST MT 08”的英汉语料，共 1859 句，包含四份参考译文。汉语语料采用北京语言大学 GPWS 分词系统进行分词。翻译系统采用 Google 在线翻译。采用 Google 在线翻译，原因有二：第一，Google 在线翻译的词汇量很大，可以很大程度上减小未登录词的影响；第二，Google 在线翻译是公开使用的系统，更适于公开评价。

在调序结果的评价上，除了 BLEU 和 NIST 两种指标之外，本文还采用调序准确率来评价短语调序的精度。为了准确地度量短语调序准确率，根据参考译文的语序对所有的待测试句子的短语手工进行短语调序方式的标注。并以此标注作为调序评价的参考答案。

调序准确率的各项评价指标如下：

- 整体的正确率和错误率(OR)。
- 出现调序的短语的正确率和错误率(RR)。
- 未调序的短语的正确率和错误率(NRR)。
- 不采用任何规则的情况下的正确率(ANRR)。

2) 基于专家规则的调序实验

在所有的评测语料中，共有 620 个句子出现了“NP of NP”结构，268 个句子的“NP of NP”结构发生了调序。调序之后的自动评价如下：

	NIST	BLEU
不调序	7.7949	0.2337
基于规则调序	7.8365	0.2440

表 1: NP of NP 的自动评价

从表 1 中可以看到，通过运用专家规则，自动评价的指标有了一定程度提高。这说明专家规则发挥了正面的调序效果。

调序准确率结果如下：

短语	OR	RR	NRR	ANRR
NP of NP	68.4	70.6	67.1	47.1

表 2: NP of NP 调序准确率

从表 2 中可以看到, 在不进行任何调序的情况下, “NP of NP” 结构的调序准确率为 47.1%。通过应用专家规则, 调序准确率上升到 68.4%, 准确率的提升幅度比较明显。

在评测语料中, 其它 “NP₁ IN NP₂” 结构出现总次数与 “NP of NP” 相近, 基于上面的专家规则进行实验, 自动评价结果如下:

	NIST	BLEU
不调序	8.2557	0.2301
基于规则调序	8.3446	0.2479

表 3: NP IN NP 的自动评价

调序准确率结果如下:

短语	OR	RR	NRR	ANRR
NP of NP	73.7	76.5	62.2	38.7

表 4: NP IN NP 调序准确率

“NP IN NP” 结构的调序准确率提高幅度比较高, 这与该结构偏向于调序有关。调序准确率的提升幅度几乎翻倍。尽管如此, 自动评价的提升幅度并不太明显。这与翻译系统有很大关系。

3) 基于专家规则和实例规则的实验

由于调序实例均是反例, 因此在使用时, 首先查看输入短语的特征是否在实例库中出现。如果在实例库中出现, 则按照实例的调序方式处理如果没有在实例库中出现, 则根据规则进行调序。把抽取的大量调序实例与其面的调序规则一起使用。

通过应用专家规则和实例规则, 对所有 “NP IN NP” 结构进行短语预调序, 自动评价指标如下:

	NIST	BLEU
不调序	8.2016	0.2267
基于规则调序	8.2893	0.2473

表 5: NP IN NP 的自动评价

从实验结果可以可以看到, 加入调序实例以后自动评价指标的提高比较明显, BLEU 指标从 0.2267 提高到 0.2473。在前面的实验中, BLEU 值的提高幅度分别为 1.1 和 1.7(按绝对值算), 在这里提高幅度为 2.1。这说明短语调序与实例的关系非常密切。

4) 调序错误举例

- 固定翻译难以处理。例如:
 - i. the assistant secretary of state: 助理国务卿
 - ii. [at] a record high of 22,159.52 points: 高达 22,159.52 点
- 短语存在调序歧义, 或者短语的组成部分的意义比较模糊。
 - i. the performance of its policy: 政策的 表现; 执行 政策
 - ii. the affection of their loved ones: 他们 亲人的 怀抱
- 调序正确, 翻译错误。
 - i. the terrorists of the FARC (原句子未调序): 恐怖主义分子的革命武装力量
 - ii. of the FARC the terrorists (原句子调序后): 在哥伦比亚革命武装力量恐怖分子

