博士论文答辩报告

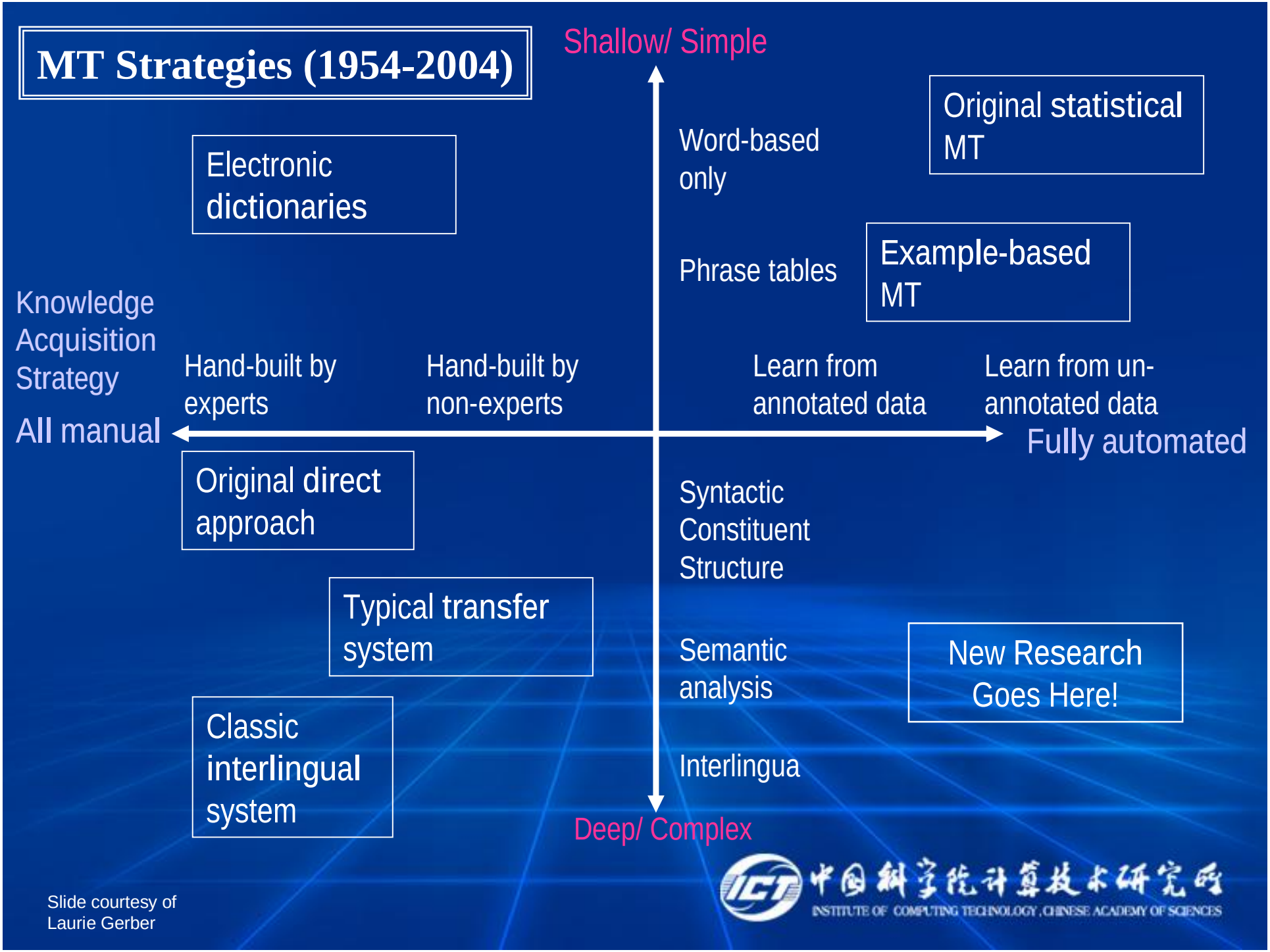
树到串统计翻译模型研究

答辩人：刘洋
指导教师：林守勋 研究员
时间：2007年6月16日

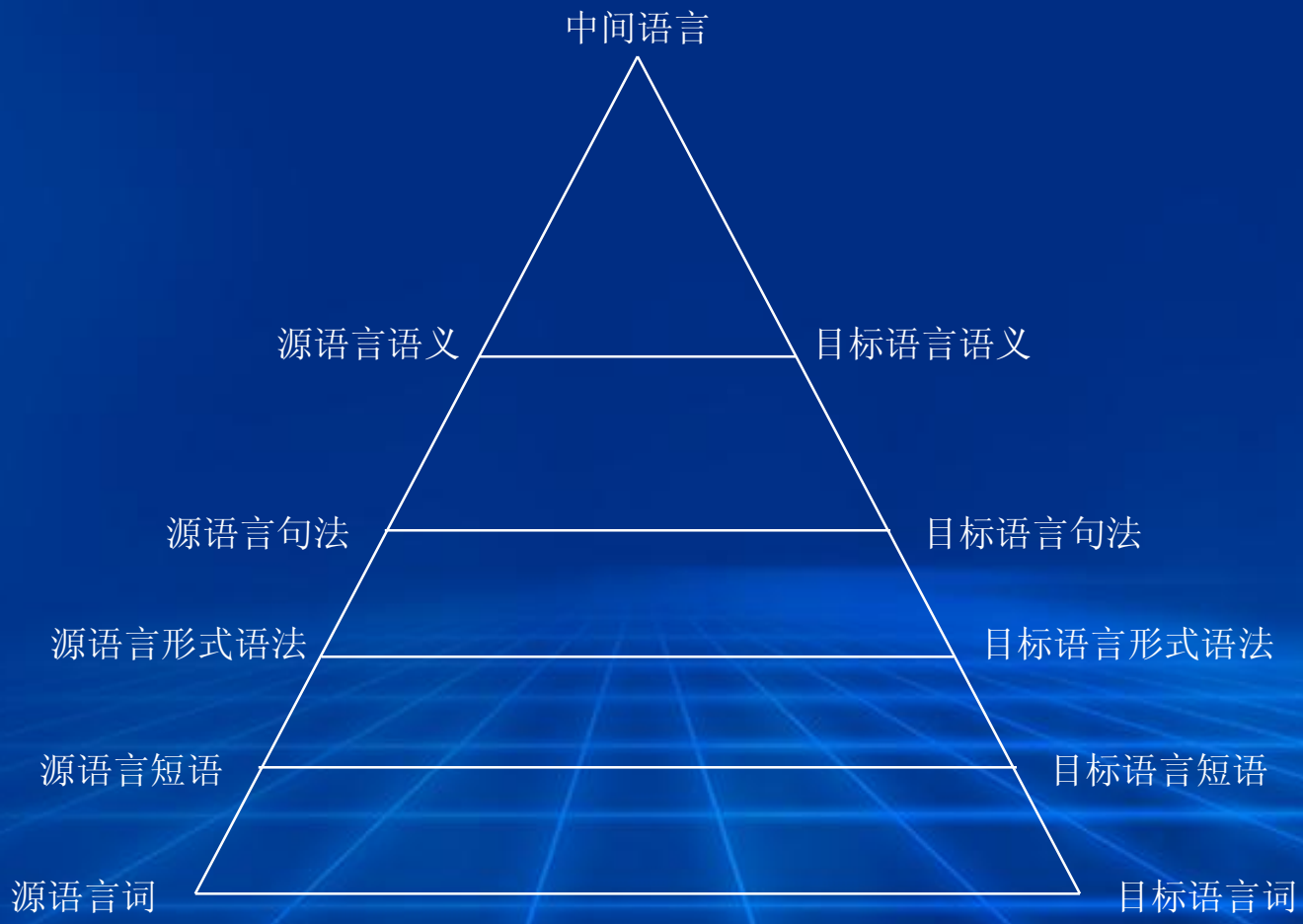
中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

提纲

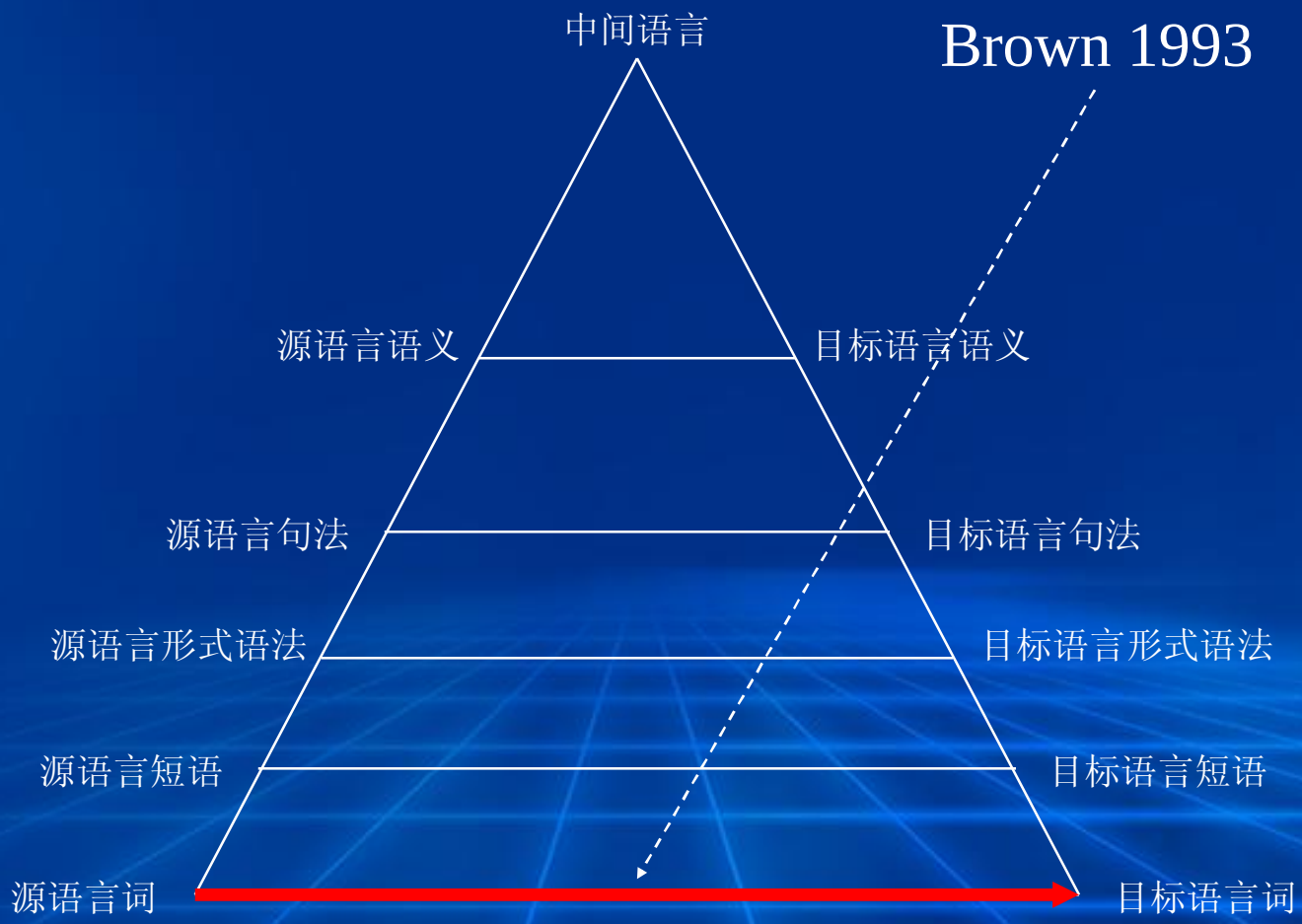
- 引言
- 词语对齐的对数线性模型
- 树到串统计翻译模型
  - 模型1
  - 模型2
  - 模型3
  - 实验
- 总结


MT Strategies (1954-2004)
Shallow/ Simple
Word-based only
Original statistical MT
Electronic dictionaries
Phrase tables
Example-based MT
Knowledge Acquisition Strategy
All manual
Hand-built by experts
Hand-built by non-experts
Learn from annotated data
Learn from un-annotated data
Fully automated
Original direct approach
Syntactic Constituent Structure
Typical transfer system
Semantic analysis
New Research Goes Here!
Classic interlingual system
Interlingua
Deep/ Complex
Slide courtesy of Laurie Gerber
中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

