第十章 机器翻译 (3/3)

3、CASIA SMT系统

10.3.1 Models

- Many researches have proven that the phrase-based SMT may outperforms other models
- CASIA Chinese-to-English SMT system:
  - Phrase-based
  - Using templates with variables
  - Different tracing back algorithms in decoding

10.3.2 CASIA SMT System

- **The phrase-based SMT model**

$$p(e|c) = p_T(c|e)^{\lambda_t} \times p_L(e)^{\lambda_l} \times p_D(e,c)^{\lambda_d}$$

Where, $p_T(c|e)$ is the translation model;

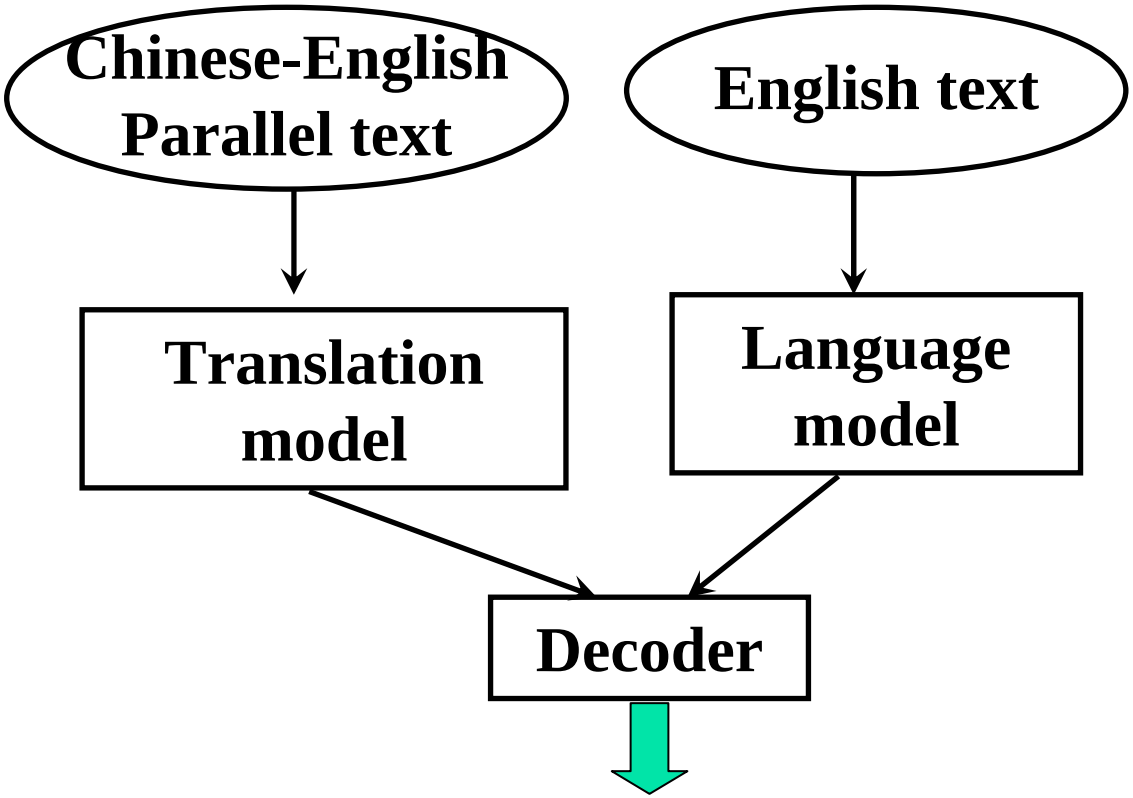
$p_L(e)$ is the target language model;

$p_D(e, c)$ is the distortion model and

$p_D(e, c) = \lambda |a_i - b_{i-1} - 1|$

10.3.2 CASIA SMT System

- Components: Translation model, language model, decoder



10.3.2 CASIA SMT System

◆ Phrase Translation Model

Methods for learning phrase translation:

❖ Integrated segmentation and phrase alignment (ISA) [Zhang, 2003]

❖ HMM word alignment model [Vogel, 1996]

❖ GIZA++ toolkit

10.3.2 CASIA SMT System

- **e.g., Phrase translation for “我 要 买”**

English Phrase	$\phi(f/e)$
I want to buy	0.386
I would like to buy	0.234
I will buy	0.119
I wanna buy	0.108
I wan to get	0.101
……	……

10.3.2 CASIA SMT System

- **Language Model**
  - Standard Technique: Trigram Model
  - SRI language modeling toolkit from http://www.speech.sri.com/projects/srilm/

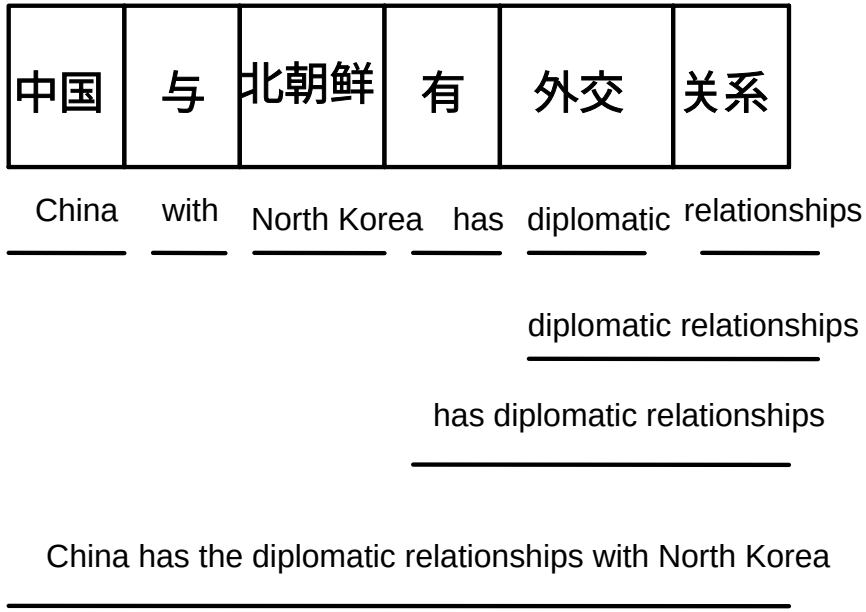
10.3.2 CASIA SMT System

◆ Beam-search for decoding

-Look up possible phrase translations [Koehn, 2003]

