中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

机器翻译原理与方法

第四讲 基于规则和基于实例的机器翻译方法

刘群
中国科学院计算技术研究所
liuqun@ict.ac.cn

中国科学院计算技术研究所2011年秋季课程

内容提要

基于规则的方法

基于实例的方法

例子

- 考虑设计一个中译英系统和一个英译中系统，能够实现以下两个句子的互译：
  - 小王在联想工作了三年。
  - Xiao Wang has worked in Lenovo for three years.

词典

汉语词语	汉语词性	英语词语	英语词性
小	a	small	a
王	nr	Wang	nr
王	n	king	n
在	p	in	p
在	p	on	p
联想	v	associate	v
联想	n	association	n
联想	nz	Lenovo	nz
工作	v	work	v
工作	n	job	n
了	u		

词典

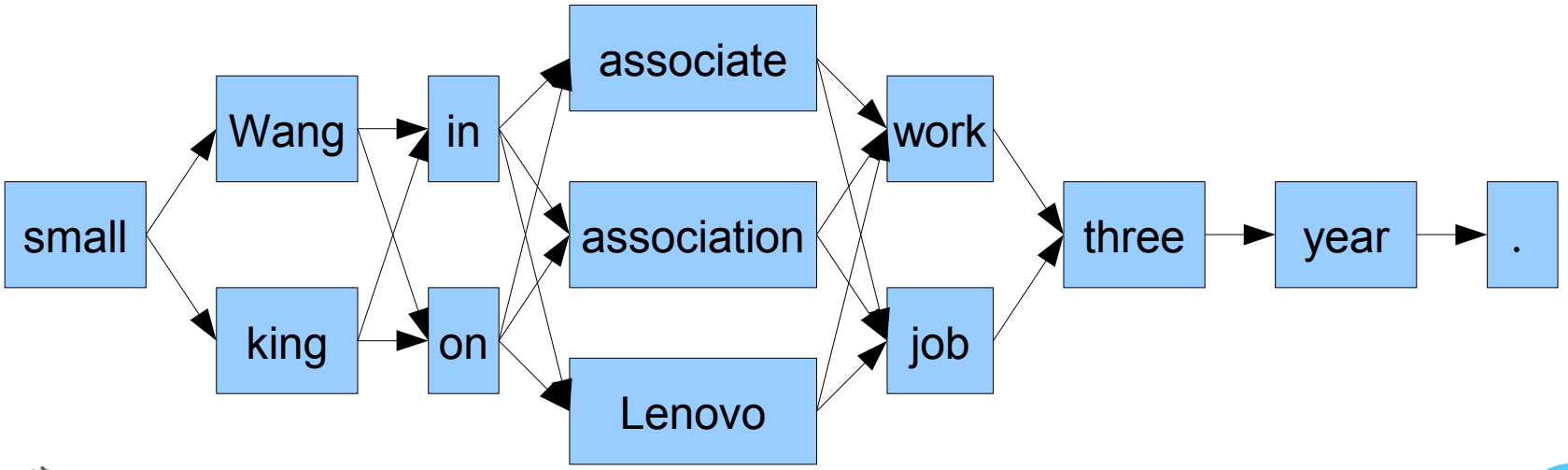
汉语词语	汉语词性	英语词语	英语词性
三	m	three	m
年	qt	year	n
有	v	have	v
		have	aux
为	p	for	p
。	w	.	w

词语切分

小王在联想工作了三年。

词语切分

小　王　在　联想　工作　了　三　年　。

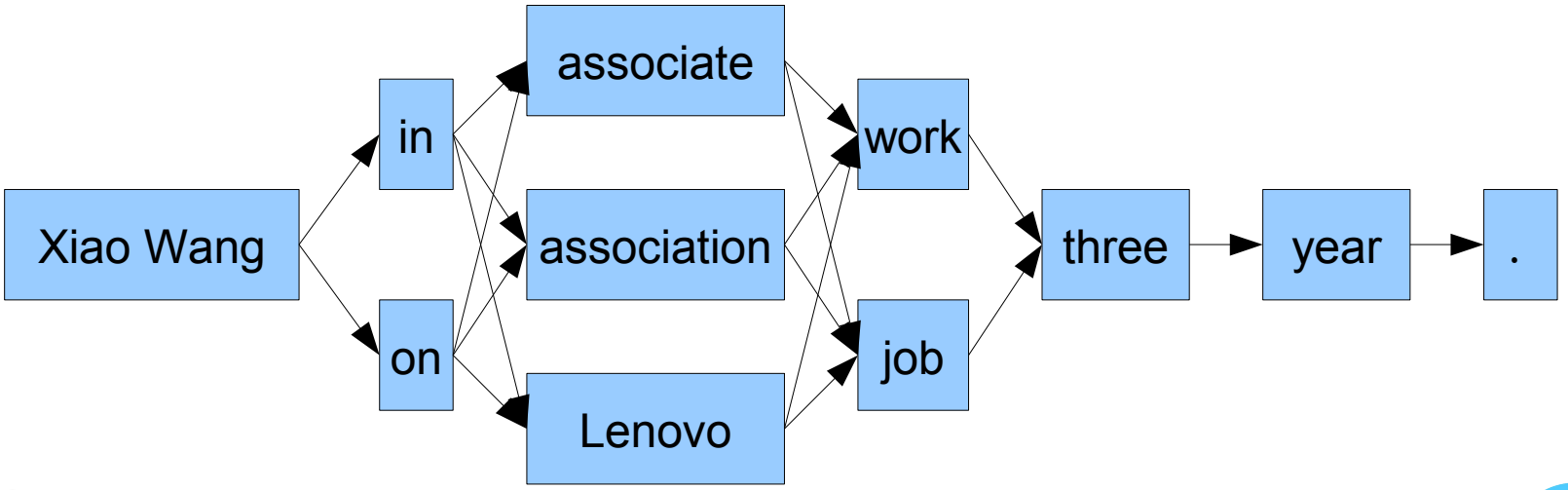


引入人名识别与翻译

小　王　在　联想　工作　了　三　年　。

人名识别

小王/ nr　在　联想　工作　了　三　年　。

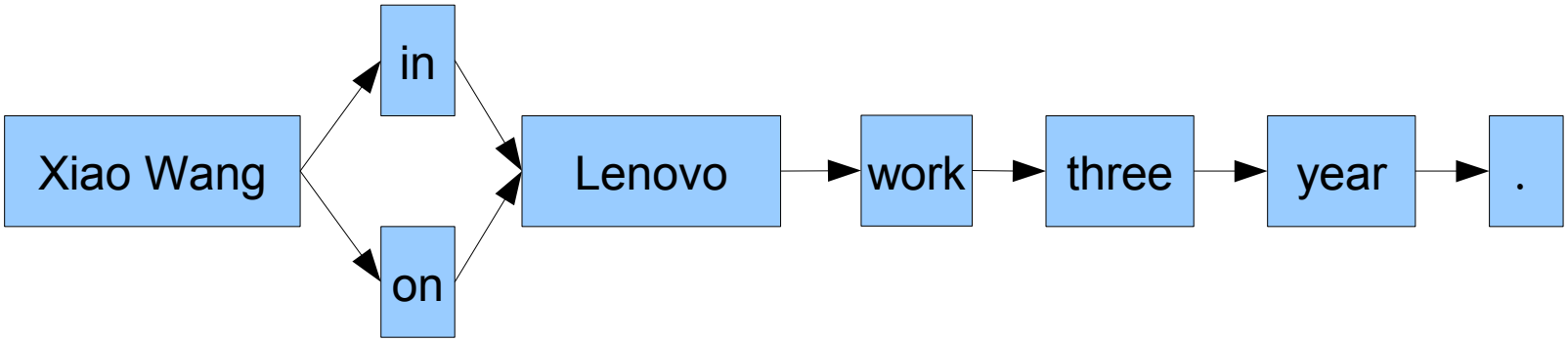


引入词性标注

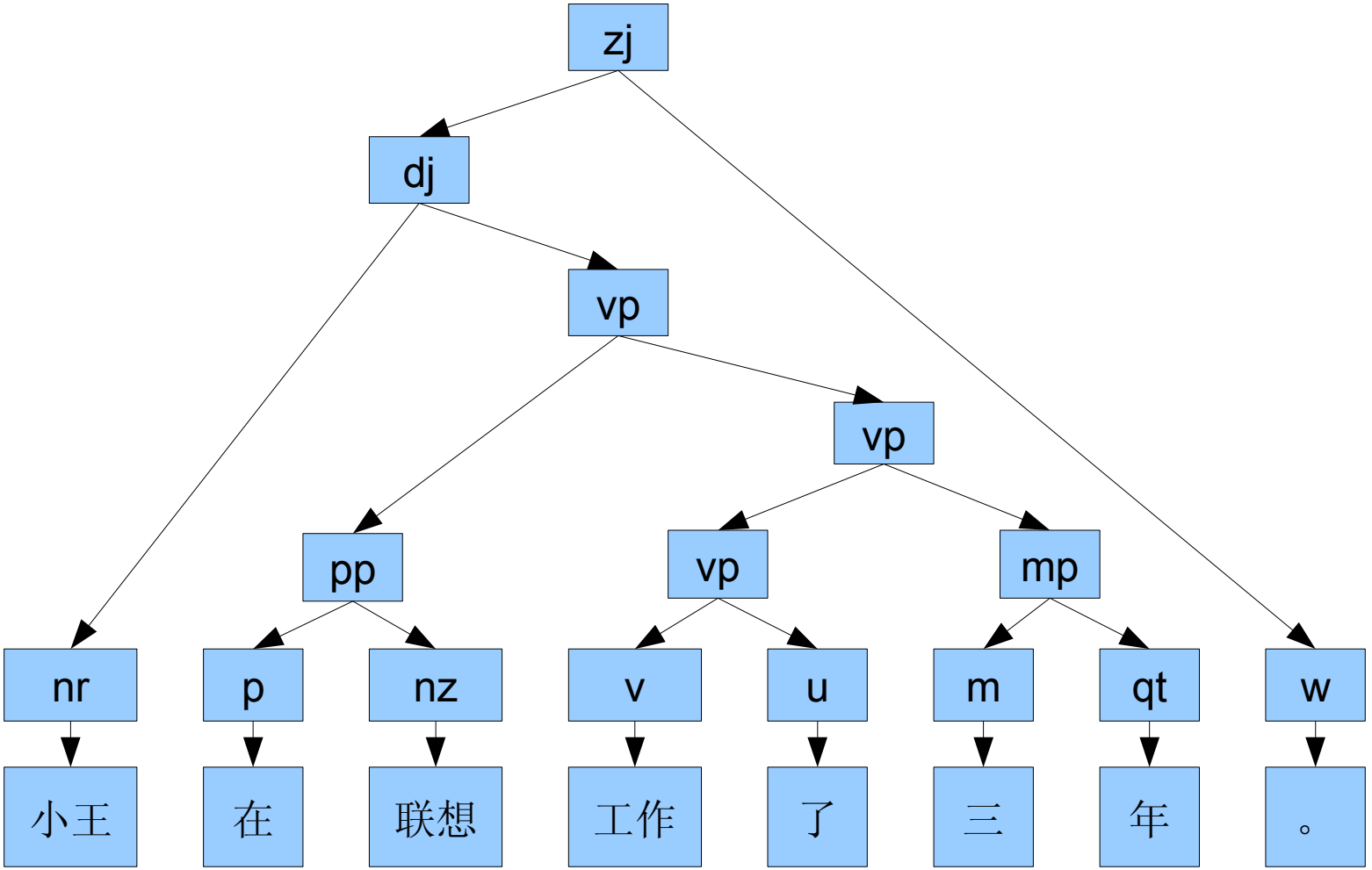
小王/ nr　在　联想　工作　了　三　年　。

词性标注

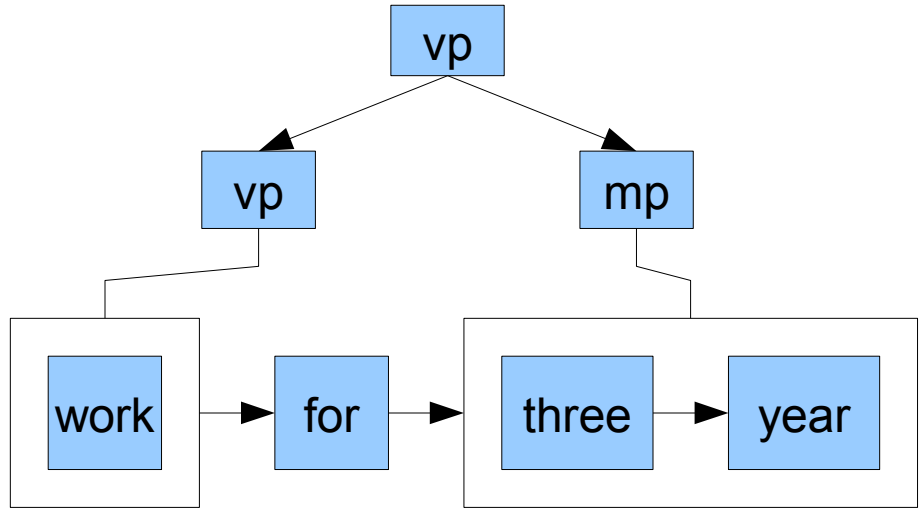
小王/ nr　在/ p　联想/ nz　工作/ v　了/ u　三/ m　年/ qt　。/w