统计机器翻译

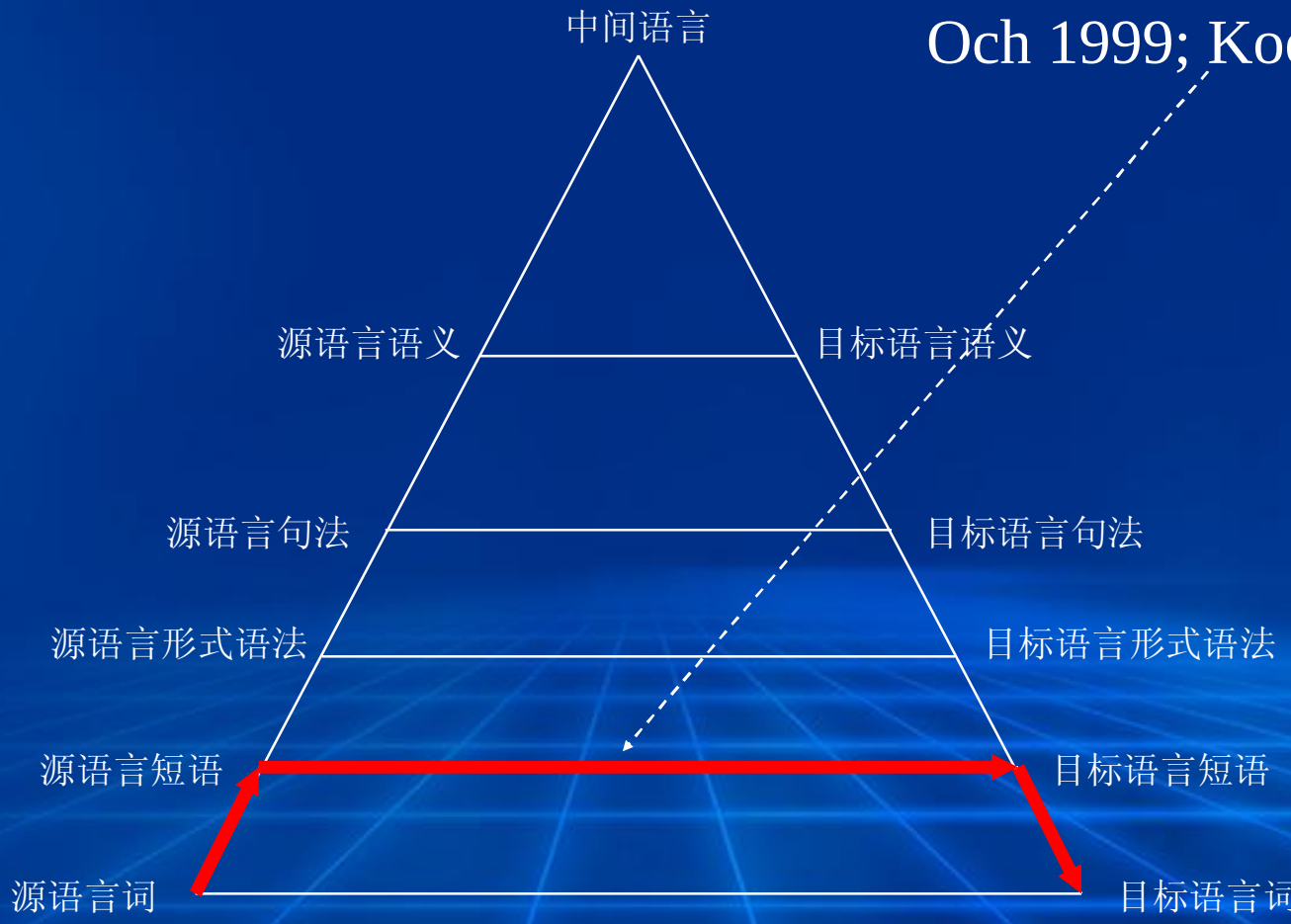


统计机器翻译

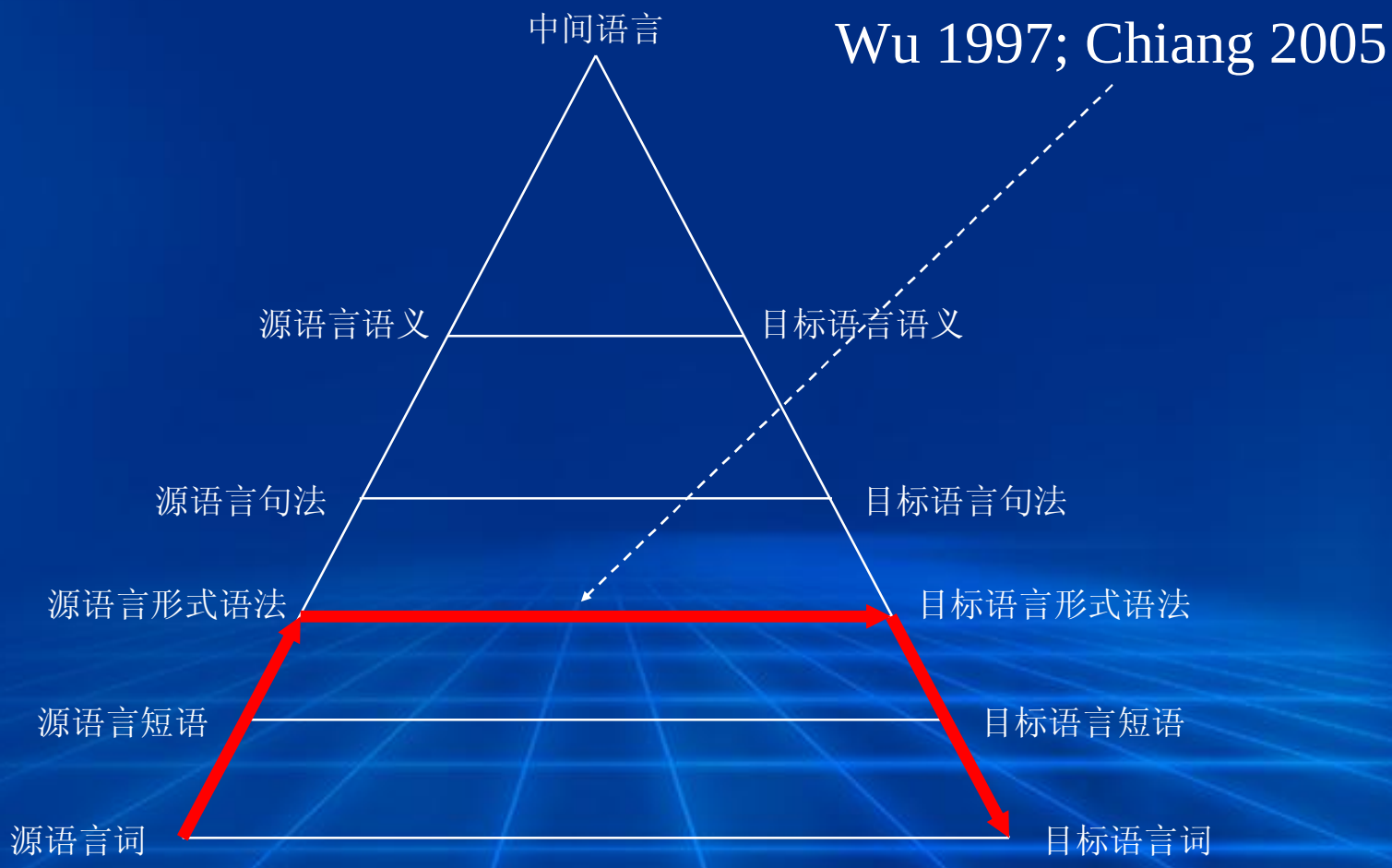


统计机器翻译

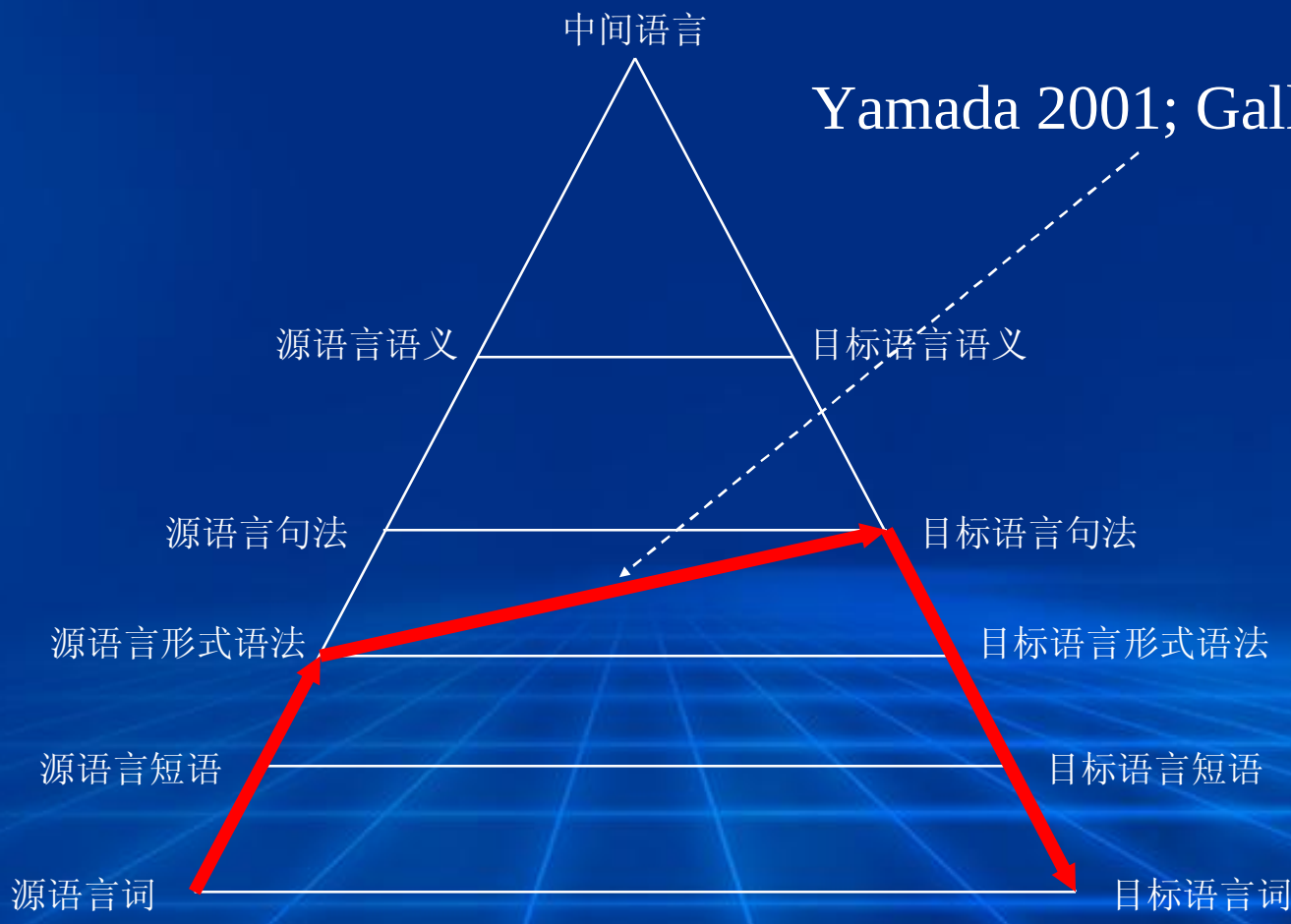
Och 1999; Koehn 2003



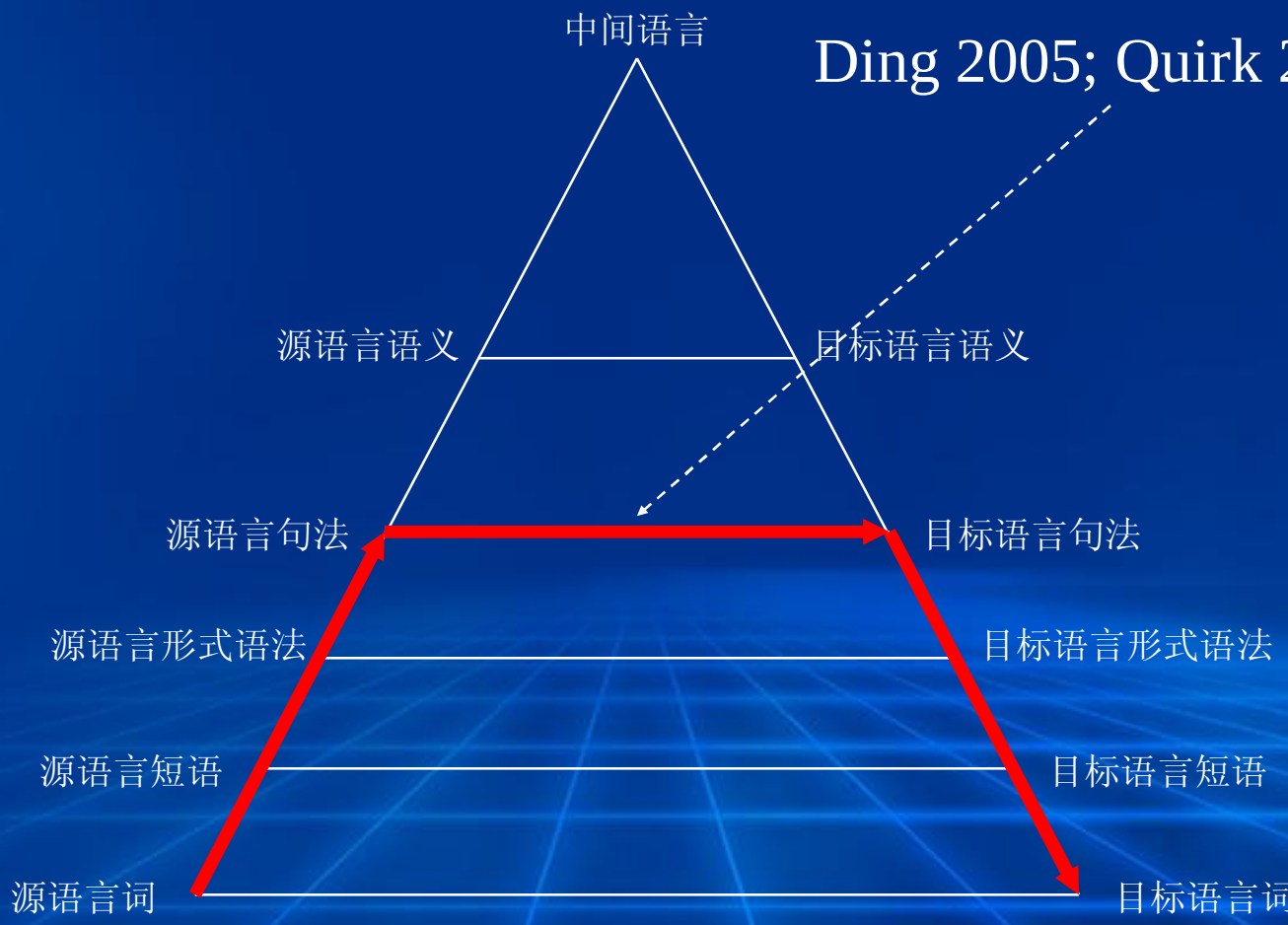
统计机器翻译



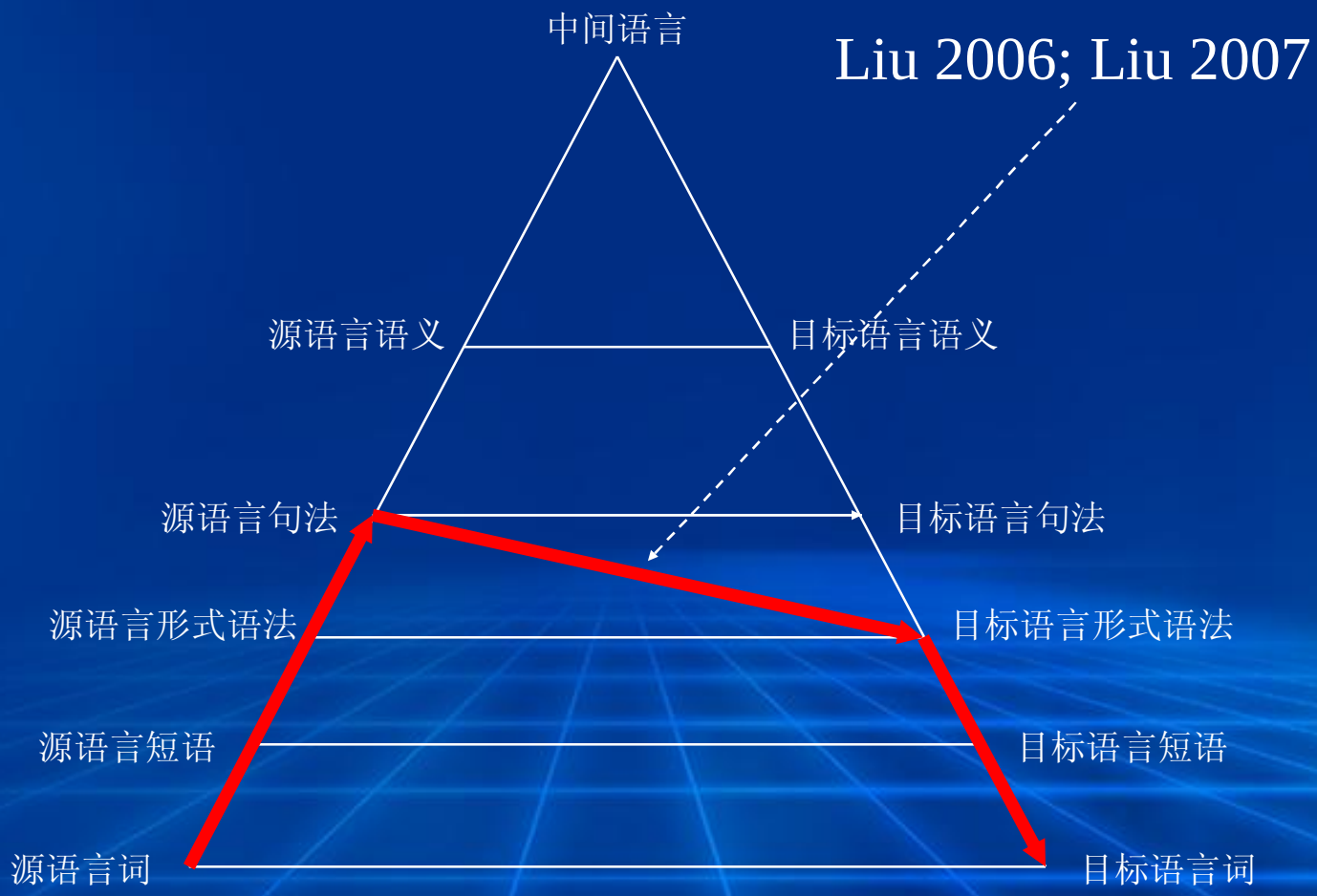
统计机器翻译



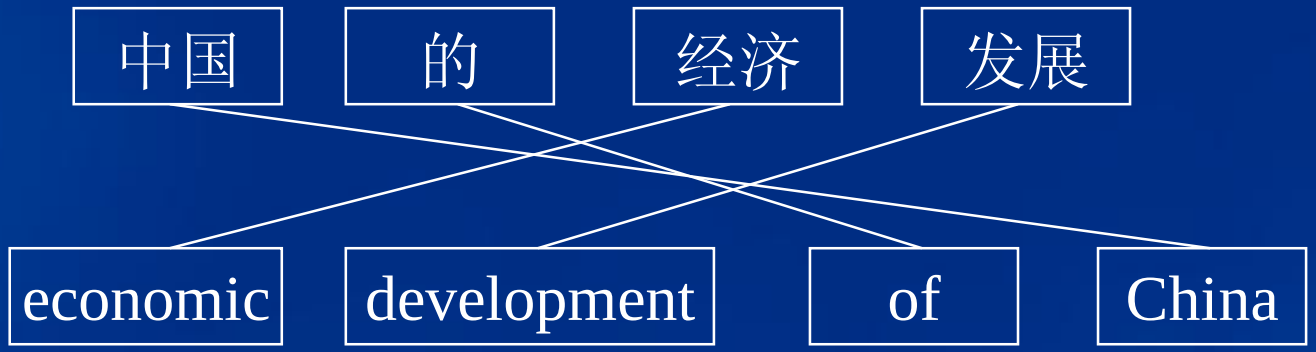
统计机器翻译



统计机器翻译



词语对齐



- 词语对齐是统计机器翻译中最重要的一种语料库标注

研究内容

- 词语对齐的对数线性模型
- 树到串统计翻译模型
  - 嵌入句法树的基于短语的翻译模型（模型1）
  - 基于树到串对齐模板的翻译模型（模型2）
  - 融入森林到串规则的树到串翻译模型（模型3）

提纲

- *引言*
- 词语对齐的对数线性模型
- 树到串统计翻译模型
  - 模型1
  - 模型2
  - 模型3
  - 实验
- 总结

词语对齐主要方法


Blunsom 2006
Och 2003
Liu 2005
Fraser 2006
Cherry 2003
Moore 2005
Liang 2006
Brown 1993
Vogel 1996
Ker 1997
Melamed 2000
Tiedemann 2003
Taskar 2005
Lacoste-Julien 2006

IBM模型的特点

- 优点
  - 语言无关性
  - 能够处理大规模数据
- 缺点
  - 难以扩展
  - 无法充分利用具体语言特性
  - 需要启发式策略简化搜索算法
  - 需要手工调参数

我的工作

- 提出了一种词语对齐的对数线性模型。该模型首次将判别方法引入词语对齐，具有良好的可扩展性。

词语对齐的对数线性模型

模型公式

$$\Pr(a \mid e, f) = \frac{\exp\left[\sum_{m=1}^{M} l_m h_m(a, e, f)\right]}{\sum_{a'} \exp\left[\sum_{m=1}^{M} l_m h_m(a', e, f)\right]}$$

搜索公式

$$\hat{a} = \arg\max_{a} \left\{\sum_{m=1}^{M} l_m h_m(a, e, f)\right\}$$

特征函数

- IBM模型
- 词性标记转换模型
- 双语词典
- 连线计数
- 交叉计数
- 词根还原的IBM模型
- 完全匹配

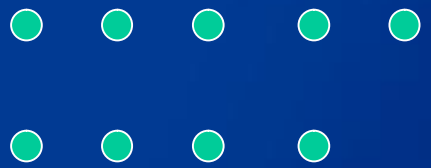
训练

- 目标：在开发集上自动学习特征权重
- 方法
  - 通用迭代算法
  - 最小错误率训练

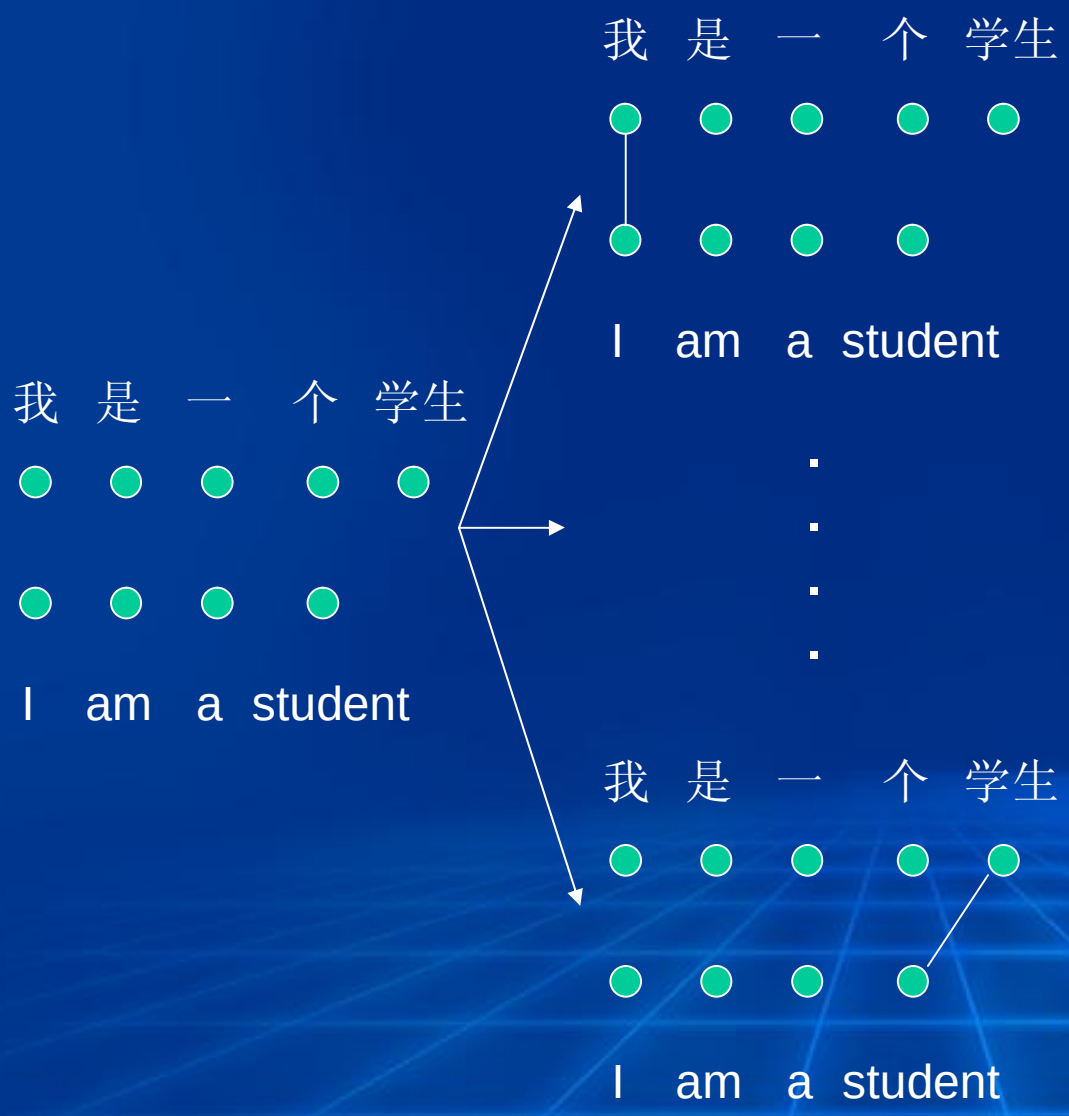
搜索

输入：源语言句子f，目标语言句子e，其他依赖关系
1. 初始化词语对齐： a=ϕ。 2. 对每个不属于a 的连线$l=(j,i)$计算增益$gain(\mathrm{a},l)$。 3. 如果对于任意的连线l，均有$gain(\mathrm{a},l)\le 0$，算法终止。 4. 向a 中添加增益$gain(\mathrm{a},l)$最大的连线$\hat{l}$。 5. 转到2。
输出：词语对齐a