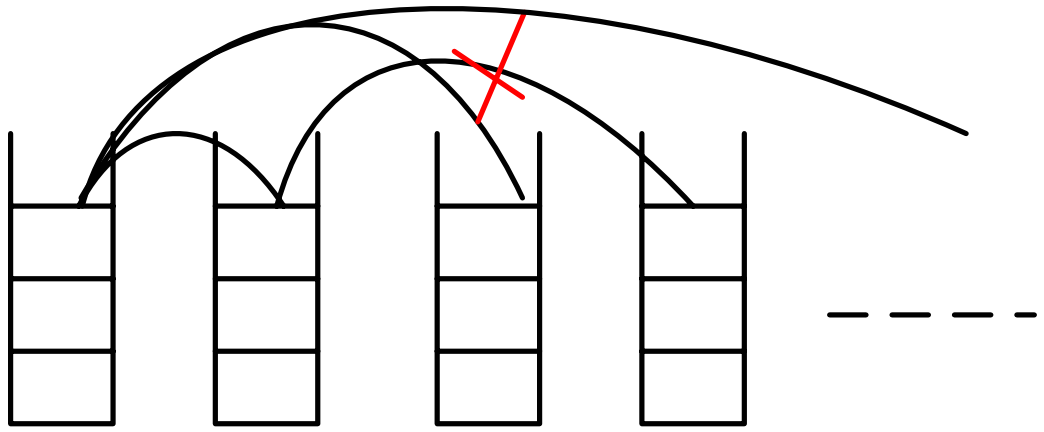
❖ Different ways to segment words into phrases

❖ Different ways to translate each phrase



10.3.2 CASIA SMT System

- The hypotheses are stored in different stacks
  - Bin label: number of Chinese words covered (One, Two)
  - Insert the Functional Words (of, the, for…..) bins.



F-Words must expand to Non F-Words Next

10.3.2 CASIA SMT System

❖*Considering there are some auxiliary words and mood words in the Chinese sentence, and these words sometimes don't have the corresponding words in the English sentence. So, the system does not require the all words in the Chinese sentence to be translated.*

Select the candidate sentences generated after L- a Chinese words have been translated.

L is the length of the input Chinese sentence;

a is an integer.

10.3.2 CASIA SMT System

For example:

L = 7 words

input: 我 想 预订 一 个 单人间

(I want to reserve a single room)

if a =1, *input* = 我 想 预订 一 个

if a =2, *input* = 我 想 预订 一

I would like to book one
I want to reserve one
… …

I want to book one
I would like to reserve one
… …

10.3.2 CASIA SMT System

- **Pre-processing of numerals using rule-based approach**

  - Arabic numbers: telephone No., room No. etc.

  - Number written in Chinese, such as “一百 one hundred”

  - Ordinal numbers:

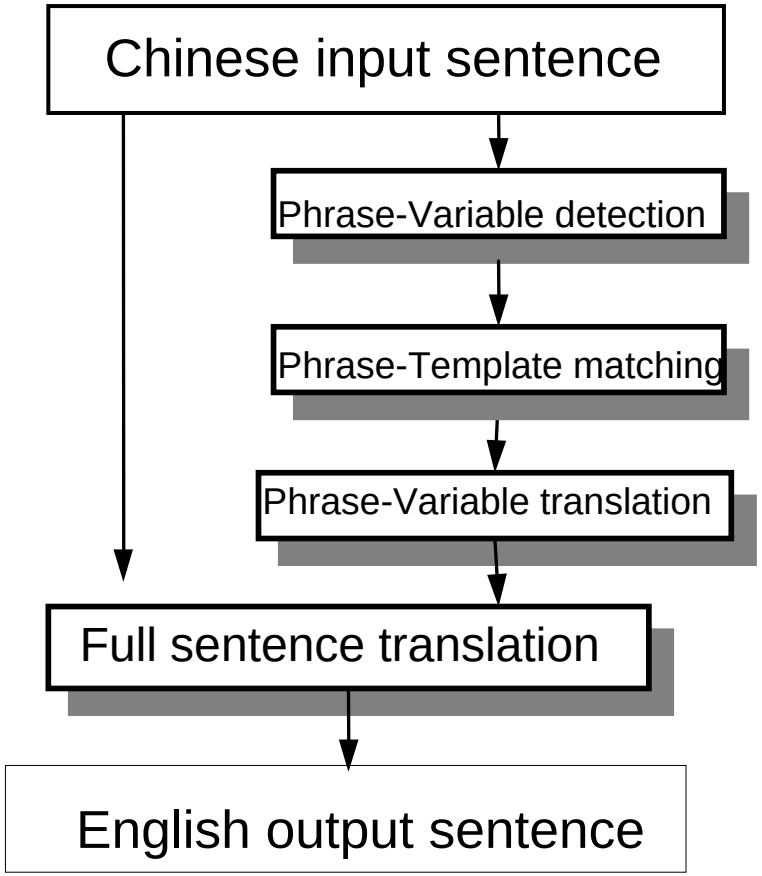
  - Dates

  - Combination of the different expressions

10.3.2 CASIA SMT System

- After pre-processing, the numerals are replaced with the specific marks (variables).
- The phrase translations become the templates with variables, e.g.,