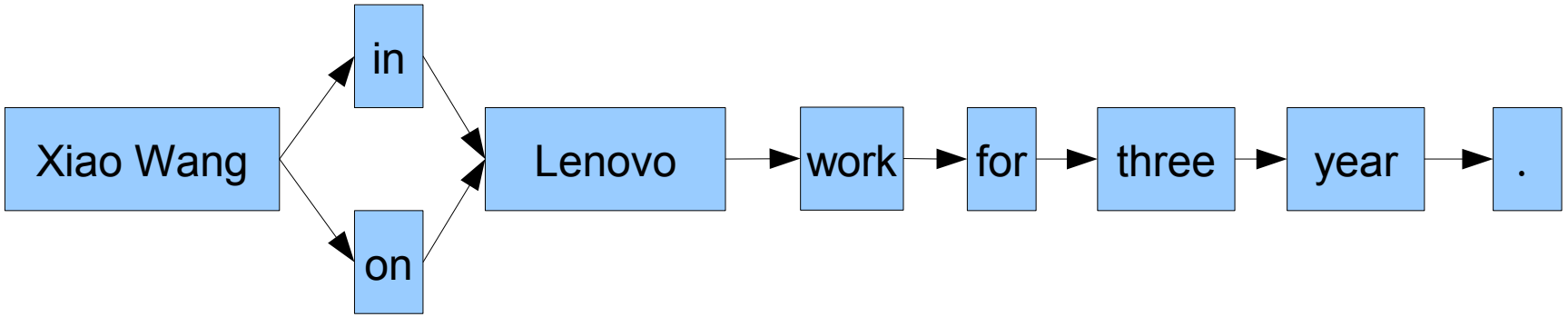
引入句法分析



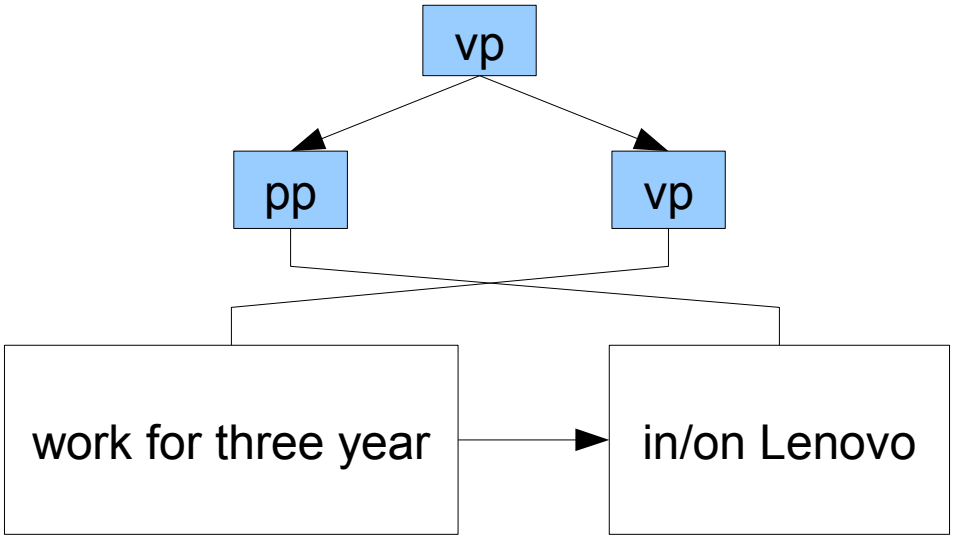
引入翻译规则：插入译文词



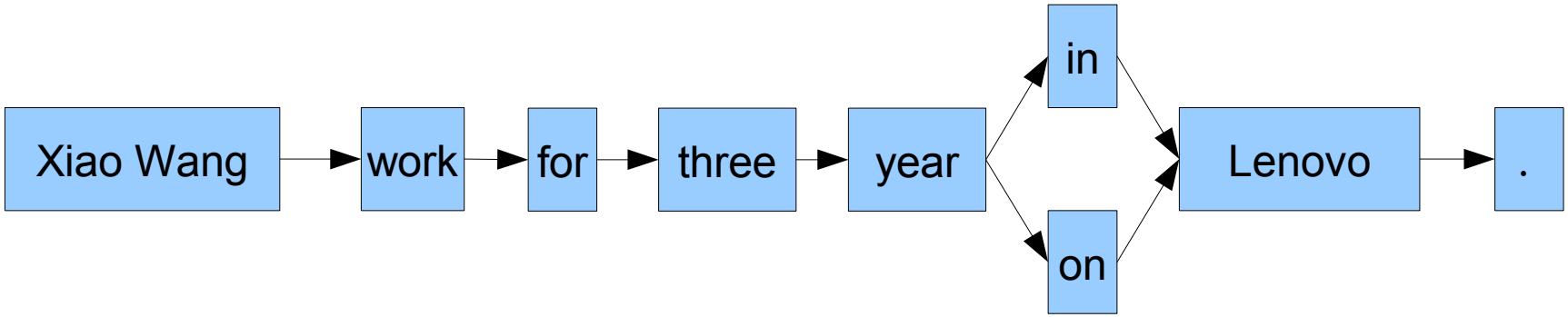
vp (vp#1 mp#2) ==> #1 for #2



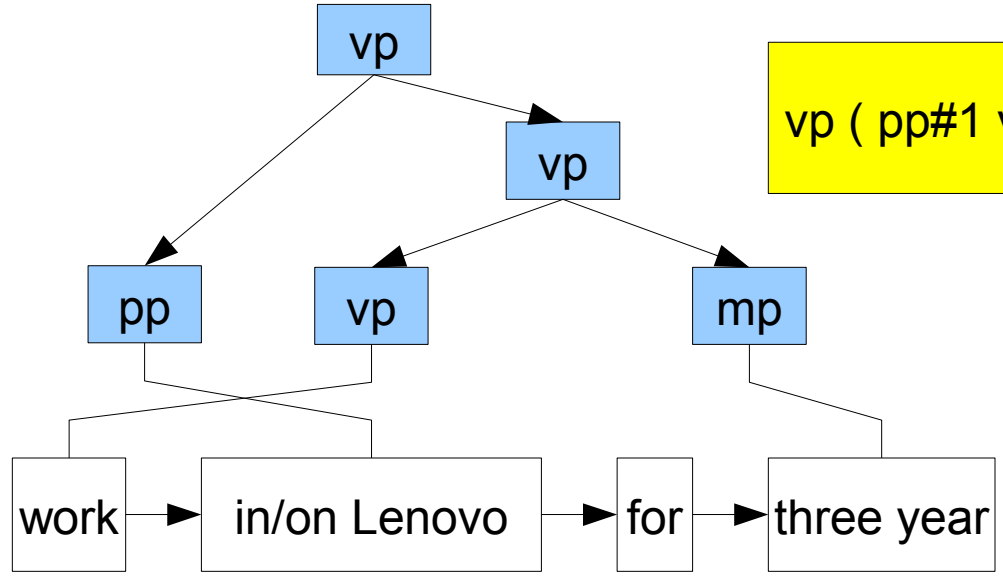
引入翻译规则：调整语序



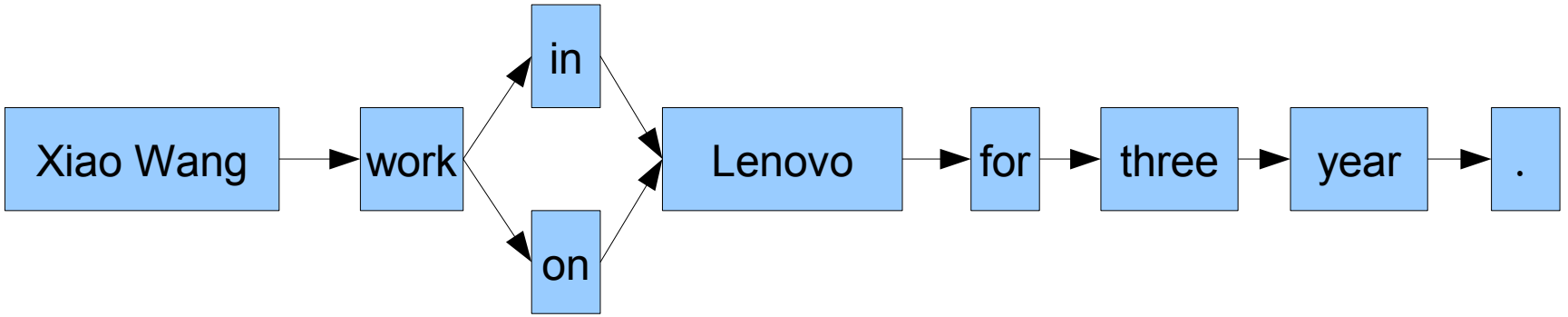
vp (pp#1 vp#2) ==> #2 #1



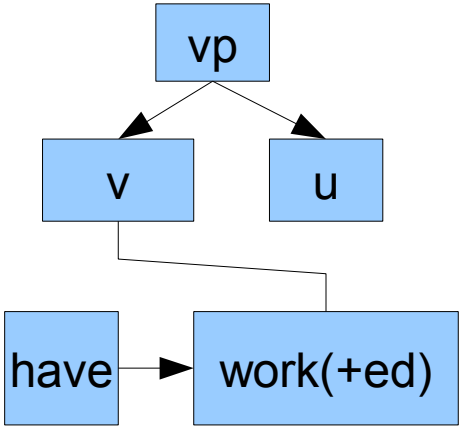
引入翻译规则：多层次调序与插入



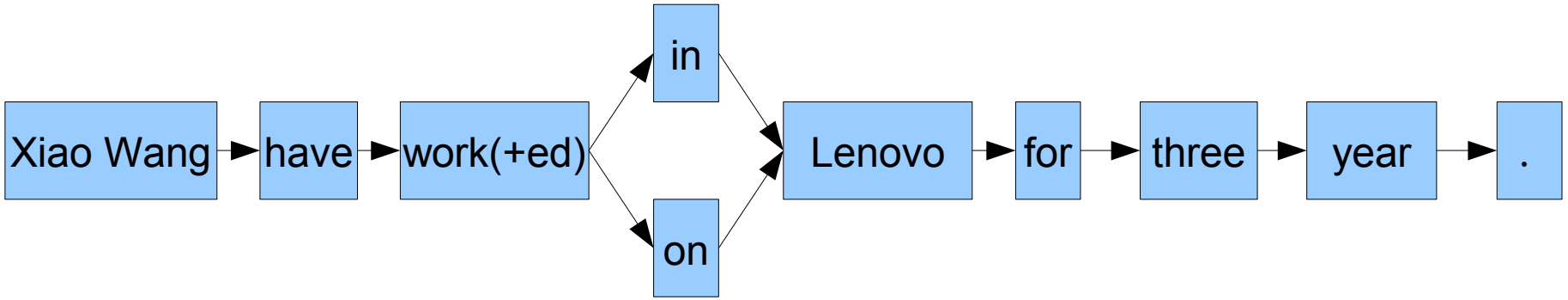
vp (pp#1 vp(vp#2 mp#3) ==> #2 #1 for #3



引入目标语言属性



vp(v#1 u#2) ==> have #1(+ed)



