ICCC2007特邀讲座

统计机器翻译研究的新进展

刘群

liuqun@ict.ac.cn

2007.10.13

目录

- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
  —基于词的IBM模型
- 最成熟的统计机器翻译方法
  —基于短语的模型
- 目前统计机器翻译研究的热点
  —基于句法的模型
- 统计机器翻译面临的问题和展望

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

统计机器翻译的研究热潮

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 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)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 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)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

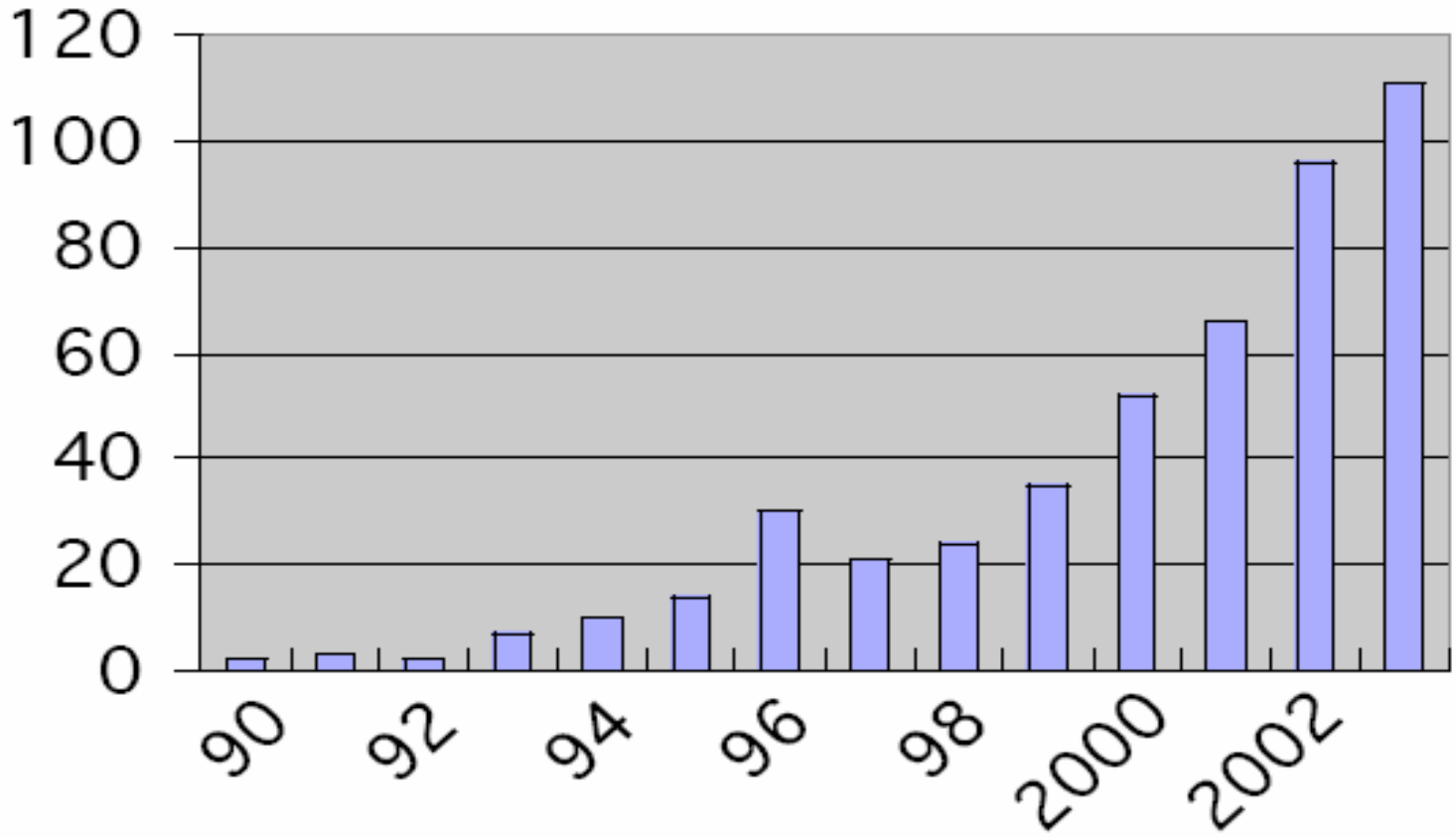
- 2003年Franz Josef Och提出对数线性模型的最小错误率训练方法
- 2004年Philipp Koehn推出Pharaoh（法老）标志着基于短语的统计翻译方法趋于成熟
- 2005年David Chiang提出层次短语模型并代表UMD在NIST评测中取得好成绩
- 2005年Google在NIST评测中大获全胜，随后Google推出基于统计方法的在线翻译工具，其阿拉伯语-英语的翻译达到了用户完全可接受的水平
- 2006年NIST评测中USC-ISI的树到串句法模型第一次超过Google（仅在汉英受限翻译项目中）
- 2007年Google推出采用统计机器翻译技术的跨语言检索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

统计机器翻译：一种新的研究范式

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

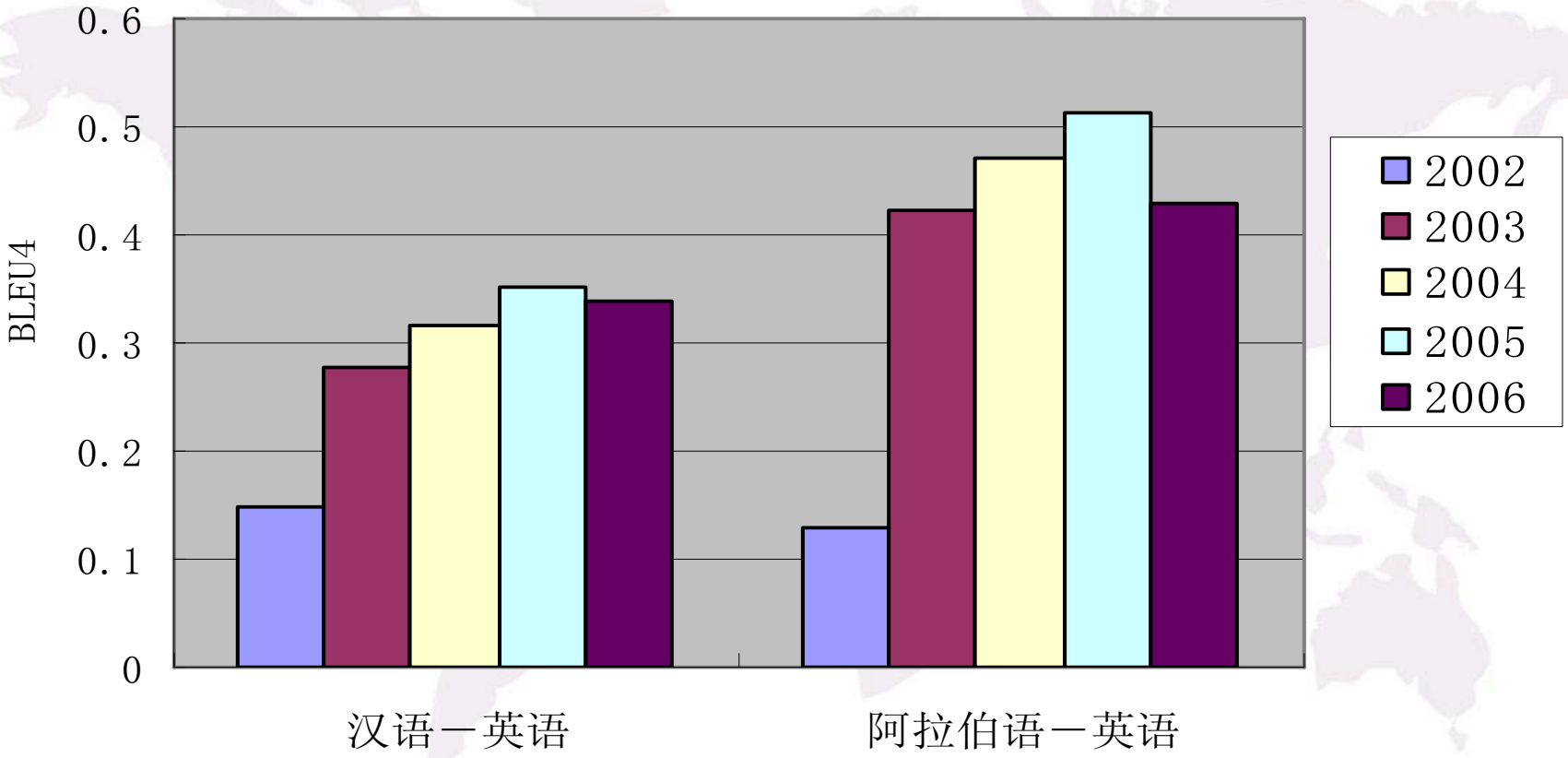
- 统计机器翻译的成功在于采用了一种新的研究范式（**paradigm**）
- 这种研究范式已在语音识别等领域中被证明是一种成功的翻译，但在机器翻译中是首次使用
- 这种范式的特点：
  - 公开的大规模的训练数据
  - 周期性的公开评测和研讨
  - 开放源码的工具

近年来统计机器翻译论文发表数量

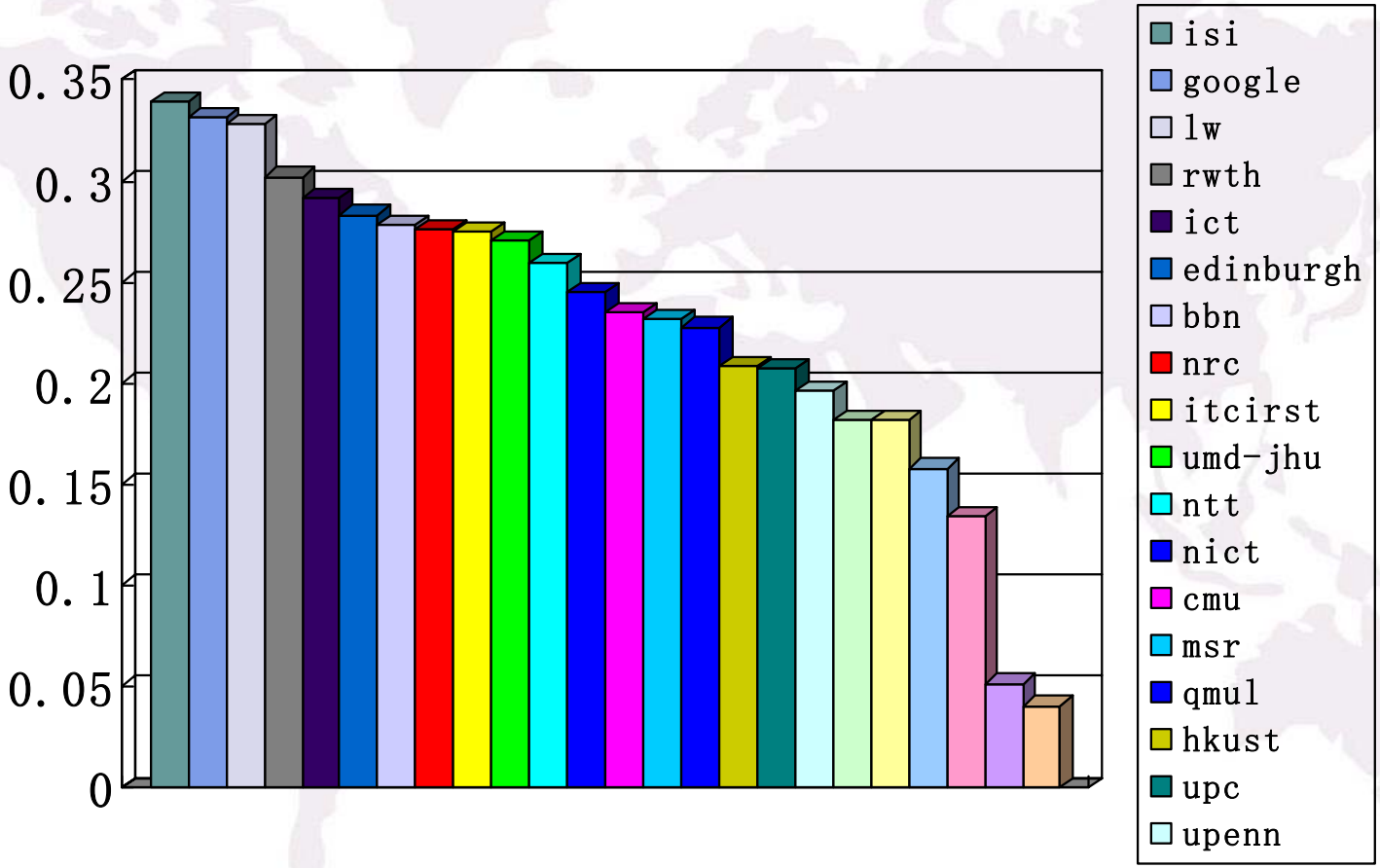


引自 Franz Josef Och, Statistical Machine Translation: Foundations and Recent Advances, Tutorials on MT Summit X, September 13-15, 2005, Phuket, Thailand

近年来国际NIST评测最好成绩



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



统计机器翻译目前的水平

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 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

INSTITUTE OF COMPUTING TECHNOLOGY

半岛电视台
网站新闻

阿拉伯语原文

الجمعة 1428/9/30 هـ- الموافق2007/10/12 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 وسحبت 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

INSTITUTE OF COMPUTING TECHNOLOGY

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

The Dalai Lama

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

- طوة خ في لاما الدلاي التبت في ذيين للبو الروحي الزعيم المقبل أسبوع ال أبيض ال البيت في بوش جورج أميركي ال الرئيس يستقبل صين ال تزعج أن المرجح من.

  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. ☺

- فعل ماغرار على،جوندروغوردون أبيض ال البيت باسم ث المتحد قال كما إعلام ال وسائل عن بعيدا ص خا إطار في ،لاما الدلاي بوش وسيلتقي السابق في.

  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،ذهبية ال الكونغرس ة ميدالي لاما الدلاي الكونغرس خلاله يقلد رسميا حفلا ط ن واشن في التالي اليوم في بوش ضر وسيح يمنحه 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 أبيض ال البيت زار أن له سبق ذي ال لاما الدلاي مع ة علاني بوش فيها ظهري التي أولى ال المر سيكون الوسام منح وحفل رسمي غير اجتماعات ة.

  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. ☺

- ش في تدخل إنه ة قائل القرار وشجبت الوسام لاما الدلاي منح أميركي ال الكونغرس قرر عندما ضب بغ صين ال وردت 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

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 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

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- **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）

目录

- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
  —基于词的IBM模型
- 最成熟的统计机器翻译方法
  —基于短语的模型
- 目前统计机器翻译研究的热点
  —基于句法的模型
- 统计机器翻译面临的问题和展望

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

基于词的统计机器翻译方法

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 统计机器翻译—为翻译建立概率模型
- IBM的信源信道模型
- 语言模型—n元语法模型
- 翻译模型—IBM模型1-5
- 搜索算法
- Candide系统

为翻译建立概率模型

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 假设任意一个英语句子**e**和一个法语句子 **f**，我们定义**f**翻译成**e**的概率为：

$$\Pr(e \mid f)$$

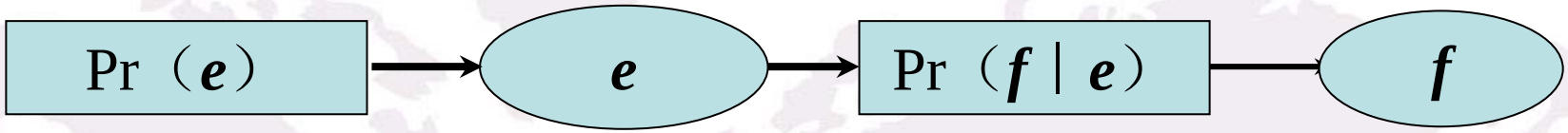
其归一化条件为：

$$\sum_{e} \Pr(e \mid f) = 1$$

- 于是将 **f** 翻译成 **e** 的问题就变成求解问题：

$$\hat{e} = \arg\max_{e} \Pr(e \mid f)$$

信源信道模型 (1)



- 假设我们看到的源语言文本F是由一段目标语言文本*E*经过某种奇怪的编码得到的，那么翻译的目标就是要将F还原成E，这也就是就是一个解码的过程。
- 注意，在信源信道模型中：
  - 噪声信道的源语言是翻译的目标语言
  - 噪声信道的目标语言是翻译的源语言

  这与整个机器翻译系统翻译方向的刚好相反

信源信道模型 (2)

$$\hat{e} = \arg\max_{e} \Pr(e)\Pr(f \mid e)$$

- **P.Brown**称上式为统计机器翻译基本方程式
  - 语言模型：***P*(E)**
  - 翻译模型：***P*(F|E)**
- 语言模型反映“ **E**像一个句子”的程度：流利度
- 翻译模型反映“**F**像**E**”的程度：忠实度
- 联合使用两个模型效果好于单独使用翻译模型，因为后者容易导致一些不好的译文。

信源信道模型 (3)

- 统计机器翻译分解为以下三个问题：
  - 语言模型的定义和参数估计
  - 翻译模型的定义和参数估计
  - 解码

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

语言模型 — n元语法模型

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 语言模型在机器翻译中具有极为重要的作用
- 到目前位置，统计机器翻译中最常用、而且最有效的模型仍然是n元语法模型
- 模型的阶数越来越高：3元、4元、5元
- 模型的训练语料越来越大：
  - Google提供了公开的Web 1T语料库，其中的n元共现词频数据是从web中得到的1T英文词的语料库中统计得到的（剪切掉了低频组合）
  - Google号称使用了2T英文词训练的语言模型
  - 大规模的数据为系统实现带来很大的困难

翻译模型

- 翻译模型$P(F|E)$反映的是一个源语言句子**E**翻译成一个目标语言句子**F**的概率
- 由于源语言句子和目标语言句子几乎不可能在语料库中出现过，因此这个概率无法直接从语料库统计得到，必须分解成词语翻译的概率和句子结构（或者顺序）翻译的概率

翻译模型与对齐

- 翻译模型的计算，需要引入隐含变量：对齐***A***:

$$P(\mathrm{F}\,|\,\mathrm{E}) = \sum_{A} P(\mathrm{F},\mathrm{A}\,|\,\mathrm{E})$$

  - 翻译概率*P*(F|E)的计算转化为对齐概率*P*(F,A|E)的估计
  - 对齐：建立源语言句子和目标语言句子的词与词之间的对应关系和句子结构之间的对应关系

词语对齐的表示 (1)

- 图形表示
  - ✓ 连线
  - ✓ 矩阵（见下页）
- 数字表示
  - ✓ 给每个目标语言单词标记其所有对应的源语言单词

词语对齐的表示 (2)

	中国	十四	个	边境	开放	城市	经济	建设	成就	显著
achievement								■	■	
economic							■			
marked										■
cities						■				
board				■						
open					■					
14		■	■							
‘s	■									
China	■									

