中文自然语言处理的现状与展望

刘群
liuqun@ict.ac.cn
2009.01.06
于洛阳解放军外国语大学

目录

引言：中文信息处理技术发展线路图

- 文字处理阶段
  - 编码：**GB2312、BIG5、UNICODE、GB18030**
  - 输入：
    - 键盘输入法：五笔字型、拼音输入……
    - 手写输入：印刷体识别（**OCR**）、联机手写识别、脱机手写识别
    - 语音输入：孤立词→连续语音、小词汇量→大词汇量、特定人→非特定人、朗读语音→自然语音
  - 输出：字库、打印、显示、语音合成…

引言：中文信息处理技术发展线路图

- 语言处理阶段
  - 基础技术
    - 词处理：词语切分、词性标注、未定义词识别、词义排歧
    - 句处理：句法分析、语义角色标注
    - 篇章处理：指代消解、篇章分析
  - 应用技术
    - 信息检索：分类聚类、搜索引擎、话题检测与跟踪
    - 信息抽取：命名实体、实体关系、事件抽取
    - 自动文摘
    - 自动问答
    - 机器翻译

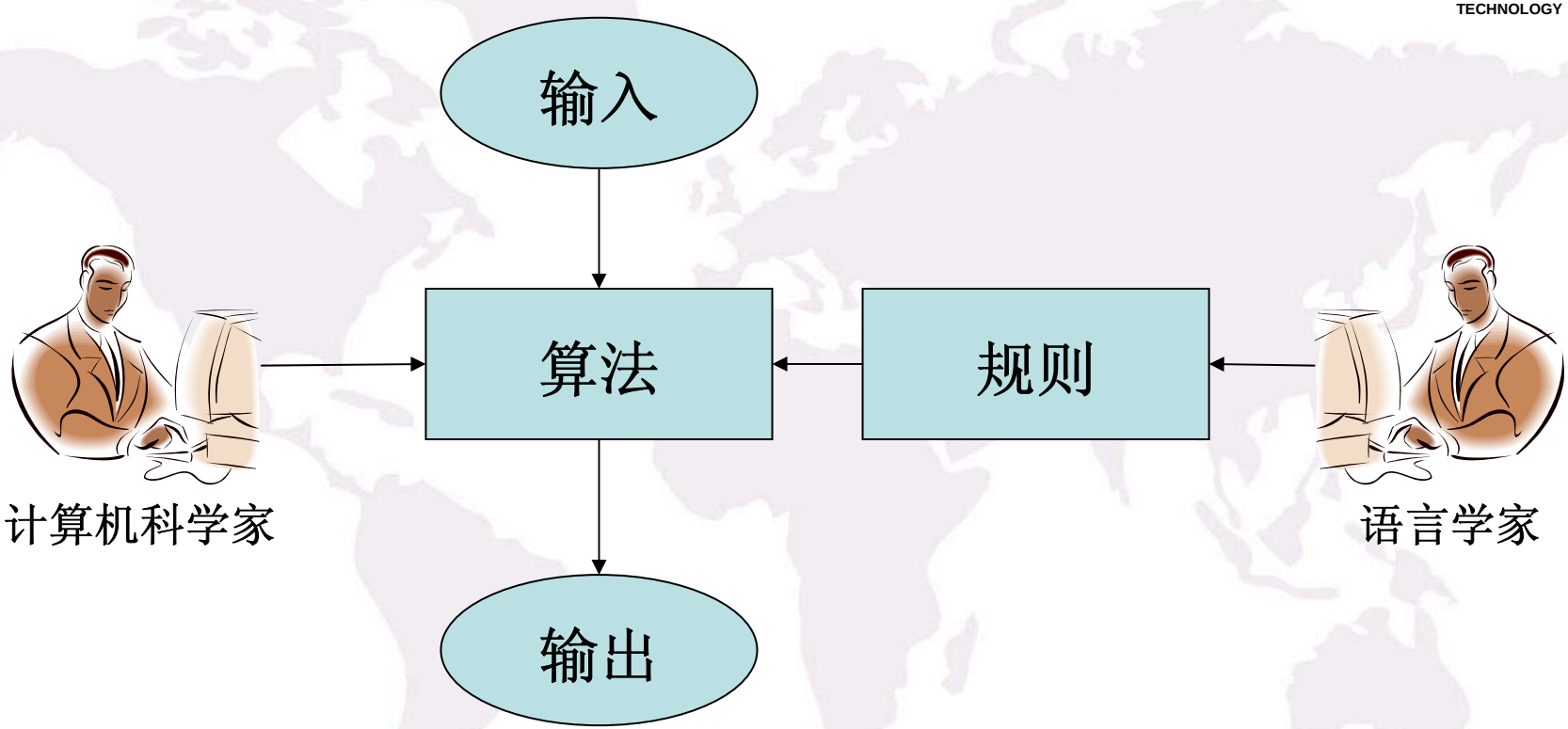
引言：文字处理和语言处理的关系

- 文字处理技术是语言处理技术的基础
  - 统一的数据编码使得数据交换成为可能
  - 大规模的数据输入技术为语言信息处理提供了语言数据
- 语言处理技术可以为使文字处理技术更上一层台阶
  - 键盘输入法、语音识别、语音合成等技术在语言处理技术的支持下性能都有了大幅度的提高

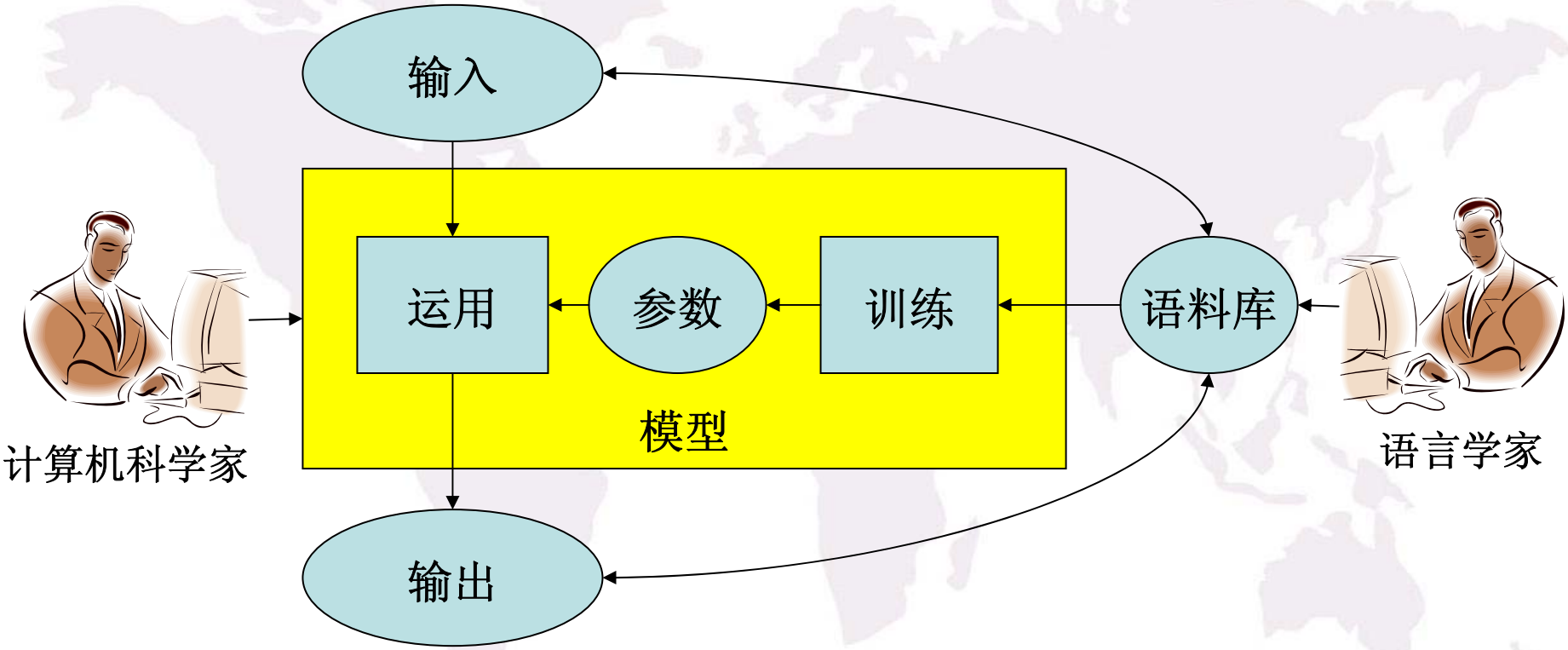
引言：自然语言处理研究的两个阶段

- 规则方法阶段
  - 语言学家：撰写“规则库”（包括“词典”）
  - 计算机科学家：编写算法程序，对“规则库”进行解释和执行
- 统计方法阶段
  - 语言学家：建立“语料库”
  - 计算机科学家：
    - 建立统计模型
    - 利用语料库训练模型参数
    - 编写算法解决问题

规则方法的流程



统计方法的流程



统计方法与规则方法的比较

- 规则方法
  - 优点：
    - 语言知识的表示直观、灵活
    - 易于表达复杂的语言知识
  - 缺点：
    - 语言知识的覆盖率低
    - 语言知识的冲突缺乏统一解决机制

统计方法与规则方法的比较

- 统计方法
  - 优点：
    - 统计模型提供了统一的冲突解决机制
    - 大规模数据提高了语言知识的覆盖率
  - 缺点：
    - 不善于表示复杂的、深层次的语言知识
    - 对于数据稀缺的语言（小语种）没有好的解决办法

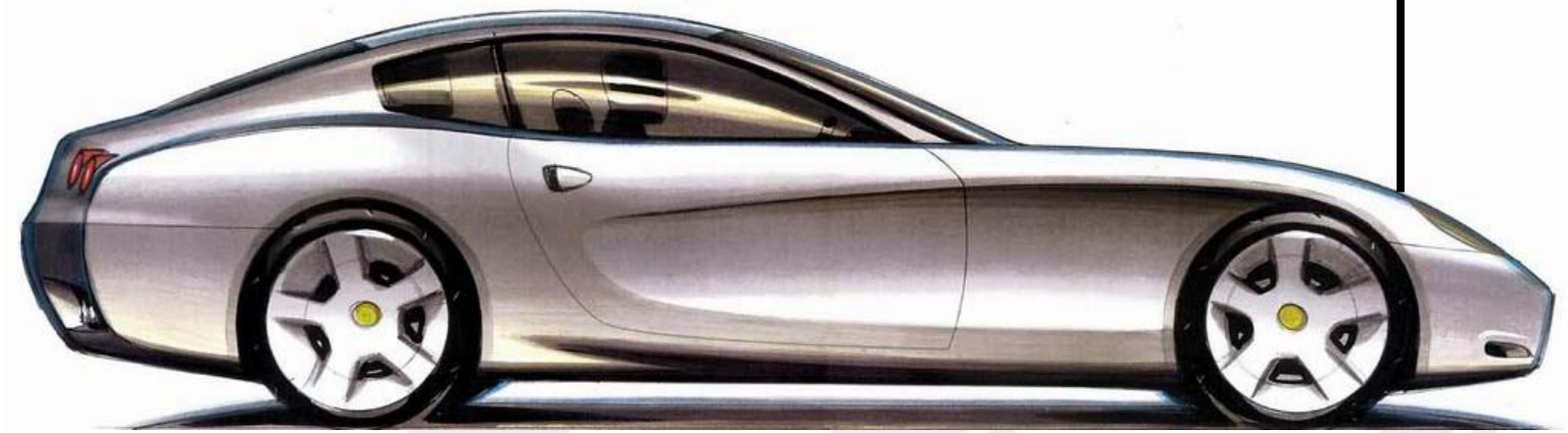
统计方法与规则方法的融合

- 近年来，统计方法成为自然语言处理处理研究的主流
- 单纯的规则方法，如果不借助统计工具，很难超越统计方法
- 统计方法在发展的过程中不断改进，逐渐吸收了传统规则方法的优点，模型趋于复杂，一些统计模型直接建立在规则表示的基础上，可以表示很复杂的语言知识

统计自然语言处理的发展

统计自然语言处理的两大驱动

统 计 自 然 语 言 处 理



大规模共享
语言资源

公开周期性
技术评测与交流

目录

- 引言
- **数据资源**
- 技术评测
- 基础技术
- 应用技术
- 总结与展望

语言资源的类型 1/2

- 词典
  - 根据信息类型进行分类
    - 词表》+拼音》+概率：词语切分、拼音输入法
    - 语法信息词典》+概率：句法分析
    - 语义词典：语义分析、机器翻译、信息检索
      - 语义聚合：同义词典（**Thesaurus**）、知识本体（**Ontology**，**WordNet**），概念词典、同义词林、复杂主题词表
      - 语义组合：配价词典、语义框架词典（**FrameNet**）、知网（**HowNet)**
  - 其他分类维度：单语与多语、通用与专用……

语言资源的类型 2/2

- 语料库
  - 单语语料库
    - 文本语料库
    - 切分标注语料库
    - 词义标注语料库
    - 句法树库
    - 语义角色标注语料库
  - 多语语料库：不同语言、不同方言、不同编码
    - 平行语料库、可比语料库
    - 句子对齐、词语对齐、结构对齐

