Semantics, Discourse and Machine Translation
(语义、语篇和机器翻译)

张 民，苏州大学

CIPSC战略研讨会，贵阳，2014年4月18日

大纲

- CIP的发展战略
- 基于多层次篇章语义的机器翻译

CIP发展战略

- 一个国家的科学研究和这个国家的发展状况、历史、文化、政治、经济等等密切相关……
- 我们每个人都在辛苦并快乐地工作着…
- 科学研究：
  - 原创性、基础研究、核心技术、影响力

CIP发展战略

- 未来十年中文信息处理的主要着眼点和着力点应该在哪里？
  - 中文语义计算
  - 语言文化产业
- 国际上自然语言处理领域近年来有哪些最新的前沿进展？
  - 基础资源、语言学、大数据、机器学习和算法

CIP发展战略

- 中文信息处理的重要原始创新可能在哪里？
  - 中文语义理解与计算

- 中文信息处理相关产业发展的重要方向可能是什么？
  - 基于大规模语义计算的智能信息处理系统

CIP发展战略

- 对学会工作的希望和建议？
  - ACL进入CCF的一类会议
  - 更加重视国际顶级会议，而不仅仅是SCI论文
  - 学报进入EI

大纲

- CIP的发展战略
- **基于多层次篇章语义的机器翻译**

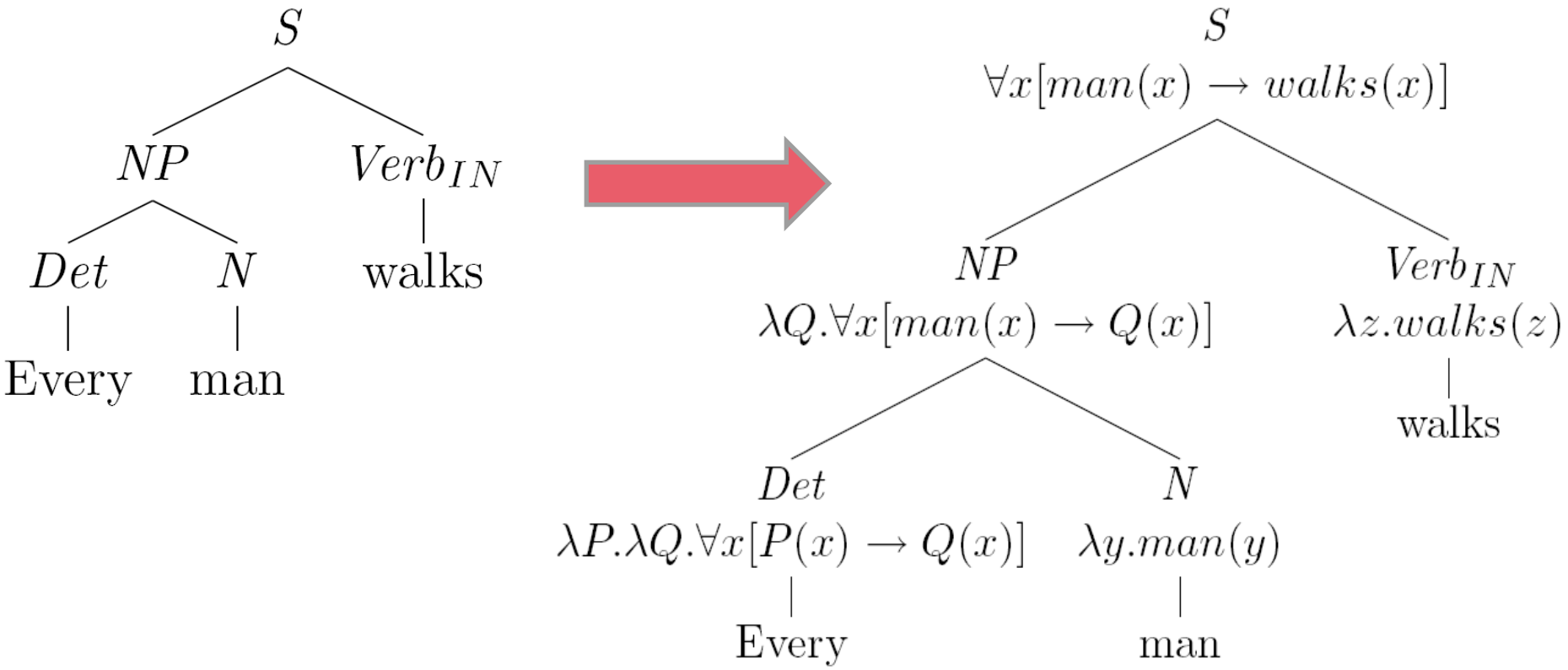
什么是多层次篇章语义？

- 词汇语义
- 句子语义
- 篇章和篇章语义

语义是怎么表示的？

- 词汇语义
- 浅层语义
  - 谓词论元
- 谓词逻辑
