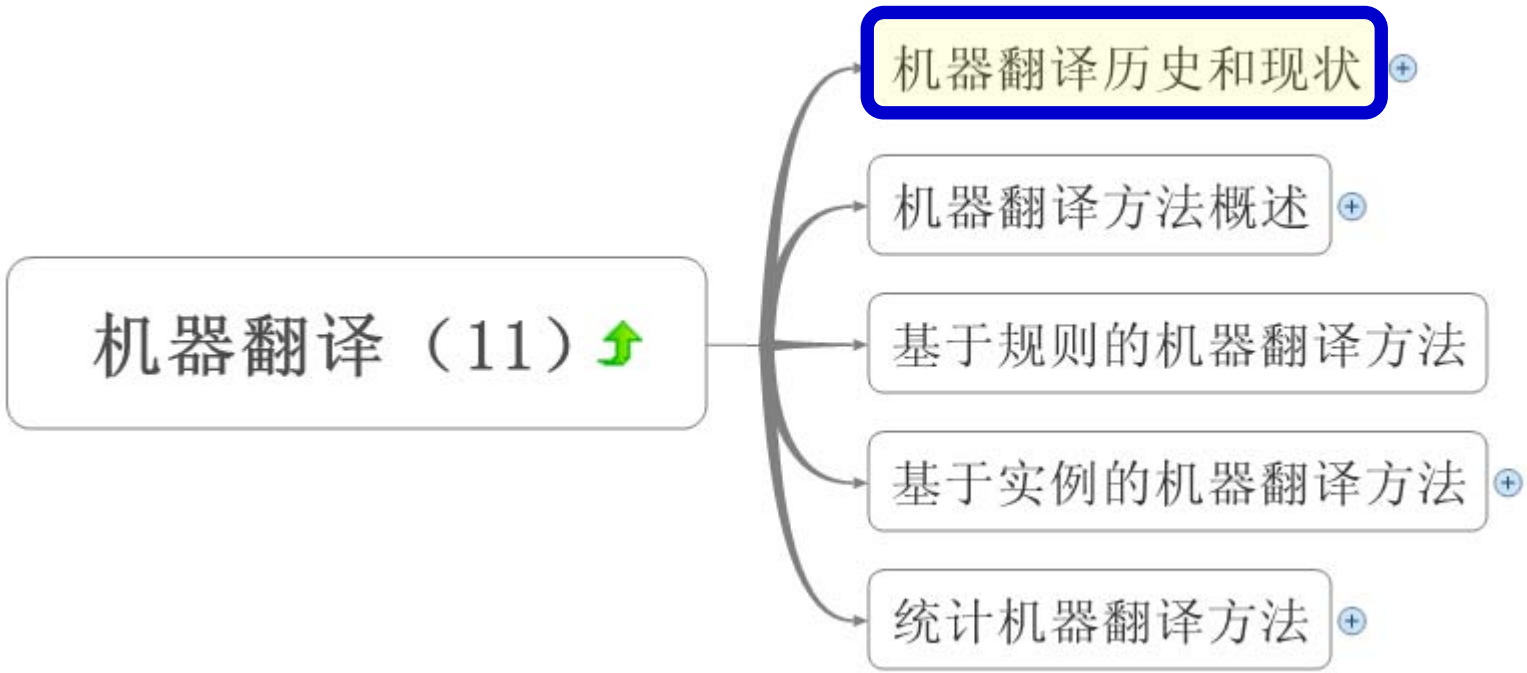
计算语言学

第11讲 机器翻译

刘群
中国科学院计算技术研究所
liuqun@ict.ac.cn

中国科学院研究生院2010年春季课程讲义

内容提要


机器翻译（11）
机器翻译历史和现状
机器翻译方法概述
基于规则的机器翻译方法
基于实例的机器翻译方法
统计机器翻译方法

机器翻译的历史

- W. J. Hutchens, latest Development in MT Technology: Beginning a New Era in MT Research. In : Proceedings of Machine Translation Summit-IV, Kobe, Japan, 1993.
- 冯志伟，自动翻译，上海知识出版社，1987年。
- 冯志伟，自然语言机器翻译新论，语文出版社，1994年。
- 冯志伟，自然语言的计算机处理，上海外语教育出版社，1996年。

机器翻译的历史

- 萌芽期（17世纪-1930年代）
- 草创期（1946-1964）
- 萧条期（1964-1960年代后期）
- 复苏期（1970年代初期）
- 繁荣期（1970年代后期-1980年代初期）
- 平台期（1980年代后期-1999年）
- 再度繁荣期（1999-现在）统计方法！

内容提要


机器翻译（11）
机器翻译历史和现状
机器翻译方法概述
基于规则的机器翻译方法
基于实例的机器翻译方法
统计机器翻译方法

机器翻译方法概述

- 机器翻译应用系统类型
- 机器翻译方法分类（按转换层面划分）
- 机器翻译方法分类（按知识表示划分）

机器翻译应用系统类型 (1)

- 理想的机器翻译
  - 全自动高质量，FAHQ MT
    Full Automatic High Quality Machine Translation
- 按人机关系分类
  - 全自动机器翻译，FAMT
    Full Automatic Machine Translation
  - 人助机译，HAMT
    Human Assisted Machine Translation
  - 机助人译，CAT
    Compute-Aided Translation

机器翻译应用系统类型(2)

- 按应用方式分类
  - 信息分发型 MT for dissemination
    - 要求高质量，不要求实时
    - 采用人机互助，或者受限领域、受限语言等方式提高翻译质量
  - 信息吸收型 MT for assimilation
    - 不要求高质量，要求方便、实时
    - 翻译浏览器、便携式翻译设备、……

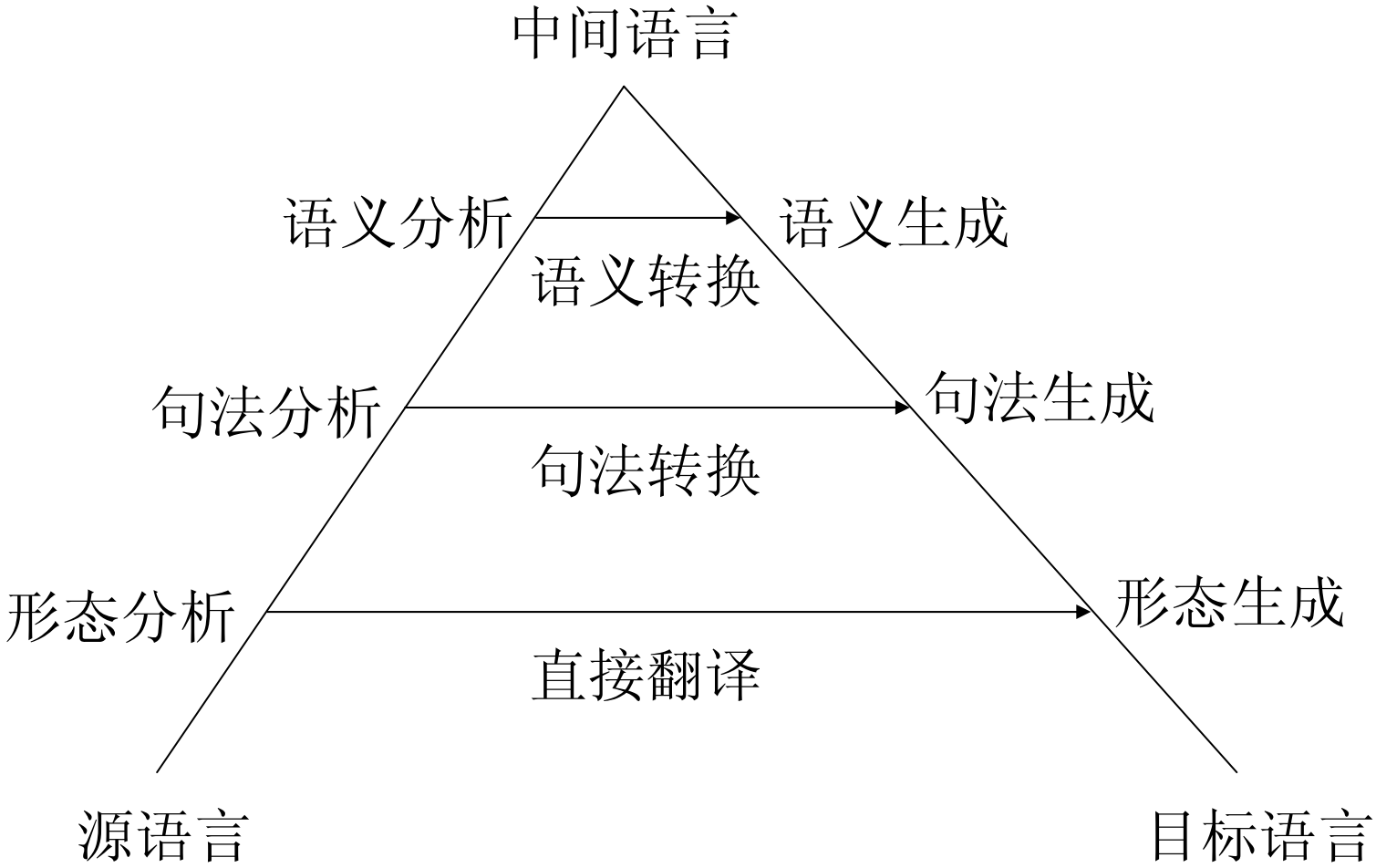
机器翻译应用系统类型(3)

- 按应用方式分类（续）
  - 信息交流型 MT for interchange
    - 不要求高质量，通常要求实时，语言随意性较大
    - 语音翻译、网络聊天翻译、电子邮件翻译
  - 信息存取型 MT for information access
    - 将机器翻译嵌入到其他应用系统中
    - 跨语言检索、跨语言信息抽取、跨语言文摘、跨语言非文本数据库的检索……

机器翻译方法概述

- 机器翻译应用系统类型
- **机器翻译方法分类（按转换层面划分）**
- 机器翻译方法分类（按知识表示划分）

机器翻译方法分类(按转换层面划分)



直接翻译方法

- 通过词语翻译、插入、删除和局部的词序调整来实现翻译，不进行深层次的句法和语义的分析，但可以采用一些统计方法对词语和词类序列进行分析
- 早期机器翻译系统常用的方法，近期IBM提出的统计机器翻译模型也可以认为是采用了这一范式
- 著名的机器翻译系统Systran早期也是采用这种方法，后来逐步引入了一些句法和语义分析

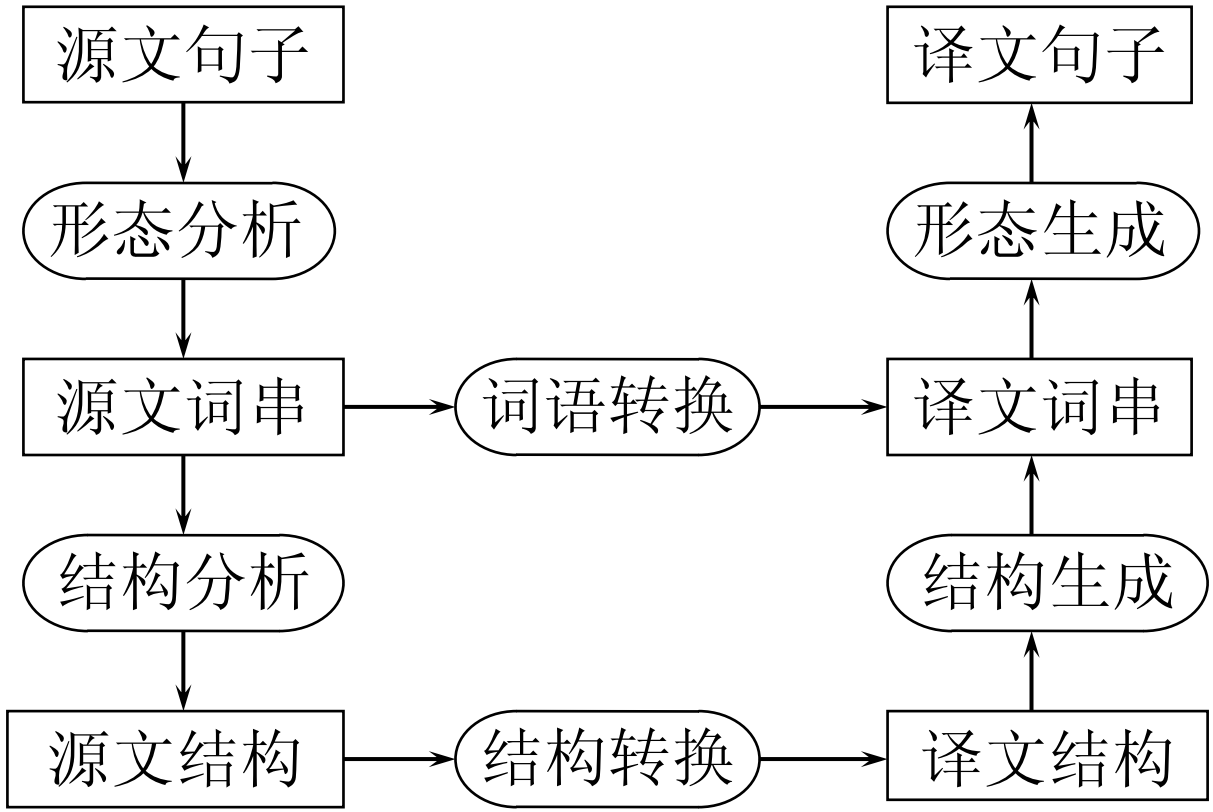
转换方法 (1)

- 整个翻译过程分为“分析”、“转换”、“生成”三个阶段；
- 分析：源语言句子➔源语言深层结构
  - 相关分析：分析时考虑目标语言的特点
  - 独立分析：分析过程与目标语言无关
- 转换：源语言深层结构➔目标语言深层结构
- 生成：目标语言深层结构➔目标语言句子
  - 相关生成：生成时考虑源语言的特点
  - 独立生成：生成过程与源语言无关

转换方法 (2)

- 理想的转换方法应该做到独立分析和独立生成，这样在进行多语言机器翻译的时候可以大大减少分析和生成的工作量；
- 转换方法根据深层结构所处的层面可分为：
  - 句法层转换：深层结构主要是句法信息
  - 语义层转换：深层结构主要是语义信息
- 分析深度的权衡
  - 分析的层次越深，歧义排除就越充分
  - 分析的层次越深，错误率也越高

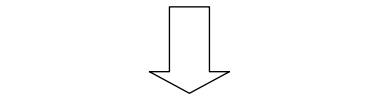
转换方法 (3)



基于转换方法的翻译流程

句法层面的转换方法 (1)

她把一束花放在桌上。⇨ She put a bunch of flowers on the table.

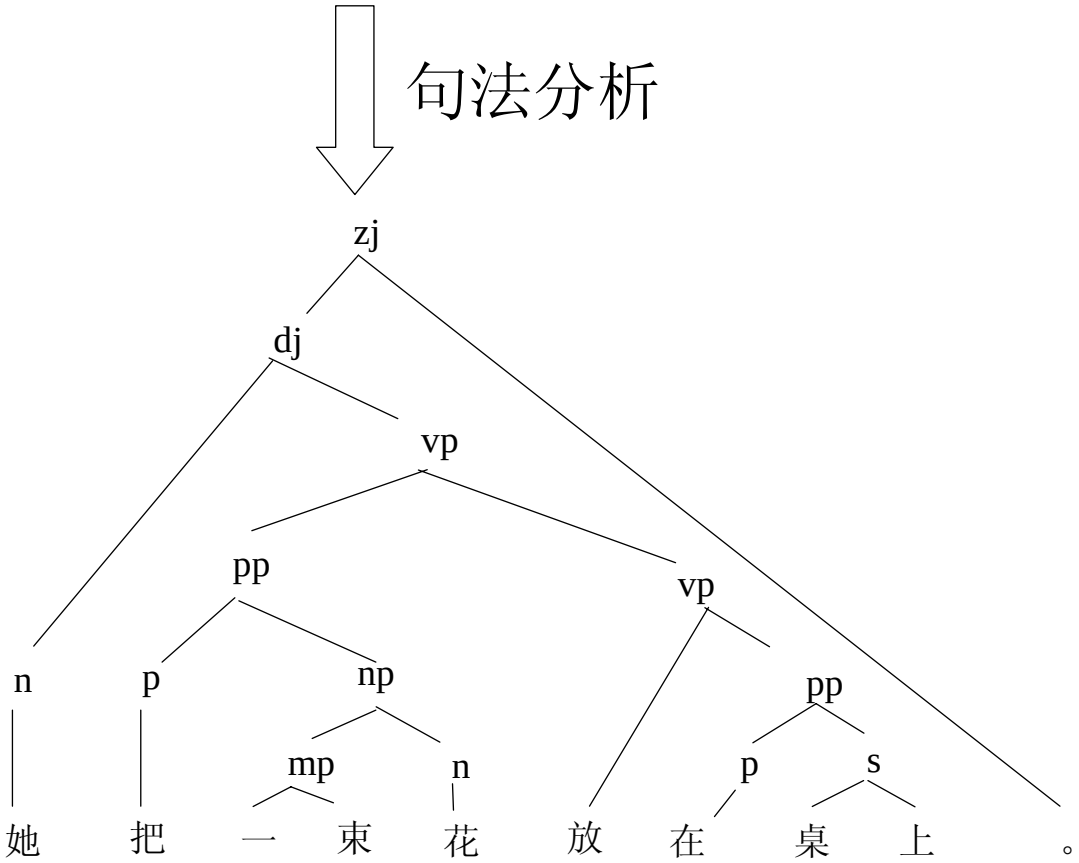


她/r 把/p-q-v-n 一/m-d 束/q 花/n-v-a 放/v 在/p-d-v 桌/n 上/f-v 。/w

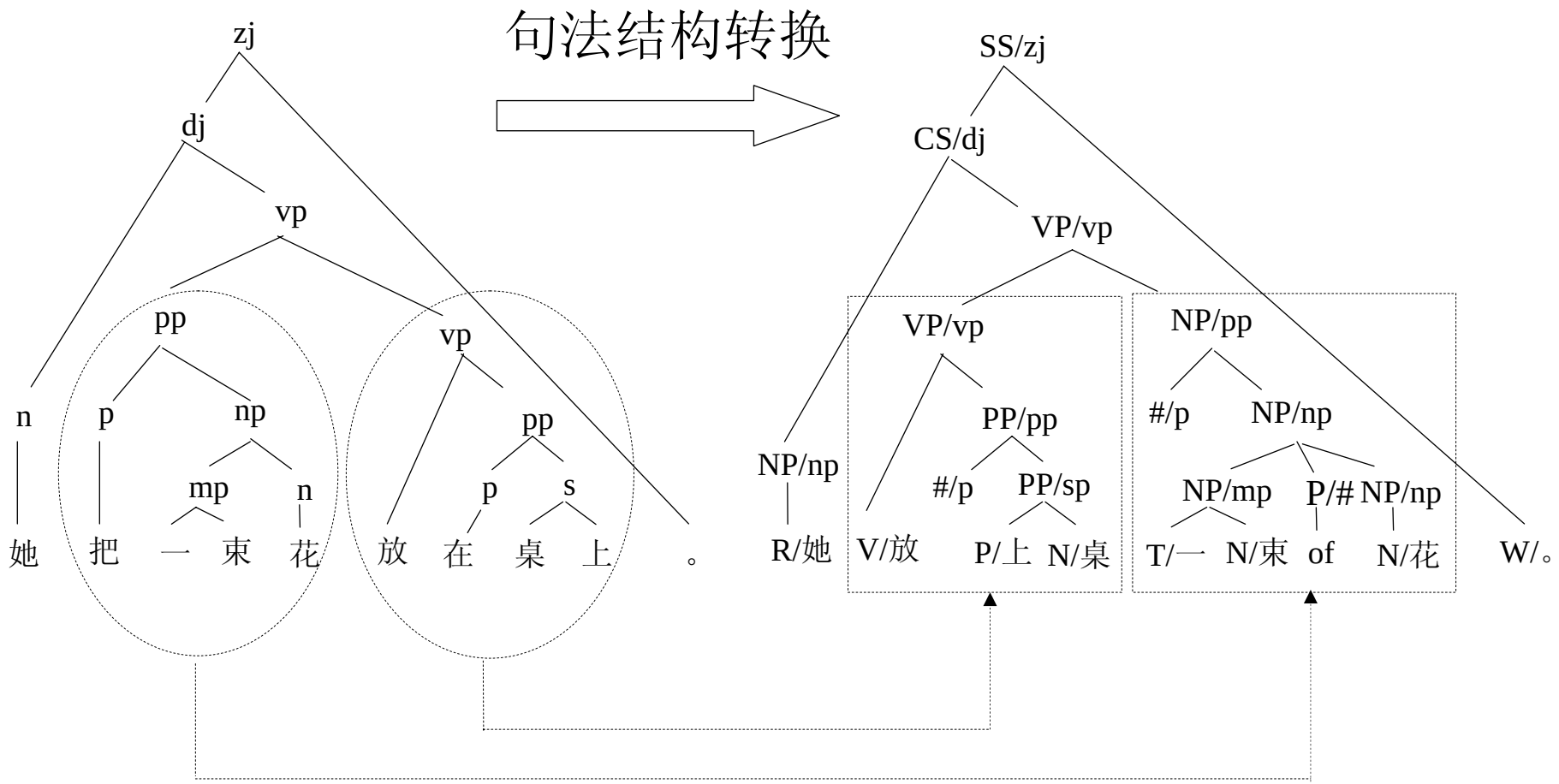
她/r 把/p 一/m-d 束/q 花/n 放/v 在/p-v 桌/n 上/f-v 。/w

句法层面的转换方法 (2)

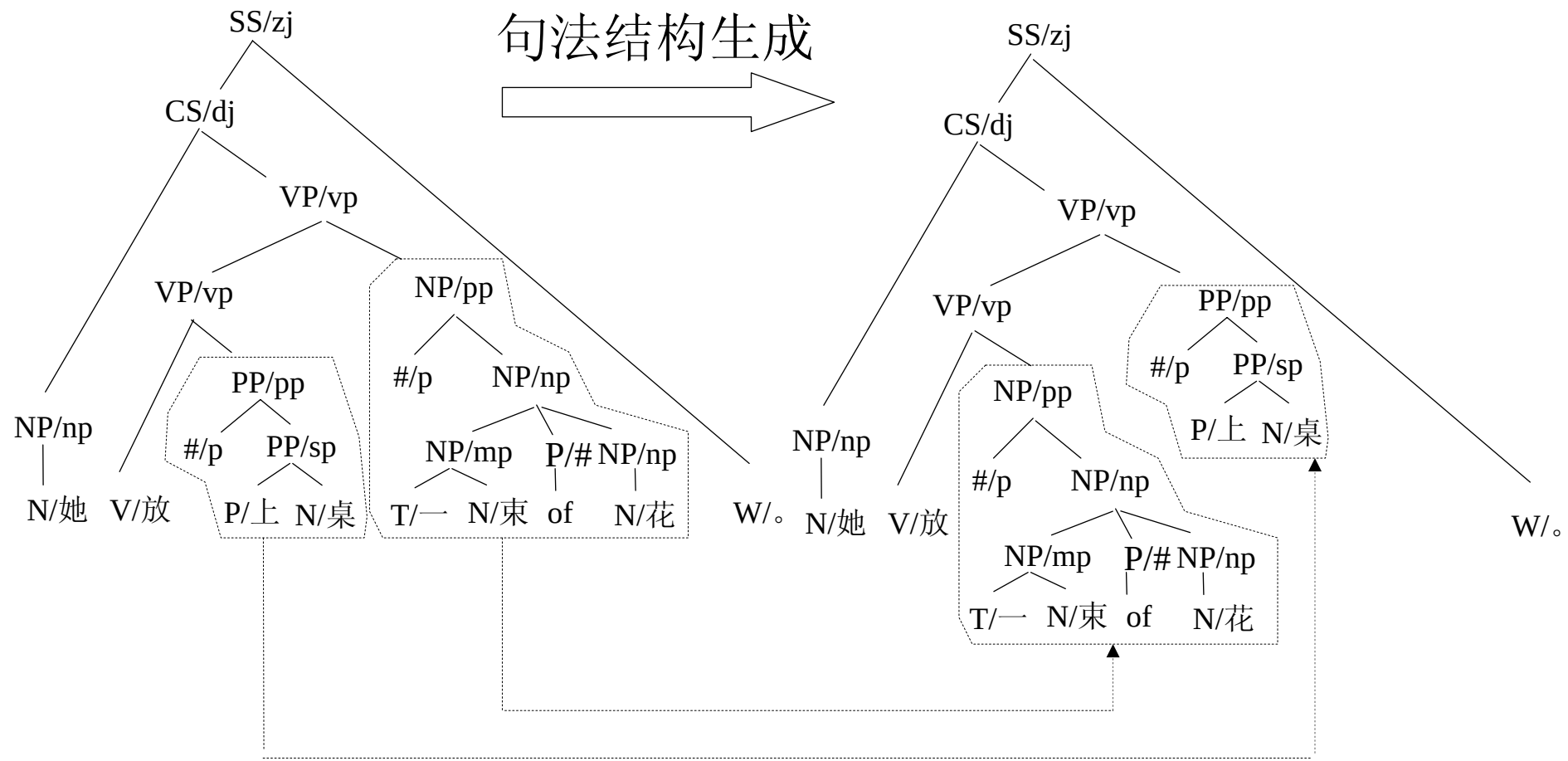
她/r　把/p　一/m-d　束/q　花/n　放/v　在/p-v　桌/n　上/f-v　。/w



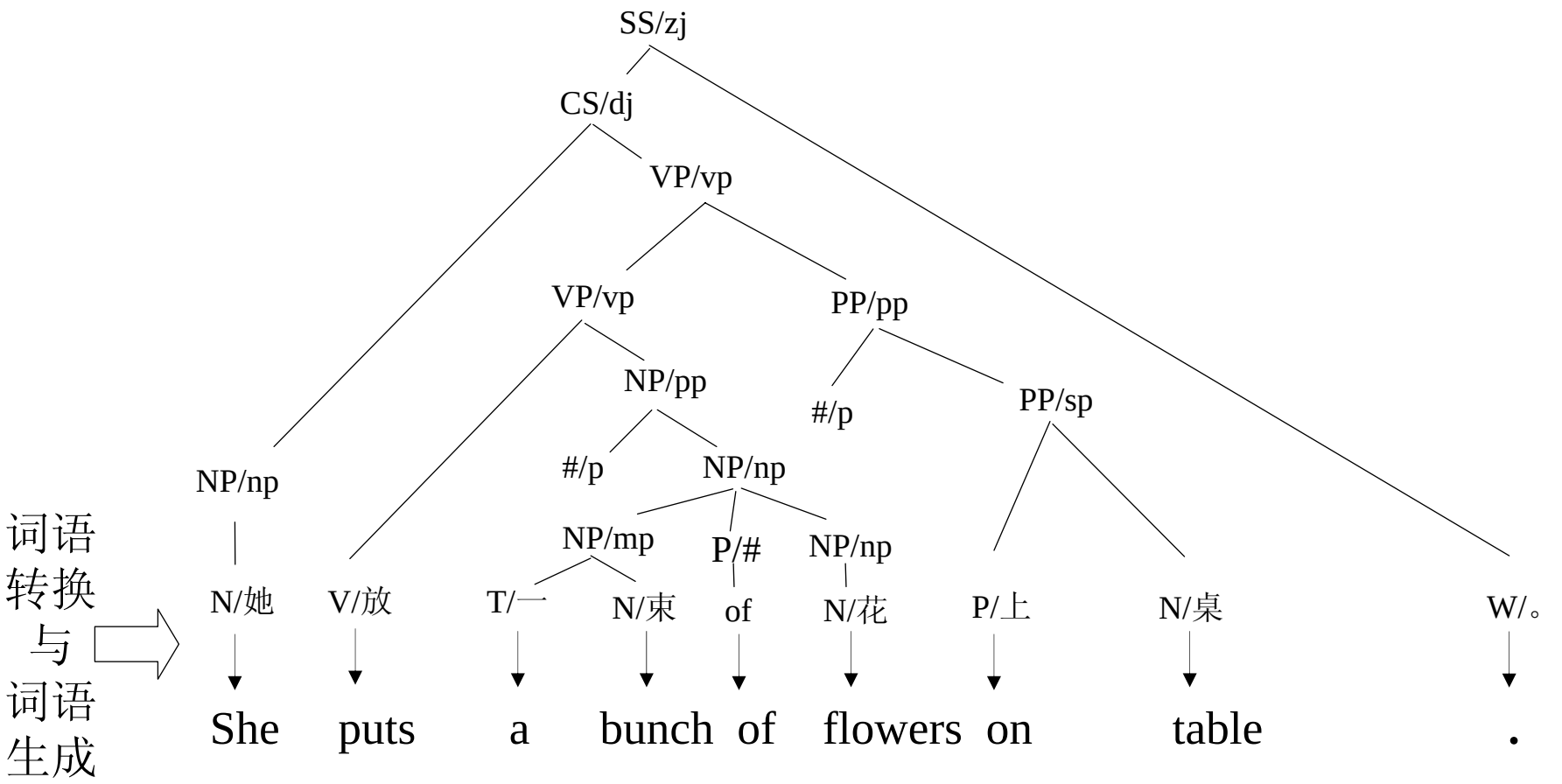
句法层面的转换方法 (3)


句法结构转换
zj
dj
vp
pp
vp
n
p
np
pp
mp
n
p
s
她
把
一
束
花
放
在
桌
上
。
SS/zj
CS/dj
VP/vp
VP/vp
NP/pp
NP/np
PP/pp
#/p
NP/np
#/p
PP/sp
NP/mp
P/#
NP/np
R/她
V/放
P/上
N/桌
T/一
N/束
of
N/花
W/。

句法层面的转换方法 (4)


句法结构生成
SS/zj
CS/dj
VP/vp
VP/vp
NP/pp
PP/pp
#/p
PP/sp
NP/np
NP/mp
P/#
NP/np
NP/np
V/放
P/上
N/桌
T/一
N/束
of
N/花
W/。
N/她
SS/zj
CS/dj
VP/vp
VP/vp
PP/pp
#/p
PP/sp
P/上
N/桌
NP/pp
#/p
NP/np
NP/mp
P/#
NP/np
T/一
N/束
of
N/花
NP/np
N/她
V/放
W/。

句法层面的转换方法 (5)


SS/zj
CS/dj
VP/vp
VP/vp
PP/pp
NP/pp
#/p
PP/sp
NP/np
#/p
NP/np
NP/mp
P/#
NP/np
词语转换与词语生成
N/她
V/放
T/一
N/束
of
N/花
P/上
N/桌
W/。
She
puts
a
bunch
of
flowers
on
table
.