一个纯基于规则的机器翻译系统

- 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—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,…

基于规则的方法—结构生成

```
## { NPMP1 } NP(T !NP(T !N))
        => NP(T/T !NP/NP(!N/N))
        /* this a kind => this kind */
## { NPATN1 } NP(AP(!A) !NP(T !N))
        => P(T/T !NP/NP(AP/AP(!A/A) !N/N))
        /* red this book => this red book */
```

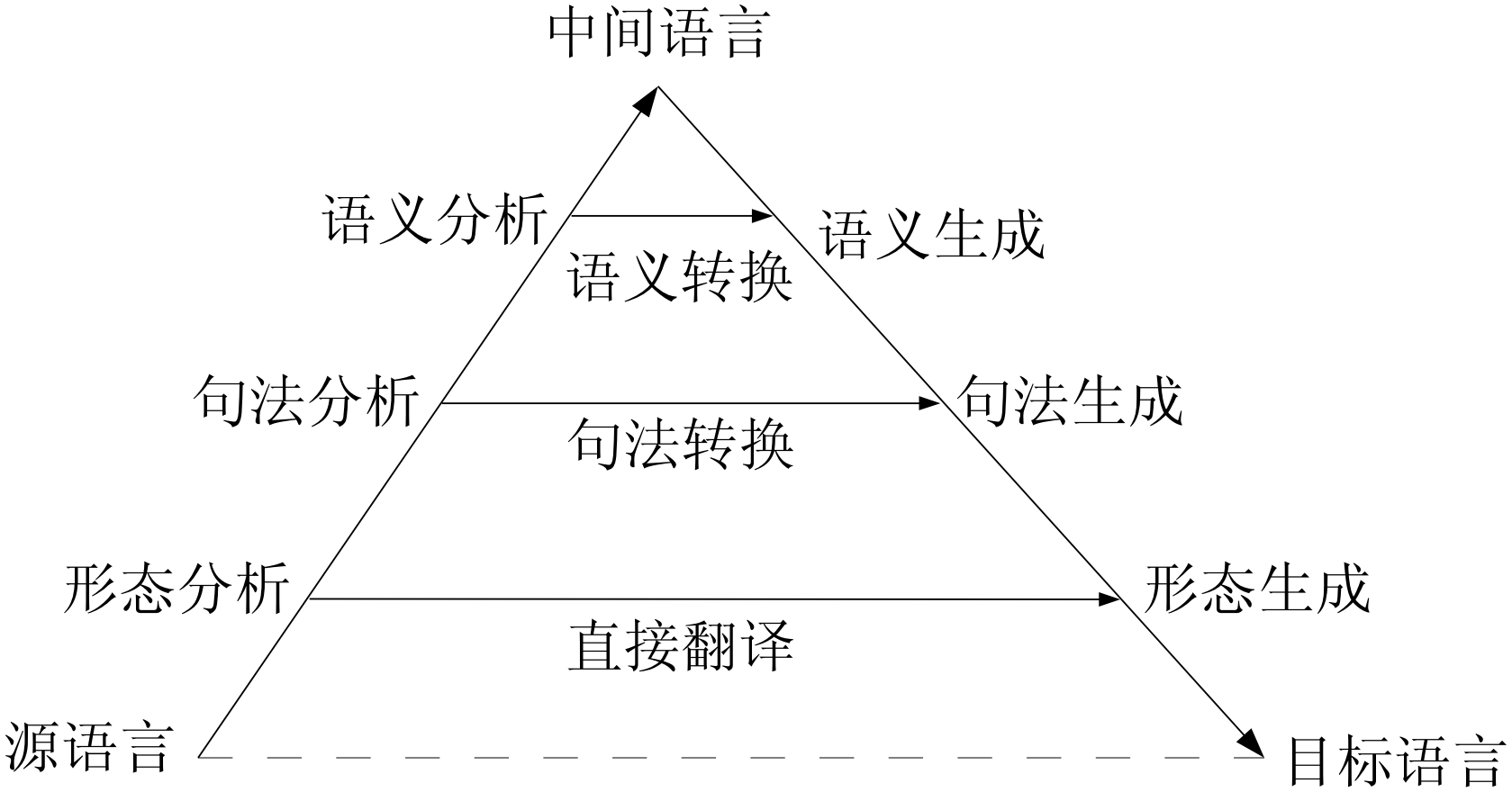
基于规则的方法：优点

- 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
- 大颗粒度的规则具有较强的系统适应性，不依赖于具体的训练语料

基于规则的方法：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）
- 规则库的调试极其枯燥乏味
- 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

机器翻译的转换层面



机器翻译的转换层面

直接翻译方法

句法转换方法

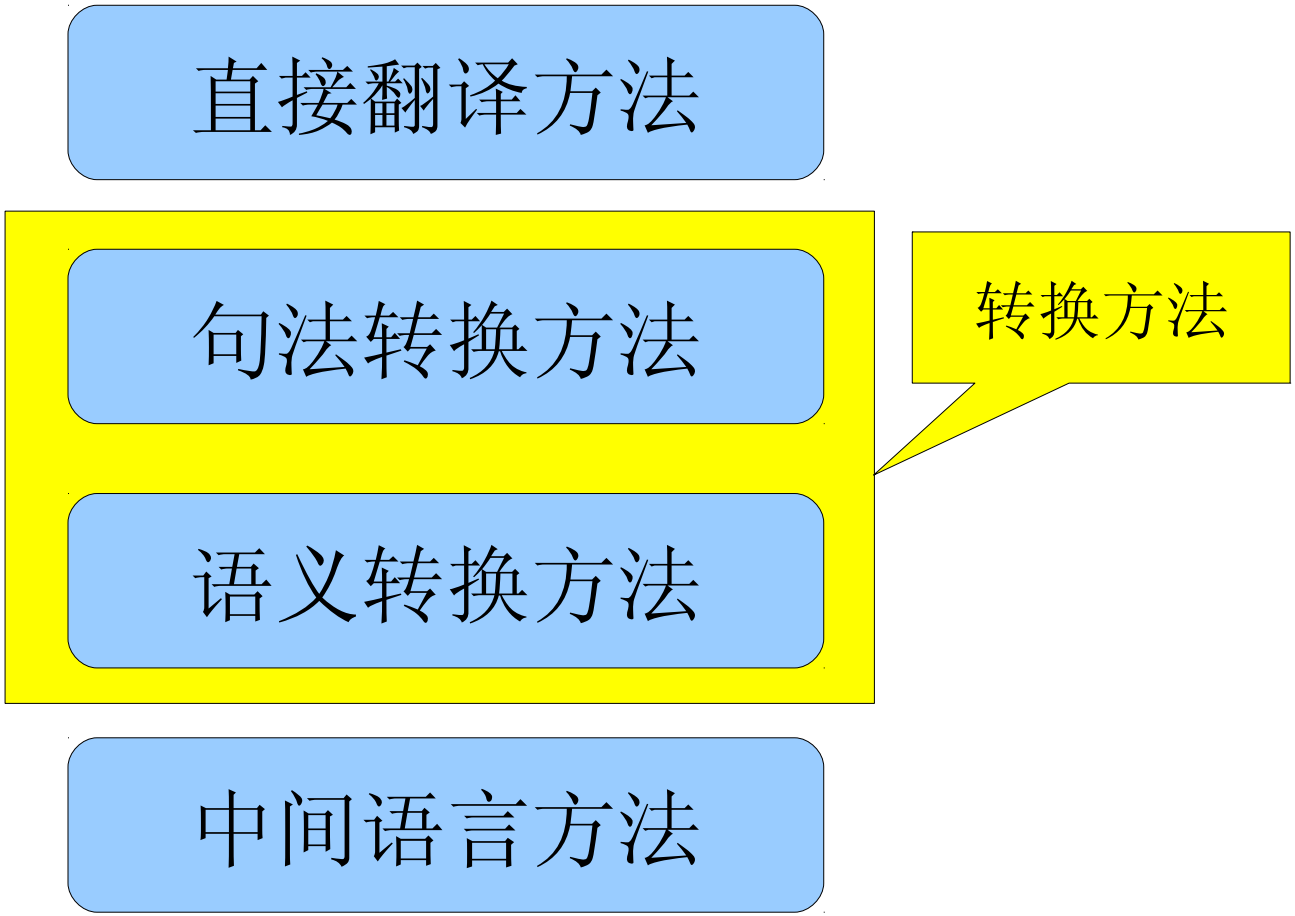
语义转换方法

中间语言方法

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，后来 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