IBM Model 1

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

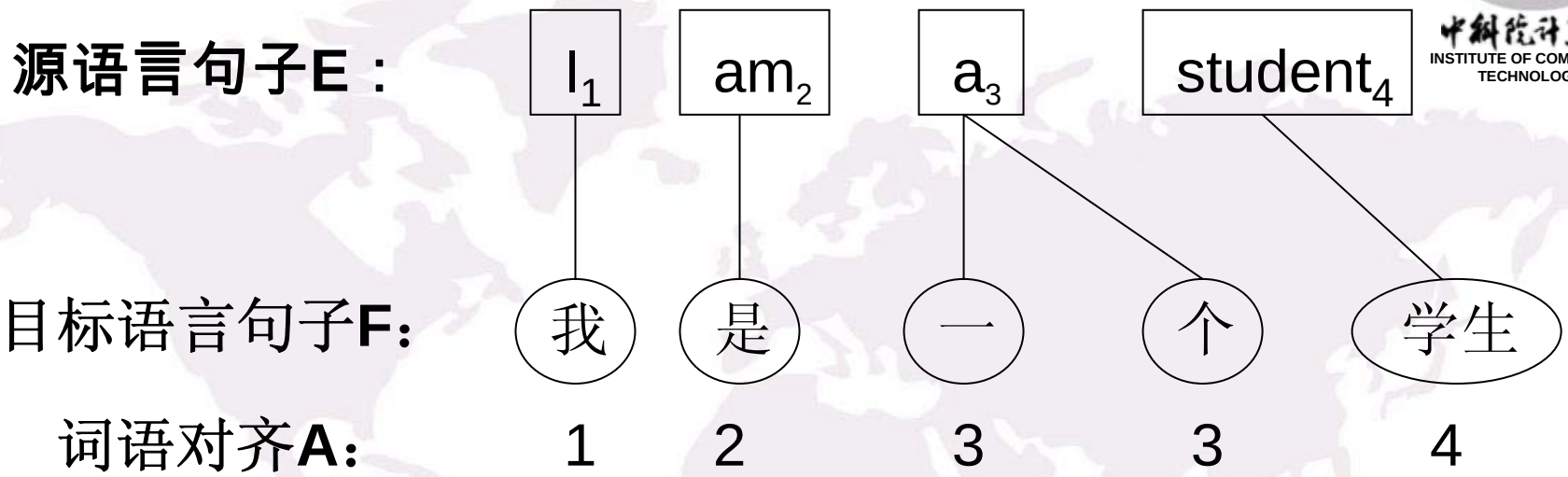
- 最简单的理解，可以句子**e**翻译成**f**的概率，就是**e**中每一个词语翻译成**f**中对应词语的概率的乘积
- 这就是**IBM Model 1**的基本思想
- **IBM**提出了复杂度递增的**5**个统计翻译模型，**IBM Model 1**是其中最简单的模型

IBM Model 1-5

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- IBM Model 1仅考虑词对词的互译概率
- IBM Model 2加入了词的位置变化的概率
- IBM Model 3加入了一个词翻译成多个词的概率
- IBM Model 4
- IBM Model 5

IBM Model 1 & 2 推导方式 (1)



IBM模型1&2的推导过程：

1. 猜测目标语言句子长度；
2. 从左至右，对于每个目标语言单词：
   - 首先猜测该单词由哪一个源语言单词翻译而来；
   - 再猜测该单词应该翻译成什么目标语言词。

IBM Model 1 & 2 推导方式 (2)

中科院计算所 INSTITUTE OF COMPUTING TECHNOLOGY

假设翻译的目标语言句子为： $\mathrm{F} = f_1^m = f_1 f_2 \cdots f_m$

假设翻译的源语言句子为： $\mathrm{E} = e_1^l = e_1 e_2 \cdots e_l$

假设词语对齐表示为：

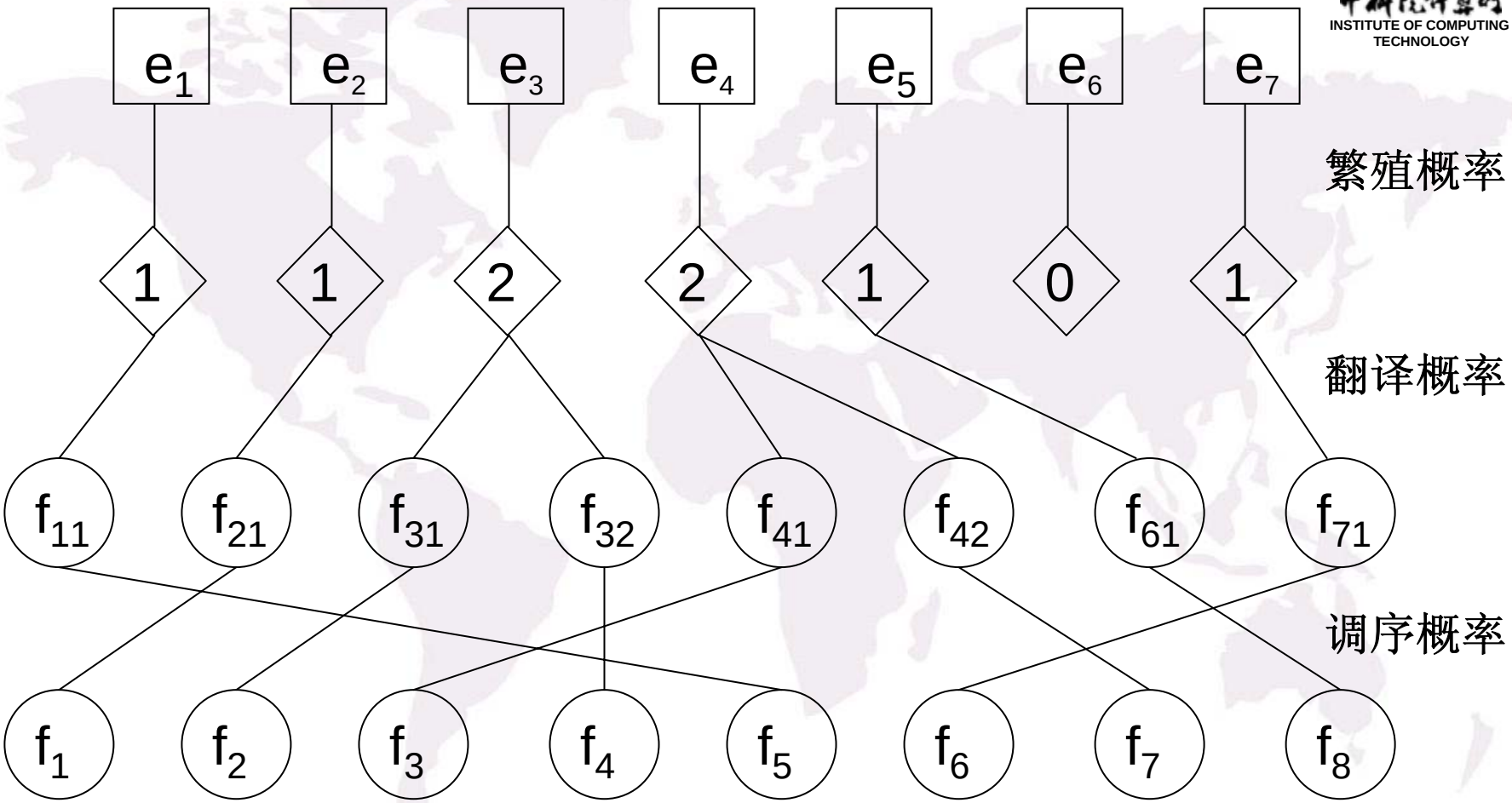
$$\mathrm{A} = a_1^m = a_1 a_2 \cdots a_m, \forall i \in \{1, \cdots, m\}, a_i \in \{0, \cdots, l\}$$

那么词语对齐的概率可以表示为：

$$\Pr(\mathrm{F}, \mathrm{A} \mid \mathrm{E}) = \Pr(m \mid \mathrm{E}) \prod_{j=1}^{m} \Pr(a_j \mid a_1^{j-1}, f_1^{j-1}, m, \mathrm{E}) \Pr(f_j \mid a_1^j, f_1^{j-1}, m, \mathrm{E})$$

注意：在**IBM Model**中，词语对齐只考虑了源语言到目标语言的单向一对多形式，不考虑多对一和多对多的形式。

IBM Model 3 & 4 & 5 推导方式 (1)



IBM Model 3 & 4 & 5 推导方式 (2)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

1. 首先根据源语言词语的繁殖概率，确定每个源语言词翻译成多少个目标语言词；
2. 根据每个源语言词语的目标语言词数，将每个源语言词复制若干次；
3. 将复制后得到的每个源语言词，根据翻译概率，翻译成一个目标语言词；
4. 根据调序概率，将翻译得到的目标语言词重新调整顺序，得到目标语言句子。

IBM模型的参数训练：EM算法

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- EM参数训练算法是经典的无指导学习的算法：
  1. 给定初始参数；
  2. E步骤：用已有的参数计算每一个句子对的所有可能的对齐的概率；
  3. M步骤：用得到的所有对齐的概率重新计算参数；
  4. 重复执行E步骤和M步骤，直到收敛。
- 由于EM算法的E步骤需要穷尽所有可能的对齐，通常这会带来极大的计算量，除非我们可以对计算公式进行化简（就像前面IBM Model 1所做的那样），否则这种计算量通常是不可承受的。

IBM模型的参数训练：Viterbi训练

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- **Viterbi**参数训练算法：
  1. 给定初始参数；
  2. 用已有的参数求概率最大（**Viterbi**）的词语对齐；
  3. 用得到的概率最大的词语对齐重新计算参数；
  4. 回到第二步，直到收敛为止。
- 在对参数计算公式无法化简的情况下，采用**Viterbi**参数训练算法是一种可行的做法，这种算法通常可以迅速收敛到一个可以接受的结果。

IBM模型的参数训练

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- **IBM Model 1**
  - 任何初始值均可达到全局最优
- **IBM Model 2~5：**
  - 存在大量局部最优，任意给定的初值很容易导致局部最优，而无法到达全局最优的结果
  - **IBM**的训练策略：
    - 依次训练**IBM Model 1-5**
    - 对于与上一级模型相同的参数初始值，直接取上一个模型训练的结果；
    - 对于新增加的参数，取任意初始值。

统计机器翻译的解码

- 给定**F**，求**E**，使得**P(E)*P(F|E)**最大
- 解码问题实际上是一个搜索问题，搜索空间巨大，不能保证总能找到全局最优，但通常一些局部最优也是可以接受的
- 如果考虑所有的词语对齐可能性，那么这个问题是一个**NP**完全问题 **[Knight 99]**
- 经典的算法：
  - 单调解码（不调整词序）
  - 堆栈搜索
  - 贪婪算法
  - ……

IBM公司的Candide系统(1)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 基于统计的机器翻译方法
- 分析—转换—生成
  - 中间表示是线性的
  - 分析和生成都是可逆的
- 分析（预处理）：

  1.短语切分 2.专名与数词检测

  3.大小写与拼写校正

  4.形态分析 5.语言的归一化

IBM公司的Candide系统(2)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 转换（解码）：基于统计的机器翻译
- 解码分为两个阶段：
  - 第一阶段：使用粗糙模型的堆栈搜索
    - 输出140个评分最高的译文
    - 语言模型：三元语法
    - 翻译模型：**EM Trained IBM Model 5**
  - 第二阶段：使用精细模型的扰动搜索
    - 对第一阶段的输出结果先扩充，再重新评分
    - 语言模型：链语法
    - 翻译模型：最大熵翻译模型（选择译文词）

IBM公司的Candide系统(3)

- ARPA的测试结果：

	Fluency		Adequacy		Time Ratio	
	1992	1993	1992	1993	1992	1993
Systran	.466	.540	.686	.743		
Candide	.511	.580	.575	.670		
Transman	.819	.838	.837	.850	.688	.625
Manual		.833		.840		

目录

- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
  —基于词的IBM模型
- 最成熟的统计机器翻译方法
  —基于短语的模型
- 目前统计机器翻译研究的热点
  —基于句法的模型
- 统计机器翻译面临的问题和展望

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

基于短语的统计机器翻译方法

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 从信源信道模型到对数线性模型
- 翻译模型的发展—基于短语的模型
- 短语的自动抽取
- 短语翻译概率的计算
- 短语语序的调整
- 几个基于短语的开源系统

统计机器翻译的对数线性模型(1)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- Och于ACL2002提出，思想来源于Papineni提出的基于特征的自然语言理解方法，该论文获得ACL2002的最佳论文称号
- 是一个比信源－信道模型更具一般性的模型，信源－信道模型是其一个特例
- 原始论文的提法是“最大熵”模型，现在通常使用“对数线性（Log-Linear）模型”这个概念。“对数线性模型”的含义比“最大熵模型”更宽泛，而且现在这个模型通常都不再使用最大熵的方法进行参数训练，因此“对数线性”模型的提法更为准确。
- 与NLP中通常使用的最大熵方法的区别：使用连续量（实数）作为特征，而不是使用离散的布尔量（只取0和1值）作为特征

统计机器翻译的对数线性模型(2)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

假设***e***、***f***是机器翻译的目标语言和源语言句子，$h_1(e,f), \ldots, h_M(e,f)$分别是***e***、***f***上的**M**个特征，$\lambda_1, \ldots, \lambda_M$是与这些特征分别对应的**M**个参数，那么翻译概率可以用以下公式模拟：

$$\Pr(e \mid f) \approx p_{\lambda_1 \ldots \lambda_M}(e \mid f)$$

$$= \frac{\exp[\sum_{m=1}^{M} \lambda_m h_m(e, f)]}{\sum_{e'} \exp[\sum_{m=1}^{M} \lambda_m h_m(e', f)]}$$

统计机器翻译的对数线性模型(3)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

对于给定的***f***，其最佳译文***e***可以用以下公式表示：

$$\hat{e} = \arg\max_{e}\{\Pr(e \mid f)\}$$

$$\approx \arg\max_{e}\{\sum_{m=1}^{M}\lambda_m h_m(e, f)\}$$

对数线性模型vs.噪声信道模型

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 取以下特征和参数时，对数线性模型等价于噪声信道模型：
  - 仅使用两个特征
  - $h_1(e,f)=\log p(e)$
  - $h_2(e,f)=\log p(f|e)$
  - $\lambda_1=\lambda_2=1$

对数线性模型：Och的实验 (1)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 方案
  - 首先将信源信道模型中的翻译模型换成反向的翻译模型，简化了搜索算法，但翻译系统的性能并没有下降；
  - 调整参数λ_1和λ_2，系统性能有了较大提高；
  - 再依次引入其他一些特征，系统性能又有了更大的提高。

对数线性模型： Och的实验 (2)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 其他特征
  - 句子长度特征(WP)：对于产生的每一个目标语言单词进行惩罚；
  - 附加的语言模型特征(CLM)：一个基于类的语言模型特征；
  - 词典特征(MX)：计算给定的输入输出句子中有多少词典中存在的共现词对。

对数线性模型： Och的实验 (3)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 实验结果

	objective criteria [%]					subjective criteria [%]	
	SER	WER	PER	mWER	BLEU	SSER	IER
Baseline($\lambda_m = 1$)	86.9	42.8	33.0	37.7	43.9	35.9	39.0
ME	81.7	40.2	28.7	34.6	49.7	32.5	34.8
ME+WP	80.5	38.6	26.9	32.4	54.1	29.9	32.2
ME+WP+CLM	78.1	38.3	26.9	32.1	55.0	29.1	30.9
ME+WP+CLM+MX	77.8	38.4	26.8	31.9	55.2	28.8	30.9

对数线性模型的优点

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 噪声模型只有在理想的情况下才能达到最优，对于简化的语言模型和翻译模型，取不同的参数值实际效果更好；
- 对数线性模型大大扩充了统计机器翻译的思路；
- 特征的选择更加灵活，可以引入任何可能有用的特征。

对数线性模型的参数训练

- 目的是得到各个特征的参数 $\lambda_1 \ldots\ldots \lambda_n$
- 可用的训练算法
  - GIS（最大熵模型的训练算法）
  - 感知机
  - 最小错误率(MER)：直接以评测指标（如BLEU）最好为训练目标
  - 最大互信息(MMI)：把导致总体BLEU值最高的译文定义为好的译文，其他译文定义为不好的译文，进行判别式训练
  - 单纯形算法
- 目前通常使用最小错误率训练算法或单纯形算法

对数线性模型的特征 (1)

- 无论在噪声信道模型还是在对数线性模型中，语言模型和翻译模型都是两个最主要的特征
- 对于语言模型，目前主流的做法都还是采用**n**元语法，还没有发现哪些方法能够超过这种简单的模型
- 对于翻译模型，研究者进行了大量的尝试
  - 最早期的**IBM Model 1-5**是基于词的翻译模型
  - 目前最成熟和稳定的模型是基于短语的翻译模型
  - 基于句法的翻译模型近年来也取得了较大进展
- 在对数线性模型中，多个翻译模型和语言模型可以同时使用

对数线性模型的特征 (2)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 其他特征
  - 词典特征
  - 长度特征：句子单词数。这个特征可以一定程度上避免由于使用语言模型导致的过于偏向短句子的倾向
  - ……

翻译模型的发展—基于短语的模型

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 基于词的IBM翻译模型有明显的缺陷：一个词在翻译的时候基本上不考虑上下文，孤立地进行翻译，导致了大量的错误；词序调整模型近乎无礼，很难准确调整词序，对词序差别较大的语言之间的翻译效果太差。
- 人们很容易想到，将一个短语捆绑起来进行翻译，可以大大提高翻译的准确率
- 很多不同的研究人员尝试了各种各样的基于短语的翻译模型，最终形成了目前比较成熟的基于短语的翻译模型

基于短语的翻译模型 (1)

- 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？

基于短语的翻译模型 (2)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 假设**f**和**e**之间存在一个短语对齐***B***，而且这个短语对齐是一对一的，那么：

$$\Pr(f_1^J \mid e_1^I) = \sum_B \Pr(f_1^J, B \mid e_1^I) = \sum_B \Pr(B \mid e_1^I) \Pr(f_1^J \mid B, e_1^I)$$

- 假设短语划分的概率$\Pr(B \mid e_1^I)$为均匀分布：

$$\Pr(B \mid e_1^I) = \alpha(e_1^I)$$

- 于是：

$$\Pr(f_1^J \mid e_1^I) = \alpha(e_1^I) \sum_B \Pr(f_1^J \mid B, e_1^I)$$

基于短语的翻译模型 (3)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 假设短语的翻译是互相独立的，并且各种短语顺序调整的概率完全相同，那么：

$$\Pr(f_1^J \mid e_1^I) = \alpha(e_1^I) \sum_B \prod_k p(\tilde{f}_k \mid \tilde{e}_k)$$

这里$\tilde{f}_k$和$\tilde{e}_k$是在B对齐下源语言和目标语言的短语

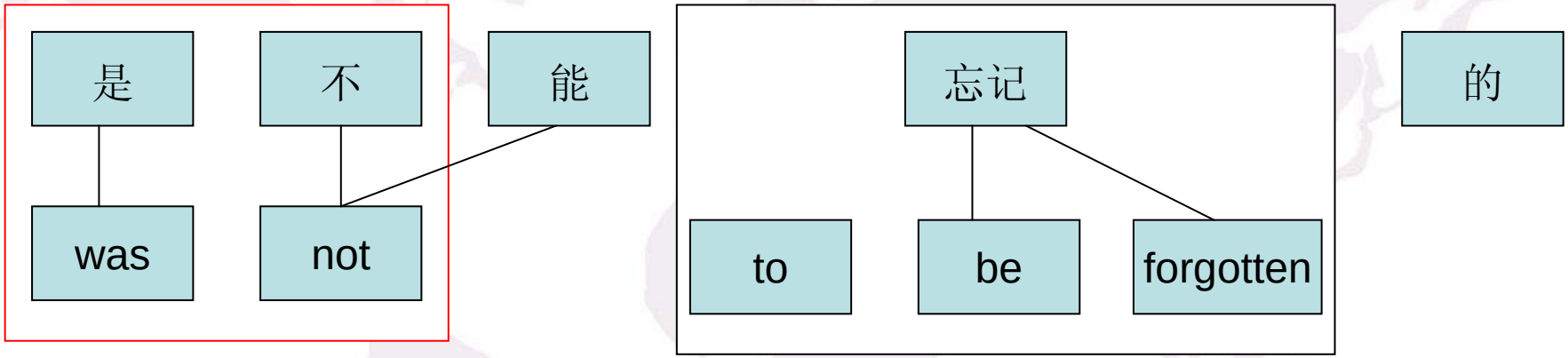
而$\Pr(\tilde{f}_k \mid \tilde{e}_k)$可以通过对短语对齐的语料库统计得到：

$$p(\tilde{f}_k \mid \tilde{e}_k) = \frac{N(\tilde{f}_k, \tilde{e}_k)}{N(\tilde{e}_k)}$$