  X个单人间 -> X single rooms



10.3.3 Experiments

- **Experiment-1**

  - Comparison of the different methods for learning phrase translations using 100,000 sentence pairs

Methods	Num of Phrases	BLEU4
ISA	187,011	0.2751
HMM	278,770	0.2637
GIZA++	695,486	0.2882

10.3.3 Experiments

❖ **Experiment-2**

- Comparison of the different searching algorithms using 100,000 sentence pairs

Marks: *M means word-based translation model;*

+NF0 means the Functional-zero words are not applied;

+F0 means the Functional-zero words are applied;

+NUM means the numerals are pre-processed;

+BACK1 stands for our decoder;

+BACK2 stands for Koehn's decoder.

10.3.3 Experiments

Methods	Bleu (4-gram)
M+NF0+BACK2	0.1833
M+NF0+BACK1	0.1919
M+F0+BACK2	0.2372
M+F0+BACK1	0.2663
GIZA++ +NF0+BACK2	0.2730
GIZA++ +NF0+BACK1	0.2864
GIZA++ +F0+BACK2	0.2763
GIZA++ +F0+BACK1	0.2882
EDM+NF0+BACK1	0.1978
EDM+F0+BACK1	0.2683
GIZA++ +F0+BACK1+NUM	**0.3177**

0.0086

0.0291

10.3.3 Experiments

❖ Experiment-3

- Comparison of the number of translation options and decoding time using 900,000 sentence pairs

Methods	Bleu (4-gram)	Decoding time
G+F0+BACK1	0.3418	2H6Min
G+F0+BACK1_Top100	0.3452	40Min
G+F0+BACK1_Top150	0.3446	54Min
G+F0+BACK1_Top200	0.3423	64Min
G+F0+BACK1_Top50	0.3366	23Min

10.3.3 Experiments

- IWSLT2005 evaluation
- Training Corpus
  - 1,000K sentence pairs in the specific domain of C-Star, including BTEC corpus and CJK corpus and CASIA corpus
  - 500K sentence pairs in the general domain (news) from Chinese LDC

10.3.3 Experiments

❖**Results from IWSTL'2005**

Track (C-E)	Data condition	Bleu4	NIST	Meteor	WER	PER
Manual transcription	unrestricted	0.5279	10.2499	0.7214	0.4160	0.3366
ASR Output	unrestricted	0.3845	8.0406	0.5802	0.5788	0.4770