机器翻译的转换层面



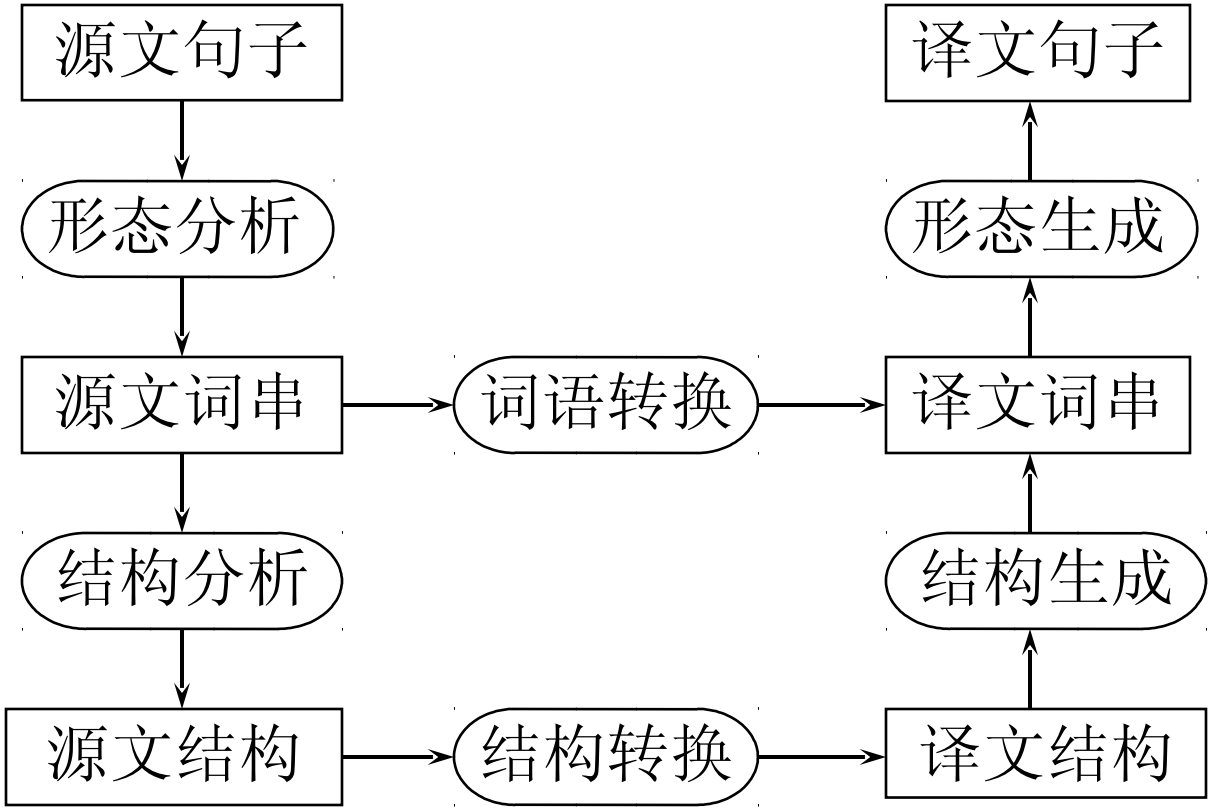
转换方法（1）

- 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

机器翻译的转换层面

直接翻译方法

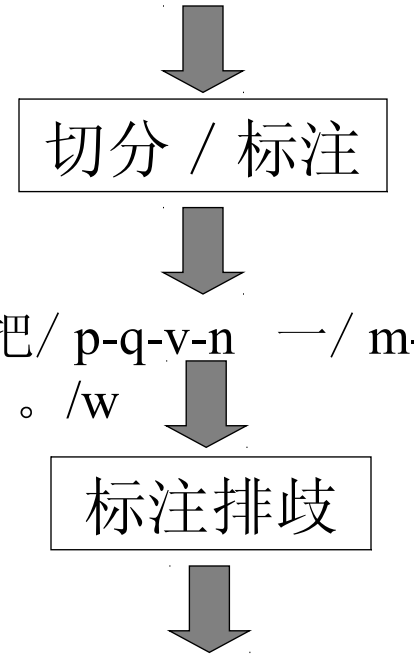
句法转换方法

语义转换方法

中间语言方法

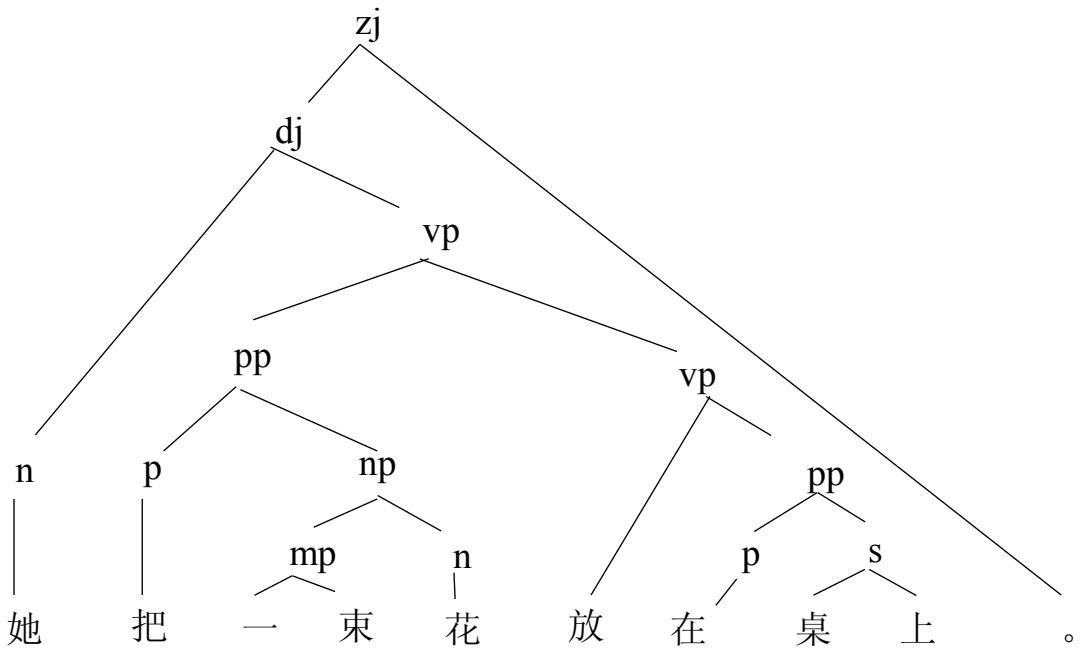
句法层面的转换方法（1）

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

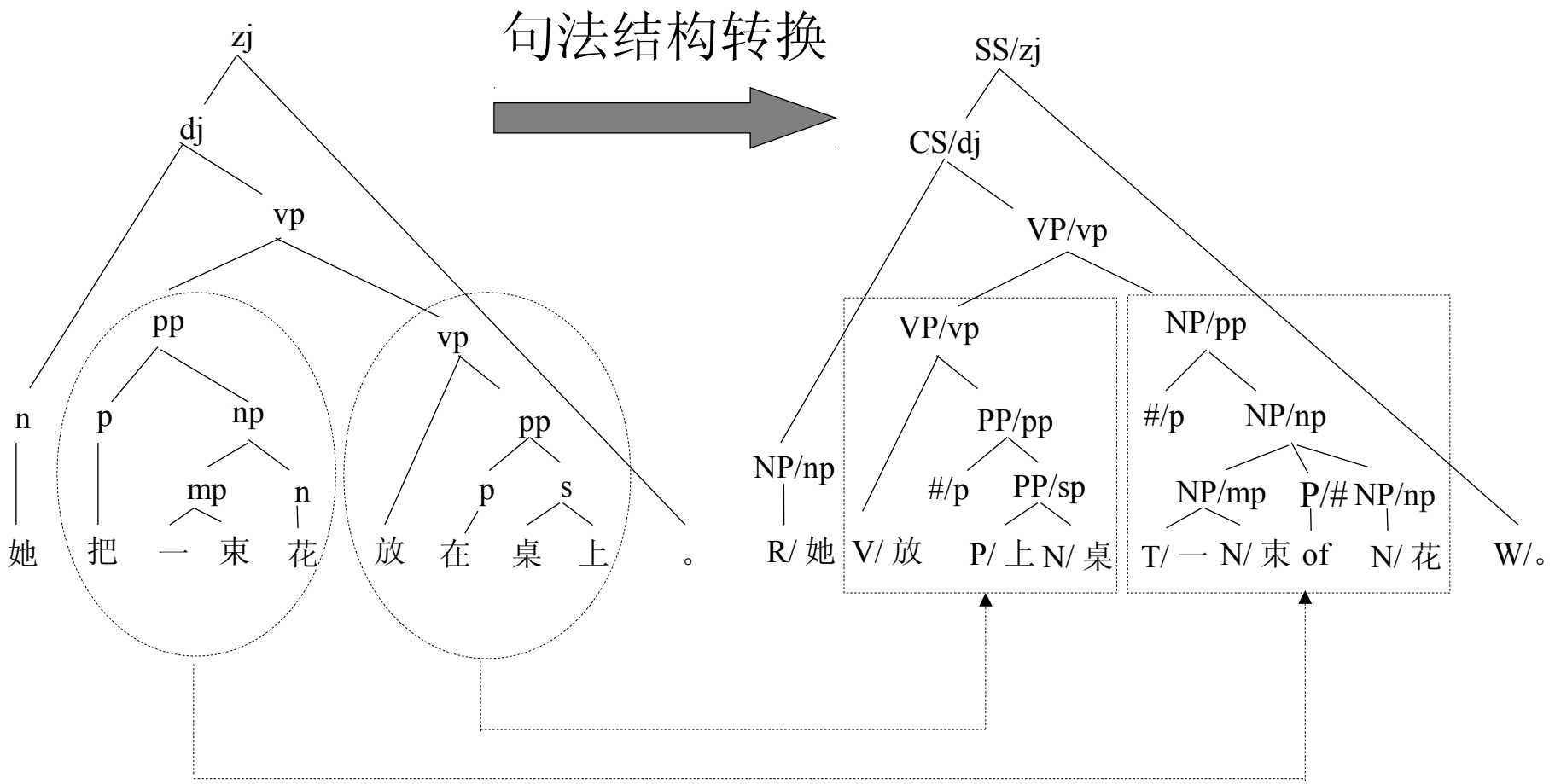


句法层面的转换方法（2）

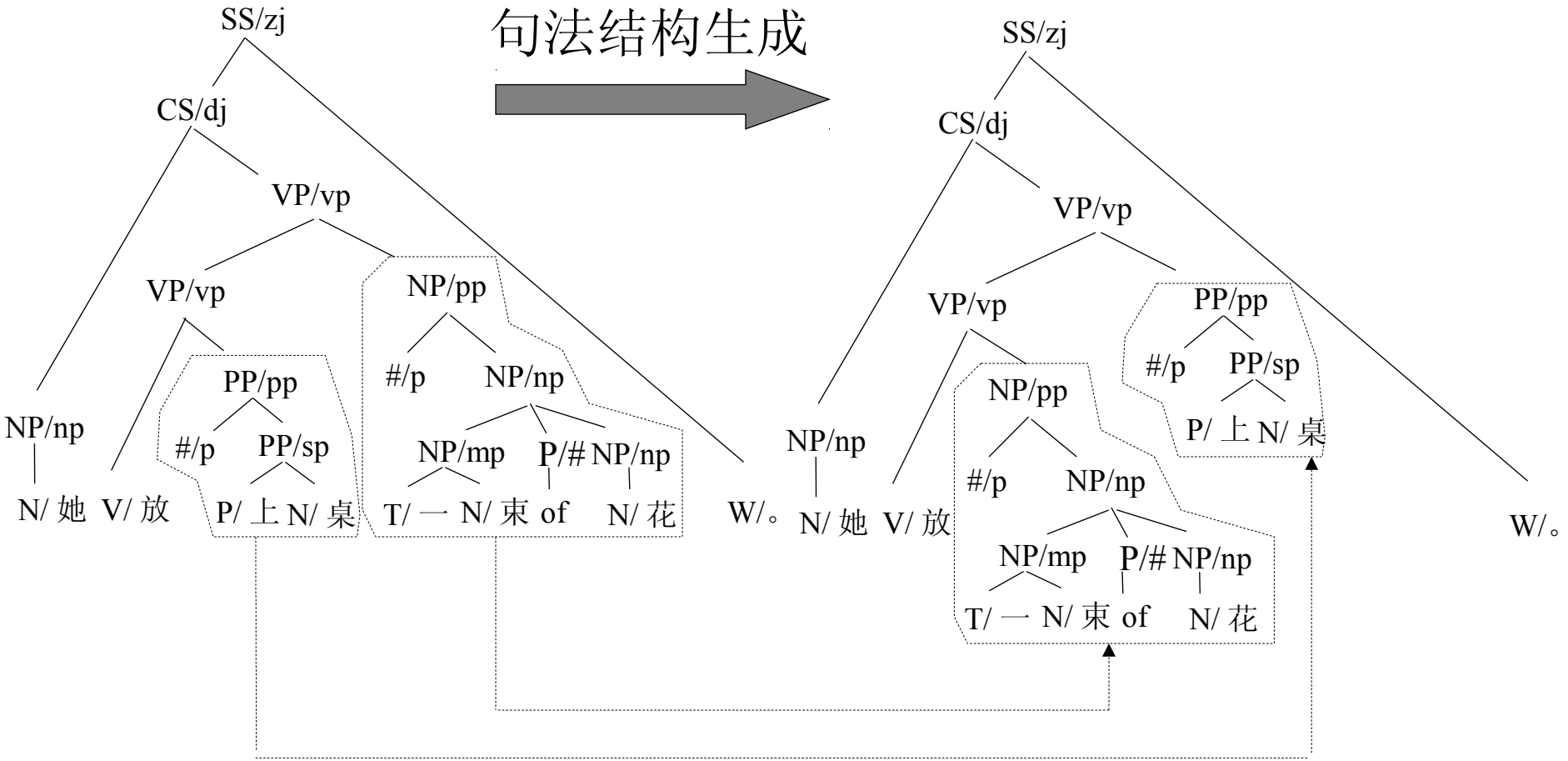
她/ r　把/ p　一/ m-d　束/ q　花/ n　放/ v　在/ p-v　桌/ n　上/ f-v　。/w