基于短语的翻译模型 (4)

- 实际上，目前在计算短语翻译概率的时候，通常并不去真正生成一个短语对齐的语料库，而是直接从词语对齐的语料库上，去产生所有可能的短语对齐
- 所以，需要先利用**IBM Model**进行词语对齐，但：
  - **IBM Model**只能产生单向一对多的对齐
  - 为了产生更合理的对齐，需要实现多对多对齐，通常的做法是：
    - 先用**IBM Model**对两个方向分别进行一对多对齐
    - 将两个对齐进行某种合并（交集、并集、部分并集），这个操作称为“平衡化”
- 根据词语对齐的结果抽取短语并计算概率

基于词语对齐的短语自动抽取

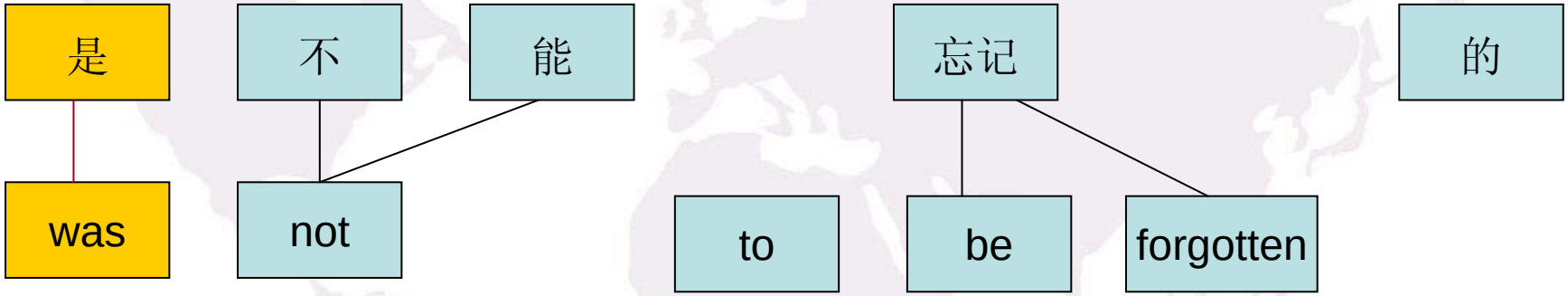


不相容

相容

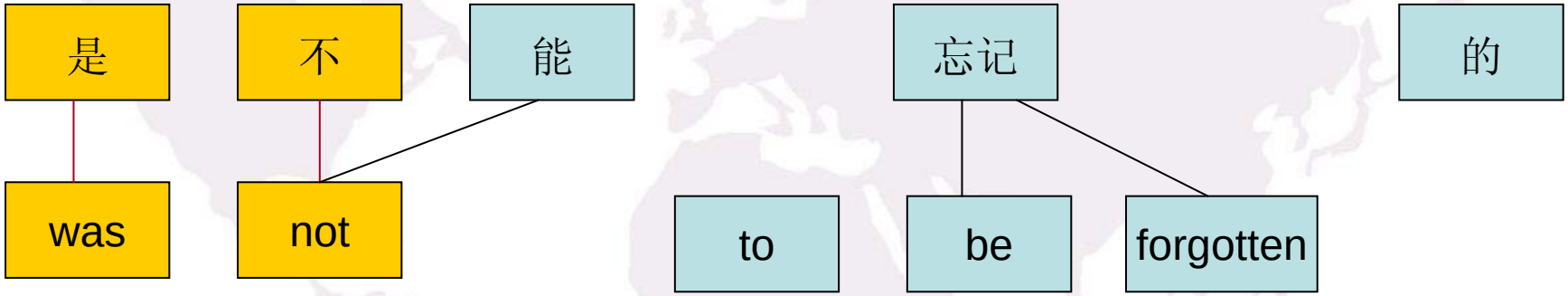
短语自动抽取算法运行示例 (1)

- 列举源语言所有可能的短语，根据对齐检查相容性



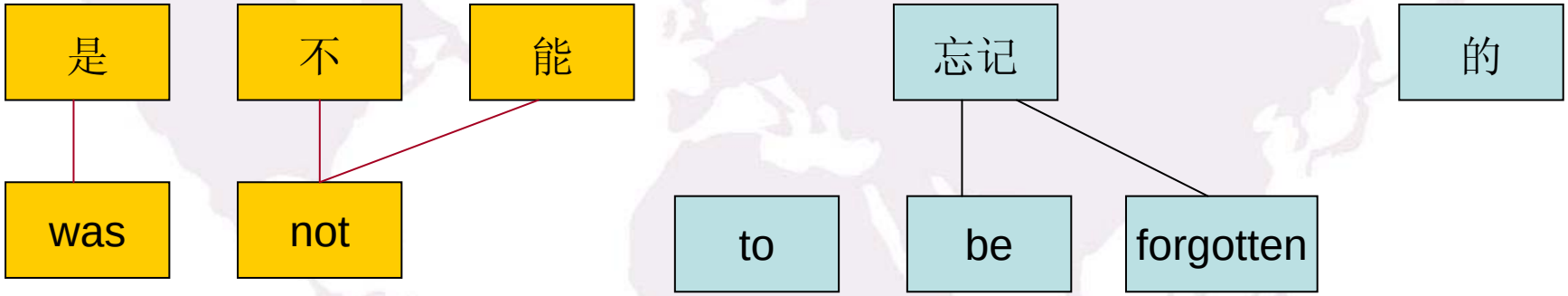
（是，was）

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)

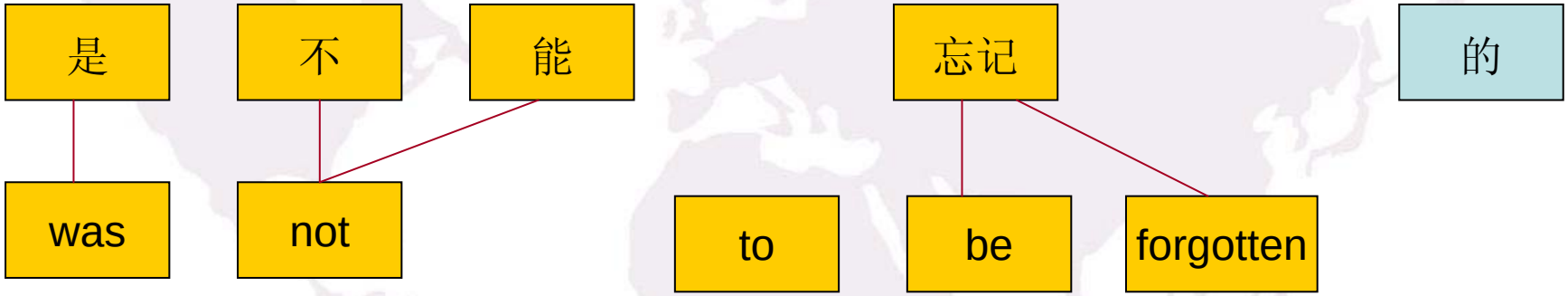


（是不能，was not)

短语自动抽取算法运行示例(4)

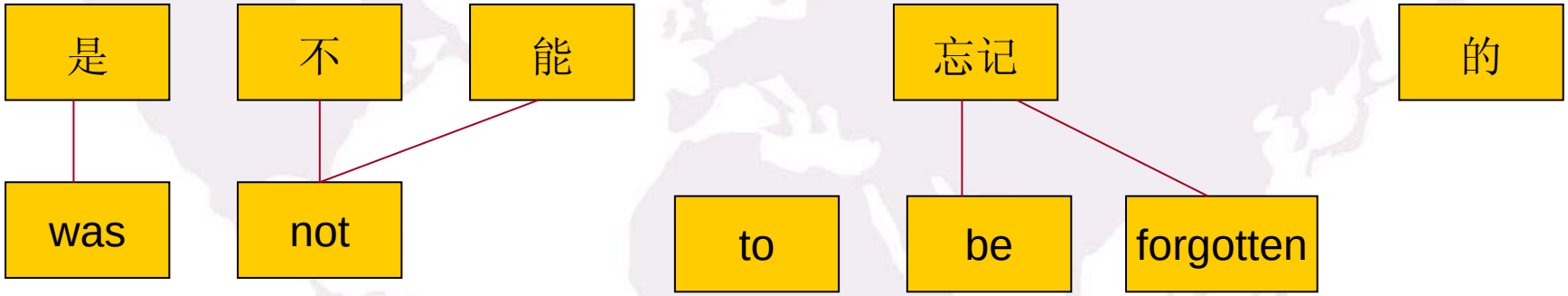
（是不能，was not to)

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

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

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

双语短语的概率计算 (1)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 利用共现次数计算概率

$$p(\tilde{f} \mid \tilde{e}) = \frac{N(\tilde{f}, \tilde{e})}{\sum_{f'} N(\tilde{f}', \tilde{e})}$$

其中 $N(\tilde{f}, \tilde{e})$ 表示短语 $(\tilde{f}, \tilde{e})$ 在语料库中出现的次数

如果 $\tilde{e}$ 的一次出现对应有N个可能的翻译，那么每一个翻译对 $N(\tilde{f}, \tilde{e})$ 的贡献是1/N

双语短语的概率计算 (2)

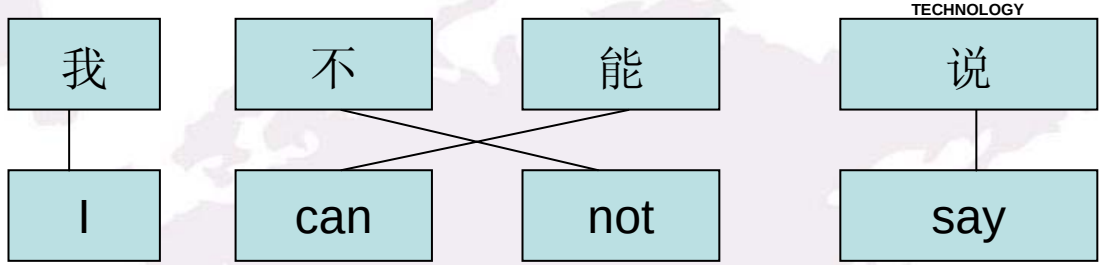
- 前例：

不能 **not**

不能 **not to**

N(不能, not) = ½

N(不能, not to) = ½



如果语料库中另外有一句话：

N(不能, can not) = 1

则

$$p(not\ to \mid 不能) = \frac{1/2}{(1/2+1/2+1)} = 1/4$$

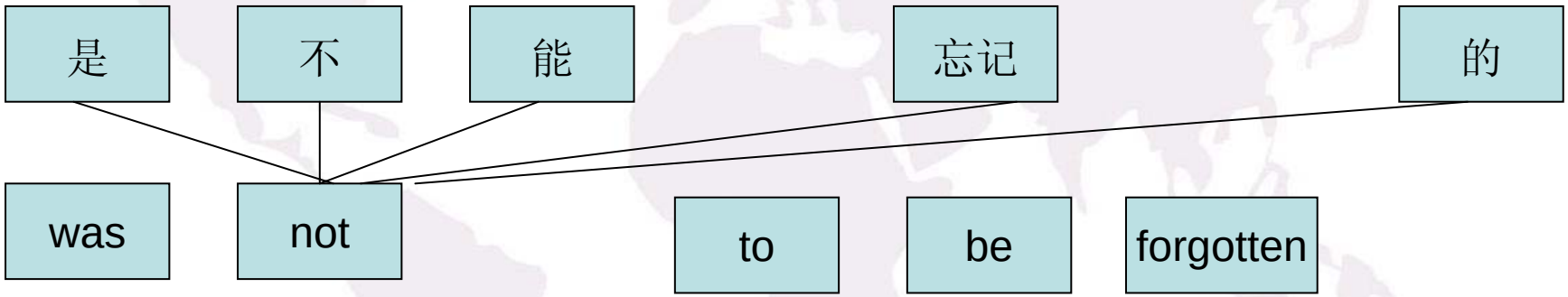
$$p(not \mid 不能) = \frac{1/2}{(1/2+1/2+1)} = 1/4$$

$$p(can\ not \mid 不能) = \frac{1}{(1/2+1/2+1)} = 1/2$$

双语短语的概率计算 (3)

- 仅利用共现次数计算概率信息不全面

  较短的短语（如单词）出现次数多，概率不集中，较长的短语概率比较大



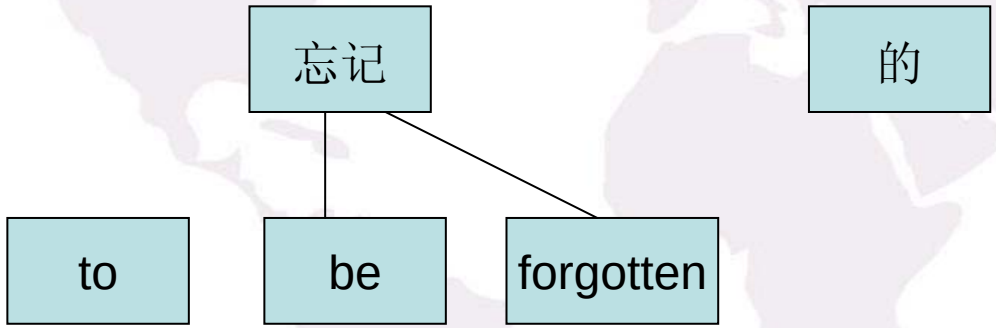
$$p(not \mid 是不能忘记的) = 1$$

对齐错误使得概率计算不准确，影响解码

双语短语的概率计算 (4)

引入短语的词汇化翻译概率，利用**IBM Model 1**训练得到的词语翻译表计算双向的词语翻译概率

$$lex(\tilde{f} \mid \tilde{e}, a) = \prod_{j=1}^{n} \frac{1}{|\{i \mid (j,i) \in a\}|} \sum_{\forall (i,j) \in a} p(f_j \mid e_i)$$



$$lex(\text{忘记 的} \mid to\ be\ forgotten) = \frac{1}{2} \times (p(\text{忘记} \mid be) + p(\text{忘记} \mid forgotten)) \times p(\text{的} \mid NULL)$$

双语短语的概率计算 (5)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 通常都使用双向的短语翻译概率和双向的词汇化翻译概率，一个四个概率作为特征，与其他特征一起，利用最小错误率算法调整特征参数

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

短语语序的调整

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 在基于短语的模型中，短语内部的顺序无需调整，只需要调整短语之间的顺序
- 短语的调序模型类似于基于词的模型，允许任意的语序调整
- 为了避免搜索空间的过于膨胀，通常限制语序调整的距离

基于短语的统计机器翻译开源系统

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 法老（**Pharaoh**）
- 丝路（**SilkRoad**）
- 摩西（**Moses**）

法老（Pharaoh）

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 由**Philipp Koehn**开发
- 最经典的开源的基于短语的统计机器翻译系统
- 效果远远好于基于词的系统
- 性能稳定
- 推出后很快成为相关研究的基准（**baseline**）
- 缺点：解码器没有开放源代码

丝路（SilkRoad）

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 由国内五家单位联合开发的基于短语的开放源代码的统计机器翻译系统
- 完全开放源代码，包括训练部分和解码部分
- 多个解码器，基本原理与法老类似
- 有完整的中文文档，便于学习

摩西（Moses）

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 最新的开放源代码的基于短语的统计机器翻译系统
- 完全开放源码，包括训练部分和解码部分
- 在基于短语的模型中加入了要素模型（Factored Model）
- 采用了词汇化的短语语序调整模型
- 代码优化非常出色
- 性能比法老又有了明显提高

目录

- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
  —基于词的IBM模型
- 最成熟的统计机器翻译方法
  —基于短语的模型
- 目前统计机器翻译研究的热点
  —基于句法的模型
- 统计机器翻译面临的问题和展望

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

基于句法的模型

- 翻译模型的发展—基于句法的模型
- 基于句法的模型概述
- 形式上基于句法的模型
- 语言学上基于句法的模型
- 搜索算法—CYK形式的堆栈搜索

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

翻译模型的发展—基于句法的模型

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 基于短语的模型比基于词的模型性能有了较大提高，但对于短语之间的语序调整，仍然没有提供合理的解决方案
- 经验表明，在基于短语的统计机器翻译系统中，绝大多数匹配的短语长度都是**2-3**个词，**1**个词的短语也占相当大的比例
- 要解决长距离语序的调整，引入句法信息是个必然的选择

基于句法的统计翻译模型 (1)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

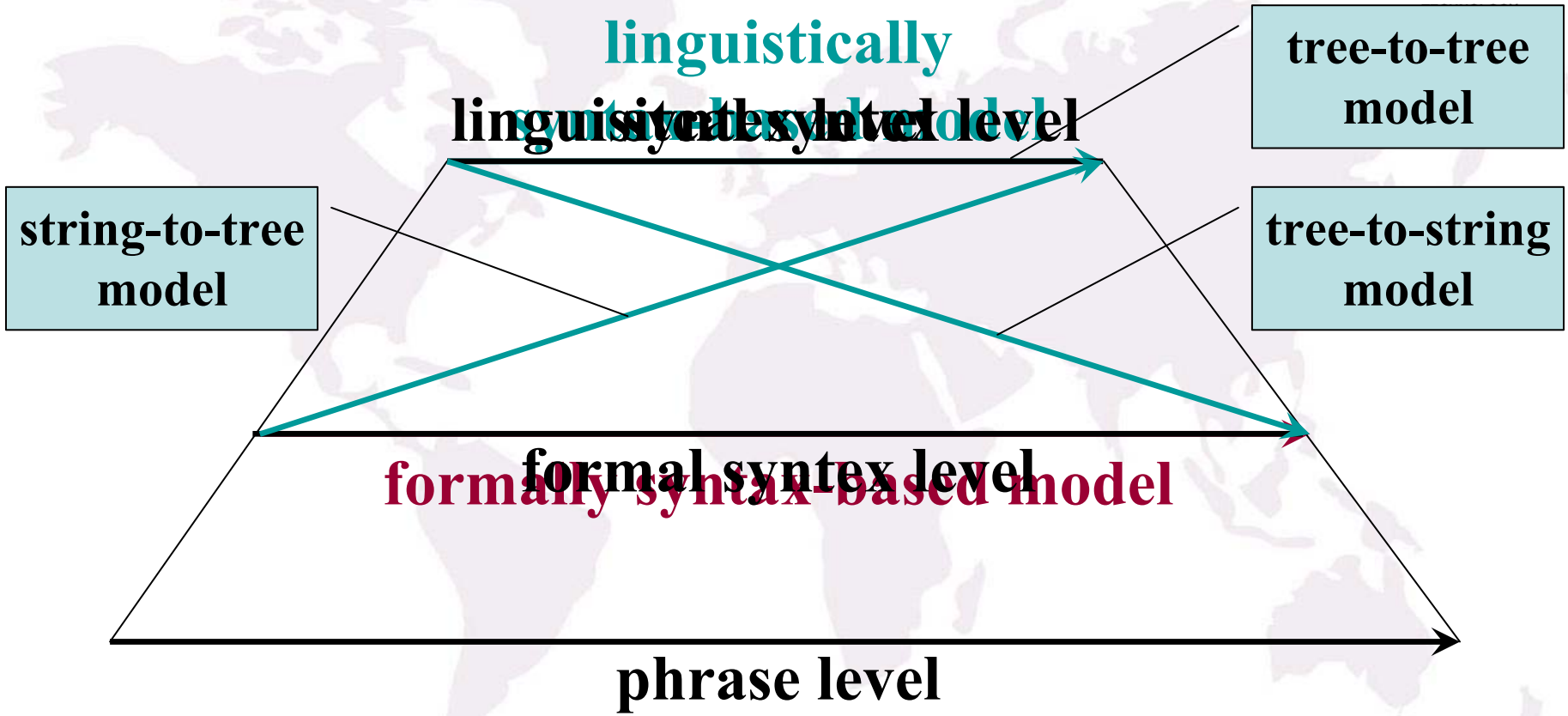
- 基于句法的统计翻译模型，通常的做法都是分别为源语言和目标语言句子建立某种句法结构，并在这两种句法结构之间建立某种对应关系
- 基于句法的统计翻译模型有两种不同的做法
  - 形式上基于句法的统计翻译模型：并不采用语言学上的句法分析，而是从词语对齐的双语语料库中自动获取某种双语平行的句法结构
  - 语言学上基于句法的统计翻译模型：利用语言学上的句法分析，为源语言句子和目标语言句子建立句法结构，并借助词语对齐建立句法结构的对应关系