Top 20 ASR results

4、口语翻译

10.4.1 概述

- 话语是人类社会信息交流最直接、最方便、最普通的使用方式

— *人类渴望打破上帝制造的语言障碍实现畅通无阻的自由通讯*

- 人类在商贸、文化、体育、旅游等领域交流的日益广泛，渴望话语翻译技术的帮助

— 消除21世纪人类社会全球化最大的障碍需要语音翻译技术

- 语音识别、语音合成和机器翻译技术的快速发展为计算机实现人类话语的自动翻译提供了可能

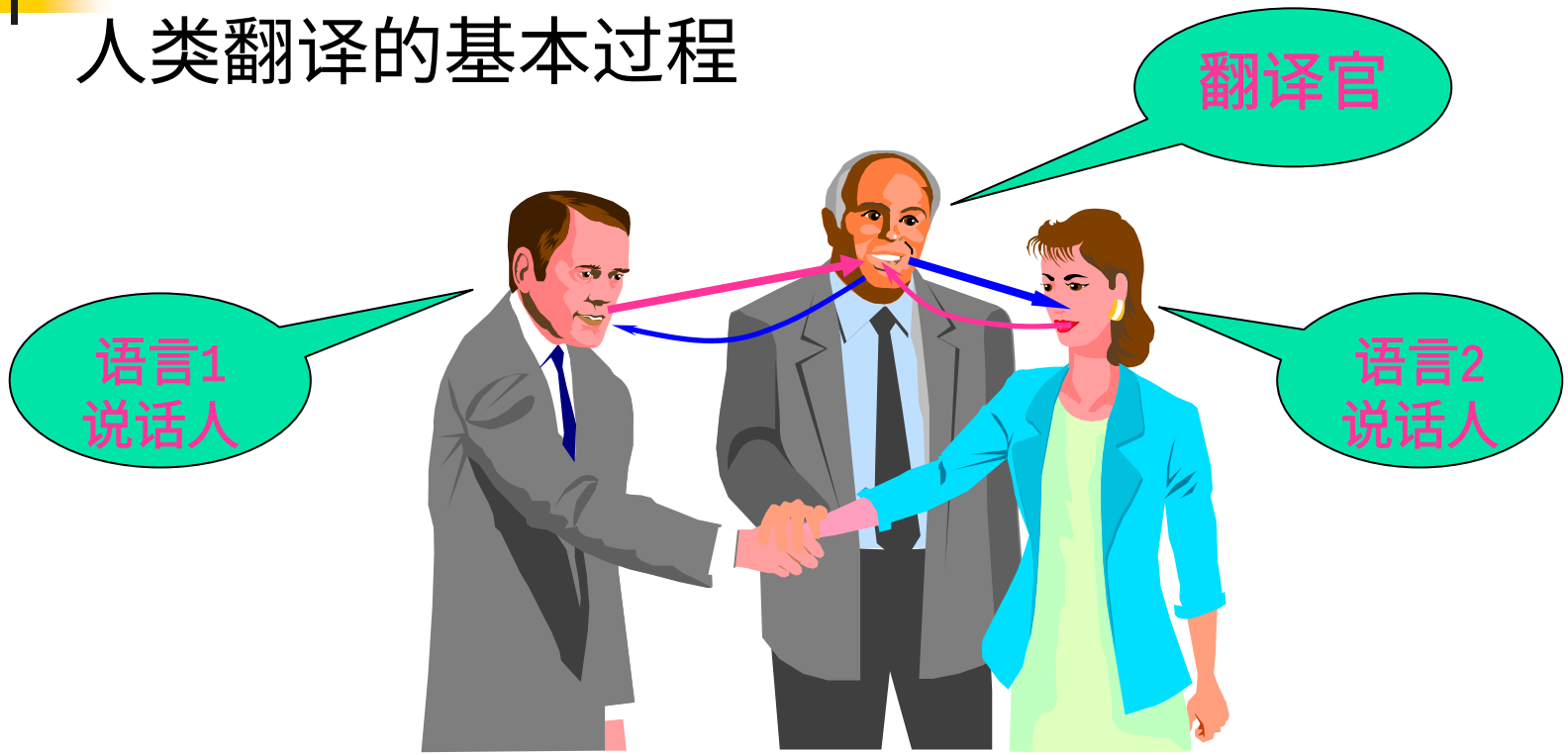
— *信息技术日新月异的发展促使了语音翻译技术的诞生*

10.4.1 概述

- 语音翻译（Speech-to-speech translation）就是用计算机系统实现持不同语言的说话人之间的话语翻译的过程。
- 由于语音识别技术和口语理解技术的提高，语音翻译研究越来越倾向于探索真实对话场景下自然口语风格的话语翻译技术，因此，语音翻译一般又称口语翻译（Spoken language translation, SLT），或对话翻译（Spoken dialogue translation）。
- 里程碑：1989年美国CMU，Speech Trans，English-Japanese 医生与病人对对话。

10.4.2 口语翻译基本原理

- 人类翻译的基本过程



源语言(Source language)1

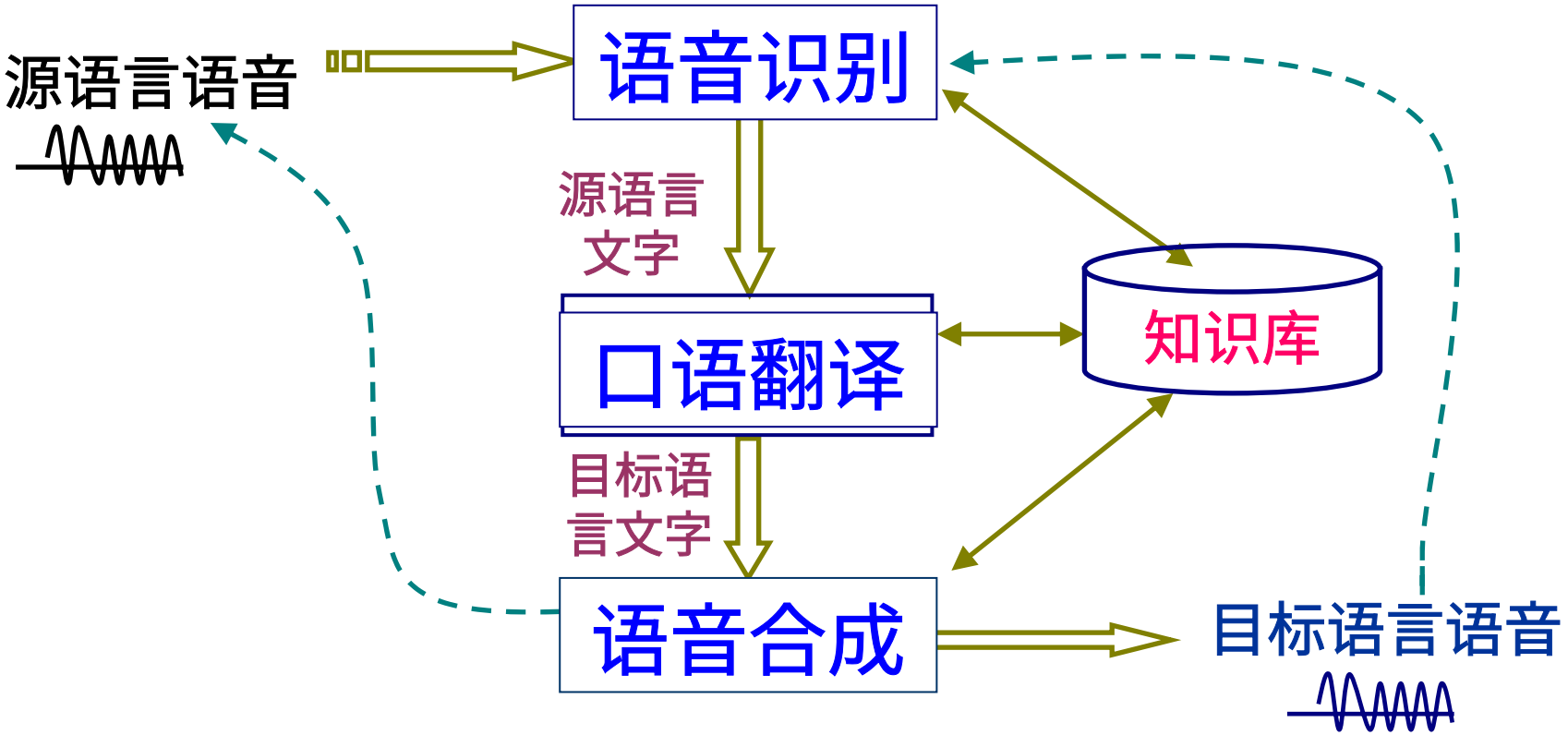
源语言(Source language)2

翻译器
(Translator)