语言资源的共享

- 语言资源的共享是自然语言处理研究的主要驱动力之一
- 语言资源本身就是成果，而不是副产品，这种成果的价值必须必须通过广泛共享才能体现出来，用的人越多，价值越大
- 很多著名的研究机构都是通过其语言资源的广泛共享和传播，累积了崇高的学术声誉（宾州树库、北大语料库、北大词典）

语言资源的共享

- 国际上主要的语言资源共享机构
  - **LDC**：语言数据联盟
    Language Data Consortium
  - **ELDA/ELRA**：欧洲语言数据联盟
  - **ChineseLDC**：中文语言资源联盟
    - 挂靠在中文信息学会，由语言资源工委会管理
    - 中立的机构
    - 负责数据共享、宣传推广、法律手续
    - 已经产生广泛的影响：**Google**、**Nokia**开始提供数据

对语言资源建设的一些建议 1/4

- 语言知识库建设是一种语言的自然语言处理研究最好的起步工作
  - 没有数据，NLP无法深入展开
  - 算法和软件容易随着技术的进步被淘汰，而语言知识库是永远有价值的
  - 语言知识库可以被所有研究者使用，可以扩大在研究界的学术影响

对语言资源建设的一些建议 2/4

- 早期应多关注规模

  就小语种而言，现有语料库的规模远远小于其他主流语言的语料库的规模

  汉英句子对齐语料库：数百万句子对

  汉语和英语的文本语料库：**Giga Word，Tera Word**

  语料库规模太小，难以满足应用需求

  语料库深度加工应谨慎

  深度加工的语料库，应有较成熟的理论指导，并且制订较详尽合理的标注规范

对语言资源建设的一些建议 3/4

- 语言资源应尽可能公开
  - 使用的人越多，语料库的价值越大
  - 使用的人越多，学术影响越大
  - 语言资源共享可以收费，但最好不要太看重短期经济效益，可以采用薄利多销的形式，学术影响扩大带来的潜在经济效益远远大于近期的经济效益
  - 国内研制的语言资源建议通过**ChineseLDC**公开

对语言资源建设的一些建议 4/4

- 资源建设可多关注语言特有的语言现象
  - 例如：蒙古语的多种编码平行语料库
- 语言资源开发应尽可能跟技术评测相结合
  - 技术评测是促进**NLP**技术发展的重要推动力
  - 技术评测和语言资源建设可以互相促进
    - 技术评测可以促进语言资源的应用
    - 技术评测反过来可以对语言资源建设提出新的要求

目录

自然语言处理技术评测的意义

- 自然语言处理研究总要面临各种各样的评测
- 自然语言处理是一项评测驱动的学科
- 评测成绩成了除发表论文以外衡量学术水平的一个重要依据
- 评测甚至有点太多太滥了，但仍然有很多新的评测不断出现
- 为什么？

宏观角度—从可重复性说起

- “可重复性”是实验性科学研究的一个重要原则
- 只有可重复的研究，才能被同行所广泛接受
- 任何一篇有学术价值的研究论文，其必要条件是文章中给出了足够的细节，使得其他研究者可以重复其实验结果

自然语言处理研究中的可重复性

- 在自然语言处理研究中，可重复性面临严重问题
- 由于这个领域的实验需要使用大量的数据，这些数据不可能在一篇论文中给出来，而这些数据的采集通常具有非常大的偶然性，同一个方法在不同的数据条件下得到的结果差异可能会非常之大。这样，一篇论文所介绍的研究方法对另一个研究者来说，就很难重复。

自然语言处理研究中的可重复性

- 对于语音识别而言，影响实验结果的因素可能有：说话人的性别、年龄、口音，录音的环境、噪音，话筒的质量，说话的方式（自然方式还是朗读）等等。有时，人们通常容易忽略的一些看起来似乎微不足道的因素都可能会对实验的结果造成重大的影响。可想而知，如果没有共同的数据，一个研究者的实验是很难被另一个研究者所重复的。而不同的研究者如果采用不同的数据进行实验，其结果几乎不具备可比性。这样整个研究领域的进展就变得非常困难。

可重复性与技术评测

- 为了解决这个问题，一些公开的技术评测应运而生。由评测的组织者给出共同的数据集，制定统一的测试方法和评价标准，不同的研究者就可以在相同的条件下进行实验比较，并且可以重复别人的实验结果，从而得到可比的数据。
- 这种评测受到研究人员的普遍欢迎，对研究工作起到了很好的促进作用。这种公开的、周期性的评测的出现，使得这一领域的研究工作出现了一种新的模式，而且这种模式逐渐成为了这一领域研究人员开展研究工作的最主要的模式。

微观角度—用评测指导研究（1/2）

- 自然语言处理是一门实验科学，任何理论设想都必须通过实验的检验才能被接受
- 在自然语言处理研究中，我们无时无刻不在尝试各种新的选择，达到模型、算法的改变，小到增加一条规则、调整一个参数
- 新的选择通常在带来有些优点的同时，也带来一些缺点，因而我们需要从总体上评价一个新的选择是带着我们往好的方向走还是往坏的方向走，评测可以起到这个作用

微观角度—用评测指导研究（2/2）

- 对某些研究，结果的评价是简单而且显而易见的，比如语音识别、词性标注等
- 对另一些研究，结果的评价是很困难的，比如机器翻译、自动文摘、语音合成等
- 对于后一类研究，研究高效的自动评测方法，就成为了非常迫切的任务。
  - 没有好的自动评测方法，我们好像要在黑暗中摸索，不知道前进的方向
  - 好的自动评测方法好比指南针，为我们的研究指明方向

评测驱动的研究（1/2）

- 定期举办公开的技术评测，在评测中，给定公共的训练数据集、测试数据、测试方法和评价标准；
- 各研究单位根据评测要求自行开发系统，并提交结果；
- 评测组织者对各单位提交的结果进行评价，并在一定范围内公开评测的结果；
- 各参加评测的单位提交参加评测的系统报告，详细介绍系统所采用的方法；
- 评测组织者组织研讨会，在会上参评单位报告各自参加评测的技术细节，并进行学术交流；

评测驱动的研究（2/2）

- 重复以上过程。往往上一届评测中表现出色的理论和方法会被其他研究单位所重复或者模仿，而只有当一种方法被多个研究单位重复并被证明有效时，才能被这个科研共同体所普遍接受，一些无法被重复的方法逐渐被淘汰。
- 通过反复多次评测，一些原先被认为是很困难的课题逐步被解决，科研共同体的研究兴趣将发生转移，一些新的研究课题被提出来。这时旧的评测任务通常会被新的评测任务所取代。
- 也有的时候，新的评测任务并非来自研究共同体内部，而是来自外部的需求，比如政府部门或者企业界的需求。这种需求可以具体化为某种评测任务，如果定义得当，这种评测同样可以对研究工作起到很好的促进和引导作用，甚至可能导致一个新的研究方向的诞生。

NIST系列评测

- **NIST**
  - **National Institute of Standard and Technology**
  - 美国国家标准技术研究所
- **NIST**在美国国防先进技术研究计划署（**DARPA， Defense Advanced Research Projects Agency**）等部门支持下，开展了一系列周期性的技术评测工作，在全球范围内吸引了大量的研究工作者参加，产生了巨大的影响。这也是到目前为止国际上影响力最大的系列评测。

NIST系列评测

- 语音识别系列评测
- 文本检索评测**TREC**
- 机器翻译评测（**Open MT Evaluation**）
- 信息提取评测（**MUC**、**ACE**）
- 话题检测与跟踪评测（**TDT** ）
- 多文档文摘评测（**DUC**）

其他国际评测

- 中文分词：**SIGHAN Chinese Language Processing Bakeoff**
- 跨语言检索：**NTCIR，CLEF**
- 机器翻译：**IWSLT，TCSTAR**
- 语言分析：**CoNLL Shared Task**
- 语义处理：**SemEval**

863评测

- 我国最有影响的系列中文信息处理技术评测是由国家863计划组织的中文信息处理与智能人机接口评测，简称863评测。该项评测涵盖了中文信息处理和人机交互技术的大部分研究领域。
- 从1991年到2005年，一共举办过8次863评测，涉及9个大的评测类别，国内相关领域的研究机构大部分都参加过863评测。这项评测对国内中文信息处理研究起到了巨大的推动作用。

863评测—项目设置

- 语音识别（Automatic Speech Recognition，ASR）
- 语音合成（Machine Translation，TTS）
- 机器翻译（Machine Translation，MT）
- 汉语分词（Chinese Word SEGmentation，SEG）（含词性标注和命名实体识别）
- 信息检索（Information Retrieval，IR）
- 文本分类（Text Categorization，TC）
- 文本摘要（Text Summary，TS）
- 文字识别（Character Recognition，CR）
- 人脸检测与识别（Face Recognition，FR）

863评测—项目详情

	1990	1991	1992	1994	1995	1998	2003	2004	2005
ASR	√	√	√	√	√	√	√	√	√
TTS				√	√	√	√	√	
MT				√	√	√	√	√	√
SEG					√	√	√	√	
IR							√	√	√
TC							√	√	
TS					√	√	√	√	
CR		√	√	√	√	√	√		
FR								√	

863评测—参评系统

年份	1990	1991	1992	1994	1995	1998	2003	2004	2005
届别	Pre	1st	2nd	3rd	4th	5th	6th	7th	8th
评测类别数	1	2	2	4	6	6	8	8	3
参评系统数	5	16	17	39	65	43	46	113	45

863评测—黄昌宁教授评价

- 国家863计划智能计算机专家组曾对语音识别、汉字（印刷体和手写体）识别、文本自动分词、词性自动标注、自动文摘和机器翻译译文质量等课题进行过多次有统一测试数据和统一计分方法的全国性评测，对促进这些领域的技术进步发挥了非常积极的作用。但是这期间也遇到了一些阻力，有些人试图用各种理由来抵制这样的统一评测，千方百计用‘自评’来取代统评。其实，废除了统一的评测，就等于丧失了可比的基础。这个损失使得上述任何理由都变得异常苍白。”

—黄昌宁，统计语言模型能做什么？，语言文字应用，2002年第1期，第77-84页

其他国内评测

- **973**项目评测
- 全国搜索引擎和网上信息挖掘学术研讨会（**SEMW**）系列评测

中文信息学会系列评测

- 进入“十一五”以后，由于863计划管理体制上的变化，863计划暂时没有对中文信息处理方面的技术评测提供直接支持。为了保持国内中文信息处理评测的连续性，中文信息学会语言资源建设和管理工作委员会决定举办“中文信息学会系列评测”。中文信息学会语言资源建设和管理工作委员会2007年开始筹备成立评测工作组，并协调有关“中文信息学会系列评测”的组织工作。

中文信息学会评测

- 目前这些在中文信息学会名义下的评测有：
  - 机器翻译评测：
    - 第三届统计机器翻译研讨会评测**SSMT2007**
    - 第四届全国机器翻译研讨会评测**CWMT2008**
  - 汉语处理评测
    - 第一届中国中文信息学会汉语处理评测（**CIPS-CLPE2007**），与第四届**SIGHAN**汉语处理评测联合举办
    - 第二届中国中文信息学会汉语处理评测正在筹备中，有清华大学和东北大学承办
  - 第一届情感计算评测：已完成
  - ……希望会有更多……

中文信息学会评测

- 这些评测与以往的863评测相比
  - 更加注重参评单位的自主性
  - 更加强调评测本身的学术性
  - 评测之后的技术交流更加深入

目录

- 引言
- 数据资源
- 技术评测
- **基础技术**
- 应用技术
- 总结与展望

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

词语切分

- 中文词语切分经过这些年的发展已经比较成熟了
  - 早期基于规则的方法只有宋柔老师开发了一个性能较高的系统，由于其中包含大量的技术诀窍，别人无法复制
  - 中科院计算所开发了第一个开源的基于层叠隐马尔科夫模型的汉语词法分析系统**ICTCLAS**，该系统在第一届**Sighan**国际中文切词比赛中获得多项第一，开放源代码以后，被普遍采用
  - 近年来的研究和评测表明，采用判别式机器学习方法的“以字组词”的中文分词系统性能超过了基于生成模型的方法
  - 大规模测试表明，在数据类型差别不大的情况下，基于统计的汉语切词系统准确率可以达到**97%**以上