7 总结

从前面的实验可以看到, 短语预调序在一定程度上提高了短语调序的精度, 增强了系统的翻译表现。但是, 通过诸多实验发现, 短语预调序这种方法以及围绕该方法的研究中存在一些问题, 这些问题在一定程度上影响预调序的表现。

- 调序后的短语组合形式。在已有的预研究中, 几乎没有提到短语调序后的组合形式, 而这对最终翻译结果会有影响。如果调序节点限于两个, 则不存在组合顺序的问题, 如果短语调序节点大于等于 3 个, 则短语调序顺序就很重要了。例如: “the foundation of algorithm”, 调序后的词语组合形式如下:
 - ◆ of algorithm the foundation
 - ◆ algorithm of the foundation

◆ algorithm the foundation of

本文对所有该结构进行实验，并对三种短语组合形式进行比较后发现，第一种要优于后两种。这与短语翻译模型的本质相符。在短语翻译模型中，短语被定义为连续的词语串。因此，在上面的实验中，如果该结构出现调序，则 of 一般与后部结合形成短语。所以，第一种短语组合形式会召回更多的短语译文，从而提高最终的翻译效果。

- 短语预调序的系统差异。在实验中发现，不同翻译系统下预调序的翻译效果存在一定差异。比较好的翻译系统能够在很大程度上保证翻译的忠实度，这有助于获得较好的预调序效果。在翻译效果不好的翻译系统下，译文的忠实度无法得到保证，因此，即使调序正确，也同样无法提高自动评价的表现。

尽管短语预调序存在种种问题，但仍不失为一种提高短语调序表现的方法。本文通过分入分析英汉句法结构的差异，对名词短语中的“NP IN NP”结构的调序因素进行了深入分析，发现了一批指导短语调序的规则。实验结果显示，这些规则在一定程度上提高了翻译结果和调序准确率。由于专家规则比较有限，且覆盖面太宽，因此从双语句对中抽取了大量短语调序实例。通过使用调序规则和调序实例，进一步提高了翻译效果和调序准确率。

参 考 文 献

- [1] M. Popović and H. Ney. POS-based Word Reorderings for Statistical Machine Translation. In Proceedings of the 5th International Conference on Language Resources and Evaluation, 2006.
- [2] Y. Zhang, R. Zens and H. Ney. Chunk-Level Reordering of Source Language Sentences with Automatically Learned Rules for Statistical Machine Translation. In Proceedings of The Conference of the North American Chapter of the Association for Computational Linguistics, 2007.
- [3] M. Collins, P. Koehn and I. Kucerová. Clause Restructuring for Statistical Machine Translation. In Proceedings of the 43rd Annual Conference of the Association for Computational Linguistics, 2005.
- [4] F. Xia and M. McCord. Improving a Statistical MT System with Automatically Learned Rewrite Patterns. In Proceedings of the 42nd Annual Conference of the Association for Computational Linguistics, 2004.
- [5] C. Wang, M. Collins and P. Koehn. Chinese Syntactic Reordering for Statistical Machine Translation. In Proceedings of joint conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007.
- [6] C. Li, D. Zhang, M. Li, M. Zhou, M. Li and Y. Guan. A Probabilistic Approach to Syntax-based Reordering for Statistical Machine Translation. In Proceedings of the 45th Annual Conference of the Association for Computational Linguistics, 2007.
- [7] N. Habash. Syntactic Preprocessing for Statistical Machine Translation. In Proceedings of MT SUMMIT XI, 2007.
- [8] E. Charniak. A Maximum-Entropy-Inspired Parser. In Proceedings of the 1st meeting of the North American Chapter of the Association for Computational Linguistics, 2000.