目标语言(Target language)2

目标语言(Target language)

10.4.2 口语翻译基本原理



10.4.2 口语翻译基本原理

- 系统组成
  - 语音识别
  - 口语翻译
  - 语音合成
  - 知识库（语音库、语料库、规则库等）
- 问题
  - 各模块本身的问题
  - 系统集成方法

10.4.3 口语翻译特点

- 系统面对的是复杂多变的口语输入
  - 重复
  - 冗余
  - 省略
  - 修正
  - 词序颠倒
  - 长时间停顿等
  - 无标点

  例子：*那个 我想问一下 那个 你们这里 那里 有没有那个 房间 就是单人间 噢 不 双人间 便宜点儿的 星期三住*

- 系统工作环境复杂
  - 周围环境噪音
  - 语音传输过程中产生的非语声信号

- 系统翻译机制处理的是含有错误信息的字符串输入
  - 含有错误的字或词

10.4.3 口语翻译特点

- 系统尚无法有效地获得和利用对话过程中的非语言信息
  - 语气
  - 手势或动作
  - 表情
- 人们希望系统输出自然流畅的口语语音
  - 语音合成器需要模拟对话双方话语的韵律特征

10.4.4 语音翻译代表系统简介

- 国外部分代表系统

系统名称	开发单位	时间	领域	语种	方法	词汇量
SpeechTrans	CMU	1989	医生与病人对话	日英	RB	--
JANUS-III	CMU, Karlsruhe	1997	旅馆预定	德英日西	ME	开放
ATR-MATRIX	ATR	1998	旅馆预定	日英韩德等	EB	2000
Head-Trans	AT&T	1996	航空旅游	英汉西班牙	SB	1300
Verbmobil	BMBF	90’s	会晤日程	德英等	ME	2500-

10.4.4 语音翻译代表系统简介

- ATR 的口语翻译研究

1986－1992 (Transfer-driven MT, TDMT), J-E

ATR-ITL: Interpreting Telephony Research Labs

1993－1999 (Example-Based, MATRIX), J-E, J-K, J-G

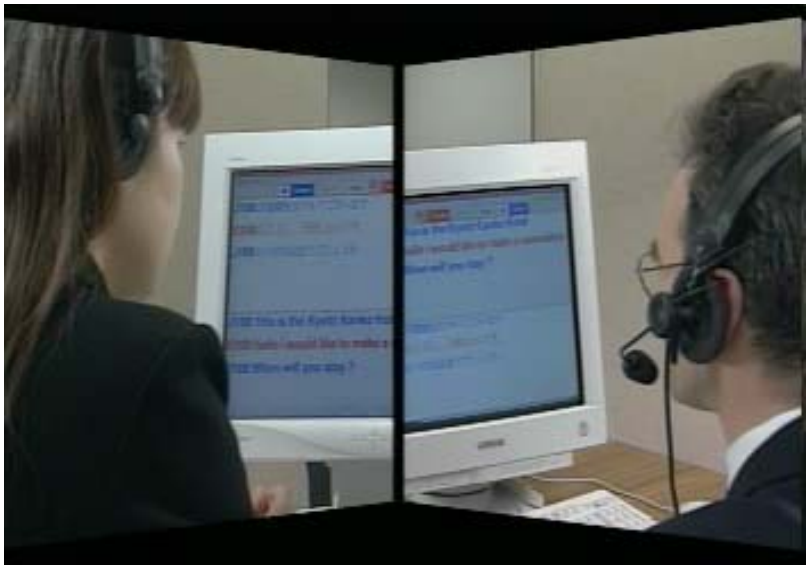
ATR-ITL: Interpreting Telecommunications Research Labs

2000－2005 (MATRIX, Statistical MT), J-E, J-C, J-K…

ATR-SLT: Spoken Language Translation Research Labs

投入300多亿日圆，几百人参与。

10.4.4 语音翻译代表系统简介



基于PC机的ATR-MATRIX
多语言口语翻译系统

基于移动终端的ATR
口语翻译系统

10.4.4 语音翻译代表系统简介

- Verbmobil 系统
  - 受德国联邦教育研究部（German Ministry for Education and Research，BMBF）资助
    - －第一阶段（1993－1996）：$33M
    - －第二阶段（1997－2000）：$28M
    - 工业界：$17M
    - 其它： $11M
    - 共计： $89M
  - 23个参加单位，900多位全职研究人员和学生。

10.4.4 语音翻译代表系统简介

- Verbmobil 系统构成:
  - 10,175 德语单词，6871 英语词汇
  - 统计翻译引擎：58,332 德英语句对训练
  - 基于格的翻译（Case-based）引擎
  - 基于转换的翻译（Transfer-based）引擎：22,783格转换规则；13,640个微观规划规则
  - 基于对话意图(Dialogue-Act)的翻译器
  - 基于子串的(Substring-based)翻译器

10.4.4 语音翻译代表系统简介

- **The** system in ongoing field use at present is the **Phraselator**, a handheld system developed for the US military, and reportedly undergoing field testing in the Middle East (www.phraselator.com).