- 分布式表示

谓词演算: 实例



句法表示

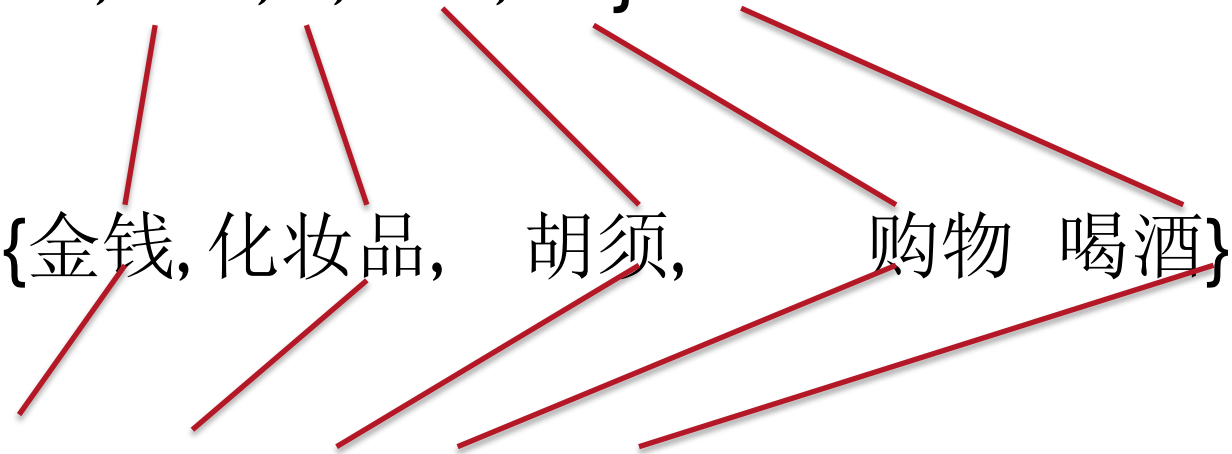
语义表示

分布式表示

女人= (100, 200, 0, 500, 50}

词汇表: {金钱, 化妆品,　胡须,　　购物　喝酒}

男人= (150, 10, 200, 50, 150}



Why now: Semantics-based SMT

SOOCHOW UNIVERSITY

- 短语和句法模型愈来愈难以依靠增加数据和模型复杂度来提高性能
- 大规模语义资源（如**WordNet/HowNet，Propbank**，知识图谱，基于互联网数据挖掘的语义资源和本体资源等）不断建设和完善
- 自然语言分析技术的进展（词法、句法、语义和篇章）
- 统计机器学习技术（特别是深度学习）在语音识别和某些自然语言处理任务上的成功应用

篇章的定义

- **Terminology: 话语？篇章？语篇？文本？**
- **Taxonomy:**
  - **From:** 词、形态、词性、短语、句法、语义
  - **To: 篇章**
- **Definition (from Wiki):**
  - **A unit of text used by linguists for the analysis of linguistic phenomena that range over more than one sentence (even within a single sentence).**