词语切分

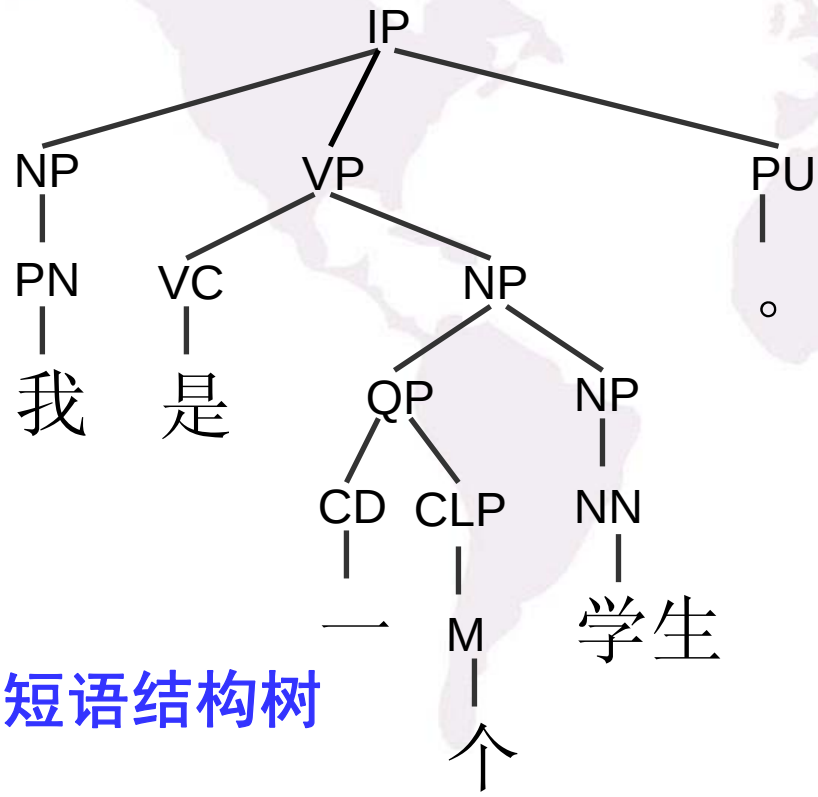
- 中文词语切分技术面临的问题
  - 领域适应性问题：由于领域的不同（比如将新闻领域语料训练出来的切词系统用在化学领域），切词系统的正确率会有较大幅度的下降
  - 多粒度切词问题：实际的应用（比如信息检索、机器翻译等）都表明，多粒度的切词对系统性能的提高有比较好的效果，现在在这方面工程实现上基本可以做到，但研究方面工作还不多

词性标注

- 英文的词性标注可以达到很高的准确率（97%以上）
- 中文的词性标注准确率仍然较低（92%左右）
  - 主要的原因可能是中文缺乏形态变化，仅仅依靠文本的局部上下文信息，很难准确地标注出词语的词性
  - 中文词性的定义本身在语言学家还存在争议，人工标注的一致性也比英语更低

句法分析

- 句法分析通常分为短语结构分析和依存分析两种类型



短语结构树

依存树

句法分析

- 早期的研究工作集中于短语结构分析方面，这方面的工作以**Collins**和**Charniak**的工作最为著名，主要的方法都是词汇化的概率短语结构语法（**Lexicalized PCFG**）
- 短语结构分析目前主要采用的数据是宾州树库
  - 在英文宾州树库上，句法分析的准确率可以达到**92-93%**
  - 在中文宾州树库上，句法分析的准确率不到**80%**（基于自动的词语切分和词性标注）

句法分析

- 近几年，依存句法分析研究的热度逐年升高，**CoNLL**会议多次将依存分析作为共享任务（**Shared Task**）进行评测
- 目前依存分析方法主要有基于图的分析算法和基于转换的分析算法，以及二者相结合的方法
- 目前英语依存分析的准确率最高在**90%**左右，而中文依存分析的准确率在**84%**左右

句法分析

- 可以看到，目前中文句法分析的准确率比英文有较大的差距，这严重影响了中文自然语言处理的深层次应用
- 中文句法分析准确率低的原因，一般还是认为是由于中文缺乏明确的词性标记，可以用于帮助句法分析的信息太少

语义角色标注

- 语义角色标注的目的是为句子中的每个动词标注出其相关的名词及其语义角色

语义角色标注

- 近年来语义角色标注也引起了较多的关注，**CoNLL**会议也多次以此作为共享任务（**Shared Task**）进行了评测
- 语义角色标注通常都基于**PropBank**和**FrameNet**等语料库进行，中文的相关工作主要是基于**PropBank**
- 目前基于正确的句法树，英文和中文的语义角色标注准确率都可以达到**91%**左右
- 但基于自动的句法分析，英文的语义角色标注准确率大约在**80%**，而中文语义角色标注的准确率不到**70%**

语义角色标注

- 可以看到，中文句法分析是语义角色标注最主要的瓶颈问题
- 我们认为（很多学者也有类似的认识），这种将句法分析和语义分析分成两阶段进行的做法不适合于汉语，汉语分析需要在句法分析阶段就引入语义信息进行一体化分析，才有可能达到较高的准确率

篇章处理

- 指代消解
- 篇章分析

目录

- 引言
- 数据资源
- 技术评测
- 基础技术
- **应用技术**
- 总结与展望

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

应用技术

- 信息检索：发展迅速，搜索引擎已经深刻地影响了我们的日常生活
- 信息提取：从非结构化或者半结构化的文本中提取结构化的信息，如命名实体、实体关系、时间信息、空间信息、事件信息、事件角色等等
- 自动文摘：近年来多文档文摘取得了长足的进步，可以从一批相关的文档中生成简单的摘要
- 自动问答：基于开放文本的自动问答同样取得了很大的进展，对于简单的知识性问题准确率可以达到**70-80%**，目前英语自动问答已经相对成熟，已经催生了像**PowerSet**这样的高技术公司
- 机器翻译：下面主要介绍机器翻译技术的进展

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

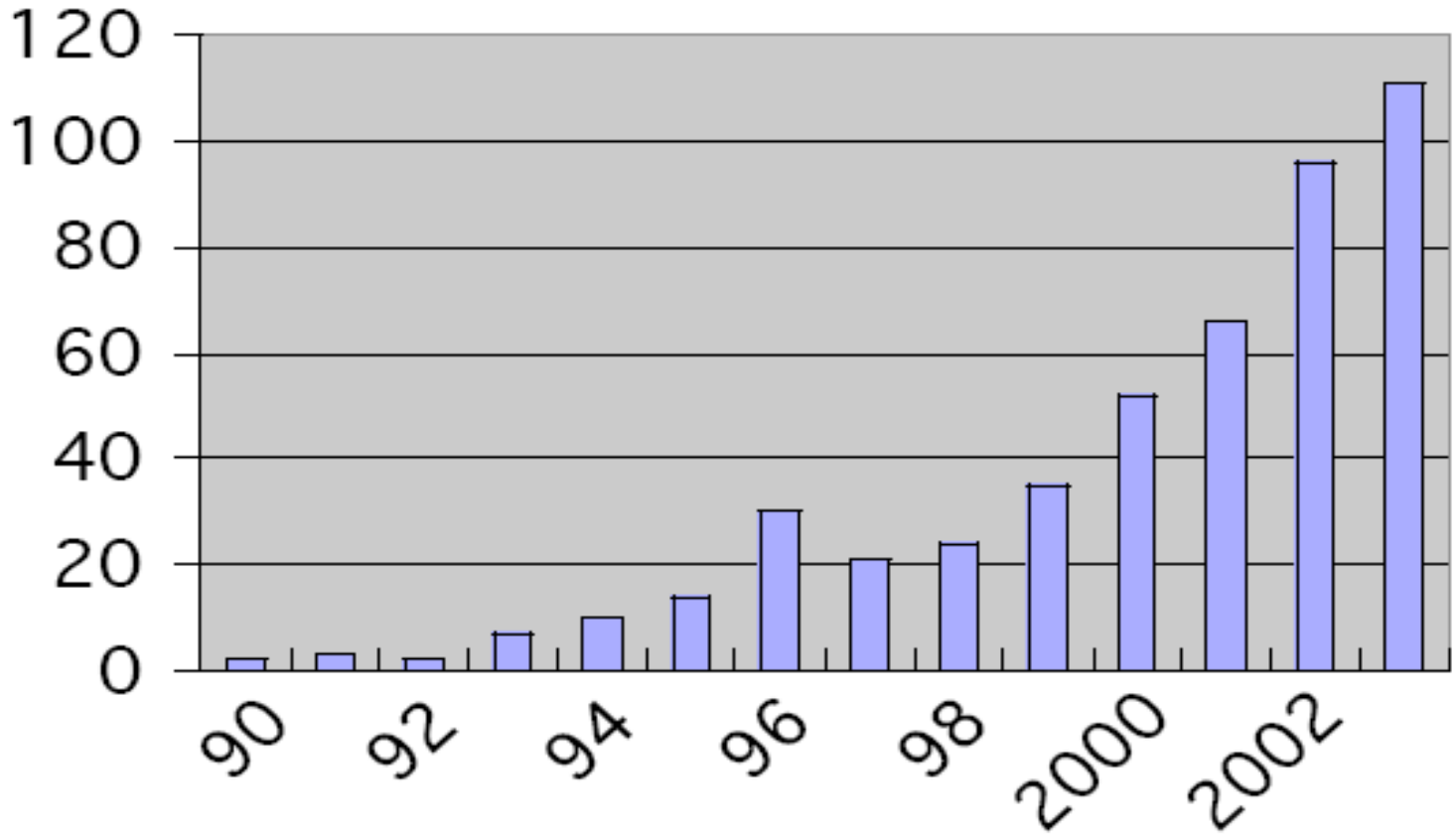
历史回顾：一些重要事件回放 (1)

- **1990**年代初**IBM**首次开展统计机器翻译研究
- **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
- **2002**年**Franz Josef Och**提出统计机器翻译的对数线性模型

历史回顾：一些重要事件回放 (2)

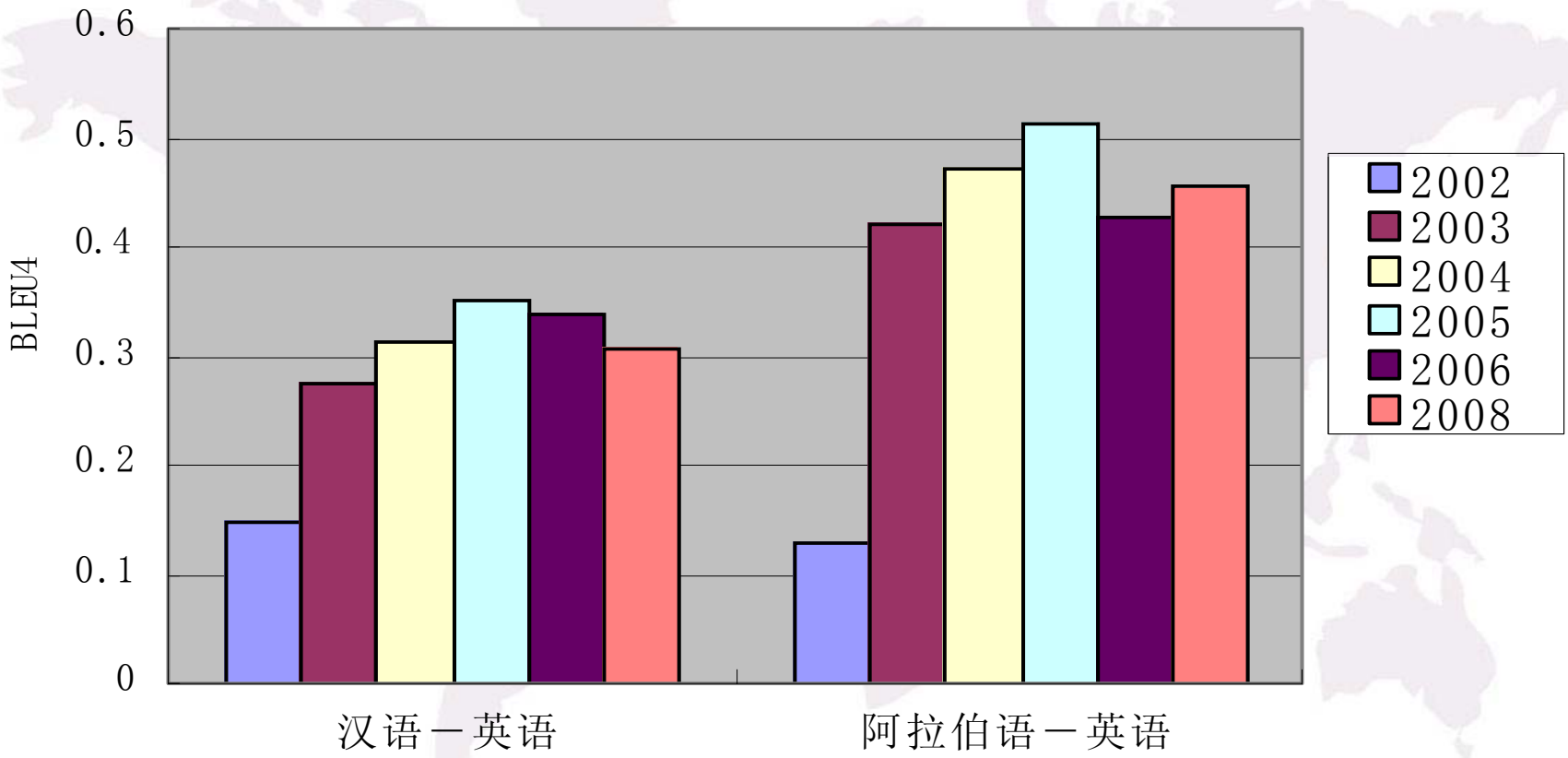
- **2003年Franz Josef Och**提出对数线性模型的最小错误率训练方法
- **2004年Philipp Koehn**推出**Pharaoh**（法老）标志着基于短语的统计翻译方法趋于成熟
- **2005年David Chiang**提出层次短语模型并代表**UMD**在**NIST**评测中取得好成绩
- **2005年Google**在**NIST**评测中大获全胜，随后**Google**推出基于统计方法的在线翻译工具，其阿拉伯语-英语的翻译达到了用户完全可接受的水平
- **2006年NIST**评测中**USC-ISI**的树到串句法模型第一次超过**Google**（仅在汉英受限翻译项目中）
- **2007年Google**推出采用统计机器翻译技术的跨语言检索网站

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



统计机器翻译目前的水平

- 以**Google Translator**为例，实地考察一下统计机器翻译的水平
  - 阿拉伯语—英语
  - 汉语—英语
  - 英语—汉语

Google Translator 阿拉伯语-英语

半岛电视台
网站新闻

阿拉伯语原文

الجمعة 30/ 9/ 1428 هـ- الموافق12/ 10/ 2007 م (آخر تحديث) الساعة 7:14 (مكة المكرمة)، 4:14 (غرينتش)

الصفحة الرئيسية: دولي

بوش يستقبل الدلاي لاما قبيل توشيح الكونغرس له

الدلاي لاما

يستقبل الرئيس الأميركي جورج بوش في البيت الأبيض الأسبوع المقبل الزعيم الروحي للبوذيين في التبت الدلاي لاما في خطوة من المرجح أن تزعج الصين.