基于句法的统计翻译模型 (2)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 语言学上基于句法的统计翻译模型又有三种不同的做法
  - 树到串模型：在源语言端进行句法分析并得到源语言句法结构，然后根据词语对齐建立对应的目标语言句法结构（可称为伪句法结构）
  - 串到树模型：在目标语言端进行句法分析并得到目标语言句法结构，然后根据词语对齐建立对应的源语言句法结构（也是伪句法结构）
  - 树到树模型：在源语言端和目标语言端分别进行句法分析并得到双语的句法结构，然后根据词语对齐建立这两种句法结构之间的对应关系

基于句法的统计翻译模型 (3)



形式上基于句法的模型

- 反向转录语法（ITG）和括号转录语法（BTG）
  Inversion (Bracketing) Transduction Grammar (ITG,BTG), Dekai Wu 1997
- 有限状态中心词转录机
  Finite-State Head Transducer, Alshawi 2000
- 基于层次短语的翻译模型
  Hierarchical Phrase-based Model, David Chiang 2005
- 最大熵括号匹配语法的翻译模型
  Maximal Entropy Bracket Transduction Grammar (ME-BTG), Deyi Xiong 2006

ICT
中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

语言学上基于句法的模型

- 串到树模型 **String-to-Tree Model**
  - 美国南加州大学信息科学研究所（**ISI/CSU**）的工作

    Yamada 2001, Galley 2006, Marcu 2006
- 树到串模型 **Tree-to-String Model**
  - 中科院计算所的工作

    Tree-to-string Alignment Template Model (TAT), Liu Yang 2006
  - 微软研究院的工作（依存模型）

    Dependency Treelet Translation, Quirk 2005
- 树到树的模型 **Tree-to-Tree Model**

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

最大熵括号转录语法模型ME-BTG

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 什么是ITG
- 基于ITG的机器翻译
- 什么是BTG
- 基于BTG建立统计翻译模型

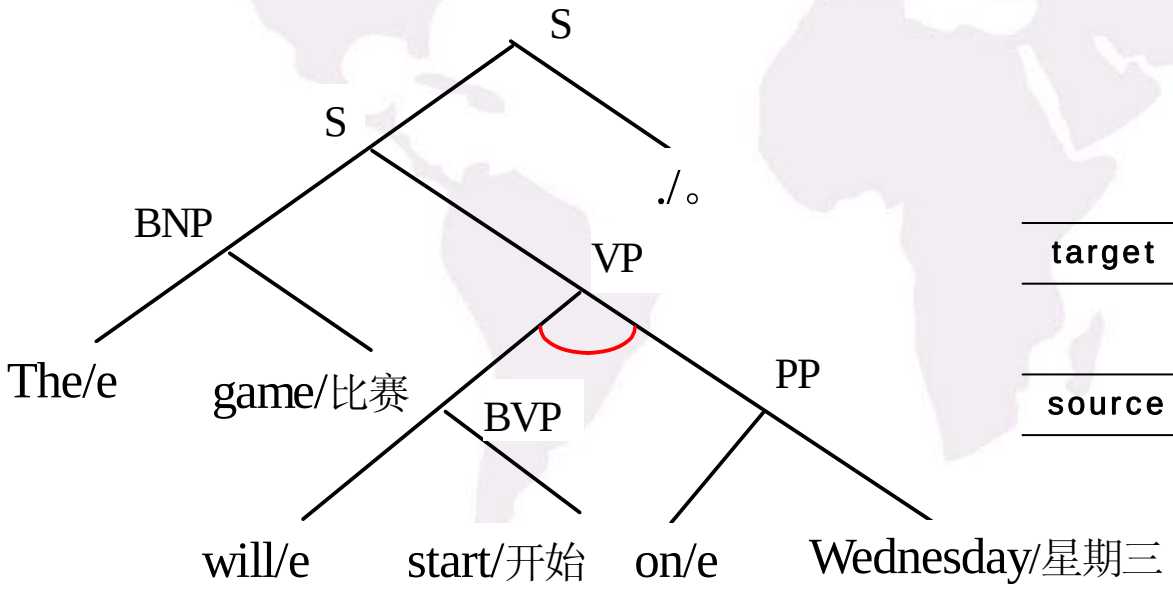
什么是ITG (1)

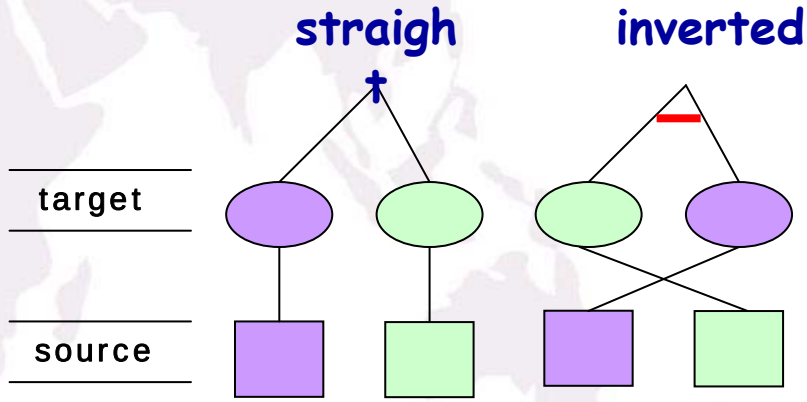
中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- **ITG：Inversion Transduction Grammar**
- **ITG**是一种**Chomsky**范式形式的同步上下文无关语法
- **ITG**的规则有两种类型：
  - 非终结符规则（语法规则）
  - 终结符规则（词典）
- **ITG**规则采用**Chomsky**范式形式，因此所有非终结符规则都是二叉的，
- **ITG**的非终结符规则中，源语言规则到目标语言规则的对应关系只有两种：保序和交换

什么是ITG (2)

ITG rules	Source	Target
A → [B C]	A→BC	A→BC
A → < B C >	A→BC	A→ CB
A → x/y	A→x	A→y





基于 ITG的机器翻译

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 在**ITG**中，仍然使用了**NP**、**VP**之类的句法标记，这对于训练语料库提出了比较高的要求
- 如果我们不考虑标记，也就是说，认为所有的标记都是相同的，只有一个非终结符标记**X**，那么**ITG**就退化成**BTG**

什么是BTG

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- **BTG：Bracketing transduction grammars**
- **BTG**是简化的**ITG**，也就是在**ITG**中，只定义一个非终结符**X**
- **BTG**为两种语言的句法结构之间的对应关系建立了一个最简单的模型
  - 没有标记，只有结构
  - 没有多叉，只有二叉
- **BTG**大大限制了统计机器翻译的解码空间

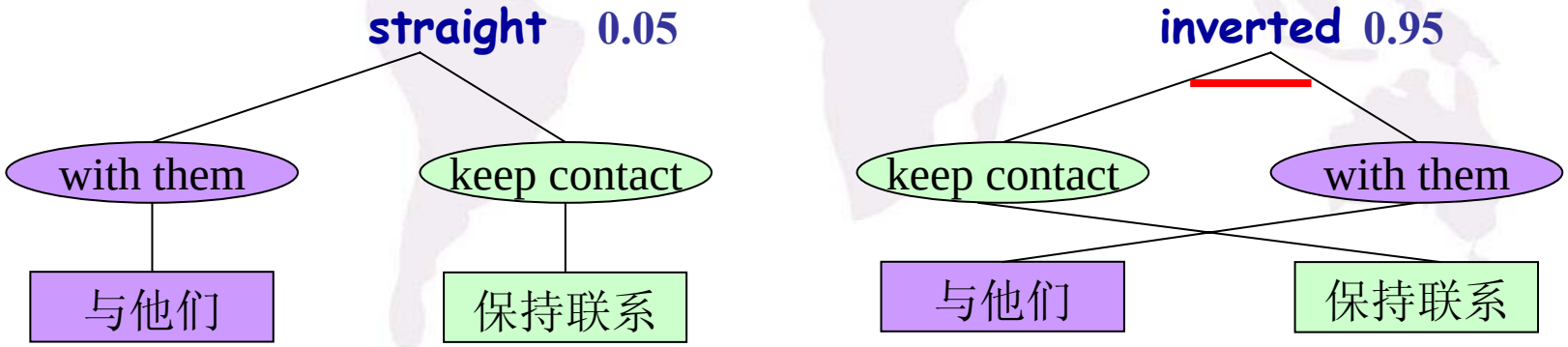
基于BTG的机器翻译

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 只有两条非终结符规则：
  A → [A A]
  A → < A A >
- 吴德凯定义的**Stochastic BTG**给每条规则赋以先验概率
- **Stochastic BTG**是一种非常粗糙的调序模型，无法在细粒度上处理词语调序问题，实际应用效果也很不理想

MEBTG：基本思想

- 在**BTG**框架下，将重排序问题看作是一个二元分类问题：
  - 条件：各种与重排序短语相关的特征
  - 类别：相邻语块的顺序 **{straight, inverted}**
- 引入最大熵模型作为分类模型，根据实际上下文计算合并规则的概率



MEBTG模型

- 模型

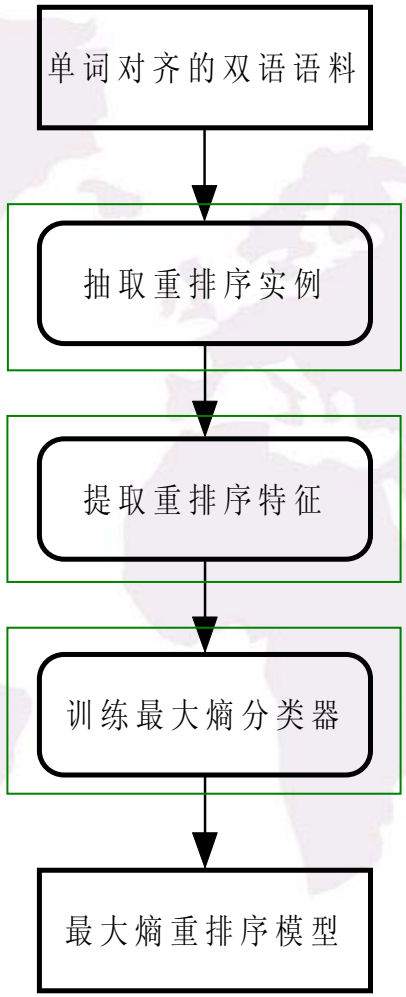
$$\Omega = p_\theta(o \mid A^1, A^2) = \frac{exp(\sum_i \theta_i h_i(o, A^1, A^2))}{\sum_{o'} exp(\sum_i \theta_i h_i(o', A^1, A^2))}$$

- 特征

$$h_i(o, A^1, A^2) = \begin{cases} 1 & \text{if } f(A^1, A^2) = T, o = O \\ 0 & \text{otherwise} \end{cases}$$

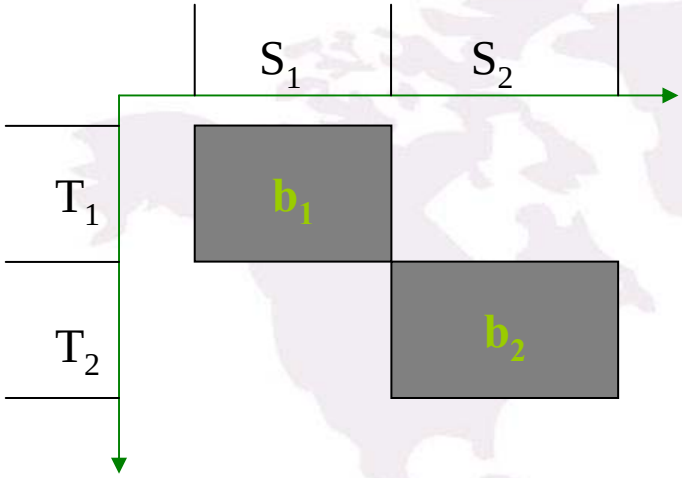
$$O \in \{straight, inverted\}$$

MEBTG训练



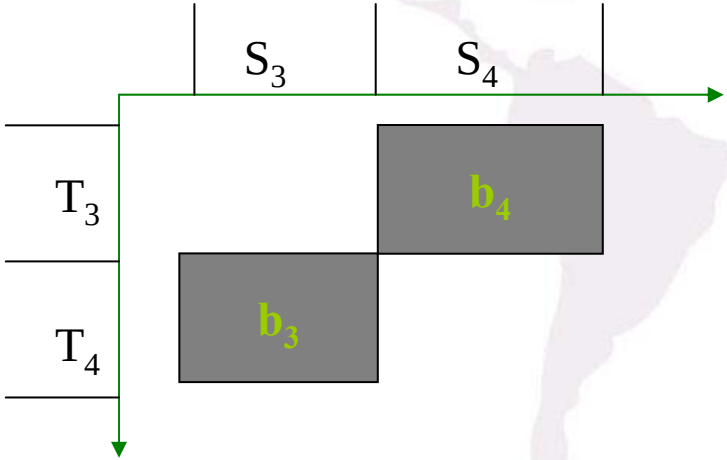
中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

重排序实例



<b_1; b_2> → STRAIGHT

E.g. <今天 有 棒球 比赛|Are there any baseball games today; 吗 ？ |?> → STRAIGHT



<b_3; b_4> → INVERTED

E.g. <澳门 政府|the Macao government; 有关 部门 |related departments of> → INVERTED

重排序特征

- 单目特征：单个源/目标语言边界单词
- 双目特征： 两个源/目标语言边界单词的组合

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

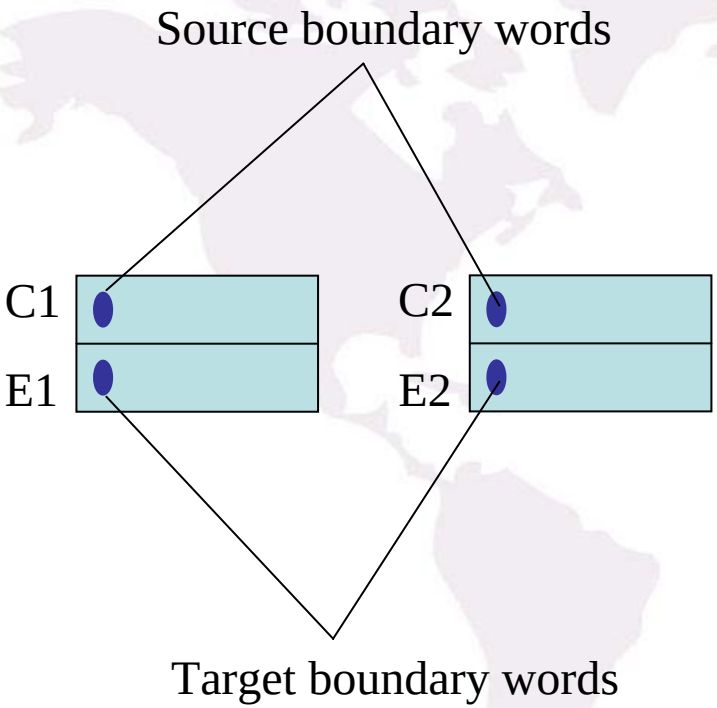
<与 他们|with them; 保持联系|keep contact> → INVERTED

特征选择

$$h_{mono}(o, A^1, A^2) = \begin{cases} 1 & \text{if } A^2.t_1 = keep,\ o = inverted \\ 0 & \text{otherwise} \end{cases}$$

$$h_{bino}(o, A^1, A^2) = \begin{cases} 1 & \text{if } A^1.t_1 = with, A^2.t_1 = keep,\ o = inverted \\ 0 & \text{otherwise} \end{cases}$$

为什么使用边界单词作为特征？



feature	IGR
Phrases	.02655
C1C2E1E2	.0263687
E1E2	.0239286
C1C2	.023363
C2E2	.0192932
C1E1	.0153117
C2	.011371
E2	.00994372
E1	.00899752
C1	.00758598

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

MEBTG：实验

Systems	NIST MT 05	IWSLT 04
Bruin with monotone search	20.1	37.8
Bruin with distance-based reordering	20.9	38.8
Bruin with flat reordering	20.5	38.7
Pharaoh	20.8	38.9
Bruin with MEBTG (单目)	22.0	42.4
Bruin with MEBTG (单目 + 双目)	22.2	42.8

MEBTG模型小结

- 形式上基于句法的模型
- 性能明显超过基于短语的模型
- 完全兼容基于短语的模型
- 采用**BTG**语法形式，只有两条规则
- 规则的选用采用最大熵方法进行取舍

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

层次短语模型 (1)

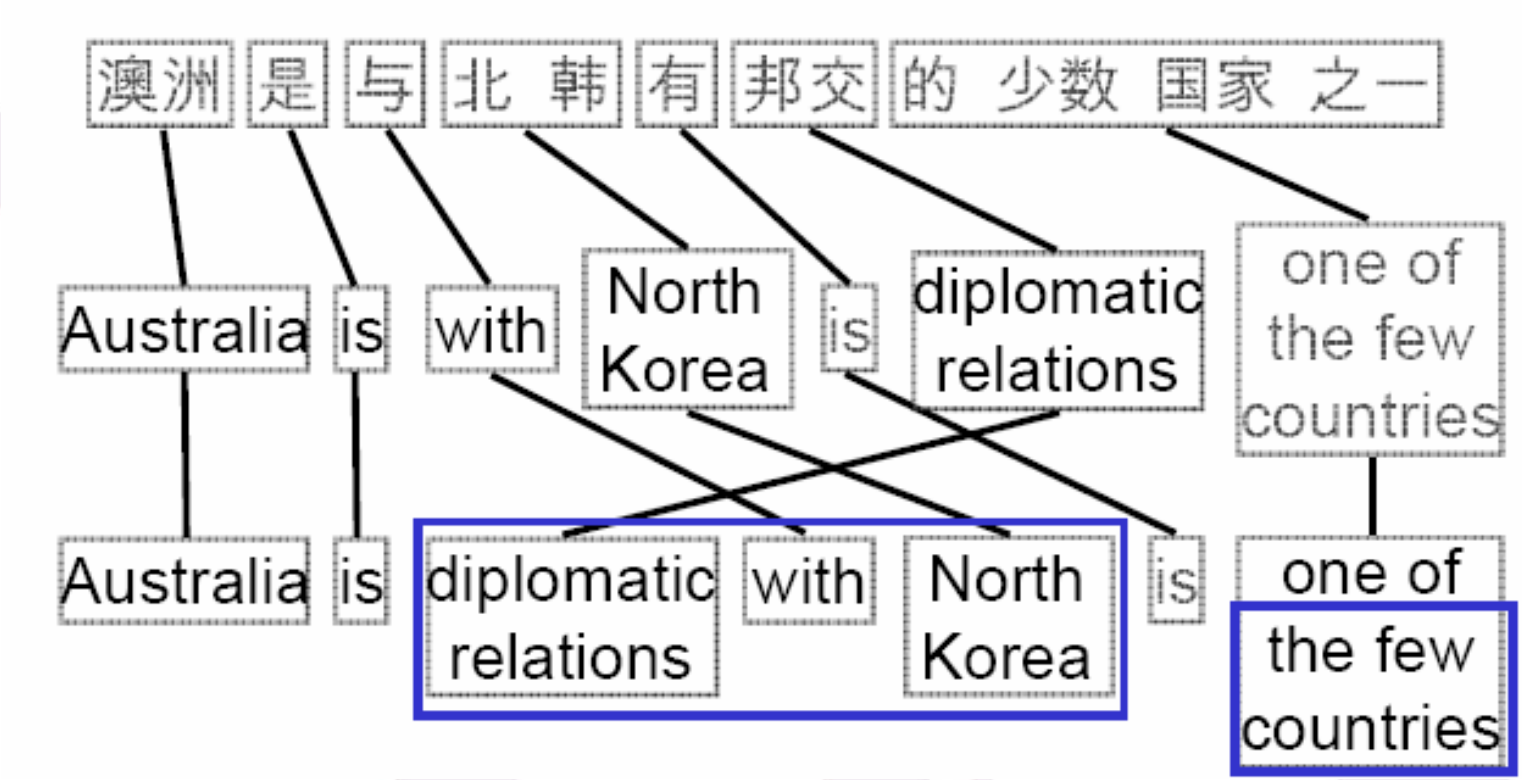
- **David Chiang. A Hierarchical Phrase-Based Model for Statistical Machine Translation. ACL2005. (Best Paper Award)**
- 本讲义这一部分内容直接引用了以下讲义的部分内容，特此说明并向原作者表示感谢：
  - **David Chiang, Hiero: Finding Structure in Statistical Machine Translation, in National University of Singapore**