篇章的话题结构

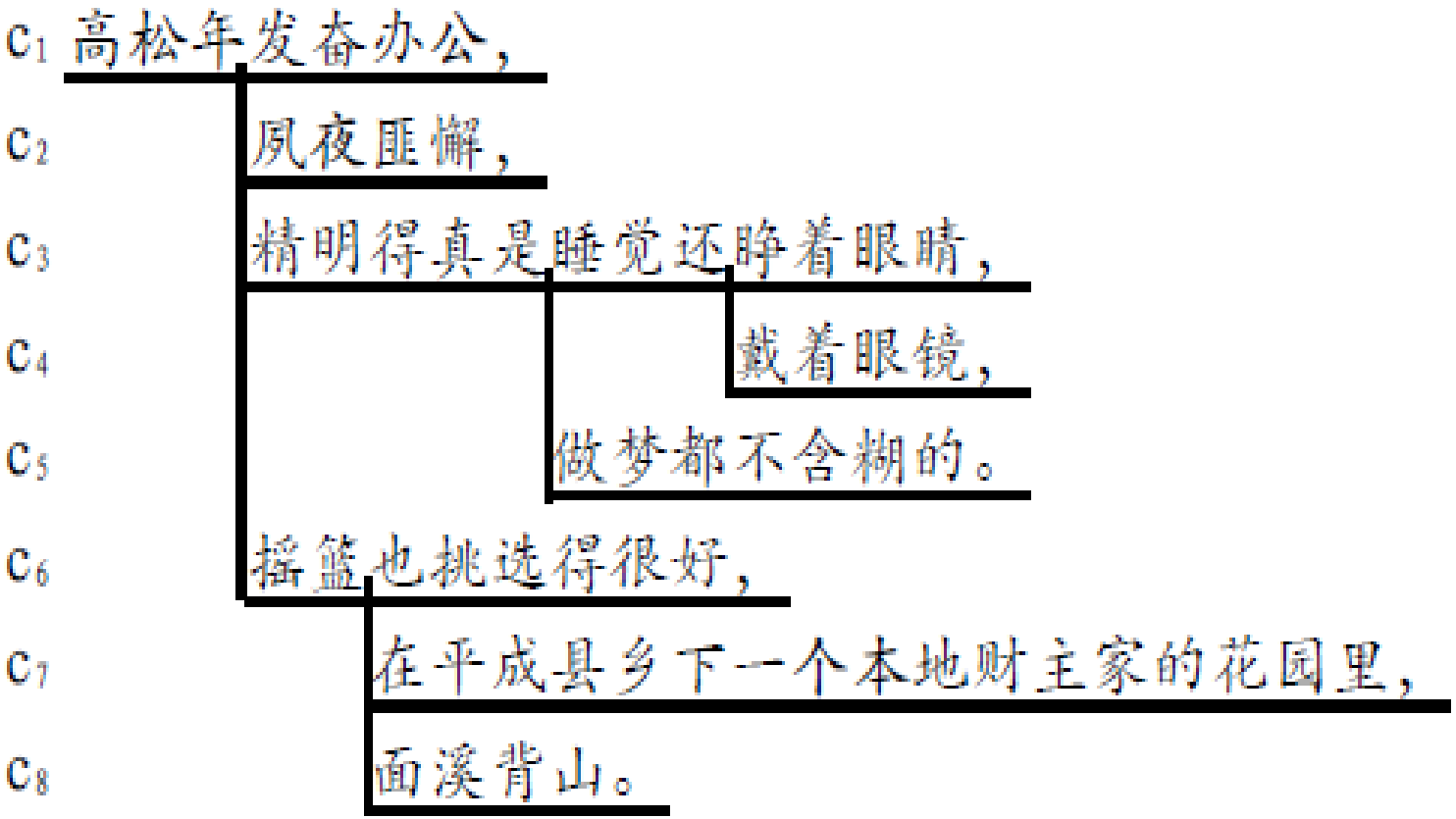


图1. **微观话题结构**实例[宋柔，2012]

篇章的逻辑语义结构

(JS(XJ1 **如果** 你 不 出面 干预 ,) (RB (XJ2 他 **即使** 把 设备 卖 了 ,) (XJ3 **也** 没 人 阻止 得 了 他 。)))

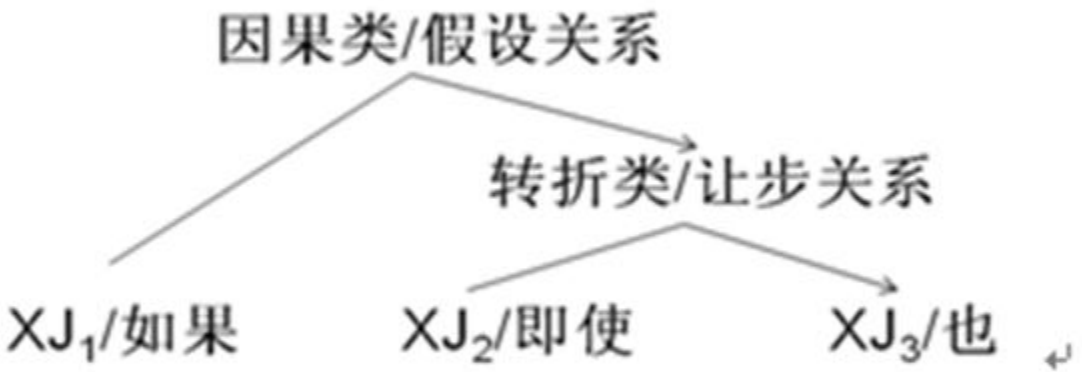


图 2. 篇章的逻辑语义结构(包括关系结构、关系类别、关联词和依存关系，JS:假设； RB:让步；XJ:小句)