وسيلتقي بوش الدلاي لاما، في إطار خاص بعيدا عن وسائل الإعلام، كما قال المتحدث باسم البيت الأبيض غوردون جوندرو، على غرار ما فعل في السابق.

وسيحضر بوش في اليوم التالي في واشنطن حفلا رسميا يقلد خلاله الكونغرس الدلاي لاما ميدالية الكونغرس الذهبية، وهي أعلى وسام يمكن للكونغرس أن يمنحه.

وحفل منح الوسام سيكون المرة الأولى التي يظهر فيها بوش علانية مع الدلاي لاما الذي سبق له أن زار البيت الأبيض لكن دائما في اجتماعات غير رسمية.

وردت الصين بغضب عندما قرر الكونغرس الأميركي منح الدلاي لاما الوسام وسجيت القرار قائلة إنه تدخل في شؤونها الداخلية.

وتعتبر الصين الدلاي لاما -الذي فر من التبت عام 1959 بعد انتفاضة فاشلة على السلطات الصينية- انفصاليا.

وتتهم الصين "بعض البلدان أو الأشخاص" باستغلال الدلاي لاما كما قال المتحدث باسم وزارة الخارجية الصينية ليو جيانشاو، قبل الإعلان عن اللقاء بين بوش والدلاي لاما.

وتؤكد الصين أنها حررت التبت من الظلم الإقطاعي لدى سيطرتها عليها عام 1949، قبل أن تقيم فيها منطقة تتمتع بالحكم الذاتي في 1965.

المصدر: وكالات

متعلقات

متعلقات من الأرشيف

- آلاف يحتشدون بنيويورك للاستماع للدلاي لاما
- مسؤولون أميركيون كبار يلتقون الدلاي لاما
- بكين تحتج على زيارة الدلاي لاما لواشنطن
- الصين تحتج على زيارة الدلاي لاما للبرتغال
- بكين تنتقد استضافة واشنطن للرئيس التايواني والدلاي لاما

أهم أخبار الصفحة الرئيسية

- الخرطوم تتهم جنوبيين بالتآمر وسلفاكير يدعو لتدخل دولي
- ملايين المسلمين يؤدون شعائر عيد الفطر اليوم
- العنف يتواصل والمارينز يضغطون لترك العراق
- عباس يلتقي ولش ويشترط تسلم غزة لحوار حماس
- أردوغان يهدد بإجراءات إضافية ضد واشنطن بشأن إبادة الأرمن

Google Translator 阿拉伯语-英语

半岛电视台网站新闻

Google 的英文译文

(Friday 30/9/1428 12/10/2007 e -approved m (Updated) at 7:14 (Mecca), 4:14 (GMT

متعلقات

متعلقات من الأرشيف

- Thousands gathered in New York to listen to Dlai Lama
- Senior American officials meet Dalai Lama
- Beijing protests the visit of the Dalai Lama to Washington
- China protests Dalai Lama's visit to Portugal
- Beijing to host criticize Washington for Taiwanese President and the Dalai Lama

أهم أخبار الصفحة الرئيسية

- Khartoum accuses of conspiring southerners and Slgakir calls for international intervention
- Millions of Muslims perform rites Eid Al-Fitr today

Internationa :Home

Bush received the Dalai Lama as a Touchih Congress

Receives American President George Bush at the White House next week, the spiritual leader of Buddhists in Tibet the Dalai Lama in .a step likely to upset China

And Bush will meet with the Dalai Lama, in the framework of special away from the media, as White House spokesman Gordon Jondro, as it .did in the past

Bush will attend the next day in Washington, a ceremony which officially imitate Congress Dalai Lama Congressional Gold Medal, th .highest decoration that can be granted by Congress

The granting of the medal ceremony will be the first time that Bush is displayed publicly with the Dalai Lama who has previously visited the .White House, but always in informal meetings

China replied angrily when the American Congress decided to grant the Dalai Lama medal and denounced the resolution, saying that .interference in their internal affairs

China considers the Dalai Lama - who fled from Tibet in 1959 after a .failed uprising on the Chinese authorities separatist

China accuses "certain countries or persons" to exploit the Dalai Lama as the spokesman of the Chinese Foreign Ministry Liu Jianchao, before the announcement of the meeting between Bush and .the Dalai Lama

The China Tibet liberated it from feudal injustice to the control by the .year 1949, prior to evaluate the autonomous regions in 1965

Agencies :Source

Google Translator 阿拉伯语-英语

- في لاما الدلاي التبت في ذيينللبو الروحي الزعيم المقبل أسبوعال أبيضال البيت في بوش جورج أميركيال الرئيس يستقبل خطوهة أن المرجح من ال تزعج صين.
  Receives American President George Bush at the White House next week, the spiritual leader of Buddhists in Tibet the Dalai Lama in a step likely to upset China.☺
- ماغرار على،جوندروغوردون أبيضال البيت باسم ثالمتحد قال كما إعلامال وسائل عن بعيدا ص خاإطار في ،لاما الدلاي بوش وسيلتقي السابق في فعل.
  And Bush will meet with the Dalai Lama, in the framework of special away from the media, as White House spokesman Gordon Jondro, as it did in the past. ☺
- وسام أعلى وهي ،ذهبيةال الكونغرس ةميدالي لاما الدلاي الكونغرس خلاله يقلد رسميا حفلا طنواشن في التالي اليوم في بوش ضروسيح بمنحه أن للكونغرس يمكن.
  Bush will attend the next day in Washington, a ceremony which officially imitate Congress Dalai Lama Congressional Gold Medal, the highest decoration that can be granted by Congress. ☺
- في دائما لكن أبيضال البيت زار أن له سبق ذيال لاما الدلاي مع ةعلاني بوش فيها ظهري التي أولىال ةالمر سيكون الوسام منح وحفل رسمي غير اجتماعات.
  The granting of the medal ceremony will be the first time that Bush is displayed publicly with the Dalai Lama who has previously visited the White House, but always in informal meetings. ☺
- وردت ال صين بغ ضب عندما قرر الكونغرس ال أميركي منح الدلاي لاما الوسام وشجبت القرار قائل ة إنه ش في تدخل وؤونها ة الداخلي.
  China replied angrily when the American Congress decided to grant the Dalai Lama medal and denounced the resolution, saying that interference in their internal affairs. ☺

Google Translator 汉语-英语

新浪新闻

中文原文

土耳其抗议美国会有关亚美尼亚大屠杀议案

http://www.sina.com.cn 2007年10月12日01:23 新京报

10月11日，土耳其，伊斯坦布尔，土耳其反对者举着旗帜和标语反对美“亚美尼亚”大屠杀议案。

点击观看本新闻视频

据亚美尼亚方面的史料记载，1915年至1923年期间，土耳其奥斯曼帝国对其统治的亚美尼亚人实施种族灭绝，导致150万亚美尼亚人死亡。

土耳其历届政府均对此予以否认，认为这是奥斯曼帝国崩溃过程中出现的非正常死亡。土耳其认为，那些人死于当时的内战和社会动荡，而且这一数字被夸大了。

美国国会众议院外交事务委员会定于10日表决通过一项关于“亚美尼亚大屠杀”的议案，并准备提交众院全体会议表决。

由于这项议案可能损害美国与重要盟友土耳其的关系，总统乔治·W·布什当天呼吁众院拒绝表决。土耳其已就这项议案向白宫提出抗议。

Google Translator 汉语-英语

新浪新闻

Google 的英文译文

Turkey, the United States will protest the Armenian massacre motion

Http://www.sina.com.cn 2007, 12 October 01:23 Xin Jing Bao

October 11, Turkey, Istanbul, Turkey, opponents holding banners and placards opposed to the "Armenian" massacre motion.

Click on the video to watch the news

According to the Armenian side of historical records, in 1915 to 1923, its rule of the Turkish Ottoman Empire implementation of the Armenian genocide, leading to the death of 1.5 million Armenians.

Turkey, successive governments have denied this, believing that this is the collapse of the Ottoman Empire appeared in the process of unnatural deaths. Turkey believes that those who died at that time of social unrest and civil war, but that figure has been exaggerated.

The U.S. House of Representatives Foreign Affairs Committee is scheduled to vote on the adoption of a on the 10th on the "Armenian Massacre," the motion and to be submitted to the House plenary vote.

As a result of this motion could damage the United States and Turkey, an important ally of the relationship between President George W. Bush appealed to the House of Representatives refused to vote on the same day. Turkey has been on the motion to the White House protest.

Google Translator 汉语-英语

- 土耳其历届政府均对此予以否认，认为这是奥斯曼帝国崩溃过程中出现的非正常死亡。
  Turkey, successive governments have denied this, believing that this is the collapse of the Ottoman Empire appeared in the process of unnatural deaths.（语序混乱） ☹☹
- 土耳其认为，那些人死于当时的内战和社会动荡，而且这一数字被夸大了。
  Turkey believes that those who died at that time of social unrest and civil war, but that figure has been exaggerated. ☹☺
- 美国国会众议院外交事务委员会定于10日表决通过一项关于“亚美尼亚大屠杀”的议案，并准备提交众院全体会议表决。
  The U.S. House of Representatives Foreign Affairs Committee is scheduled to vote on the adoption of a on the 10th on the "Armenian Massacre," the motion and to be submitted to the House plenary vote. ☺☺

Google Translator 英语-汉语

CNN新闻

英文原文

updated 3 hours, 34 minutes ago

EMAIL SAVE PRINT

Turkey recalls ambassador over genocide resolution

STORY HIGHLIGHTS

- NEW: Top Dem on defense says resolution could hinder redeployment from Iraq
- Turkish Ambassador Nabi Sensoy will head home after U.S. House panel vote
- Vote calls killing of Armenians during World War I genocide
- Condoleezza Rice set to call Turkish leaders to express "deep disappointment"

Next Article in Politics »

READ VIDEO MAP

TEXT SIZE

WASHINGTON (CNN) -- Turkey on Thursday recalled its ambassador to the United States and warned of repercussions in a growing dispute over congressional efforts to label the World War I era killings of Armenians by Ottoman Turkish forces "genocide."

Members of the Workers Party protest the U.S. House resolution Thursday in Istanbul, Turkey.

The U.S. House Committee on Foreign Affairs passed the measure 27-21 Wednesday. President Bush and key administration figures lobbied hard against the measure, saying it would create unnecessary headaches for U.S. relations with Turkey.

Turkey -- now a NATO member and a key U.S. ally in the war on terror -- accepts Armenians were killed but call it a massacre during a chaotic time, not an organized campaign of genocide.

The full House could vote on the genocide resolution as early as Friday. A top Turkish official warned Thursday that consequences "won't be pleasant" if the full House approves the resolution.

"Yesterday some in Congress wanted to play hardball," said Egemen Bagis, foreign policy adviser to Turkish Prime Minister Recep Tayyip Erdogan. "I can assure you Turkey knows how to play hardball."

Asked about Ambassador Nabi Sensoy's recall after the news broke, a State Department spokesman said he could not confirm it. "People are sometimes called back for consultation; sometimes they're called back for other reasons," said spokesman Tom Casey.

"If they wanted to bring their ambassador back for consultations or do something else, that is their decision. I certainly think that it will not do anything to limit our efforts to continue to reach out to Turkish officials, to explain our views, to engage them on this issue and again to make clear that we intend to work on this with Congress."