层次短语模型 (2)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

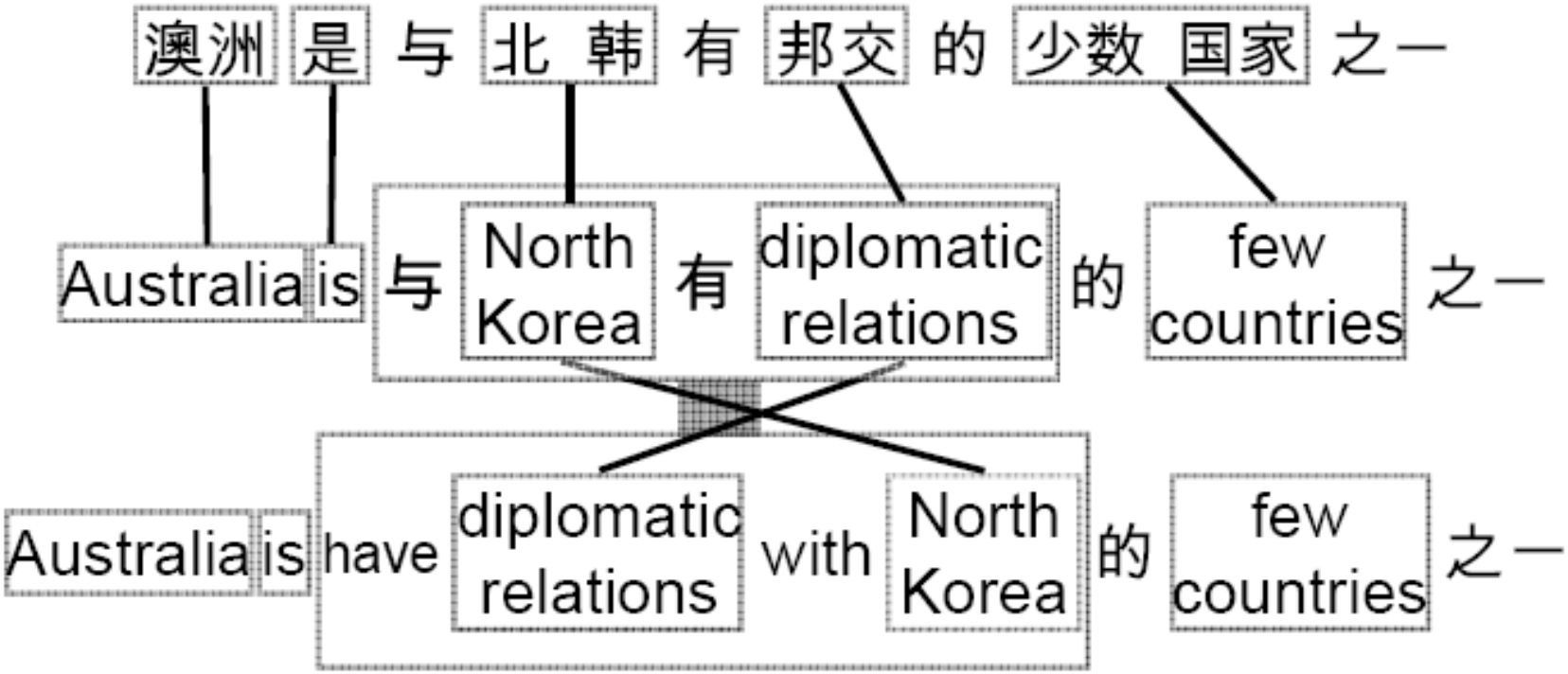
- 传统的基于短语的翻译模型中，短语是平面的，不能嵌套
- 在层次短语模型中，引入了嵌套的层次短语
- 采用平行上下文无关语法作为理论基础，但只使用唯一的非终结符标记
- 效果比传统的短语模型有很大提高

平面的短语

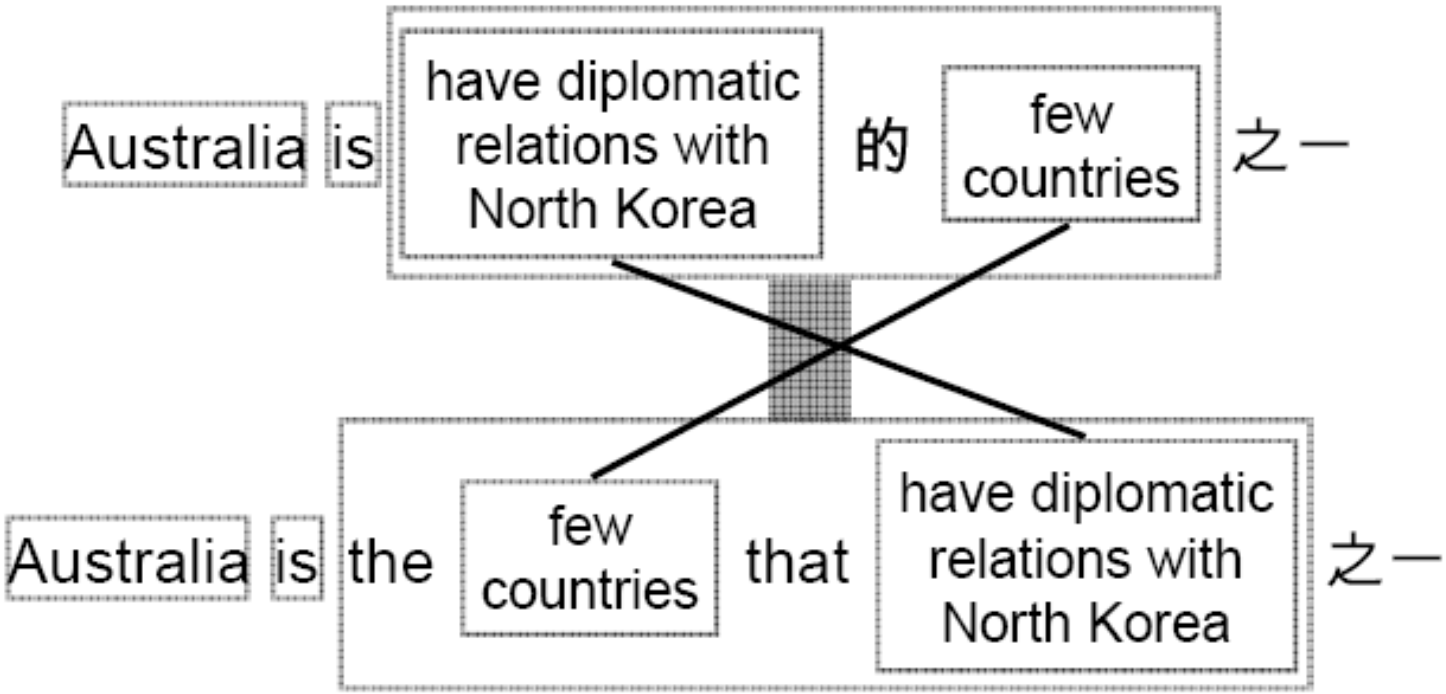


可以观察到短语是有层次的。

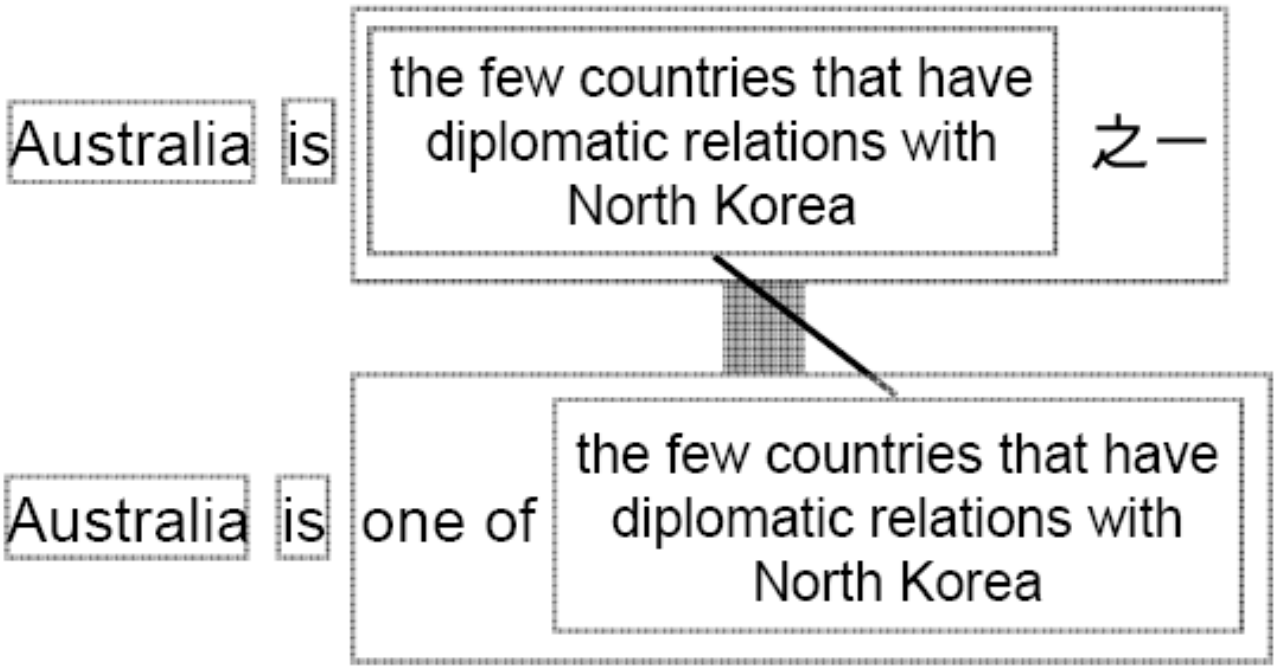
层次短语 (1)


澳洲 是 与 北 韩 有 邦交 的 少数 国家 之一
Australia is 与 North Korea 有 diplomatic relations 的 few countries 之一
Australia is have diplomatic relations with North Korea 的 few countries 之一

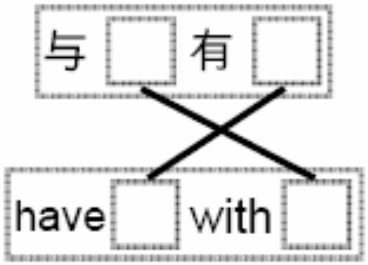
层次短语 (2)


Australia
is
have diplomatic relations with North Korea
的
few countries
之一
Australia
is
the
few countries
that
have diplomatic relations with North Korea
之一

层次短语 (3)

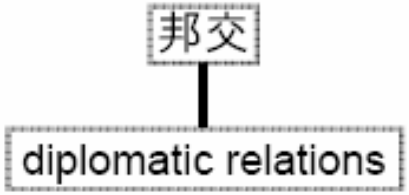

Australia
is
the few countries that have diplomatic relations with North Korea
之一
Australia
is
one of
the few countries that have diplomatic relations with North Korea

用同步语法表示层次短语 (1)



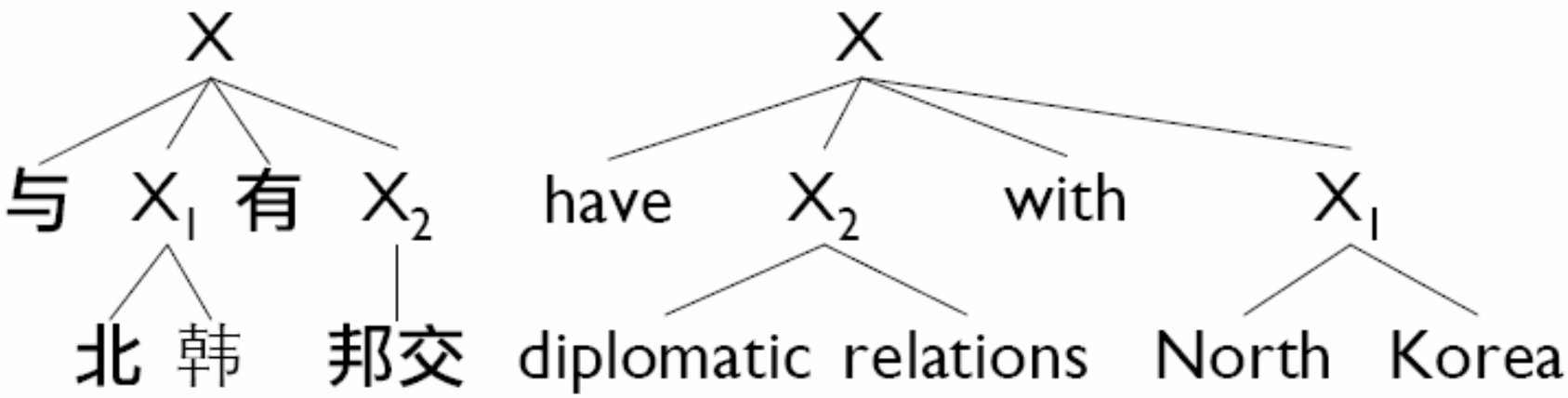
(X → 与 X_1 有 X_2, X → have X_2 with X_1)

(X → 北 韩, X → North Korea)

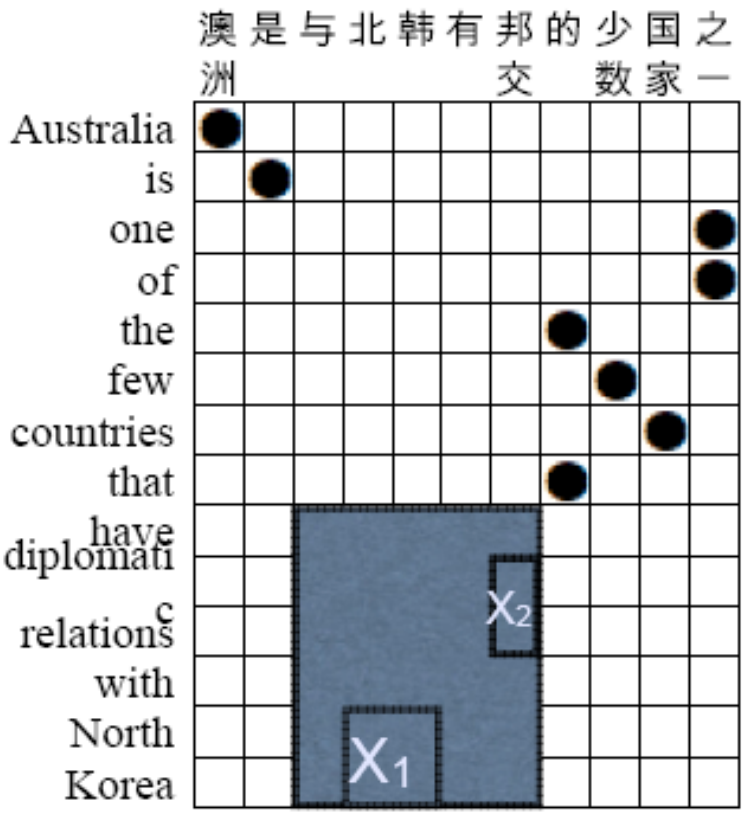


(X → 邦交, X → diplomatic relations)

用同步语法表示层次短语 (2)



层次短语的抽取



(与 北 韩 有 邦交, have diplomatic relations with North Korea)

(邦交, diplomatic relations)

(北 韩, North Korea)

(X → 与 X_1 有 X_2, X → have X_2 with X_1)

规则举例

$X \rightarrow$ 的	$X \rightarrow$'s
$X \rightarrow X_1$ 的 X_2	$X \rightarrow$ the X_2 of X_1
$X \rightarrow X_1$ 的 X_2	$X \rightarrow$ the X_2 that X_1
$X \rightarrow$ 在	$X \rightarrow$ in
$X \rightarrow$ 在 X_1 下	$X \rightarrow$ under X_1
$X \rightarrow$ 在 X_1 前	$X \rightarrow$ before X_1
$X \rightarrow$ 今年 X_1	$X \rightarrow X_1$ this year
$X \rightarrow X_1$ 之一	$X \rightarrow$ one of X_1
$X \rightarrow X_1$ 总统	$X \rightarrow$ president X_1

模型

- 直接利用同步上下文无关语法的概率模型
- 通过对数线性模型融合其他特征，如传统短语模型的各种特征

模型特征

- Language model $p(e)$
- Phrase translation probabilities $p(\bar{f} \mid \bar{e}), p(\bar{e} \mid \bar{f})$
- PCFG-like probability $p(\bar{f})$ (since all rules are $X \rightarrow \bar{f}$)
- Probability for glue rule $S \rightarrow SX$
- Word penalty, phrase penalty
- Constituent reward (optional)

解码

INSTITUTE OF COMPUTING TECHNOLOGY

- 类似于句法分析，在对源语言分析的同时，产生目标语言的结构。
- 算法复杂度$O(n^3)$

层次短语模型小结

- 形式上基于句法的模型
- 性能明显超过基于短语的模型
- 完全兼容基于短语的模型
- 规则采用同步上下文无关语法形式，但只有一个非终结符X
- 所有规则可以自动抽取
- 规则数量极为庞大

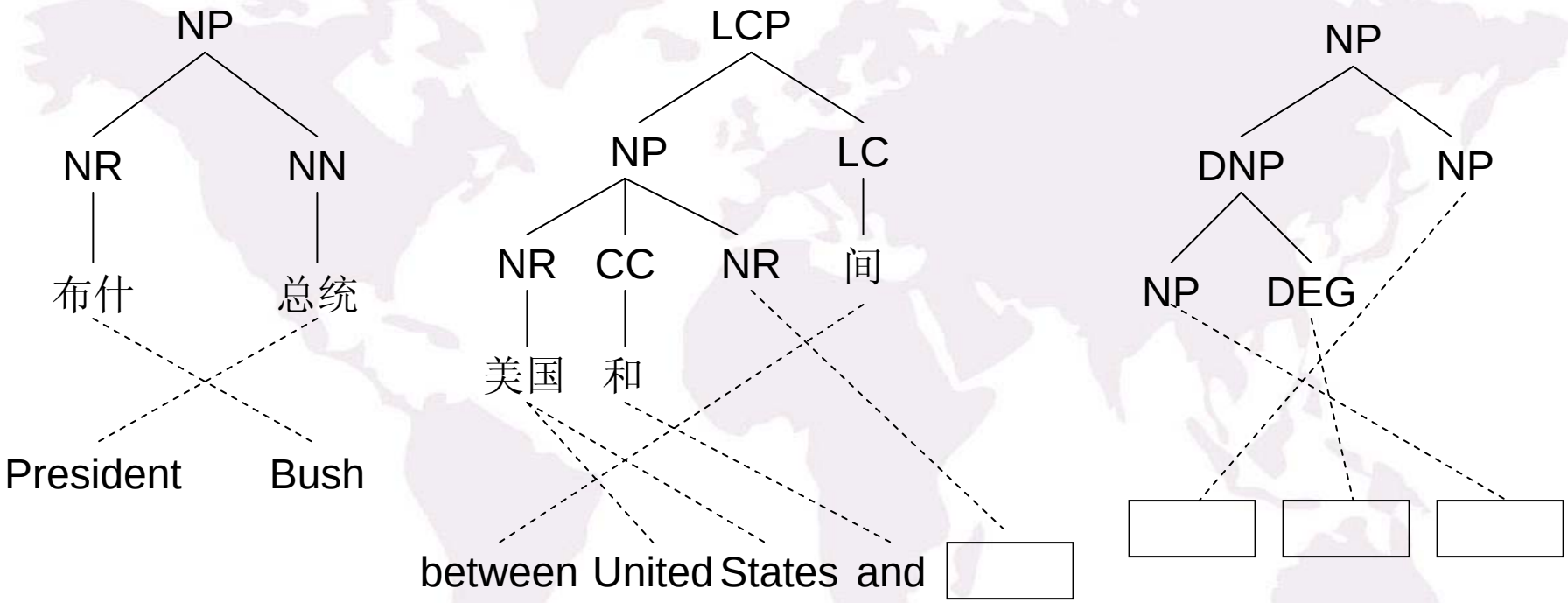
中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

