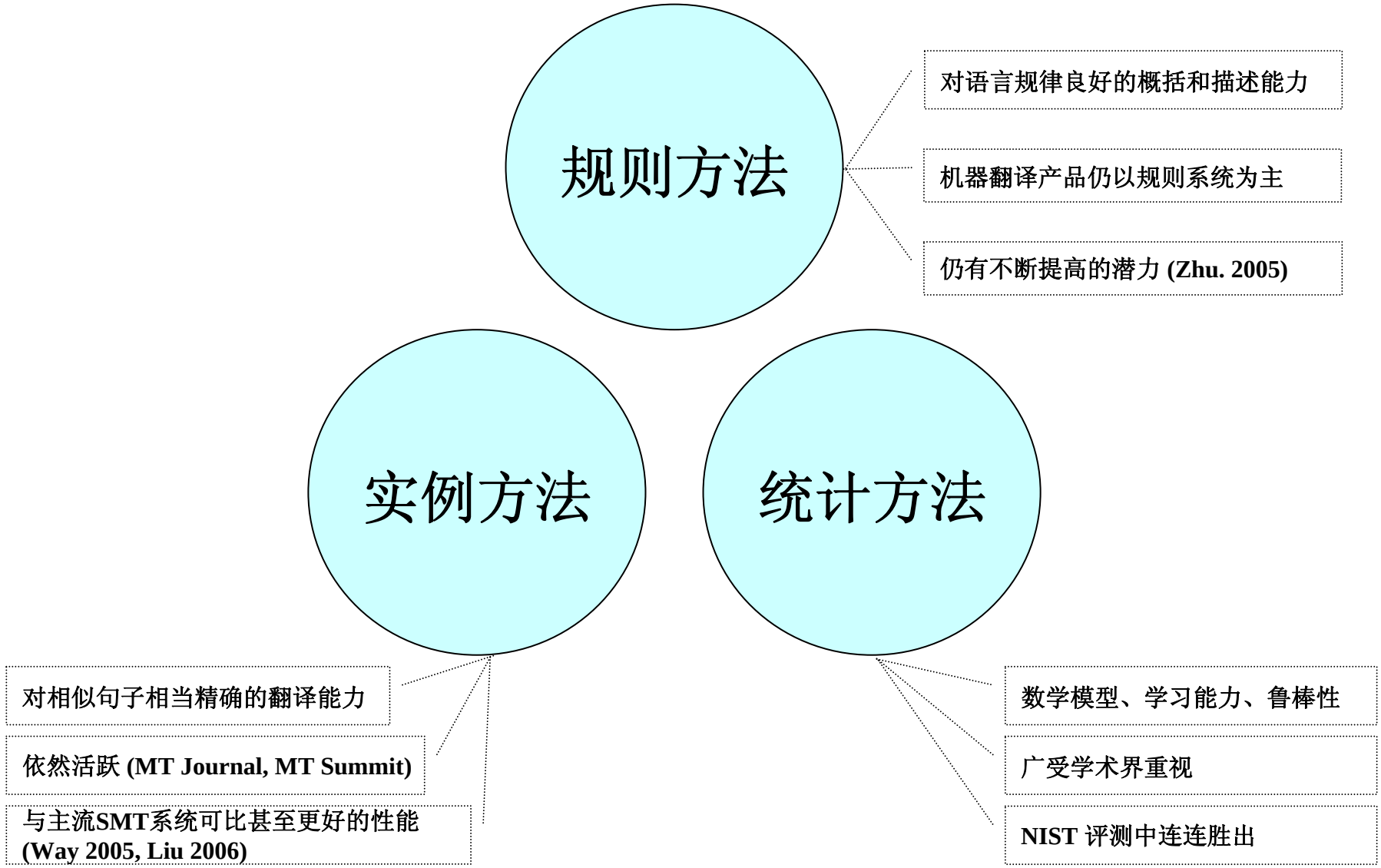
多策略的机器翻译

王海峰
东芝（中国）研究开发中心
2006年11月21日

提纲

- 概述
- 东芝的机器翻译研究
  - 概况
  - 规则方法
  - 实例方法
  - 统计方法
  - 其它
  - 多种方法的融合
- 讨论与展望

三种主流机器翻译方法


规则方法
对语言规律良好的概括和描述能力
机器翻译产品仍以规则系统为主
仍有不断提高的潜力 (Zhu. 2005)
实例方法
对相似句子相当精确的翻译能力
依然活跃 (MT Journal, MT Summit)
与主流SMT系统可比甚至更好的性能 (Way 2005, Liu 2006)
统计方法
数学模型、学习能力、鲁棒性
广受学术界重视
NIST 评测中连连胜出

东芝的机器翻译研究

- 历史
  - 28年前，开始研发规则方法
  - 5年前，中国研成立，中文相关的翻译，各种方法的研发及融合
- 应用系统
  - 自动翻译
  - 辅助翻译
  - 语音翻译
  - 基于机器翻译的跨语言信息检索
- 产品形态
  - 软件包
  - 翻译引擎授权
  - 翻译服务
  - 与东芝硬件捆绑
- 语言
  - 中、日、英互译