被过滤广告

Most Popular

STORIES

Most Viewed | Most Emailed | Top Searches

1. Paparazzi 'snapped Diana in crash'
2. Turkey recalls ambassador
3. Raid finds U.S. soldiers' weapons
4. Writer suspect in dismembering...
5. Madonna to sign $120M record deal
6. Plane fire after emergency landing
7. Raid finds U.S. soldiers' weapons
8. Report: Myanmar prisoners abused
9. Iraqi families sue Blackwater
10. Spears wins expanded visitation

VIDEOS

TOPICS

We Recommend

Stories you may be interested in based on past browsing

Google Translator 英语-汉语

CNN新闻

Google的
中文译文

updated 3 hours, 34 minutes ago

EMAIL SAVE PRINT

土耳其驻华大使回忆种族灭绝决议

故事聚焦

- **新：顶层**数字高程模型对国防部说，决议案可能会阻碍调配从伊拉克撤军
- 土耳其驻华大使纳比申索伊会回家后，美国众议院小组投票
- 投票呼吁杀害亚美尼亚人在第一次世界大战种族灭绝
- 赖斯定致电土耳其领导人表示"极度失望"

明年第从政»

阅读 视频 地图

TEXT SIZE

华盛顿（美国） -土耳其对周四召回其驻美国大使，并警告反响，在越来越多的争议，美国国会努力标签第一次世界大战时期屠杀亚美尼亚人，由奥斯曼土耳其军队的"种族灭绝"

委员对工人党为抗议美国众议院决议案周四在土耳其的伊斯坦布尔。

美国众议院外交委员会通过这项措施队医周三。 布什总统及主要政府人物努力游说反对这项措施，认为这将有助于制造不必要的烦恼，为美国与土耳其的关系。

土耳其-现在是北约成员国和美国的重要盟友，在反恐战争-接纳亚美尼亚人被杀害，但它称作是屠杀，在一个混乱的时候，没有一个有组织的运动，灭绝种族罪的人。

充分众议院可以投票就种族灭绝决议早在周五。 顶级土耳其官员警告说，周四的后果"不会愉快"，如果充分众院通过这项决议案。

"昨天，在国会要扮演强硬，说： "埃盖曼巴厄什，外交政策顾问的土耳其总理埃尔多安。 "我可以向你保证，土耳其也知道如何发挥强硬"

问大使彩蝶申索伊召回的消息传出后，打破了，美国国务院发言人表示，他无法证实这一消息。 "的人有时被称为回谘询公众;有时，他们正在召回其他原因，说： "发言人汤姆凯西。

"如果他们希望把他们的大使回国磋商或去做别的事，那是他们自己的决定。当然，我认为它不会做任何限制，我们的努力，以继续接触，以土耳其官员，以解释我们的看法，从事他们对这个问题，并再次明确表示，我们打算在这方面的工作与国会合作。 "

Most Popular

STORIES

最受欢迎	大多数电子邮件	搜索排名
一	海仔队'室慕安娜在坠机'	
	召回大使	
三	空袭认定美军士兵的武器	
四	作家嫌疑人在肢解... ...	
五	麦当娜签署$ 1.2记录处理	
六	架飞机火警后紧急降落	
7日	空袭认定美军士兵的武器	
	：缅甸囚犯受虐待	
9日	，伊拉克家屋控告黑水	
十	布兰妮赢得扩大探视	

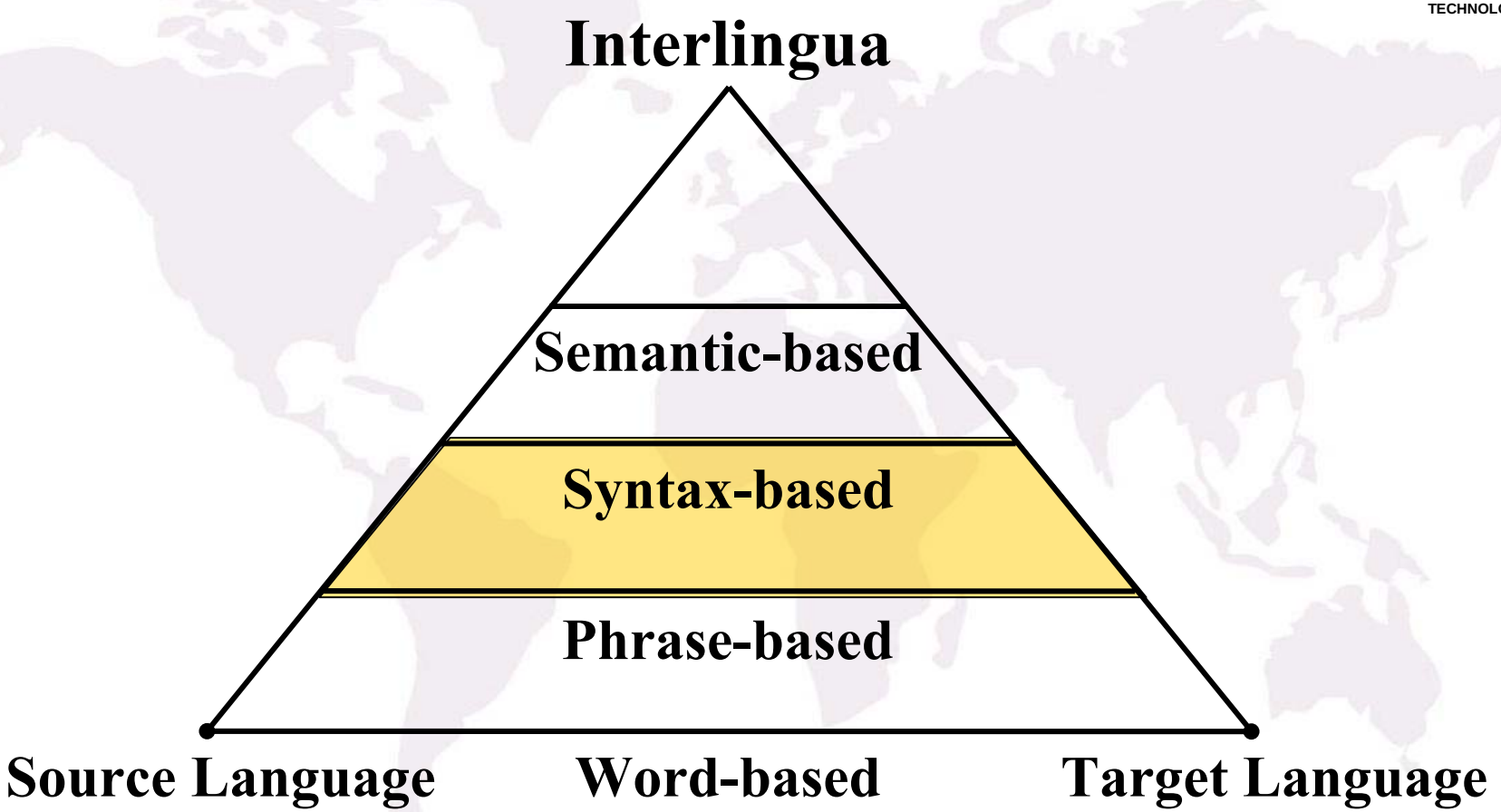
VIDEOS

TOPICS

Google Translator 英语-汉语

- **Turkey recalls ambassador over genocide resolution**
  土耳其驻华大使回忆种族灭绝决议 ☹ ☹
- **Members of the Workers Party protest the U.S. House resolution Thursday in Istanbul, Turkey.**
  委员对工人党为抗议美国众议院决议案周四在土耳其的伊斯坦布尔。 ☹ ☺
- **The U.S. House Committee on Foreign Affairs passed the measure 27-21 Wednesday.**
  美国众议院外交委员会通过这项措施队医周三。 ☺ ☺
- **President Bush and key administration figures lobbied hard against the measure, saying it would create unnecessary headaches for U.S. relations with Turkey.**
  布什总统及主要政府人物努力游说反对这项措施，认为这将有助于制造不必要的烦恼，为美国与土耳其的关系。 ☺ ☺
- **Turkey -- now a NATO member and a key U.S. ally in the war on terror -- accepts Armenians were killed but call it a massacre during a chaotic time, not an organized campaign of genocide.** ☹ ☹
  土耳其-现在是北约成员国和美国的重要盟友，在反恐战争-接纳亚美尼亚人被杀害，但它称作是屠杀，在一个混乱的时候，没有一个有组织的运动，灭绝种族罪的人。 （语序混乱）

统计翻译模型的进展


Interlingua
Semantic-based
Syntax-based
Phrase-based
Source Language
Word-based
Target Language

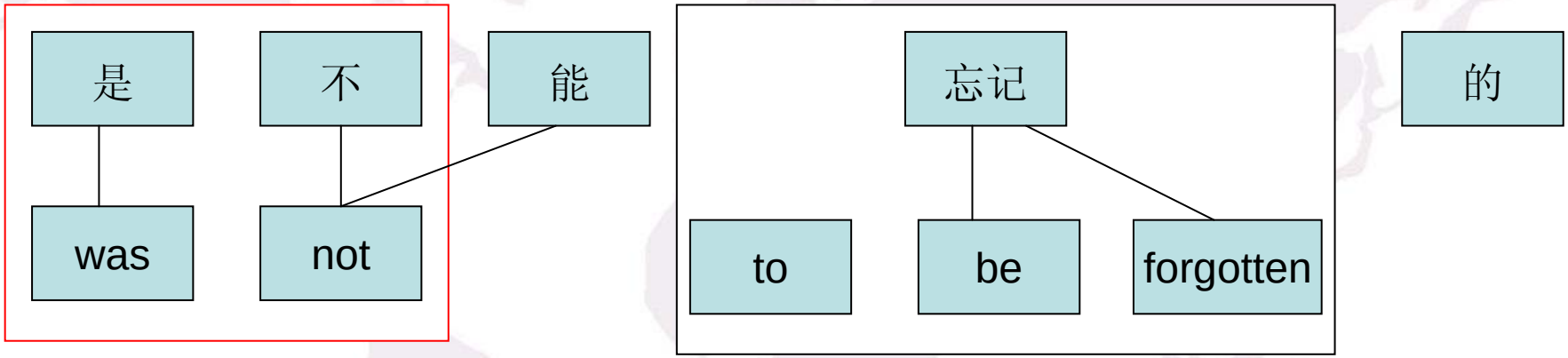
统计翻译模型的进展

- 目前，基于短语的模型是最成熟、也是最稳定的模型
- 利用开源的统计机器翻译模型工具（如摩西**Moses**），在足够大的语料库上，经过简单训练，即可以达到比较满意的效果
- 基于句法的统计机器翻译模型是目前的研究热点，一些方法已经超过的基于短语的模型

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

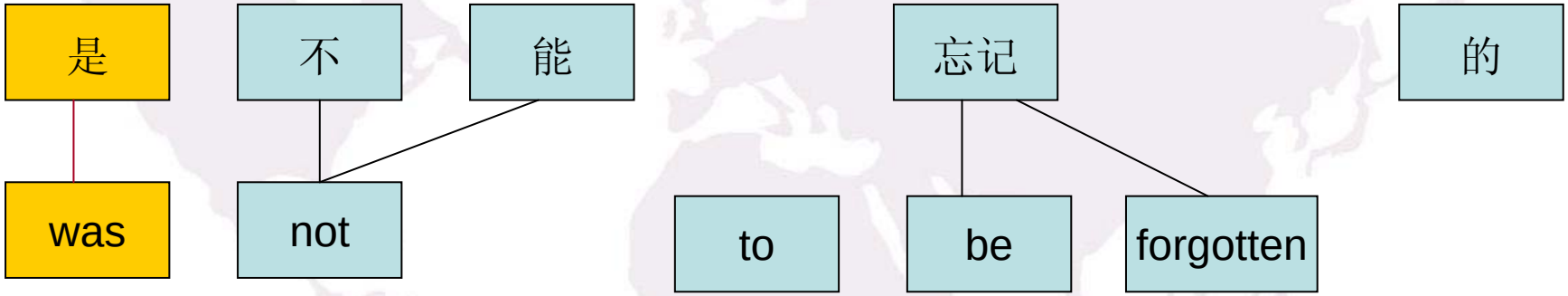


不相容

相容

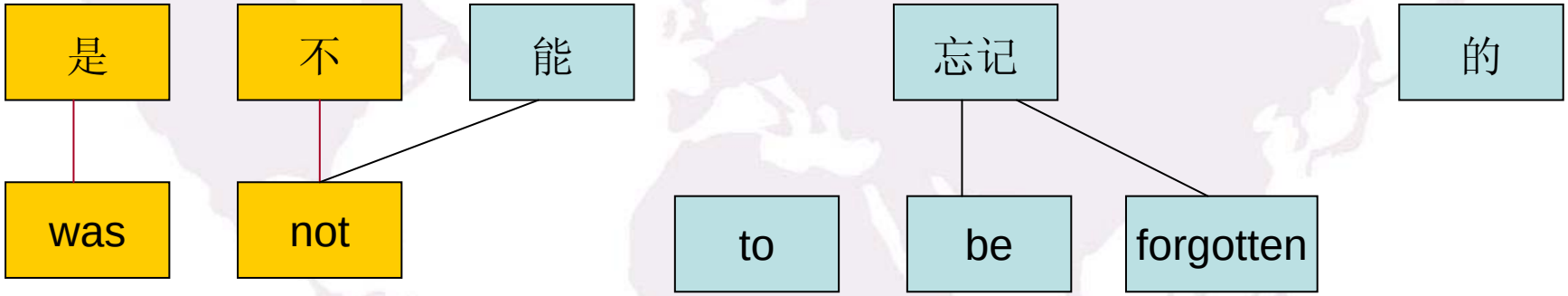
短语自动抽取算法运行示例 (1)

- 列举源语言所有可能的短语，根据对齐检查相容性



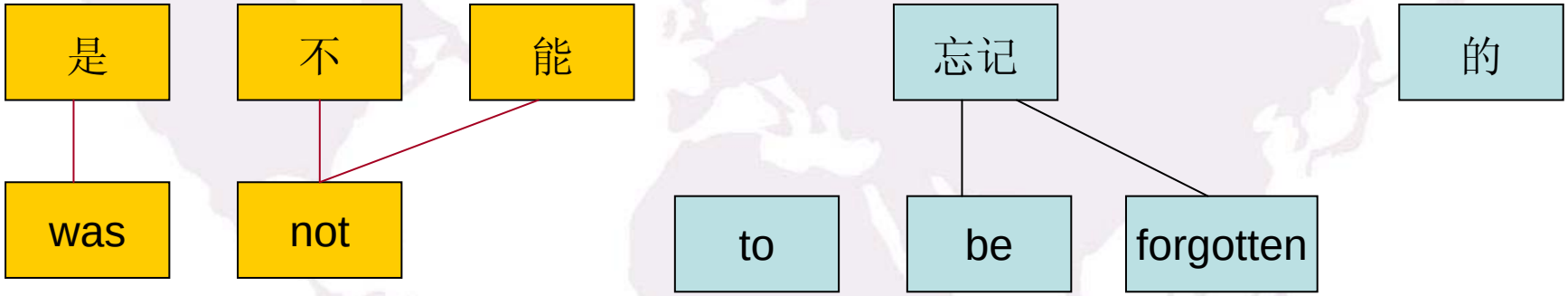
（是，was）

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)