基于树到串对齐模板的翻译模型

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 基于树到串对齐模板（简称**TAT**）的统计翻译模型是一种在源语言进行句法分析的基于语言学句法结构的统计翻译模型
- 树到串对齐模板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


NP
NR
NN
布什
总统
President
Bush
LCP
NP
LC
NR
CC
NR
间
美国
和
between
United
States
and
NP
DNP
NP
NP
DEG

模型特征

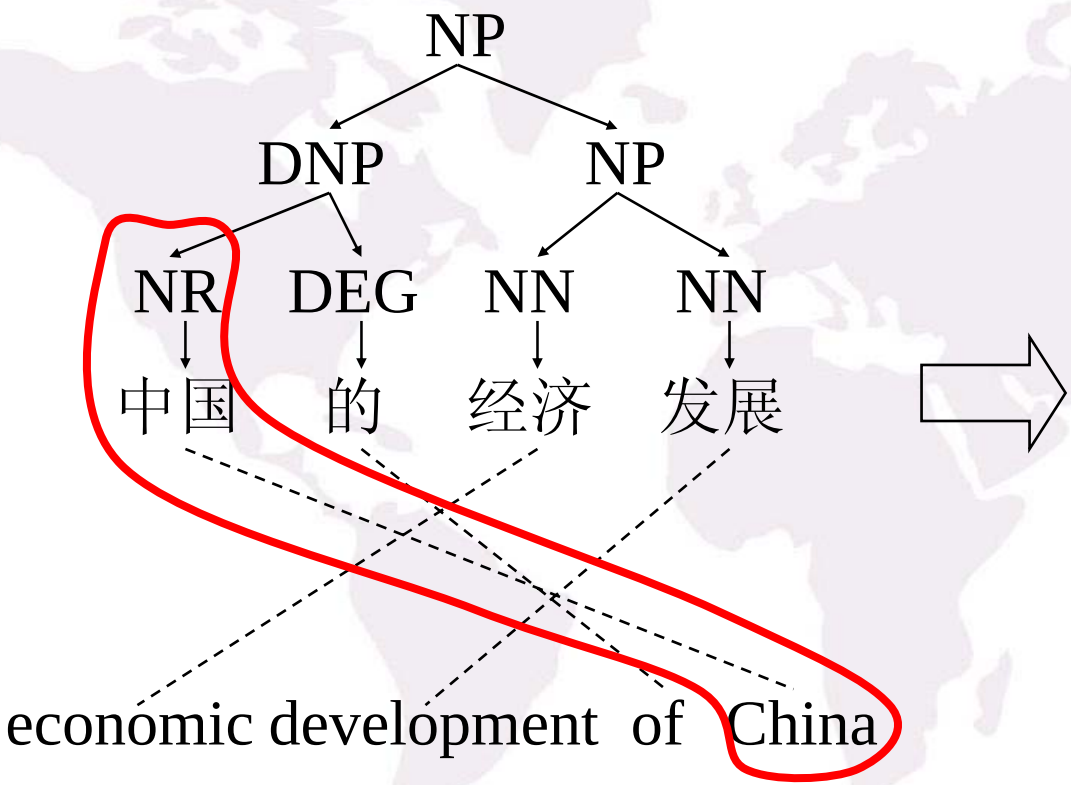
$$
\begin{aligned}
h_1(e_1^I, f_1^J) &= \log \prod_{k=1}^{K} \frac{N(z) \cdot \delta(T(z), \tilde{T}_k)}{N(T(z))} \\
h_2(e_1^I, f_1^J) &= \log \prod_{k=1}^{K} \frac{N(z) \cdot \delta(T(z), \tilde{T}_k)}{N(S(z))} \\
h_3(e_1^I, f_1^J) &= \log \prod_{k=1}^{K} lex(T(z)|S(z)) \cdot \delta(T(z), \tilde{T}_k) \\
h_4(e_1^I, f_1^J) &= \log \prod_{k=1}^{K} lex(S(z)|T(z)) \cdot \delta(T(z), \tilde{T}_k) \\
h_5(e_1^I, f_1^J) &= K \\
h_6(e_1^I, f_1^J) &= \log \prod_{i=1}^{I} p(e_i|e_{i-2}, e_{i-1}) \\
h_7(e_1^I, f_1^J) &= I
\end{aligned}
$$

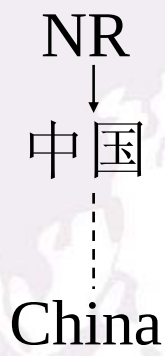
模板的约束

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

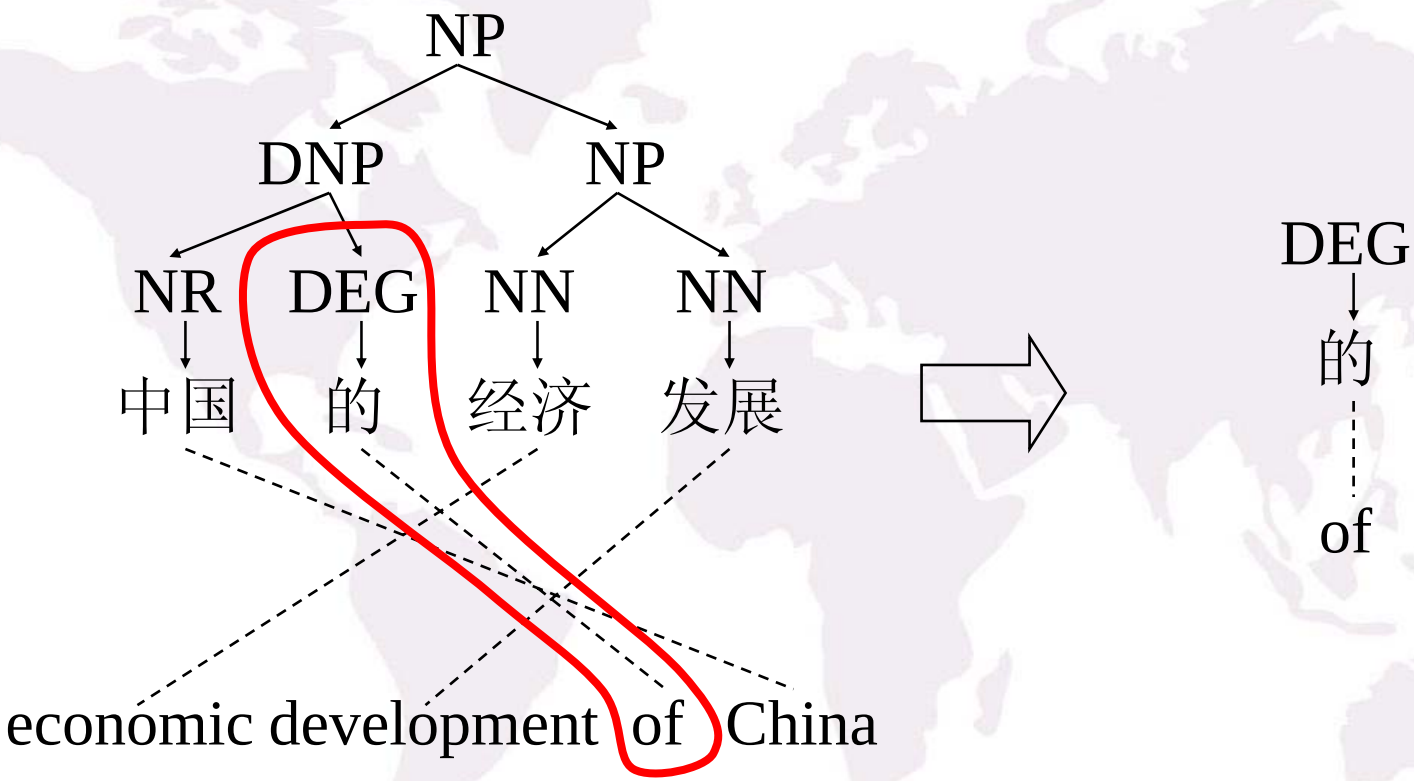
- **Constraints**
  - 源语言子树必须是完整的子树，也就是说，不能只取一个父节点的部分子节点，而忽略另一部分子节点
  - 源语言子树和目标串之间必须满足对齐约束，也就是不能对齐到外部的节点
- **Restrictions on extraction**
  - 目标串的第一个词和最后一个词都必须有对齐
  - 对源语言子树的高度加以限制
  - 对源语言子树中一个父节点的子节点个数加以限制

树到串对齐模板的抽取 (1)

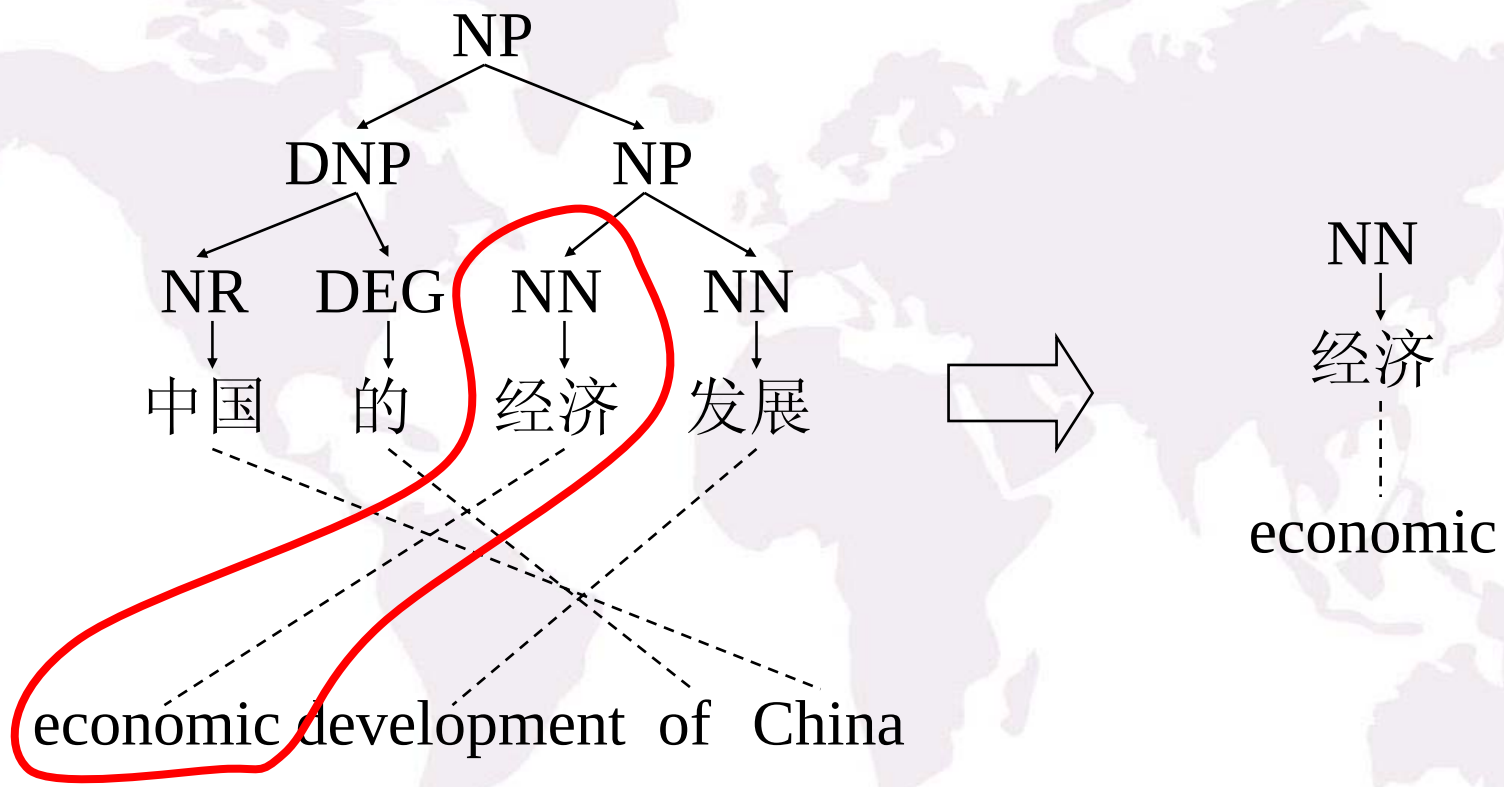




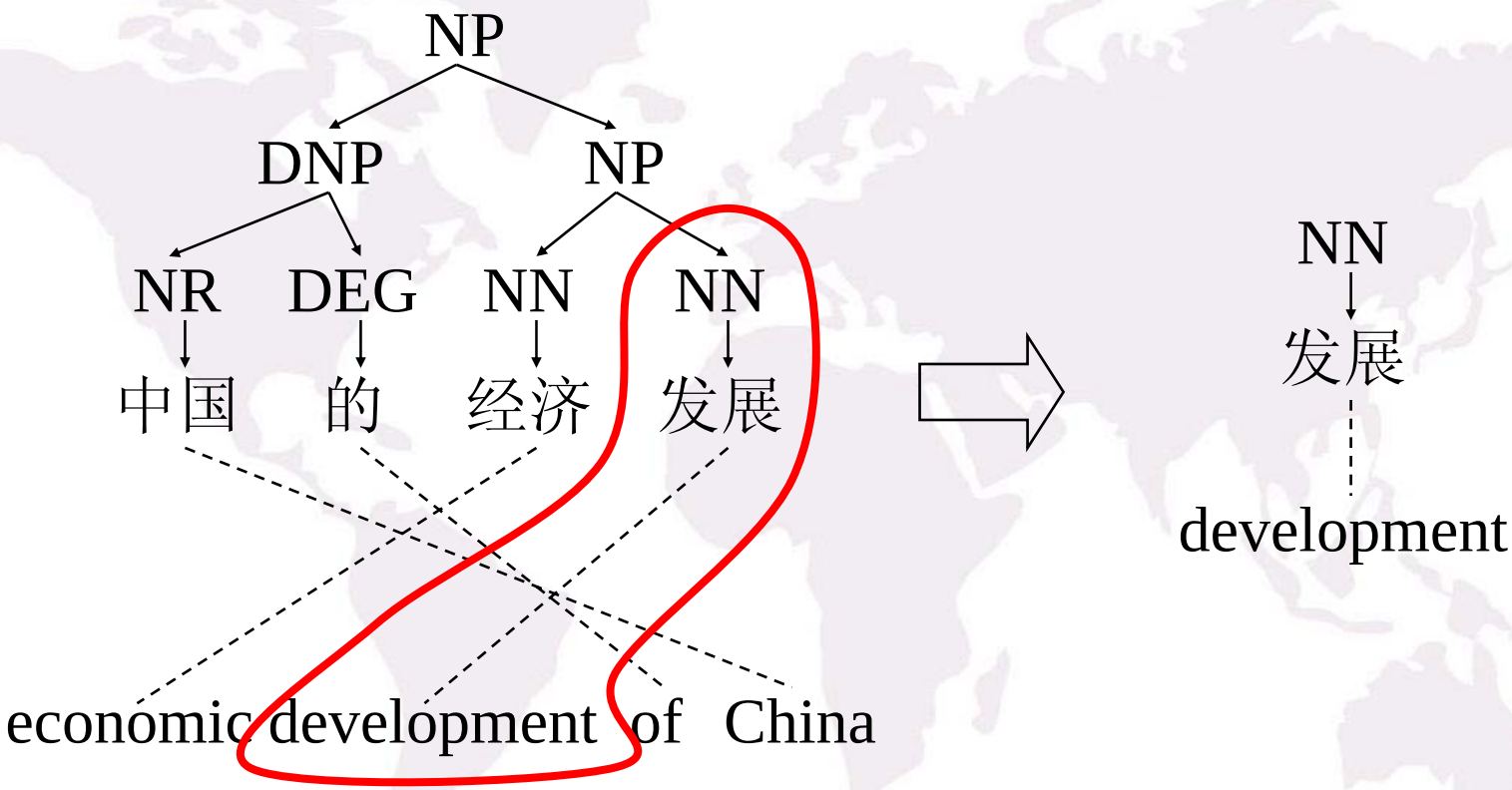
树到串对齐模板的抽取 (2)



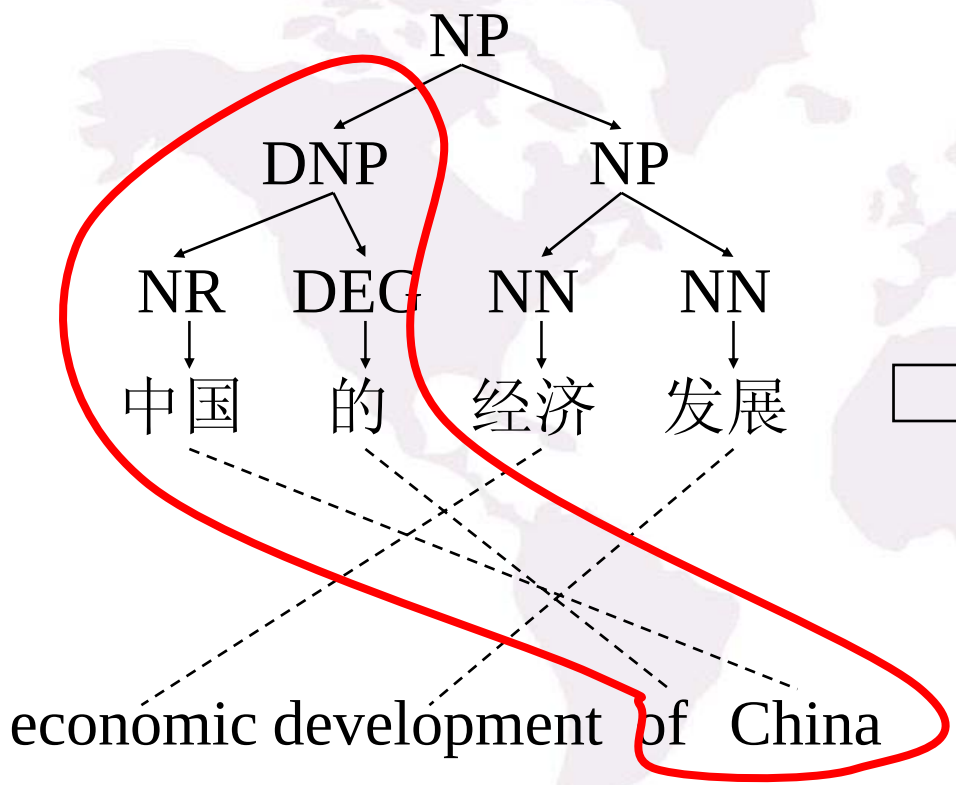
树到串对齐模板的抽取 (3)