方法

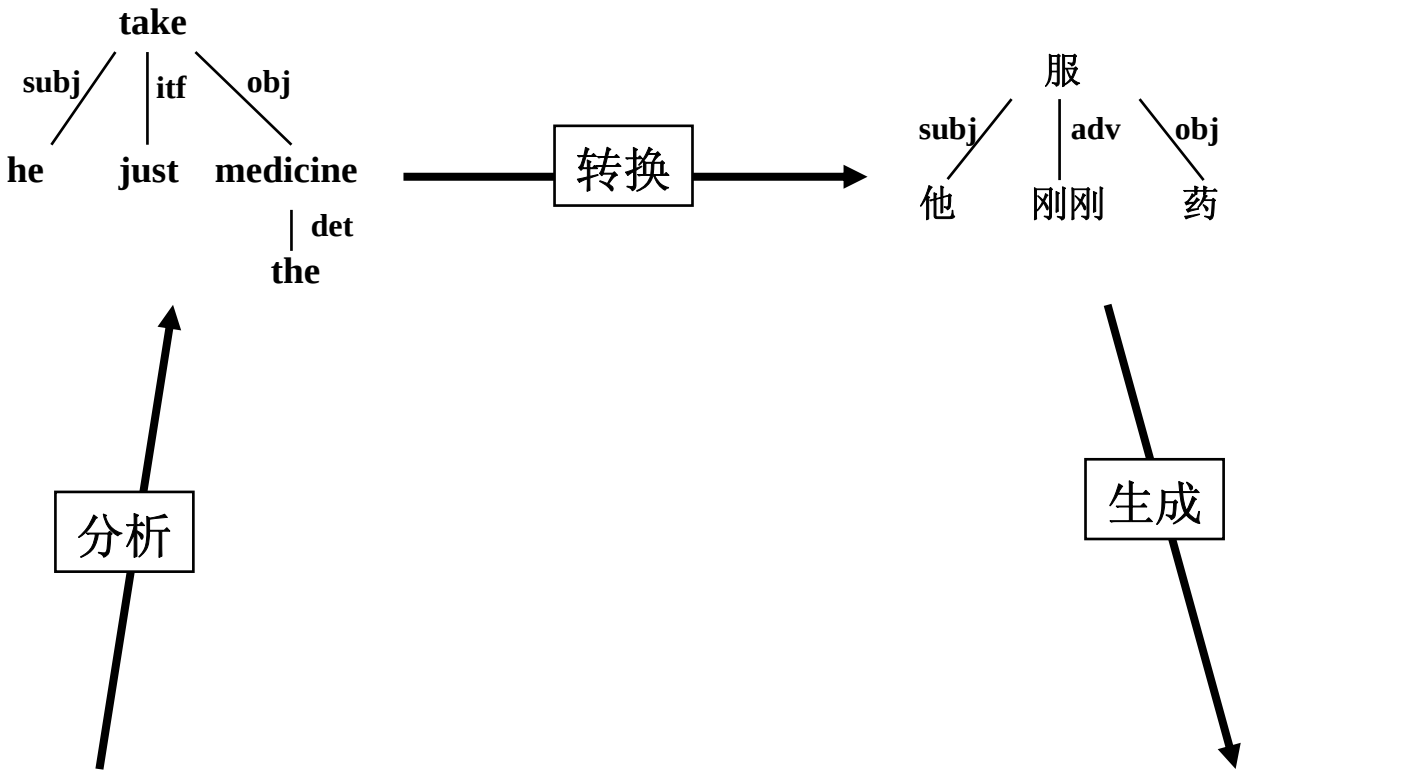
- 基于规则的机器翻译方法
- 基于实例的机器翻译方法
- 统计机器翻译方法
- 翻译记忆方法
- 相关研究
  - 词义消歧
  - 语言模型
  - 基于机器翻译的跨语言信息检索

  ……

基于规则的机器翻译

系统设计

基于转换的系统



28年的积累

特征 – 多层次, 细粒度, 可扩展

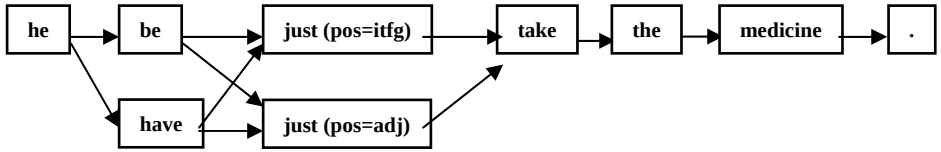
分析

源语言句子

He’s just taken the medicine.