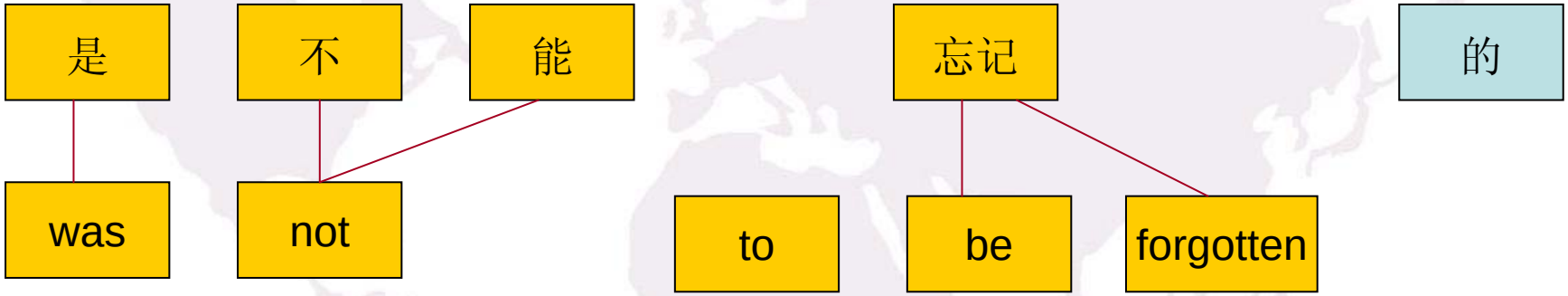


（是不能，was not)

短语自动抽取算法运行示例(4)

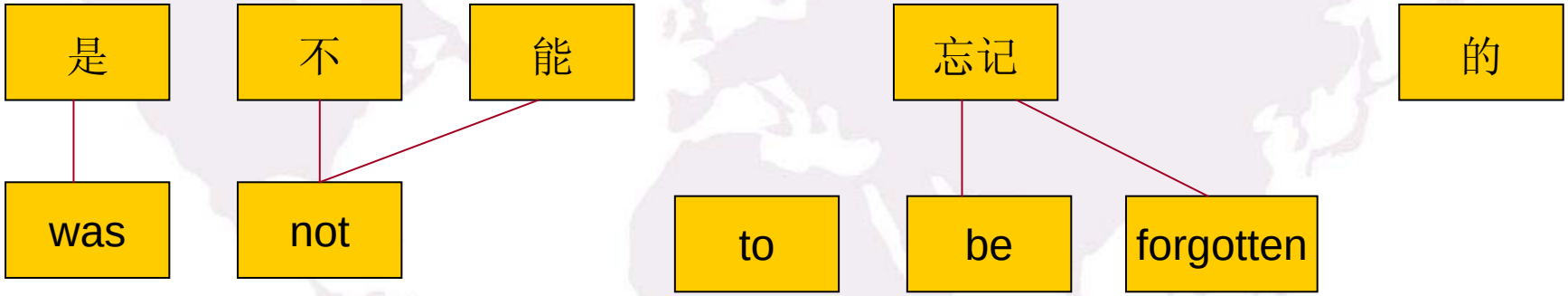
（是不能，was not to)

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

短语翻译概率表

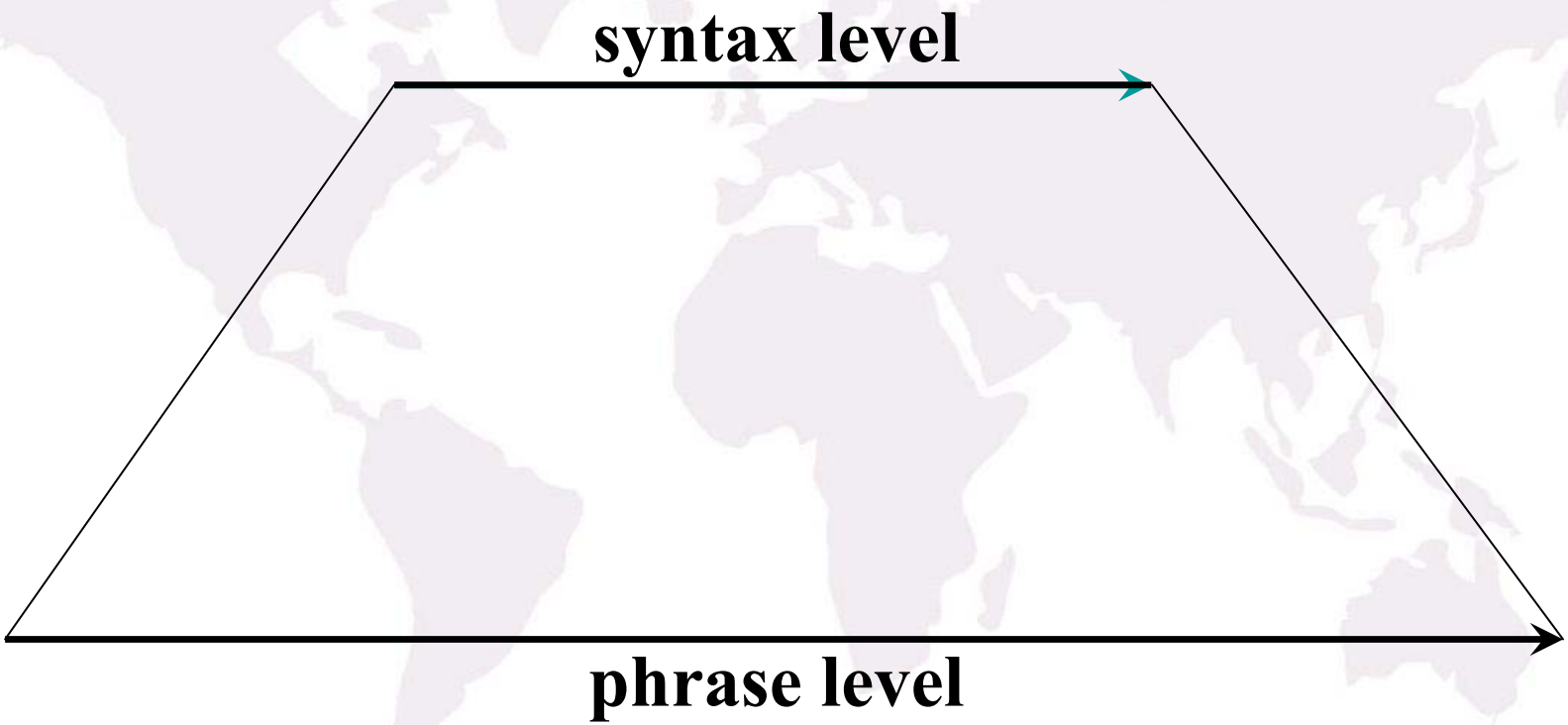
f e $p(f|e)$ $lex(f|e)$ $p(e|f)$ $lex(e|f)$

没有 达成 共识 ||| no consensus was reached ||| 1 0.00210153 1 8.87474e-05 2.718
没有 达成 共识 。 ||| no consensus was reached . ||| 1 0.0017517 1 8.83361e-05 2.718
没有 得到 澄清 ||| clarified ||| 1 0.000592593 1 0.036396 2.718
没有 得到 南方 的 响应 ||| no response ||| 0.5 1.49065e-06 1 0.00921419 2.718
没有 得到 证实 ||| no evidence ||| 0.5 0.000178961 1 0.0021538 2.718
没有 兑现 ||| has sent ||| 0.2 6.64599e-05 1 0.00346412 2.718
没有 发生 变化 ||| had not changed ||| 0.5 0.000141333 1 8.84361e-05 2.718
没有 发现 明显 ||| is no obvious ||| 1 0.00114645 1 0.000308419 2.718
没有 犯罪 ||| no criminal ||| 1 0.0613205 1 0.0251376 2.718
没有 犯罪 纪录 ||| no criminal record ||| 1 0.0196688 1 0.0123866 2.718
没有 放弃 ||| has not given up its ||| 1 0.000278368 1 8.34878e-06 2.718
没有 改变 ||| There is no change ||| 0.5 0.0148622 0.333333 1.50262e-05 2.718
没有 改变 ||| has not changed ||| 0.5 0.00505152 0.333333 0.00145408 2.718
没有 改变 ||| is no change ||| 1 0.0283586 0.333333 0.00201351 2.718
没有 改变 。 ||| is no change . ||| 1 0.0236378 1 0.00200418 2.718
没有 改变 ， ||| has not changed , and ||| 1 0.000628986 1 8.72846e-05 2.718
没有 改变 ， 如果 ||| has not changed , and if ||| 1 0.000308107 1 5.71048e-05 2.718
没有 工作 ||| without work ||| 1 0.0559111 1 0.0130721 2.718
没有 工作 许可证 ||| without work permits ||| 1 0.00559111 1 0.000344003 2.718
没有 归还 ||| not repaid till now ||| 1 0.00498227 1 6.22208e-05 2.718
没有 和平 ||| without a peaceful ||| 1 0.0398149 1 7.62298e-06 2.718

