统计机器翻译的问题与思考

张家俊
中科院自动化研究所

1, 句法翻译模型的发展

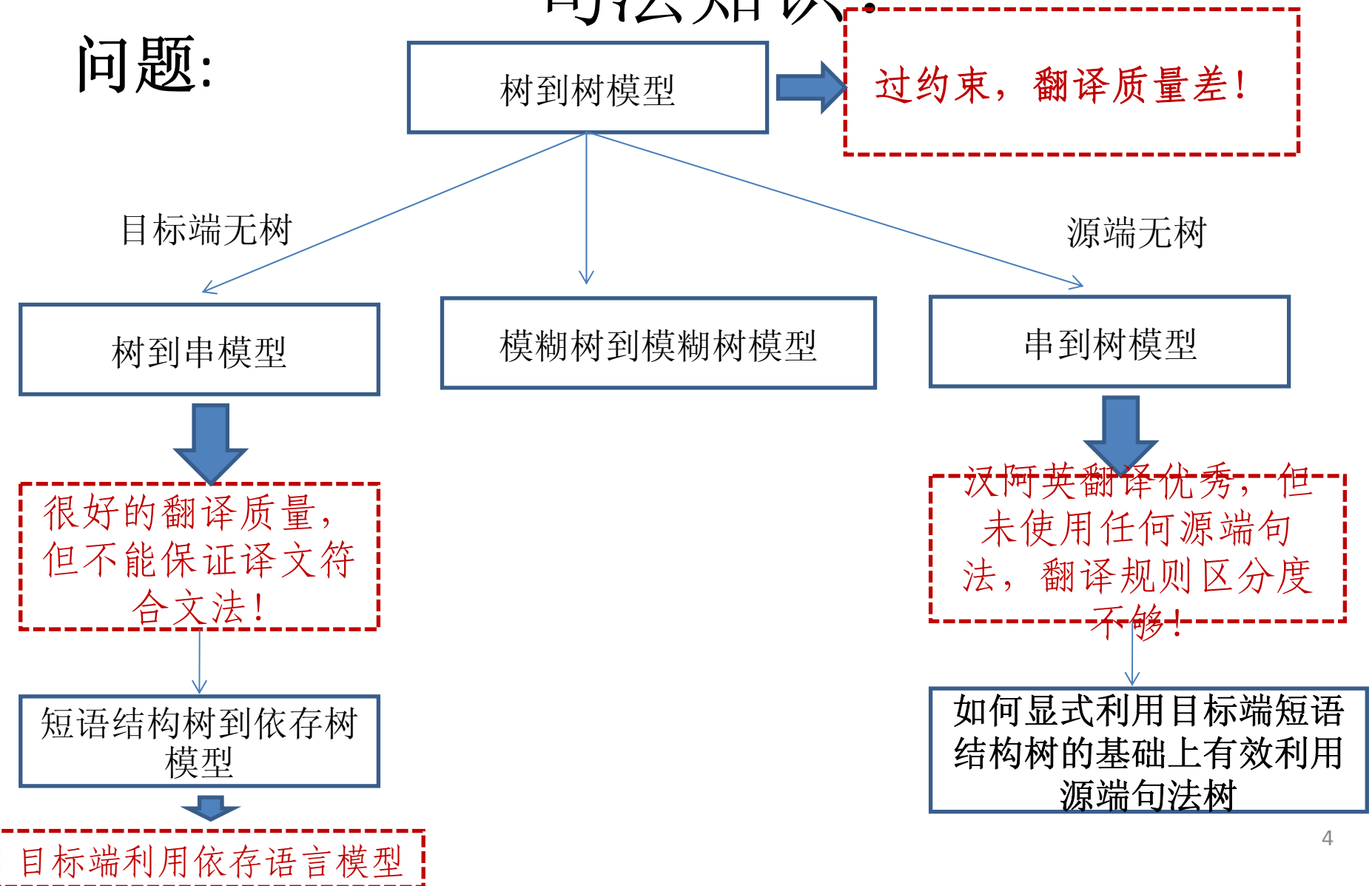
2, 非平行语料的翻译知识获取

1, 句法翻译模型的发展

- 句法翻译模型的问题（局限性）
  - 如何有效地同时利用源端与目标端的句法知识？
  - 如何解决规则中非终结符的过分泛化现象？
  - 如何跨越句法分析模型与翻译模型之间的鸿沟？