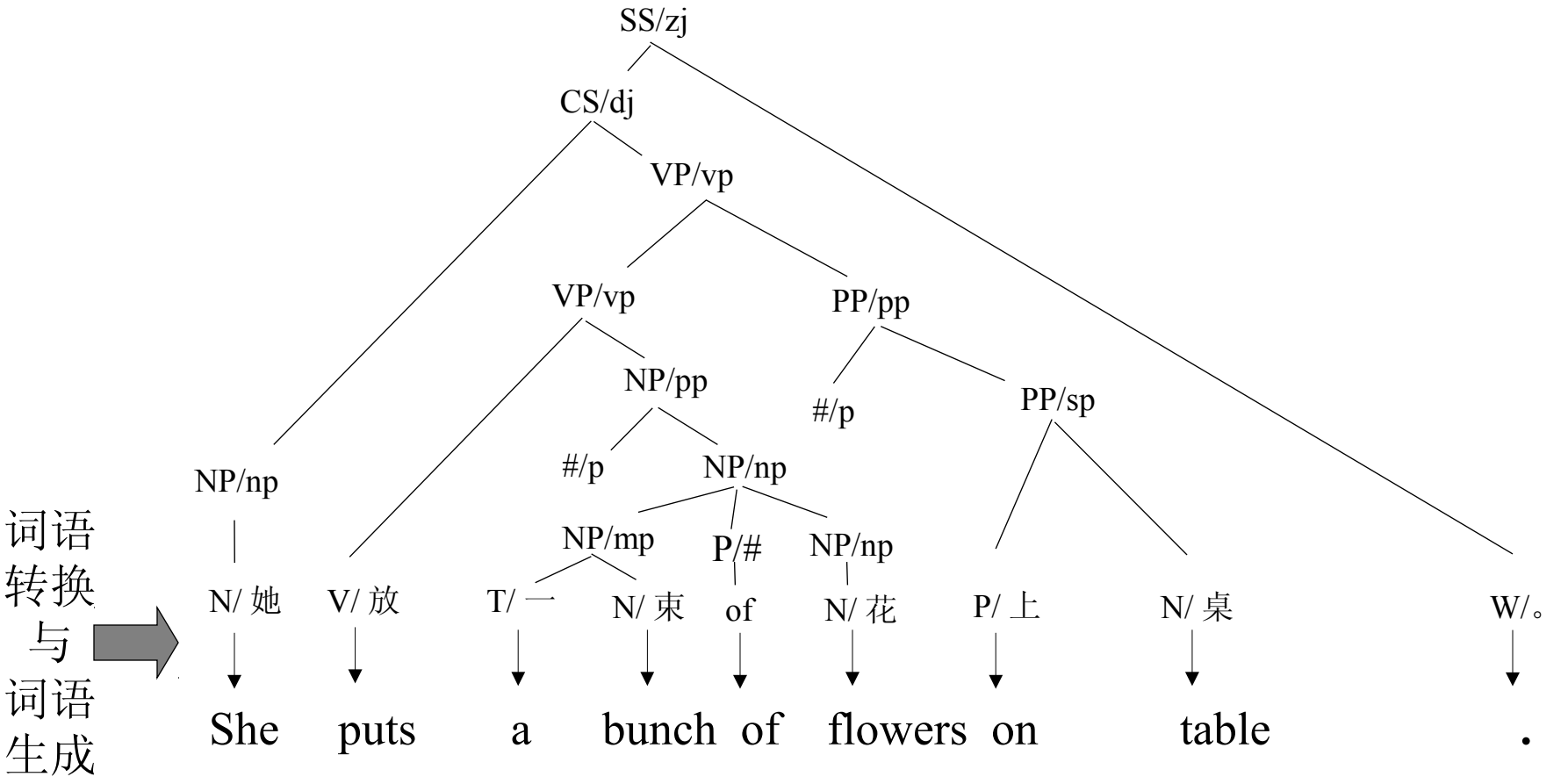
句法层面的转换方法（3）


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

句法层面的转换方法（5）


词语转换与词语生成
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
N/ 她
V/ 放
T/ 一
N/ 束
of
N/ 花
P/ 上
N/ 桌
W/。
She
puts
a
bunch
of
flowers
on
table
.

机器翻译的转换层面

直接翻译方法

句法转换方法

语义转换方法

中间语言方法

语义层面的转换方法

机器翻译的转换层面

直接翻译方法

句法转换方法

语义转换方法

中间语言方法

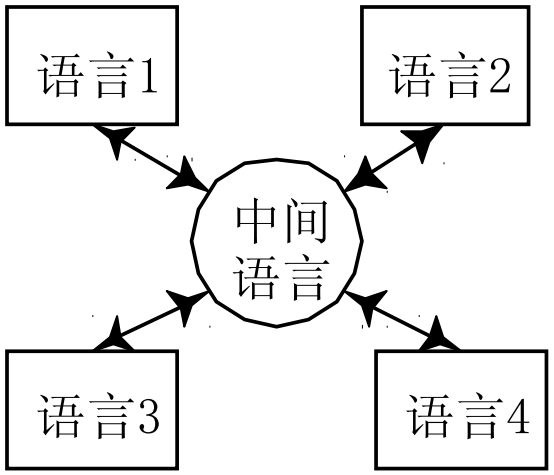
中间语言方法（1）

- 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

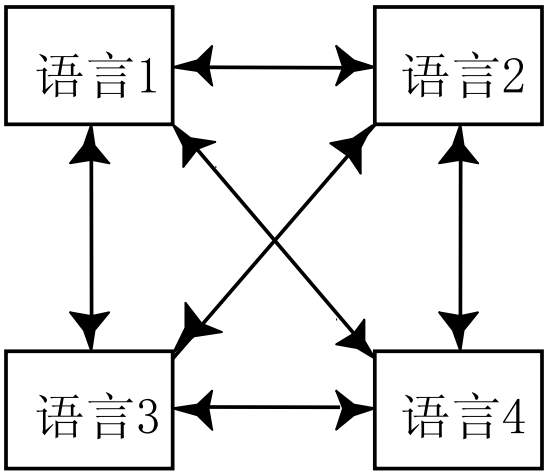
中间语言方法（2）

- 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（interlingua）通常是一种独立于语言的语义或者知识表示形式，常见的形式有：
  - 语义网络（Semantic Network）（注意：这与语义网 Semantic Web 是完全不同的概念）
  - 框架（Frame）
  - 逻辑（Logic）
- 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)

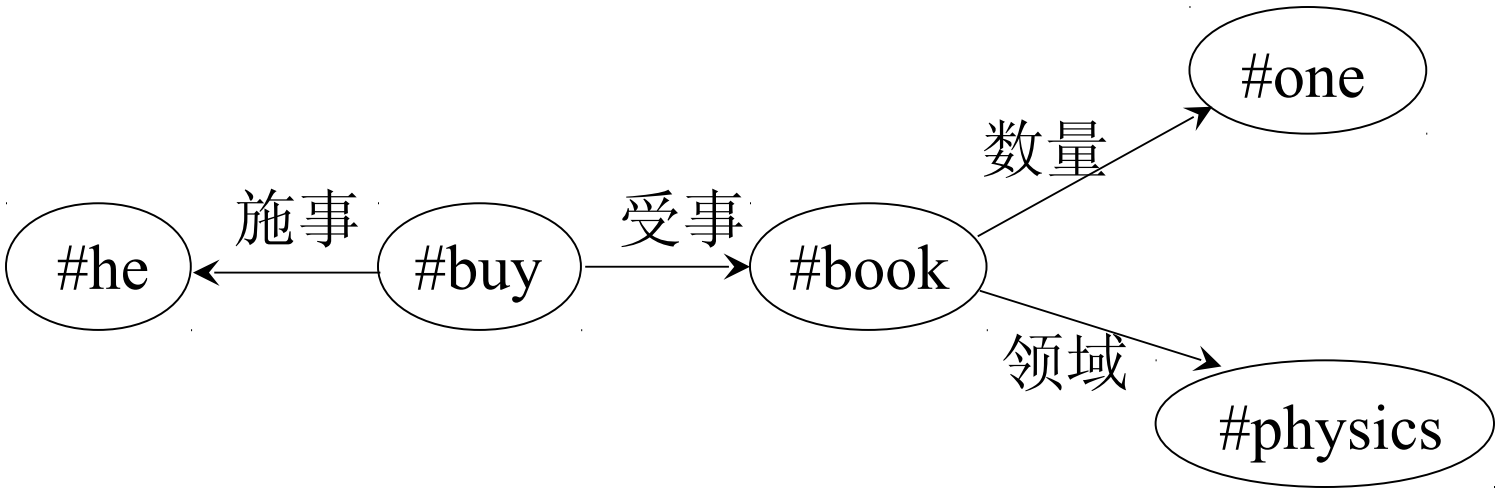
中间语言方法（6）

- 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，采用某种人工定义的知识表示形式作为中间语言进行多语言机器翻译都不太成功，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 多国语口语机器翻译系统中，曾经采用了一种中间语言方法，其中间语言是一种带话语信息的语义表示形式，由于语音翻译都限制在非常狭窄的领域中（如旅游领域或机票预定），语义描述可以做到比较精确，因此采用中间语言方法有一定的合理性。但该方法最终也不成功。
- 实际上，领域特别窄的场合可以采用中间语言方法。一个适合于中间语言方法的例子是数词的翻译，采用阿拉伯数字作为中间语言显然是非常合理的。

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。

中间语言示例—框架

英语： He bought a book on physics.

汉语：他买了一本关于物理学的书。

谓词	概念	#buy	
	施事	概念	#he
	受事	概念	#book
		数量	#one
		领域	#physics

说明：这里#后面表示的是概念，而不是英语词。

中间语言示例—概念词典

概念	语义类	中文词	英文	格框架
#he	指代词	他	he	
#buy	获得	买	buy	施事，受事
#book	出版物	书	book	
#physics	学科	物理	physics	
#one	数量	一	one	

中介语言方法

- 在多语言机器翻译中，很多研究人员开始采用某种自然语言作为中介语言（这时又称“枢纽语言”或“桥接语言”，英文是Pivot Language）。 中介语言也可以是人造语言，如世界语。
- 这种方法不同于前述的中间语言方法，因为这种方法中经过了两个独立的翻译过程，而这两个过程可能有各自独立的分析、转换、生成模块。而中间语言方法只有一个分析过程和一个转换过程。
- 也有文献把中介语言方法归入到中间语言方法，阅读文献的时候请注意区分这些概念的具体含义。
- 多语言统计机器翻译（如 Google 翻译）比较适合采用中介语言方法。主要原因是英语到其他语言的双语语料库比较容易获得，而其他语言之间的双语语料库很难获得。