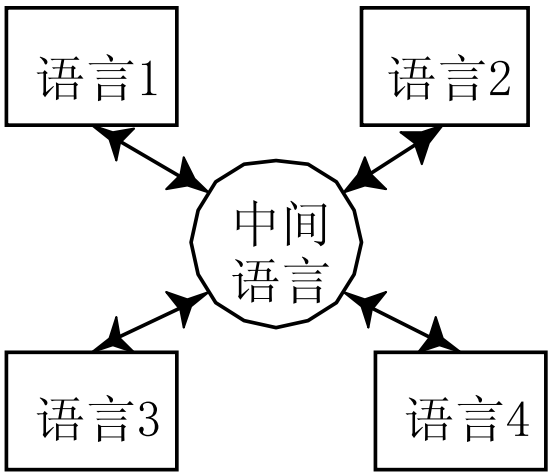
中间语言方法 (1)

- 利用一种中间语言（interlingua）作为翻译的中介表示形式；
- 整个翻译的过程分为“分析”和“生成”两个阶段
- 分析：源语言➔中间语言
- 生成：中间语言➔目标语言
- 分析过程只与源语言有关，与目标语言无关
- 生成过程只与目标语言有关，与源语言无关

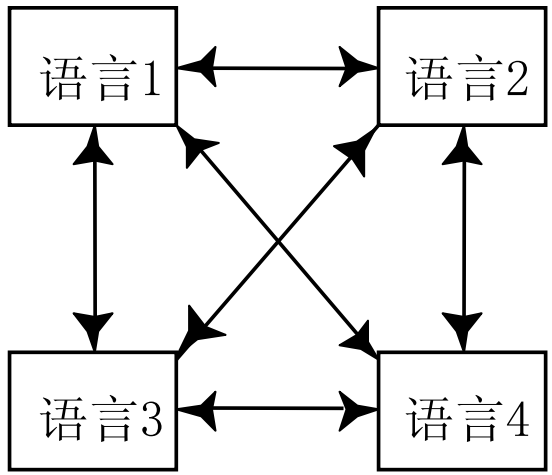
中间语言方法 (2)

- 中间语言方法的优点在于进行多语种翻译的时候，只需要对每种语言分别开发一个分析模块和一个生成模块，模块总数为2*n，相比之下，如果采用转换方法就需要对每两种语言之间都开发一个转换模块，模块总数为n*(n-1)

中间语言方法 (3)



中间语言方法



转换方法

中间语言方法 (4)

- 中间语言的类型
  - 自然语言：如英语、汉语
  - 人工语言：如世界语
  - 某种知识表示形式：如语义网络
- 以某种知识表示形式作为中间语言的机器翻译方法有时也称为基于知识的机器翻译方法

中间语言方法 (5)

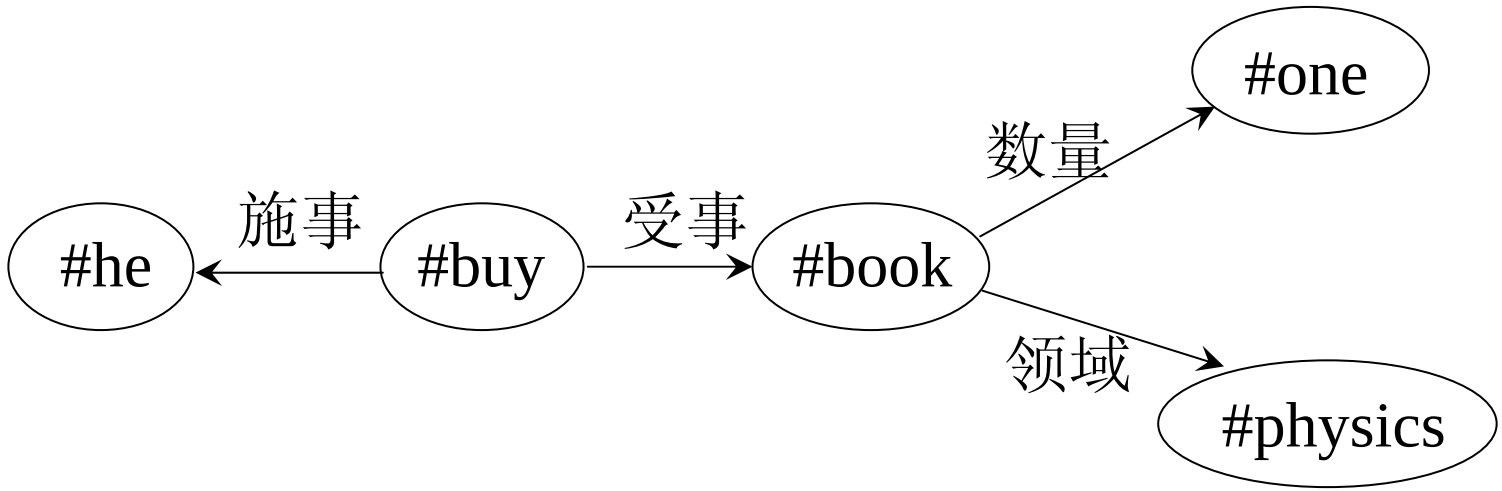
- Makoto Nagao (Kyoto University) said: “.. when the pivot language [i.e. interlingua] is used, the results of the analytic stage must be in a form which can be utilized by all of the different languages into which translation is to take place. This level of subtlety is a practical impossibility.” (Machine Translation, Oxford, 1989)
- Patel-Schneider (METAL system) said: ”METAL employs a modified transfer approach rather than an interlingua. If a meta-language [an interlingua] were to be used for translation purposes, it would need to incorporate all possible features of many languages. That would not only be an endless task but probably a fruitless one as well. Such a system would soon become unmanageable and perhaps collapse under its own weight.” (A four-valued semantics for terminological reasoning, Artificial Intelligence, 38, 1989)

中间语言方法 (6)

- 基于中间语言方法一般都用于多语言的机器翻译系统中；
- 从实践看，基于中间语言的机器翻译系统还没有比较成功的先例，如日本主持的亚洲五国语言机器翻译系统，总体上是失败的；
- 在CSTAR多国语语音机器翻译系统中，曾经采用了一种中间语言方法，其中间语言是一种语义表示形式，由于语音翻译都限制在非常狭窄的领域中（如机票预定），语义描述可以做到非常精确，因此采用中间语言方法有一定的合理性。

中间语言示例—语义网络

英语：He bought a book on physics.
汉语：他买了一本关于物理学的书。



说明：这里#后面表示的是概念，而不是英语词。

机器翻译方法概述

- 机器翻译应用系统类型
- 机器翻译方法分类（按转换层面划分）
- **机器翻译方法分类（按知识表示划分）**

机器翻译方法分类(按知识表示划分)

- 基于规则的机器翻译方法
- 基于实例的机器翻译方法
- 基于统计的机器翻译方法

内容提要


机器翻译（11）
机器翻译历史和现状
机器翻译方法概述
基于规则的机器翻译方法
基于实例的机器翻译方法
统计机器翻译方法

基于规则的方法 (1)

- 采用规则作为知识表示形式
  - 重叠词规则
  - 切分规则
  - 标注规则
  - 句法分析规则
  - 语义分析规则
  - 结构转换规则（产生译文句法语义结构）
  - 词语转换规则（译词选择）
  - 结构生成规则（译文结构调整）
  - 词语生成规则（译文词形生成）

基于规则的方法 (2)

- 优点
  - 直观，能够直接表达语言学家的知识
  - 规则的颗粒度具有很大的可伸缩性
    - 大颗粒度的规则具有很强的概括能力
    - 小颗粒度的规则具有精细的描述能力
  - 便于处理复杂的结构和进行深层次的理解，如解决长距离依赖问题
  - 系统适应性强，不依赖于具体的训练语料

基于规则的方法 (3)

- 缺点
  - 规则主观因素重，有时与客观事实有一定差距
  - 规则的覆盖性差，特别是细颗粒度的规则很难总结得比较全面
  - 规则之间的冲突没有好的解决办法（翘翘板现象）
  - 规则一般只局限于某一个具体的系统，规则库开发成本太高
  - 规则库的调试极其枯燥乏味

基于规则的方法—译词选择

$$ 开

**{v} v $=[…]

|| $.主体=是,$.主体.语义类=植物

→ V<bloom> $=[…]

|| $.客体=是,$.客体.汉字=灯|机|器

→ V(!V<turn> D<on>) $=[…]

|| $.客体=是,$.客体.语义类=交通工具

=> V<drive> $=[…]

|| OTHERWISE

=> V<open> $=[…]

基于规则的方法—结构转换

&& {mp7} mp->r !mp :: $.内部结构=组合定中,…

|| %mp.定语.内部结构=单词, %mp.定语.yx=一,%mp.量词子类=集体|种类|容量|时量|度量|成形

=> NP(T/r !NP/mp) %T.TNNUM=%NP.NNUM /*这一年*/

|| %mp.定语.内部结构=单词, ,%mp.定语.yx=一,%mp.量词子类=个体

=> T(T/r M<one>) /*这一个 哪一个*/

|| %r.yx=这|那, IF %mp.定语.内部结构=单词,%mp.定语.yx=一 FALSE

=> NP(T/r !M/mp) %T.TNNUM=PLUR,$.NNUM=PLUR /*这两张*/

=> NP(T/r !NP/mp) %T.TNNUM=PLUR,$.NNUM=PLUR

|| %r.yx=~这~那,IF %mp.定语.内部结构=单词,%mp.定语.yx=一 FALSE

=> NP(T/r !M/mp) $.NNUM=%M.NNUM

=> NP(T/r !NP/mp) %T.TNSUB=%NP.NSUBC,…

内容提要


机器翻译（11）
机器翻译历史和现状
机器翻译方法概述
基于规则的机器翻译方法
基于实例的机器翻译方法
统计机器翻译方法

基于实例的机器翻译方法

- 基于实例的机器翻译方法
- 基于实例的机器翻译方法的扩展
  - 基于翻译记忆的机器翻译方法
  - 基于模板（模式）的机器翻译方法
- 语料库对齐技术

基于语料库的机器翻译方法

- 机器翻译的实例方法和统计方法都是基于语料库的机器翻译方法
- 优点
  - 使用语料库作为翻译知识来源，无需人工编写规则，系统开发成本低，速度快
  - 从语料库中学习到的知识比较客观
  - 从语料库中学习到的知识覆盖性比较好
- 缺点
  - 系统性能依赖于语料库
  - 数据稀疏问题严重
  - 语料库中不容易获得大颗粒度的高概括性知识

基于实例的机器翻译（1）

- 长尾真(Makoto Nagao)在1984年发表了《采用类比原则进行日-英机器翻译的一个框架》一文，探讨日本人初学英语时翻译句子的基本过程，长尾真认为，初学英语的日本人总是记住一些最基本的英语句子以及一些相对应的日语句子，他们要对比不同的英语句子和相对应的日语句子，并由此推论出句子的结构。参照这个学习过程，在机器翻译中，如果我们给出一些英语句子的实例以及相对应的日语句子，机器翻译系统来识别和比较这些实例及其译文的相似之处和相差之处，从而挑选出正确的译文。
- 长尾真指出，人类并不通过做深层的语言学分析来进行翻译，人类的翻译过程是：首先把输入的句子正确地分解为一些短语碎片，接着把这些短语碎片翻译成其它语言的短语碎片，最后再把这些短语碎片构成完整的句子，每个短语碎片的翻译是通过类比的原则来实现的。
- 因此，我们应该在计算机中存储一些实例，并建立由给定的句子找寻类似例句的机制，这是一种由实例引导推理的机器翻译方法，也就是基于实例的机器翻译。

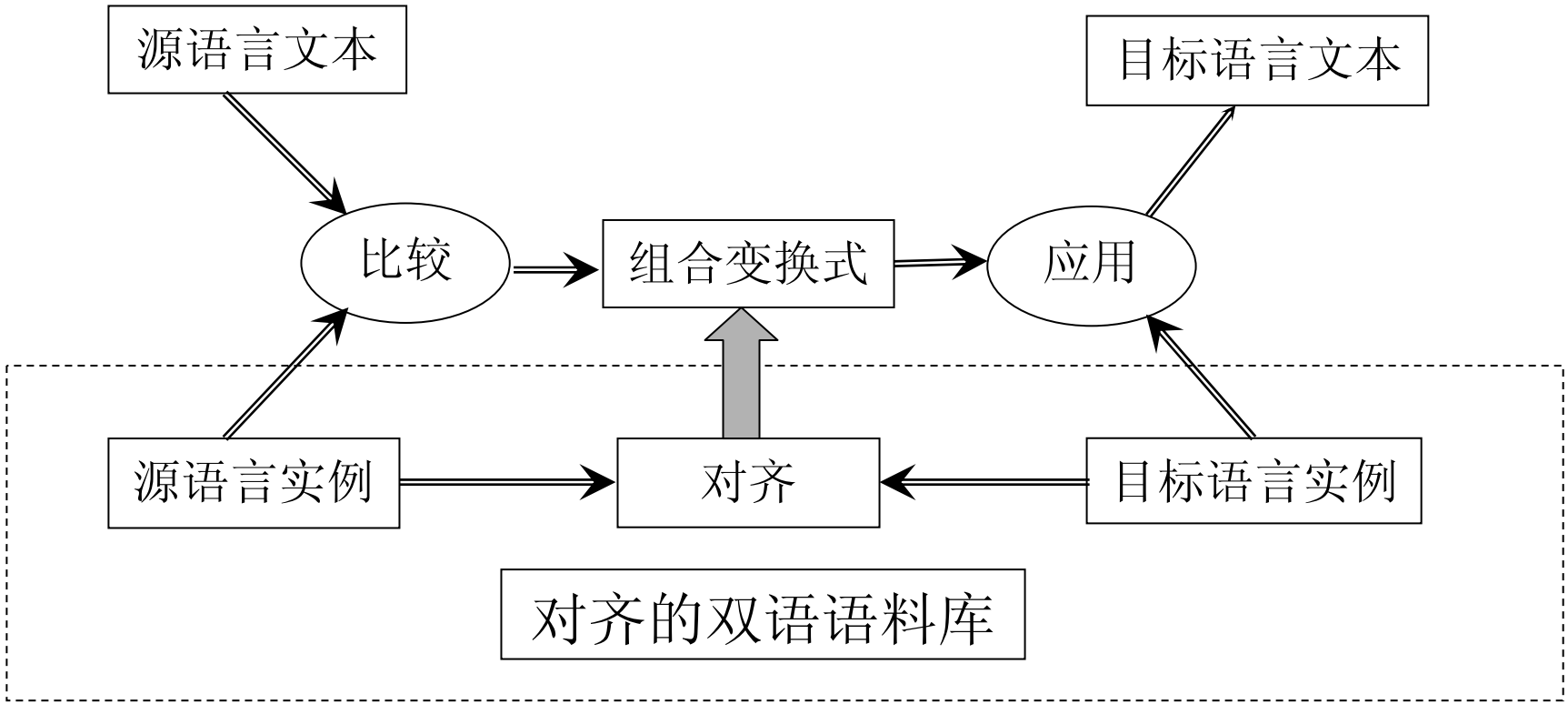
基于实例的机器翻译（2）

- 在基于实例的机器翻译系统中，系统的主要知识源是双语对照的翻译实例库，实例库主要有两个字段，一个字段保存源语言句子，另一个字段保存与之对应的译文，每输入一个源语言的句子时，系统把这个句子同实例库中的源语言句子字段进行比较，找出与这个句子最为相似的句子，并模拟与这个句子相对应的译文，最后输出译文。
- 基于实例的机器翻译系统中，翻译知识以实例和义类词典的形式来表示，易于增加或删除，系统的维护简单易行，如果利用了较大的翻译实例库并进行精确的对比，有可能产生高质量译文，而且避免了基于规则的那些传统的机器翻译方法必须进行深层语言学分析的难点。在翻译策略上是很有吸引力的。

基于实例的机器翻译（3）

- 优点
  - 直接使用对齐的语料库作为知识表示形式，知识库的扩充非常简单
  - 不需要进行深层次的语言分析，也可以产生高质量的译文
- 缺点
  - 覆盖率低，实用的系统需要的语料库规模极大（百万句对以上）

基于实例的机器翻译系统结构



基于实例的机器翻译—举例

要翻译句子：

(E1) He bought a book on physics.

在语料库中查到相似英语句子及其汉语译文是：

(E2) He wrote a book on history.

(C2) 他写了一本关于历史的书。

比较(E1)和(E2)两个句子，我们得到变换式：

(T1) replace(wrote, bought) and replace(history, physics)

将这个变换式中的单词都换成汉语就变成：

(T2) replace(写,买) and replace(历史,物理)

将(T2)作用于(C2)

(C1)他买了一本关于物理学的书。

基于实例的机器翻译需要研究的问题

- 正确地进行双语自动对齐(alignment)：在实例库中要能准确地由源语言例句找到相应的目标语言例句，在基于实例的机器翻译系统的具体实现中，不仅要求句子一级的对齐，而且还要求词汇一级甚至短语一级的对齐。
- 建立有效的实例匹配检索机制：很多研究者认为，基于实例的机器翻译的潜力在于充分利用短语一级的实例碎片，也就是在短语一级进行对齐，但是，利用的实例碎片越小，碎片的边界越难于确定，歧义情况越多，从而导致翻译质量的下降，为此，要建立一套相似度准则(similarity metric)，以便确定两个句子或者短语碎片是否相似。
- 根据检索到的实例生成与源语言句子相对应的译文：由于基于实例的机器翻译对源语言的分析比较粗，生成译文时往往缺乏必要的信息，为了提高译文生成的质量，可以考虑把基于实例的机器翻译与传统的基于规则的机器翻译方法结合起来，对源语言也进行一定深度的分析。
- 开展浅层句法分析(shallow parsing)的研究：浅层句法分析以建立语段(chunk)之间的依附关系为目标，进行语段的识别，分析语段之间的依附关系。由于分析的语言单位的颗粒度比较大，歧义就比较少，有利于提高双语对齐的准确度。

实例库的匹配 (1)

- 实例匹配的目的是将输入句子分解成语料库中实例片断的组合，这是基于实例的机器翻译的关键问题之一，实例匹配的各种方法有很大的差异，还没有那种做法显示出明显的优势；
- 实例库匹配的效率问题：由于实例库规模较大，通常需要建立倒排索引；
- 实例库匹配的其他问题：
  - 实例片断的分解：
  - 实例片断的组合：

实例库的匹配 (2)

- 实例片断的分解
  - 实例库中的句子往往太长，直接匹配成功率太低，为了提高实例的重用性，需要将实例库中的句子分解为片断
  - 几种通常的做法：
    - 按标点符号分解
    - 任意分解
    - 通过组块分析进行分解

实例库的匹配 (3)

- 实例片断的组合
  - 一个被翻译的句子，往往可以通过各种不同的实例片断进行组合，如何选择一个最好的组合？
  - 简单的做法：
    - 最大匹配
    - 最大概率法：选择概率乘积最大的片断组合
  - 有点像汉语词语切分问题

片断译文的选择

- 由于语料库中一个片断可能有多种翻译方法，因此存在片断译文的选择问题；
- 常用的方法：
  - 根据片断上下文进行排歧；
  - 根据译文的语言模型选择概率最大的译文片断组合

基于实例的机器翻译系统

- MBT1和MBT2系统：由日本京都大学长尾真和佐藤研制。该系统的翻译过程分为分解(decomposition)、转换(transfer)、合成(composition)三步。在分解阶段，系统根据提交的源语言词汇依存树检索实例库，并利用检索到的实例碎片来表示该源语言句子的依存树，形成源匹配表达式；在转换阶段，系统利用实例库中的对齐信息将源匹配表达式转换成目标匹配表达式；在合成阶段，将目标匹配表达式展开成为目标语言词汇依存树，输出译文。
- PANGLOSS系统：由美国卡内基-梅隆大学研制，这是一个多引擎机器翻译系统(Multi-engine Machine Translation)。这个系统的主要引擎是基于知识的机器翻译系统，基于实例的机器翻译系统只是它的一个引擎，为整个多引擎机器系统提供候选结果。
- ETOC和EBMT系统：由日本口语翻译通信研究实验室 ATR研制。 ETOC系统能够检索出与给定的源语言句子相似的实例，EBMT系统能够利用实例库来消解歧义，这两个基于实例的机器翻译系统还不完整。
- 我国清华大学计算机系的基于实例的日汉机器翻译系统。

基于实例的机器翻译方法

- 基于实例的机器翻译方法
- **基于实例的机器翻译方法的扩展**
  - **基于翻译记忆的机器翻译方法**
  - 基于模板（模式）的机器翻译方法
- 语料库对齐技术

翻译记忆方法 (1)

- 翻译记忆方法（Translation Memory）是基于实例方法的特例；
- 也可以把基于实例的方法理解为广义的翻译记忆方法；
- 翻译记忆的基本思想：
  - 把已经翻译过的句子保存起来
  - 翻译新句子时，直接到语料库中去查找
    - 如果发现相同的句子，直接输出译文
    - 否则交给人去翻译，但可以提供相似的句子的参考译文

翻译记忆方法 (2)

- 翻译记忆方法主要被应用于计算机辅助翻译（CAT）软件中
- 翻译记忆方法的优缺点
  - 翻译质量有保证
  - 随着使用时间的增加匹配成功率逐步提高
  - 特别适用于重复率高的文本翻译，例如公司的产品说明书的新版本翻译
  - 与语言无关，适用于各种语言对
  - 缺点是匹配成功率不高，特别是刚开始使用时

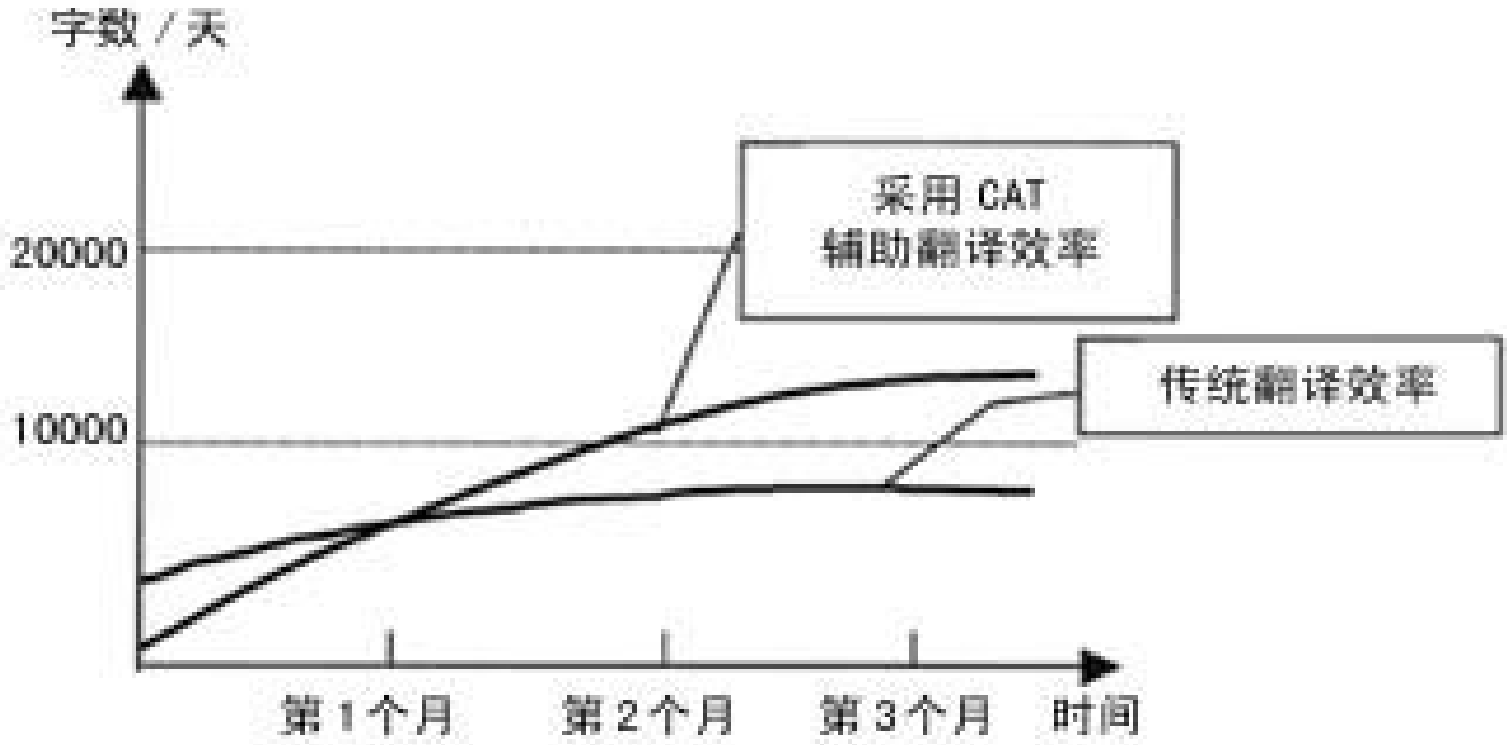
翻译记忆方法 (3)

- 计算机辅助翻译（CAT）软件已经形成了比较成熟的产业
  - TRADOS
    - 号称占有国际CAT市场的70%
    - Microsoft、Siemens、SAP等国际大公司和一些著名的国际组织都是其用户
  - 雅信CAT
    - 适合中国人的习惯
    - 产品已比较成熟
  - 国际组织： LISA（Localisation Industry Standards Association）
- 面向用户：专业翻译人员
- 数据交换：LISA制定了TMX（Translation Memory eXchange）标准。

翻译记忆方法 (4)

- 完整的计算机辅助翻译软件除了包括翻译记忆功能以外，还应该包括以下功能
  - 多种文件格式的分解与合成
  - 术语库管理功能
  - 语料库的句子对齐（历史资料的重复利用）
  - 项目管理：
    - 翻译任务的分解与合并
    - 翻译工作量的估计
  - 数据共享和数据交换

翻译记忆方法 (5)