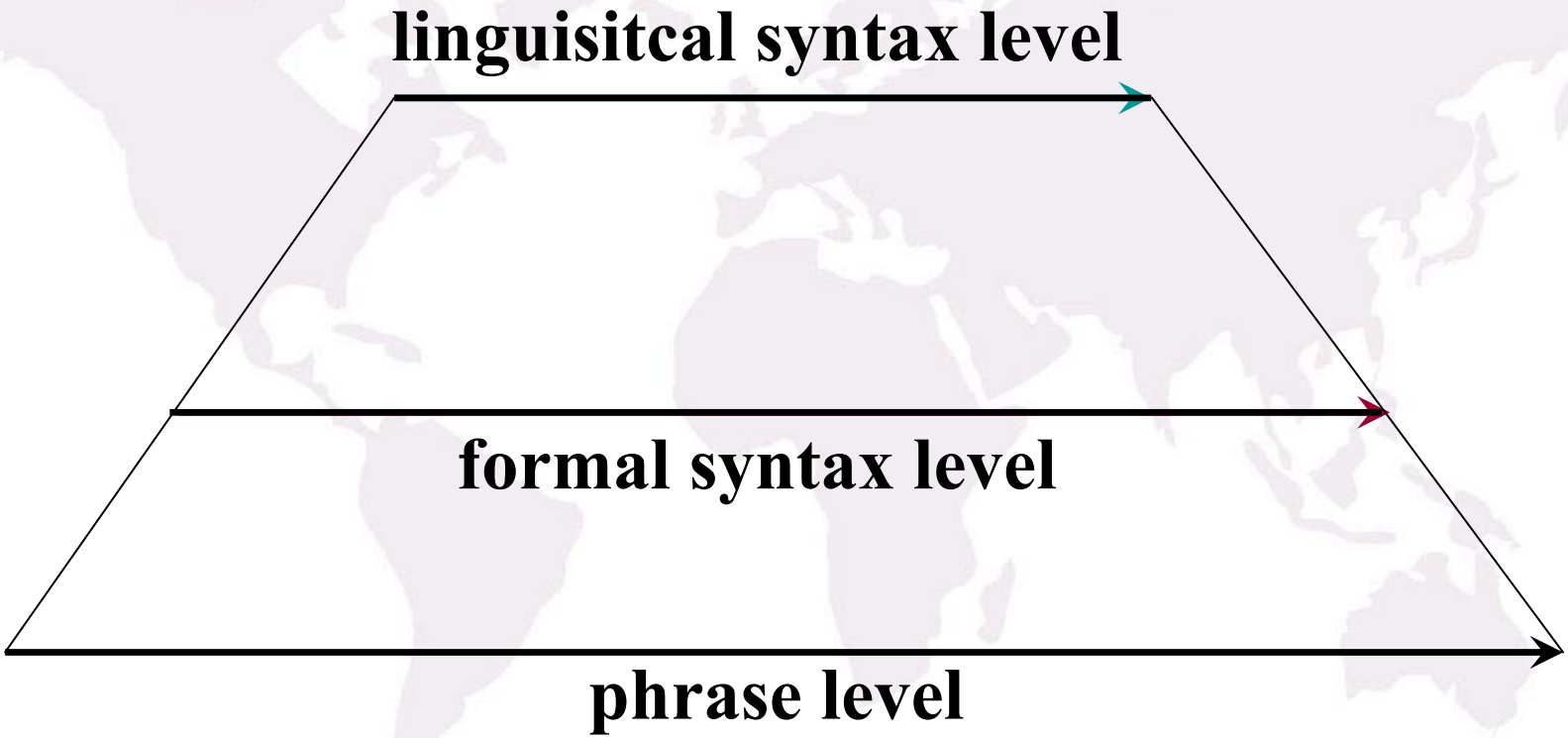
我们的贡献 —— 统计翻译建模

- 基于句法的统计模型是目前机器翻译的研究热点
- 我们提出了统计机器翻译的树到串模型，是目前统计机器翻译研究中的主要几个句法模型之一

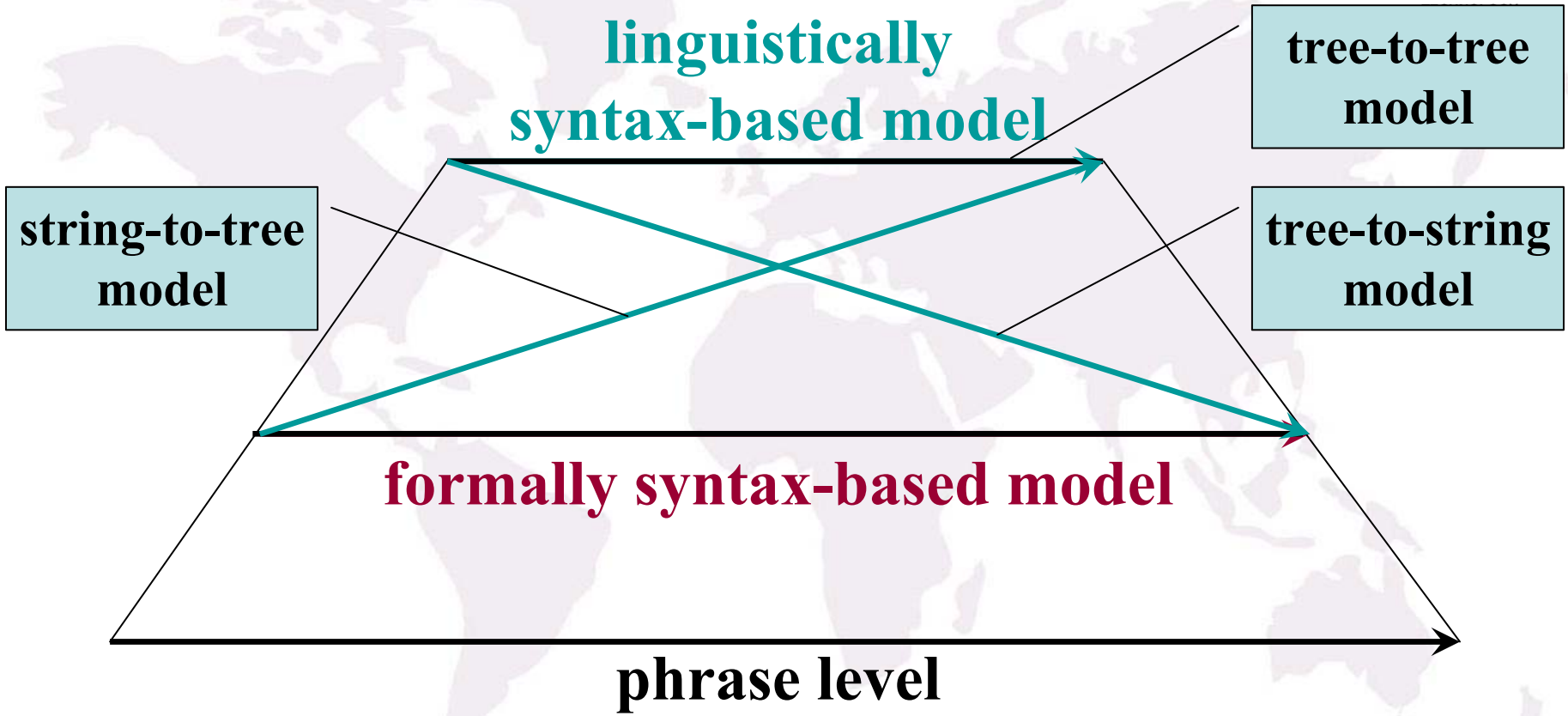
基于句法的统计翻译模型



基于句法的统计翻译模型



基于句法的统计翻译模型



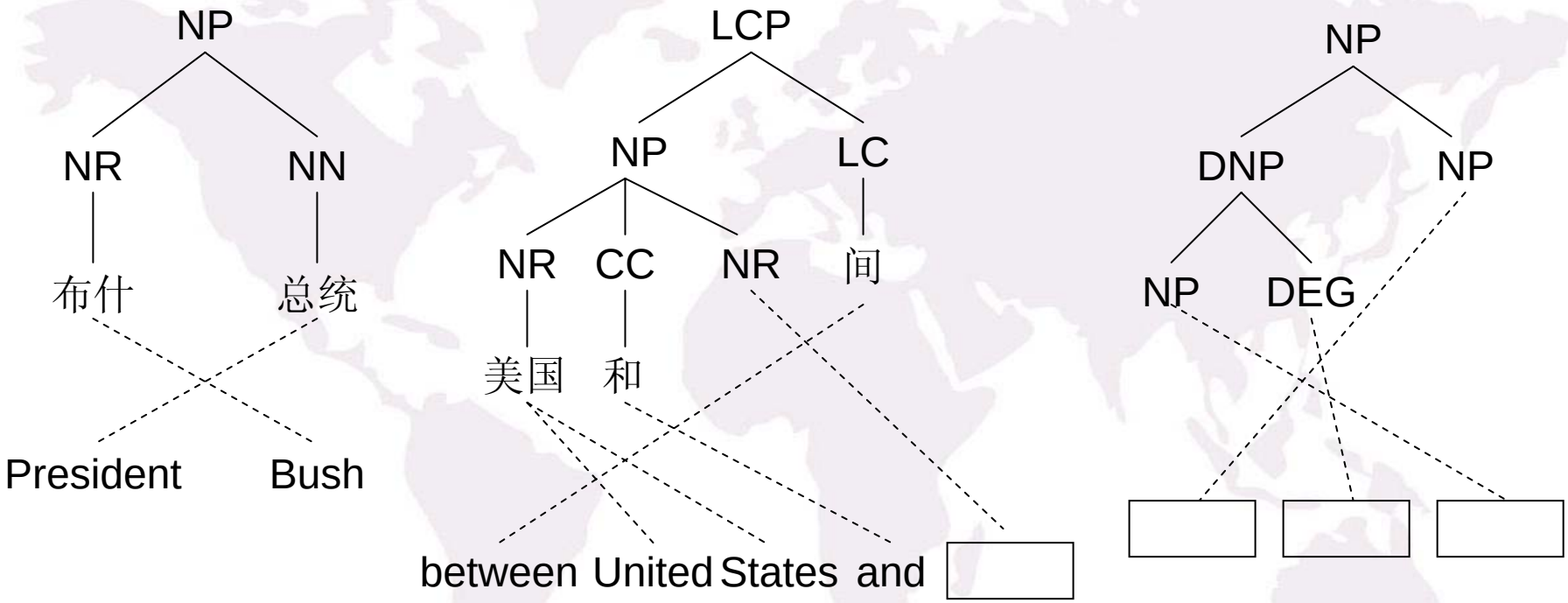
树到串统计翻译模型

我们在计算语言学领域的国际顶级会议上发表了一系列关于树到串统计翻译模型的论文

- **LIU Yang, et al., ACL2006**
- **LIU Yang, et al., ACL2007**
- **MI Haitao, et al., ACL2008**
- **MI Haitao & HUANG Liang, EMNLP2008**
- **LIU Qun, et al., EMNLP2008**