内容提要

基于规则的方法

基于实例的方法

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。

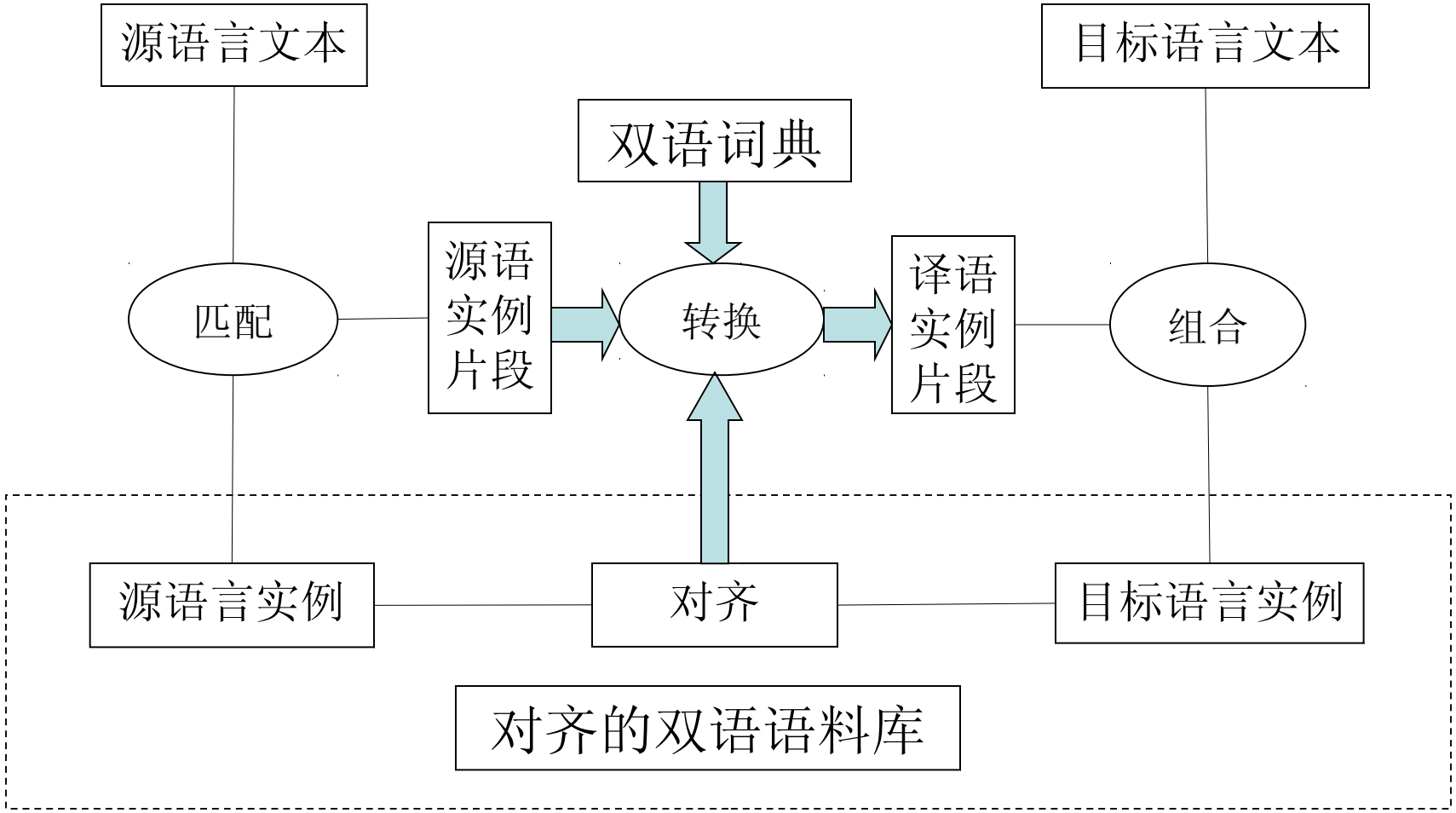
基于实例的机器翻译（2）

- 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—举例

- [Sato & Nagao 1990] 的方法：
  - 将实例按照**词语依存树**配对的形式进行存储，同时保存结点对应关系链接的集合

He eats vegetables.

e([e1,[eat,v],
  [e2,[he,[pron]],
  [e3,[vegetable,n]]]).

Kare ha yasai wo taberu.

e([j1,[taberu,v],
  [j2,[ha,p],
  [j3,[kare,pron]]],
  [j4,[wo,p],
  [j5,[yasai,n]]]]).

clinks([[e1, j1], [e2, j3], [e3, j5]]).

基于实例的机器翻译—举例

- [Sato & Nagao 1990] 的方法：
  - 在翻译的过程中，每一个输入句子都被表示为一个或多个匹配表达式。
  - 每一个匹配表达式表示在实例库中找到的某个依存子树的特定结点上所进行的某种操作（即插入、删除和替换）。
  - 利用这些操作，可以通过数据库中找到的实例片段来组合得到输入的句子。

输入英语句子：“ He eats mashed potatoes.”
匹配表达式为：[e1, [r, e3, [e^x]]]
这里 r 表示替换，整个表达式的意思是
“ 在实例 e1 中，用结点 e^x 替换结点 e3”

基于实例的机器翻译需要研究的问题（1/2）

- 双语自动对齐（alignment）：在实例库中要能准确地由源语言实例和实例片段找到相应的目标语言实例和实例片段，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或句子结构一级的对齐。
- 实例片段的定义：实例片段可以定义在句子级别、子句级别、短语级别，或者定义为某种句法结构的片段。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，从而又会导致翻译质量的下降。需要在二者之间取得平衡。

基于实例的机器翻译需要研究的问题（2/2）

- 实例匹配检索：由于实例库规模巨大，为了在实例库中迅速找到与要翻译的句子匹配的实例或者实例片段，需要建立高效的检索机制。另外，实例和实例片段的匹配通常都不是精确匹配，而是模糊匹配，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。
- 译文片段的选择：对于一个源文片段，可能有多个译文片段与其对应，为此需要选择恰当的译文片段。这实际上也是一个排歧问题。
- 实例片段的组合：得到实例片段的译文后，需要将实例片段重新组合成目标语言句子。这里通常涉及词序调整问题。

双语自动对齐

- 自动对齐技术简介
- 句子对齐
- 词语对齐

平行语料库的对齐

- 实例库又称双语语料库（Bilingual Corpus）或平行语料库（Parallel Corpus）
- 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 $A=\{A_1,\ldots,A_k\}$， 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：

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

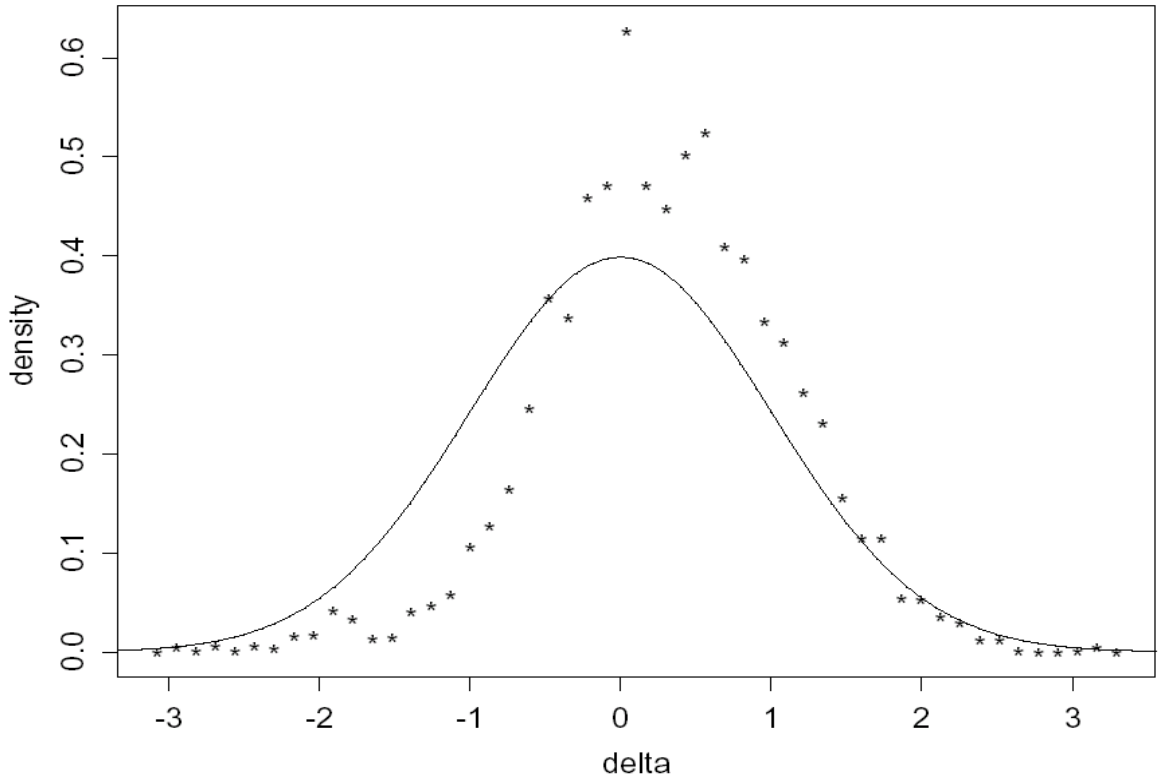
基于长度的句子对齐（1）

- 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）

$$P(A_i) = p(S_i, T_i)$$
$$\approx p(l_{S_i}, l_{T_i}) \times p(m_{S_i}, m_{T_i})$$

基于长度的句子对齐（2）

- 根据统计，随机变量 $X = l_{T_i} / l_{S_i}$ 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：

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

- 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

句子对齐搜索算法

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

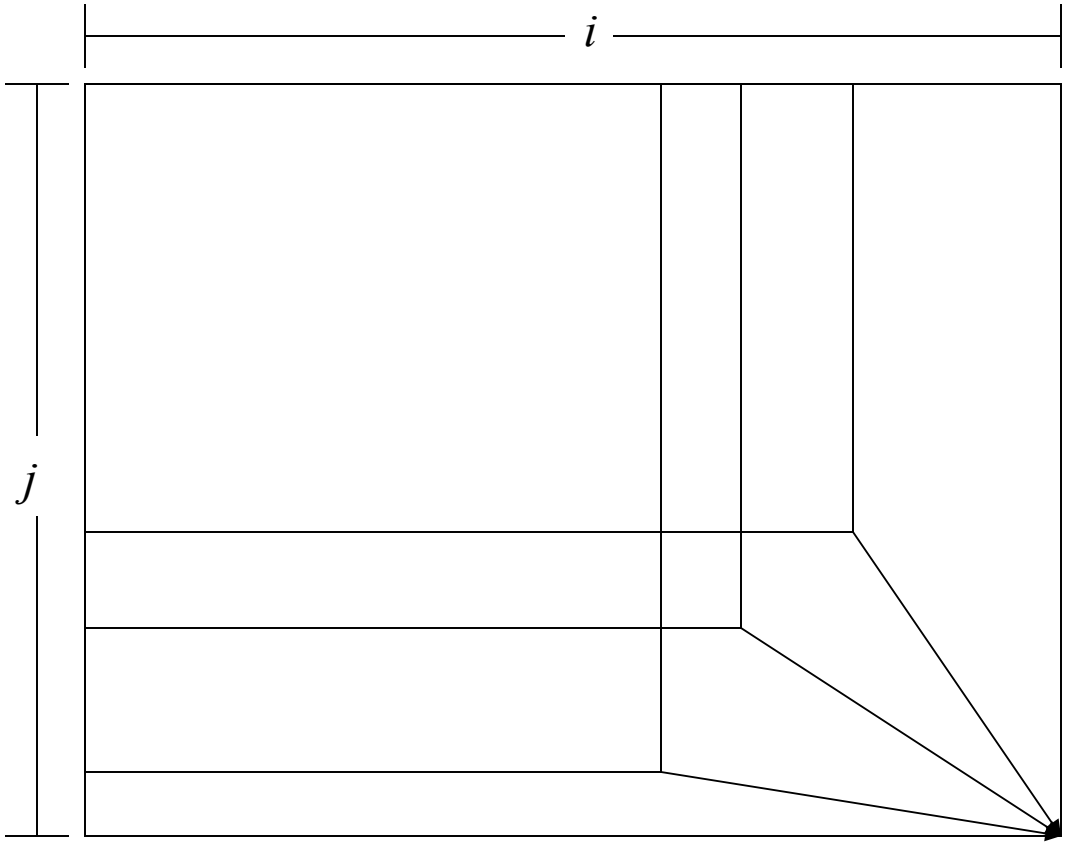
$$score(i,j)=\max_{x,y=0}^{k} score(i-x, j-y) + \log p(s_{i-x+1} \ldots s_i, t_{j-y+1} \ldots t_j)$$