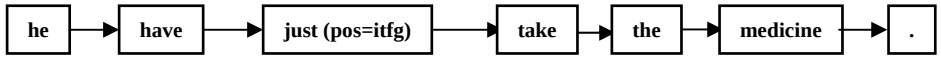
词形态分析

词网格



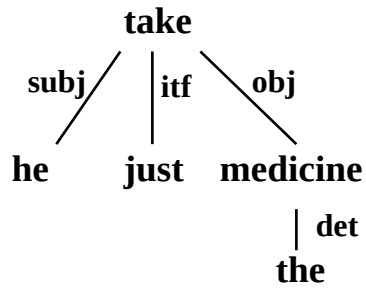
过滤

词序列



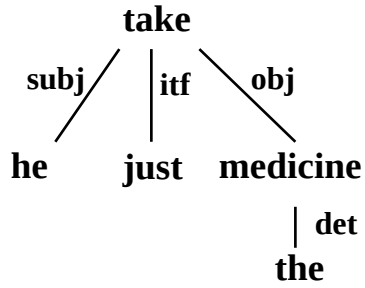
语法分析

语义树



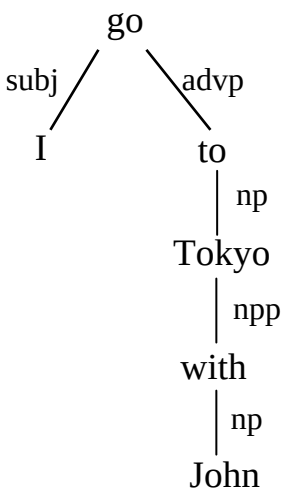
语义分析

语法树

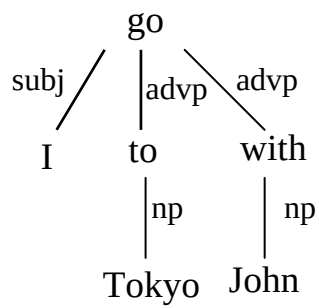


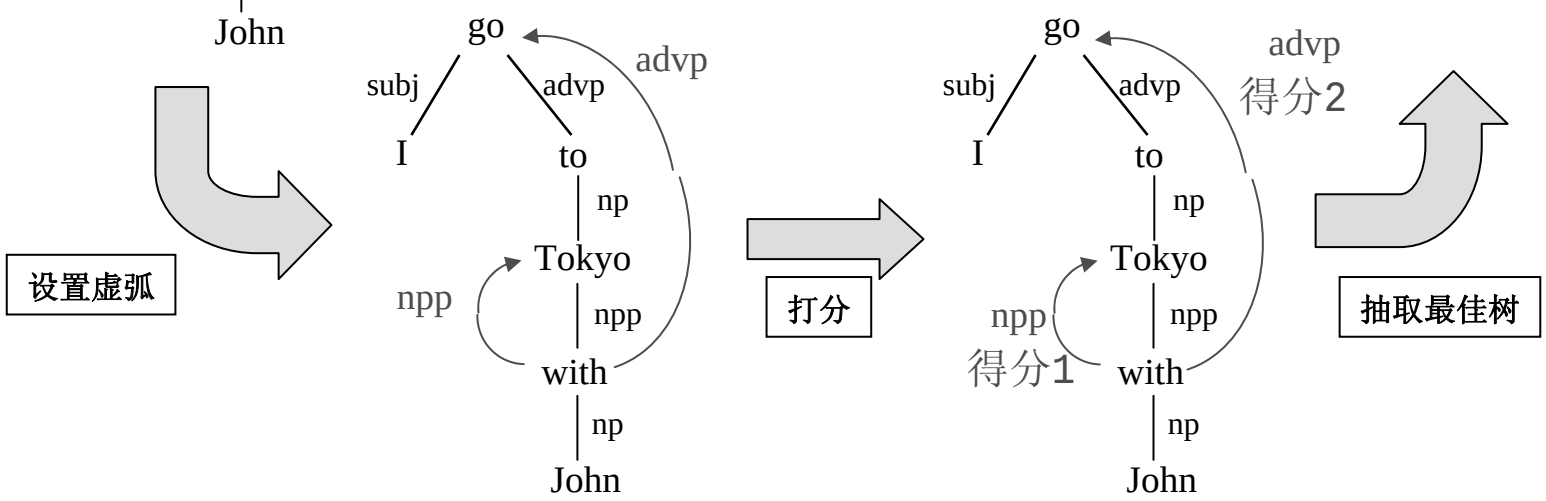
语义分析

语法树

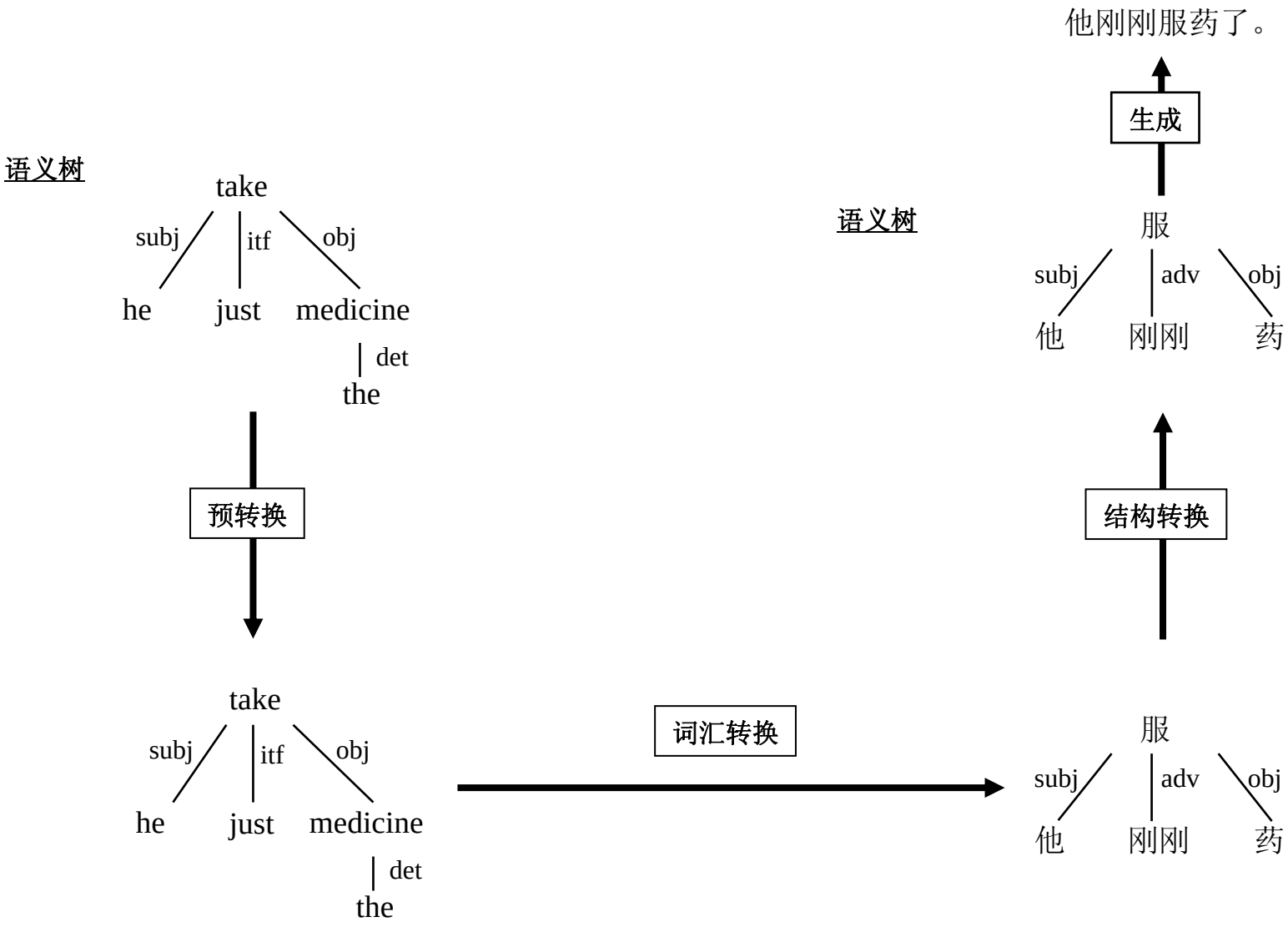

go
subj
advp
I
to
np
Tokyo
npp
with
np
John

语义树


go
subj
advp
advp
I
to
with
np
np
Tokyo
John

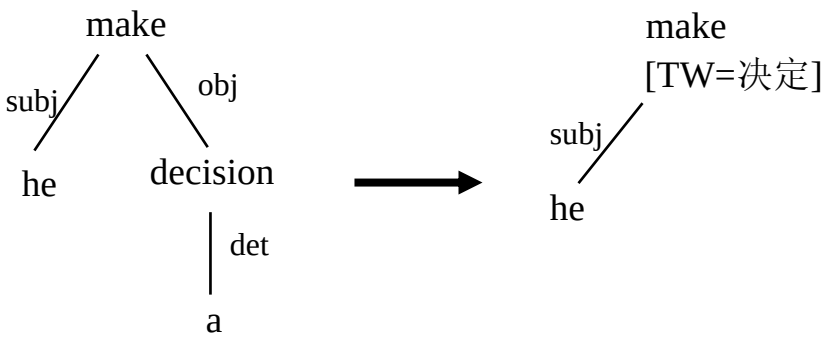

设置虚弧
go
subj