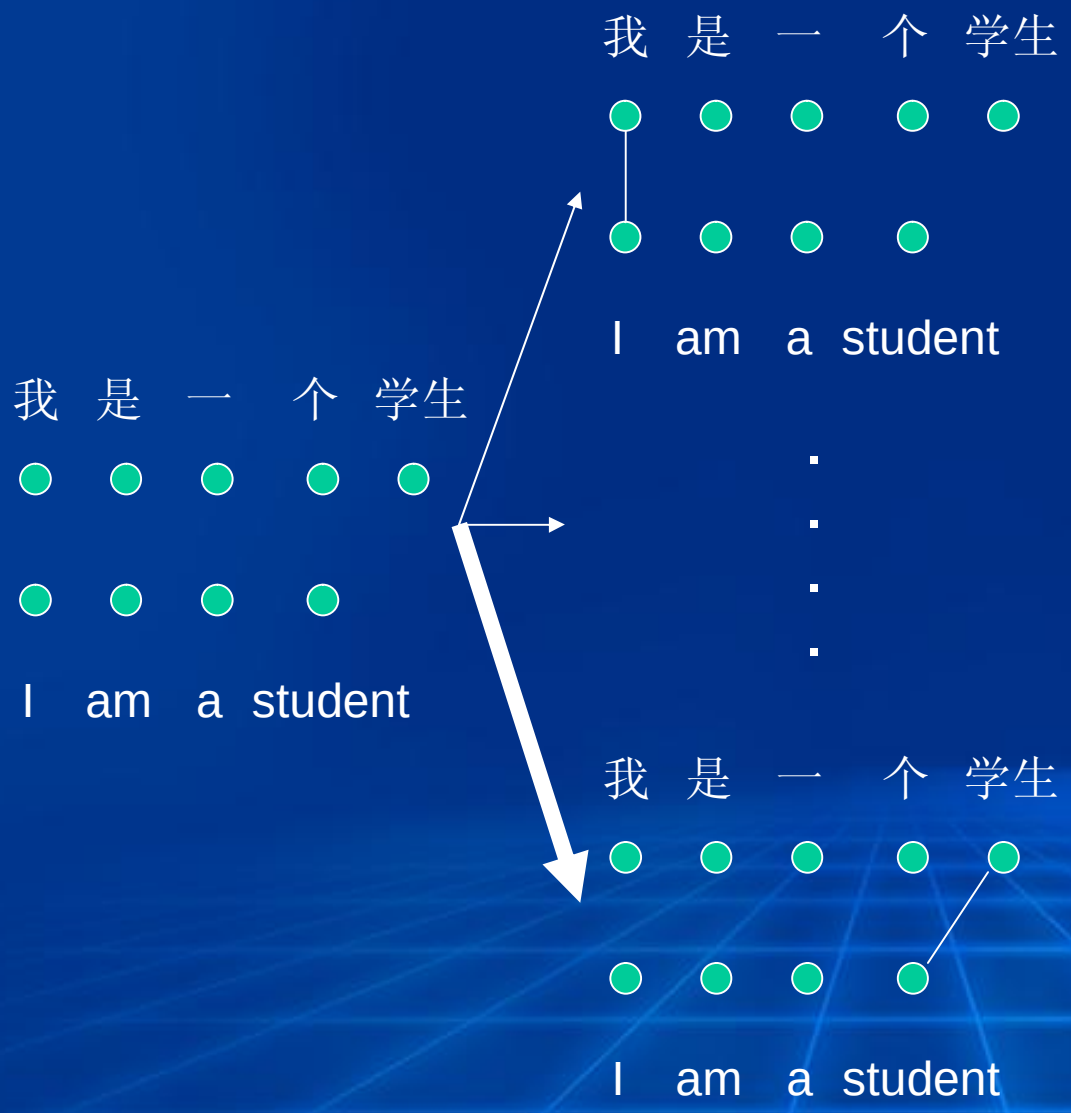
我 是 一 个 学生



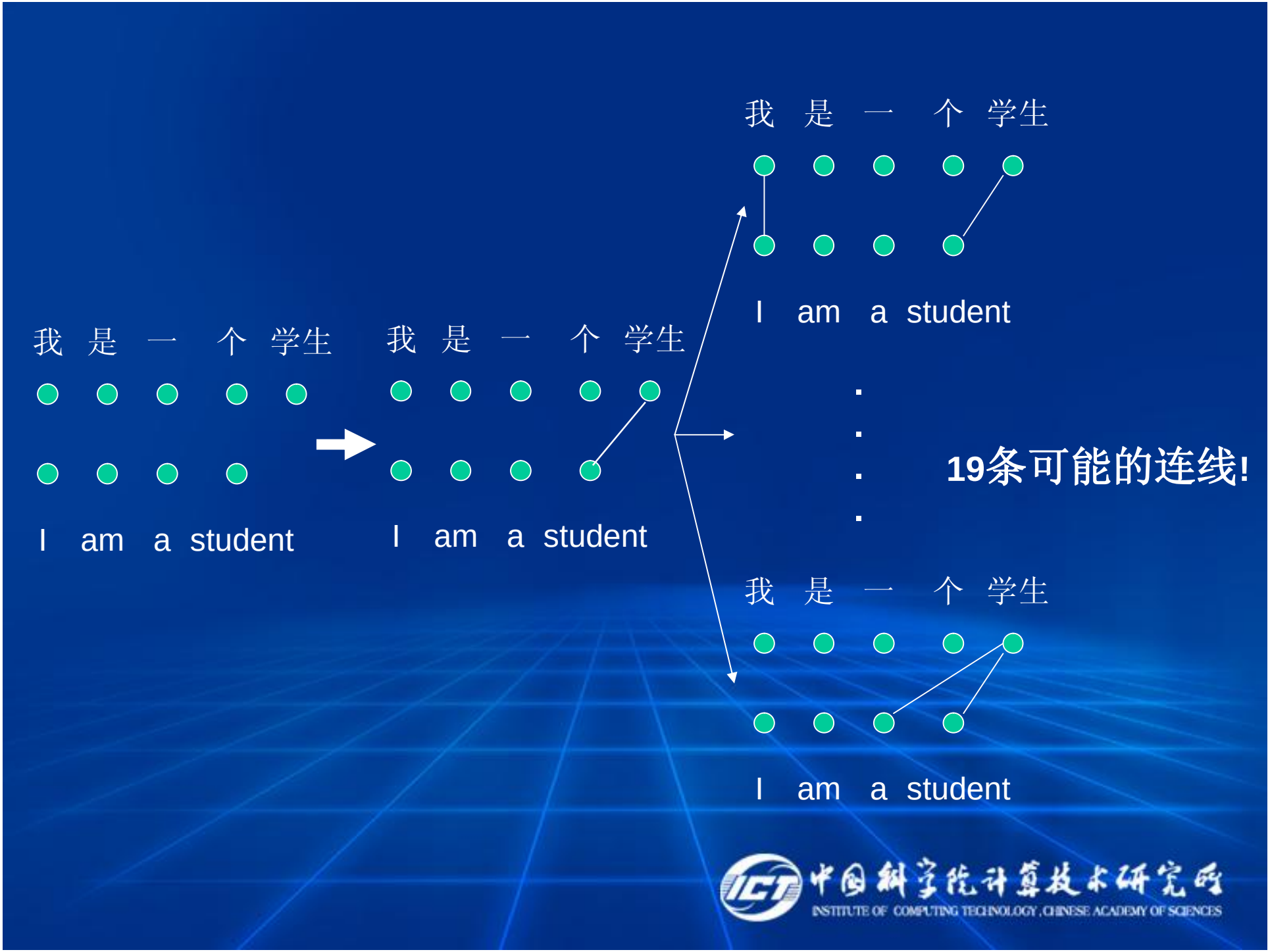
I am a student

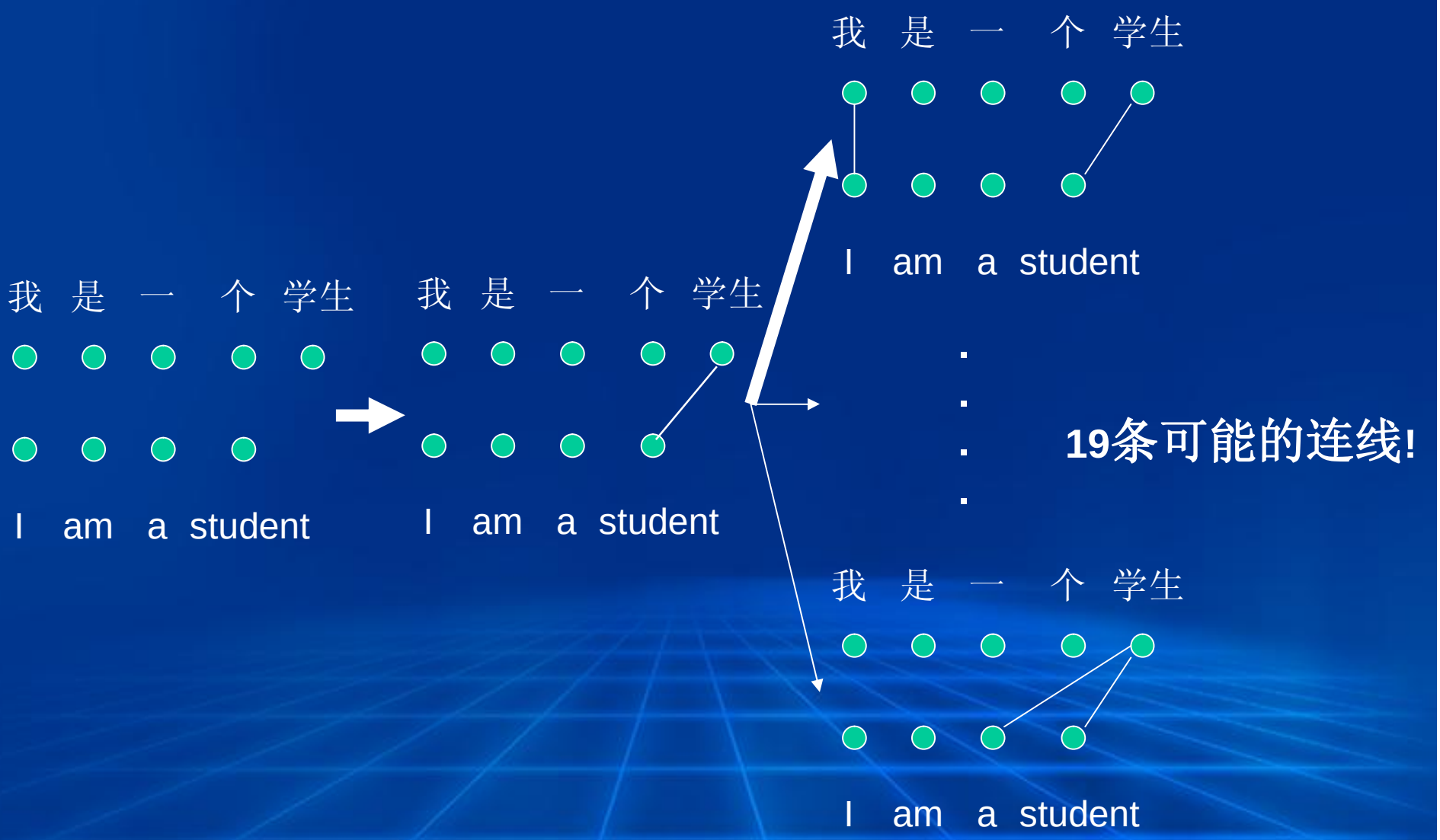


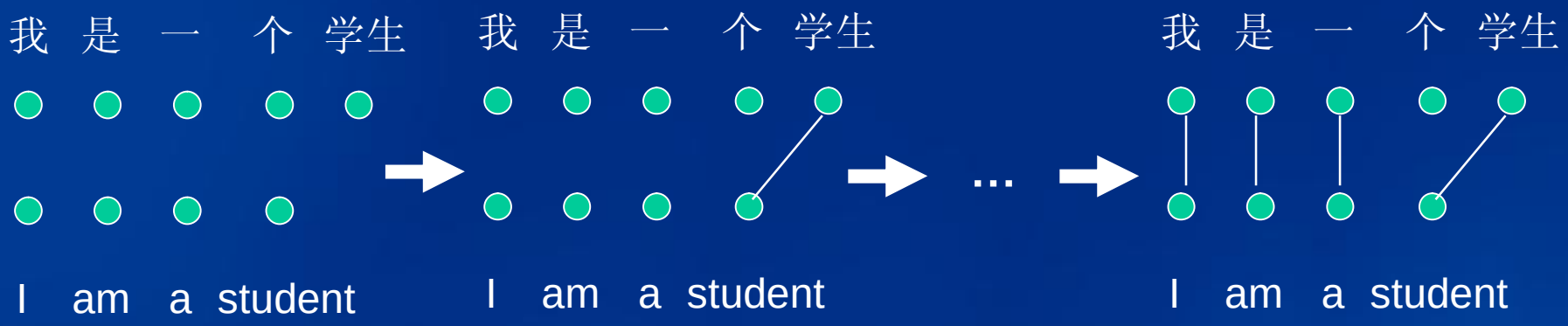
20 条可能的连线!



20 条可能的连线!







实验结果

任务	子任务	参评系统结果	我们的结果	排名
HLT/NAACL 2003 workshop on Parallel Texts	Romanian-English, non-null	0.2886-0.5267	0.2660	1/10
	Romanian-English, null	0.3741-0.5979	0.3234	1/10
	English-French, non-null	0.0853-0.2938	0.0633	1/9
	English-French, null	0.1850-0.5171	0.0633	1/9
ACL 2005 workshop on Parallel Texts	English-Inuktitut	0.0946-0.7127	0.1784	4/11
	Romanian-English	0.2655-0.4449	0.2614	1/34
	English-Hindi	0.5142	0.4764	1/2
HTRDP 2005	Chinese-English	0.2348-0.4918	0.1815	1/3

对比

特性	IBM模型	对数线性模型
可扩展性	差	好
语言无关性	支持	支持
利用具体语言特性	不支持	支持
需要手工优化参数	是	否
广泛应用于处理大规模数据	是	否

小结

- 论文提出了一种词语对齐的对数线性模型。该模型首次将判别方法引入词语对齐，具有良好的可扩展性。实验结果表明，对数线性模型在对齐质量上优于其它模型。

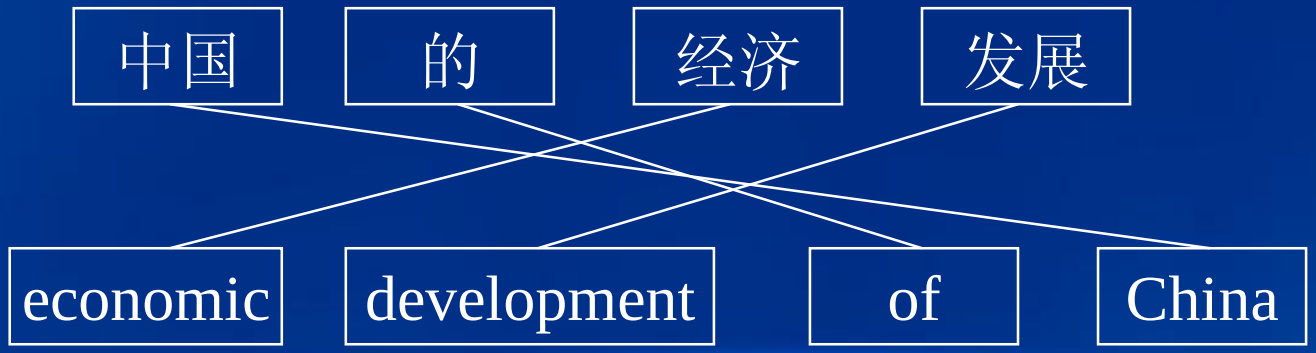
提纲

- *引言*
- *词语对齐的对数线性模型*
- 树到串统计翻译模型
  - 模型1
  - 模型2
  - 模型3
  - 实验
- 总结

基于短语的模型

- 基于短语的模型是目前统计机器翻译的主流
- 基本问题
  - 短语划分
  - 短语重排序
  - 短语翻译
- 短语重排序是基于短语的模型中最关键的部分

短语重排序



短语重排序方法

- 利用句法信息
  - Xia 2004
  - Collins 2005
- 不利用句法信息
  - Och 2002
  - Zens 2004
  - Tillmann 2005
  - Xiong 2006
  - Al-Onaizan 2006

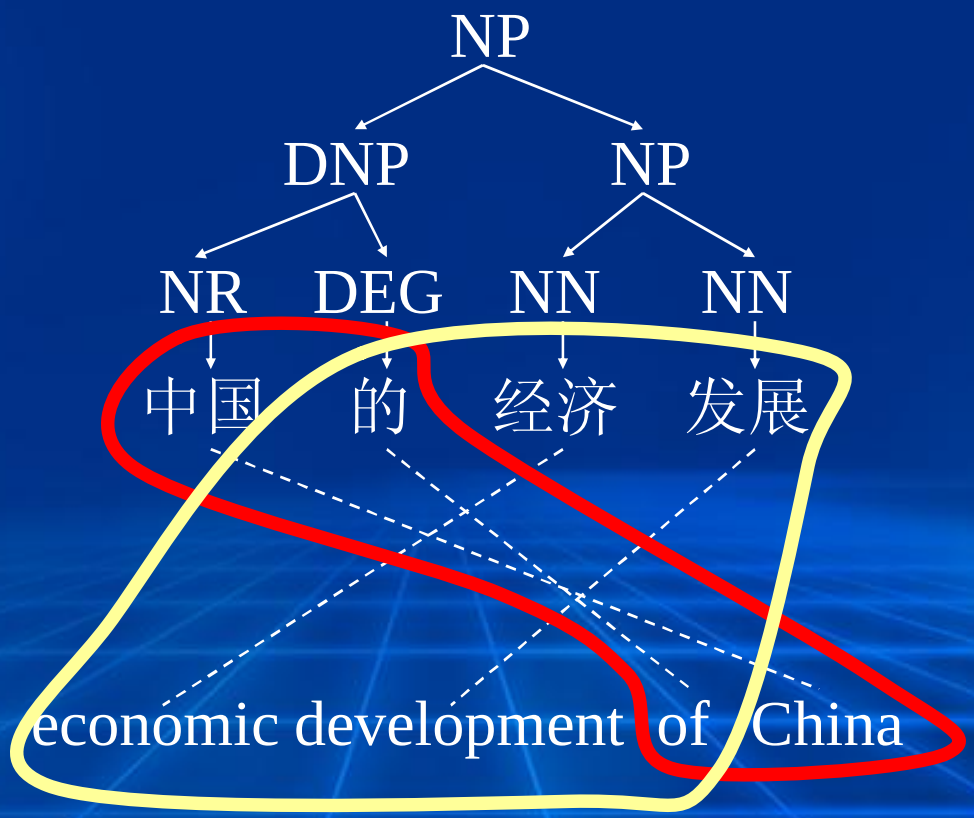
我的工作

- 提出了嵌入句法树的基于短语的翻译模型，该模型首次建模上利用句法信息指导短语重排序。

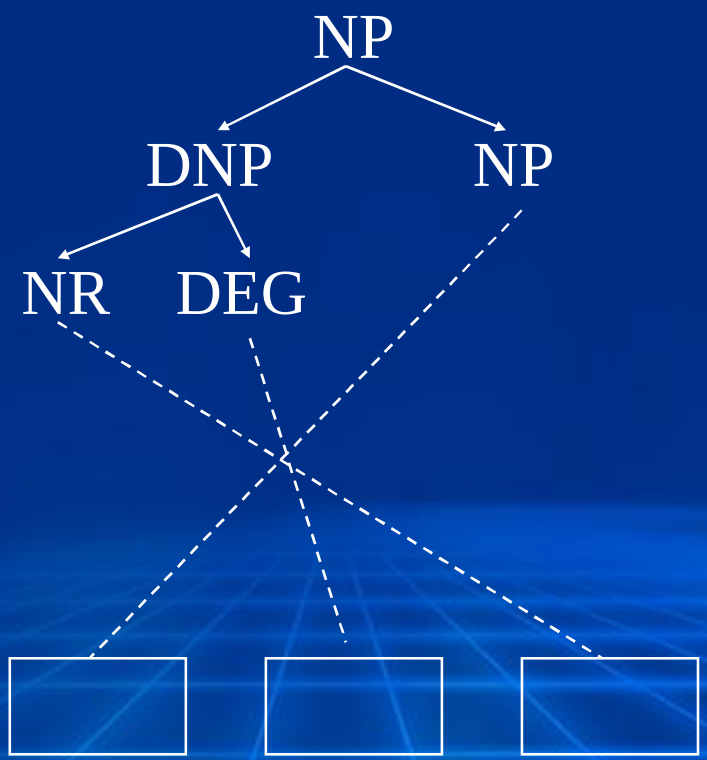
模型1

- 嵌入句法树的基于短语的翻译模型
- 只使用句法双语短语，利用树节点重排序（简称TNR）执行短语重排序
- 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TNR
- 自底向上的柱搜索算法