10.4.4 语音翻译代表系统简介

- 中科院自动化研究所的口语翻译研究

基于台式PC机的汉－日、汉英口语翻译系统

10.4.4 语音翻译代表系统简介

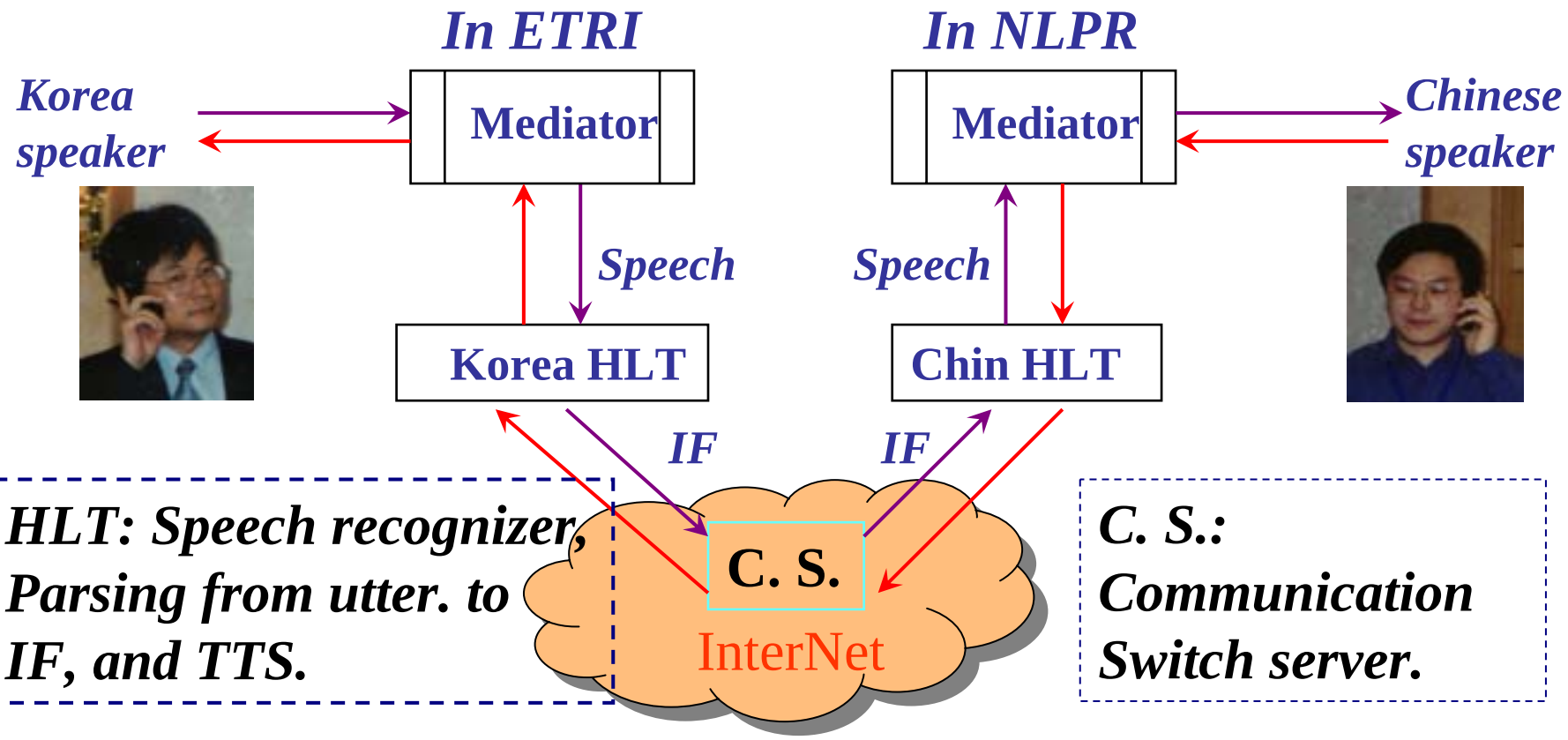
中科院自动化所、日本松下技研(株)及松下电器产业(株)2001年2月分别在北京和东京召开成果发布会，宣布共同开发成功了基于移动终端的汉、日、英双向旅游对话语音翻译系统原型。

10.4.4 语音翻译代表系统简介

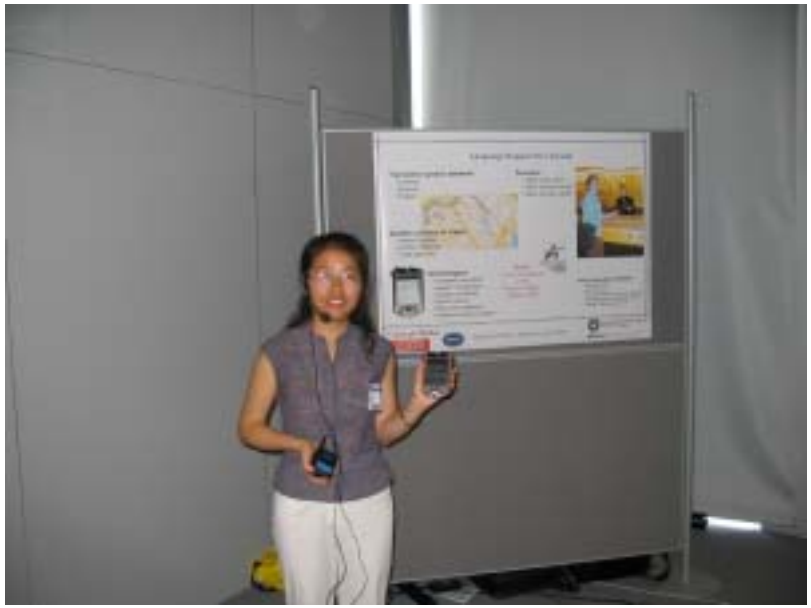
2002年3月自动化所与韩国电子通信研究院联合开发成功基于普通手机的双向中韩口语翻译系统。