advp
advp
I
to
np
Tokyo
npp
npp
with
np
John
打分
go
subj
advp
advp
得分2
I
to
np
Tokyo
npp
npp
得分1
with
np
John
抽取最佳树

转换与生成

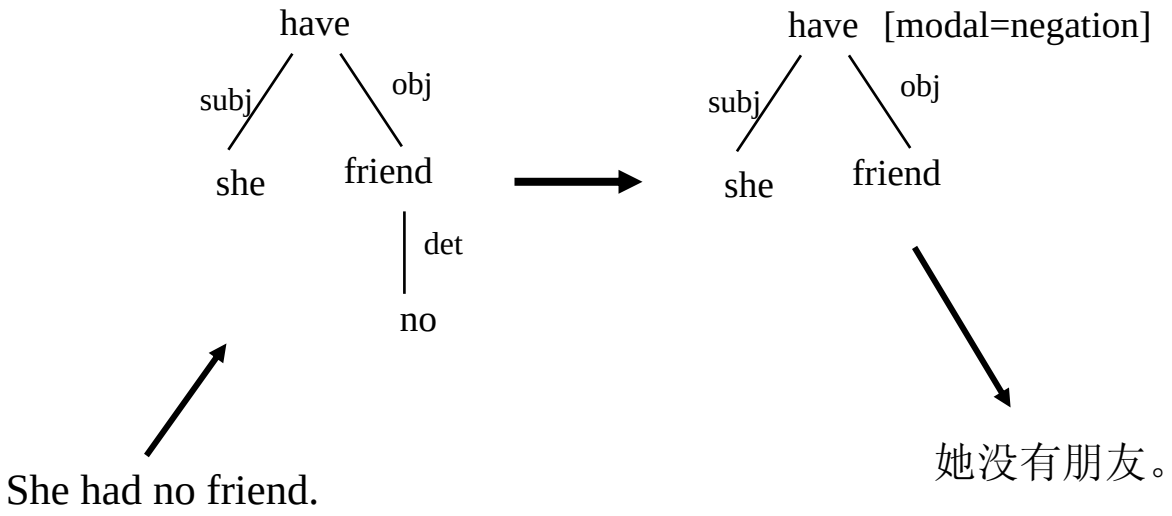

他刚刚服药了。
生成
语义树
服
subj
adv
obj
他
刚刚
药
结构转换
语义树
take
subj
itf
obj
he
just
medicine
det
the
预转换
take
subj
itf
obj
he
just
medicine
det
the
词汇转换
服
subj
adv
obj
他
刚刚
药

转换实例

(a)


make
subj
obj
he
decision
det
a
make
[TW=决定]
subj
he

(b)


have
subj
obj
she
friend
det
no
have [modal=negation]
subj
obj
she
friend
She had no friend.
她没有朋友。