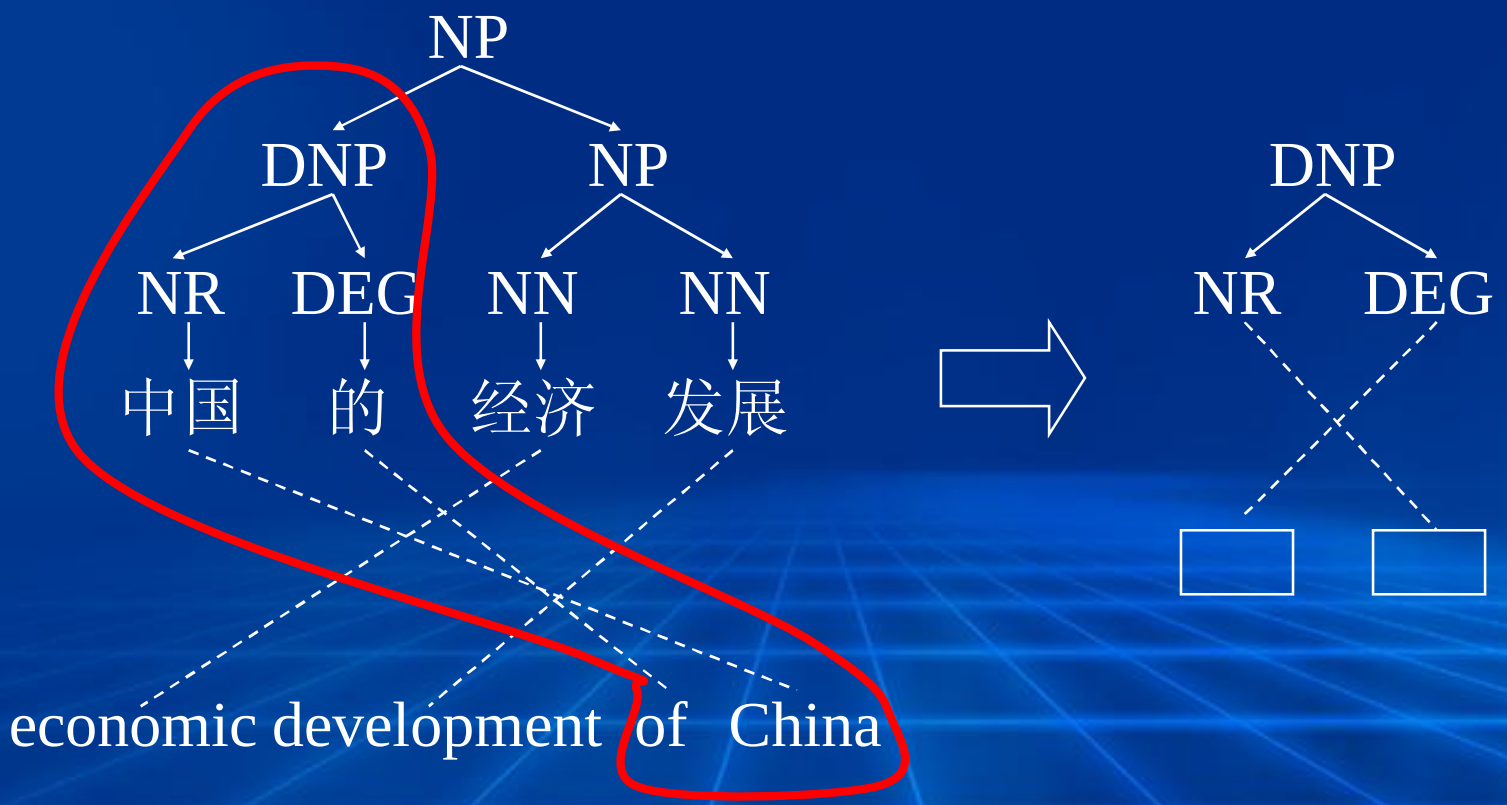
句法双语短语



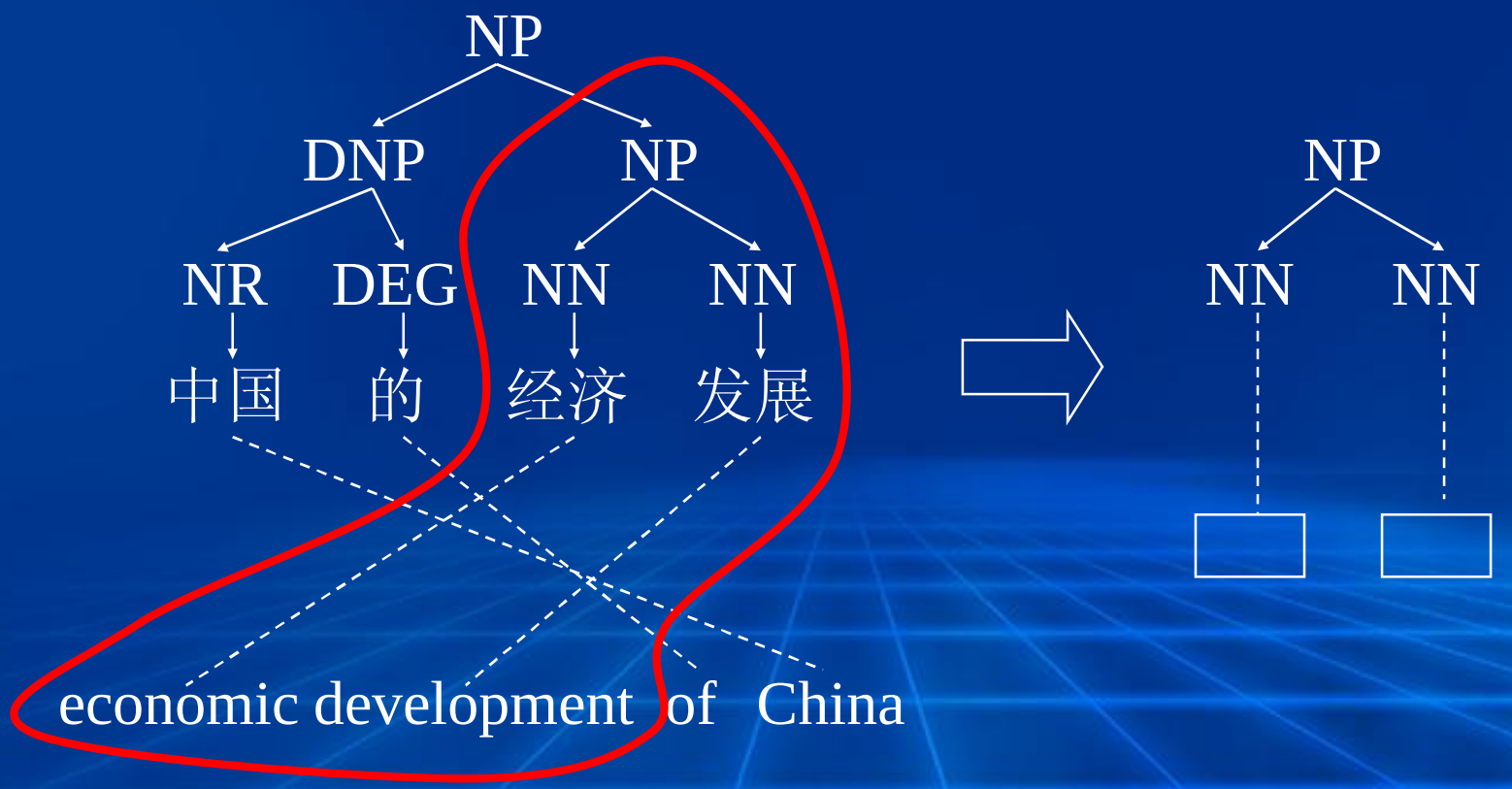
TNR



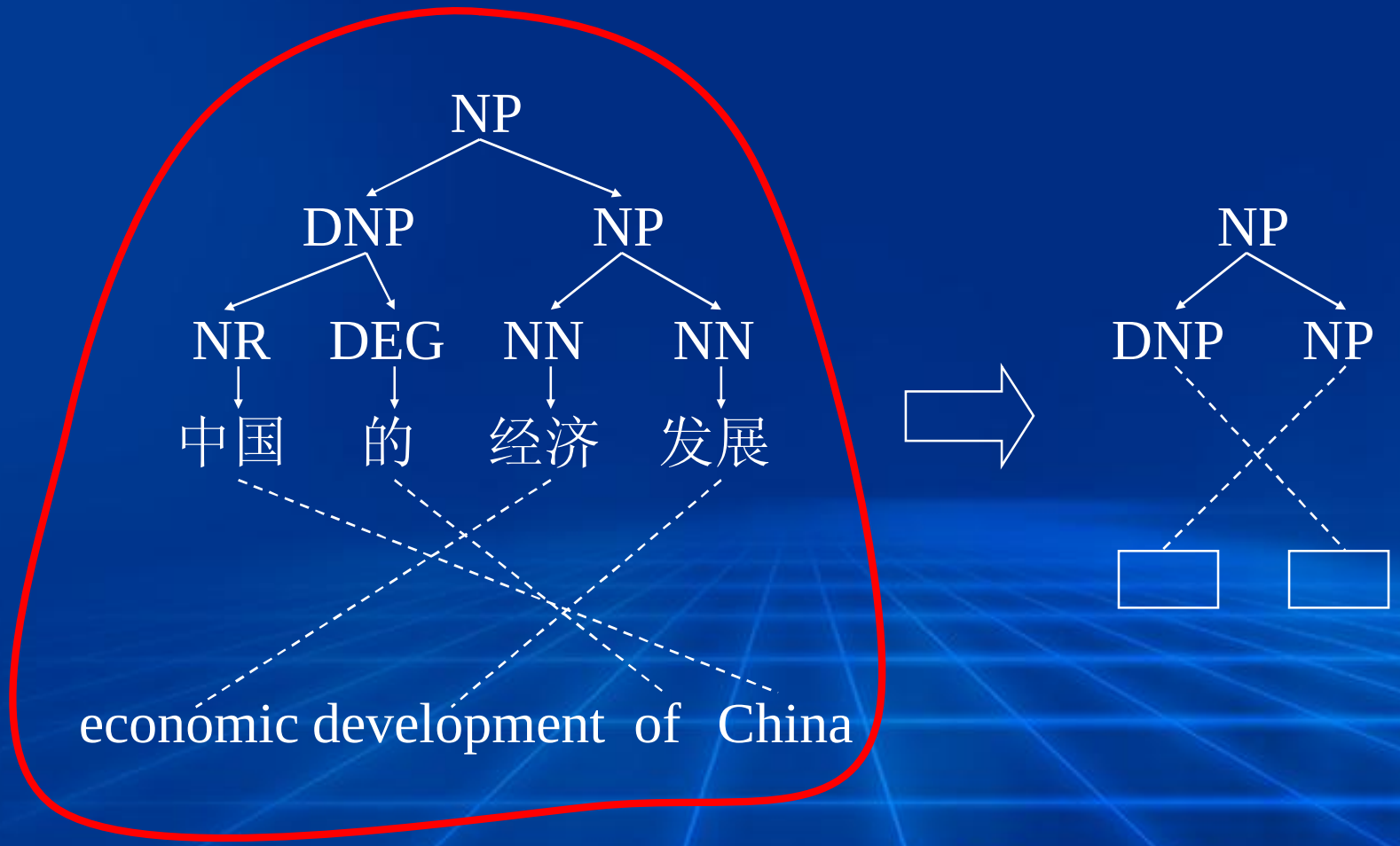
抽取TNR



抽取TNR



抽取TNR



搜索

BP

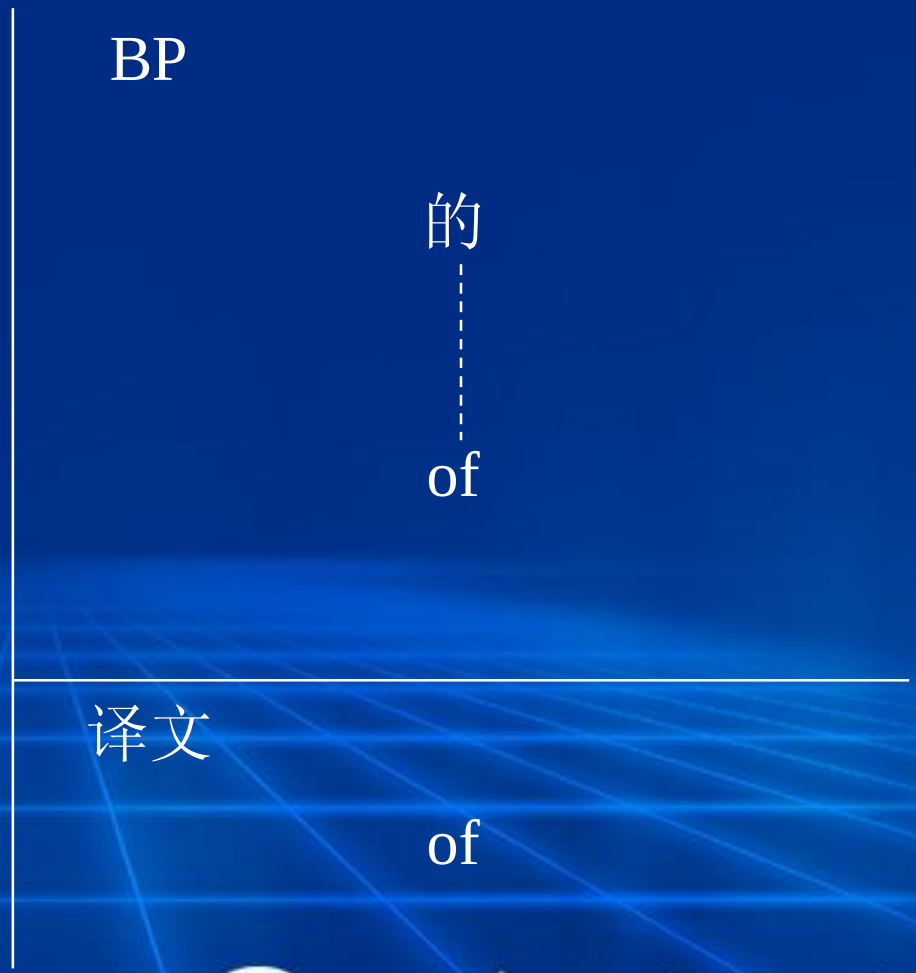
中国

China

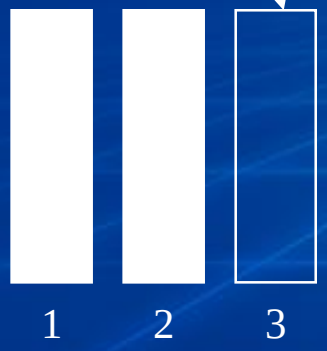
译文

China

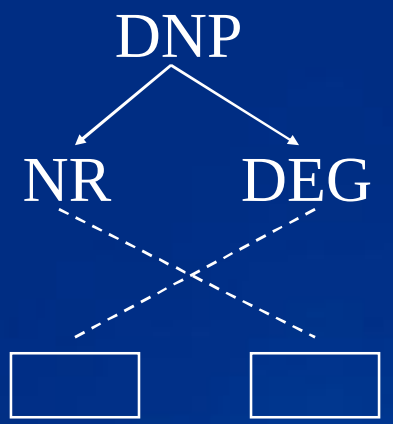
搜索



搜索



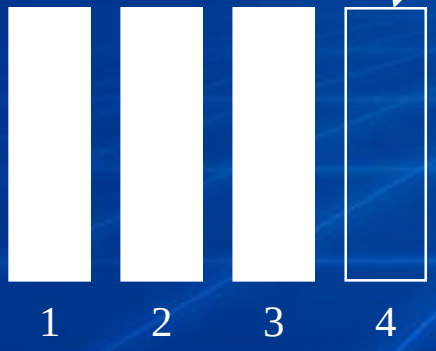
TNR



译文

of China

搜索



BP

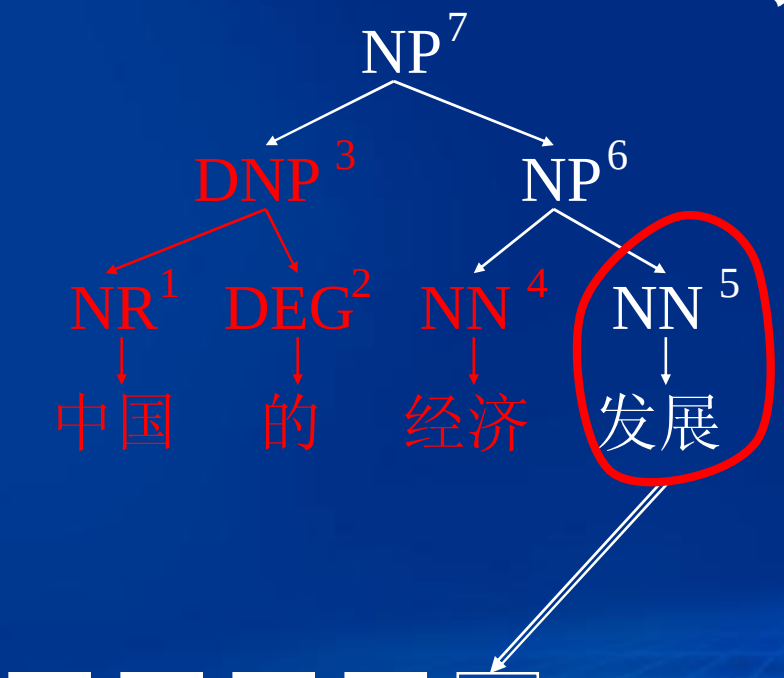
经济

economy

译文

economy

搜索



BP

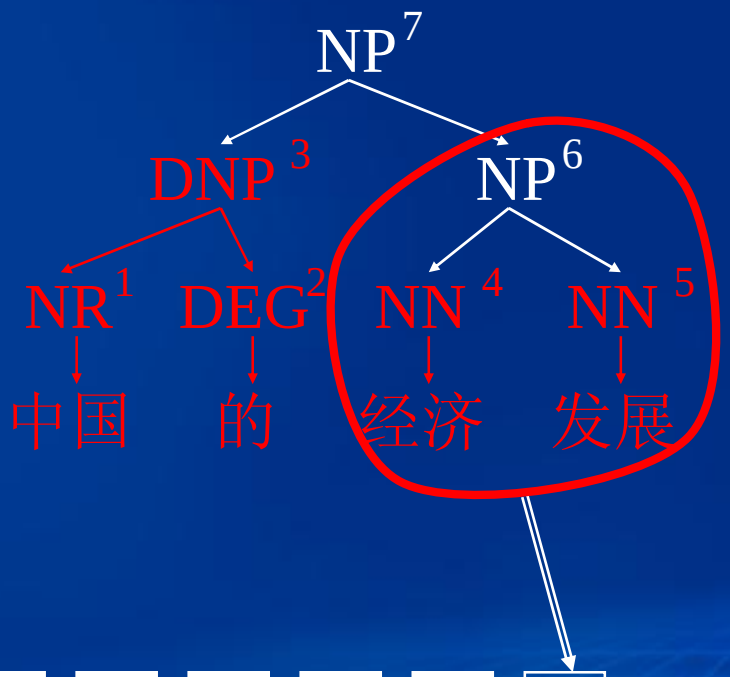
发展

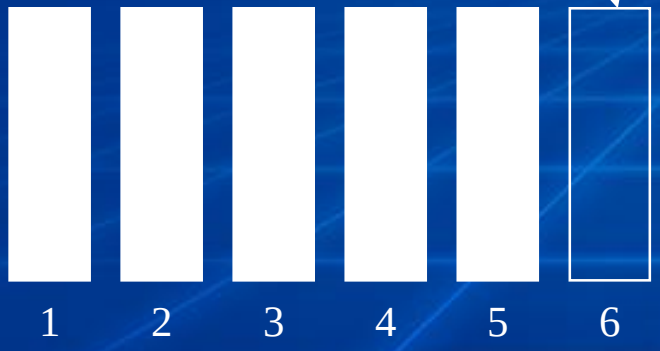
development

译文

development

搜索





BP

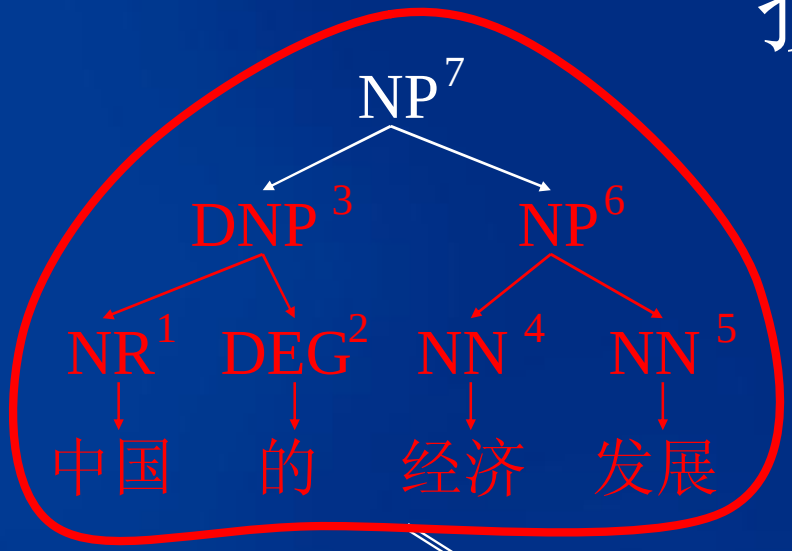
经济 发展

economic development

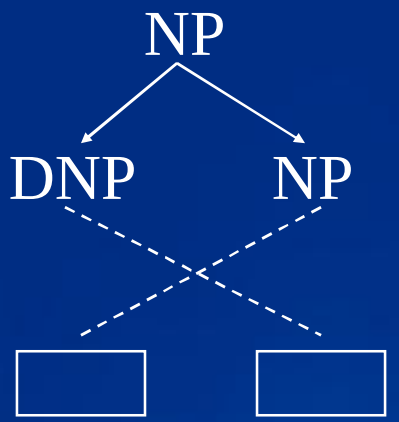
译文

economic development

搜索



TNR



译文

economic development of China

小结

- 论文提出了嵌入句法树的基于短语的翻译模型，该模型首次建模上利用句法信息指导短语重排序。

提纲

- *引言*
- *词语对齐的对数线性模型*
- 树到串统计翻译模型
  - *模型1*
  - 模型2
  - 模型3
  - 实验
- 总结

基于句法的方法

- 基于句法的方法
  - 形式化基于句法
    - SITG
      - [Wu 1997]
    - SCFG
      - [Chiang 2005]
  - 语言学基于句法
    - 串到树
      - [Yamada 2001]
    - 树到树
      - [Ding 2005]
- 目前大多数基于句法的方法没有在实际评测中明显超过基于短语的方法，原因可能在于：
  - 复杂度过高
  - 难以处理非同构性问题