10.4.4 语音翻译代表系统简介

❖基于普通手机的中韩双向语音翻译系统



10.4.4 语音翻译代表系统简介



2004年5月22～26自动化所与美国CMU、德国UKA联合开发的中英双向语音实时互译系统在北京国际科技博览会上成功展示。该系统于2004年7月16日－18日在西班牙巴塞罗那国际文化论坛上展示。

10.4.5 国际语音翻译先进研究联盟

- 国际语音翻译先进研究联盟(Consortium for Speech Translation Advanced Research International, C-STAR)

  - 成立于 1991年，由日本 ATR, 美国CMU, 德国 Univ. of Karlsruhe (UKA) 和 Siemens 等单位发起；

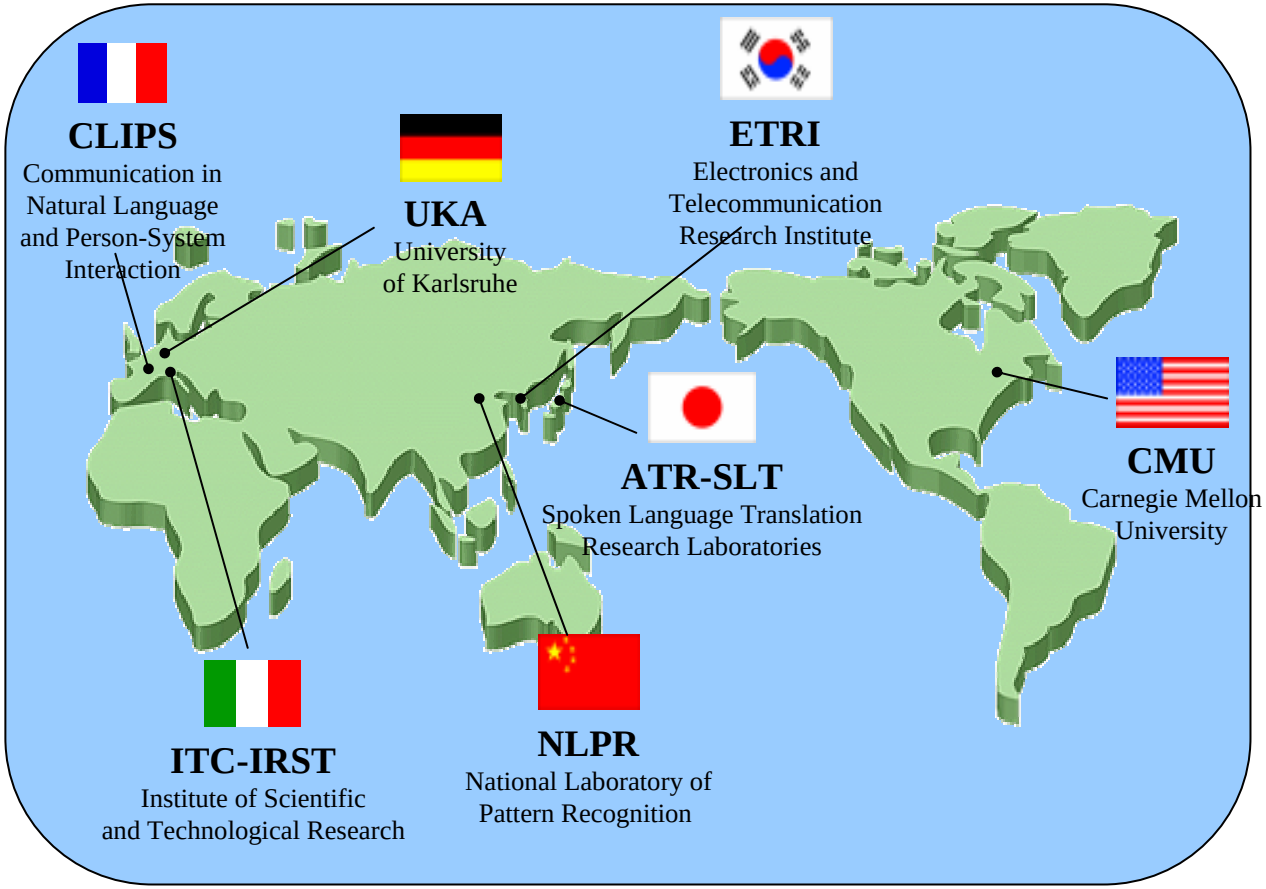
  - 目前拥有12个国家的20多个联系成员 (affiliate members), 7个核心成员(partners)；

  - 1996年中科院自动化研究所成为C-STAR联系成员，2000年10月正式成为C-STAR核心成员。

10.4.5 国际语音翻译先进研究联盟

2000年10月10日C-STAR在日本ATR召开会议，正式接纳自动化所成为第七个核心成员，将汉语正式纳入了国际语音互译研究的大家庭！

10.4.5 国际语音翻译先进研究联盟



Seven Languages:

English, French, Chinese, Germany, Japanese, Italian, and Korean

10.4.5 国际语音翻译先进研究联盟

- **C-STAR 发展的三个阶段**
  - First stage: 1991 ~ 1993: C-STAR I
  - Second stage: 1993 ~ 2000: C-STAR II (6 full partners in 6 countries: ATR of Japan, ETRI of Korea, CMU of USA, UKA of Germany, ITC-irst of Italy, and GETA, CLIPS-IMAG of France)
  - Third stage: 2000 ～ (7 full partners in 7 countries)

10.4.5 国际语音翻译先进研究联盟

- 任务和目标

开展自动语音翻译实用技术的研究，最终实现任意时间(Any time)、任意地点(Any where)、任何语言(Any language)的自由通讯辅助翻译服务。