潜力

- 当前性能
- 系统性能与规则数量的关系

  线性
- 系统性能与规则添加次序的关系

  近于无关
- 参考文献

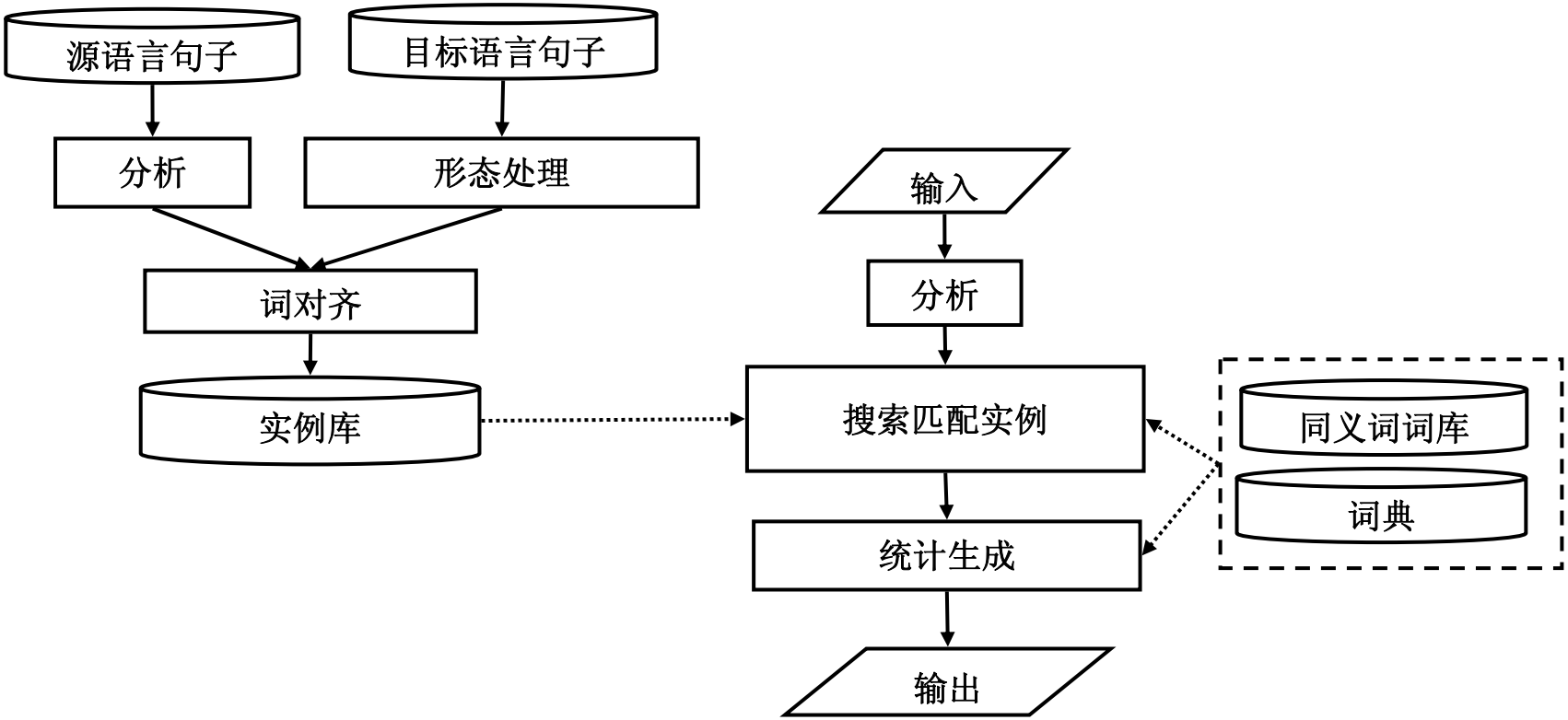
  The Effect of Adding Rules into the Rule-based MT System. MT SUMMIT 05.

基于实例的机器翻译

基于树串映射的EBMT

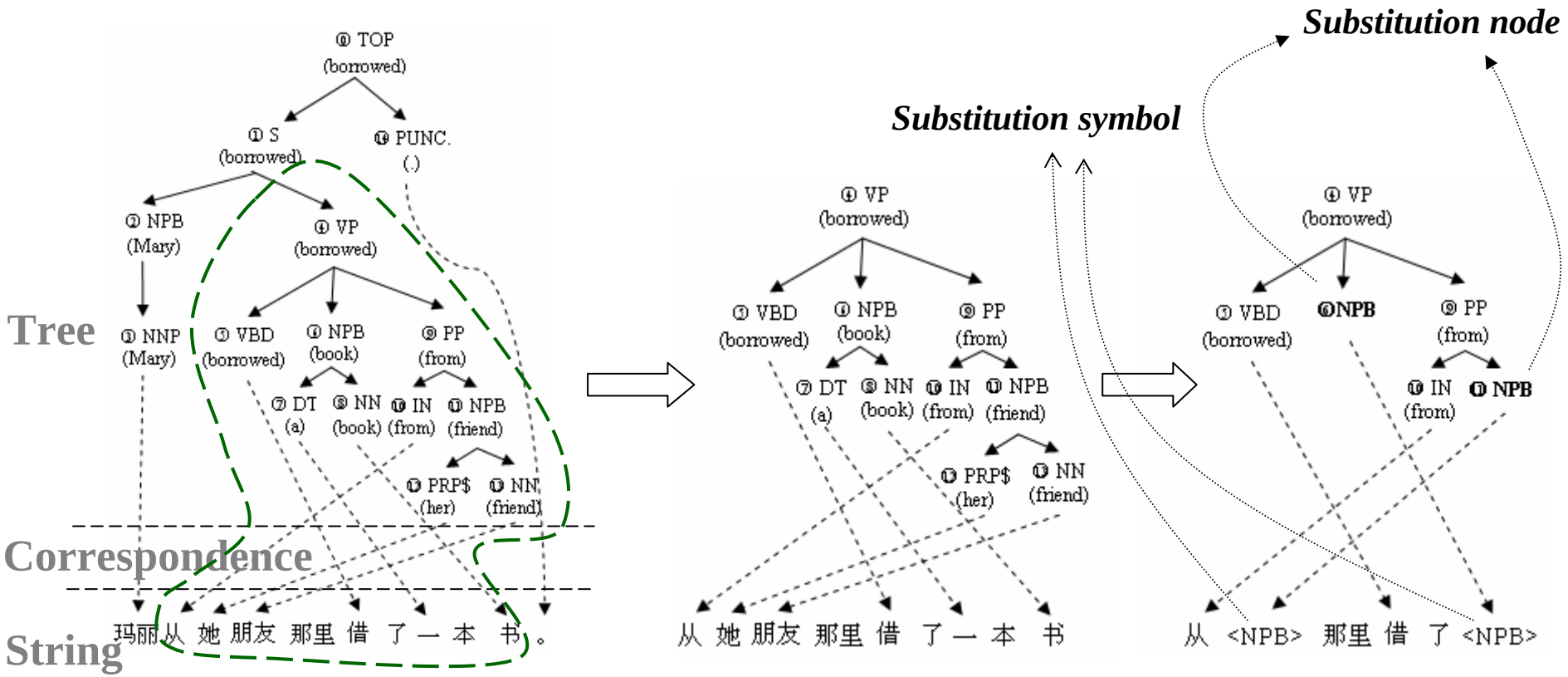
- 概要
  - 树串映射 (TSC)
    - 源语言分析树
    - 目标语言串
  - 树匹配算法
  - 统计生成
- 参考文献
  - **Machine Translation**
  - **MT SUMMIT 05**