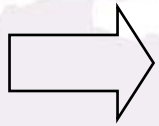


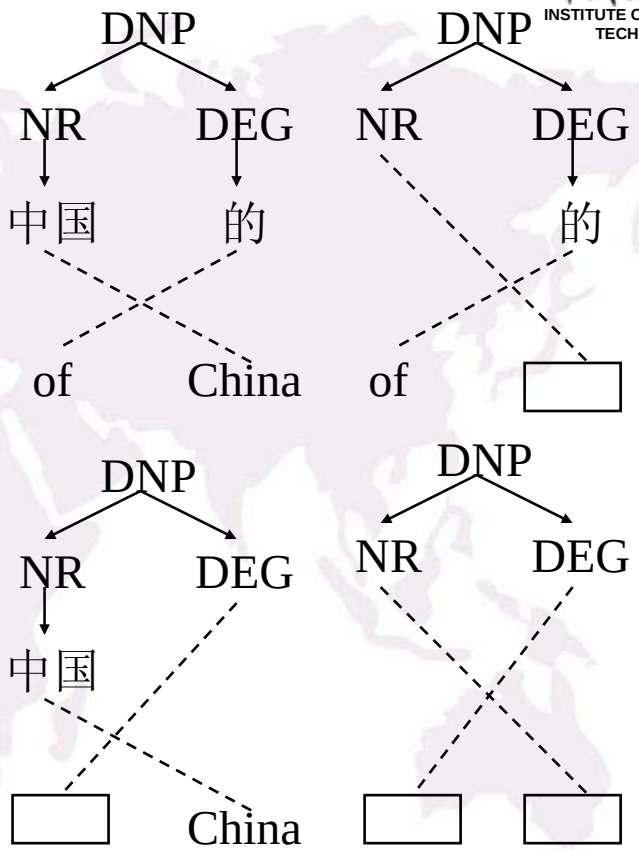
树到串对齐模板的抽取 (4)



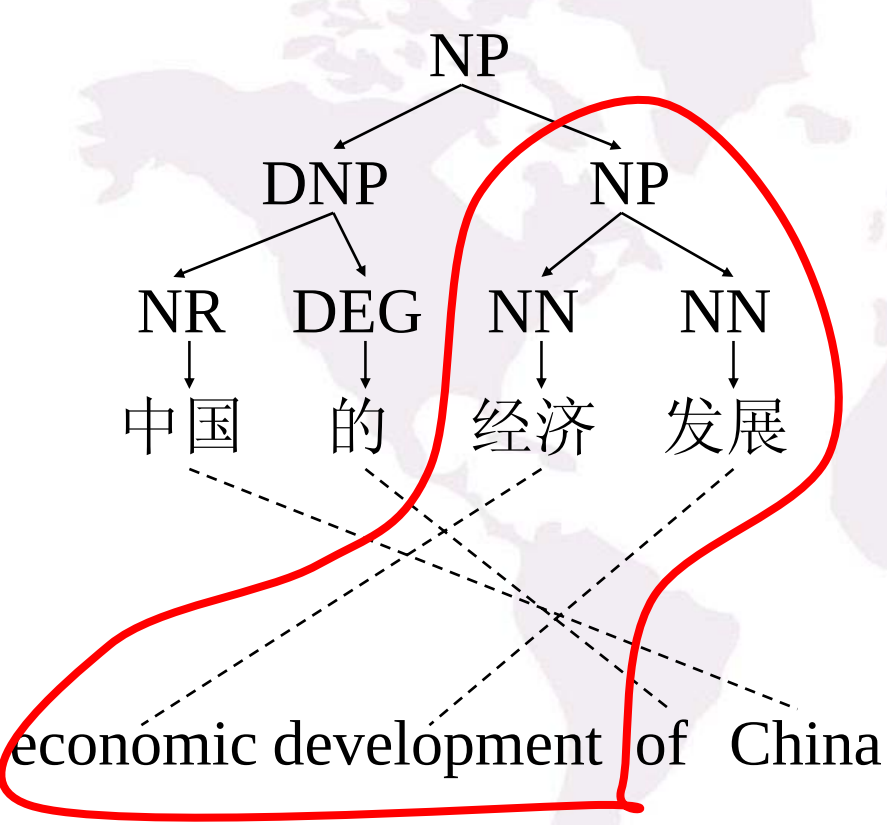
树到串对齐模板的抽取 (5)

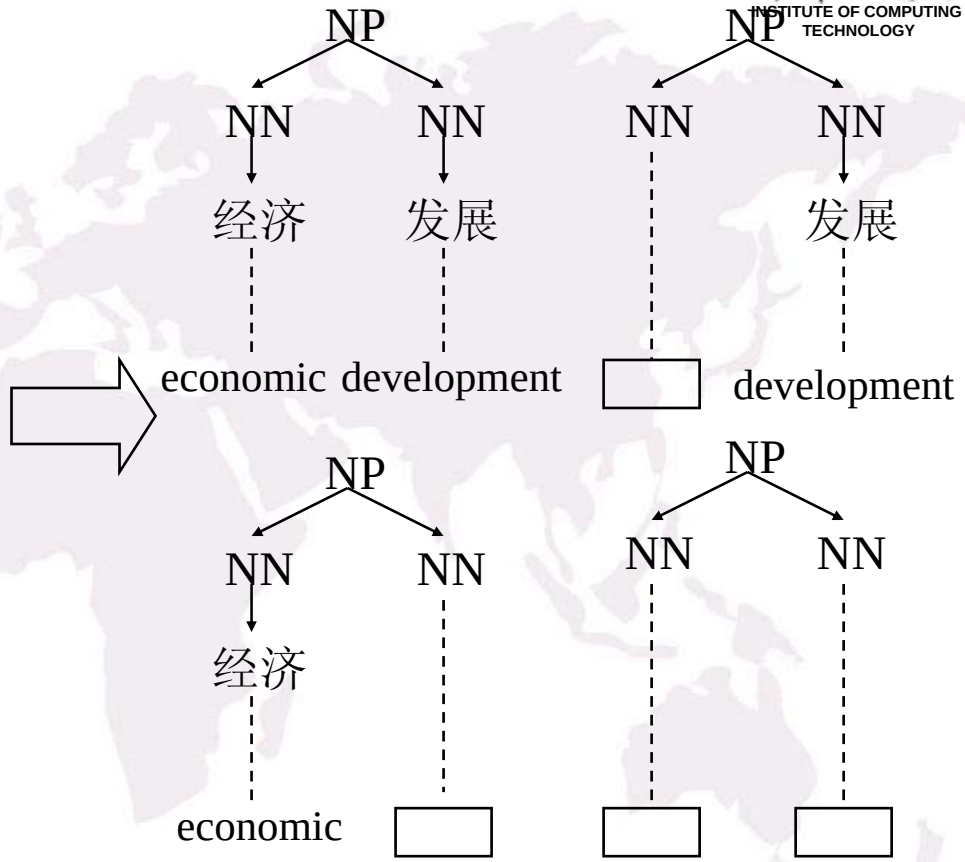




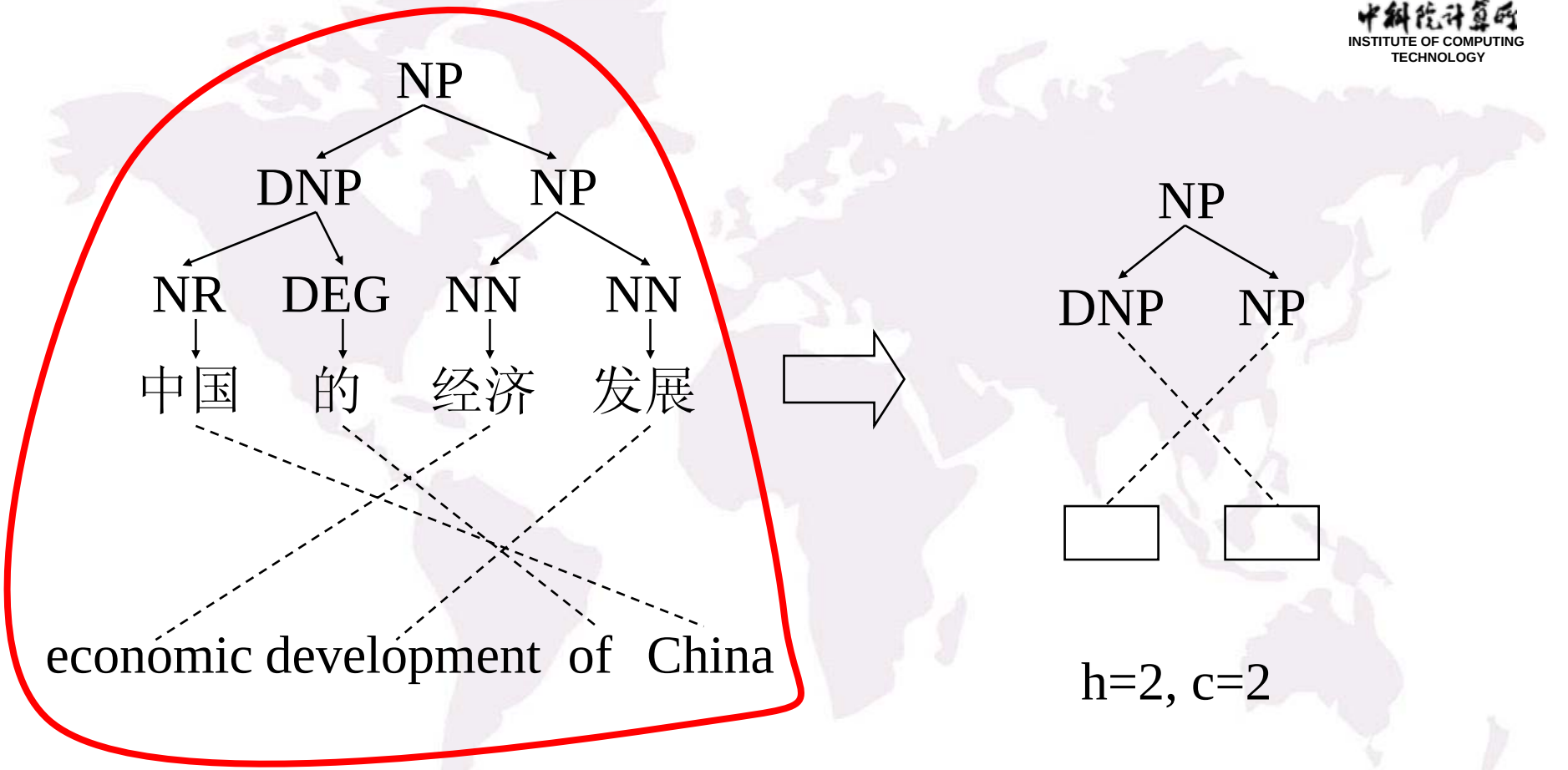


树到串对齐模板的抽取 (6)

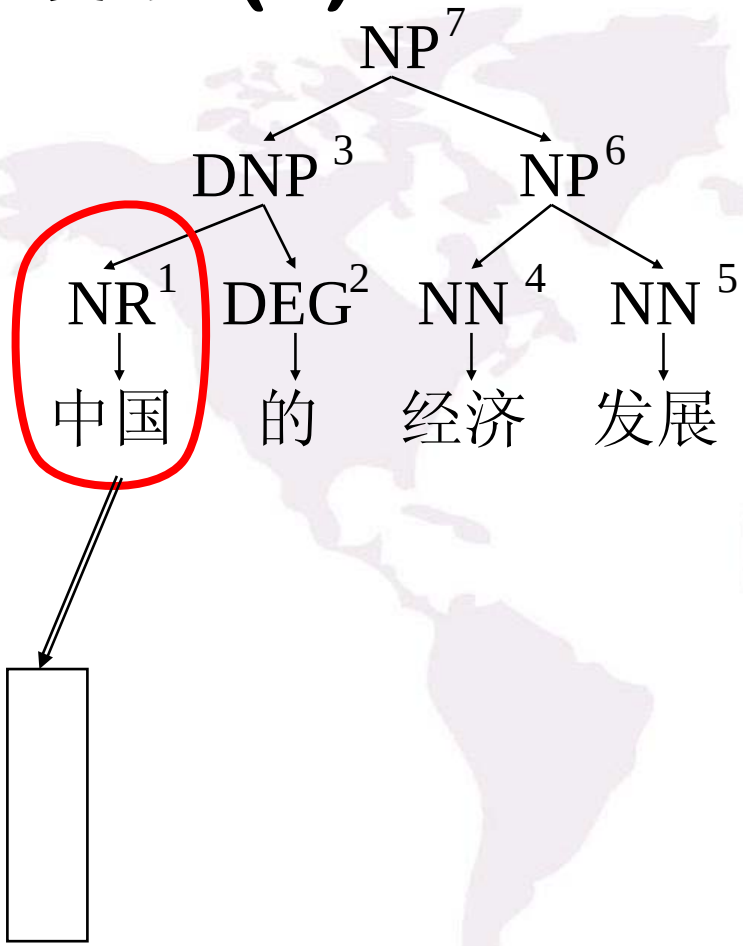




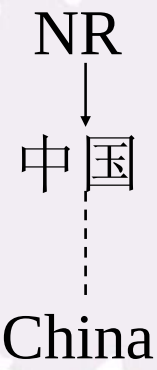
树到串对齐模板的抽取 (7)



搜索 (1)



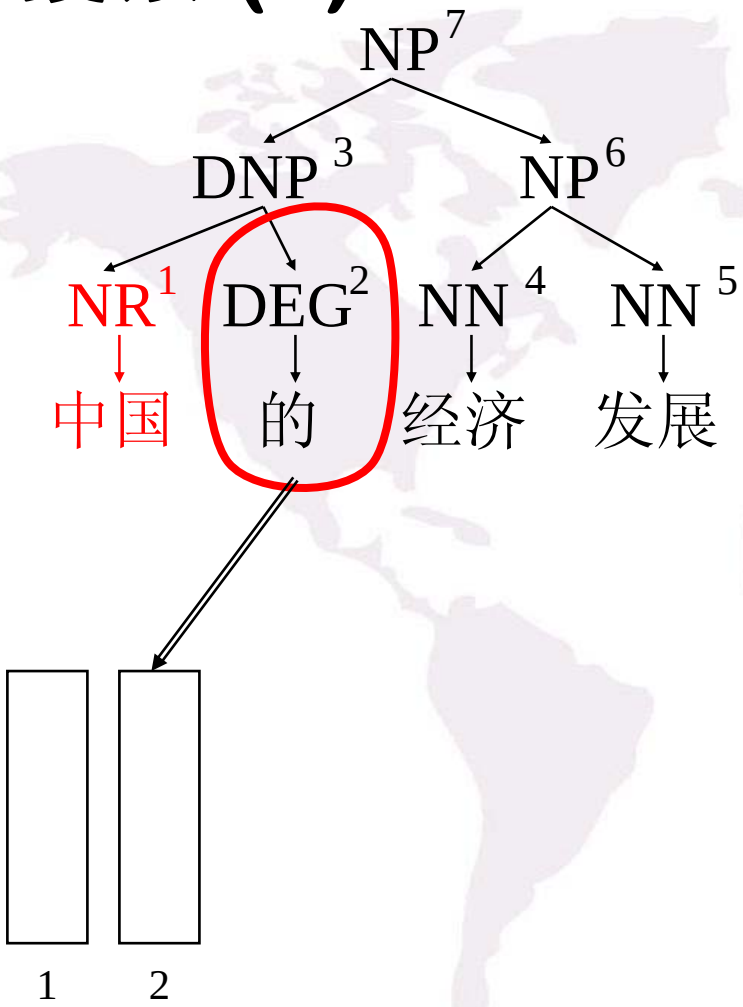
TAT



译文

China

搜索 (2)



TAT

DEG

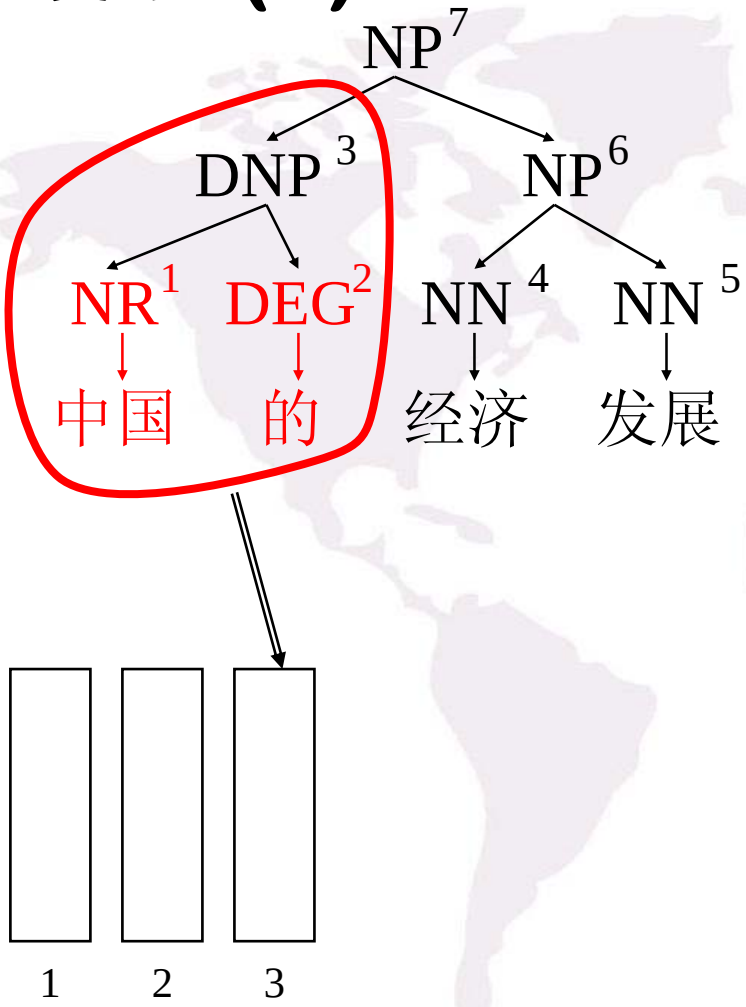
的

of

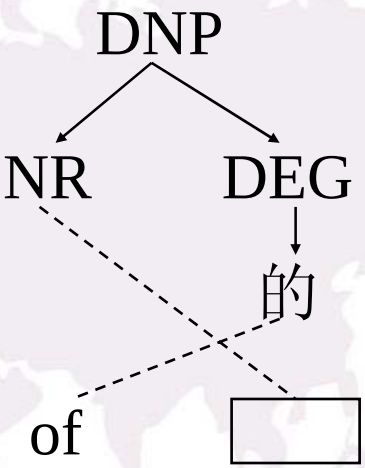
译文

of

搜索 (3)