10.4.6 语音翻译研究现状

- 基本现状
  - 限定领域（旅游信息咨询、旅店服务等）
  - 限定词汇量（1000－60000）
  - 自然口语（连续语音、非特定说话人，句型无严格限制）
  - 多语言双向翻译
  - 多方法、多策略结合

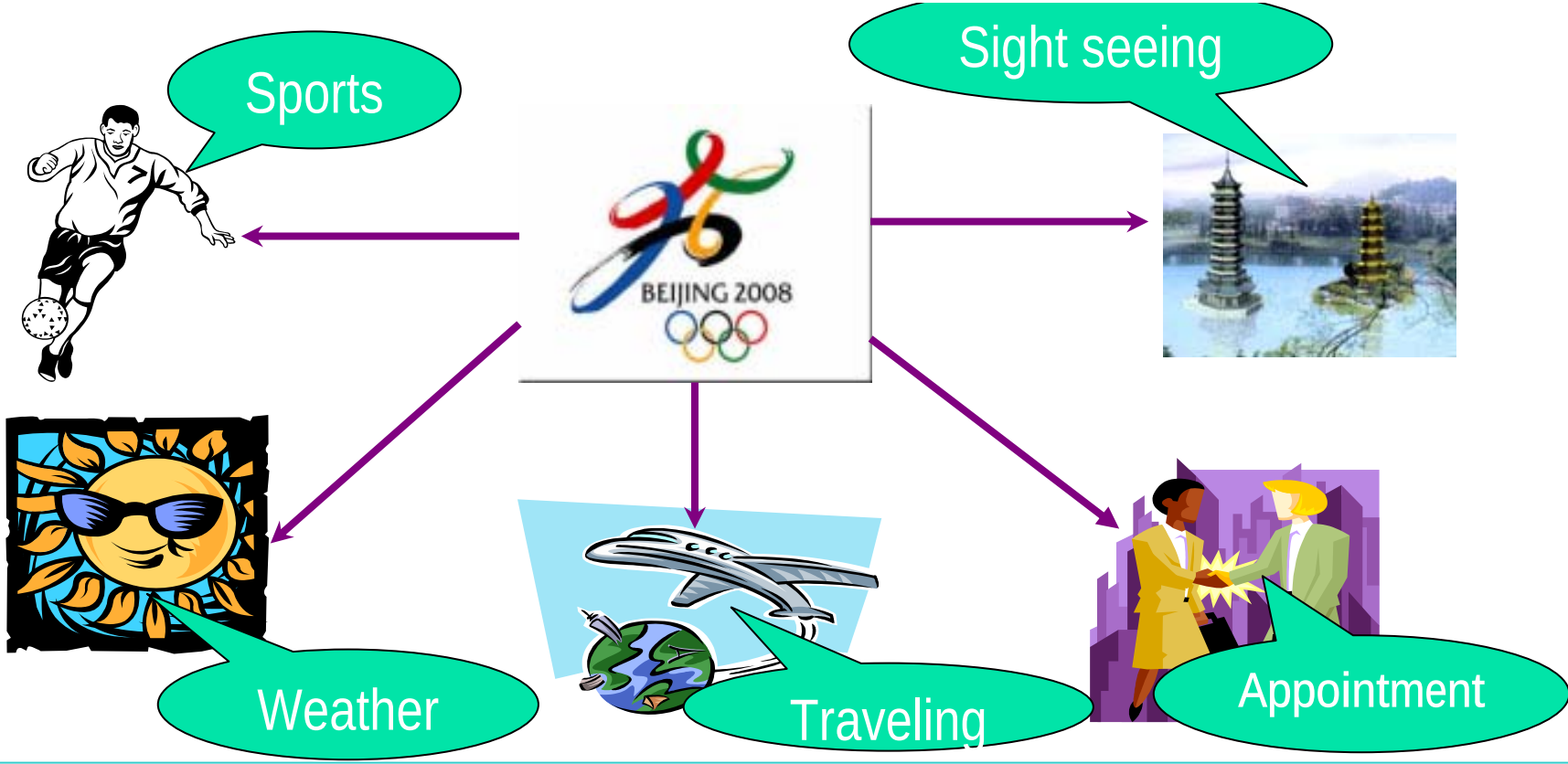
10.4.6 语音翻译研究现状

主要问题

- 口语的声学特性分析不够，声学语音层HMM模型不够精细，语音识别的鲁棒性(robustness)差；
- 口语的语言学特性有待于深入研究，非规范语言现象的处理能力差，歧义消解能力差，解析机制的鲁棒性和正确率都有待于进一步提高；
- 翻译方法不够奏效，多翻译策略的结合方法有待于深入探讨；
- 对话情景知识的表示和利用,多媒体、多模态集成翻译技术有待于深入研究；

10.4.6 语音翻译研究现状

- 面向限定领域、特定任务的语音翻译技术走向实际应用并不遥远



上个世纪的话

In the next millennium, we will find that we are talking as much or more with machines than we are with human.

在下一个千年里，我们会发现我们和机器说的话，与我们和人类说的话一样多，甚至比跟人类说的话还要多。

- N. Negroponte

from "Being Digital"

留给自己的话

为了实现任意时间、任意地点、任意语言之间的全球无障碍自由通讯的梦想我们终日绞尽脑汁，可是，如果在追求那些遥远的辉煌目标的同时，人与人之间的距离变得越来越疏远，心与心之间的交流变得越来越困难，那么，我们所苦苦追求的一切也便失去了意义。所以，无论我们在做什么，都不要丧失人类最本质、最人性的基点，用我们的无私与真诚去感动世界。计算机可以为我们做一切，但唯一不能做到的就是付出友爱与真情。

Thanks

谢谢!