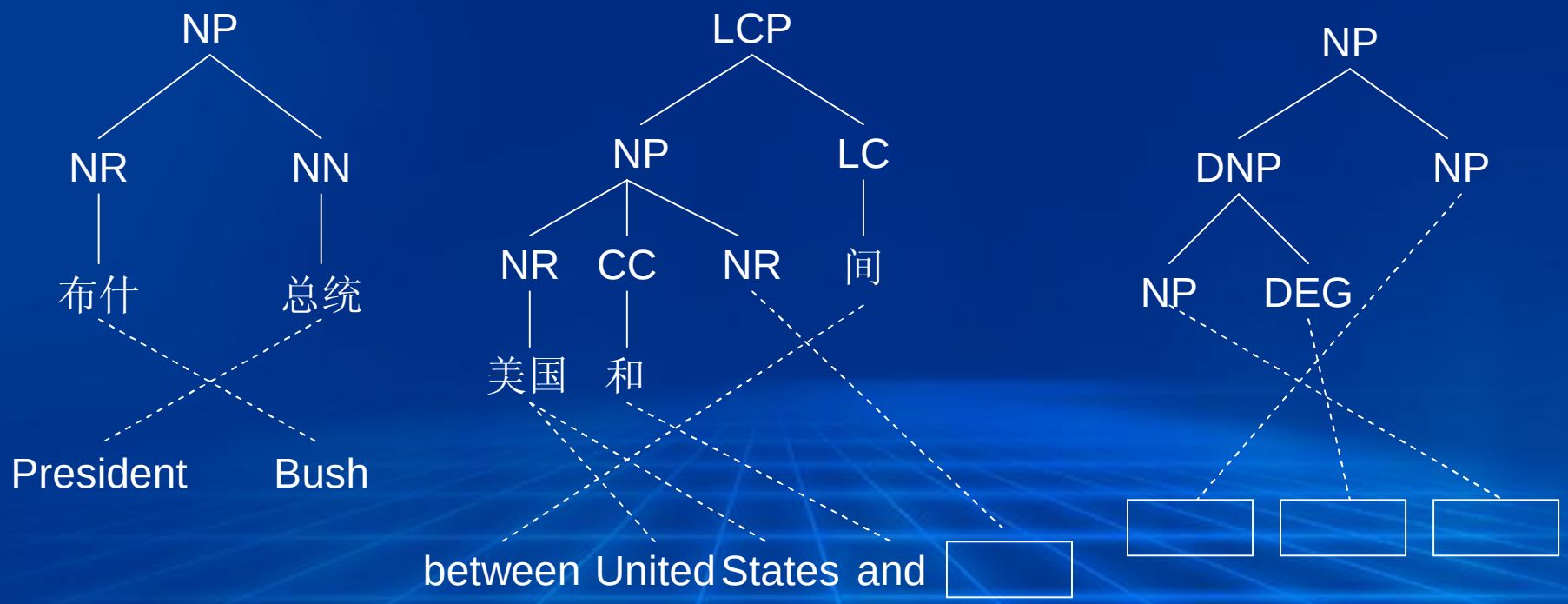
我的工作

- 提出了基于树到串对齐模板的翻译模型。该模型复杂度低，具备很强的重排序能力。

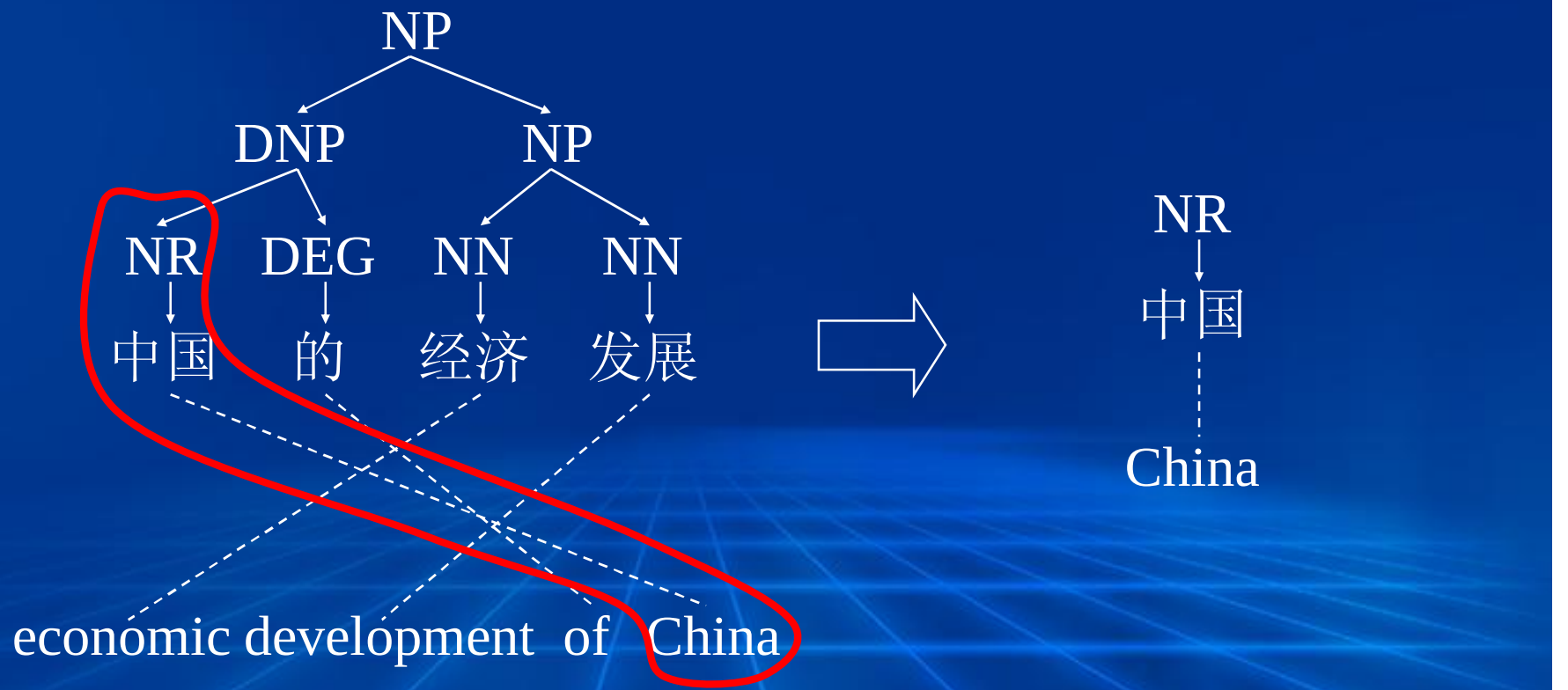
模型2

- 基于树到串对齐模板的翻译模型
- 树到串对齐模板（简称TAT）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