上面假设一个句珠中最多只有 k 个句子，

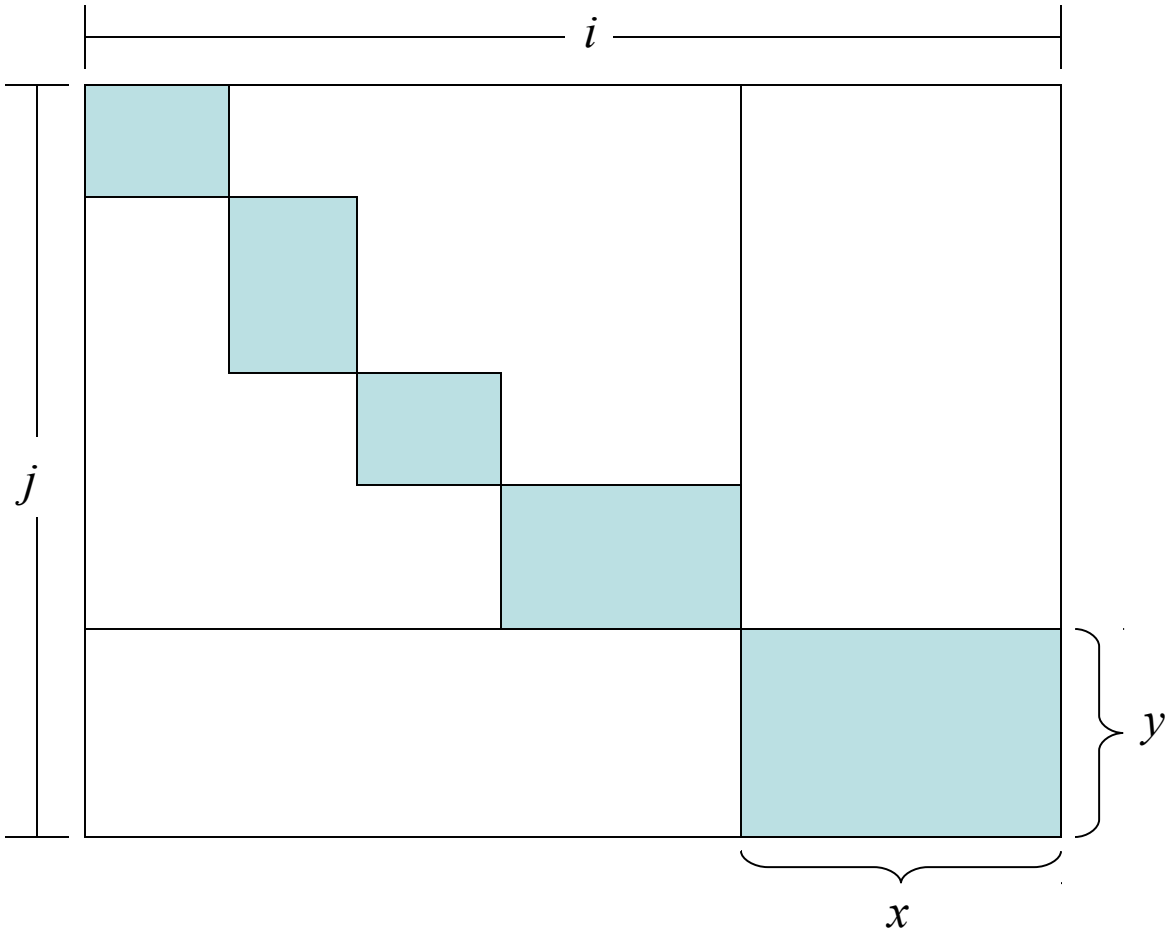
$(s_{i-x+1}s_i, t_{j-y+1}t_j)$ 为一个句珠

- 最优对齐为 $P(m,n)$ 所对应的路径

句子对齐搜索算法



句子对齐搜索算法



基于长度的句子对齐（5）

- 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）

$$p(A_i) = p(S_i, T_i)$$
$$\approx p(w_{S_i}, w_{T_i}) \times p(m_{S_i}, m_{T_i})$$

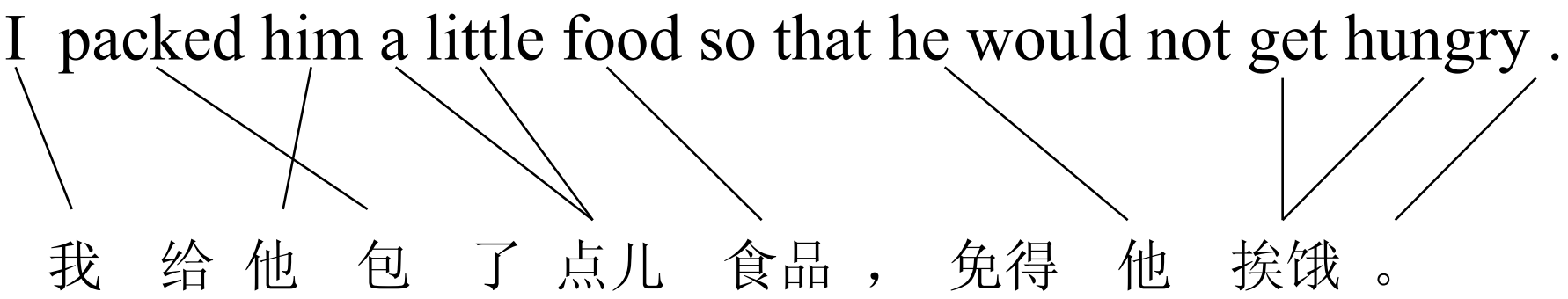
基于词的句子对齐（2）

- 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）
    两个词语的意义越相似，对齐的可能性越大
  - 位置扭曲模型（word distortion model）
    词语语序变化越小，对齐的可能性越大
- 用公式表示如下：

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

双语词语相似度计算

- 基于双语词典的方法
- 基于双语句子对齐语料库的方法

基于词典的双语词语相似度计算

$$Sim(E,C)=\max\{\max_{dict(E,C')} Sim(C',C),\ \max_{dict(E',C)} Sim(E,E')\}$$

- *Sim*(*,*) 为词语相似度
- *E*、*E'* 为英文词， *C*、*C'* 为中文词
- *dict*(*E*,*C*) 表示双语词典中存在条目（*E*,*C*)
- 上述公式可以利用一部双语词典，将双语词语的相似度计算转换为单语词语的相似度计算

单语词语相似度计算

- 基于字面相似度的方法
- 基于同义词词典(Thesaurus)的方法

基于字面相似度的单语词语相似度计算

- 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

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

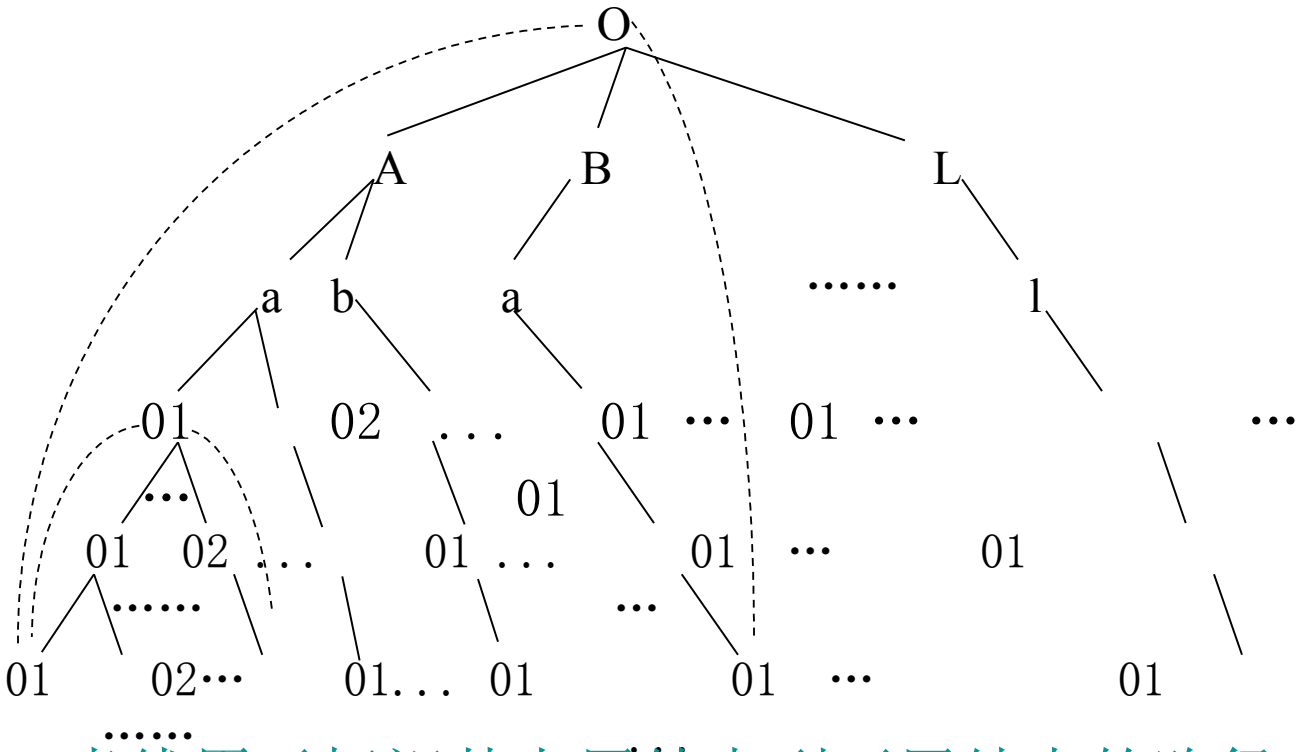
- 把汉语词理解为汉字的集合，戴斯系数就是两个词中相同的汉字占两个词汉字总数的比例。考虑到汉字表意性，这种方法在计算汉语词相似度时有较好的效果
- 英语词语相似度也可以用戴斯系数来计算，不过计算的时候两个词的交集应只考虑前缀相同的部分
- 某些双语词语相似度也可以直接利用戴斯系数进行计算，如汉语和日语、英语和法语等

基于同义词词典的单语词语相似度计算（1/3）

- 同义词词典（Thesaurus）通常将所有词语根据语义的相似性组织成一棵树的形式，这种形式通常称为一个概念层次体系或者一个知识本体（Ontology）
- 在一个概念层次体系中，两个词的距离远近，可以刻画对两个词语义相似程度，同一结点上的两个词为同义词，距离越远，相似度越低
- 常见的同义词词典
  - Wordnet（原始版本为英语，很多语言有对应版本）
  - Hownet（中英文）
  - 同义词词林（中文）
  - Roget's Thesaurus（英语）

基于同义词词典的单语词语相似度计算（2/3）

《同义词词林》的五层概念层次体系：



虚线用于标识某上层结点到下层结点的路径

基于同义词词典的单语词语相似度计算（3/3）

- 将词语距离转化为相似度：

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

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 是一个可调节的参数

基于语料库的双语词语相似度计算（1/5）

- 利用一部句子对齐的双语语料库
- 如果一对词语总是出现在对齐的双语句子中，我们就倾向于认为该对词语相似度较高

基于语料库的双语词语相似度计算（2/5）

- 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

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

- 给定中文词 C 和英文词 E
- 假设 C 和 E 在句子对齐的语料库中出现的句子对集合分别是 S_C 和 S_E
- 可以用 S_C 和 S_E 的戴斯系数来估计词语 C 和 E 的相似度

基于语料库的双语词语相似度计算（3/5）

- 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：

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

- 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。

基于语料库的双语词语相似度计算（4/5）

∀ χ^2（chi-square）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)