基于实例的机器翻译方法

- 基于实例的机器翻译方法
- **基于实例的机器翻译方法的扩展**
  - 基于翻译记忆的机器翻译方法
  - **基于模板（模式）的机器翻译方法**
- 语料库对齐技术

基于模板(模式)的机器翻译方法(1)

- 基于模板（Template）或者模式（Pattern）的机器翻译方法通常也被看做基于实例的机器翻译方法的一种延伸
- 所谓“翻译模板”或者“翻译模式”可以认为是一种颗粒度介于“翻译规则”和“翻译实例”之间的翻译知识表示形式
  - 翻译规则：颗粒度大，匹配可能性大，但过于抽象，容易出错
  - 翻译实例：颗粒度小，不易出错，但过于具体，匹配可能性小
  - 翻译模板（模式）：介于二者之间，是一种比较合适的知识表示形式
- 一般而言，单语模板（或模式）是一个常量和变量组成的字符串，翻译模板（或模式）是两个对应的单语模板（或模式），两个模板之间的变量存在意义对应关系

基于模板(模式)的机器翻译方法(2)

- 模板举例：
  - 这个 X 比 Y 更 Z。
  - The X is more Z than Y.
- 模板方法的主要问题
  - 对模板中变量的约束
  - 模板抽取
  - 模板的冲突消解

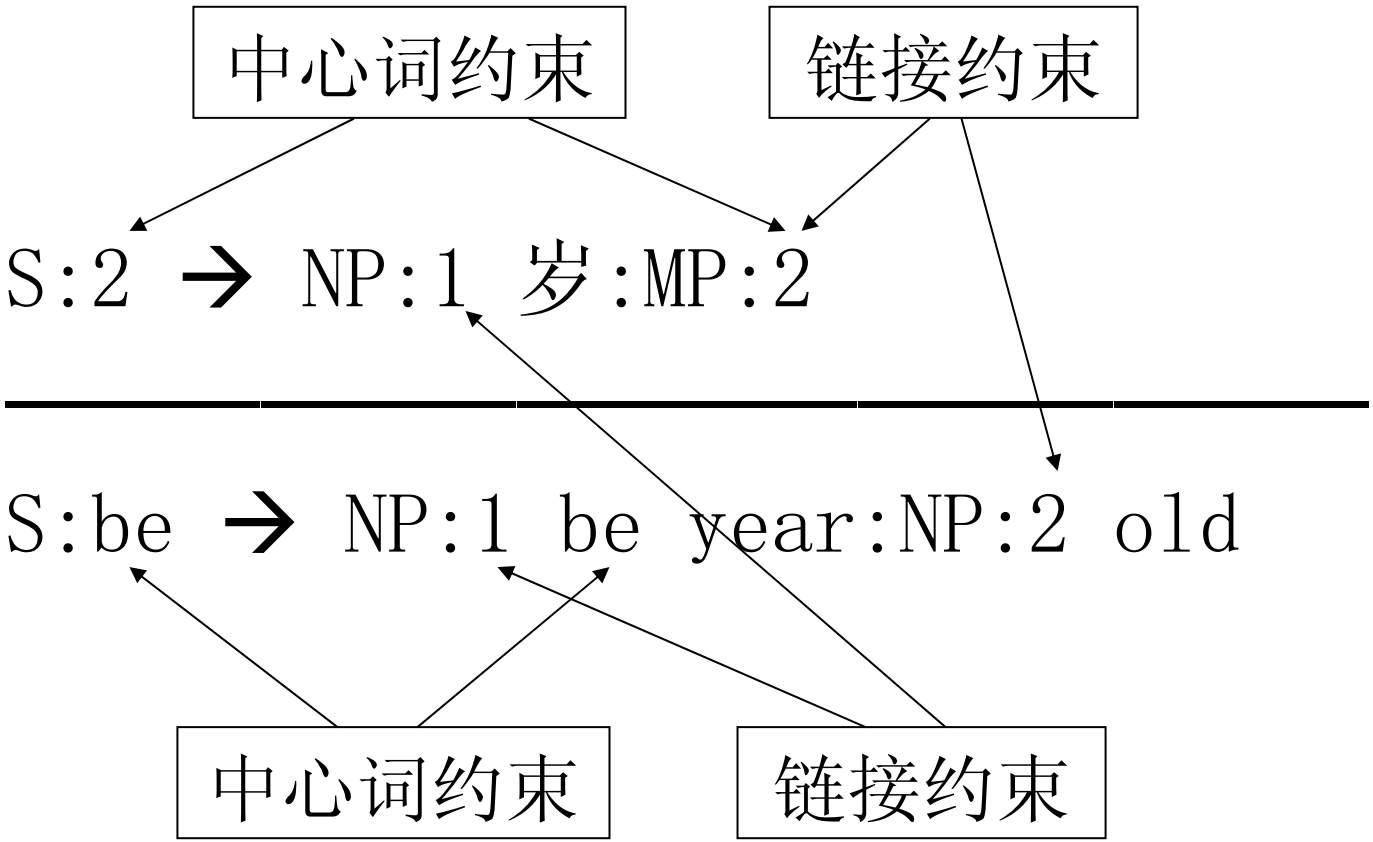
Pattern-Based CFG for MT(1)

- Koichi Takeda, Pattern-Based Context-Free Grammars for Machine Translation, Proc. of 34th ACL, pp. 144-- 151, June 1996
- 给出了翻译模式的一种形式化定义，并给出了相应的翻译算法以及算法复杂性的理论证明

Pattern-Based CFG for MT(2)

- 每个翻译模板由一个源语言上下文无关规则和一个目标语言上下文无关规则（这两个规则称为翻译模板的骨架），以及对这两个规则的中心词约束和链接约束构成；
- 中心词约束：对于上下文无关语法规则中右部（子结点）的每个非终结符，可以指定其中心词；对于规则左部（父结点）的非终结符，可以直接指定其中心词，也可以通过使用相同的序号规定其中心词等于其右部的某个非终结符的中心词；
- 链接约束：源语言骨架和目标语言骨架的非终结符子结点通过使用相同的序号建立对应关系，具有对应关系的非终结符互为翻译。

Pattern-Based CFG for MT(3)



Pattern-Based CFG for MT(4)

- 翻译的过程分为三步：
  - 使用源语言CFG骨架分析输入句子s
  - 应用源语言到目标语言的CFG骨架的链接约束，生成一个译文CFG推导序列
  - 根据译文CFG推导序列产生译文
- 模板排序的启发式原则：
  - 对于源文CFG骨架相同的模板，有中心词约束的模板优先于没有中心词约束的模板；
  - 对于同一跨度上的两个结点，比较其对应的模板的源文CFG骨架，非终结符少的模板优先于非终结符多的模板；
  - 中心词约束被满足的结点优先于中心词约束不被满足的结点；
  - 对于一个输入串而言，分析步骤越短（推导序列越短）越优先。

Pattern-Based CFG for MT(5)

- 模板库的获取：假设T是一组翻译模板，B是双语语料库，<s,t>是一对互为翻译的句子
  - 如果T能够翻译句子s为t，那么do nothing；
  - 如果T将s译为t'（不等于t），那么：
    - 如果T中存在<s,t>的推导Q，但这个推导不是最优解，那么给Q中的模板进行实例化；
    - 如果不存在这种推导，那么加入适当的模板，使得推导成立；
  - 如果根本无法翻译s（分析失败），那么将<s,t>直接加入到模板库中。

模板的自动提取

- 利用一对实例进行泛化
  - Jaime G. Carbonell, Ralf D. Brown, Generalized Example-Based Machine Translation
    http://www.lti.cs.cmu.edu/Research/GEBMT/
- 利用两对实例进行比较
  - H. Altay Guvenir, Ilyas Cicekli, Learning Translation Templates from Examples Information Systems, 1998
  - 张健，基于实例的机器翻译的泛化方法研究，中科院计算所硕士论文，2001

通过泛化实例得到翻译模板

- 已有实例：
  - Karl Marx was born in Trier, Germany in May 5, 1818.
  - 卡尔·马克思于1818年5月5日出生在德国特里尔城。
- 泛化：
  - <Person> was born in <City> in <Date>
  - <Person>于<Date>出生在<City>
- 对齐
  - <Person> ⇔ <Person>
  - <City> ⇔ <City>
  - <Date>⇔<City>

通过比较实例得到翻译模板

- 已有两对翻译实例：
  - 我给玛丽一支笔 ⇔ I gave Mary a pen.
  - 我给汤姆一本书 ⇔ I gave Tom a book.
- 双侧单语句子分别比较，得到：
  - 我 给 #X 一 #Y #Z ⇔ I give #W a #U.
- 查找变量的对应关系：
  - #X ⇔ #W
  - #Y ⇔ ϕ
  - #Z ⇔ #U

基于实例的机器翻译方法

- 基于实例的机器翻译方法
- 基于实例的机器翻译方法的扩展
  - 基于翻译记忆的机器翻译方法
  - 基于模板（模式）的机器翻译方法
- 语料库对齐技术

双语语料库的对齐

- 双语语料库（Bilingual Corpus）或平行语料库（Parallel Corpus），在EBMT中又称为实例库
- 双语语料库对齐的级别
  - 篇章对齐
  - 段落对齐
  - 句子对齐
  - 词语对齐
  - 短语块对齐
  - 句法结构对齐
- 基于实例的机器翻译中实例库必须至少做到句子级别的对齐

不同对齐级别的差异

- 段落对齐和句子对齐
  - 要求保持顺序（允许局部顺序的调整）
  - 只有一个层次
- 词语对齐和短语块对齐
  - 不要求保持顺序
  - 只有一个层次
- 句法结构对齐
  - 不要求保持顺序
  - 多层次对齐

句子对齐 (1)

汉语	英语	模式
1995年初我来成都的那天，没想到会是在一个冬季的漆黑的日子。	I little thought when I arrived in Chengdu in the dark, dark days of winter, early in 1995, that I would still be here more than five years later.	1:1
那时我也根本没有想到会在这儿呆上五年，也不知道我会遇到一位成都的女儿，并且后来还娶她为妻。一个完全陌生的家庭接纳了我，我也因此成为成都的一部分。	I little knew that I would meet one of Chengdu's daughters, and later marry her, thus acquiring a whole new family who embraced me as one of them, and thus I became part of this place.	2:1

句子对齐 (2)

对于篇章对齐（或者段落对齐）的一对文本(S,T)：

$$S = s_1 ... s_m, T = t_1 ... t_n$$

定义其对齐为A={A_1,…,A_k}，其中A_i称为一个句珠(Bead)：

$$A_i = (S_i, T_i) = (s_{a_{i-1}+1} ... s_{a_i}, t_{b_{i-1}+1} ... t_{b_i}),$$

其中$a_0 = 0 < ... < a_{i-1} < a_i < ... < a_k = m, b_0 = 0 < ... < b_{i-1} < b_i < ... < b_k = n$

整个对齐的概率为：

$$P(A) = \prod_{i=1}^{k} P(A_i)$$

基于长度的句子对齐 (1)

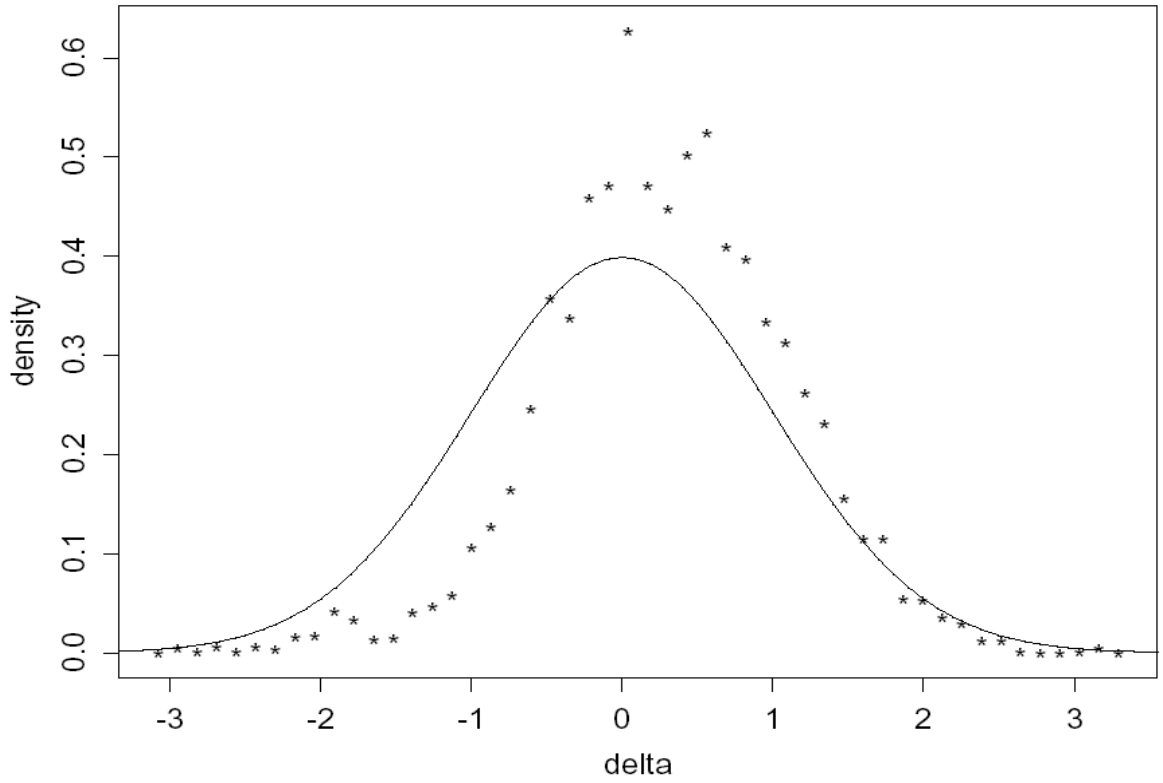
- 基本思想：源语言和目标语言的句子长度存在一定的比例关系
- 用两个因素来估计一个句珠的概率
  - 源语言和目标语言中句子的长度
  - 源语言和目标语言中的句子数（对齐模式）

$$P(A_i) = P(S_i, T_i)$$

$$\approx P(l_{Si}, l_{Ti}) \times P(m_{Si}, m_{Ti})$$

基于长度的句子对齐 (2)

- 根据统计，随机变量$X=l_{Ti}/l_{Si}$服从正态分布



基于长度的句子对齐 (3)

- 设通过语料库统计得到X的期望为c，方差为v^2，那么随机变量 δ 将服从[0,1]正态分布：

$$\delta = \frac{X - c}{v} = \frac{l_T - cl_S}{vl_S} \sim N(0,1)$$

- 根据正态分布公式可以计算出(直接查表)：

$$P(l_S, l_T) = P(\delta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\delta^2}{2}}$$

基于长度的句子对齐 (4)

- 对齐模式的概率$P(m_S,m_T)$可以通过对语料库的统计得到。
- 下面是Gale & Church根据UBS语料库的统计结果：

Category	Frequency	Prob(match)
1-1	1167	0.89
1-0 or 0-1	13	0.0099
2-1 or 1-2	117	0.089
2-2	15	0.011
	1312	1.00

基于长度的句子对齐 (5)

- 最优路径的搜索：采用动态规划算法
- 定义P(i,j)=P($s_1 \ldots s_i, t_1 \ldots t_j$)

$$P(i, j) = \max_{x,y}\{P(i-x, j-y) + Score(s_{i-x+1} \ldots s_i, t_{j-y+1} \ldots t_j)\}$$

- 最优对齐为P(m,n)所对应的路径

基于长度的句子对齐 (6)

- 优点
  - 不依赖于具体的语言；
  - 速度快；
  - 效果好
- 缺点
  - 由于没有考虑词语信息，有时会产生一些明显的错误
- 讨论
  - 长度计算可以采用词数或者字节数，没有明显的优劣之分

基于词的句子对齐 (1)

- 基本思想：互为翻译的句子对中，含有互为翻译的词语对的概率，大大高于随机的句子对
- 用两个因素来估计一个句珠的概率
  - 源语言和目标语言中互译词语的个数
  - 源语言和目标语言中的句子数（对齐模式）

$$P(A_i) = P(S_i, T_i)$$

$$\approx P(w_{Si}, w_{Ti}) \times P(m_{Si}, m_{Ti})$$

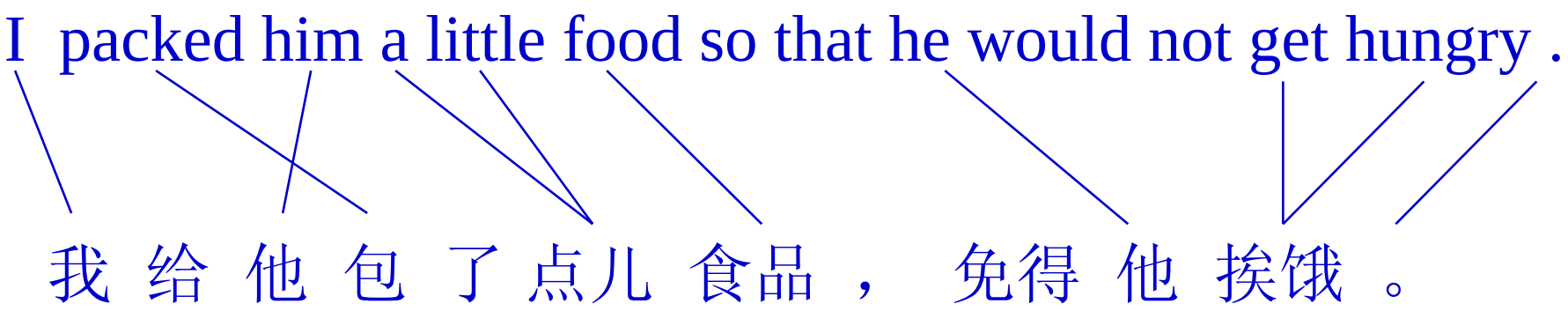
基于词的句子对齐 (2)

- 优点
  - 可以充分利用词语互译信息，提高正确率
- 缺点
  - 单独使用时，正确率有时低于基于长度的方法（取决于词典的规模质量等）
  - 时空开销大
- 讨论
  - 对于同源的语言（英语和法语，汉语和日语）可以利用词语同源信息而不使用词典

句子对齐小结

- 句子对齐的语料库是基于语料库的机器翻译的基础；
- 综合采用基于长度的方法和基于词汇的方法可以取得较好的效果；
- 句子对齐可以取得很高的正确率，已经达到实用水平。

词语对齐 (1)



- 特点：
  - 保序性不再满足
  - 对齐模式复杂：一对多、多对一、多对多都非常普遍

词语对齐 (2)

- 困难：
  - 翻译歧义：一个词出现两个以上的译词
  - 双语词典覆盖率有限：非常普遍的现象
  - 位置歧义：出现两个以上相同的词
  - 汉语词语切分问题
  - 虚词问题：虚词的翻译非常灵活，或没有对译词
  - 意译问题：根本找不到对译的词

词语对齐 (3)

- 一般而言，一个单词对齐的模型可以表述为两个模型的乘积：
  - 词语相似度模型(word similarity model)
  - 位置扭曲模型(word distortion model)

  用公式表示如下：

$$Score(e_i, c_j) = S(e_i, c_j) \times D(i, j)$$

词语相似度模型 (1)

- T-Score：

$$T-score(\mathbf{e},\mathbf{c})=\frac{N_{ec}\times Total-N_{c}\times N_{e}}{Total\times\sqrt{Total}}$$

N_c：语料库中单词c出现的词数

N_e：语料库中单词e出现的词数

N_{ec}：语料库中单词e和单词c互译的词数

Total：语料库中句子对的数量

词语相似度模型 (2)

- 戴斯系数（dice coefficient）

  设S_1和S_2分别是两个集合，则这两个集合的戴斯系数可以通过如下公式计算

$$Dice(S_1, S_2) = \frac{2|S_1 \cap S_2|}{|S_1| + |S_2|}$$

- 把汉语词理解为汉字的集合，戴斯系数就是两个词中相同的汉字占两个词汉字总数的比例。考虑到汉字表意性，这种方法在计算汉语词相似度时有较好的效果
- 计算汉语词c和英语词e的相似度：
  - 先用英语词e查英汉词典，得到所有的汉语对译词；
  - 计算所有对译词和c的戴斯系数，取其中的最大值。

词语相似度模型 (3)

- 互信息（mutual information）

  通过两个事件X和Y各自出现的概率为p(X)和p(Y)，他们联合出现的概率为p(X,Y)，这两个事件之间共同的互信息量定义为：

$$I(X,Y) = -\log_2 \frac{p(X)p(Y)}{p(X,Y)}$$

- 当两个事件相互独立时，互信息量为0；
- 当两个事件倾向于同时出现时，互信息量为正；
- 当两个事件倾向于互相排斥时，互信息量为负；
- 利用互信息作词语相似度计算效果较差。

词语相似度模型 (4)

- ϕ^2方法：利用联立表（contingency table）

	Wt+	Wt-
Ws+	31,950(a)	12,004(b)
Ws-	4,793(c)	848,330(d)

$$\phi^2 = \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

- ϕ^2(phi-square)方法的效果比较好

词语相似度模型 (5)

- 对数似然比（ Log Likelihood Ratio,LLR ）

$$LLR = \log L(p_1, k_1, n_1) + \log L(p_2, k_2, n_2) - \log L(p, k_1, n_1) - \log L(p, k_2, n_2)$$

其中：$\log L(p, n, k) = k \log p + (n - k) \log(1 - p)$

$$k_1 = f(w_t, w_s), k_2 = f(w_t, \neg w_s), n_1 = f(w_s), n_2 = f(\neg w_s)$$

$$p_1 = p(w_t \mid w_s) = \frac{k_1}{n_1}, p_2 = p(w_t \mid \neg w_s) = \frac{k_2}{n_2}, p = p(w_t) = \frac{k_1 + k_2}{n_1 + n_2}$$

对数似然比在使用中比较有效，在训练语料库规模较小时尤为明显

词语相似度模型 (6)

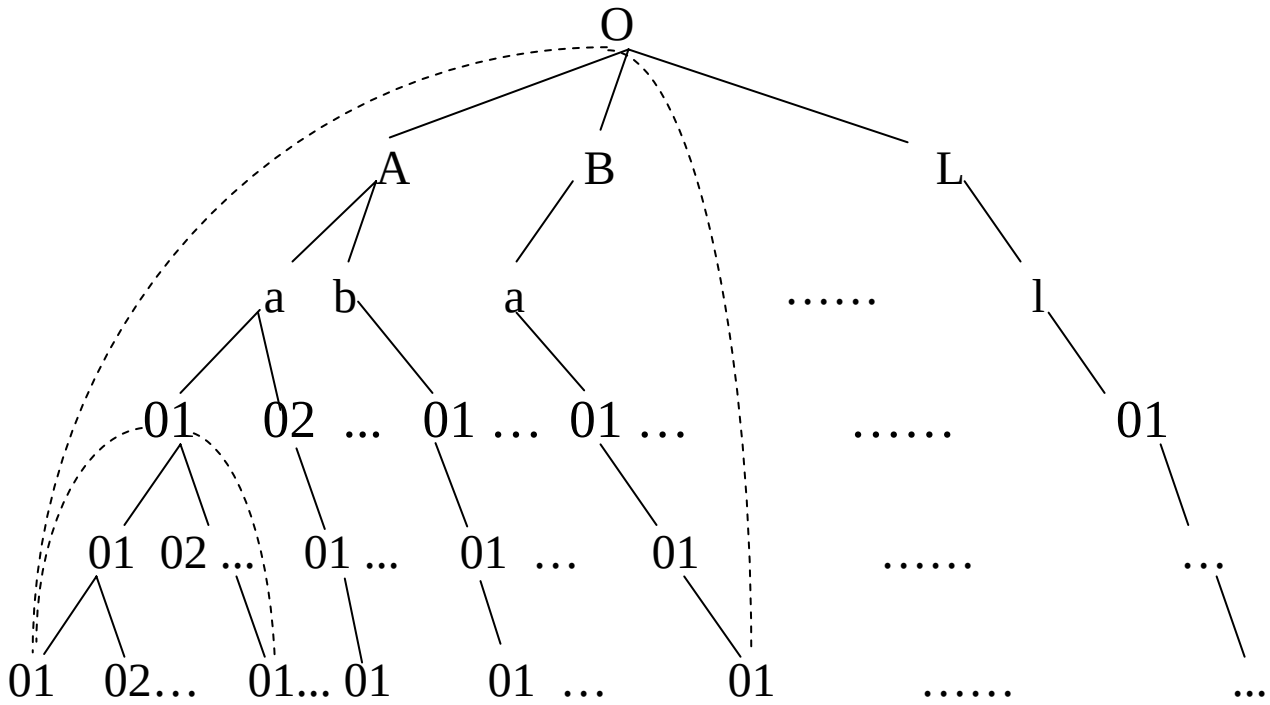
- 概念相似度
  利用某种形式的义类词典（Thesaurus），计算两个词语对应的概念之间的相似度

$$Sim(p_1, p_2) = \frac{\alpha}{d + \alpha}$$

其中d是概念p_1、p_2之间的距离，一般用概念层次体系中两个结点之间的距离来计算
α 是一个可条件的参数

词语相似度模型 (7)

《同义词词林》的概念层次体系



虚线用于标识某上层结点到下层结点的路径

位置扭曲模型

- 相对偏移模型

$$dis(i,j) = \min(|L|,|R|)$$

$$L = |s_i - s_{i-1}| - |t_j - t_{j-1}|$$

$$R = |s_i - s_{i+1}| - |t_j - t_{j+1}|$$

$$d(i,j) = \begin{cases} d1 & if\ dis(i,j) = 0 \\ d2 & if\ dis(i,j) = 1 \\ d3 & if\ dis(i,j) = 2 \\ d4 & if\ dis(i,j) >= 3 \end{cases}$$

s_i是源语言e_i单词的位置

t_j是目标语言单词c_j的位置

s_i跟t_j对齐

s_{i-1}是s_i左侧最近的一个对齐的单词

s_{i-1}是s_i左侧最近的一个对齐的单词

t_{j-1}是跟s_{i-1}对齐的单词

t_{j+1}是跟s_{i+1}对齐的单词

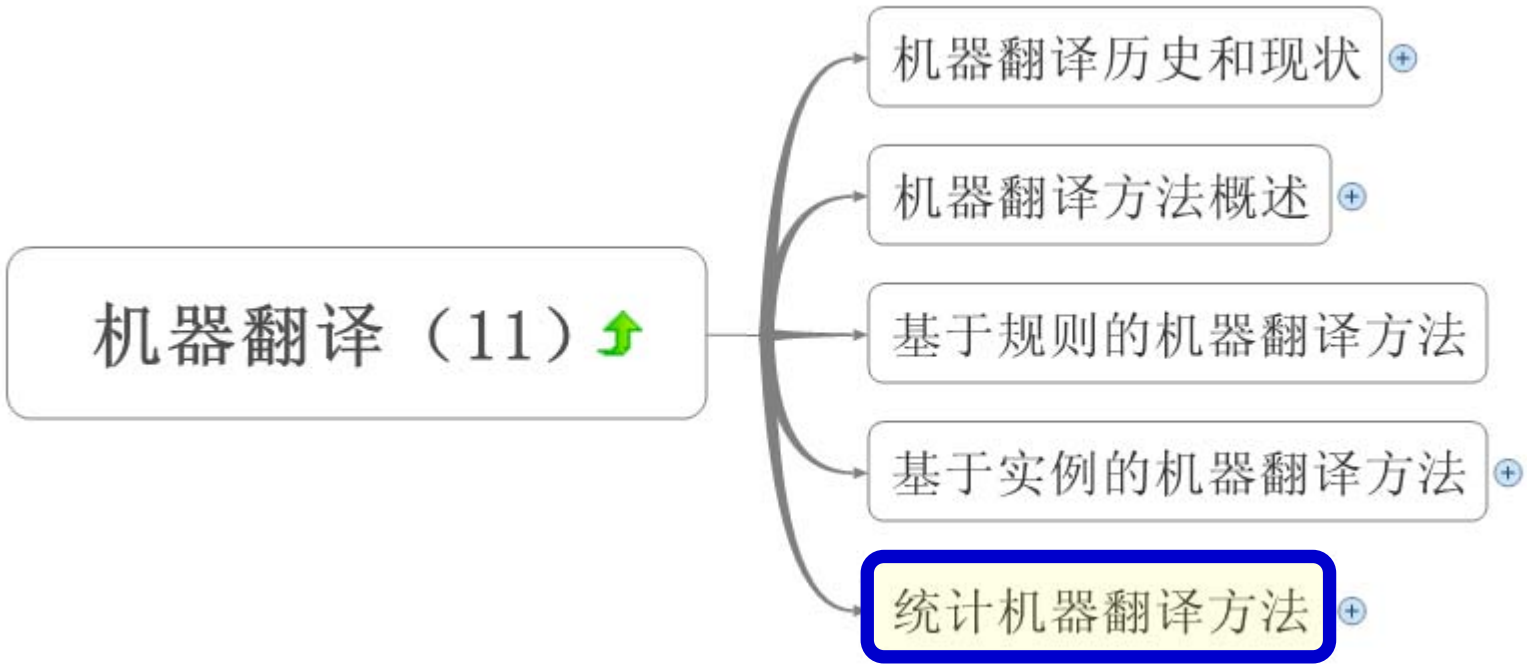
词语对齐的搜索算法

- 贪心法
  1.定义对齐评价函数
  2.把两种语言单词集合的笛卡儿积作为候选集合
  3.计算所有候选词对儿的评价函数
  4.找出最好的对齐词对儿，从候选集合中删除
  5.删除与刚找出的词对儿冲突的词对儿
  6.重复以上3～5，直到评分低于某个阈值
- 搜索法