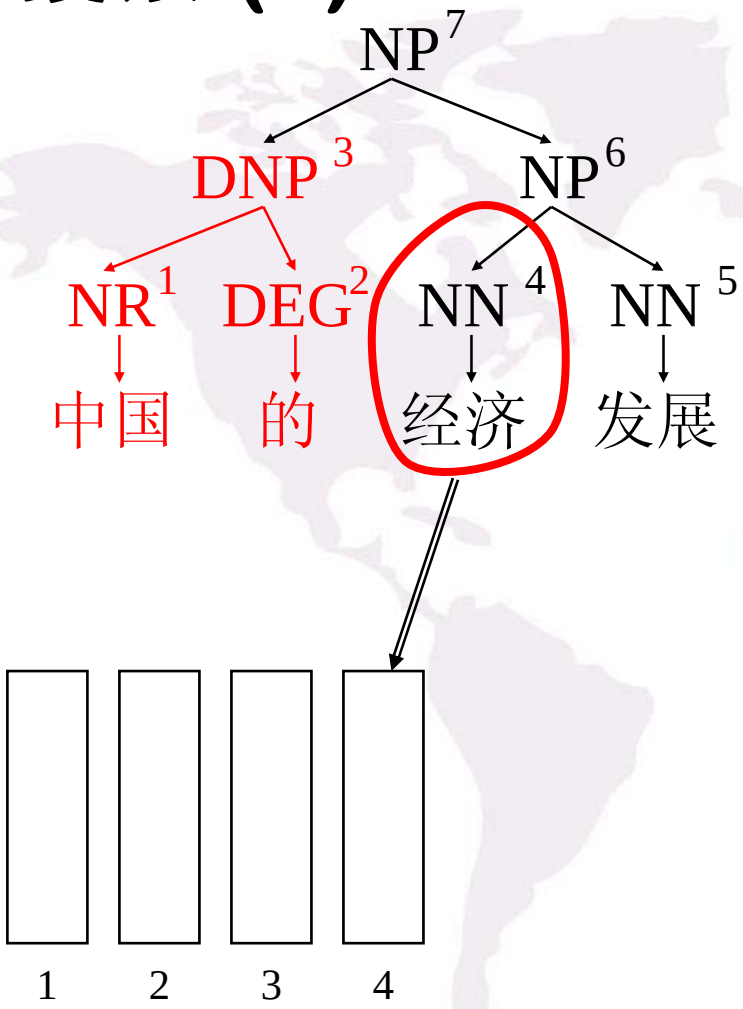
TAT



译文

of China

搜索 (4)



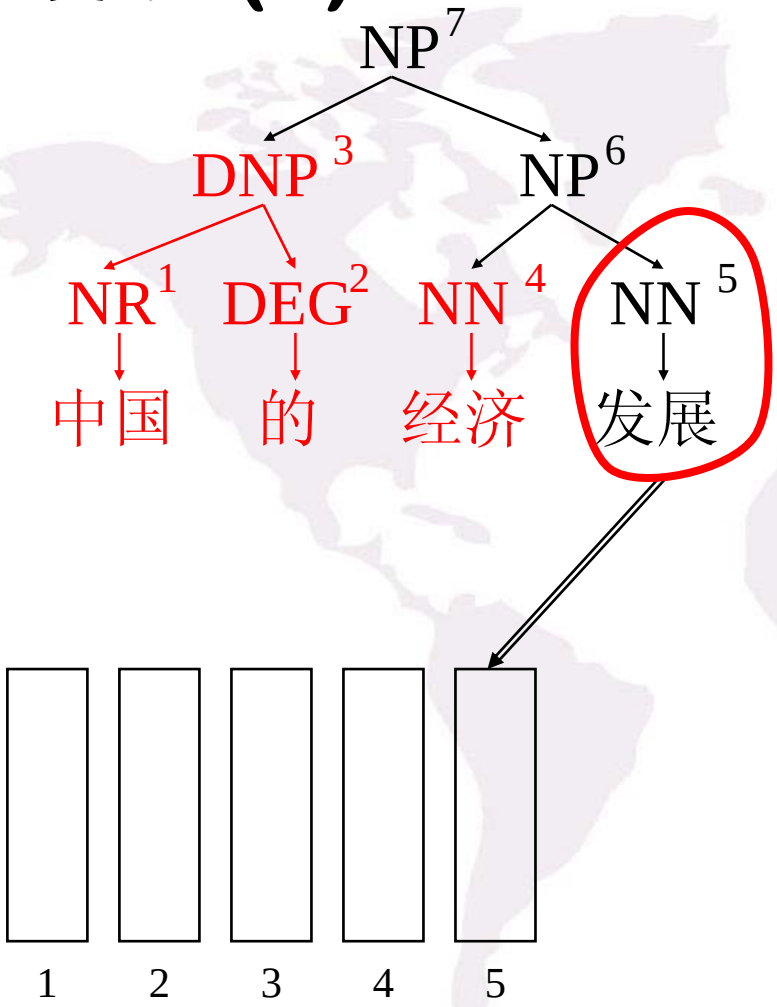
TAT

economy

译文

economy

搜索 (5)



TAT

NN

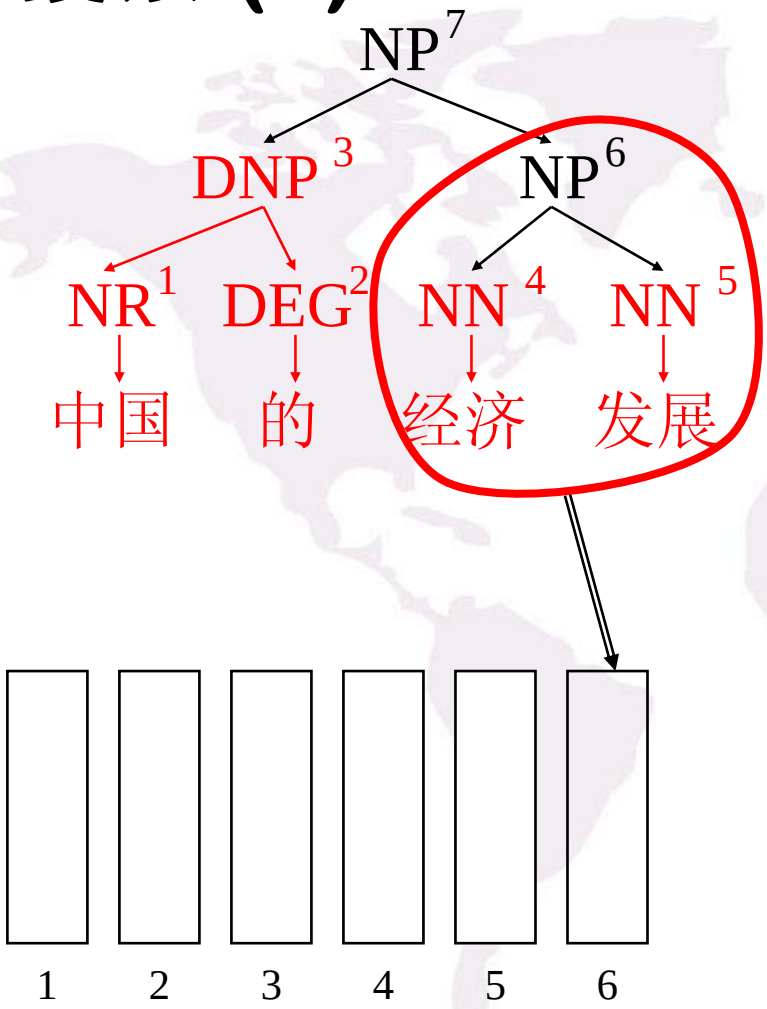
发展

development

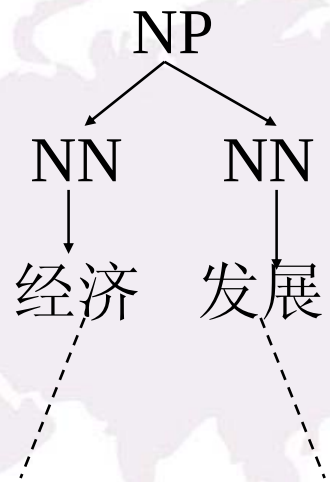
译文

development

搜索 (6)



TAT

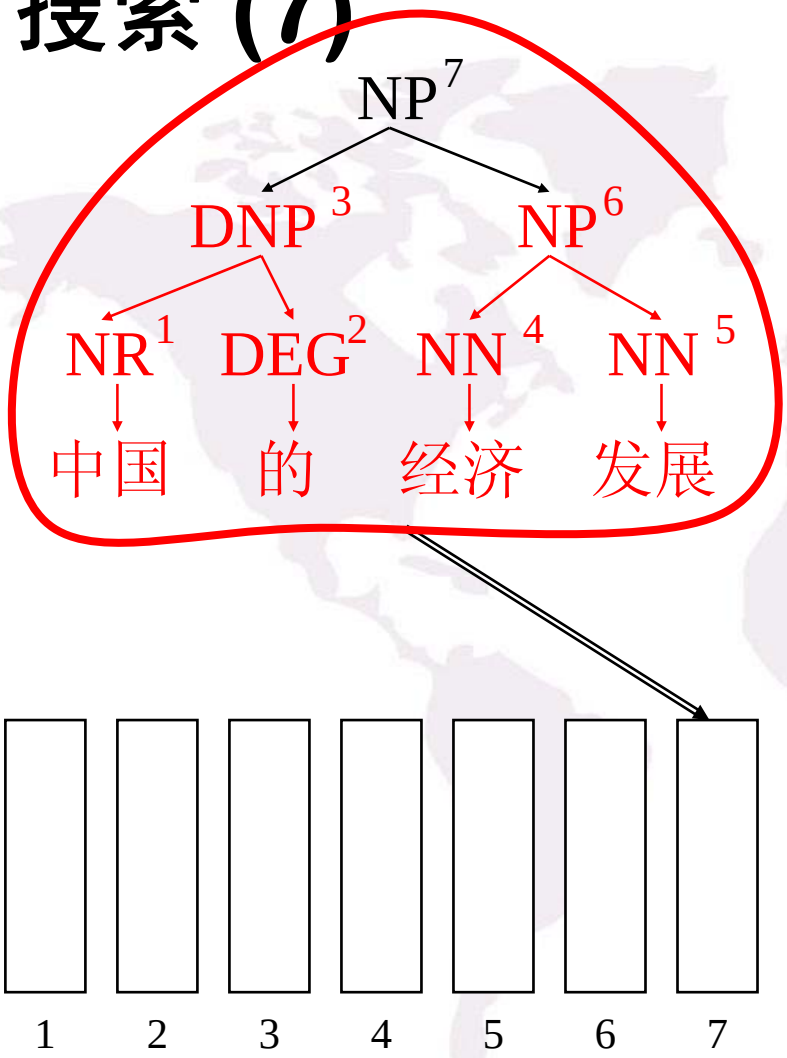


economic development

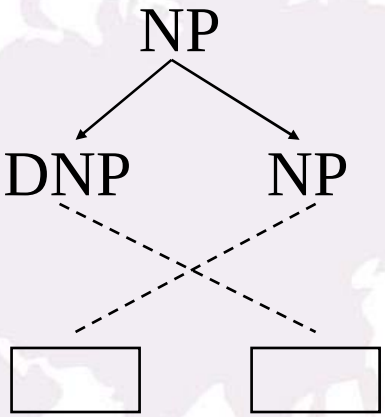
译文

economic development

搜索 (7)



TAT



译文

economic development of China

基于树到串模板的统计翻译模型小结

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 一种语言学上基于句法的模型
- 训练时除了双语词语对齐外，还要对源语言进行句法分析
- 树到串对齐模板**TAT**的源语言端是一个子树，不是一条上下文无关语法的规则，等价于同步数替换语法（**STSG**）
- 不完全兼容于基于短语的模型（其扩展形式“基于森林到串模板的统计翻译模型”可以兼容于基于短语的模型）
- 所有规则全自动抽取
- 规则数量极为庞大
- 解码时要先利用传统的句法分析器进行源语言句法分析，然后采用基于句法树的堆栈搜索
- 性能比基于短语的模型有显著提高

串到树的统计翻译模型 (1)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- **USC-ISI**的系列工作
- 发表了大量论文，但还没有一个完整的论述
- 性能优异，在**NIST2006**汉英项目平常中超过了**Google**（**Google**使用的语言模型规模比**ISI**大得多）

串到树的统计翻译模型 (2)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 基本思想
  - 在目标语言端进行句法分析
  - 根据目标语言端的句法结构，和词语对齐，建立源语言端的句法结构（伪树）
  - 利用两个句法结构自动抽取带概率的平行上下文无关语法
  - 对平行上下文无关语法进行二叉化
  - 解码时类似规则方法，复杂度等价于句法分析
    - 源文分析
    - 规则映射
    - 译文生成

串到树的统计翻译模型小结

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 一种语言学上基于句法的模型
- 训练时除了双语词语对齐外，还要对目标语言进行句法分析
- 规则形式是同步上下文无关语法形式
- 不完全兼容于基于短语的模型所有规则全自动抽取（修改后可兼容）
- 规则数量极为庞大
- 不需要利用传统的句法分析器进行句法分析
- 解码过程等价于句法分析过程
- 性能比基于短语的模型有显著提高

目录

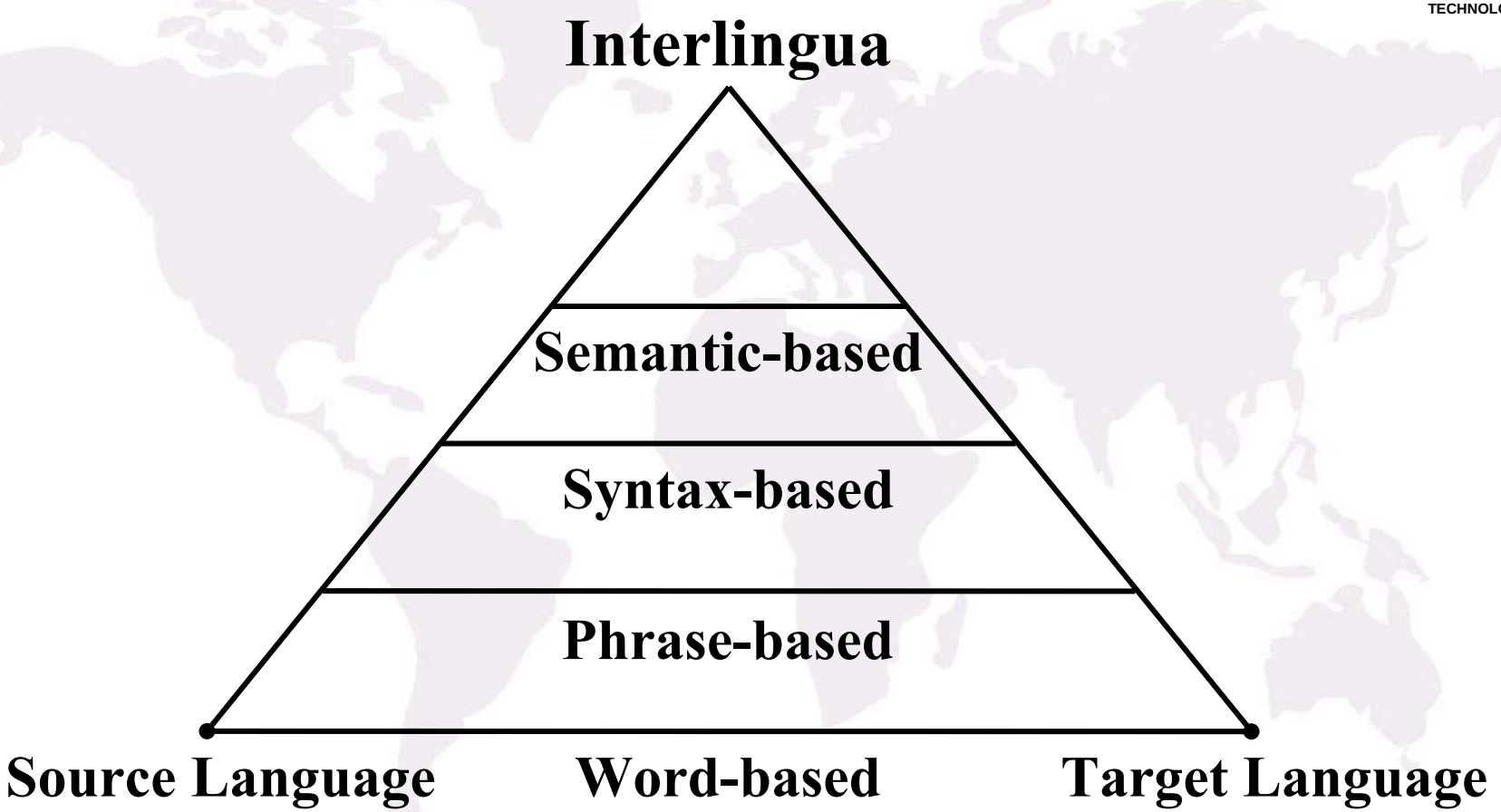
- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
  —基于词的IBM模型
- 最成熟的统计机器翻译方法
  —基于短语的模型
- 目前统计机器翻译研究的热点
  —基于句法的模型
- 统计机器翻译面临的问题和展望

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

总结

- 综上所述，统计机器翻译的发展可以清理出两条主线的进展
  - 框架模型的进展
    - 信源信道模型
    - 对数线性模型
  - 翻译模型的进展
    - 基于词的模型
    - 基于短语的模型
    - 基于句法的模型

统计翻译模型的进展 (1)


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

统计翻译模型的进展 (2)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 我们看到，统计机器翻译与基于规则的机器翻译一样，也存在一个金字塔形的发展过程
- 基于词的模型是**1990**年前后由**IBM**公司提出来的
- 基于短语的模型是**Och**、**Zens**、**Koehn**等人**2003**年前后提出来的，是目前最成熟稳定、也是最普遍采用的模型
- 基于句法的模型是目前的研究热点，虽然很早就有人开展研究，但真正取得较好的结构是在**2005**年以后，典型的工作有**[Chiang 2005]**、**ISI**的工作、中科院计算所的工作。

统计翻译模型的进展 (3)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 对于基于规则的方法
  - 在理想的情况下，如果语言分析（主要是句法分析和语义分析）完全正确，那么，转换的层次越深，可以利用的信息就越多，应该可以达到更好的翻译效果。
  - 由于语言分析的过程中总会引入各种各样的错误，而且这些错误还会随着分析层次的加深而逐渐积累起来，从而导致翻译的错误，因此，并不是在越深层次上进行转换所获得的翻译系统性能就越好，而是需要取一个折衷的“最优”。
  - 目前，大部分基于规则的机器翻译系统都是在句法或者语义层面进行转换。

统计翻译模型的进展 (4)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 对于统计方法
  - 在统计方法中，沿金字塔向上攀升是一个艰难的过程，从基于词的方法到基于短语的方法，经过了十几年的发展才成熟起来，而基于短语的方法到基于句法的方法，经过几年的发展，虽然成为了目前研究的热点，但还没有成为主流的做法
  - 早期很多基于短语的模型和基于句法的模型研究都已失败告终，有些虽然报告取得了较好的结果，但在没有人能够重复的情况下，依然无法被研究界所接受

统计翻译模型的进展 (5)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 语言知识与统计模型的融合
  - 语言知识与语言模型结合成为研究的趋势
  - 统计模型中如果不引入语言知识，其性能很难再有大的提高
  - 语言知识如果不能与统计模型相结合，就回到了基于规则的老路上去，所面临的知识获取和冲突的问题无法从根本上得到解决

统计翻译模型的进展 (6)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 在统计模型中加入语言知识是一个非常困难的工作。语言知识的加入会大大增加模型的复杂性，使得系统的性能和能够处理的规模都大大下降；在大部分情况下，由于语言知识的覆盖率和正确率不够，使得系统的整体性能反而下降了。一种融入语言知识的统计模型，只有经过精心设计、反复实验、不断改进，才能取得成功。而且这种成功只有经过其他研究人员的重复才能被学术界广泛接受。
- 在统计模型中加入语言知识很难一蹴而就，只能由浅层知识到深层知识，由简单到复杂，循序渐进地实现

统计翻译模型的进展 (7)

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

- 基于句法的统计翻译模型目前还是研究热点
- 在统计翻译中应用语义知识也有人开始研究，他们的研究结果表明**WSD**对统计翻译能够有所贡献，但并不太大
  - **Dekai Wu, ACL2007**
  - **Hwee Tou Ng, EMNLP2007**

展望

- 能否建立更有效的基于句法的统计翻译模型？
- 能否建立有效的基于语义的统计翻译模型？
- 更多的丰富多彩、形式多样的语言知识如何融入统计翻译模型？
- 在将语言知识融入统计模型时，如何克服语言知识的错误对机器翻译带来的反面影响？
- 框架模型是否还有更新的可能？
- 语言模型是否还能进行改进？（比如建立基于句法的语言模型）

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY

谢谢！