流程



树串映射 (TSC)

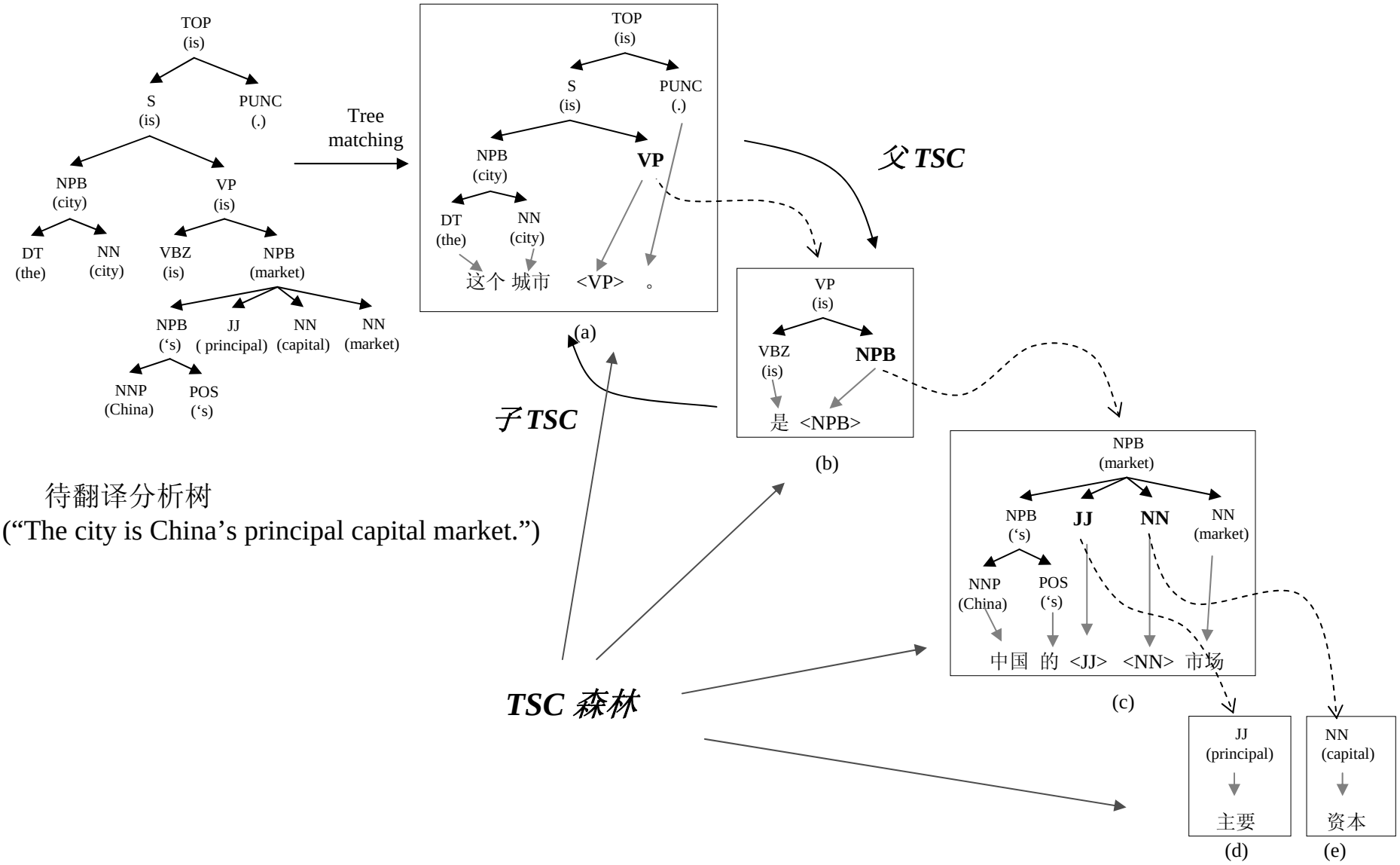
TSC = <*t*, *s*, *c*>



匹配

- **TSC** 与树的匹配
  - **t** 的非终结节点与树中的对应节点的词性和类型都一致
  - **t** 的叶节点与树中的对应节点的至少要类型一致
- 匹配算法
  - 匹配得分
  - **TSC** 森林
  - 贪心算法