词语对齐小结

- 词语对齐比句子对齐困难得多
- 词语对齐主要使用一个词语相似度模型和一个位置扭曲模型
- 词语对齐算法常见的有迭代法和贪心法
- 词语对齐的副产品：双语词典抽取

内容提要


机器翻译（11）
机器翻译历史和现状
机器翻译方法概述
基于规则的机器翻译方法
基于实例的机器翻译方法
统计机器翻译方法

统计机器翻译

- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
  —基于词的IBM模型
- 最成熟的统计机器翻译方法
  —基于短语的模型
- 目前统计机器翻译研究的热点
  —基于句法的模型
- 机器翻译的自动评价

统计机器翻译的研究热潮

- 历史回顾：一些重要事件回放
- 一种新的研究范式
- 统计机器翻译论文发表数量的增长
- 近年来国际机器翻译评测的最好成绩
- 统计机器翻译目前的水平

历史回顾：一些重要事件回放 (1)

- 1980年代末IBM首次开展统计机器翻译研究
- 1992年IBM首次提出统计机器翻译的信源信道模型
- 1993年IBM提出五种基于词的统计翻译模型IBM Model 1-5
- 1994年IBM发表论文给出了Candide系统与Systran系统在ARPA评测中的对比测试报告
- 1999年JHU夏季研讨班重复了IBM的工作并推出了开放源代码的工具
- 2001年IBM提出了机器翻译自动评测方法BLEU
- 2002年NIST开始举行每年一度的机器翻译评测
- 2002年第一个采用统计机器翻译方法的商业公司Language Weaver成立

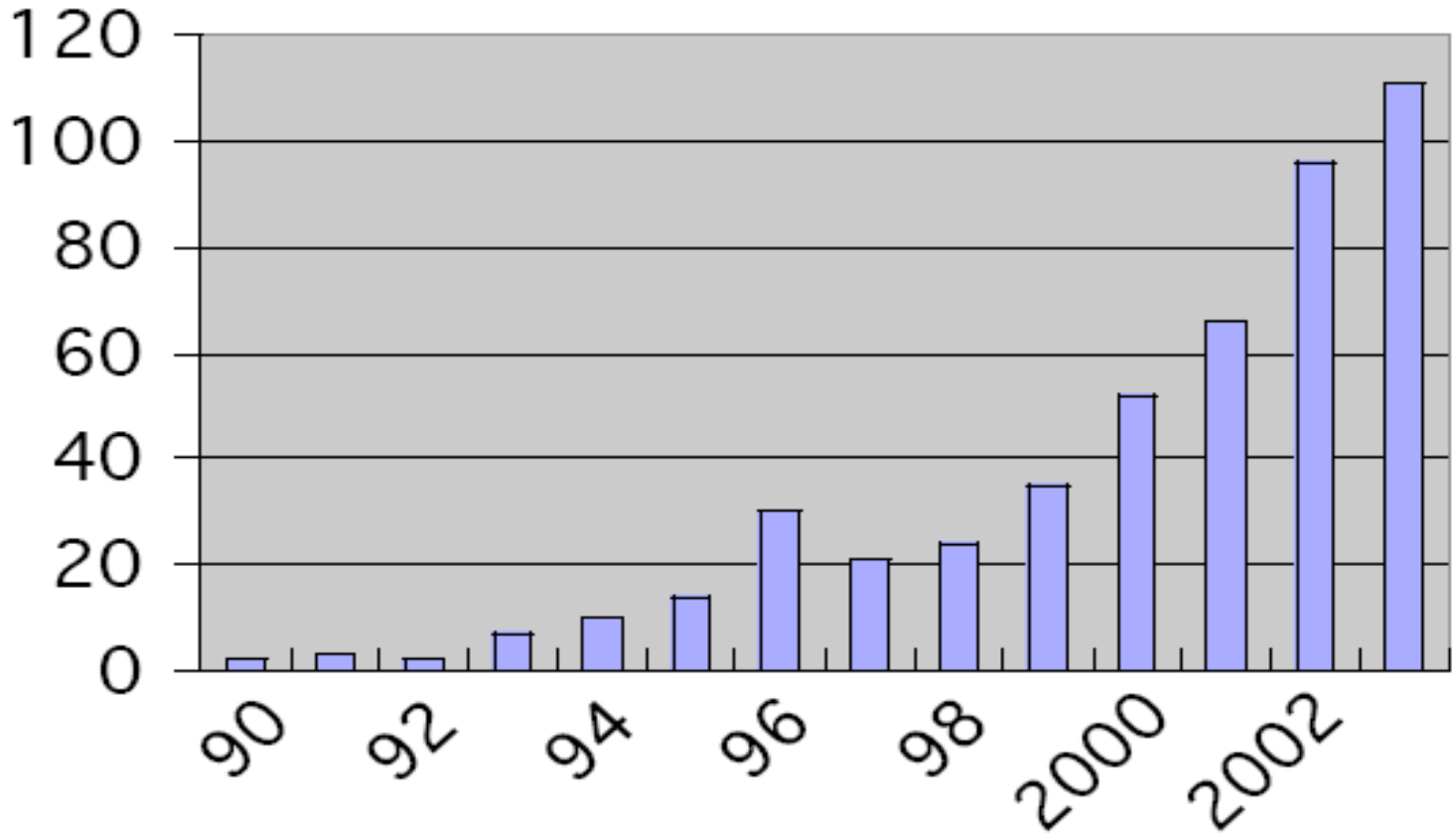
历史回顾：一些重要事件回放 (2)

- 2002年Franz Josef Och提出统计机器翻译的对数线性模型
- 2003年Franz Josef Och提出对数线性模型的最小错误率训练方法
- 2004年Philipp Koehn推出Pharaoh（法老）标志着基于短语的统计翻译方法趋于成熟
- 2005年David Chiang提出层次短语模型并代表UMD在NIST评测中取得好成绩
- 2005年Google在NIST评测中大获全胜，随后Google推出基于统计方法的在线翻译工具，其阿拉伯语-英语的翻译达到了用户完全可接受的水平，目前已经可以支持40多种语言的互译
- 2006年NIST评测中USC-ISI的串到树句法模型第一次超过Google（仅在汉英受限翻译项目中）

统计机器翻译：一种新的研究范式

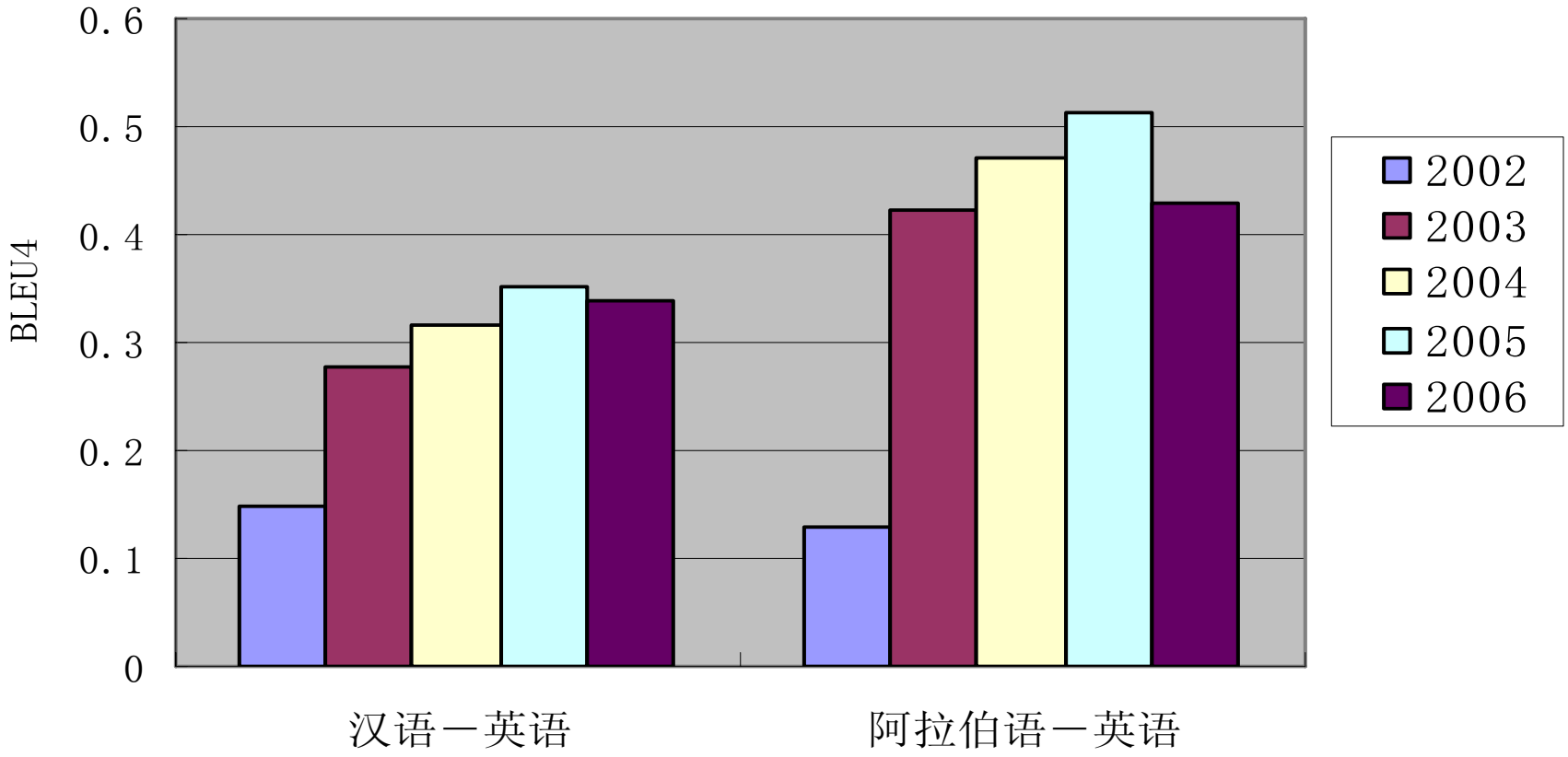
- 统计机器翻译的成功在于采用了一种新的研究范式（paradigm）
- 这种研究范式已在语音识别等领域中被证明是一种成功的翻译，但在机器翻译中是首次使用
- 这种范式的特点：
  - 公开的大规模的训练数据
  - 周期性的公开评测和研讨
  - 开放源码的工具

近年来统计机器翻译论文发表数量

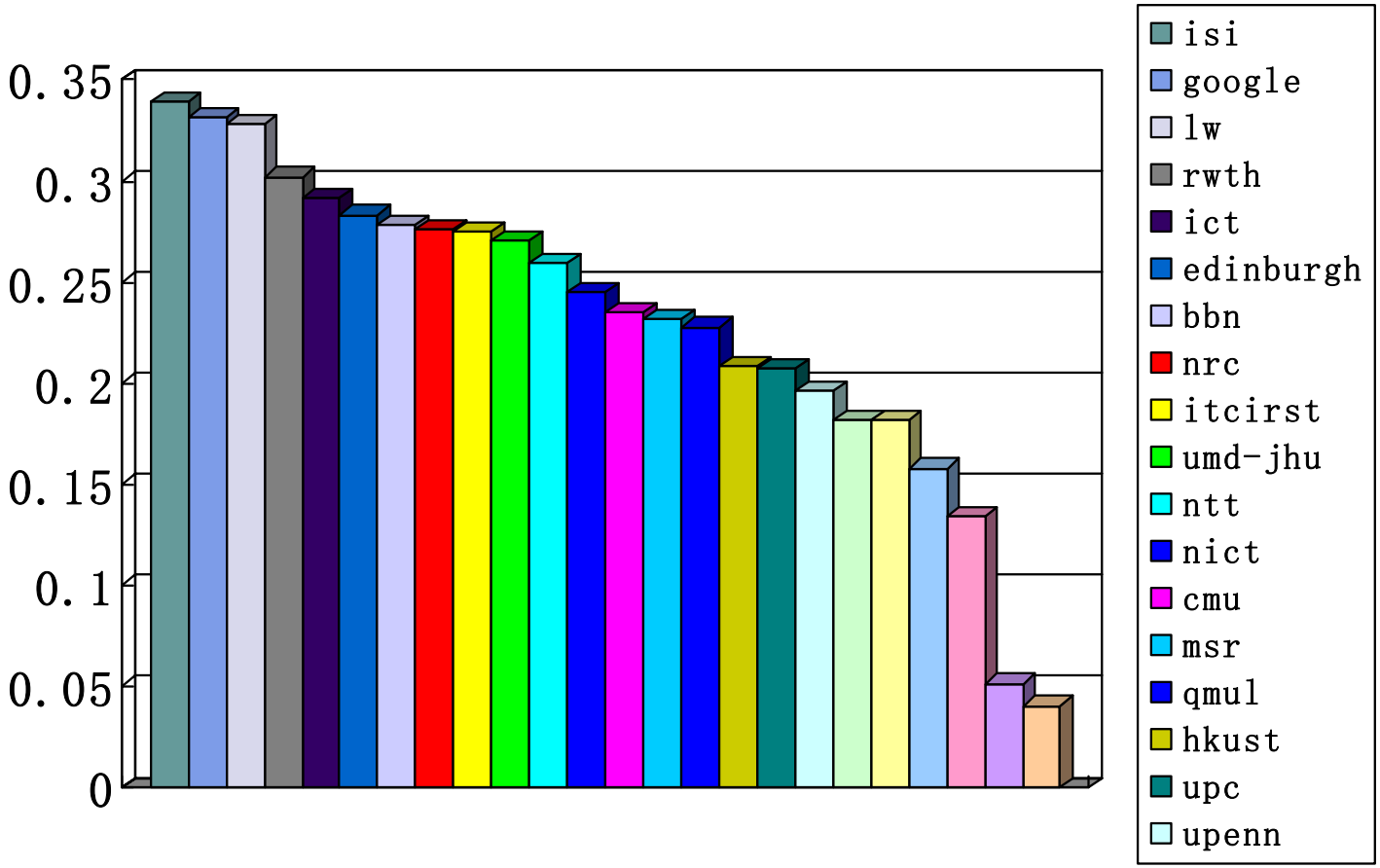


引自 Franz Josef Och, Statistical Machine Translation: Foundations and Recent Advances, Tutorials on MT Summit X, September 13-15, 2005, Phuket, Thailand

近年来国际NIST评测最好成绩



Results on NIST 2006 Evaluation: Large Data Track, NIST Subset



统计机器翻译

- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
  —基于词的IBM模型
- 最成熟的统计机器翻译方法
  —基于短语的模型
- 目前统计机器翻译研究的热点
  —基于句法的模型
- 机器翻译的自动评价

基于词的统计机器翻译方法

- 统计机器翻译—为翻译建立概率模型
- IBM的信源信道模型
- 语言模型—n元语法模型
- 翻译模型—IBM模型1-5
- Candide系统

为翻译建立概率模型

- 假设任意一个英语句子e和一个法语句子 f，我们定义f翻译成e的概率为：

$$\Pr(\boldsymbol{e} \mid \boldsymbol{f})$$

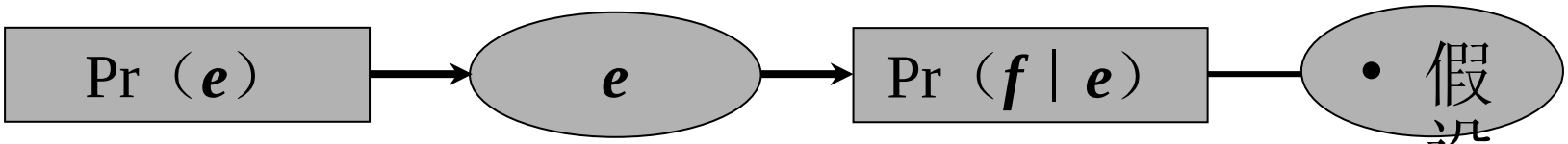
其归一化条件为：

$$\sum_{\boldsymbol{e}} \Pr(\boldsymbol{e} \mid \boldsymbol{f}) = 1$$

- 于是将 f 翻译成 e 的问题就变成求解问题：

$$\hat{\boldsymbol{e}} = \underset{\boldsymbol{e}}{\operatorname{argmax}} \Pr(\boldsymbol{e} \mid \boldsymbol{f})$$

信源信道模型 (1)



- 假设我们看到的源语言文本f 是由一段目标语言文本e 经过某种奇怪的编码得到的，那么翻译的目标就是要将f 还原成e，这也就是就是一个解码的过程。
- 注意，在信源信道模型中：
  - 噪声信道的源语言是翻译的目标语言
  - 噪声信道的目标语言是翻译的源语言

这与整个机器翻译系统翻译方向的刚好相反

信源信道模型 (2)

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e}} \Pr(\boldsymbol{e}) \Pr(\boldsymbol{f} \mid \boldsymbol{e})$$

- P.Brown称上式为统计机器翻译基本方程式
  - 语言模型：*P*(E)
  - 翻译模型：*P*(F|E)
- 语言模型反映“ E像一个句子”的程度：流利度
- 翻译模型反映“F像E”的程度：忠实度
- 联合使用两个模型效果好于单独使用翻译模型，因为后者容易导致一些不好的译文。

信源信道模型 (3)

- 统计机器翻译分解为以下三个问题：
  - 语言模型的定义和参数估计
  - 翻译模型的定义和参数估计
  - 解码

语言模型 — n元语法模型

- 语言模型在机器翻译中具有极为重要的作用
- 到目前位置，统计机器翻译中最常用、而且最有效的模型仍然是n元语法模型
- 模型的阶数越来越高：3元、4元、5元
- 模型的训练语料越来越大：
  - Google提供了公开的Web 1T语料库，其中的n元共现词频数据是从web中得到的1T英文词的语料库中统计得到的（剪切掉了低频组合）
  - Google号称使用了2T英文词训练的语言模型
  - 大规模的数据为系统实现带来很大的困难

翻译模型

- 翻译模型$P(F|E)$反映的是一个源语言句子E翻译成一个目标语言句子F的概率
- 由于源语言句子和目标语言句子几乎不可能在语料库中出现过，因此这个概率无法直接从语料库统计得到，必须分解成词语翻译的概率和句子结构（或者顺序）翻译的概率

翻译模型与对齐

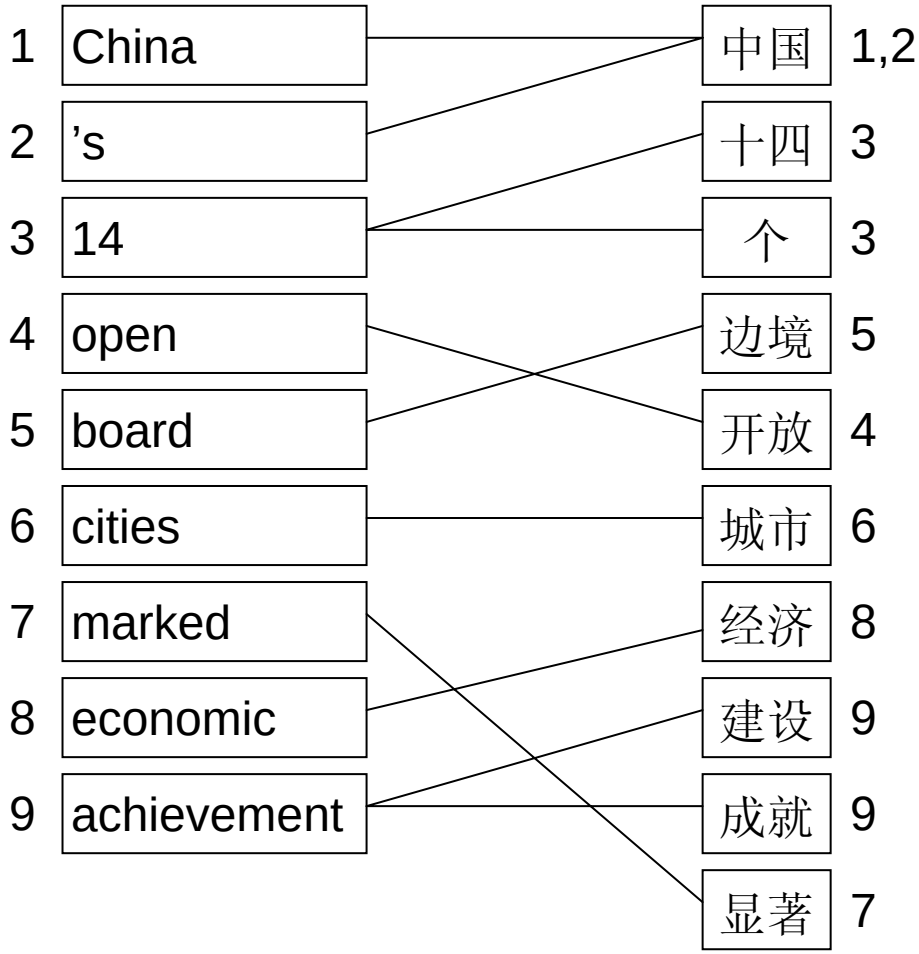
- 翻译模型的计算，需要引入隐含变量：对齐***A***：

$$P(\mathrm{F}\,|\,\mathrm{E}) = \sum_{A} P(\mathrm{F}, \mathrm{A}\,|\,\mathrm{E})$$

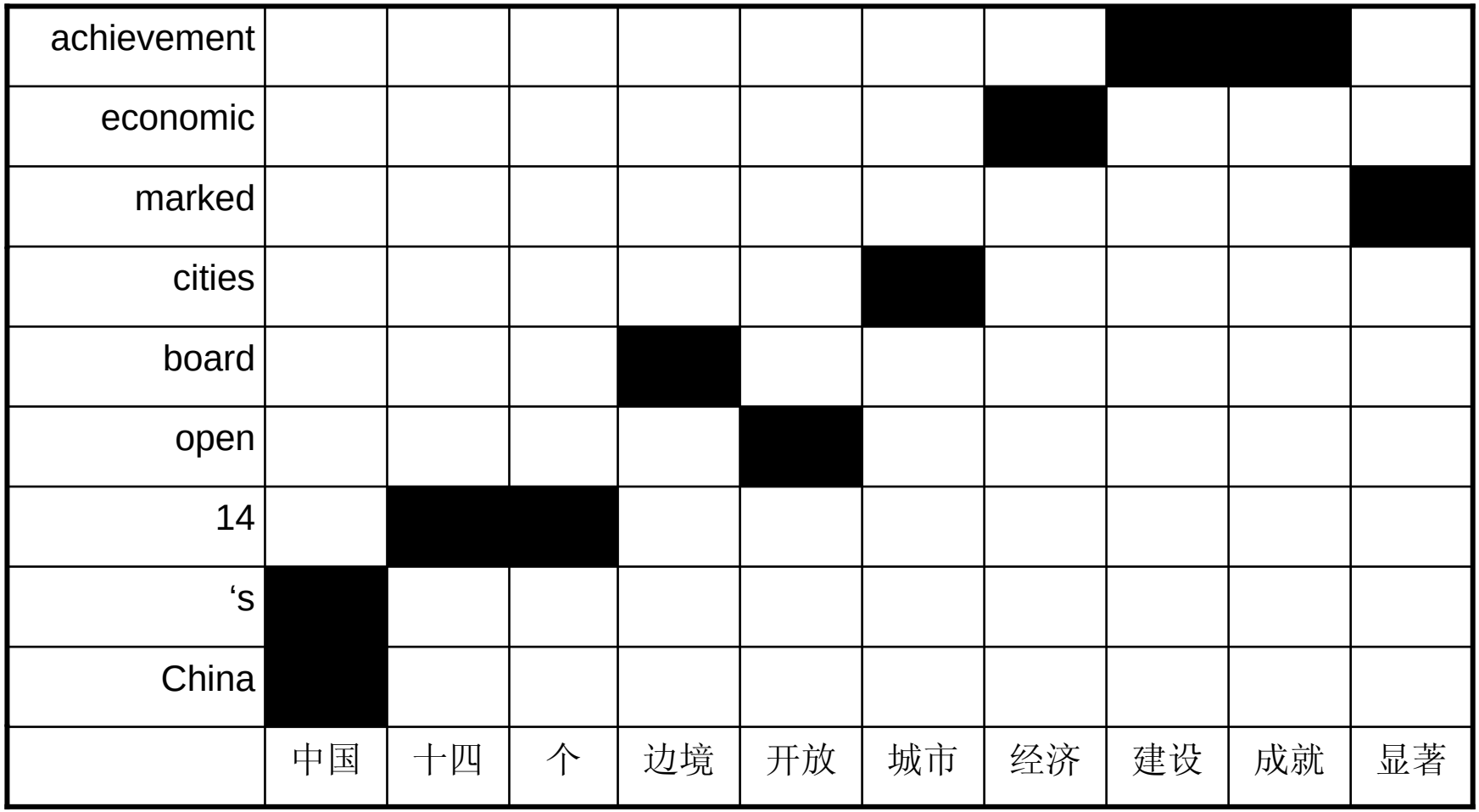
- 翻译概率*P*(F|E)的计算转化为对齐概率*P*(F,A|E)的估计
- 对齐：建立源语言句子和目标语言句子的词与词之间的对应关系和句子结构之间的对应关系

词语对齐的表示 (1)

- 图形表示
  - ✓ 连线
  - ✓ 矩阵（见下页）
- 数字表示
  - ✓ 给每个目标语言单词标记其所有对应的源语言单词



词语对齐的表示 (2)


achievement
economic
marked
cities
board
open
14
‘s
China
中国
十四
个
边境
开放
城市
经济
建设
成就
显著

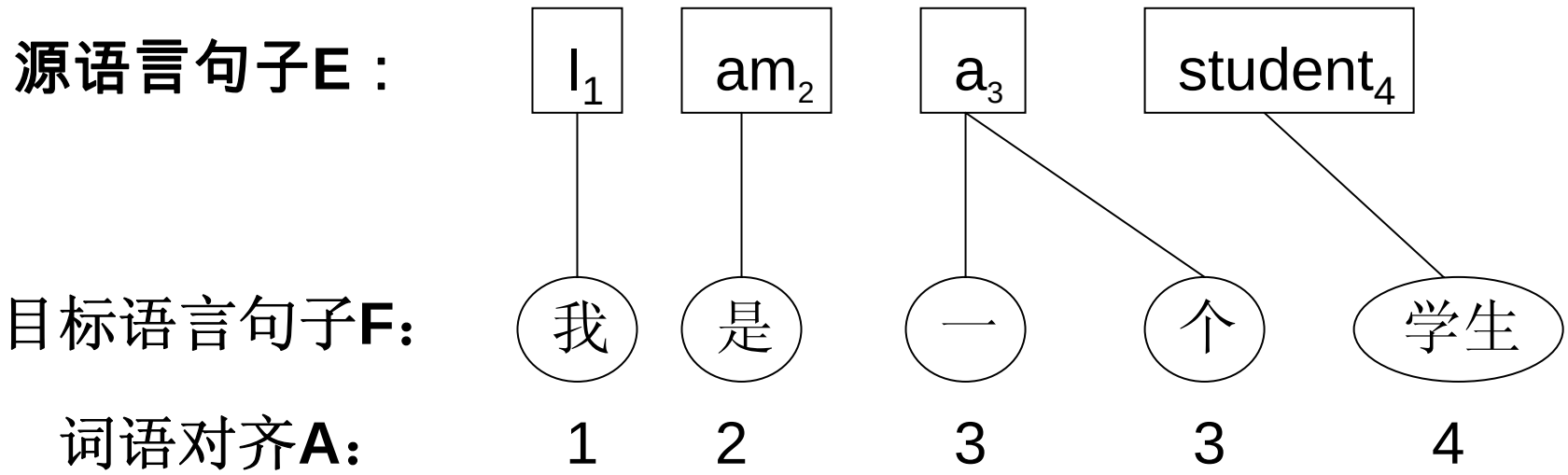
IBM Model 1

- 最简单的理解，可以句子e翻译成f的概率，就是e中每一个词语翻译成f中对应词语的概率的乘积
- 这就是IBM Model 1的基本思想
- IBM提出了复杂度递增的5个统计翻译模型，IBM Model 1是其中最简单的模型