What is Discourse: 核心问题

- 篇章研究核心问题的可计算性：结构和特征
  - 篇章的基本结构分析：

    篇章结构指的是篇章内部关系的不同结构化表达形式，主要包括逻辑语义结构(Discourse Relations)、指代结构(Co-reference)、话题结构(Topics)、功能结构(Functions)、事件结构(Eventualities)等范畴

  - 篇章的基本特征的研究：

    衔接性(cohesion)、连贯性(coherence)、意图性(intentionality)、可接受性(acceptability)、信息性(informativity)、情景性(situationality)和跨篇章性(intertextuality)等七个基本特征

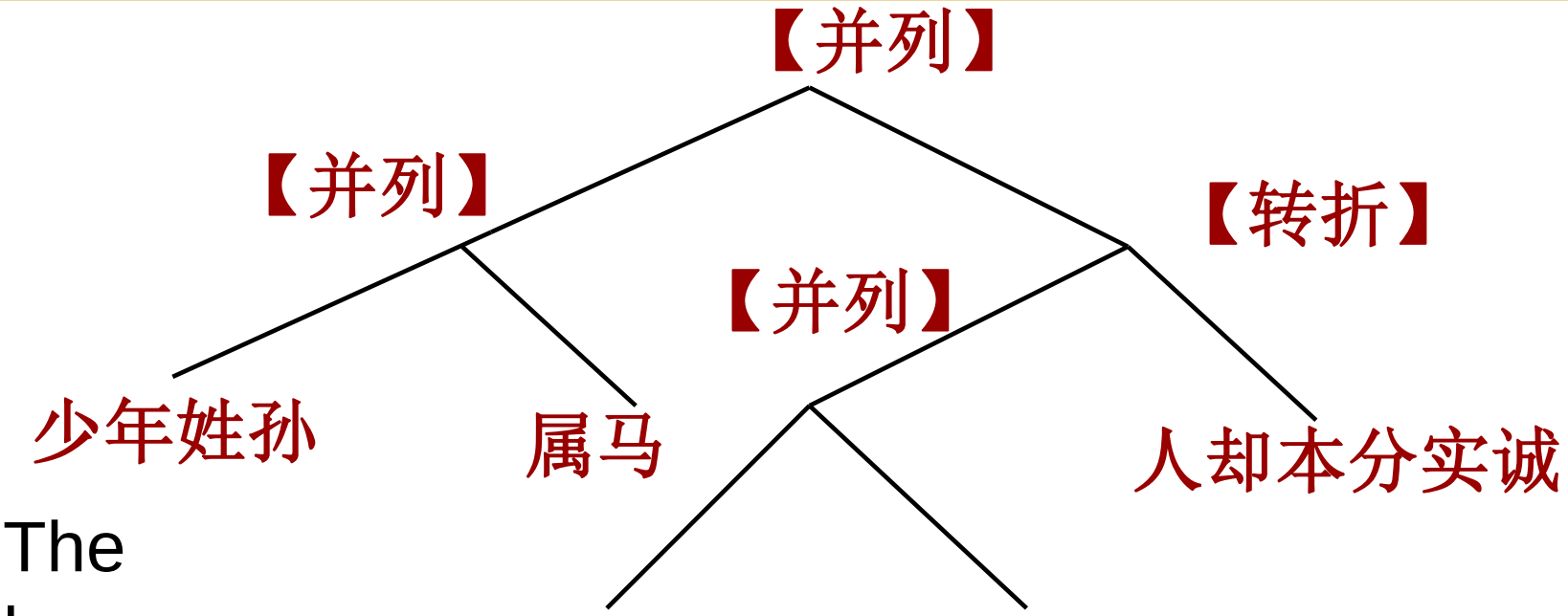
Why Discourse-based SMT

- **Leverage on Discourse-level Knowledge for SMT: 时态, 格, 词汇选择, 话题一致性, 连接词, 句子结构, 篇章结构， 指代, 评价…**
- **International research community?**
- **Now it is a good timing: an emerging, promising and important research …**

Why Discourse-based SMT

- 少年姓孙，//[并列]属马，/[并列]比小水小着一岁，///[并列]个头也没小水高，//[转折]人却本分实诚。(贾平凹《浮躁》)
- The boy, a member of the Sun family, //[并列]was born in the year of the horse. /[并列]Although he was a year younger ///[并列]and a head shorter than Water Girl, //[转折]he was honest and sincere . (Goldblatt, 1991)