匹配实例



待翻译分析树

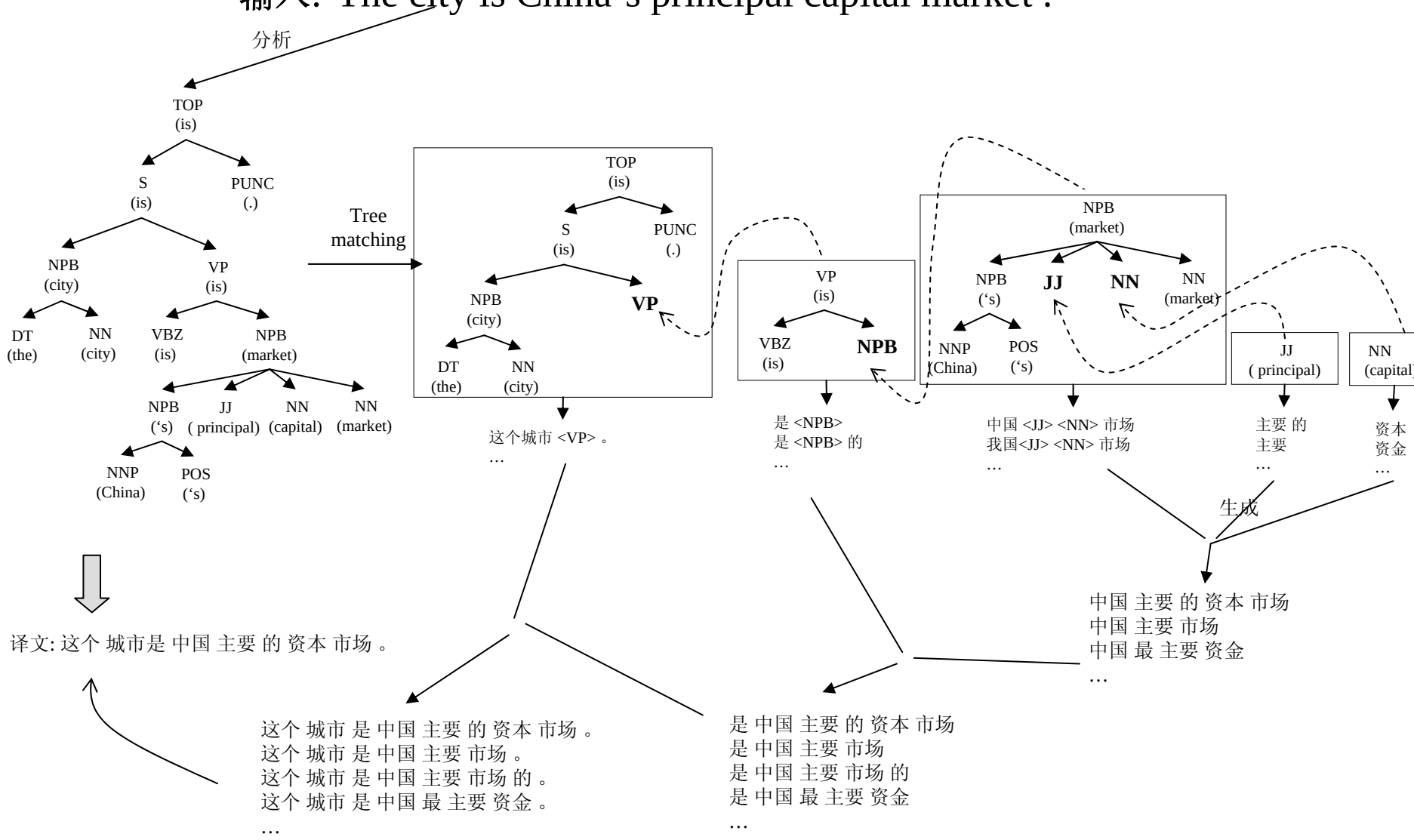
("The city is China's principal capital market.")

生成

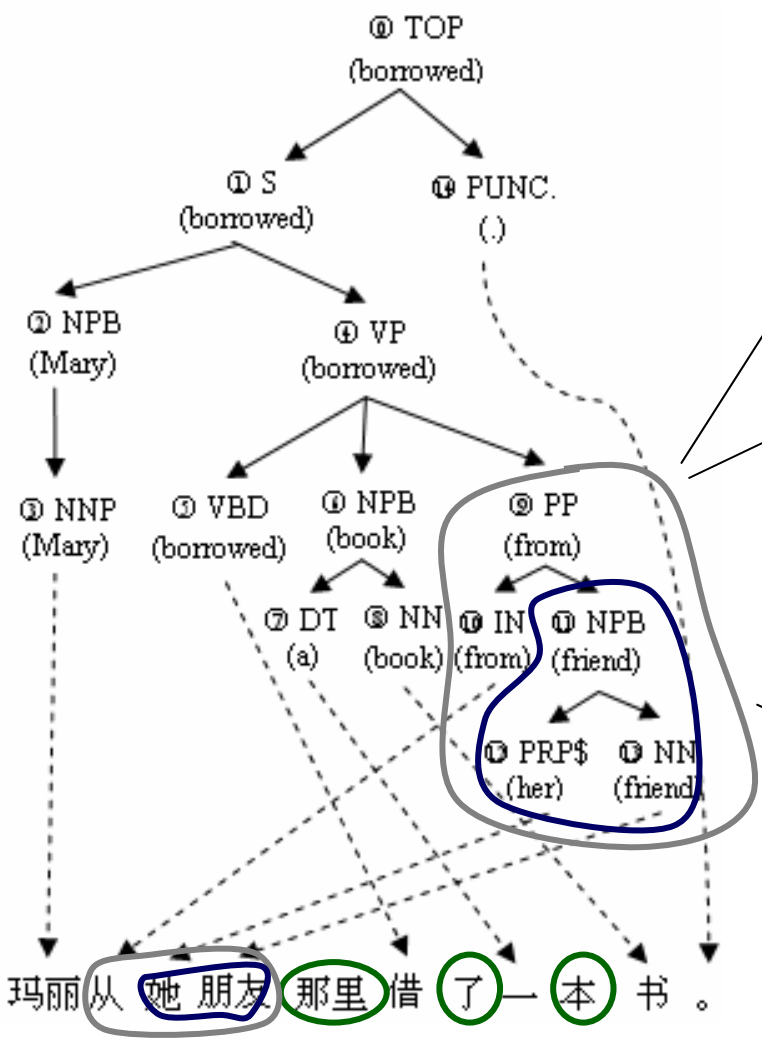
- 自底向上的生成
  - 翻译替换节点
- **TSC** 森林扩展
  - 同源 **TSC**
  - 未对齐词
- 统计生成模型
  - 语义相似度
  - 翻译概率
  - 语言模型