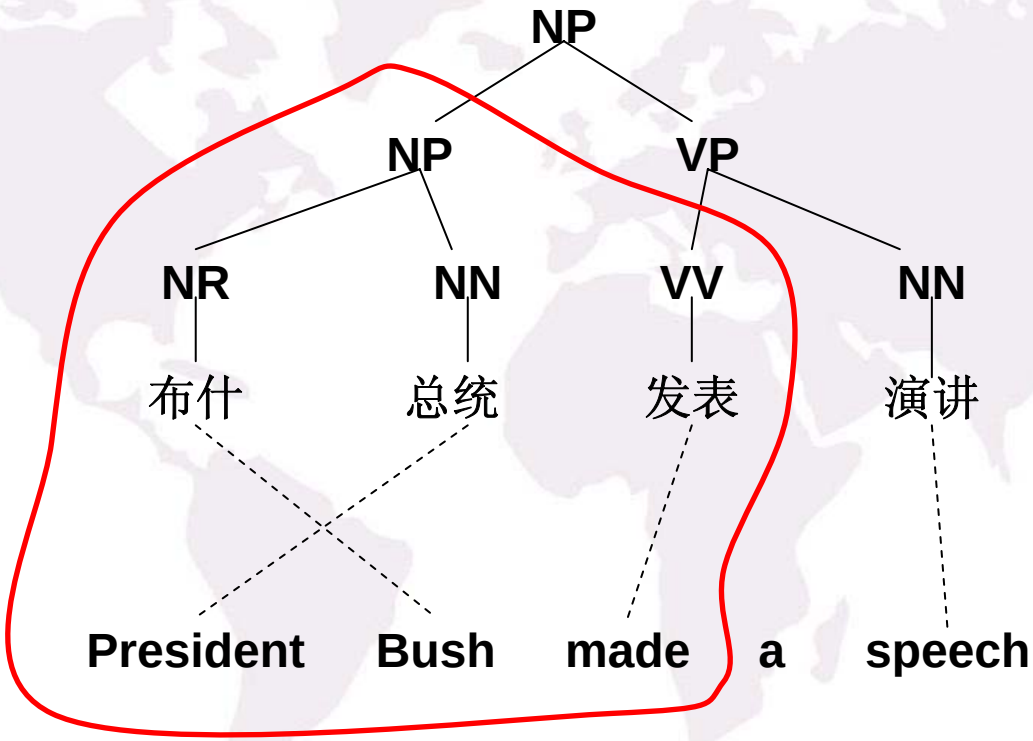
基于树到串对齐模板的统计翻译模型

LIU Yang, et al., ACL2006



加入树序列到串规则的树到串模型

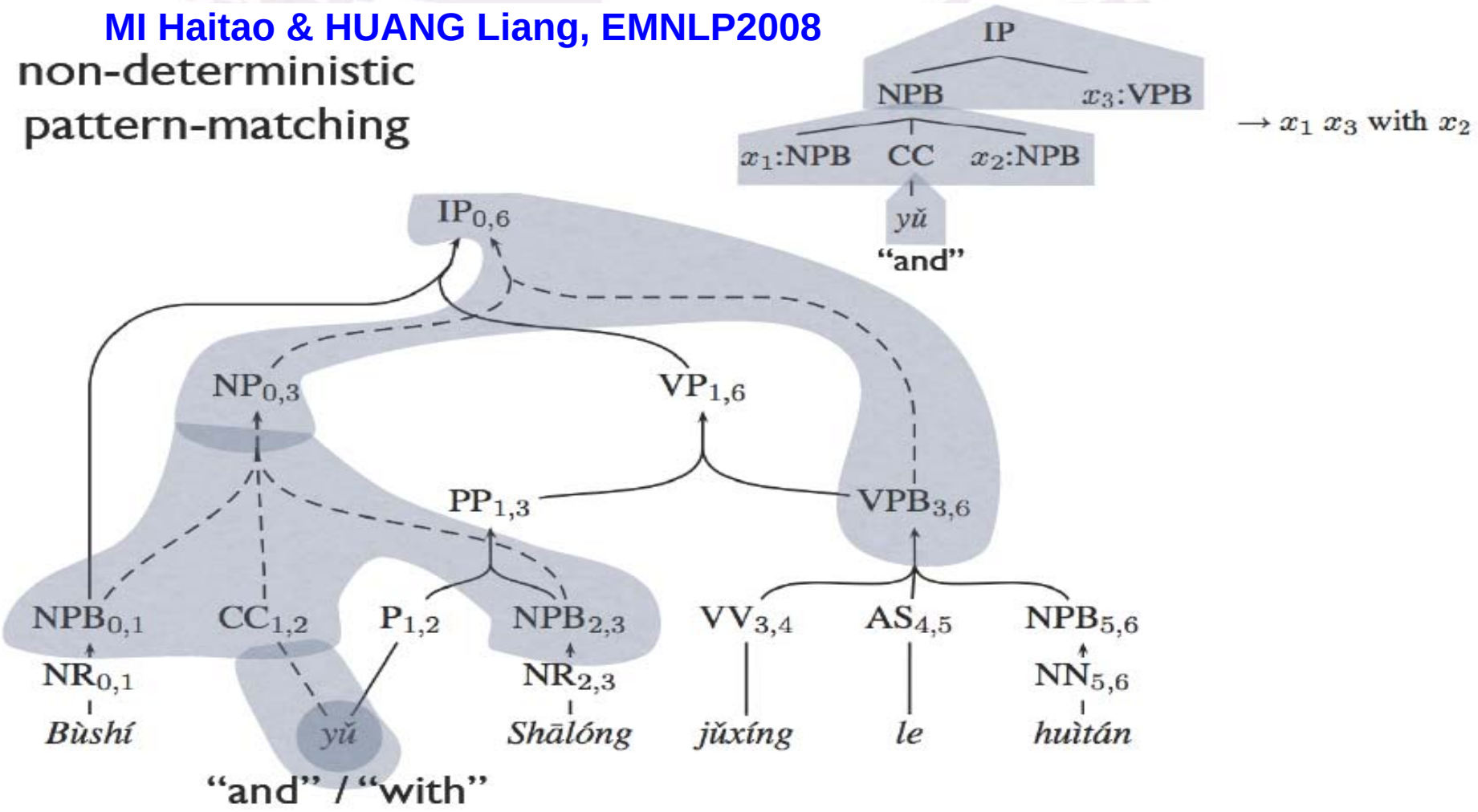
LIU Yang, et al., ACL2007



基于句法森林的统计翻译模型

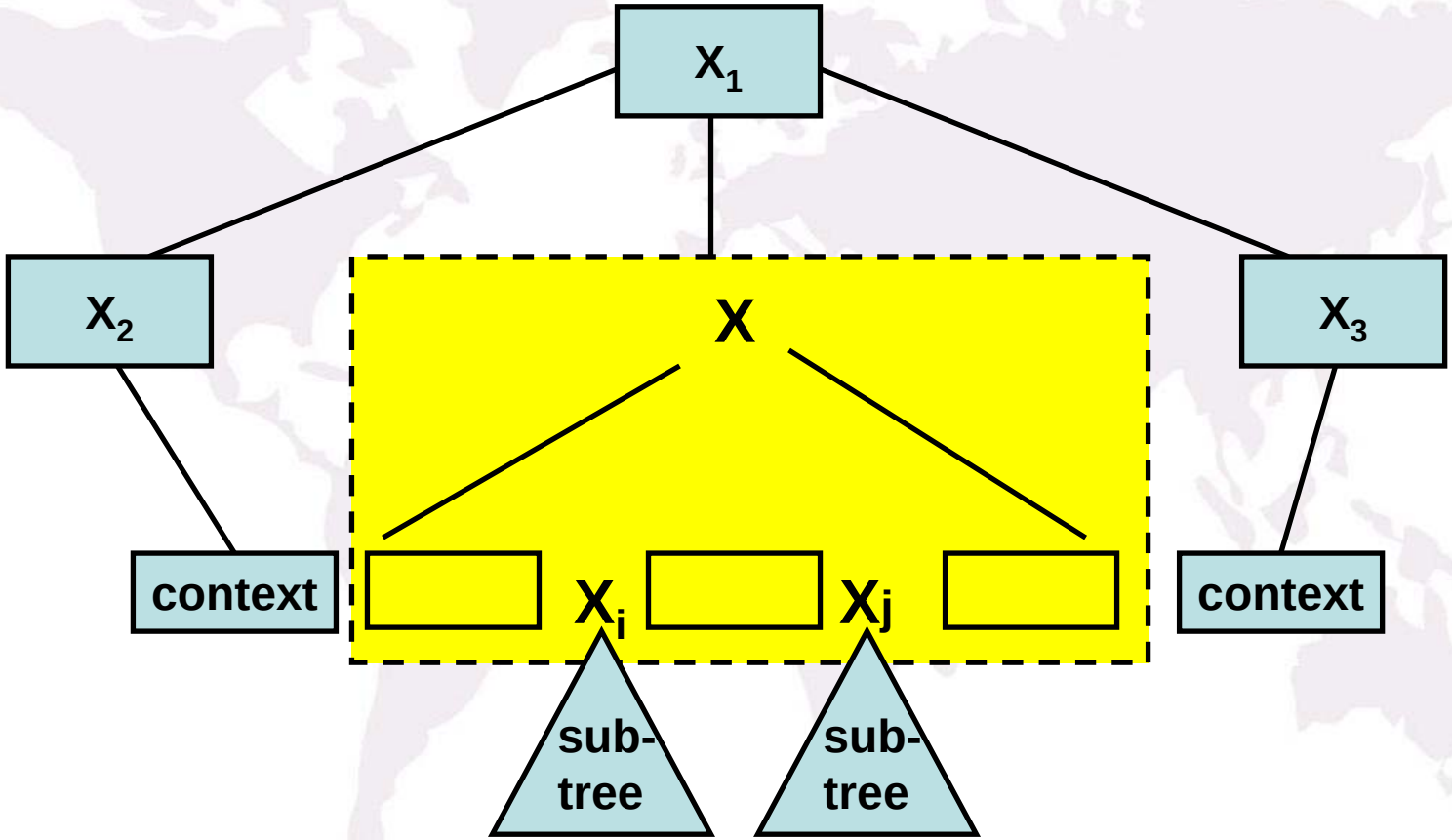
MI Haitao, et al., ACL2008

MI Haitao & HUANG Liang, EMNLP2008



基于最大熵的规则选择模型

LIU Qun, et al., EMNLP2008



本领域顶级会议上发表的论文

- ACL2005：[Liu Yang]
- ACL-COLING2006：[Liu Yang] [Xiong Deyi]
- ACL2007：[Liu Yang]
- EMNLP2007：[Lv Yajuan]
- ACL2008：[Mi Haitao] [Jiang Wenbin] [He Zhongjun]
- COLING2008：[Jiang Wenbin] [He Zhong]
- EMNLP2008：[Mi Haitao] [Liu Qun]

参加机器翻译国际评测结果

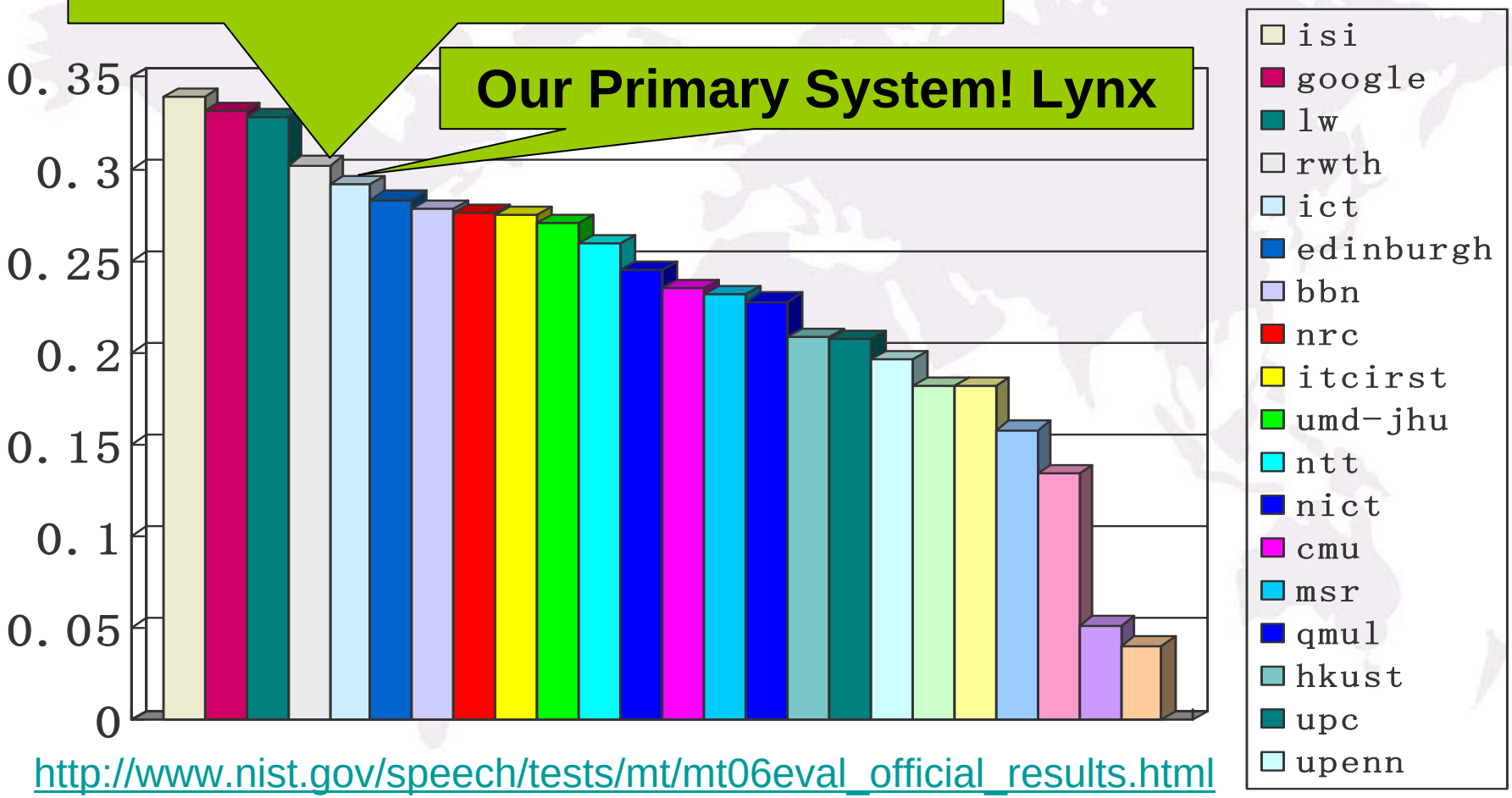
- **NIST**机器翻译评测是国际上影响最大也是竞争最激烈的机器翻译评测
- 我们在**NIST2006**的**24**个参评单位中排名第**5**

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

INSTITUTE OF COMPUTING TECHNOLOGY

One of our contrast system! Bruin

Our Primary System! Lynx



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

我们的工作 —— 应用

- 面向专利文献的机器翻译
- 面向移动设备的机器翻译

面向专利文献的机器翻译

- 计算所与东方灵盾公司合作
- 面向专利翻译的计算机辅助翻译平台
- 概况
  - 服务器-客户端结构
  - 基于短语模型的统计翻译系统
  - 八个领域，数百万句子对
  - 数千用户自定义模板
  - 用户提供的术语词典
  - 用户评价：准确率**75-85%**

面向专利文献的机器翻译

- 技术特点
  - 采用成熟的基于短语的统计机器翻译方法
  - 允许用户自定义翻译模板以改进系统，克服了统计机器翻译系统在应用中无法进行用户干预的缺点
  - 允许用户自定义多级词典、自行维护记忆库
- 优点
  - 开发周期短
  - 易于移植到不同领域
  - 易于移植到不同语种
  - 学习功能强大、越用越好用
  - 使用灵活，用户可以方便地定制系统、改进系统

专利翻译实例

- 一种治疗骨关节疾病的中药纸片
- A Chinese medicinal sheet for the treatment of bone and joint diseases

- 戒毒脱瘾药物及其配制方法
- Medicine for treating drug addiction and its preparation method

- 一种降脂保健饮料及其制备方法
- A health beverage capable of reducing blood lipid and its preparation method

- 该保健食品具有增食欲，壮肾阳、润肠、定喘、抗衰老作用;
- The health food has effects in stimulating appetite, supporting kidney yang, moisturizing the intestine, relieving asthma, and antiaging;

- （I）具有以下特性：（1）带有苦味的、微红的咖啡色的无味粉末；（2）溶于乙醇、二甲基亚砜、甲醇和丙酮，不溶于水、氯仿、乙酸乙酯、醚和丙醇；（3）不含重元素如铅、锌和砷。
- (I) have the following characteristics: (1) an acrid, or red brown odorless powder; (2) dissolved in ethanol, dimethyl sulfoxide, methanol and acetone, and insoluble in water, chloroform, ethyl acetate, ether and propanol; (3) free of heavy element such as lead, zinc and arsenic.

面向移动设备的机器翻译

- 计算所与某跨国公司合作
- 开发应用于移动翻译设备的机器翻译系统
- 旅游领域
- 中、英、韩三国语言的口语翻译
- 计算所提供核心文本翻译引擎以及汉英英汉机器翻译系统

面向移动设备的机器翻译

- 技术难点：通常统计机器翻译需要消耗大量的计算资源，而移动设备的机器翻译只能运行在很小的内存（**100M**）上，而且**CPU**性能低一个数量级以上
- 解决办法：我们探索并综合采用了多种优化算法，终于在性能没有太大降低的情况下，将系统优化到可以在给定的移动设备上运行，解决了该项目面临的最大的困难

目录

总结与展望

- 近十几年来，统计方法成为了自然语言处理研究的主流
- 近年来，统计方法有与规则方法融合的趋势，统计模型更加复杂，可以将一些复杂的语言学知识（如句法知识等）融入到统计模型中，克服了早期统计模型无法处理长距离依赖问题的缺陷
- 统计模型的深入研究，更加需要合适的语言学理论的指导，需要大规模的适用于自然语言处理的语言资源的支持

总结与展望

- 中文词语切分技术已经比较成熟，但中文切分的领域自适应技术、命名实体识别技术、多粒度切分技术等都还有深入研究的必要
- 中文的句法分析和语义角色标注距离英语都还有较大的差距，主要的困难在于中文缺乏明确的形态标记导致的汉语句法分析的困难，句法语义一体化分析也许是下一步值得努力的方向
- 自然语言处理的应用技术，包括信息检索、信息提取、机器翻译、自动文摘、自动问答等等近年来也取得了非常大的进展，一些技术已经走向实用，甚至极大地影响了我们的生活（如信息检索）

总结与展望

- **Internet**技术的普及已经世界经济社会一体化的潮流对自然语言处理技术的迫切需求，为自然语言处理研究提供了无尽的动力
- 各种新理论、新方法、新模型的出现也为自然语言处理研究带来了日新月异的崭新面貌
- 可以预期，自然语言处理还将处在一个比较长时期的快速发展的轨道上，理论上的突破将给我们带来更多的惊喜，而在应用上也将为满足我们的国家和社会的需求做出更大的贡献

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

谢谢！