Why Discourse-based SMT



The boy, a member of the Sun family

was born in the year of the horse

Although he was a year younger

and a head shorter than Water Girl

he was honest and sincere

大纲

- CIP的发展战略
- **基于多层次篇章语义的机器翻译**

Semantics-based SMT: the state-of-the-art

- 基于词汇语义的机器翻译
- 基于谓词论元结构的机器翻译
- 基于谓词演算的机器翻译
- 基于分布式表示和深度学习的机器翻译

Discourse-based SMT: the state-of-the-art

- Document level translation
- Topic-based translation
- RST structure-based translation
- Lexical cohesion-based translation
- Coherence-based translation
- Discourse connectives translation
- Discourse-aware decoding
- Discourse-based translation evaluation
- ……

Discourse-based SMT: Previous and the State-of-the-art Work

- Daniel Marcu, Lynn Carlson, and Maki Watanabe. ***The automatic translation of discourse structures.*** NAACL-2000
- Xiaodong Shi and Yidong Chen. ***Previews of approach to discourse-based machine translation***. Proc. Frontiers of Chinese Information Processing, 2006
- Bonnie Webber's work: …
- Mei Tu, Yu Zhou and Chengqing Zong. ***A novel translation framework based on rhetorical structure theory***. ACL-2013
- Mei Tu, Yu Zhou and Chengqing Zong. ***Enhancing grammatical cohesion: generating transitional expressions for SMT***. ACL-2014
- … … … … … … …

Discourse-based SMT: Our Work

SOOCHOW UNIVERSITY

- **Cache-based Document-level SMT**
- **Document-level Tense Models**
  - **N-gram-based**
  - **Classifier-based**

- Zhengxian Gong, Min Zhang, Chew Lim Tan and Guodong Zhou. *Cache-based document-level SMT*. EMNLP-2011
- Zhengxian Gong, Min Zhang, Chew Lim Tan and Guodong Zhou. *N-gram-based tense models for SMT*. EMNLP-2012
- Zhengxian Gong, Min Zhang, Chew Lim Tan and Guodong Zhou. *Classifier-based tense models for SMT*. COLING-2012

Discourse-based SMT: Our Work

SOOCHOW UNIVERSITY

- **Document-level topic model for rule selection**
  - **Similarity model**
  - **Sensitivity model**

- Xinyan Xiao, Deyi Xiong, Min Zhang, Qun Liu and Shouxun Lin. *A topic similarity model for hierarchical phrase-based translation*. ACL-2012
- Min Zhang, Xinyan Xiao, Deyi Xiong and Qun Liu. *Topic-based dissimilarity and sensitivity models for translation rule selection*. Journal of Artificial Intelligence Research, 2014

Discourse-based SMT: Our Work

SOOCHOW UNIVERSITY

- **Discourse-based SMT: cohesion and coherence**
  - **Lexical cohesion: lexical device**
  - **Lexical chain cohesion: lexical chain**
  - **Topic coherence**

- Deyi Xiong and Min Zhang. *A topic-based coherence model for SMT*. AAAI-2013
- Deyi Xiong, Guosheng Ben, Min Zhang, Yajuan Lü and Qun Liu. *Modeling lexical cohesion for document-level MT*. IJCAI-2013
- Deyi Xiong, Yang Ding, Min Zhang and Chew Lim Tan. *Lexical chain based cohesion models for document-level SMT*. EMNLP-2013

Semantic-based SMT: Our Work

SOOCHOW UNIVERSITY

- **Lexical semantics for SMT**
  - **Word sense induction**
- **Predicate-argument for SMT**
  - **Predicate translation**
  - **Argument reordering**

- Deyi Xiong and Min Zhang. *A Sense-based Translation Model for Statistical Machine Translation*. ACL-2014
- Deyi Xiong and Min Zhang. *Modeling the Translation of Predicate-Argument Structure for SMT*. ACL-2012

多层次篇章语义机器翻译

- 词汇和句子语义机器翻译：语义推导和语义合成性
- 双语篇章对齐语料的构建
  - 切分、层次结构、关系
  - 中心、角色分布
- 篇章结构的翻译
- 话题结构的翻译
- 衔接性和连贯性建模
- 篇章级机器翻译评测

感谢！

ありがとう！

감사합니다!

Thank you!

Terima kasih !

நன்றி!

ขอบคุณ!

Cảm On Bạn!

请帮忙注册：谢谢!

SOOCHOW UNIVERSITY

1. COLING-2014 Tutorial (Aug 23)

   "Dependency Parsing: Past, Now and Future"

2. ACL-2014 Tutorial (June 22)

   "Semantics, Discourse and Machine Translation"