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

∀ χ^2 方法的效果比较好

基于语料库的双语词语相似度计算（5/5）

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

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

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

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

位置扭曲模型（1/3）

- 相对偏移模型：

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

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

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

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

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

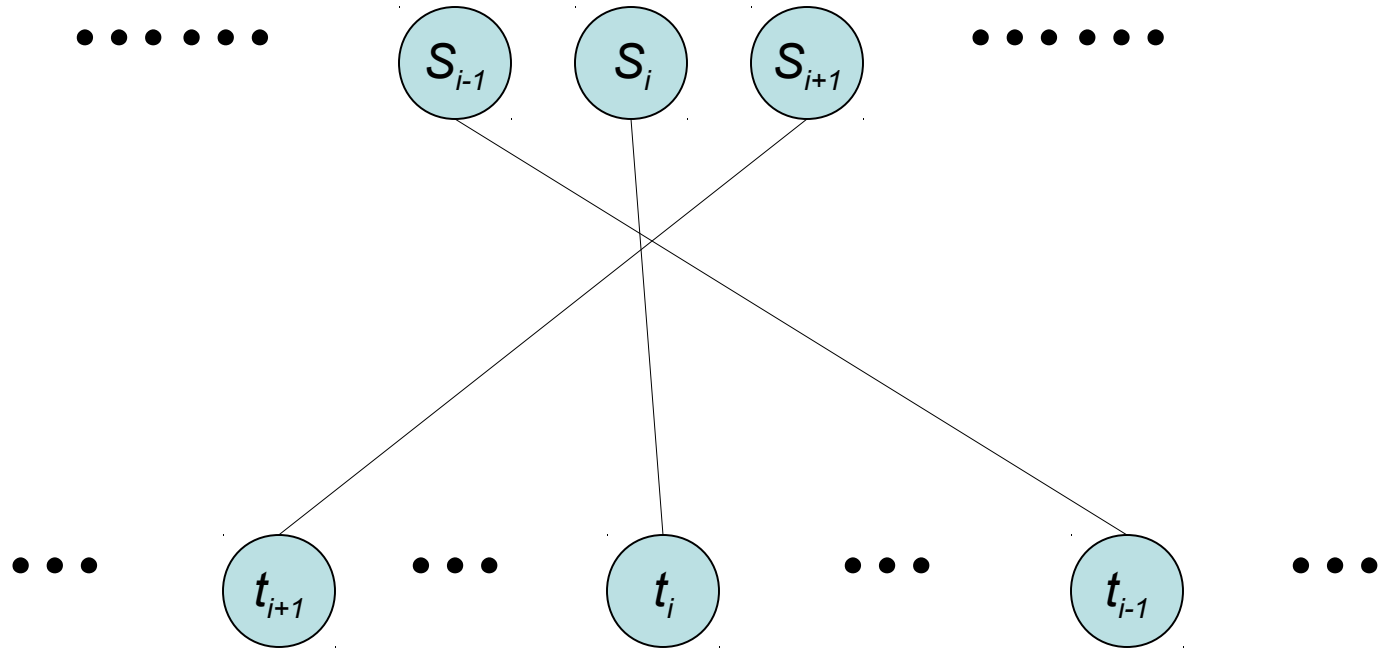
s_{i-1} 是 s_i 左侧最近的一个对齐的单词

s_{i+1} 是 s_i 右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

位置扭曲模型（2/3）

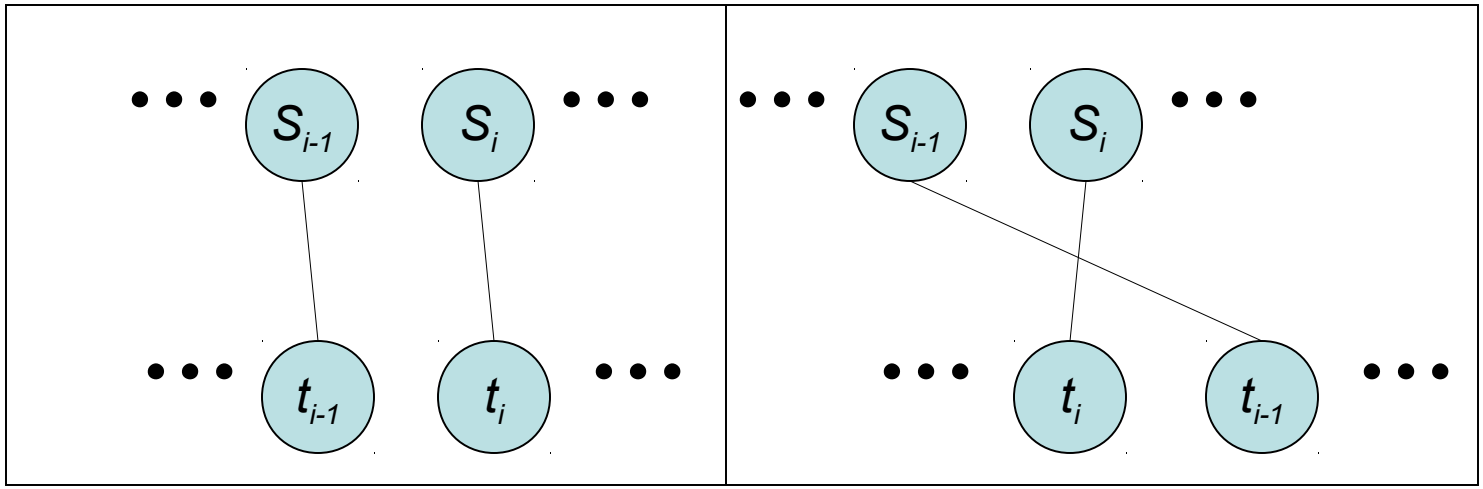


根据左右两侧相邻词的位置计算扭曲值

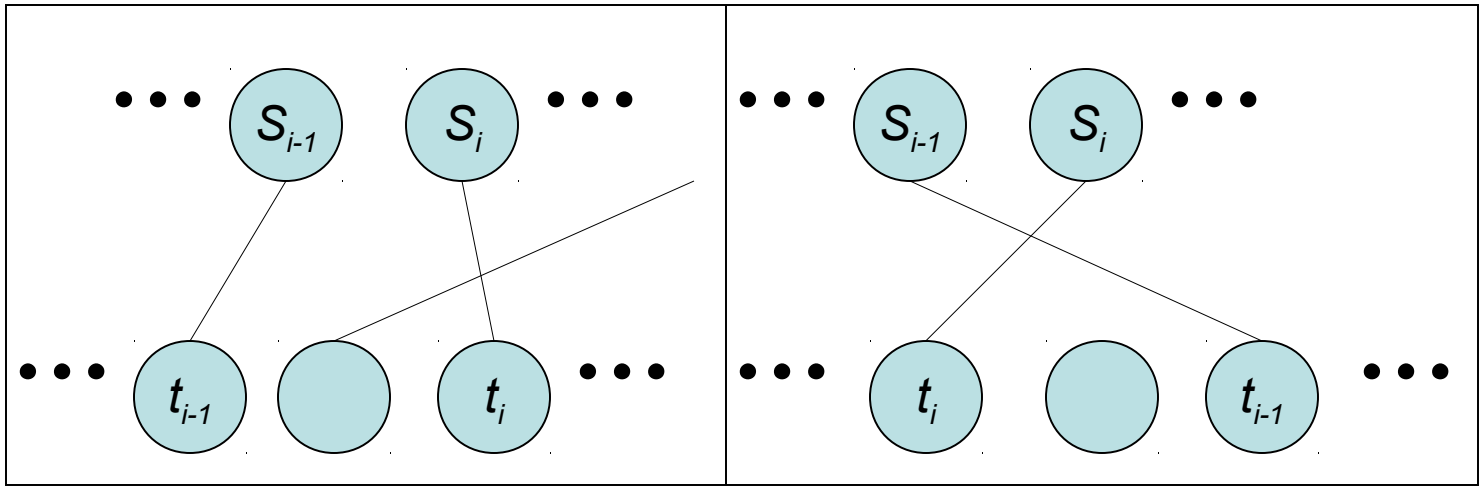
取其中较小的值

位置扭曲模型（3/3）

$dis(i,j)=0$



$dis(i,j)=1$



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. 计算所有候选词对的评价函数
  4. 找出最好的对齐词对，使得对齐总体评分最高
  5. 从候选集合中删除刚找出的词对
  6. 删除与刚找出的词对冲突的词对
  7. 重复以上3～6，直到对齐总评分不再增加

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：最直接的方法：贪心法
- 词语对齐的副产品：双语词典抽取

实例片段的定义

- 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段
- 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
  - 通过句法结构进行分解

实例库的匹配

- 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；
- 实例库匹配的相似度计算：

译文片段的选择与组合？

- 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
- 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？？？

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。

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

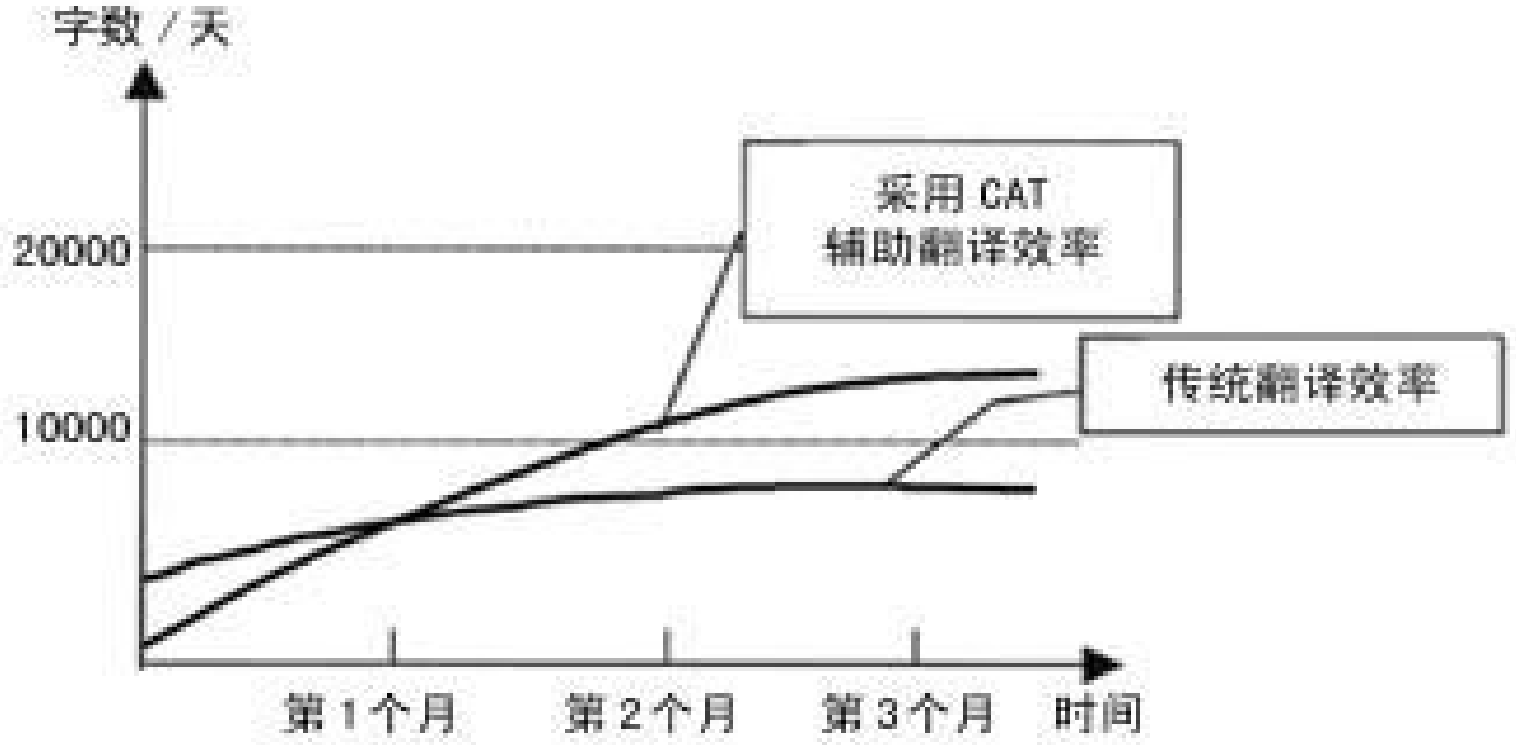
翻译记忆方法（3）

- 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

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

内容提要

- 机器翻译方法（按转换层面划分）
  - 直接翻译方法
  - 句法转换方法
  - 语义转换方法
  - 中间语言方法
- 机器翻译方法（按知识表示形式划分）
  - 基于规则的方法
  - 基于实例的方法（含模板方法、翻译记忆方法）
  - **统计方法**

统计机器翻译概述

- 统计机器翻译也是基于语料库的机器翻译方法，不需要人工撰写规则，而是从语料库中获取翻译知识，这一点与基于实例的方法相同
- 为翻译建立统计模型，把翻译理解为搜索问题，即从所有可能的译文中选择概率最大的译文。基于实例的机器翻译无需建立统计模型
- 与基于实例的方法的区别在于，基于实例的机器翻译中，语言知识表现为实例本身，而统计机器翻译中，翻译知识表现为模型参数

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

统计机器翻译的优缺点

- 优点：
  - 无需人工编写规则，利用语料库直接训练得到机器翻译系统；（但可以使用语言资源）
  - 系统开发周期短；
  - 鲁棒性好；
  - 只要有语料库，很容易适应新的领域或者语种。
- 缺点：
  - 时空开销大；
  - 数据稀疏问题严重；
  - 对语料库依赖性强；
  - 引入复杂的语言知识比较困难。

思考题

- 采用基于句法层面的转换式机器翻译系统，尝试写规则将完成以下句子翻译：
  - 原文：他对语言学很有兴趣。
  - 译文： He is very interested in linguistics.
  - 源文：买这本书花多少钱？
  - 译文： How much does it take to by this book?