如何有效地同时利用源端与目标端的句法知识？

问题：



如何有效地同时利用源端与目标端的句法知识？

- 我们采用的方法:
  - 源语言端采用模糊算法使用句法知识
  - 目标语言端精确使用句法知识

- 实验结果表明这种方法可以明显地优于传统的串到树翻译模型

如何有效地同时利用源端与目标端的句法知识？

- 一点想法:
  - 由于句法分析的影响，在实际中如何使用两端的句法信息很大程度上取决于翻译语言对以及翻译方向

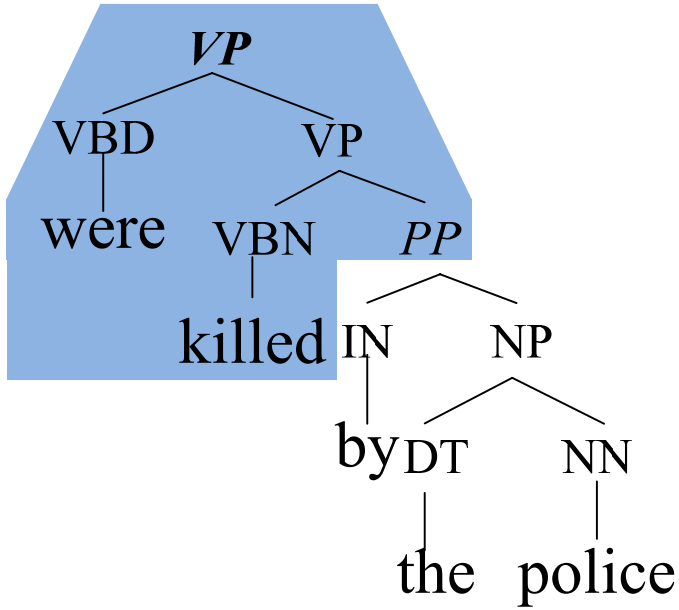
如何解决规则中终结符的过分泛化现象？

问题:

- 规则中的非终结符在翻译中起着泛化的作用，但往往会带来很多歧义问题，在解码过程中构建目标语言树的上层结构时尤其明显

$VP \rightarrow (\ x_0$ 击毙, VBD(were) VP(VBN(killed) PP:x_0))

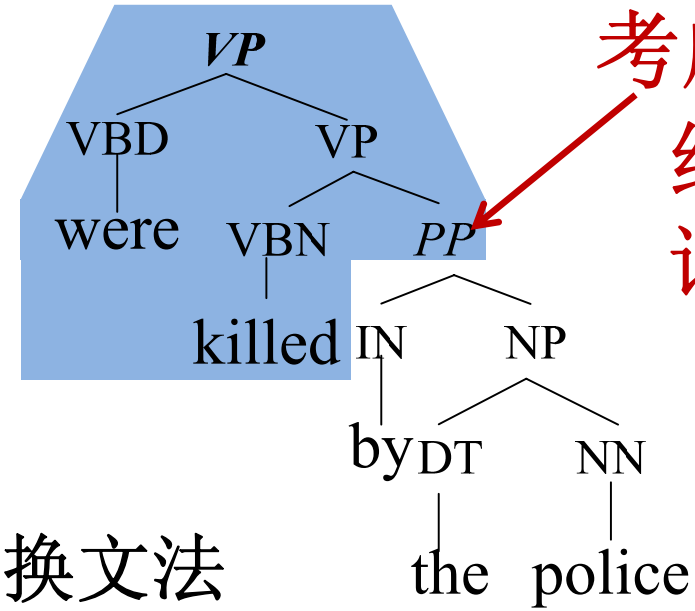
被 警方 击毙 ：



如何解决规则中终结符的过分泛化现象？

VP → (x_0 击毙, VBD(were) VP(VBN(killed) PP:x_0))