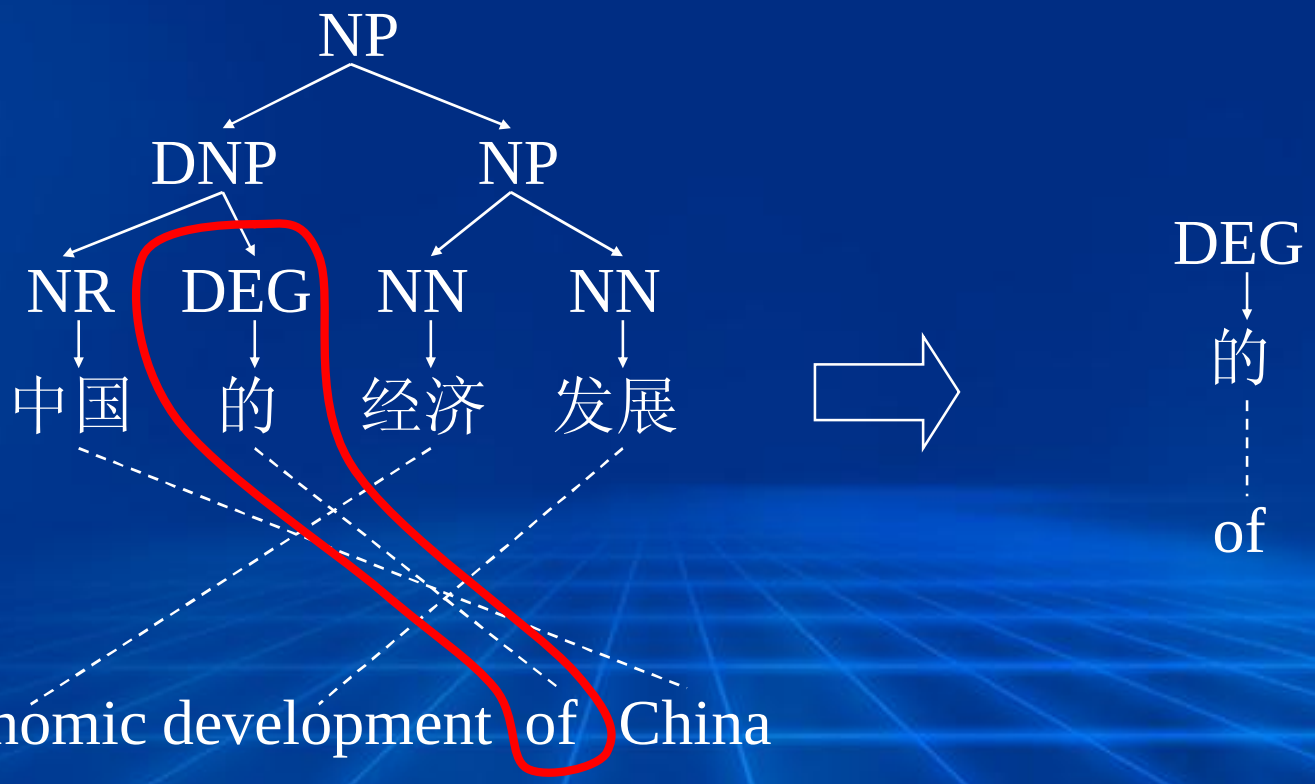
树到串对齐模板



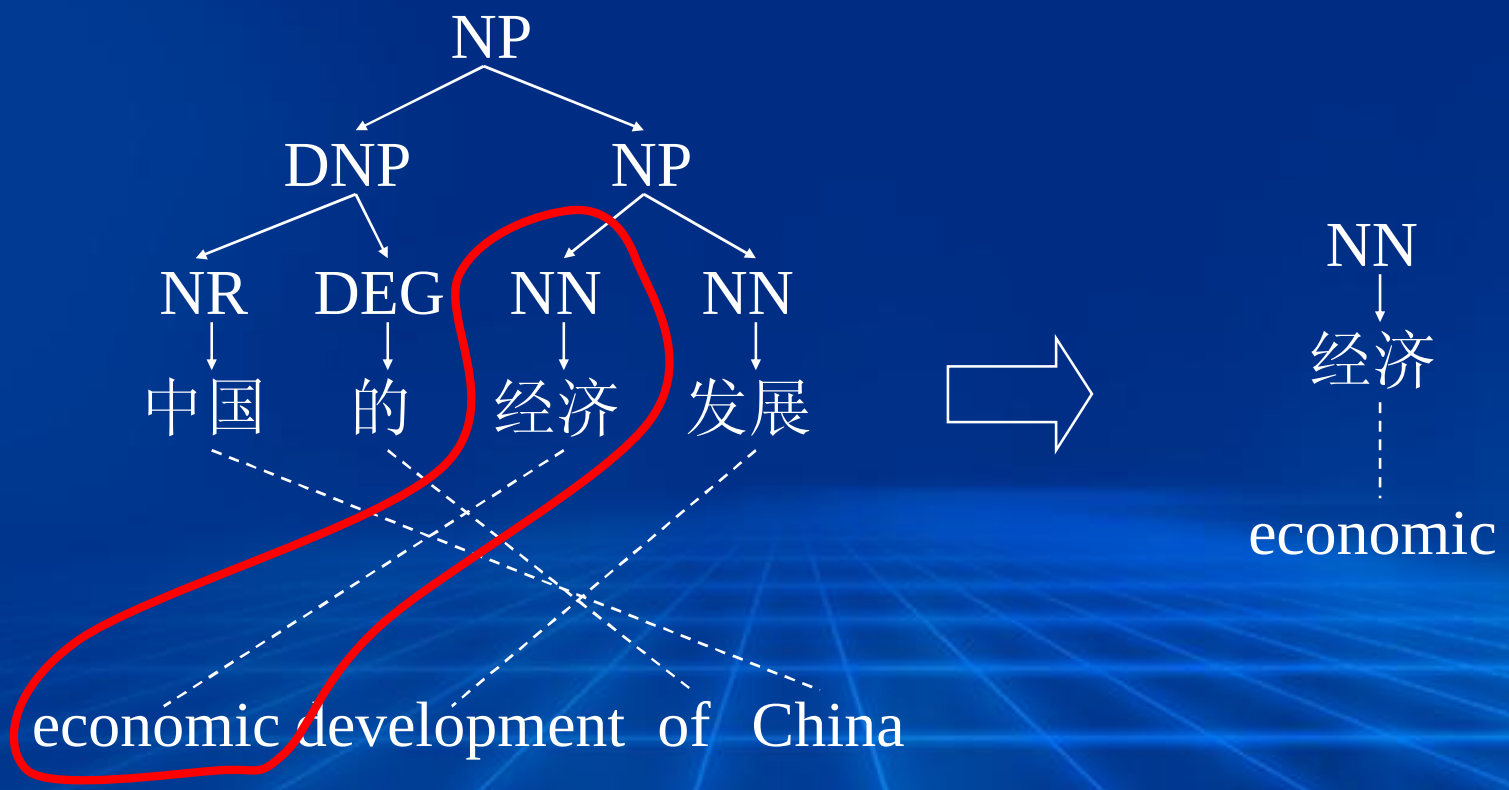
抽TAT



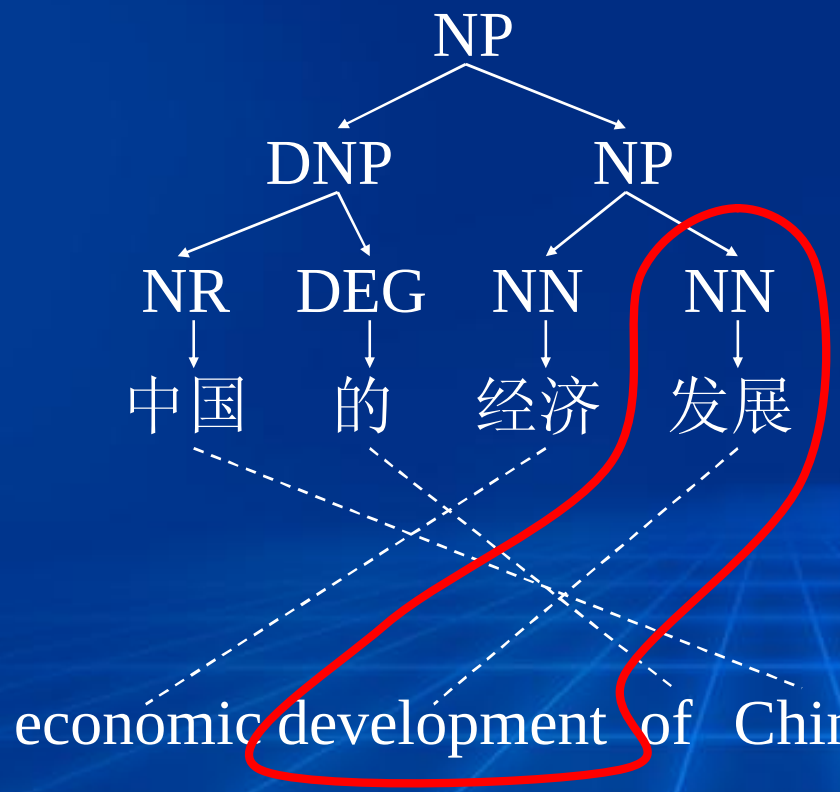
抽TAT

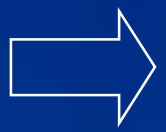


抽TAT



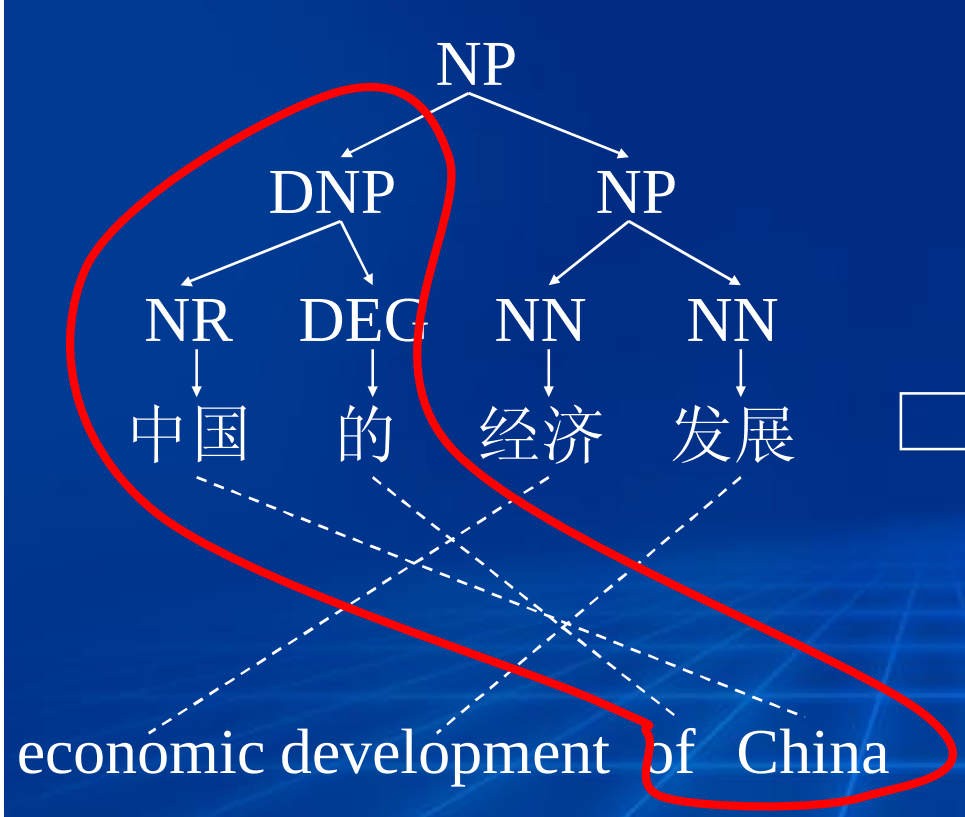
抽TAT

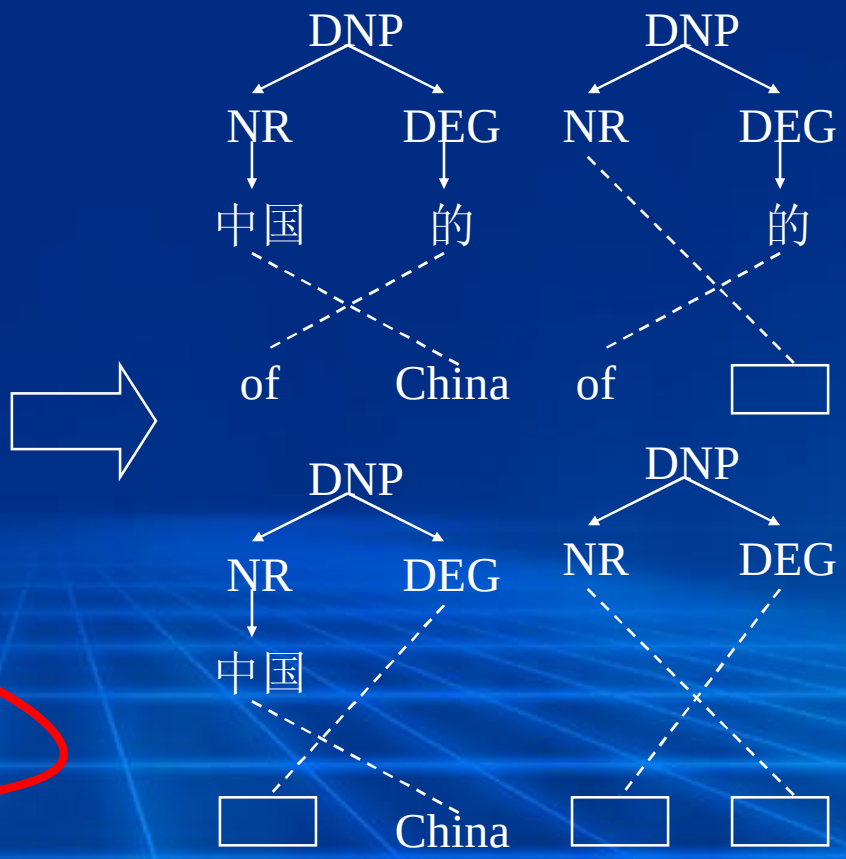




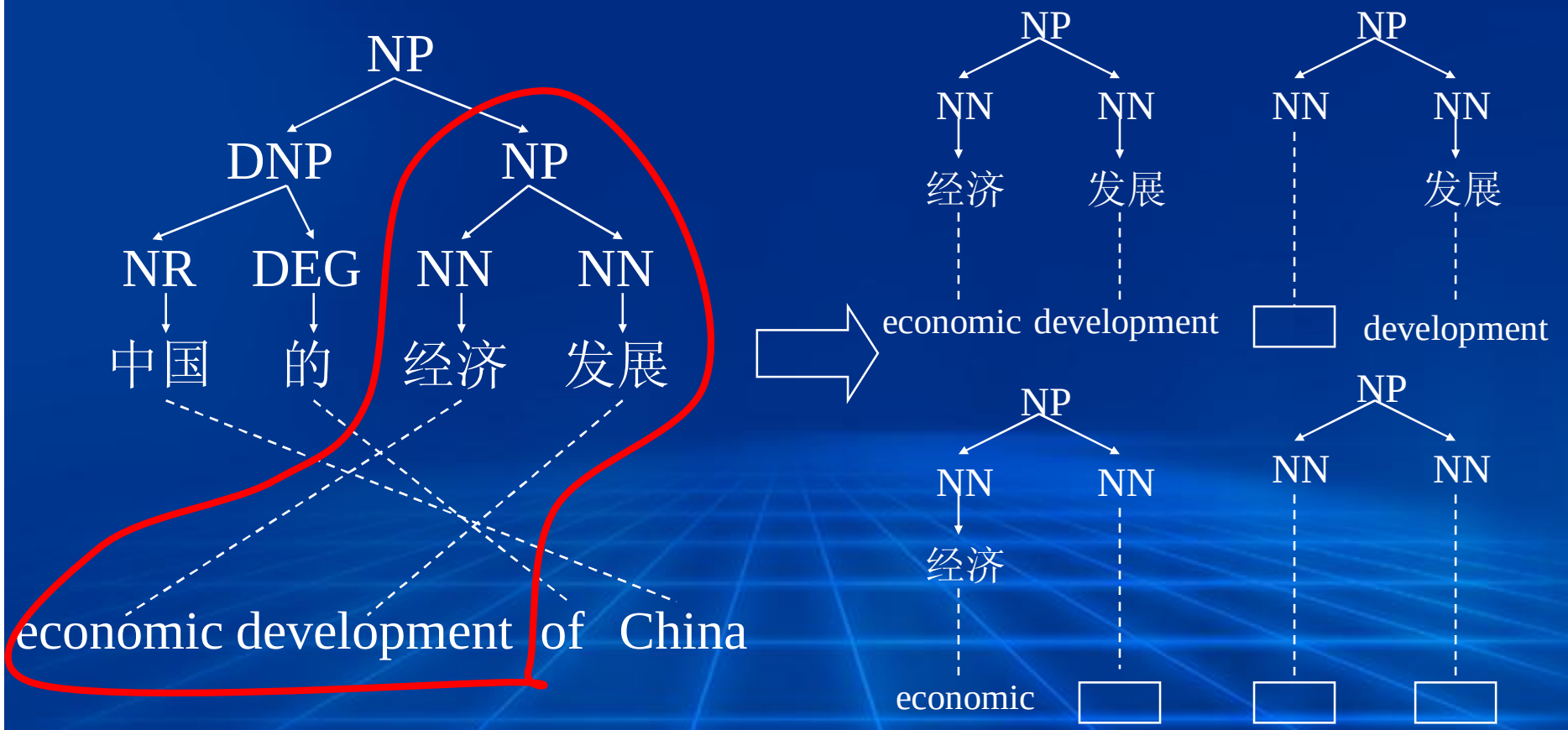
economic development of China

抽TAT

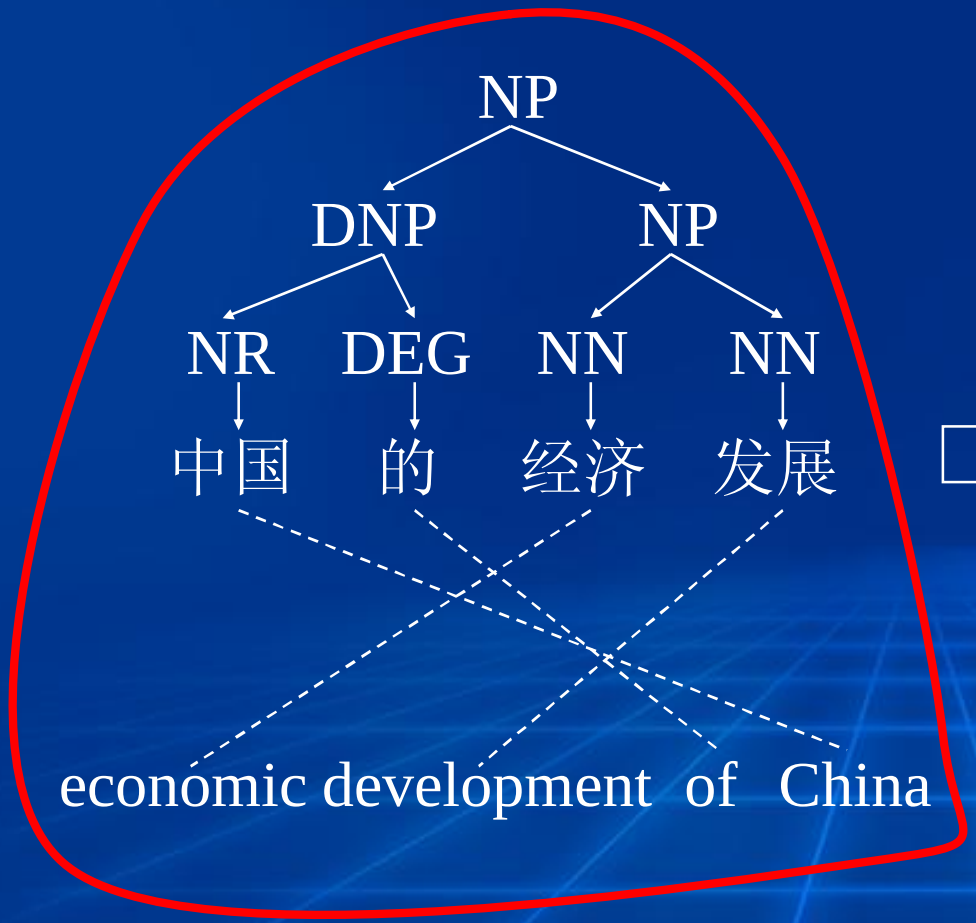


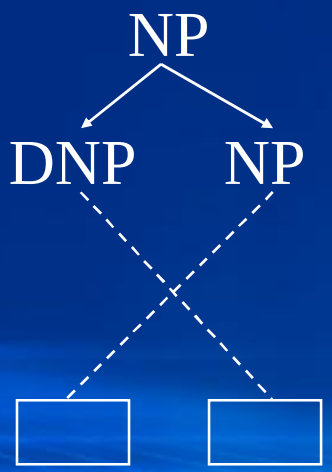


抽TAT



抽TAT



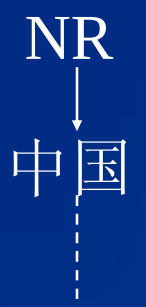


h=2, c=2

搜索

1

TAT



China

译文

China

搜索

TAT

DEG

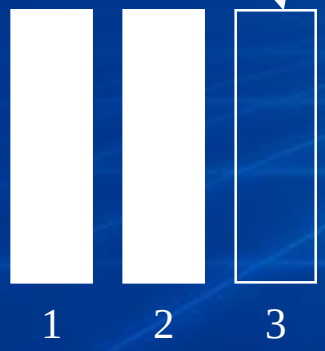
的

of

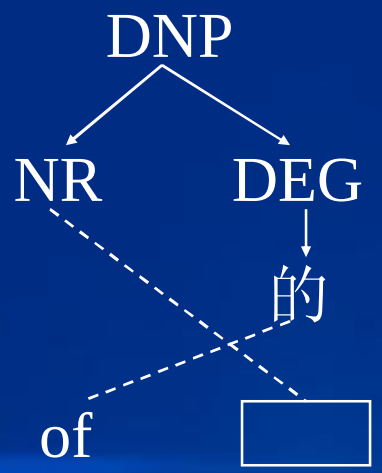
译文

of

搜索



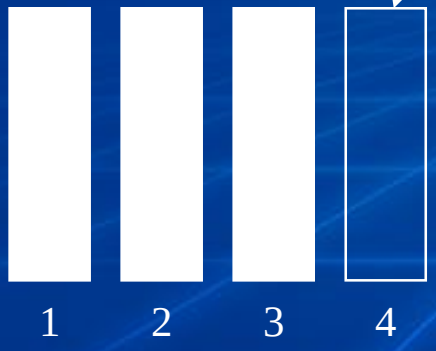
TAT



译文

of China

搜索



TAT

NN

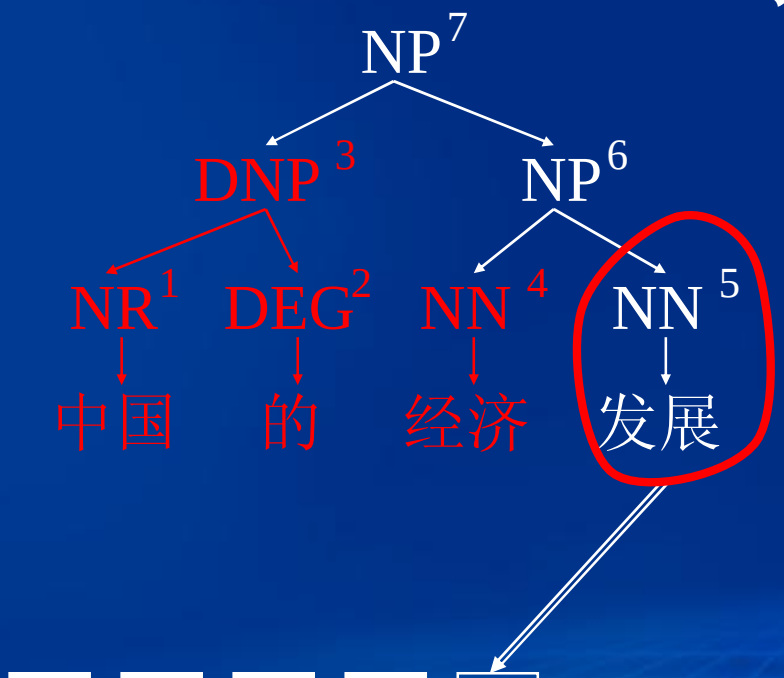
经济

economy

译文

economy

搜索



TAT

NN

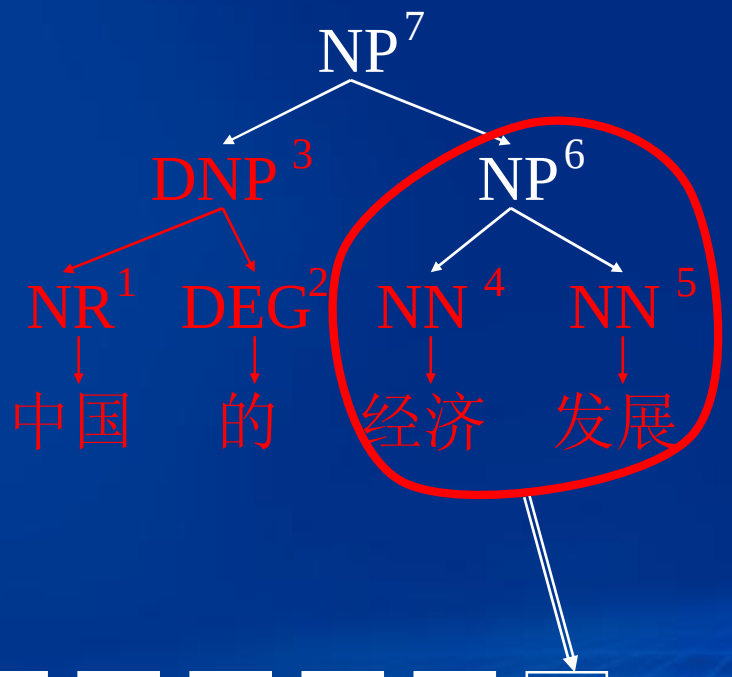
发展

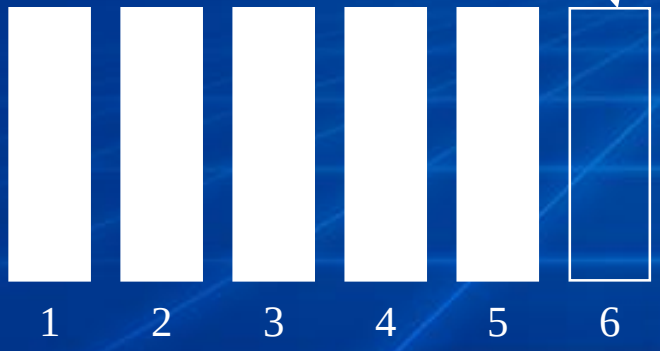
development

译文

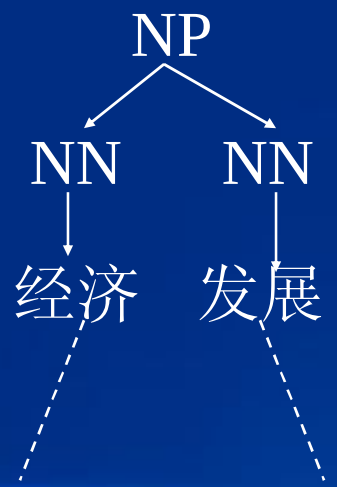
development

搜索





TAT

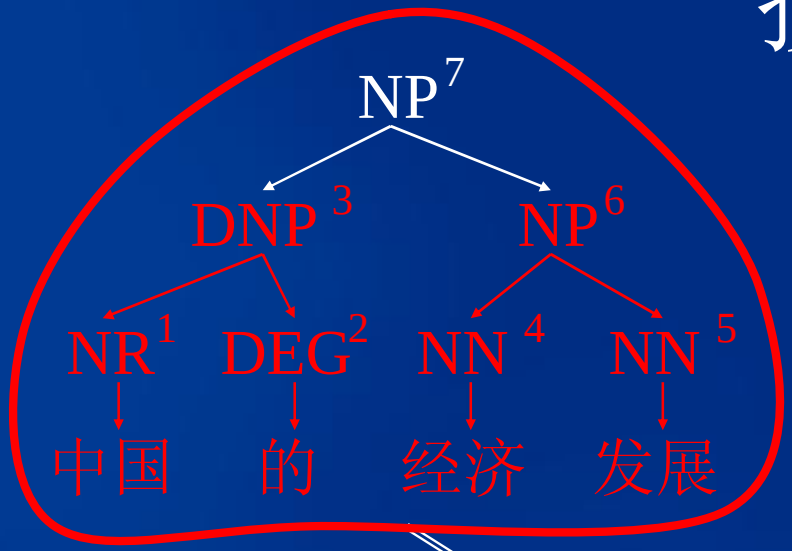


economic development

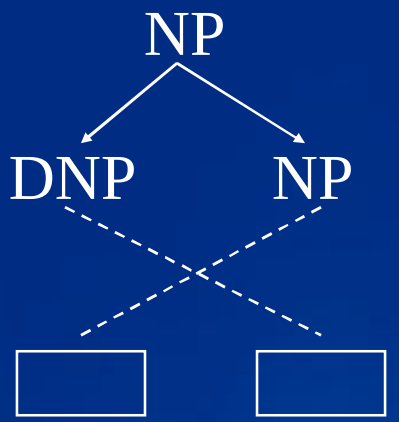
译文

economic development

搜索



TAT



译文

economic development of China

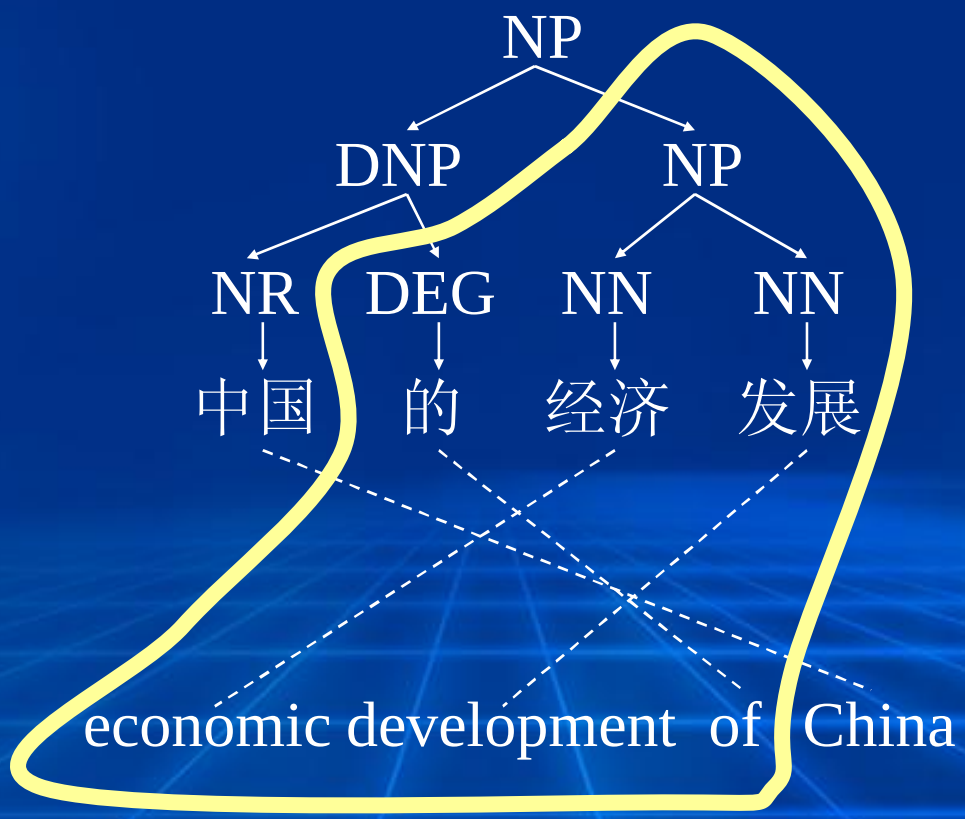
小结

- 论文提出了基于树到串对齐模板的翻译模型。该模型复杂度低，具备很强的重排序能力。

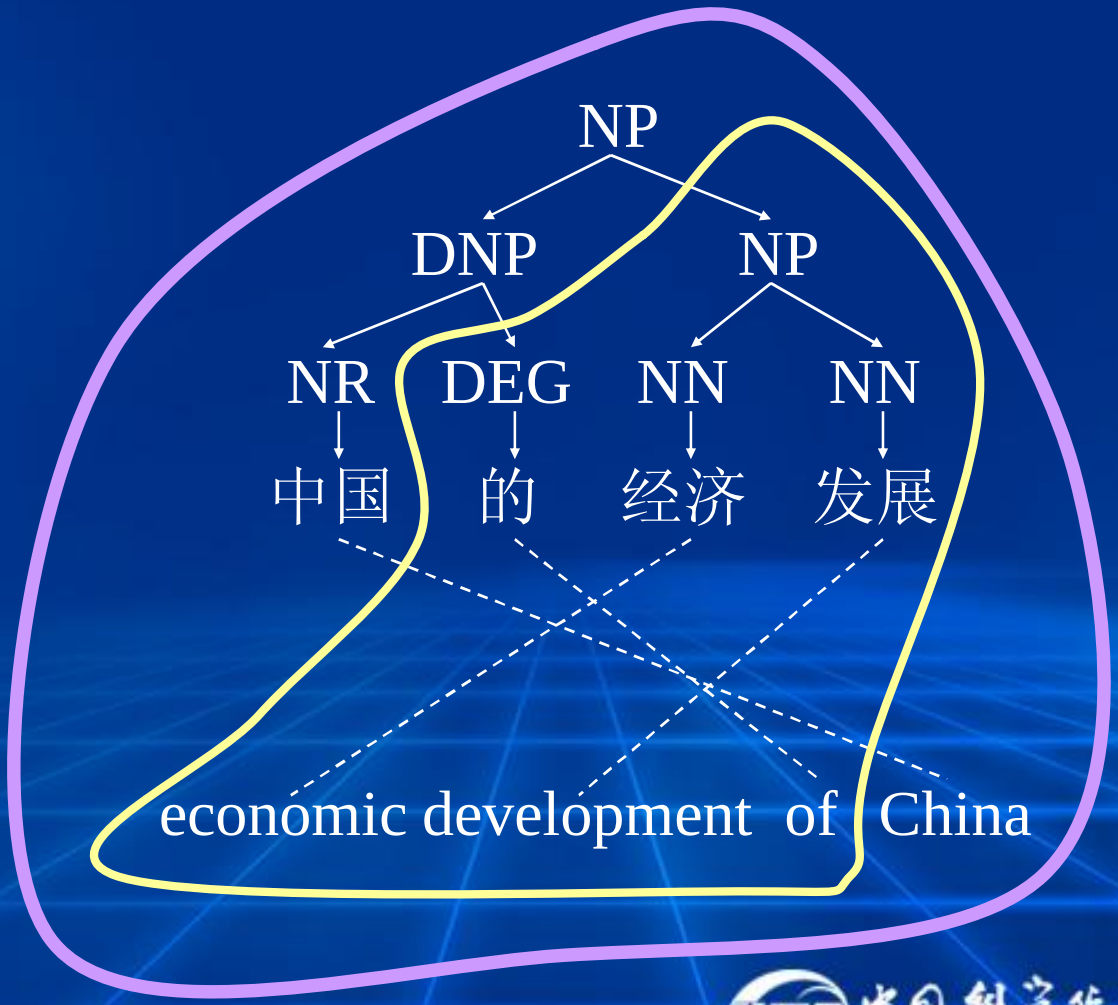
提纲

- *引言*
- *词语对齐的对数线性模型*
- 树到串统计翻译模型
  - *模型1*
  - *模型2*
  - 模型3
  - 实验
- 总结

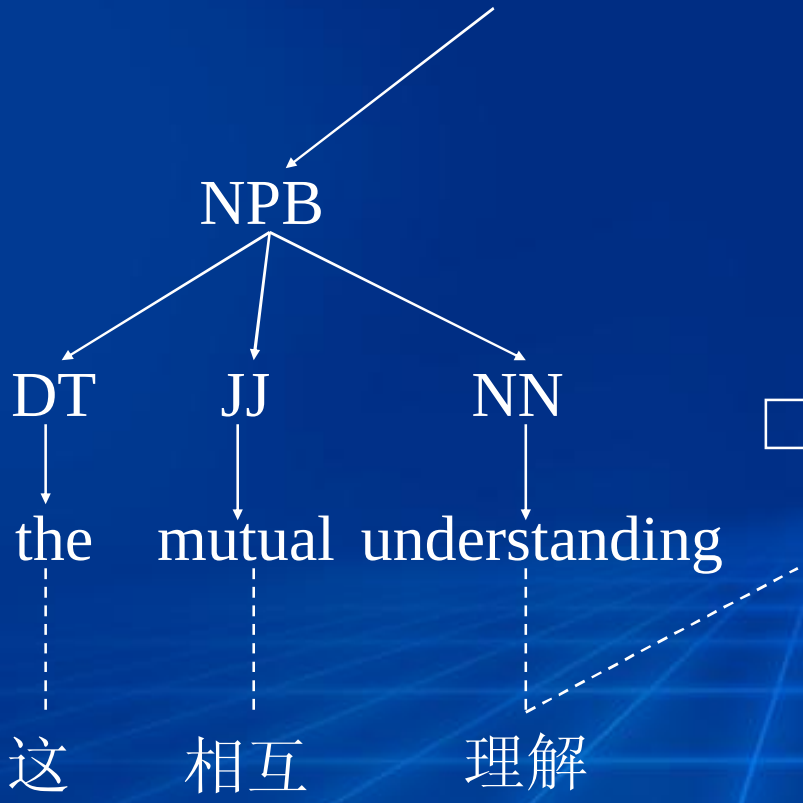