IBM Model 1-5

- IBM Model 1仅考虑词对词的互译概率
- IBM Model 2加入了词的位置变化的概率
- IBM Model 3加入了一个词翻译成多个词的概率
- IBM Model 4
- IBM Model 5

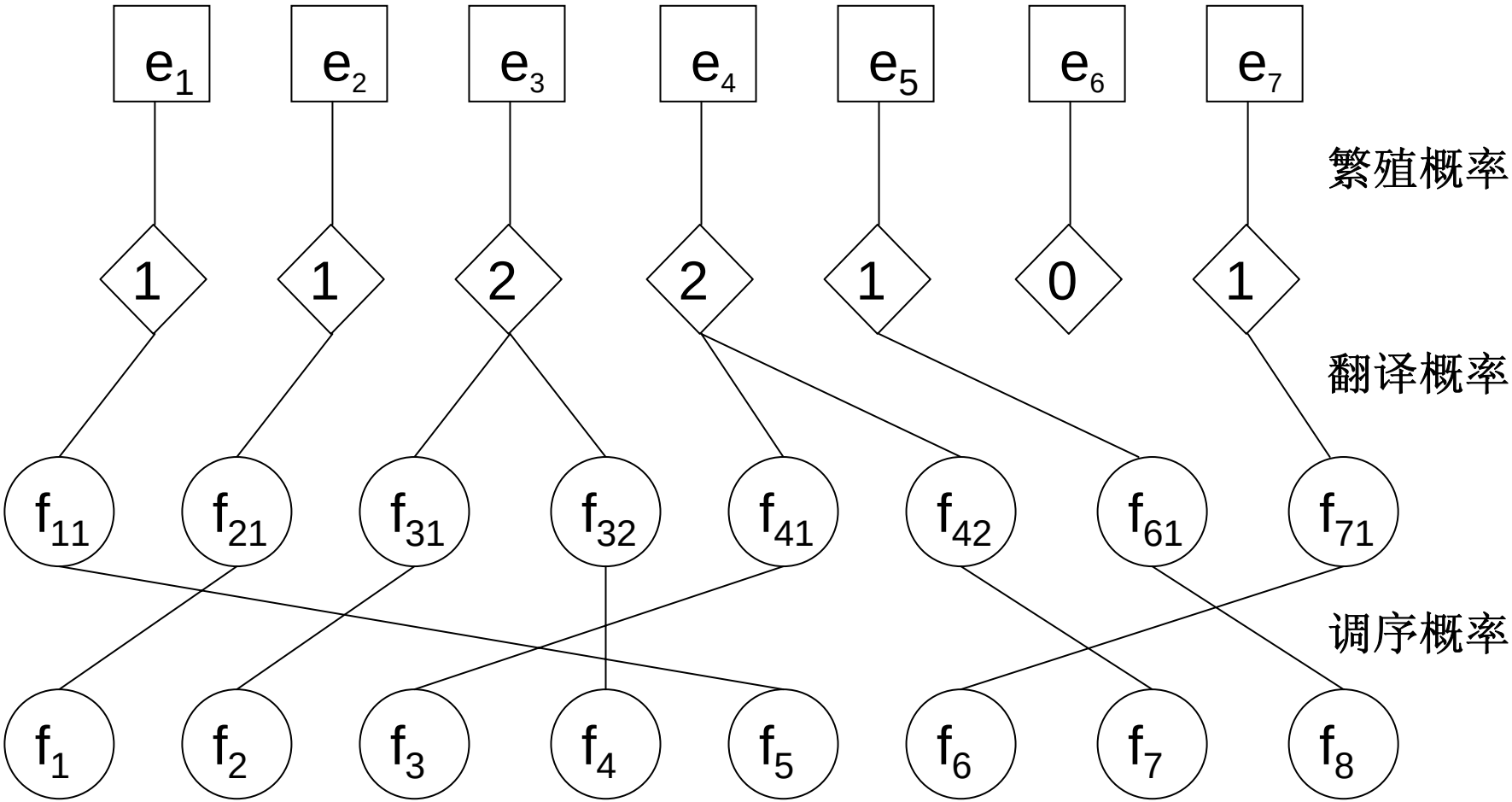
IBM Model 1 & 2 推导方式



IBM模型1&2的推导过程：

1. 猜测目标语言句子长度；
2. 从左至右，对于每个目标语言单词：
   - 首先猜测该单词由哪一个源语言单词翻译而来；
   - 再猜测该单词应该翻译成什么目标语言词。

IBM Model 3 & 4 & 5 推导方式



IBM公司的Candide系统(1)

- 基于统计的机器翻译方法
- 分析—转换—生成
  - 中间表示是线性的
  - 分析和生成都是可逆的
- 分析（预处理）：

  1.短语切分 2.专名与数词检测

  3.大小写与拼写校正

  4.形态分析 5.语言的归一化

IBM公司的Candide系统(2)

- 转换（解码）：基于统计的机器翻译
- 解码分为两个阶段：
  - 第一阶段：使用粗糙模型的堆栈搜索
    - 输出140个评分最高的译文
    - 语言模型：三元语法
    - 翻译模型：EM Trained IBM Model 5
  - 第二阶段：使用精细模型的扰动搜索
    - 对第一阶段的输出结果先扩充，再重新评分
    - 语言模型：链语法
    - 翻译模型：最大熵翻译模型（选择译文词）

IBM公司的Candide系统(3)

- ARPA的测试结果：

	Fluency		Adequacy		Time Ratio	
	1992	1993	1992	1993	1992	1993
Systran	.466	.540	.686	.743		
Candide	.511	.580	.575	.670		
Transman	.819	.838	.837	.850	.688	.625
Manual		.833		.840		

统计机器翻译

- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
  —基于词的IBM模型
- 最成熟的统计机器翻译方法
  —基于短语的模型
- 目前统计机器翻译研究的热点
  —基于句法的模型
- 机器翻译的自动评价

基于短语的统计机器翻译方法

- 从信源信道模型到对数线性模型
- 翻译模型的发展—基于短语的模型
- 短语的自动抽取
- 短语语序的调整

统计机器翻译的对数线性模型(1)

- Och于ACL2002提出，思想来源于Papineni提出的基于特征的自然语言理解方法，该论文获得ACL2002的最佳论文称号
- 是一个比信源—信道模型更具一般性的模型，信源—信道模型是其一个特例
- 原始论文的提法是“最大熵”模型，现在通常使用“对数线性（Log-Linear）模型”这个概念。“对数线性模型”的含义比“最大熵模型”更宽泛，而且现在这个模型通常都不再使用最大熵的方法进行参数训练，因此“对数线性”模型的提法更为准确。
- 与NLP中通常使用的最大熵方法的区别：使用连续量（实数）作为特征，而不是使用离散的布尔量（只取0和1值）作为特征

统计机器翻译的对数线性模型(2)

假设*e*、*f*是机器翻译的目标语言和源语言句子，$h_1(e,f), \ldots, h_M(e,f)$分别是*e*、*f*上的M个特征，$\lambda_1, \ldots, \lambda_M$是与这些特征分别对应的M个参数，那么翻译概率可以用以下公式模拟：

$$\Pr(e \mid f) \approx p_{\lambda_1 \ldots \lambda_M}(e \mid f)$$

$$= \frac{\exp[\sum_{m=1}^{M} \lambda_m h_m(e, f)]}{\sum_{e'} \exp[\sum_{m=1}^{M} \lambda_m h_m(e', f)]}$$

统计机器翻译的对数线性模型(3)

对于给定的*f*，其最佳译文*e*可以用以下公式表示：

$$\hat{e} = \arg\max_{e}\{\Pr(e \mid f)\}$$

$$\approx \arg\max_{e}\{\sum_{m=1}^{M}\lambda_m h_m(e, f)\}$$

对数线性模型vs.噪声信道模型

- 取以下特征和参数时，对数线性模型等价于噪声信道模型：
  - 仅使用两个特征
  - $h_1(e,f)=\log p(e)$
  - $h_2(e,f)=\log p(f|e)$
  - $\lambda_1=\lambda_2=1$

对数线性模型的优点

- 噪声模型只有在理想的情况下才能达到最优，对于简化的语言模型和翻译模型，取不同的参数值实际效果更好；
- 对数线性模型大大扩充了统计机器翻译的思路；
- 特征的选择更加灵活，可以引入任何可能有用的特征。

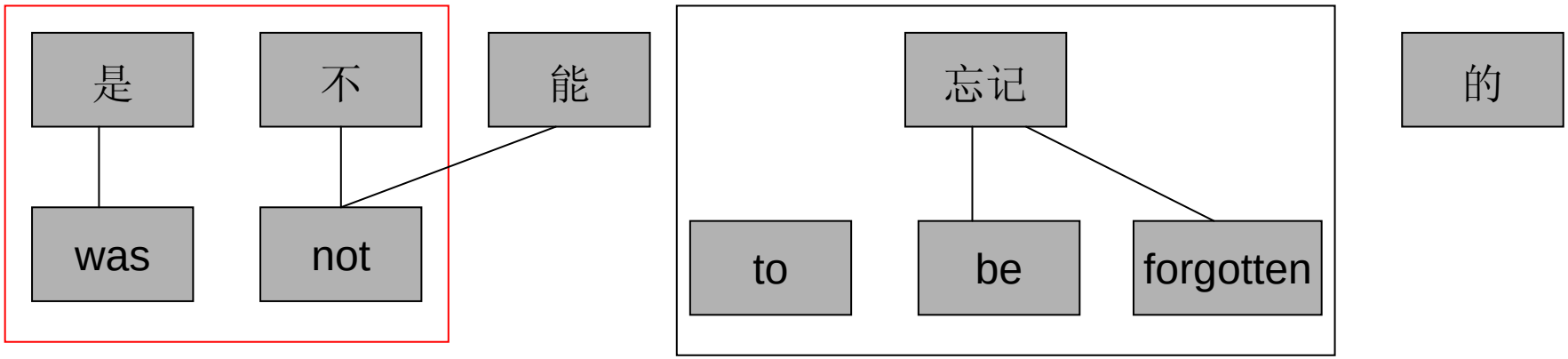
翻译模型的发展—基于短语的模型

- 基于词的IBM翻译模型有明显的缺陷：一个词在翻译的时候基本上不考虑上下文，孤立地进行翻译，导致了大量的错误；词序调整模型近乎无礼，很难准确调整词序，对词序差别较大的语言之间的翻译效果太差。
- 人们很容易想到，将一个短语捆绑起来进行翻译，可以大大提高翻译的准确率
- 很多不同的研究人员尝试了各种各样的基于短语的翻译模型，最终形成了目前比较成熟的基于短语的翻译模型

基于短语的翻译模型

- 基本思想
  - 把训练语料库中所有对齐的短语及其翻译概率存储起来，作为一部带概率的短语词典
  - 这里所说的短语是任意连续的词串，不一定是一个独立的语言单位
  - 翻译的时候将输入的句子与短语词典进行匹配，选择最好的短语划分，将得到的短语译文重新排序，得到最优的译文
- 问题：
  - 短语如何抽取？
  - 短语概率如何计算？

基于词语对齐的短语自动抽取

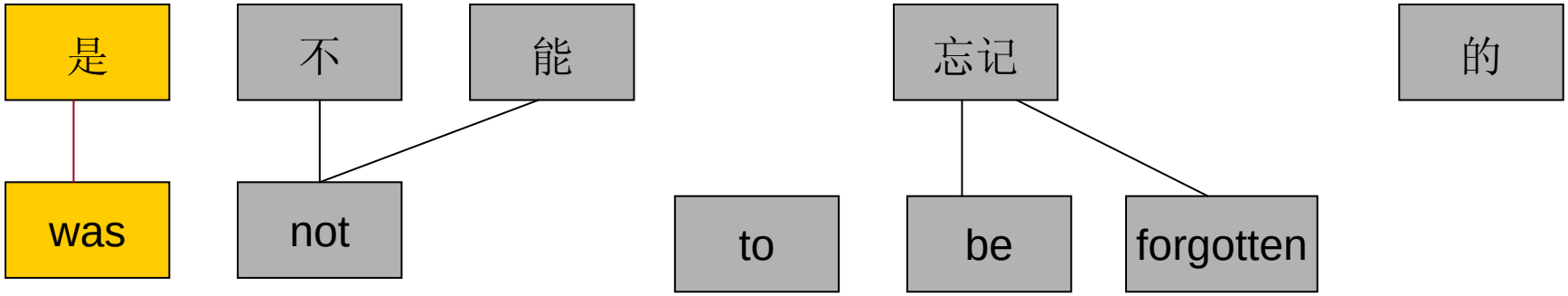


不相容

相容

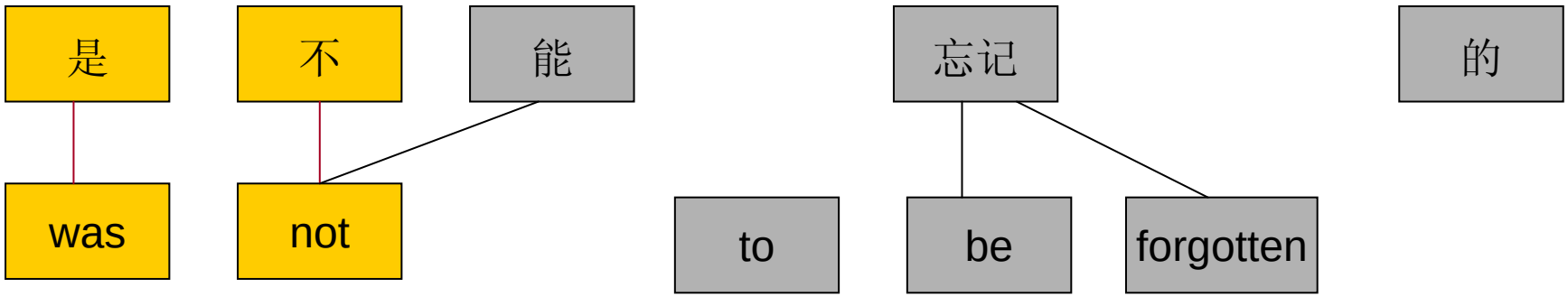
短语自动抽取算法运行示例 (1)

- 列举源语言所有可能的短语，
  根据对齐检查相容性



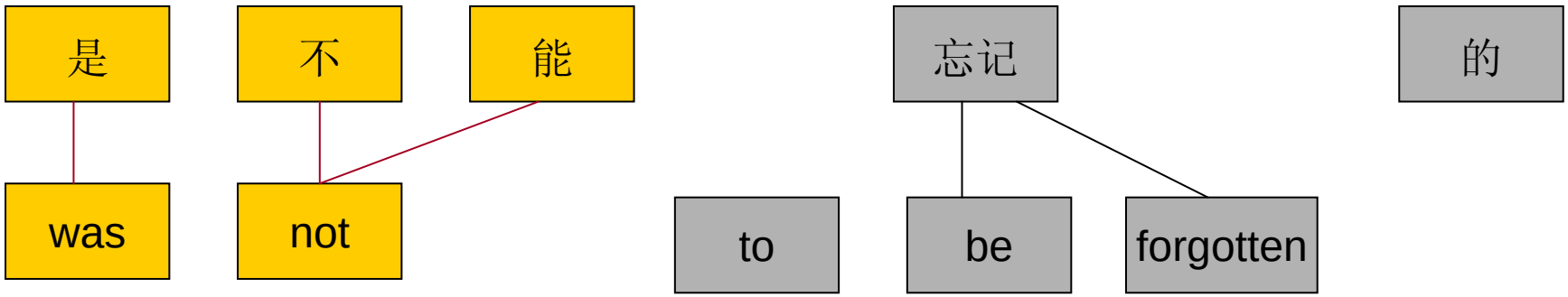
（是，was）

短语自动抽取算法运行示例(2)



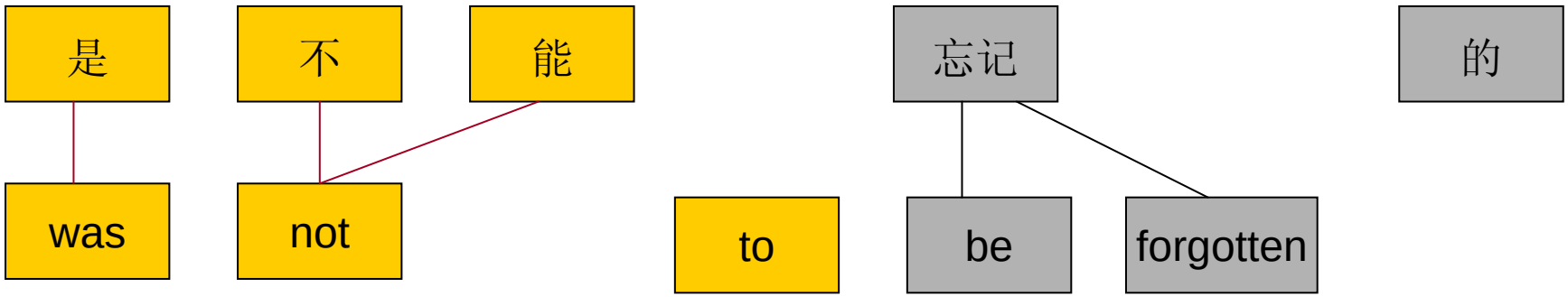
不相容

短语自动抽取算法运行示例(3)



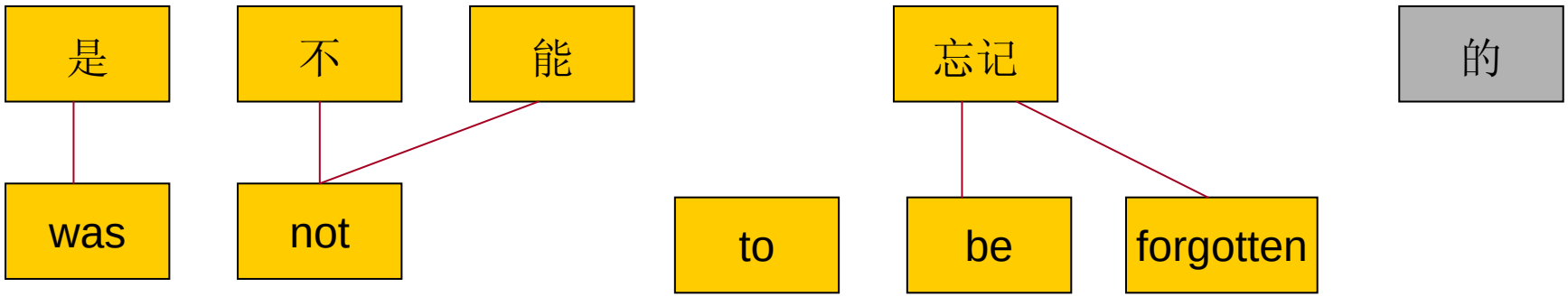
（是不能，was not)

短语自动抽取算法运行示例(4)



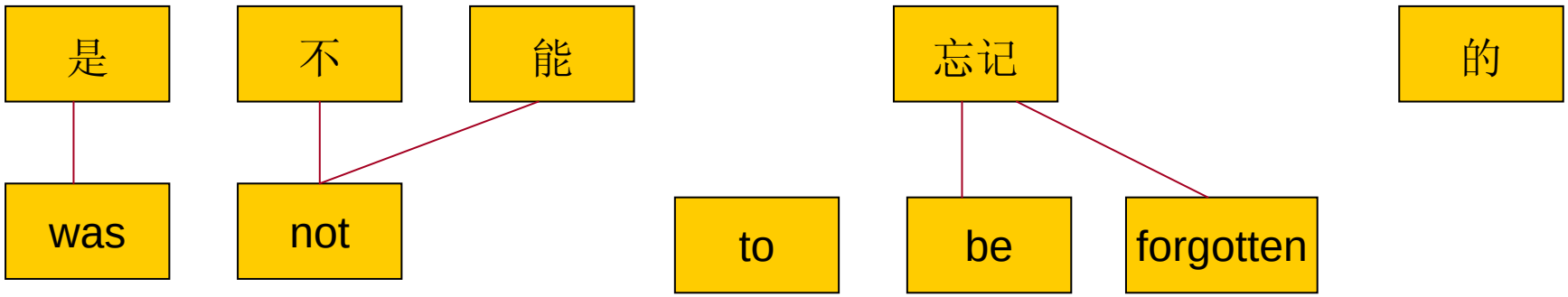
（是不能，was not to)

短语自动抽取算法运行示例(5)



（是不能忘记，was not to be forgotten)

短语自动抽取算法运行示例(6)



（是不能忘记的，was not to be forgotten)

短语表

是	was
是不能	was not
是不能	was not to
是不能忘记	was not to be forgotten
是不能忘记的	was not to be forgotten
不 能	not
不 能	not to
不 能 忘记	not to be forgotten
不 能 忘记 的	not to be forgotten
忘记	be forgotten
忘记	to be forgotten
忘记 的	be forgotten
忘记 的	to be forgotten

短语语序的调整

- 在基于短语的模型中，短语内部的顺序无需调整，只需要调整短语之间的顺序
- 短语的调序模型类似于基于词的模型，允许任意的语序调整
- 为了避免搜索空间的过于膨胀，通常限制语序调整的距离

统计机器翻译

- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
  —基于词的IBM模型
- 最成熟的统计机器翻译方法
  —基于短语的模型
- 目前统计机器翻译研究的热点
  —基于句法的模型
- 统计机器翻译面临的问题和展望

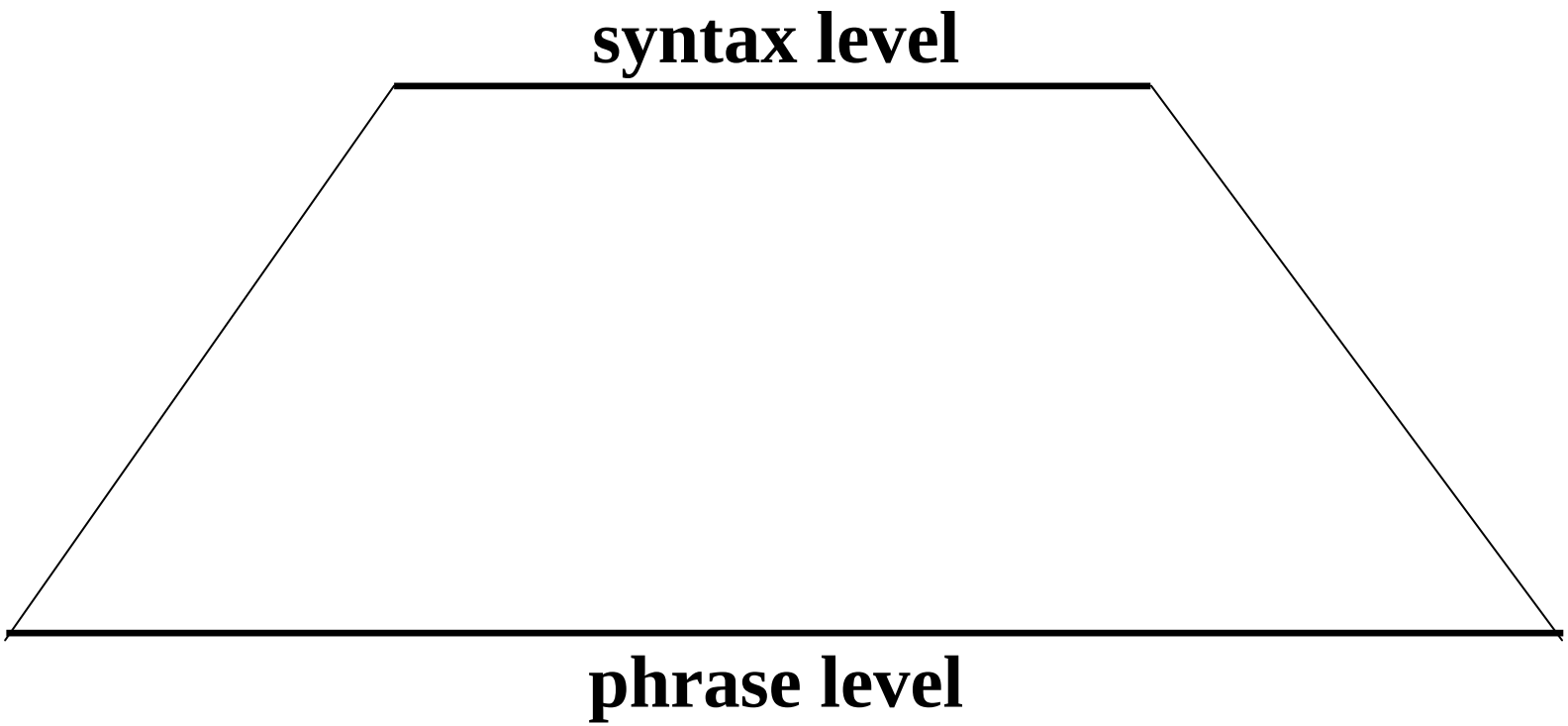
基于句法的模型

- 翻译模型的发展—基于句法的模型
- 基于句法的模型概述
- 形式上基于句法的模型
  - ITG和BTG
  - 最大熵BTG模型
  - 层次短语模型
- 语言学上基于句法的模型
  - 树到串模型
  - 串到树模型

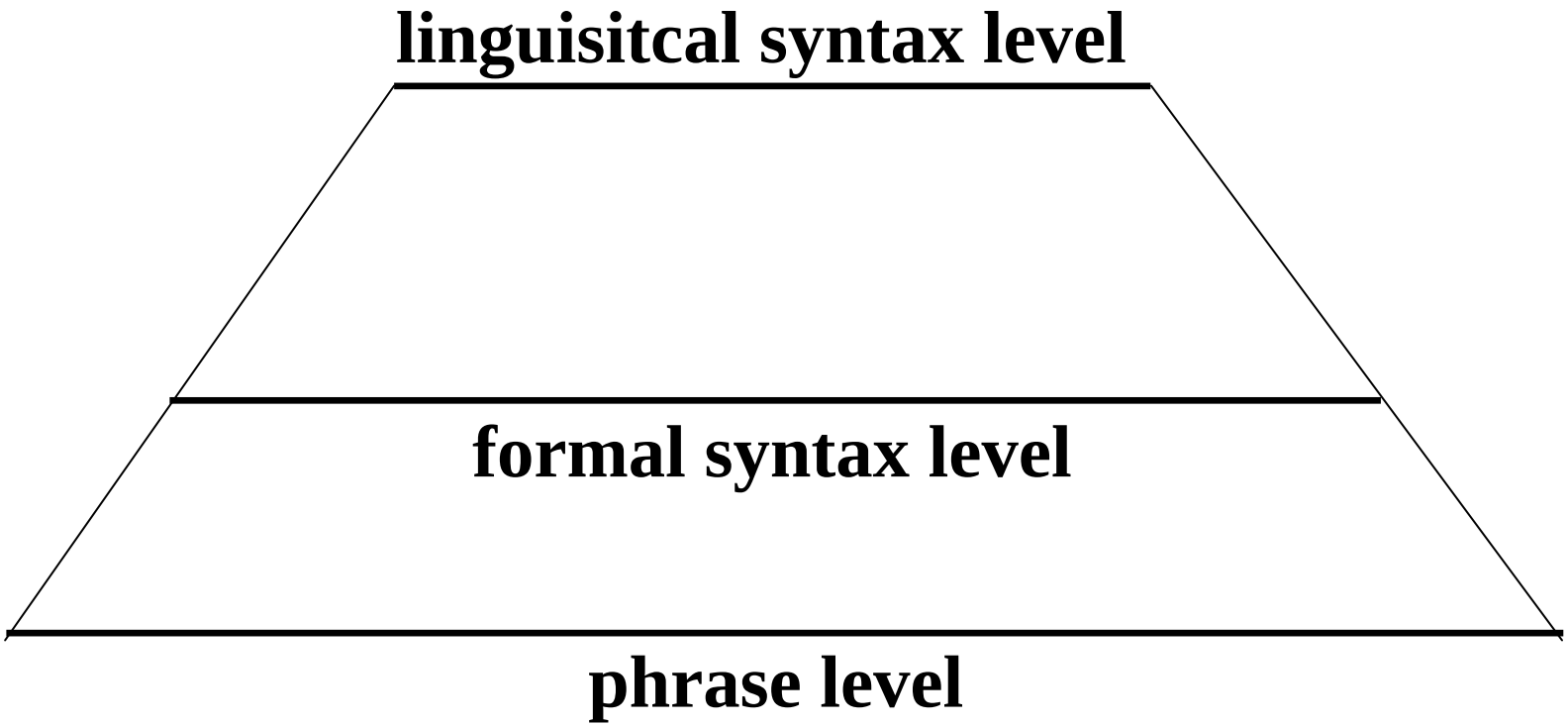
翻译模型的发展—基于句法的模型

- 基于短语的模型比基于词的模型性能有了较大提高，但对于短语之间的语序调整，仍然没有提供合理的解决方案
- 经验表明，在基于短语的统计机器翻译系统中，绝大多数匹配的短语长度都是2-3个词，1个词的短语也占相当大的比例
- 要解决长距离语序的调整，引入句法信息是个必然的选择

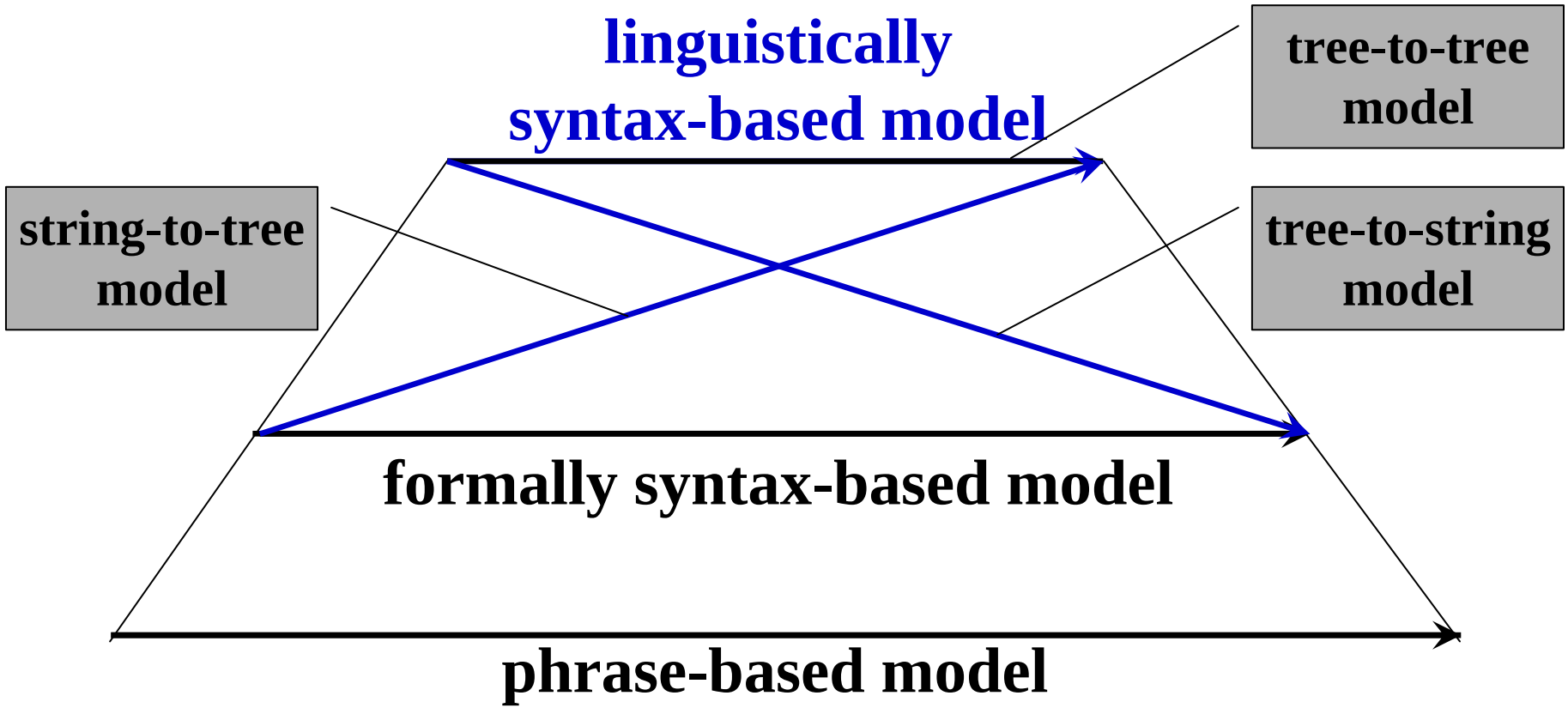
基于句法的统计翻译模型(1)



基于句法的统计翻译模型(1)



基于句法的统计翻译模型(1)



基于句法的统计翻译模型 (2)

- 基于句法的统计翻译模型，通常的做法都是分别为源语言和目标语言句子建立某种句法结构，并在这两种句法结构之间建立某种对应关系
- 基于句法的统计翻译模型有两种不同的做法
  - 形式上基于句法的统计翻译模型：并不采用语言学上的句法分析，而是从词语对齐的双语语料库中自动获取某种双语平行的句法结构
  - 语言学上基于句法的统计翻译模型：利用语言学上的句法分析，为源语言句子和目标语言句子建立句法结构，并借助词语对齐建立句法结构的对应关系

基于句法的统计翻译模型 (3)

- 语言学上基于句法的统计翻译模型又有三种不同的做法
  - 树到串模型：在源语言端进行句法分析并得到源语言句法结构，然后根据词语对齐建立对应的目标语言句法结构（可称为伪句法结构）
  - 串到树模型：在目标语言端进行句法分析并得到目标语言句法结构，然后根据词语对齐建立对应的源语言句法结构（也是伪句法结构）
  - 树到树模型：在源语言端和目标语言端分别进行句法分析并得到双语的句法结构，然后根据词语对齐建立这两种句法结构之间的对应关系

形式上基于句法的模型

- 反向转录语法（ITG）和括号转录语法（BTG）
  Inversion (Bracketing) Transduction Grammar (ITG,BTG), Dekai Wu 1997
- 有限状态中心词转录机
  Finite-State Head Transducer, Alshawi 2000
- 基于层次短语的翻译模型
  Hierarchical Phrase-based Model, David Chiang 2005
- 最大熵括号匹配语法的翻译模型
  Maximal Entropy Bracket Transduction Grammar (ME-BTG), Deyi Xiong 2006

语言学上基于句法的模型

- 串到树模型 String-to-Tree Model
  - 美国南加州大学信息科学研究所（ISI/CSU）的工作

    Yamada 2001, Galley 2006, Marcu 2006
- 树到串模型 Tree-to-String Model
  - 中科院计算所的工作

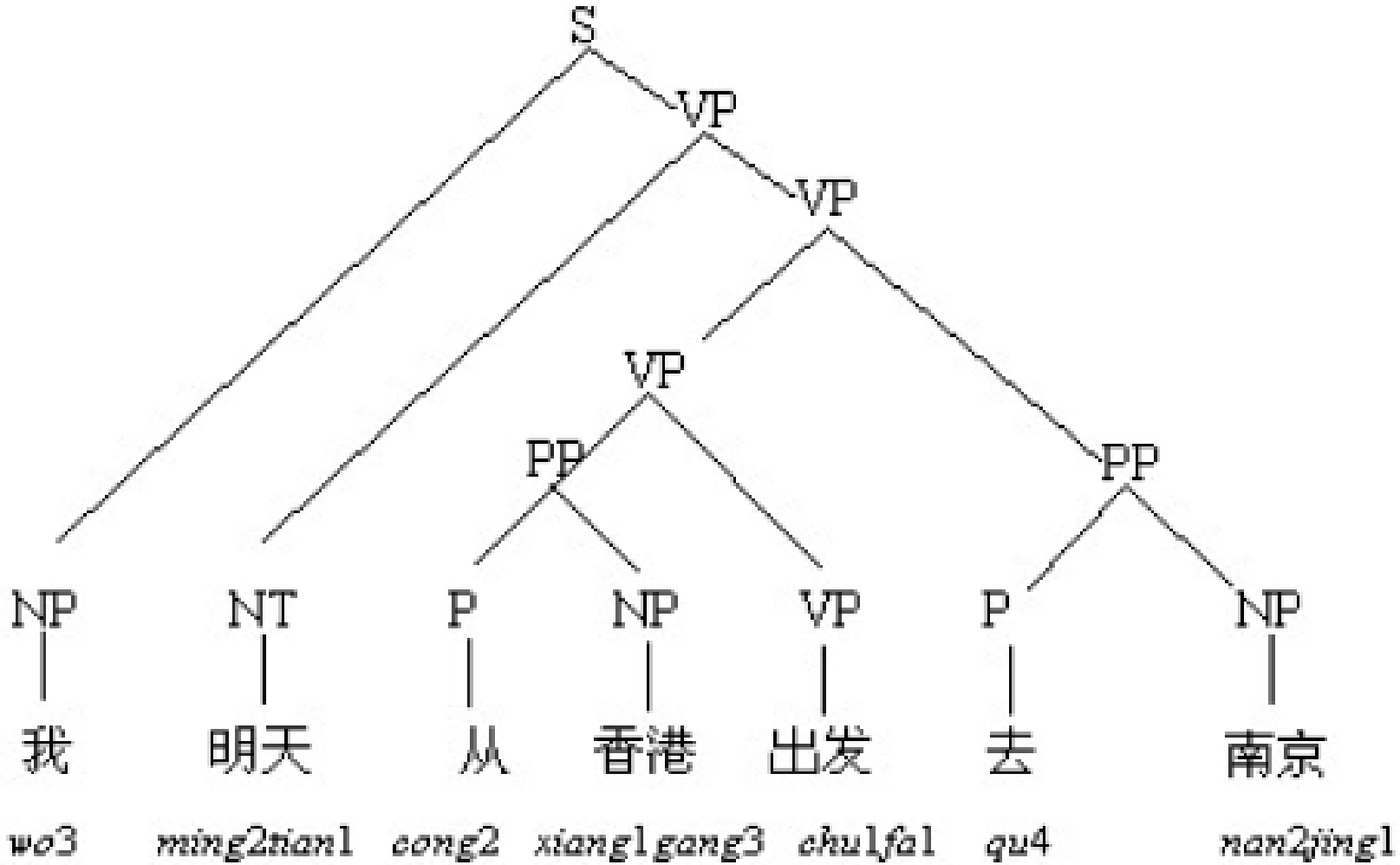
    Tree-to-string Alignment Template Model (TAT), Liu Yang 2006
  - 微软研究院的工作（依存模型）

    Dependency Treelet Translation, Quirk 2005
- 树到树的模型 Tree-to-Tree Model

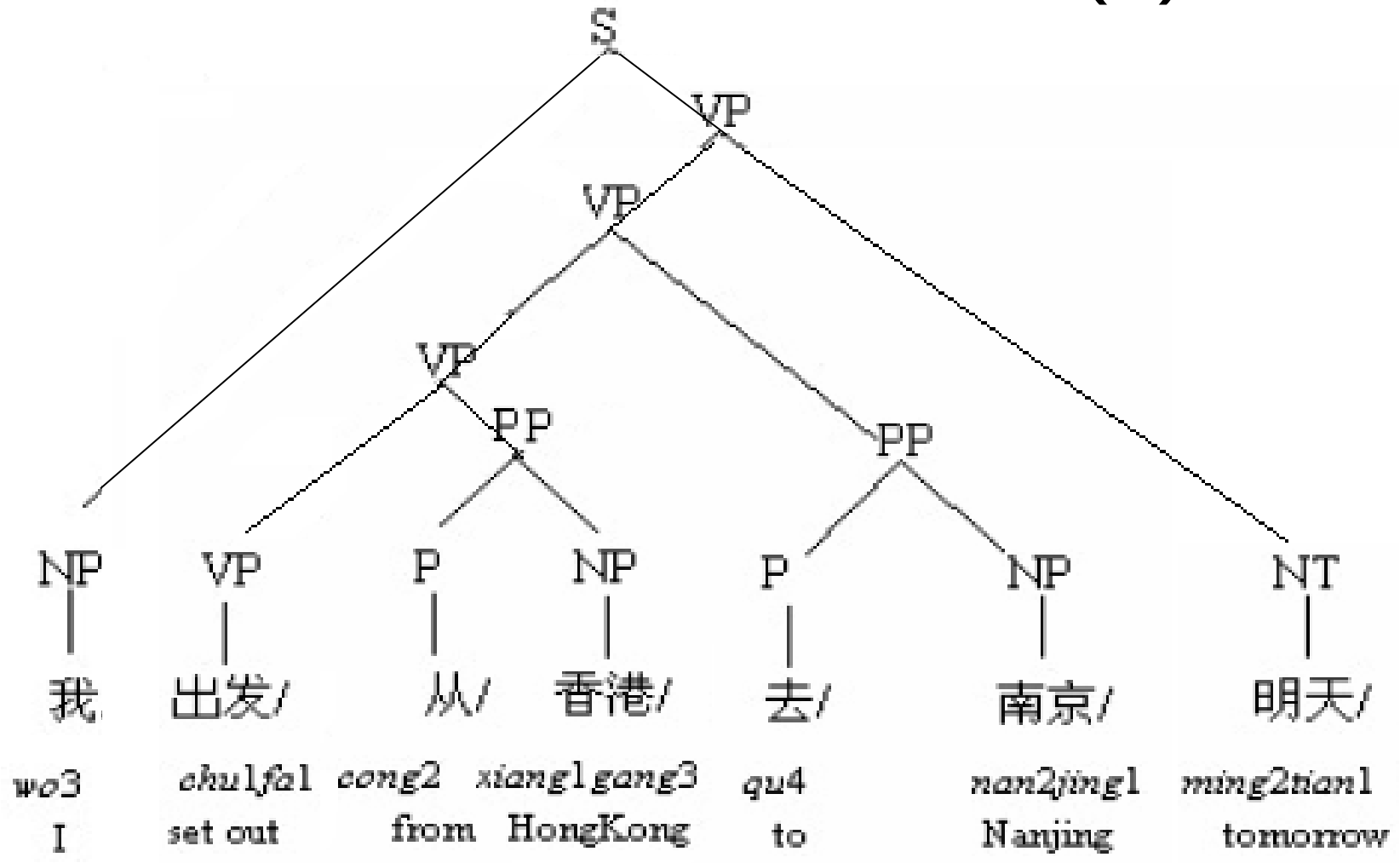
基于 ITG的机器翻译 (1)

- 训练：
  从词语对齐的双语语料库中自动获得ITG规则
- 解码：
  类似于传统的基于规则的机器翻译方法
  - 先用ITG的源语言端规则对源语言进行句法分析
  - 根据ITG规则的映射关系，确定源语言句法树中每条源语言句法规则对应的目标语言句法规则
  - 生成目标语言句法树

基于 ITG的机器翻译 (2)



基于 ITG的机器翻译 (3)



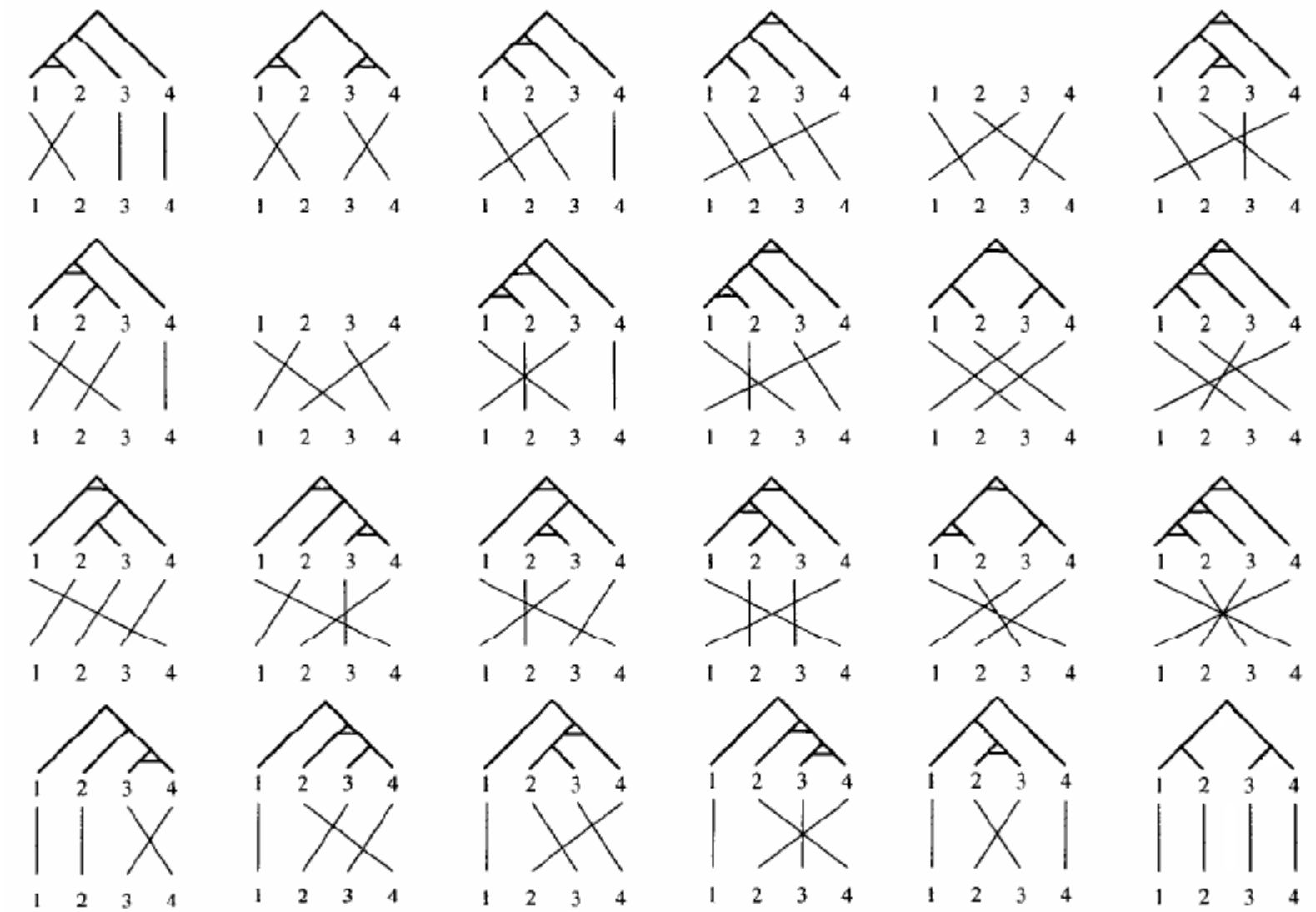
基于 ITG的机器翻译 (4)

- 在ITG中，仍然使用了NP、VP之类的句法标记，这对于训练语料库提出了比较高的要求
- 如果我们不考虑标记，也就是说，认为所有的标记都是相同的，只有一个非终结符标记X，那么ITG就退化成BTG

ITG约束

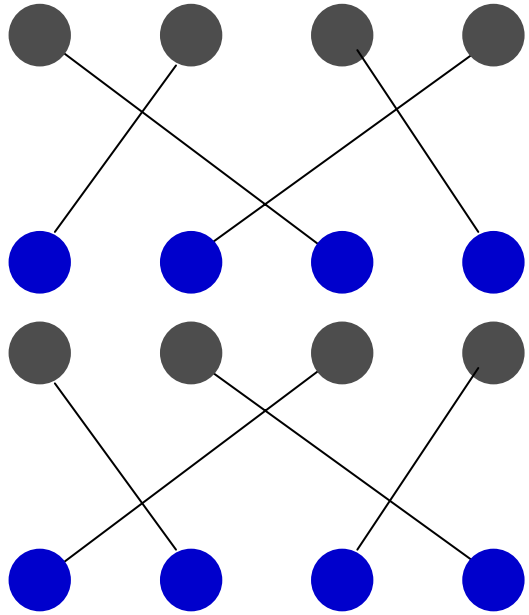
- BTG约束（BTG constraint）
  - 只有满足某种BTG对应关系的目标语言词序才是允许的，否则排除在搜索空间之外
  - 解码的时候，采用类似于CYK句法分析的方式进行解码，就可以穷尽所有可能的BTG约束下的词序
  - 在BTG约束下，可能的对齐方式是多项式级的
  - 无需限制长距离的词序调整

这里给出了四个词在BTG约束下所有可能的词序调整方案
其中有两种方案在BTG约束下是不允许的



ITG约束(3)

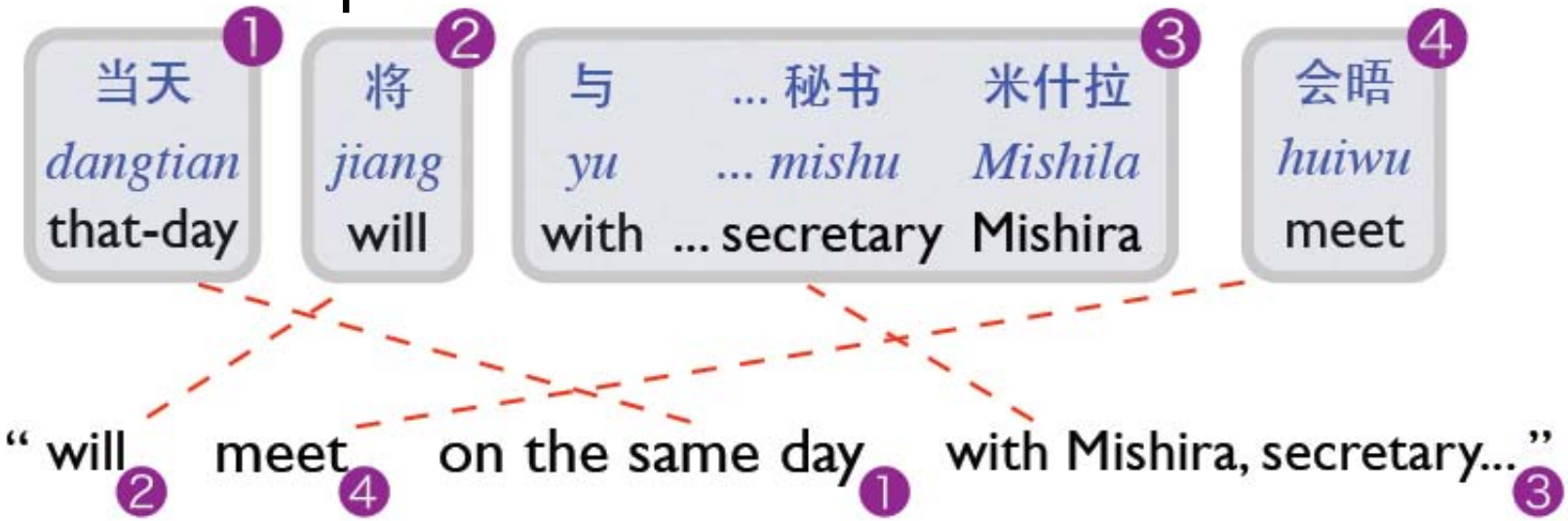
f	BTG	all matchings	ratio
0	1	1	1.000
1	1	1	1.000
2	2	2	1.000
3	6	6	1.000
4	22	24	0.917
5	90	120	0.750
6	394	720	0.547
7	1806	5040	0.358
8	8558	40320	0.212
9	41586	362880	0.115
10	206098	3628800	0.057
11	1037718	39916800	0.026
12	5293446	479001600	0.011
13	27297738	6227020800	0.004
14	142078746	87178291200	0.002
15	745387038	1307674368000	0.001
16	3937603038	20922789888000	0.000



word reordering which are not permitted in BTG

ITG约束 (4) 一个反例

- For Chinese and English, almost true.
  - an exception:



- For some other languages with free order, not true.

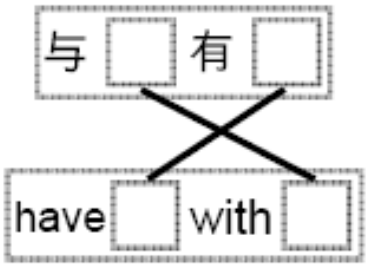
层次短语模型 (1)

- David Chiang. A Hierarchical Phrase-Based Model for Statistical Machine Translation. ACL2005. (Best Paper Award)
- 本讲义这一部分内容直接引用了以下讲义的部分内容，特此说明并向原作者表示感谢：
  - David Chiang, Hiero: Finding Structure in Statistical Machine Translation, in National University of Singapore

层次短语模型 (2)

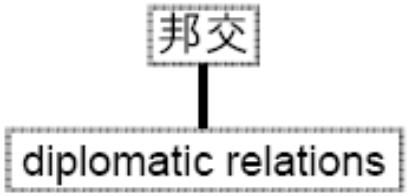
- 传统的基于短语的翻译模型中，短语是平面的，不能嵌套
- 在层次短语模型中，引入了嵌套的层次短语
- 采用平行上下文无关语法作为理论基础，但只使用唯一的非终结符标记
- 效果比传统的短语模型有很大提高

用同步语法表示层次短语 (1)



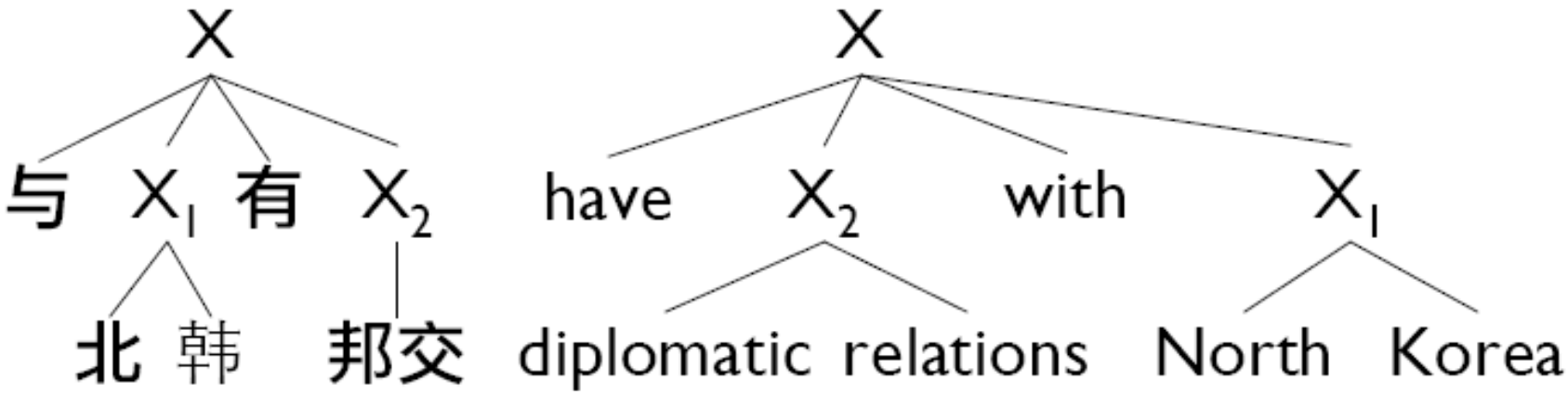
(X → 与 X_1 有 X_2, X → have X_2 with X_1)

(X → 北 韩, X → North Korea)



(X → 邦交, X → diplomatic relations)

用同步语法表示层次短语 (2)



规则举例

X → 的	X → 's
X → X_1 的 X_2	X → the X_2 of X_1
X → X_1 的 X_2	X → the X_2 that X_1
X → 在	X → in
X → 在 X_1 下	X → under X_1
X → 在 X_1 前	X → before X_1
X → 今年 X_1	X → X_1 this year
X → X_1 之一	X → one of X_1
X → X_1 总统	X → president X_1

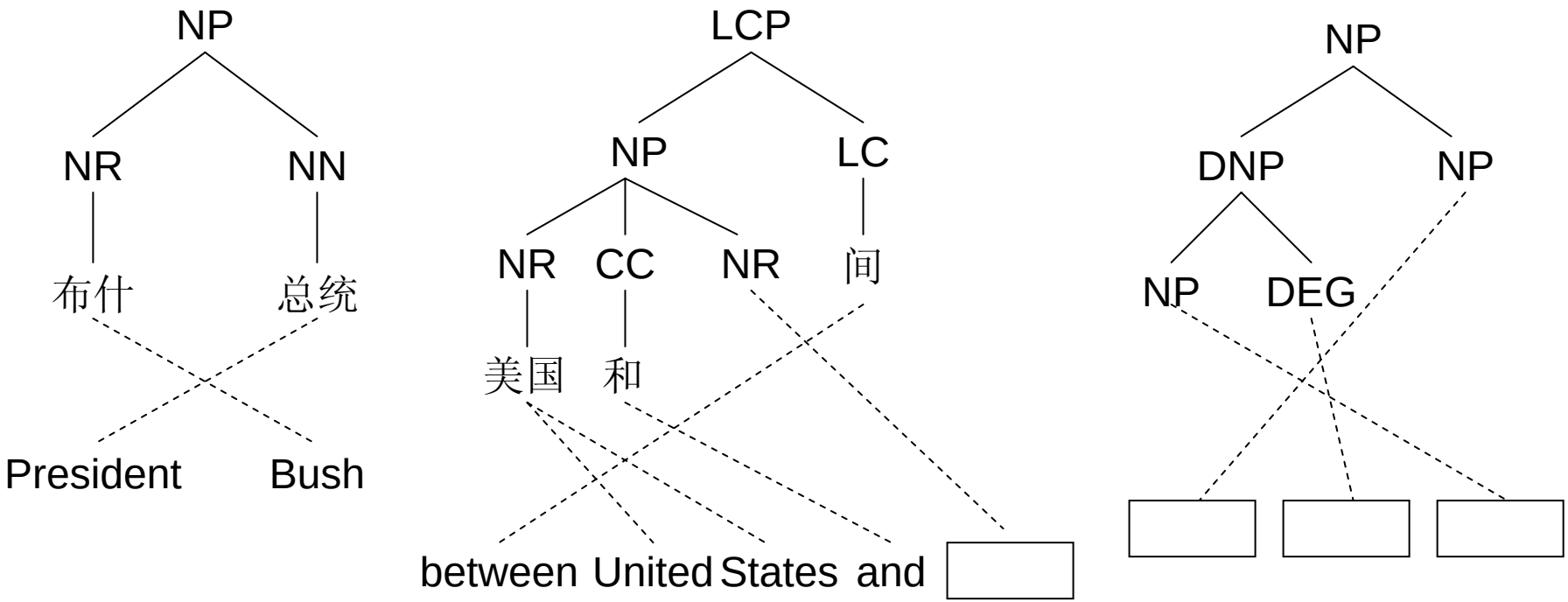
基于树到串对齐模板的翻译模型

- 基于树到串对齐模板的翻译模型（刘洋，ICT）
  A Translation Model Based on Tree-to-String Alignment Template
- Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-String Alignment Template for Statistical Machine Translation. COLING-ACL 2006, Sydney, Australia, July 17-21.
- Yang Liu, Yun Huang, Qun Liu and Shouxun Lin, Forest-to-String Statistical Translation Rules, ACL2007, Prague, Czech，June 2007

基于树到串对齐模板的翻译模型

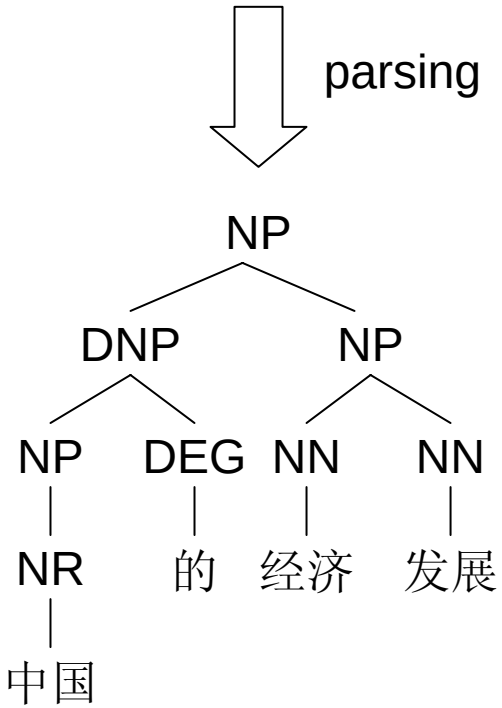
- 基于树到串对齐模板（简称TAT）的统计翻译模型是一种在源语言进行句法分析的基于语言学句法结构的统计翻译模型
- 树到串对齐模板既可以生成终结符也可以生成非终结符，既可以执行局部重排序也可以执行全局重排序
- 从经过词语对齐和源语言句法分析的双语语料库上自底向上自动抽取TAT
- 自底向上的柱搜索算法

树到串对齐模板

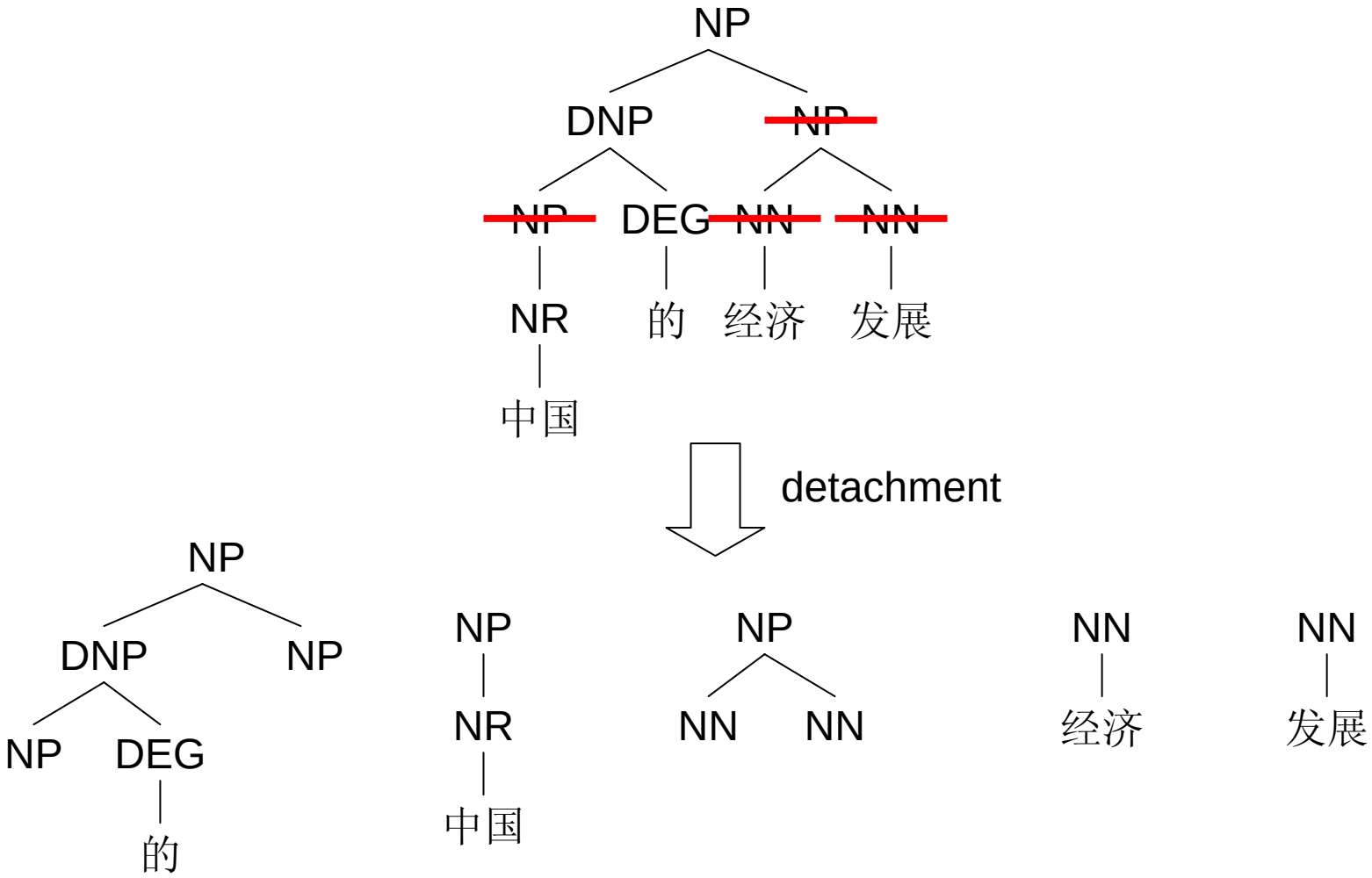

NP
NR
NN
布什
总统
President
Bush
LCP
NP
LC
NR
CC
NR
间
美国
和
between
United
States
and
NP
DNP
NP
NP
DEG

翻译过程 : Parsing

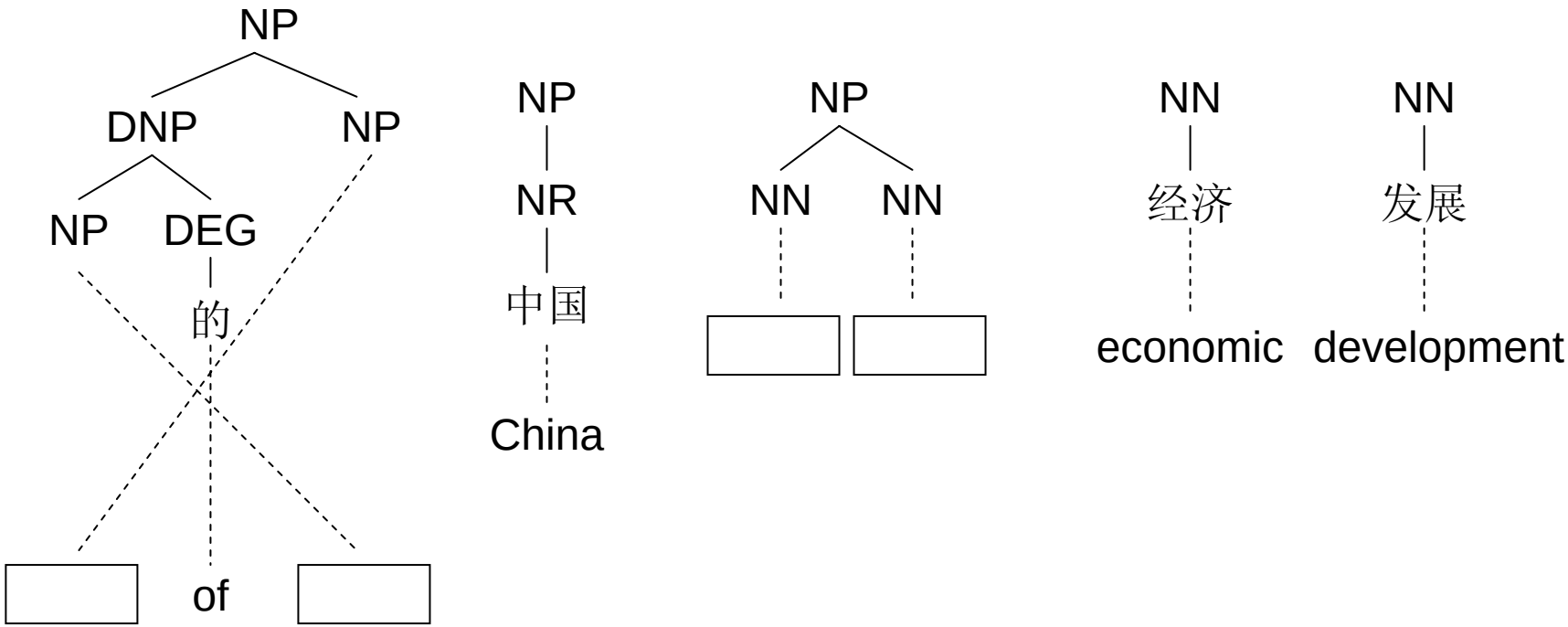
中国 的 经济 发展



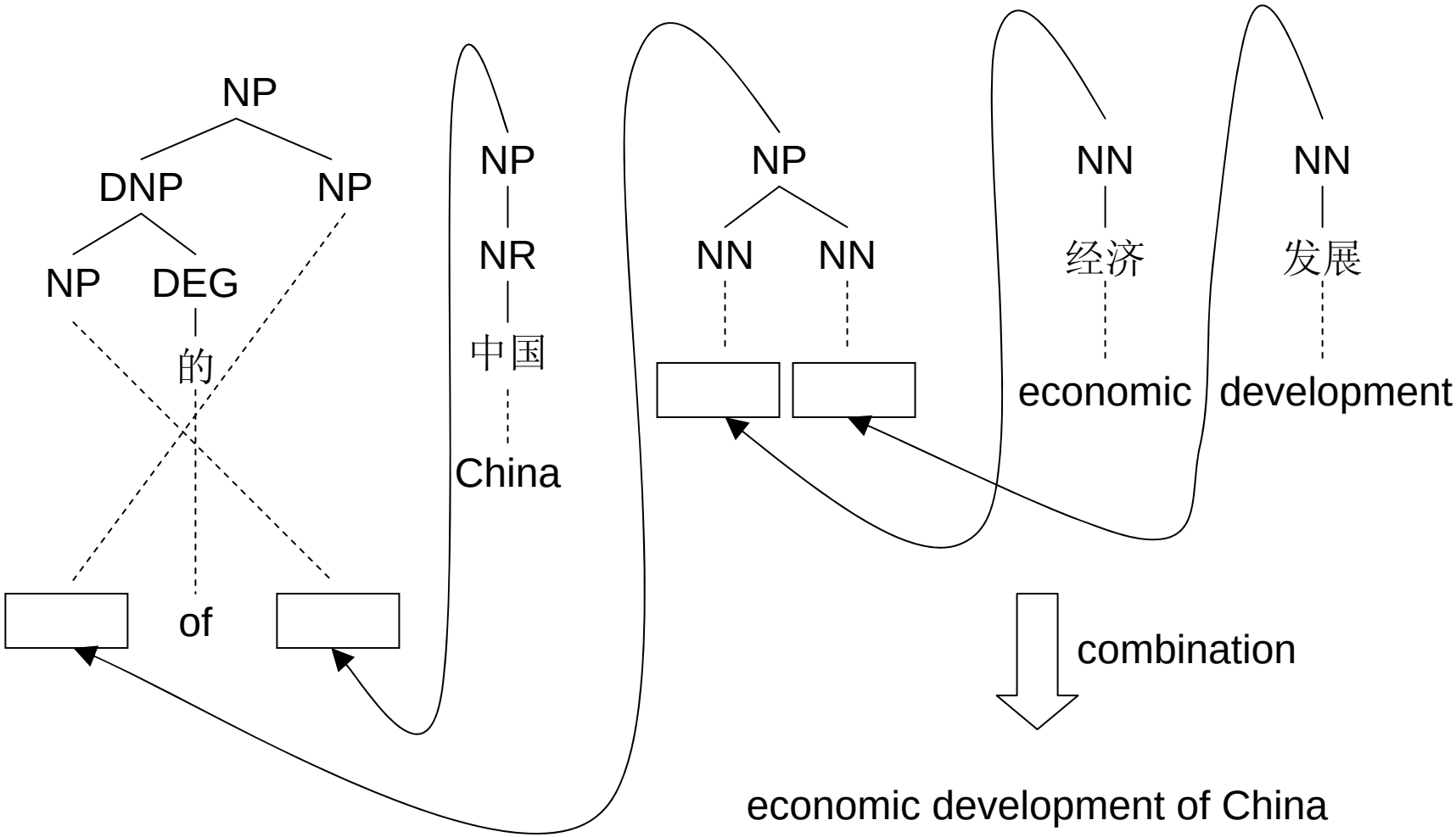
翻译过程: Detachment



翻译过程: Production



翻译过程: Combination



串到树的统计翻译模型 (1)

- USC-ISI的系列工作
- 发表了大量论文，但还没有一个完整的论述
- 性能优异，在NIST2006汉英项目平常中超过了Google（Google使用的语言模型规模比ISI大得多）

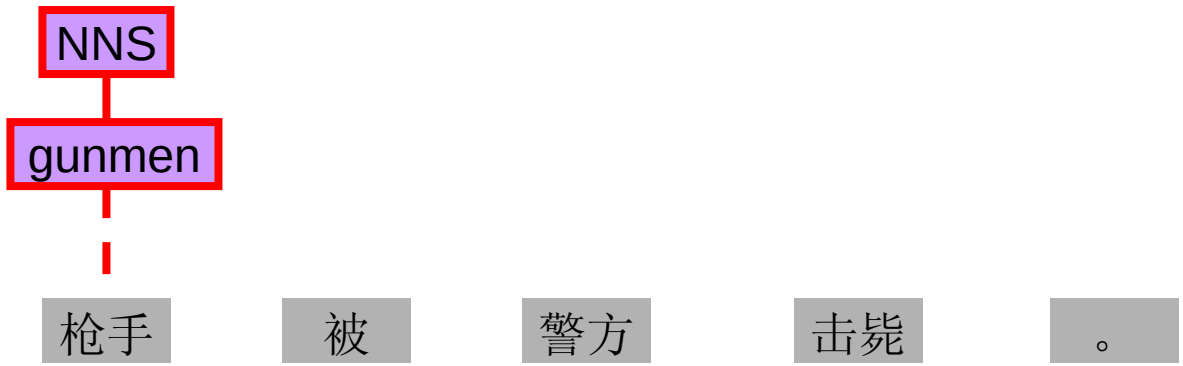
串到树的统计翻译模型 (2)

- 基本思想
  - 在目标语言端进行句法分析
  - 根据目标语言端的句法结构，和词语对齐，建立源语言端的句法结构（伪树）
  - 利用两个句法结构自动抽取带概率的平行上下文无关语法
  - 对平行上下文无关语法进行二叉化
  - 解码时类似规则方法，复杂度等价于句法分析
    - 源文分析
    - 规则映射
    - 译文生成

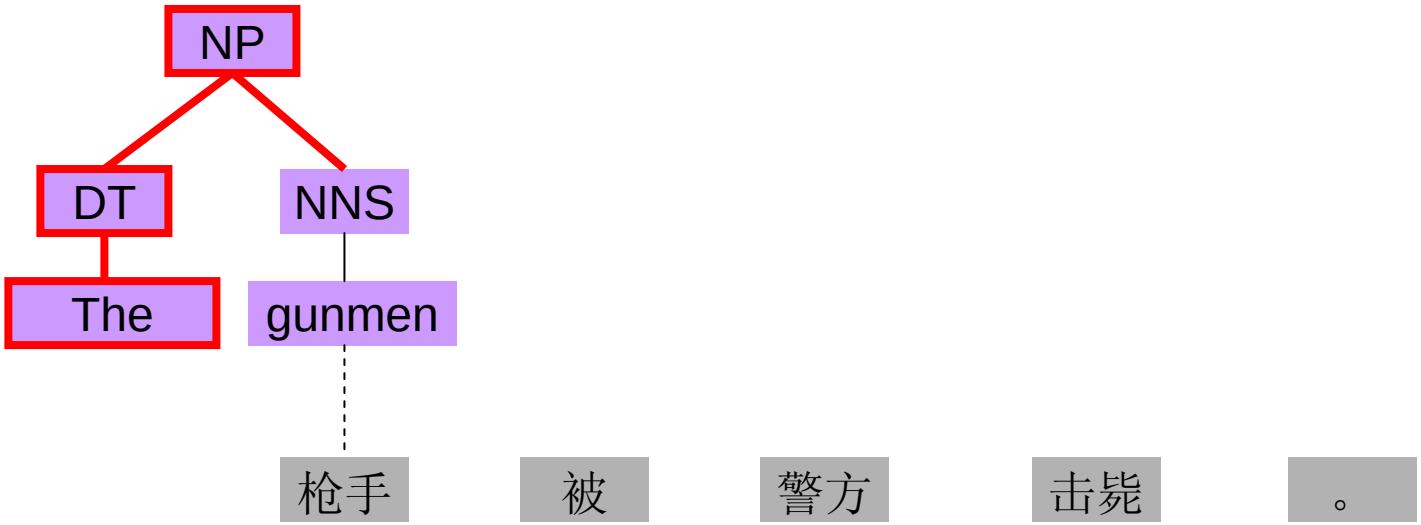
Example

枪手	被	警方	击毙	。

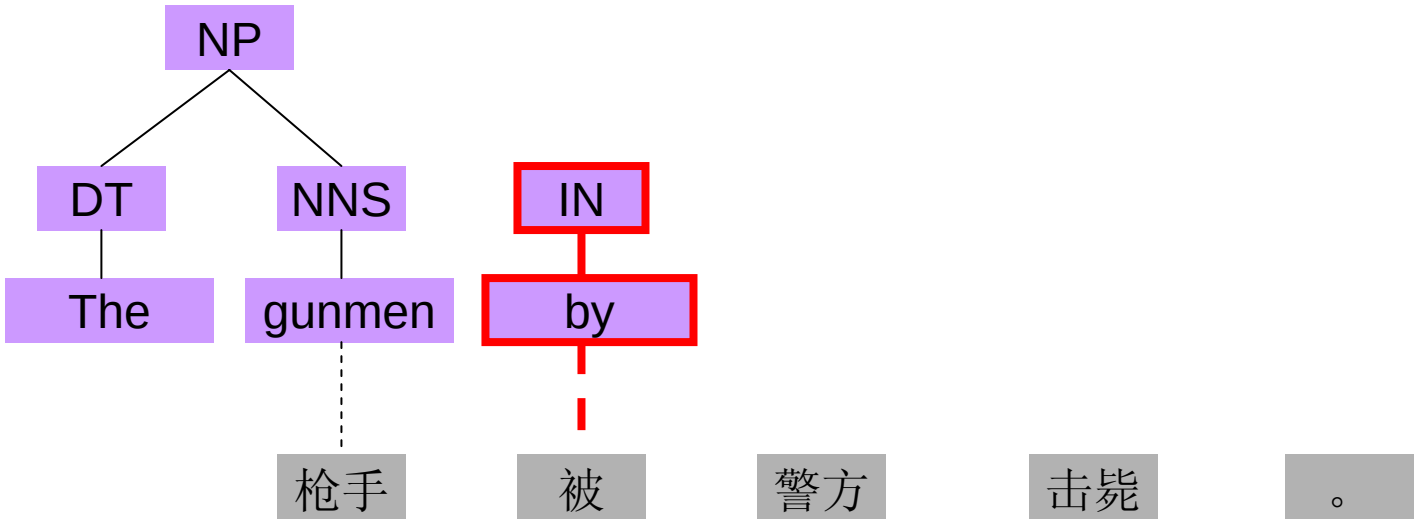
Example



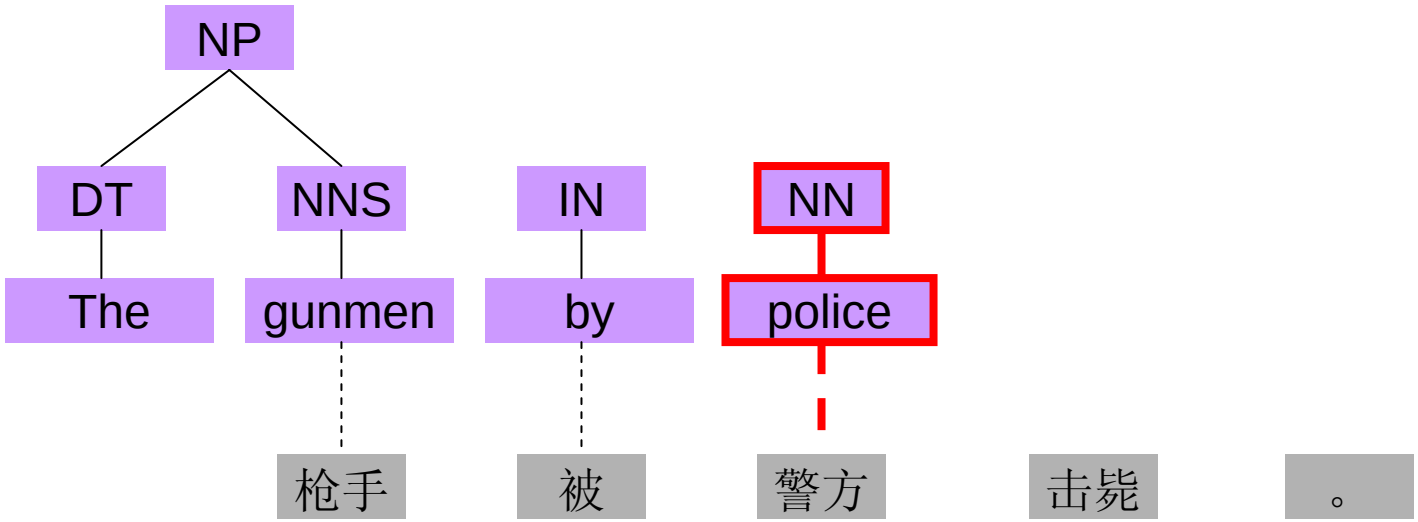
Example



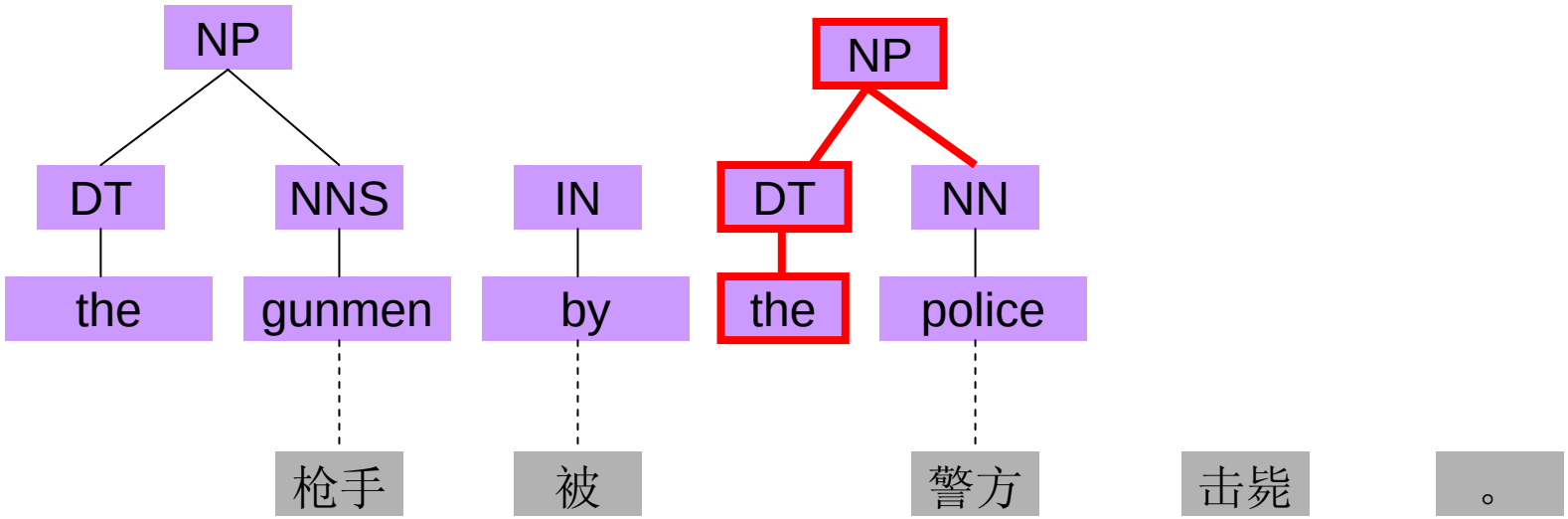
Example



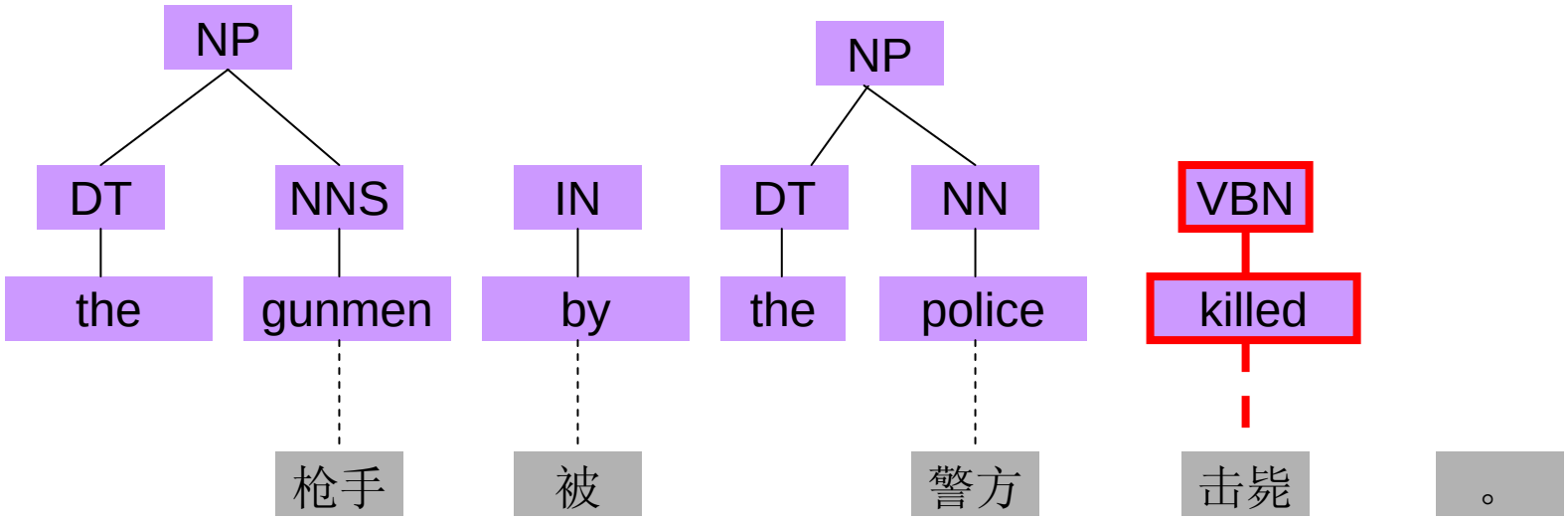
Example



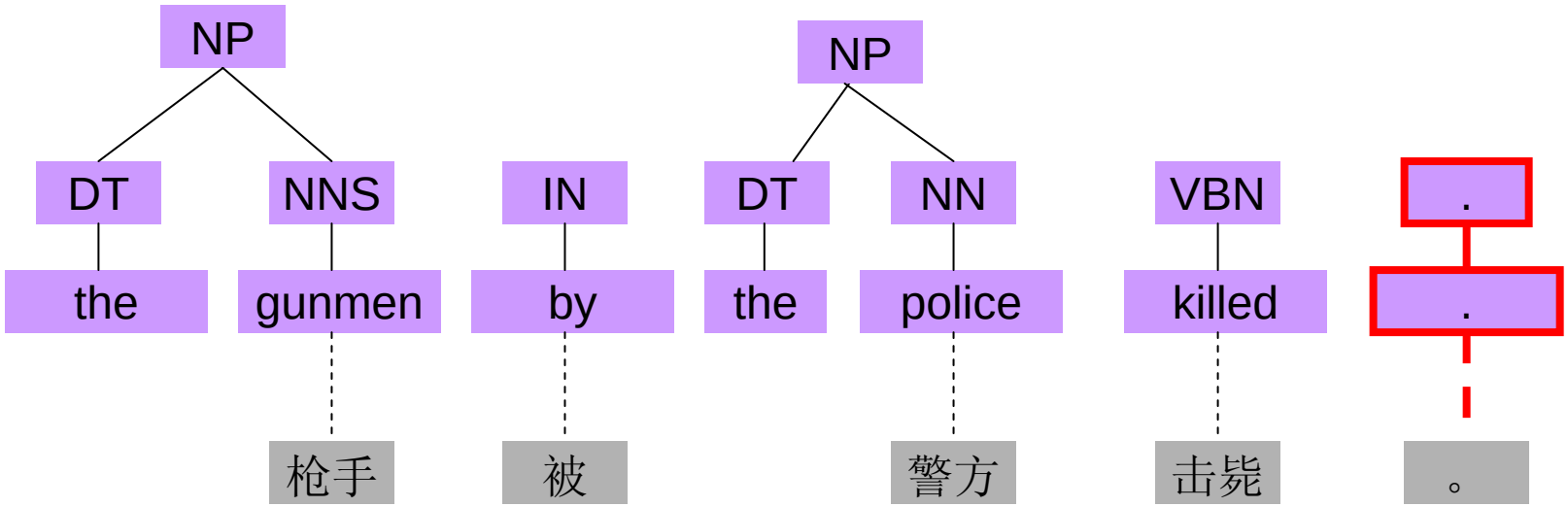
Example



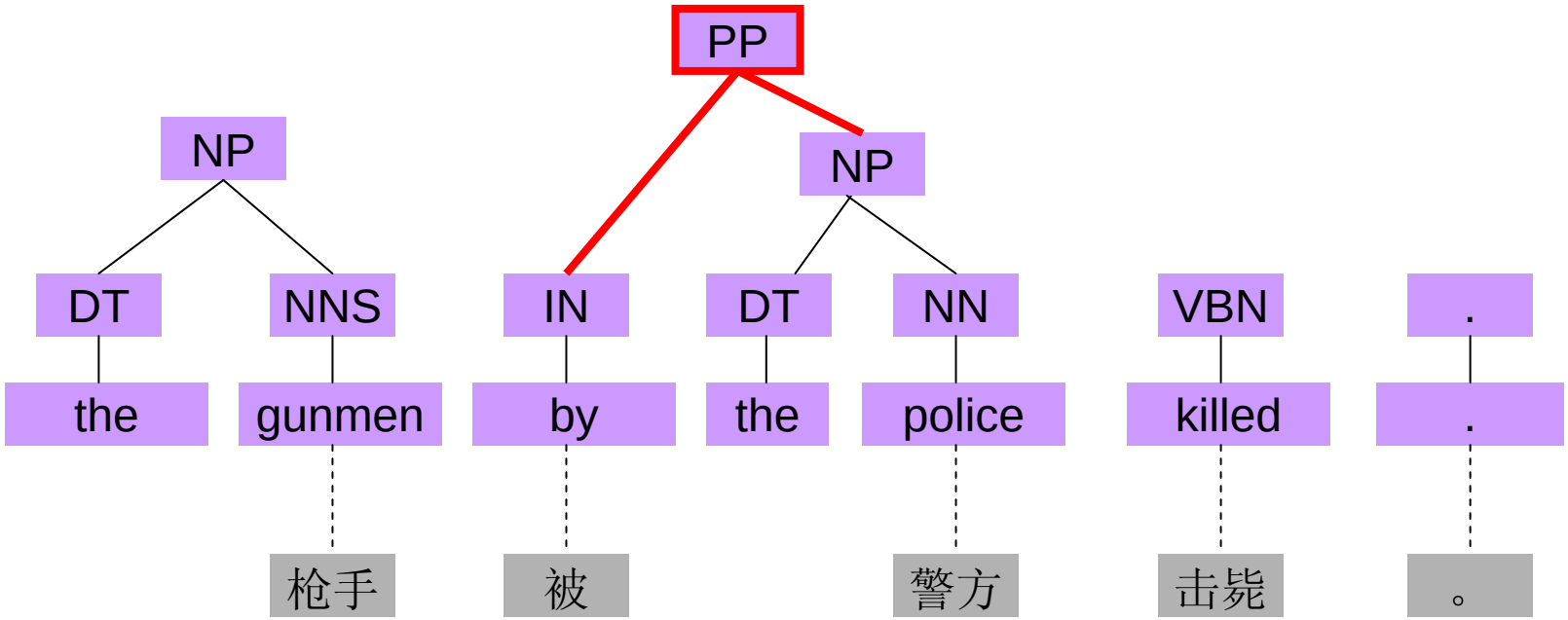
Example


NP
DT
NNS
IN
NP
DT
NN
VBN
the
gunmen
by
the
police
killed
枪手
被
警方
击毙
。

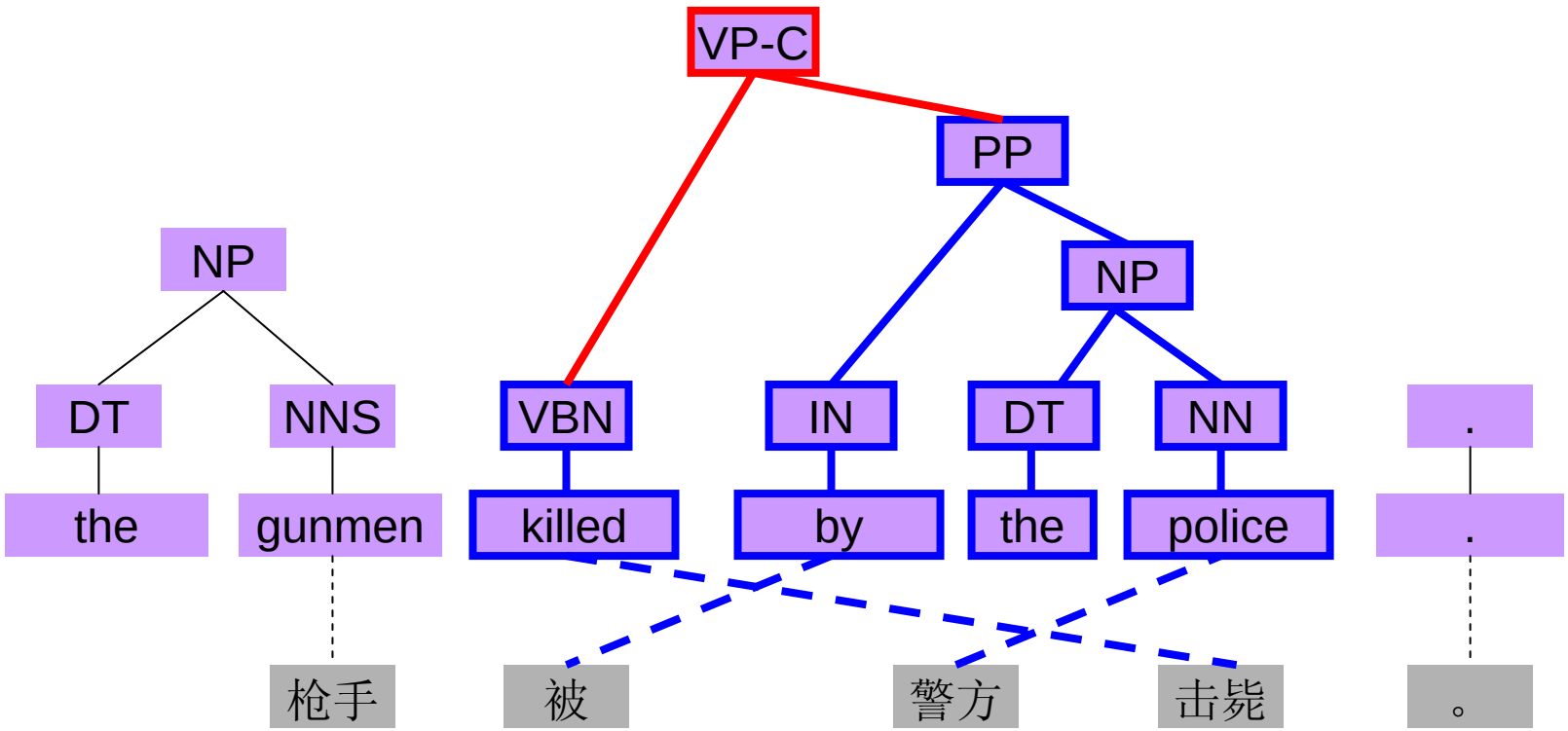
Example


NP
DT
NNS
the
gunmen
枪手
IN
by
被
NP
DT
NN
the
police
警方
VBN
killed
击毙
.
.
。

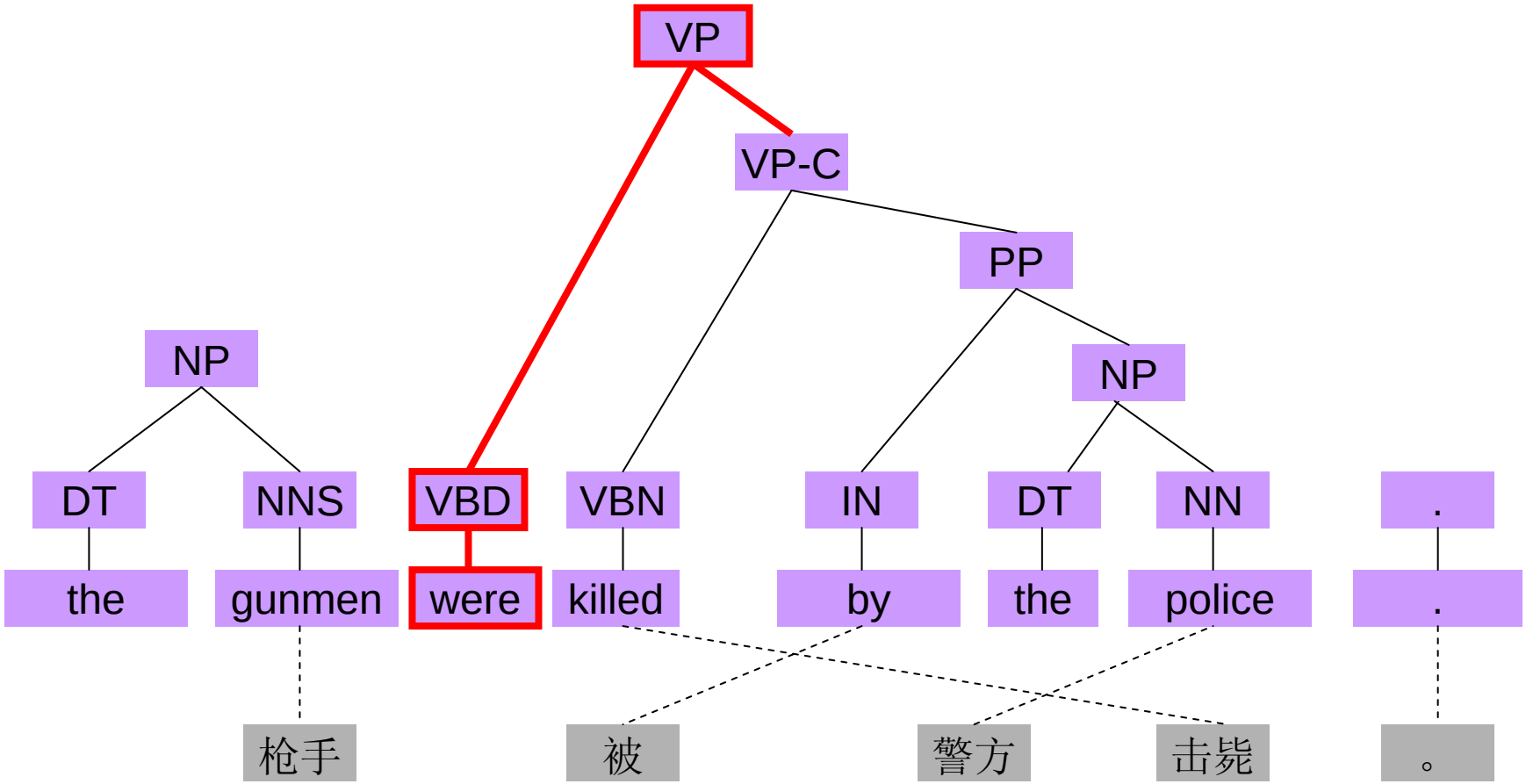
Example


PP
NP
NP
DT
NNS
IN
DT
NN
VBN
.
the
gunmen
by
the
police
killed
.
枪手
被
警方
击毙
。

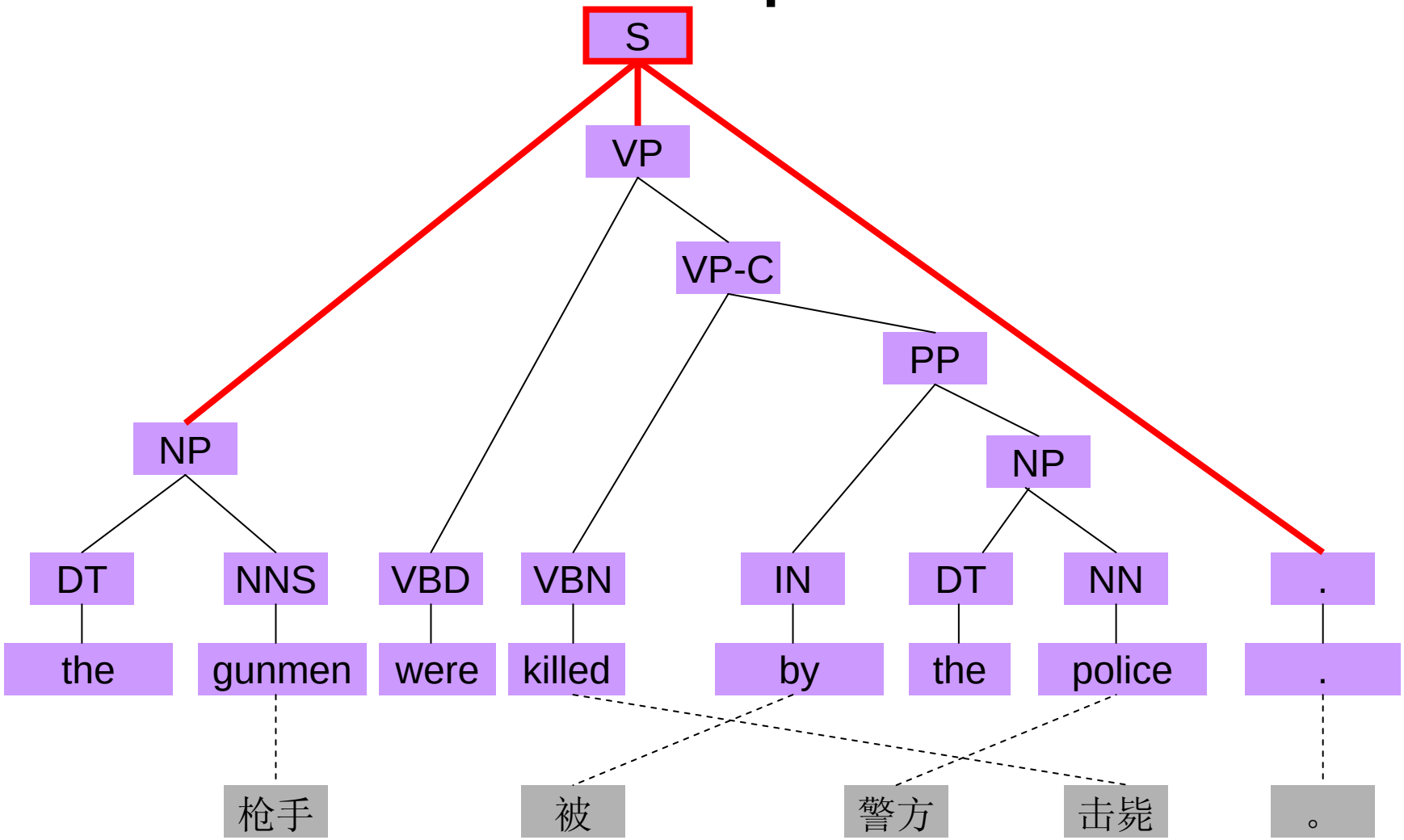
Example


VP-C
PP
NP
NP
DT
NNS
VBN
IN
DT
NN
.
the
gunmen
killed
by
the
police
.
枪手
被
警方
击毙
。

Example


VP
VP-C
PP
NP
NP
DT
NNS
VBD
VBN
IN
DT
NN
.
the
gunmen
were
killed
by
the
police
.
枪手
被
警方
击毙
。

Example


S
VP
VP-C
PP
NP
NP
DT
NNS
VBD
VBN
IN
DT
NN
.
the
gunmen
were
killed
by
the
police
.
枪手
被
警方
击毙
。

统计机器翻译

- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
  —基于词的IBM模型
- 最成熟的统计机器翻译方法
  —基于短语的模型
- 目前统计机器翻译研究的热点
  —基于句法的模型
- 机器翻译的自动评价

机器翻译评价 (1)

- 最早的机器翻译评价：ALPAC报告
- 机器翻译评价的常用指标
  - 忠实度（Adequacy）：译文在多大程度上传递了源文的内容；
  - 流利度（Fluency）：译文是否符合目标语言的语法和表达习惯；
  - 信息度（Informative）：用户可以从译文中获得信息的程度（通过选择题评分）
- 绝对评价和相对评价

机器翻译评价 (2)

- 人工评价
  - 准确
  - 成本极高
  - 不能反复使用
- 自动评价
  - 准确率低
  - 成本低
  - 可以反复使用

机器翻译评价 (3)

- 机器翻译的评价一直是机器翻译研究领域中一个备受关注的问题；
- 机器翻译的自动评价越来越引起重视
  - “评测驱动”成为自然语言处理研究的一个主要动力
  - 大规模语料库的出现、各种机器翻译算法的提出，使得开发过程中频繁的评测成为必需
  - 开发过程中频繁的评测只能通过采用自动评测方法

机器翻译的自动评测

- 完全匹配方法
  - 与参考译文完全相同的译文才被认为是正确的
  - 显然该标准过于严格，不适用
- 编辑距离方法
- 基于测试点的方法
- 基于N元语法的方法

基于编辑距离的机器翻译评测 (1)

- 编辑距离定义：

  从候选译文到参考译文，所需要进行的插入、删除、替换操作的次数

- 举例说明：
  - 源文：She is a star with the theatre company.
  - 机器译文：她 是 与 剧院 公司 的 一 颗 星 。
  - 参考译文：她 是 剧团 的 明星 。
  - 编辑距离：6
    - 删除：与 公司 一 颗
    - 替换：剧院→剧团 星→明星

基于编辑距离的机器翻译评测 (2)

- 单词错误率：编辑距离除以参考译文中单词数
  - 这个指标是从语音识别中借鉴过来的。
  - 由于语音识别的结果语序是不可变的，而机器翻译的结果语序是可变的，显然这个指标存在一定的缺陷。
- 与位置无关的单词错误率：计算编辑距离时，不考虑插入、删除、替换操作的顺序
  - 也就是说，候选译文与参考译文相比，多出或不够的词进行删除或插入操作，其余不同的词进行替换操作。
  - 这个指标与单词错误率相比，允许语序的变化，不过又过于灵活。

基于测试点的机器翻译评测 (1)

- 俞士汶等，机器翻译译文质量自动评估系统，中国中文信息学会1991年论文集，pp. 314～319
- 基本思想
  - 对于每一个句子，孤立测试点，简化测试目标（模拟人类标准化考试的办法）
  - 对于每一个句子，采用一种TDL语言描述的BNF去与译文匹配，匹配成功则正确，否则错误
  - 大批量出题，全面评价机器翻译译文质量

基于测试点的机器翻译评测 (2)

- 测试点分组：

  单词、词组、词法、语法（初、中、高级）

- 测试点示例：
  - 源文：I am a student.
  - 测试：译文中出现“学生/大学生”为正确
  - 源文：I bought a table with three dollars.
  - 测试：“买”出现在“美元”之后为正确
  - 源文：I bought a table with three legs.
  - 测试：“买”出现在“腿”之前为正确

基于测试点的机器翻译评测 (3)

- 优点：
  - 全自动
  - 实验证明，评价结果是可信的
  - 可以按照人类专家的要求进行单项评测
- 缺点
  - 题库的构造需要具有专门知识的专家，并且成本较高

基于N元语法的机器翻译评测 (1)

- 基本思想
  - 用译文中出现的N元组和参考译文中出现的N元组相比，计算匹配的N元组个数与候选译文的N元组总个数的比例
  - 允许一个源文有多个参考译文，综合评分
- 参考文献
  - Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation, IBM Research, RC22176 (W0109-022) September 17, 2001

基于N元语法的机器翻译评测 (2)

源文：党指挥枪是我党的行动指南。

候选译文：

- It is a guide to action which ensures that the military always obeys the command of the party
- It is to insure the troops forever hearing the activity guidebook that party direct

参考译文：

- It is a guide to action that ensures that the military will forever heed party commands
- It is the guiding principle which guarantees the military forces always being under the command of the party
- It is the practical guide for the army to heed the directions of the party

基于N元语法的机器翻译评测 (3)

- 两个改进：
  - 对于候选译文中某个n元接续组出现的次数，如果比参考译文中出现的最大次数还多，要把多出的次数“剪掉”（不作为正确的匹配）。
  - 为了避免“召回率”过低的问题，BLEU的评价标准又对比参考译文更短的句子设计了“惩罚因子”。

基于N元语法的机器翻译评测 (4)

- BLEU的总体评价公式如下：

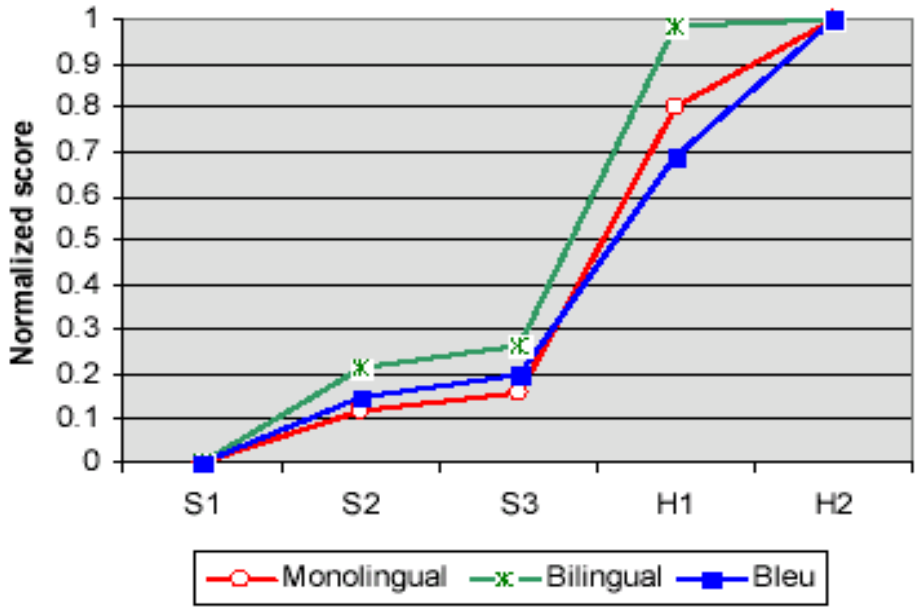
$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

其中，p_n是出现在参考译文中的n元词语接续组占候选译文中n元词语接续组总数的比例，$w_n = 1/N$，N为最大的n元语法阶数（实际取4）。

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}.$$

其中c为候选译文中单词的个数，r为参考译文中与c最接近的译文单词个数。

基于N元语法的机器翻译评测 (5)



其中S1、S2、S3分别是三个不同的机器翻译系统提供的译文，H1和H2是两个人类翻译者提供的译文。蓝线是BLEU系统评测的结果，红线是只懂目标语言的人类专家提供的评测结果，绿线是同时懂源语言和目标语言的人类专家提供的评测结果。

基于N元语法的机器翻译评测 (6)

- 这种方法比较好地模拟了人对机器翻译结果的评价
  - 对于低质量译文比高质量译文的评价更准确；
  - 评价结果与只懂目标语言的人的评价结果更接近（相对于懂双语的人而言）
- 优点
  - 全自动
  - 可以提供多种参考译文综合考虑，结果更全面
  - 容易构造测试集，不需要专门知识

复习思考题

- 访问一些知名的网上翻译网站，直观了解机器翻译
  - SYSTRAN Homepage
  - WordLingo
  - 看世界
- 尝试写一些规则，将英语句子“He wrote a book on history.”翻译成汉语句子“他写了一本关于历史的书。”
- 写一个程序实现英语数字、汉语数字和阿拉伯数字之间的互译
- 写一个程序实现英语和汉语之间时间表达式的互译
- 实现一个基于实例的机器翻译中的实例匹配模块，也就是说，将一个输入的句子分解为实例库中的句子片段的组合，并使得这种组合尽可能简单