被 警方 击毙 ：



考虑生成PP结点的双语词汇化信息

双语词汇化的同步树替换文法

1,生成式模型
2,判别式模型

取得显著的质量提升

如何解决规则中终结符的过分泛化现象？

- 一点想法
  - 词汇化增加了解码的复杂度，如何找到一个平衡点是个难题
  - 句法标记的细化 vs. 词汇化

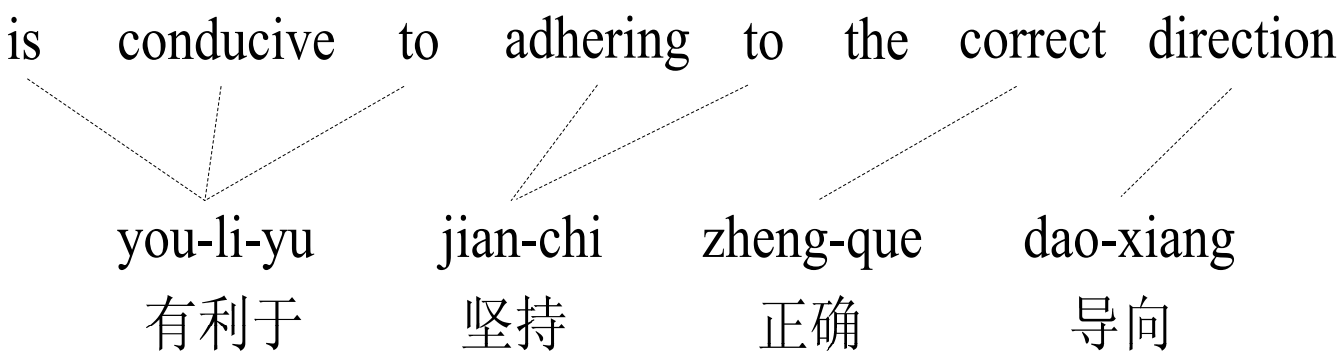
如何跨越句法分析模型与翻译模型之间的鸿沟？

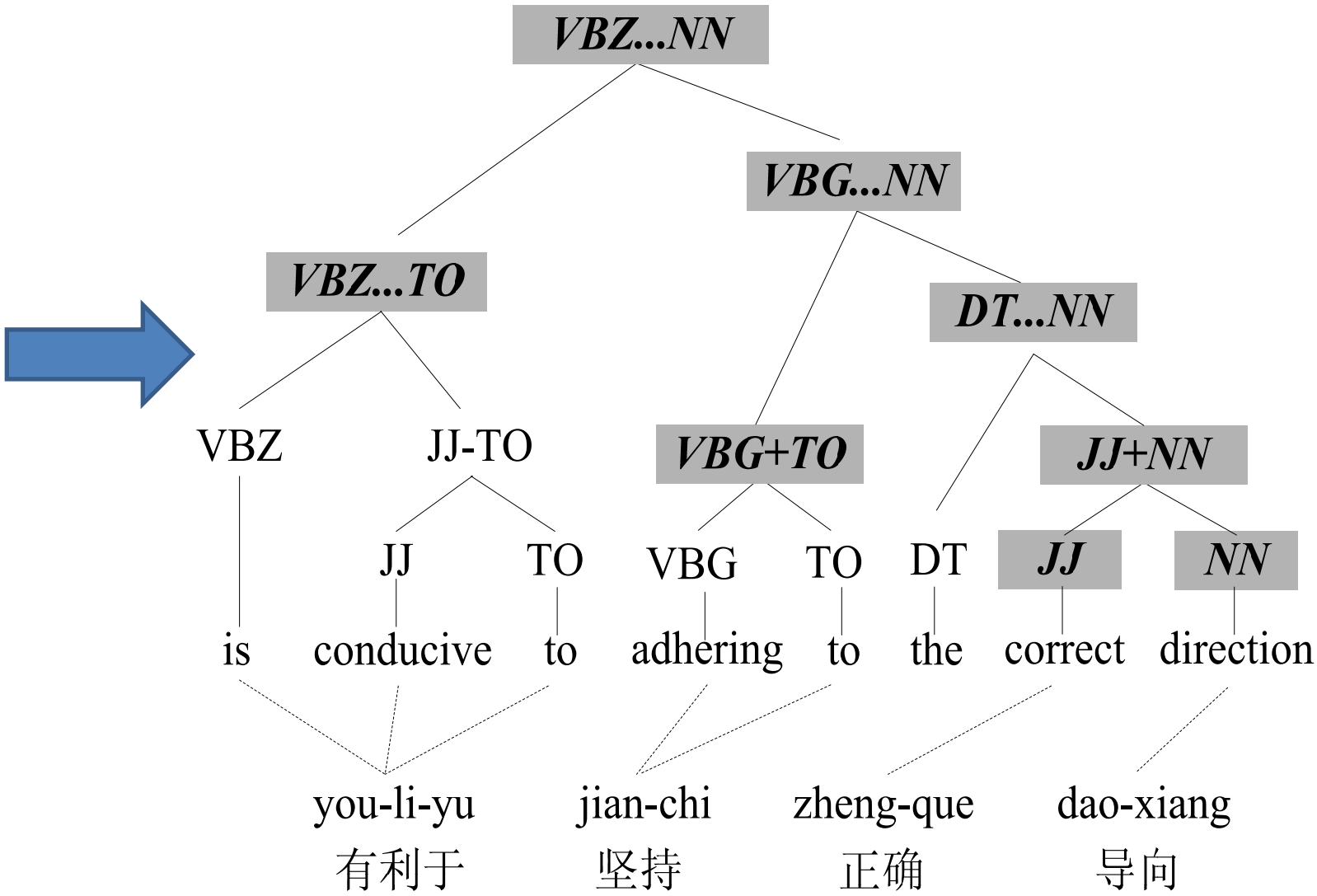
- 普遍的矛盾
  - 句法树由单语句法分析器产生
  - 词语对齐由平行语料学习获得
  - 两者之间不兼容
- 一些解决方法
  - 优化或者转化句法树结构
  - 根据句法树优化词对齐

我们采用的方法

- 以串到树翻译模型为例
  - 以词对齐的双语语料为输入，无监督地学习兼容词对齐信息的目标语言句法树结构
  - 借助了Bayesian模型和Gibbs采样

一个例子


is conducive to adhering to the correct direction
you-li-yu
jian-chi
zheng-que
dao-xiang
有利于
坚持
正确
导向



在串到树翻译模型上非常显著地优于传统句法树方法！

如何跨越句法分析模型与翻译模型之间的鸿沟？

- 一点想法
  - 句法翻译模型可以不依赖句法分析器
  - 是否可以同时学习词对齐信息与句法树结构

2, 非平行语料的翻译知识获取

- 绝大多数情况下平行语料无法获得
  - 新的领域
  - 新的语言对
- 典型的情形是源语言或目标语言的单语语料大规模存在
  - 我们能否直接从非平行语料中学习翻译模型？

目标

- 词典学习 (*)
- 翻译规则学习 (**)

翻译规则学习

词典

源端与目标端的大规模单语语料

翻译规则

请教大家的问题

- 句法翻译模型走向何方？如何迈向语义及如何建立基于语义的翻译模型？

- 如何从非平行语料中学习翻译模型？

谢谢大家!