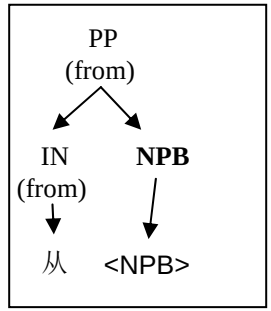
翻译实例

输入: The city is China's principal capital market .

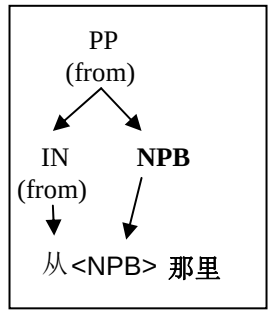


未对齐词

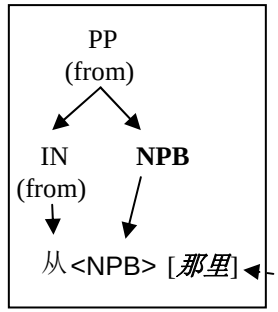




(a)



(b)



ID	Source sentence and translation	LM
1	*He got down from the bus.*	-
1a	他从公共汽车下来。	**10.98**
1b	他从公共汽车那里下来。	14.77
2	*He got the alms from the government.*	-
2a	他从政府得到救助金。	18.36
2b	他从政府那里得到救助金。	**16.39**

$$LM = -\log(P_{LM})$$

未对齐词

评价

Method	NIST
LM	4.7722
MS	4.6611
LM + MS	5.0429
LM + TM + MS	4.8174
LM + TM + MS + UW	5.2577
Word-based SMT (ReWrite)	4.5239
Phrase-based SMT (Pharaoh)	5.2214

统计机器翻译

研究内容

- 词对齐
  - 用于基于实例的翻译
  - 用于辅助翻译
  - 用于翻译知识获取
- 其它正在进行的研究
  - 统计机器翻译系统
  - 与其他方法的融合

词对齐 – 机器学习方法

- **Bagging (IJCNLP 2005)**
  - 随机抽样翻译实例
- **Boosting (MT Summit 2005)**
  - 通过对翻译实例加权来重新抽样翻译实例
  - 伪参考集
  - 保留集用于错误率计算
- **Semi-supervised boosting (COLING/ACL 2006 poster)**
  - 模型插值
  - 未标注数据的伪参考集
  - 最终的 **Ensemble**
    - 每个对齐模型的打分
    - 每个对齐弧的打分

词对齐 – 双语语料不足时的对齐问题

- 基本思想
  - 语言对 **L1-L2** 不足甚至没有双语语料
  - 语言对 **L1-L3, L2-L3** 有较大双语语料
  - 使用 **L1-L3, L2-L3** 的对齐模型来推导 **L1-L2** 的对齐模型
- 参数估计
  - 翻译概率估计
  - 繁殖概率估计
  - 位置扭曲概率估计
  - 跨语言词相似度
- 参考文献 – **COLING/ACL 2006 poster**

词对齐 – 特定领域词对齐问题

- 动机
  - 现有词对齐方法都需要较大的双语语料和/或词典
  - 领域资源有限
    - 领域语料
    - 领域词典
  - 利用两类语料
    - 通用语料 – 通用词汇
    - 领域语料中的词汇
      - 通用词汇
      - 领域词汇
- **结果级融合 (ACL 04 poster)**
- **模型级融合 (ACL 05)**

其它相关研究

- 翻译记忆
- 分析
  ICASSP 06
- 语言模型
  - 用于分词
    (COLING/ACL 06)
  - 用于机器翻译
  - 用于语音
- 词义消歧 (与哈工大合作)
  (COLING/ACL 06)
- 基于机器翻译的跨语言信息检索
  (COLING/ACL 06)

融合

- 系统级
  - 规则方法 + 实例方法
  - 机器翻译结果用于跨语言信息检索
- 模块级
  - 实例系统中的统计模块
    - 词对齐
    - 语言模型
  - 实例系统中基于规则的前处理和后处理
  - 辅助翻译系统中的统计和实例模块
  - 规则方法改进统计词对齐
- 资源处理
  - 在规则和实例系统中使用统计方法处理翻译资源
  - 基于规则的前、后处理用于各种目的

讨论

规则方法	通常用人工撰写的描述语言规律的规则进行翻译

从语言现象入手来描述语言的成分构成规律

实例方法	在翻译过程中直接使用翻译实例

从机器学习的角度强调对翻译实例的抽象程度

统计方法	使用事先训练好的统计模型进行翻译

从数学角度强调统计建模能力

谢　谢！