非句法双语短语

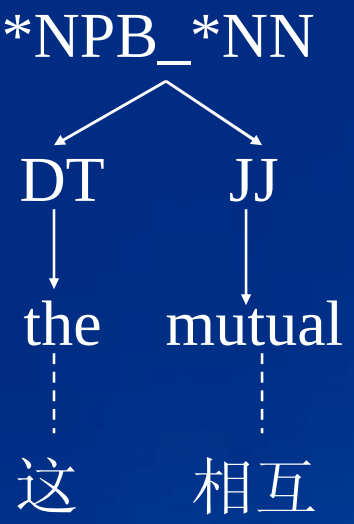


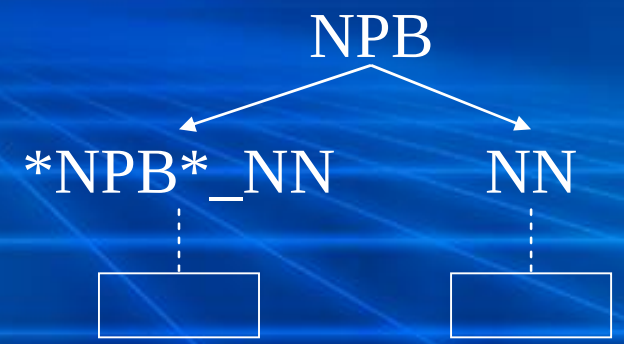
Galley 2004



Marcu 2006







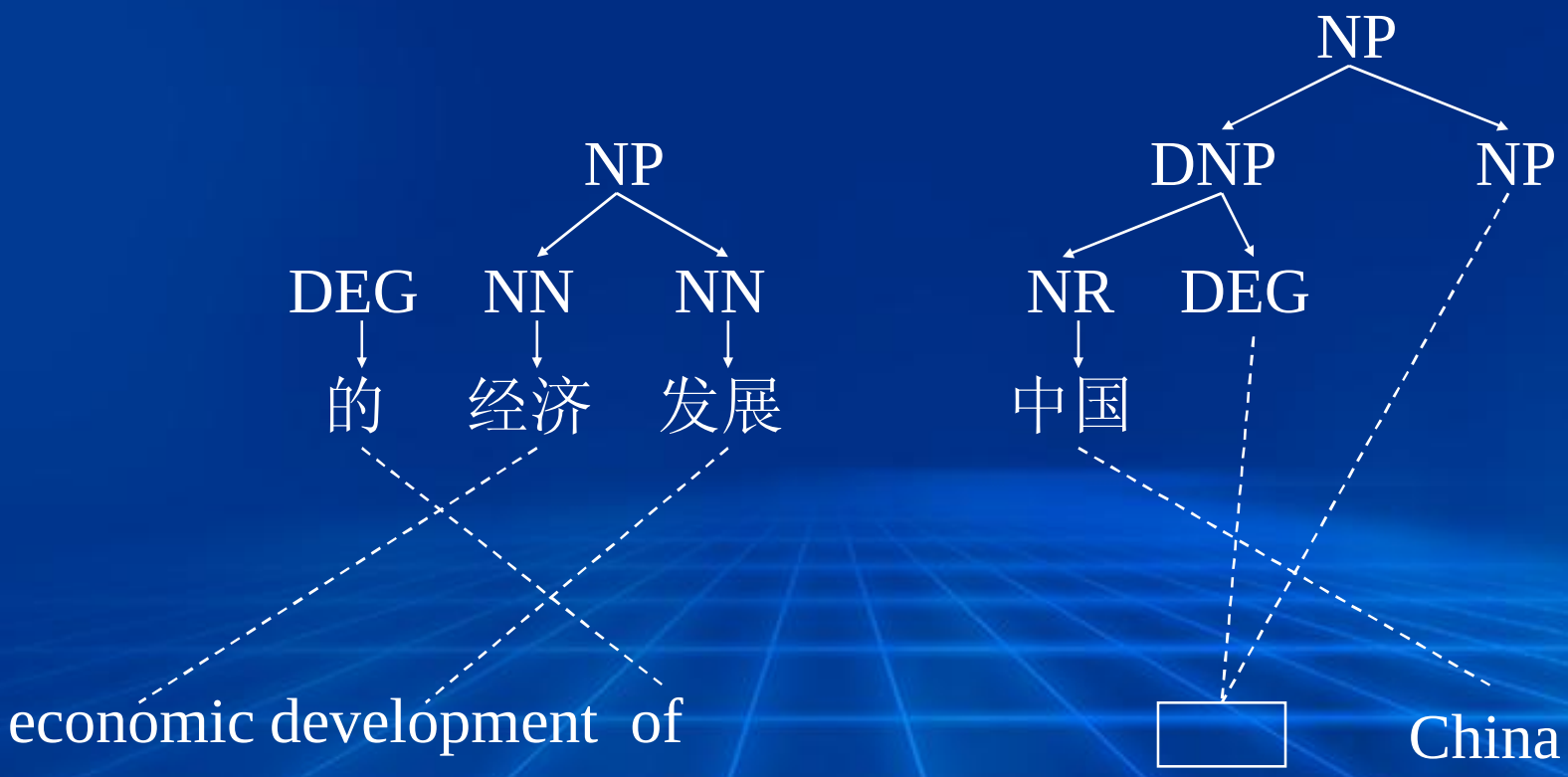
我的工作

- 提出了融入森林到串规则的树到串翻译模型，该模型为短语兼容性问题提供了良好的解决方案，极大提高了树到串翻译模型的表达能力。

模型3

- 融入森林到串规则的树到串翻译模型
- 在模型2的基础上，模型3引入两类新规则：
  - 森林到串规则：表达和泛化非句法双语短语
  - 辅助规则：将森林到串规则融入到树到串模型
- 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树到串规则和森林到串规则
- 解码时动态构造辅助规则
- 自底向上的柱搜索算法

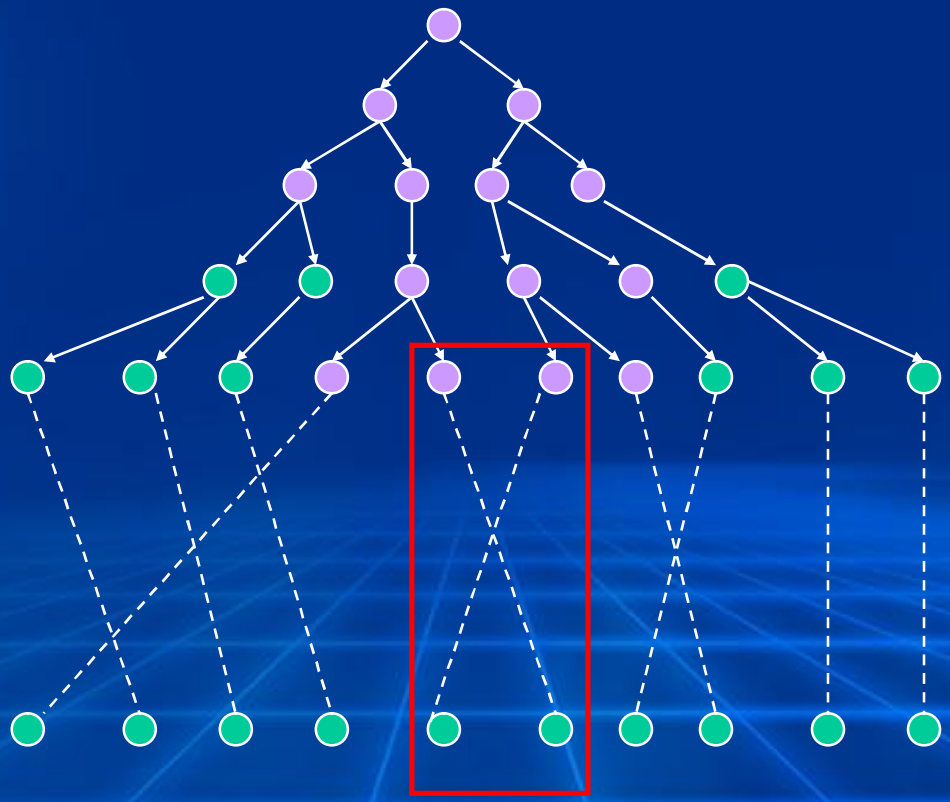
森林规则和辅助规则



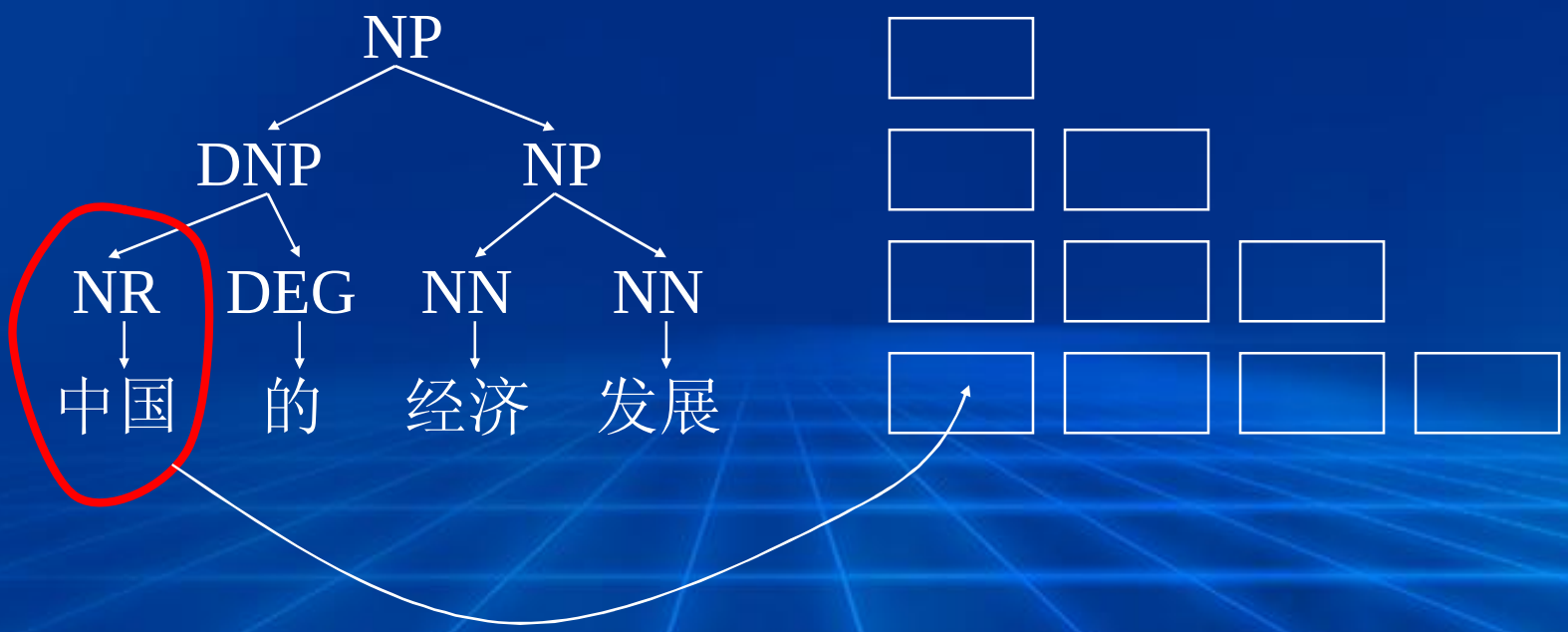
抽取算法

Input: a source tree $T = T(f_1^J)$, a target string $S = e_1^I$, and word alignment A between them
$\mathcal{R} := \emptyset$
for $u := 0$ to $J - 1$ **do**
  for $v := 1$ to $J - u$ **do**
    identify the triple set $\mathcal{T}$ corresponding to span $(v, v + u)$
    for each triple $t = \langle T', S', A' \rangle \in \mathcal{T}$ **do**
      if $\langle T', S' \rangle$ is not consistent with A **then**
        continue
      end if
      if $u = 0 \wedge node(T') = 1$ **then**
        add t to $\mathcal{R}$
        add $\langle root(T'), \text{“}X\text{”}, 1{:}1 \rangle$ to $\mathcal{R}$
      else
        compute the skeleton s of the triple t
        register rules that are built on s using rules extracted from the sub-triples of t: $\mathcal{R} := \mathcal{R} \cup build(s, \mathcal{R})$
      end if
    end for
  end for
end for
Output: rule set $\mathcal{R}$

为什么不抽取辅助规则？

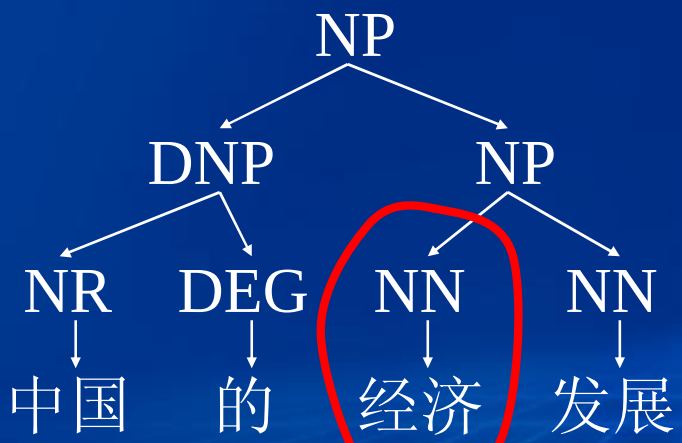


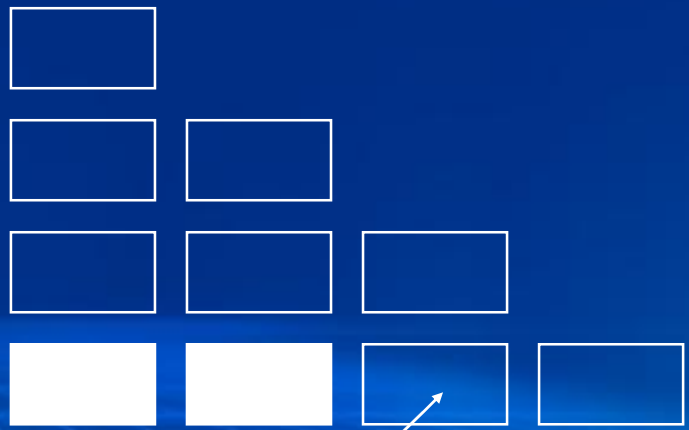
解码



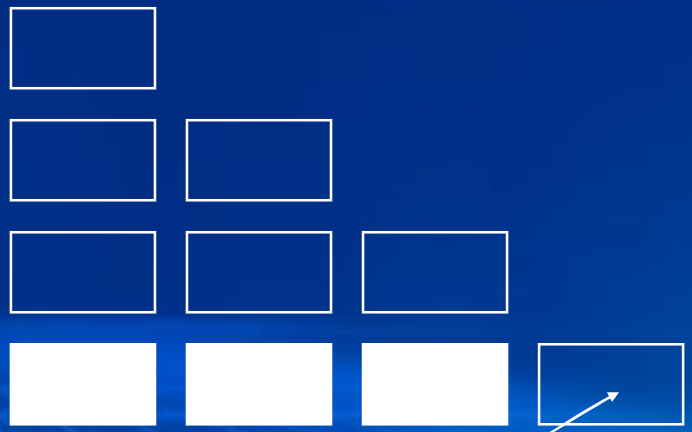
解码

解码

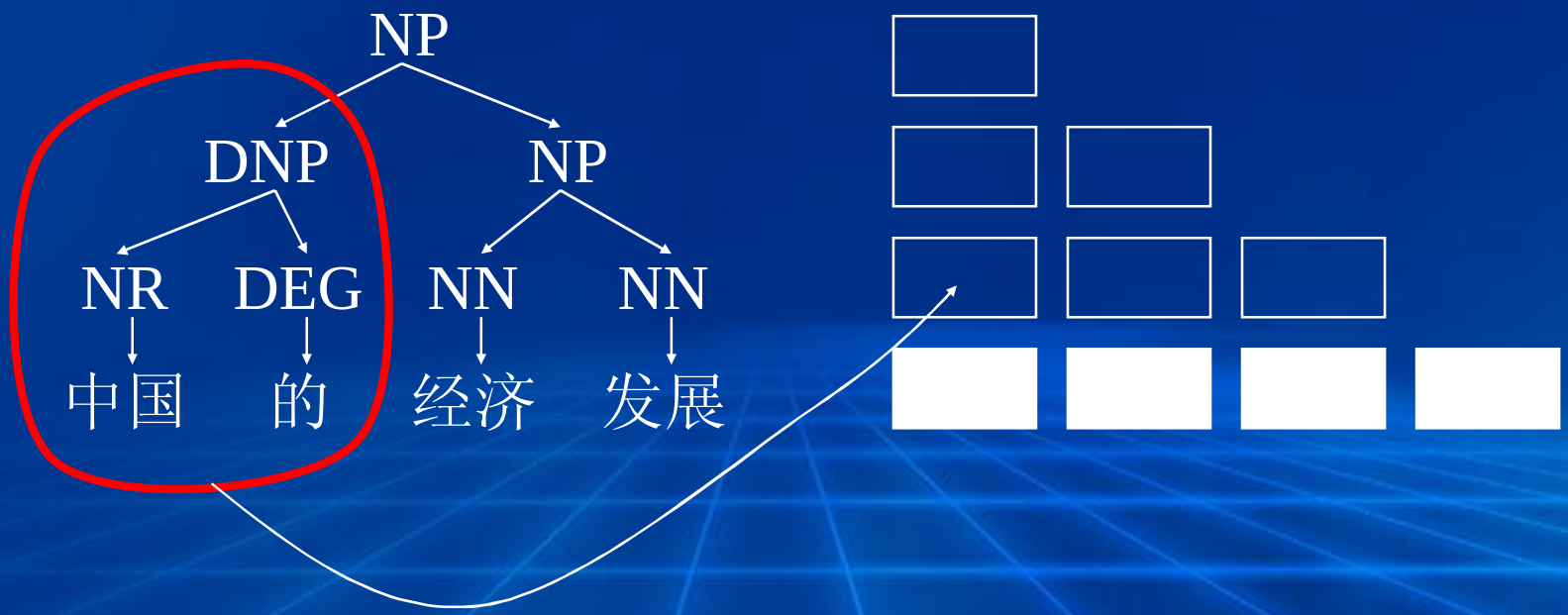




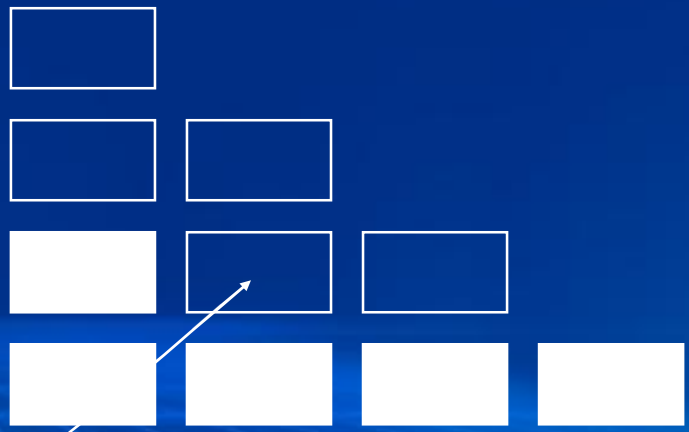
解码



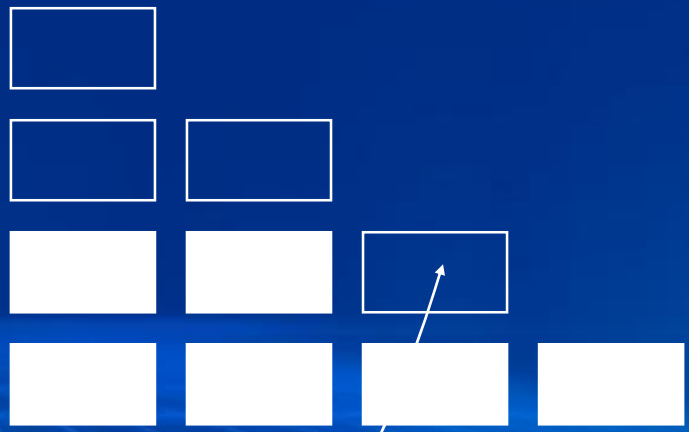
解码



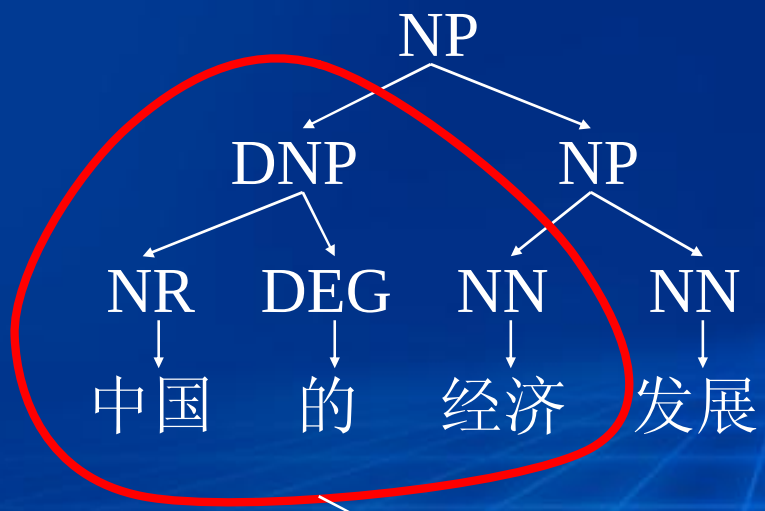
解码

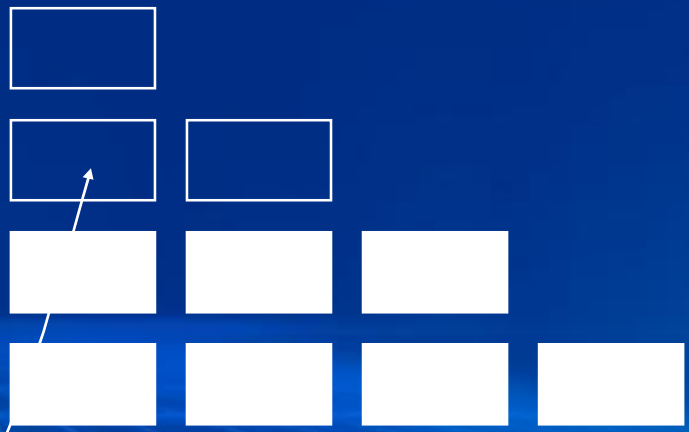


解码

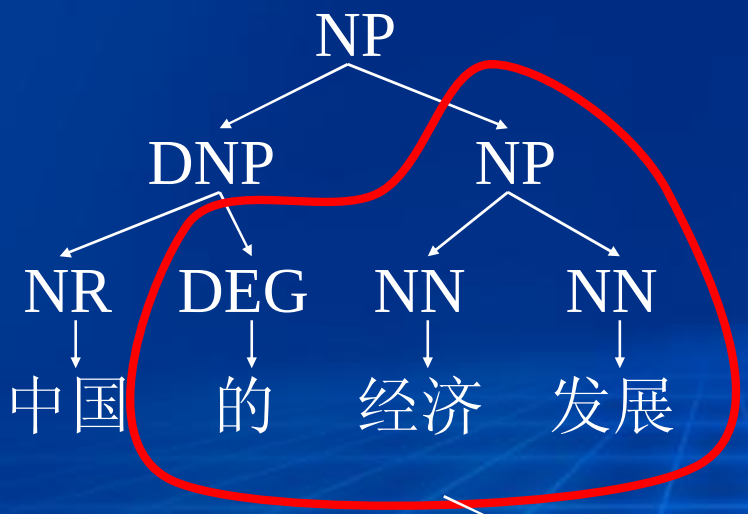


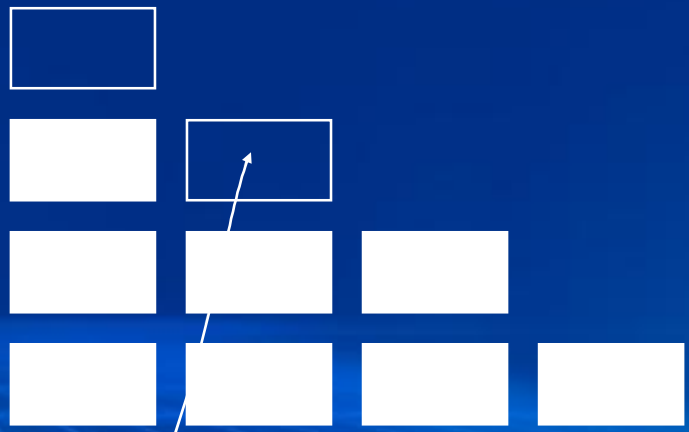
解码



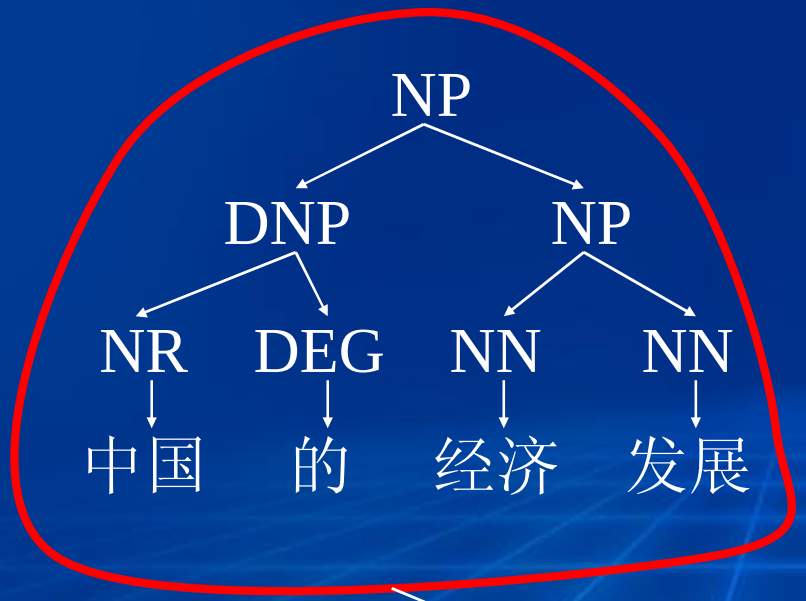


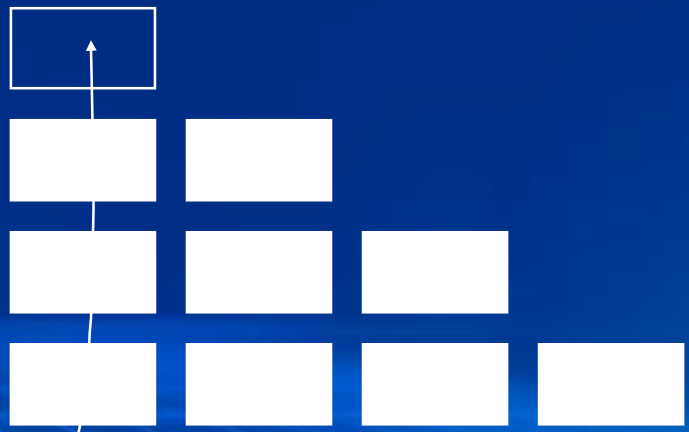
解码





解码

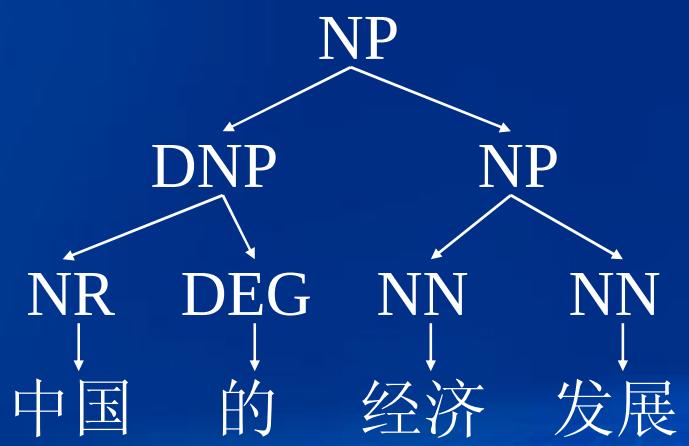




解码算法

Input: a source parse tree $T = T(f_1^J)$
for $u := 0$ to $J - 1$ **do**
  for $v := 1$ to $J - u$ **do**
    for each T' spanning from v to $v + u$ **do**
      if T' is a tree **then**
        for each usable tree-to-string rule r **do**
          for each derivation θ inferred from r and derivations in *matrix* **do**
            add θ to $matrix[v, v + u, root(T')]$
          end for
        end for
        search subcell divisions $\mathcal{D}[v, v + u]$
        for each subcell division $d \in \mathcal{D}[v, v + u]$ **do**
          if d contains at least one forest cell **then**
            construct auxiliary rule r_a
            for each derivation θ inferred from r_a and derivations in *matrix* **do**
              add θ to $matrix[v, v + u, root(T')]$
            end for
          end if
        end for
      else
        for each usable forest-to-string rule r **do**
          for each derivation θ inferred from r and derivations in *matrix* **do**
            add θ to $matrix[v, v + u, ""]$
          end for
        end for
        search subcell divisions $\mathcal{D}[v, v + u]$
      end if
    end for
  end for
end for
find the best derivation $\hat{\theta}$ in $matrix[1, J, root(T)]$ and get the best translation $\hat{S} = e(\hat{\theta})$
Output: a target string $\hat{S}$

子跨度分割



1:4

1:1 2:4

1:2 3:4

1:3 4:4

1:1 2:2 3:4

1:1 2:3 4:4

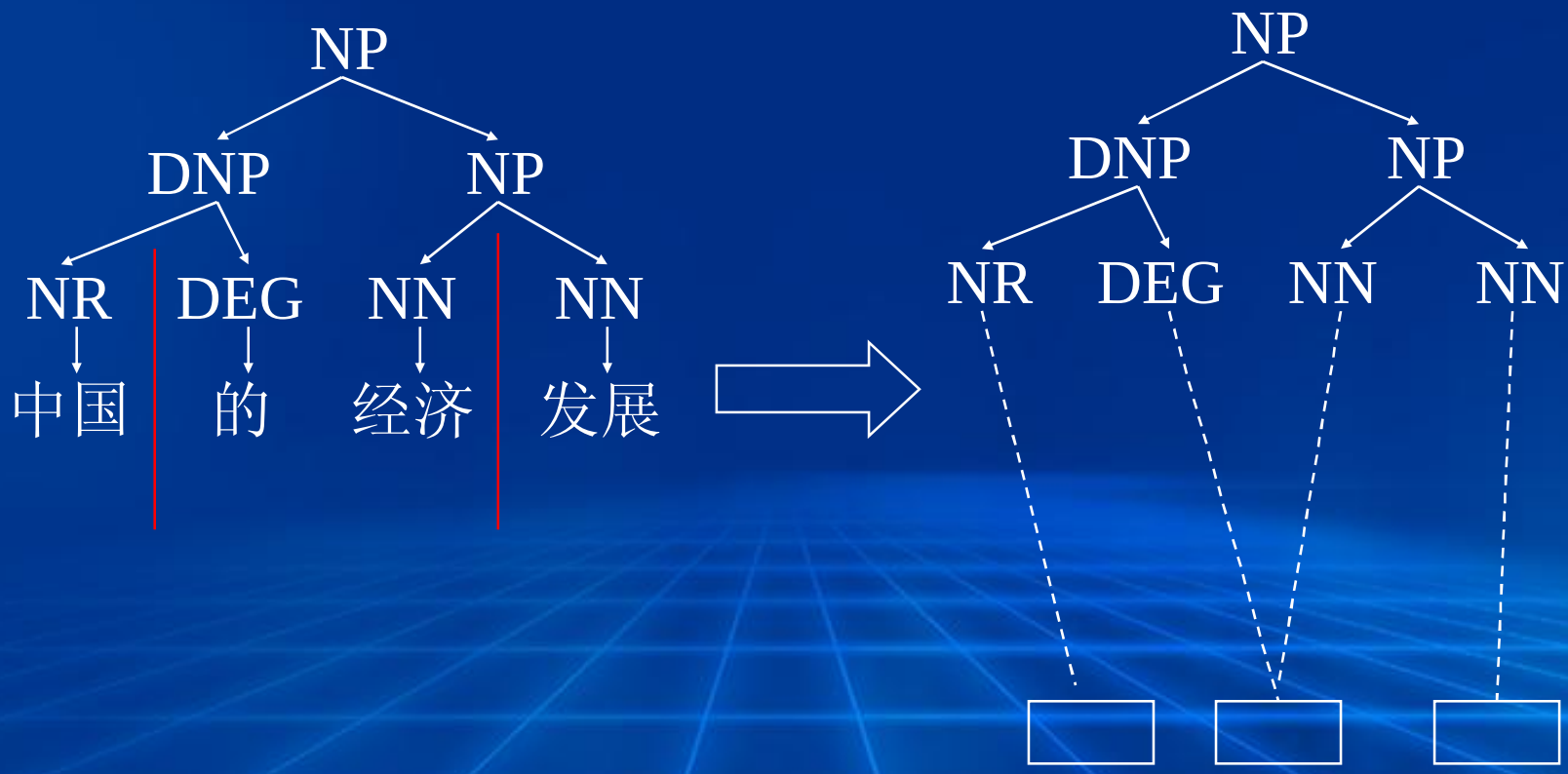
1:2 3:3 4:4

1:1 2:2 3:3 4:4

子跨度分割搜索算法

```
1:  Input: a cell [j1, j2], the derivation array matrix,
    the subcell division array D
2:  if j1 = j2 then
3:     p̂ := 0
4:     for each derivation θ in matrix[j1, j2, ·] do
5:        p̂ := max(p(θ), p̂)
6:     end for
7:     add {[j1, j2]} : p̂ to D[j1, j2]
8:  else
9:     if [j1, j2] is a forest cell then
10:       p̂ := 0
11:       for each derivation θ in matrix[j1, j2, ·] do
12:          p̂ := max(p(θ), p̂)
13:       end for
14:       add {[j1, j2]} : p̂ to D[j1, j2]
15:    end if
16:    for j := j1 to j2 − 1 do
17:       for each division d1 ∈ D[j1, j] do
18:          for each division d2 ∈ D[j + 1, j2] do
19:             create a new division: d := d1 ⊕ d2
20:             add d to D[j1, j2]
21:          end for
22:       end for
23:    end for
24: end if
25: Output: subcell divisions D[j1, j2]
```

构造辅助规则



小结

- 提出了融入森林到串规则的树到串翻译模型，该模型为短语兼容性问题提供了良好的解决方案，极大提高了树到串翻译模型的表达能力。

提纲

- *引言*
- *词语对齐的对数线性模型*
- 树到串统计翻译模型
  - *模型1*
  - *模型2*
  - *模型3*
  - 实验
- 总结

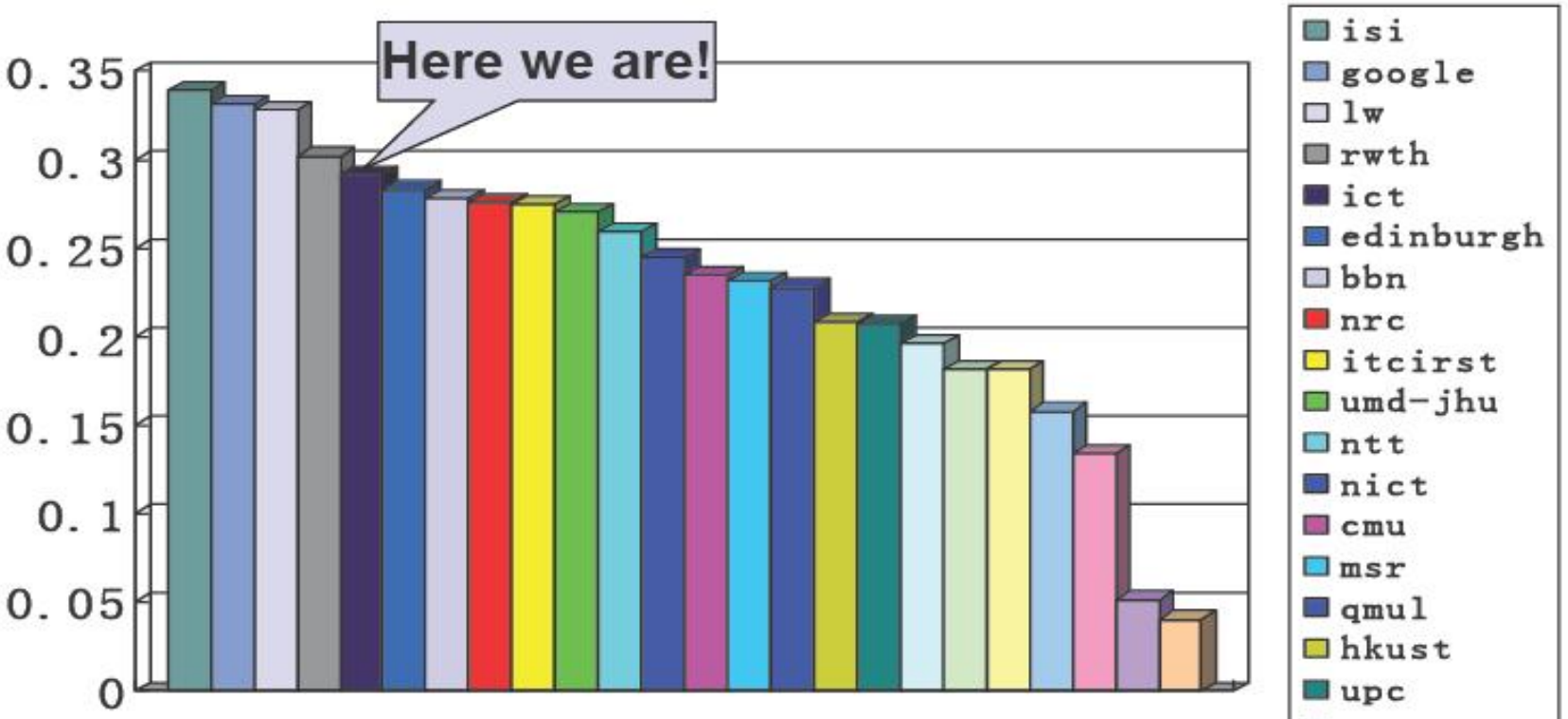
理论上的对比

特性	模型1	模型2	模型3
规则	TNR	TR	TR+FR+AR
词汇化	N	L+P+N	L+P+N
短语兼容性	S	S	S+N
复杂度	低	中	高

对比实验

系统	规则	BLEU4
Pharaoh	BP	0.2182 ± 0.0089
	SBP	0.2033 ± 0.0087
Model 1	TNR + SBP	0.2123 ± 0.0085
Model 2	SBP	0.1912 ± 0.0085
	TR	0.2302 ± 0.0089
	TR + SBP	0.2346 ± 0.0088
Model 3	BP	0.2059 ± 0.0083
	TR + FR + AR	0.2402 ± 0.0087

模型2在NIST评测中的成绩



http://www.nist.gov/speech/tests/mt/mt06eval_official_results.html

提纲

- *引言*
- *词语对齐的对数线性模型*
- *树到串统计翻译模型*
  - *模型1*
  - *模型2*
  - *模型3*
  - *实验*
- 总结

论文的研究成果（1）

- 论文提出了一种词语对齐的对数线性模型。该模型首次将判别方法引入词语对齐，具有良好的可扩展性。实验结果表明，对数线性模型在对齐质量上优于其它模型。

论文的研究成果（2）

- 论文提出了嵌入句法树的基于短语的翻译模型，该模型首次建模上利用句法信息指导短语重排序，在翻译性能上接近国际上主流的基于短语的翻译系统Pharaoh。

论文的研究成果（3）

- 论文提出了基于树到串对齐模板的翻译模型，该模型复杂性低，具备很强的重排序能力，在翻译性能上明显超过Pharaoh。

论文的研究成果（4）

- 论文提出了融入森林到串规则的树到串翻译模型，该模型为短语兼容性问题提供了良好的解决方案，极大提高了树到串翻译模型的表达能力。

下一步工作

- 将词语对齐对数线性模型应用到大规模数据处理上。
- 研究支持多对多对应关系的词语对齐模型。
- 在大规模数据上考察模型3的翻译性能。

谢谢！

中国科学院计算技术研究所

INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES