自然语言理解

(11)

宗成庆

中科院自动化研究所

模式识别国家重点实验室

cqzong@nlpr.ia.ac.cn

http://www.nlpr.ia.ac.cn/English/cip/cqzong.htm

No.95, Zhongguancun East Road
Beijing 100080, China

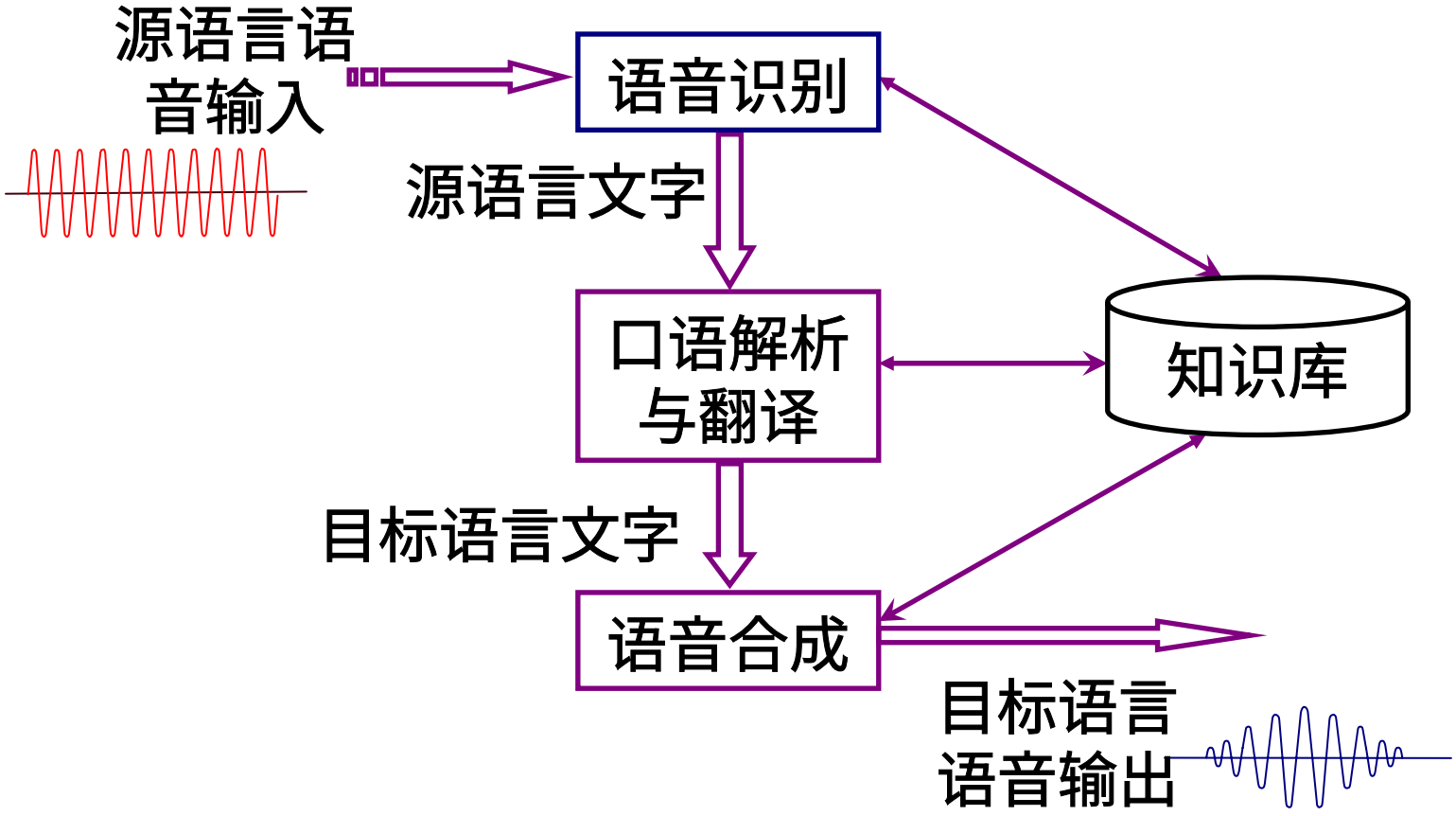
http://www.ia.ac.cn
Tel. No.: +86-10-6255 4263

第十一章 口语翻译

11.1 概述

- 里程碑：1989年 CMU，Speech Trans, English-Japanese
  医生与病人对话，基于规则方法
- 概念：语音翻译 Speech-to-speech translation, S2S
  口语翻译 Spoken language translation, SLT
  对话翻译 Spoken dialogue translation
  State-of-the-art

11.2 口语翻译基本原理



11.3 口语翻译技术的特点

❒口语翻译的难点

- 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

例子：*我想问一下 那个 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

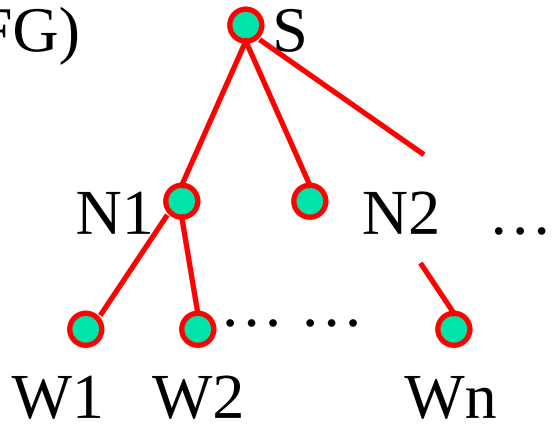
11.3 口语翻译技术的特点

- 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
- 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
  - 无标点
- 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

11.3 口语翻译技术的特点

- 口语自动翻译技术类型

1. 基于规则的翻译方法(Rule-Based)
   - 方法：输入语句→ 词法分析→ 句法分析→ 语言生成
   - 典型文法：上下文无关文法 (CFG)
   - 代表系统：SpeechTrans (CMU)



11.3 口语翻译技术的特点

2. 基于事例的翻译方法（Example-Based）

- 方法：输入语句→ 与事例相似度比较→ 翻译结果
- 代表系统：ATR-MATRIX（ATR, Japan)
- 事例库：源语言语句或成份S1 ==> 目标语言表达T1

  … …

  源语言语句或成份Sn ==> 目标语言表达Tn

输入语句S’ :: { S1, S2, … … Sn } ==> 翻译结果T’

11.3 口语翻译技术的特点

3. 基于中间语义表示的翻译方法（Interlingua-based）

- 方法：输入语句→ 分析转换→ 中间语言→ 翻译结果
- 代表系统：JANUS-III (CMU) 早期版本

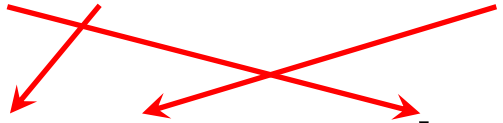
✯ 鲁棒的(Robust)口语解析器(Parser)
✯ 比较准确的中间语义描述语言(Interlingua)
✯ 目标语言生成器(Target Language Generator)

11.3 口语翻译技术的特点

4. 基于统计的翻译方法（Statistical Method）

- 方法：看作经典的噪声信道的信号恢复问题

Input: W = w1 w2 w3 … … wn



Output: T = t1 t2 … … tk

$$P(T|W) = \frac{P(W|T) \cdot P(T)}{P(W)}$$

$$P(W|T) = \prod_i P(w_i|t_i)$$

$$P(T) \approx P(t_1) \cdot P(t_2) \cdot \ldots P(t_N) \approx P(t_1) \cdot P(t_2|t_1) \cdot \ldots \cdot P(t_N|t_{N-1})$$

- 代表系统：Head-Transducer (AT&T)

11.3 口语翻译技术的特点

5. 多引擎、多策略的翻译方法（Multi-Engine）

- 方法：上述多种方法同时进行，选取最优翻译结果(并行)
  在不同的情况下执行不同的翻译引擎(串行)
- 代表系统：Verbmobil (Germany), JANUS-III

11.4 口语翻译代表系统

- 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-

11.4 口语翻译代表系统

- 自动化研究所的口语翻译技术研究

Lodestar: Chinese-to-Japanese Spoken language translation,1998

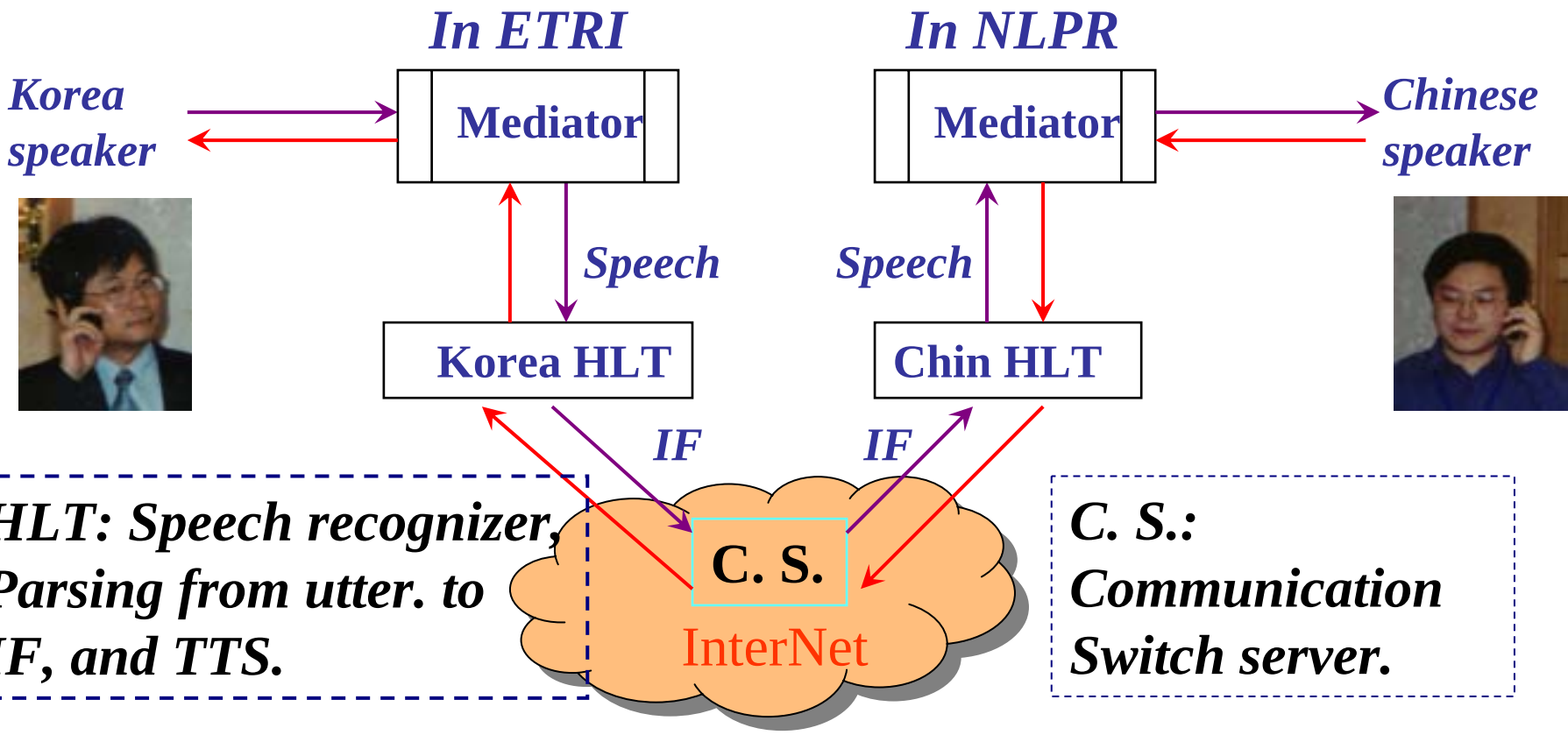
Lodestar: Chinese-to-English Spoken language translation

11.4 口语翻译代表系统

Chinese, Japanese, English S2S translation, 2002

11.4 口语翻译代表系统

❖基于普通手机的中韩双向语音翻译系统



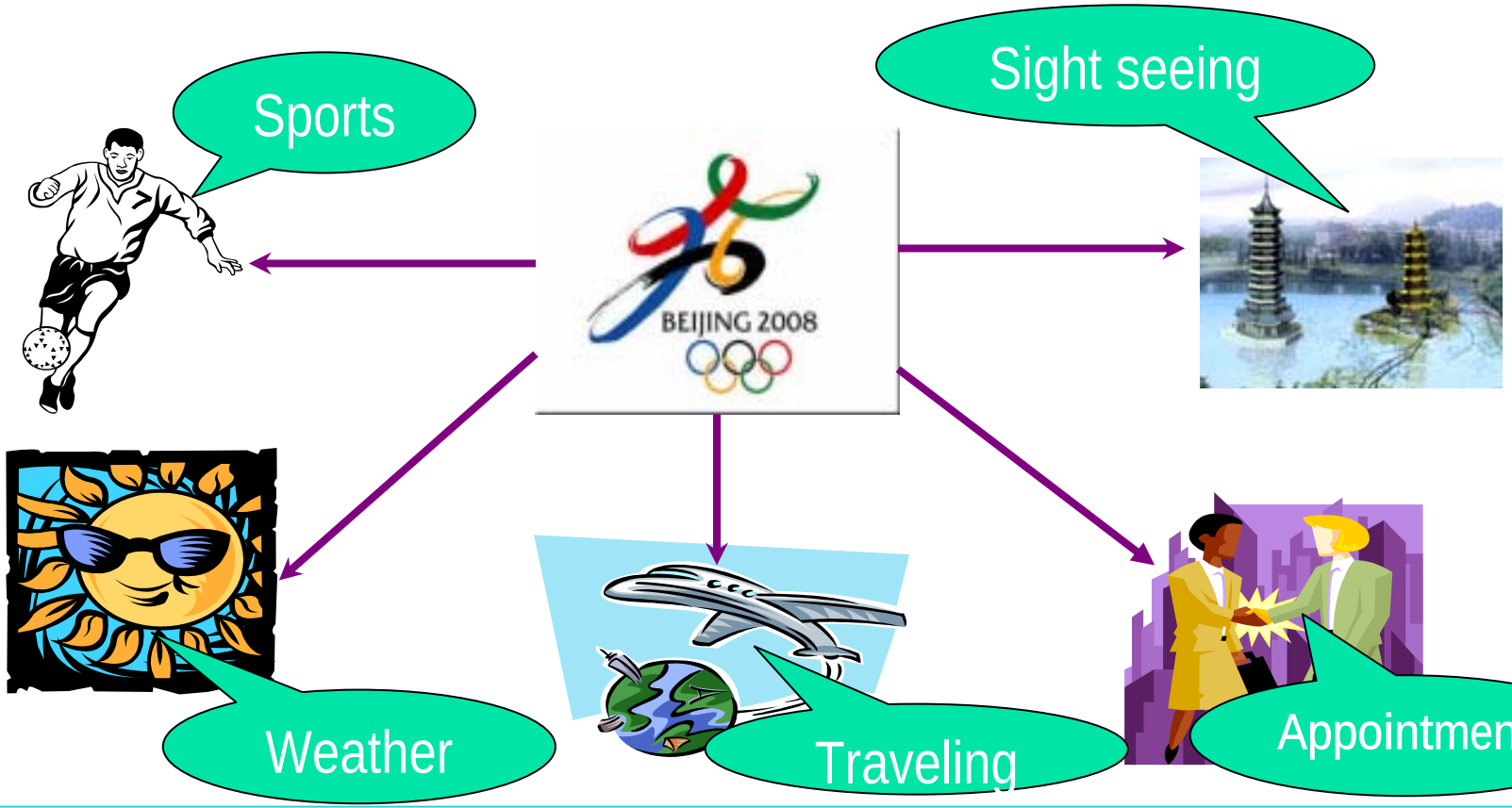
HLT: Speech recognizer, Parsing from utter. to IF, and TTS.

C. S.: Communication Switch server.

11.4 口语翻译代表系统

11.4 口语翻译代表系统

- 面向2008北京奥运会的多语言智能信息服务系统

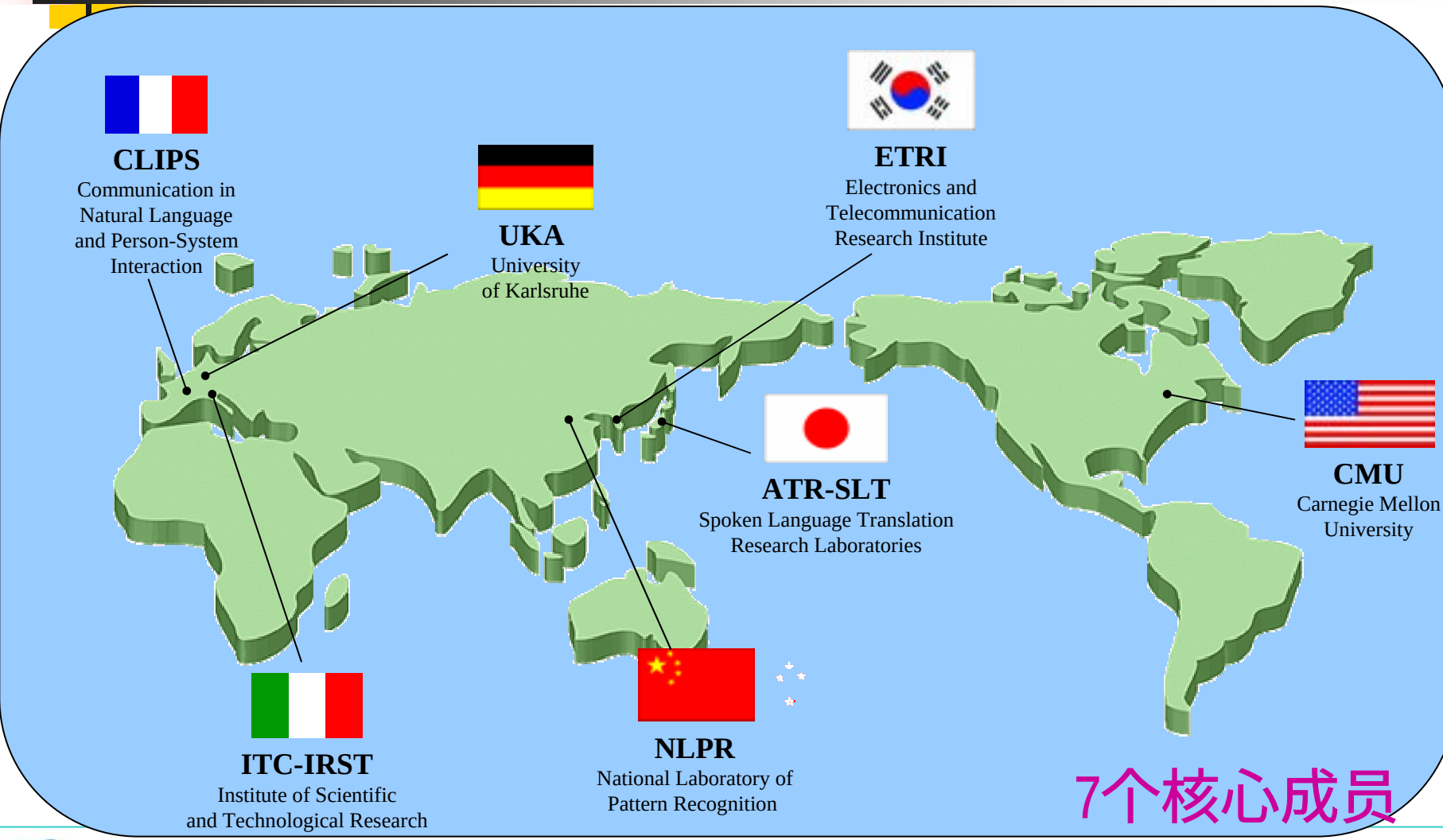


11.5 关于国际语音翻译联盟

❒C-STAR: Consortium for Speech Translation Advanced Research

- ♦ Set up in 1991
- ♦ 12 countries, over 20 affiliate members and 7 partners
- ♦ NLPR was an affiliate members since 1996
- ♦ 三个阶段： 1991－1993
  1993－2000.10
  2000.10－

11.5 关于国际语音翻译联盟



7个核心成员

11.5 关于国际语音翻译联盟

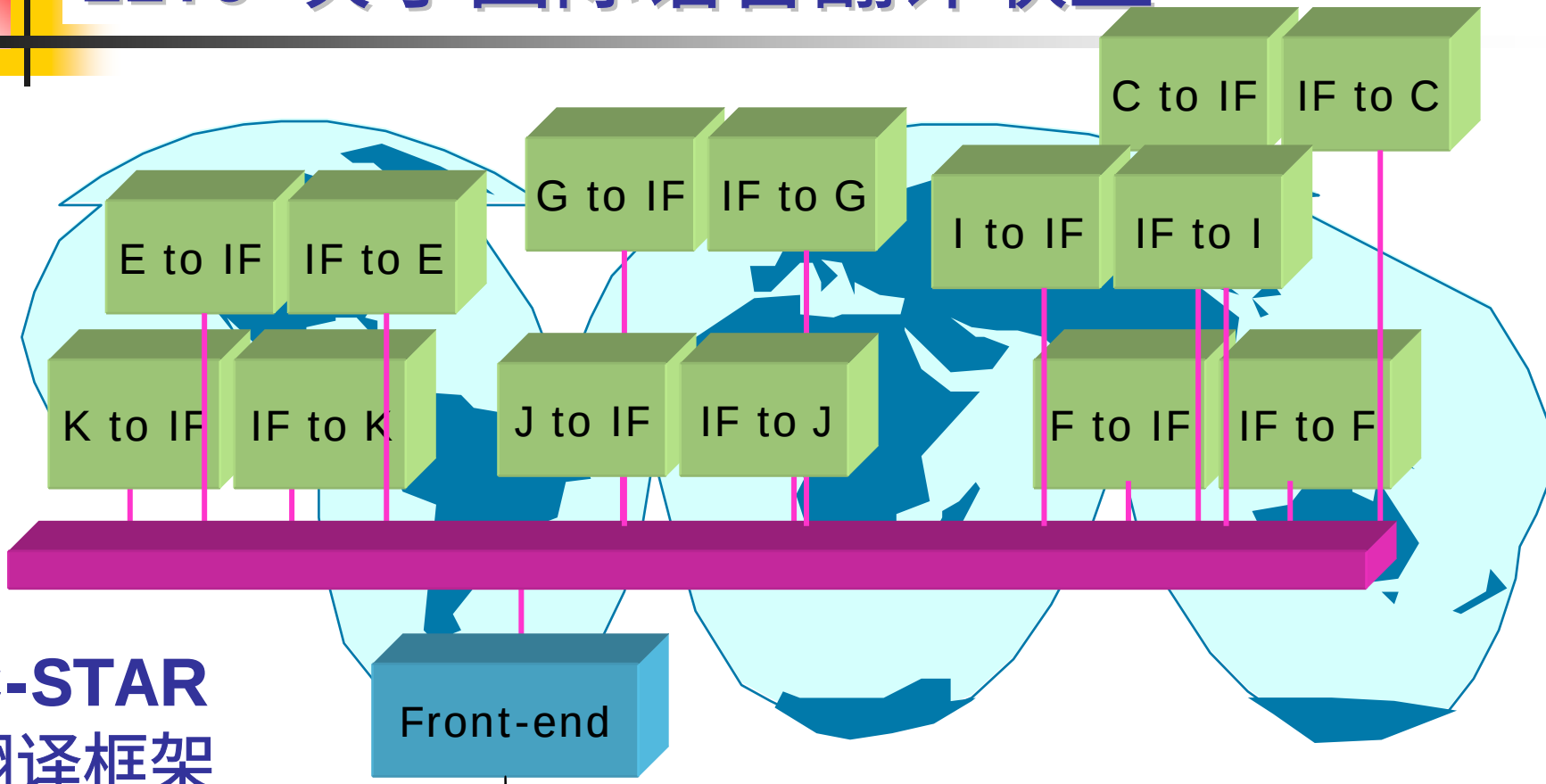
2000年10月自动化所正式成为C-STAR 核心成员。

11.5 关于国际语音翻译联盟

C-STAR III Goal

- Technology for real application
- Translating aid for traveler
- Service available anywhere, anytime

11.5 关于国际语音翻译联盟



C-STAR
翻译框架

IF: Interchange Format
or Translated Text

11.6 口语翻译技术研究现状

- 基本现状
  - 限定领域
  - 自然口语
  - 多方法、多策略结合
  - 国际合作
  - 限定词汇量
  - 多语言双向

11.6 口语翻译技术研究现状

- 有待进一步研究的课题
  - 进一步加强翻译方法的理论研究
  - 加强口语的声学特性分析，使声学语音层HMM模型精细化，提高语音识别的鲁棒性(Robustness)
  - 研究口语的语言学特征，完善语言模型，构造鲁棒的口语解析器(Parser)
  - 加强受限领域的语料库(Corpus Base)技术研究
  - 研究对话情景知识的表示和利用方法以及多媒体、多模态的集成翻译技术

11.7 关于分析方法与统计方法的思考

1、日本 ATR对两种口语翻译方法的比较*

依据的语料：

	日语	英语
句子数	204,108	204,108
词汇个数	1,689,449	1,235,747
词汇量	19,640	15,374
平均长度（单词/句子）	8.3	6.1

* Eiichiro Sumita, Corpus-Centered Computation, in *Proc. ACL Workshop: Speech-to-speech Translation*. Philadelphia, USA. July 11, 2002. pp. 1-8.

11.7 关于分析方法与统计方法的思考

实验系统测试结果：

	A	A+B	A+B+C
日英统计翻译系统[1]	25%	46%	64%
英日统计翻译系统	41%	48%	57%
基于事例的日英翻译系统－1（D^3）[2]	47%	66%	77%
基于实例的日英翻译系统－2（HPAT）[3]	50%	61%	71%

[1] 基于词的统计翻译模型

[2] Dp-match Driven transDucer [Sumita, 2001, Example-based Machine Translation Using DP-matching between Word Sequences. *Proc. of DDMT(ACL)*, pp. 1-8.]

[3] Hierarchical Phrase Alignment (HPA) based Translation (HPAT) [Imamura, 2002, Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment, *Proc. of TMI*.]

11.7 关于分析方法与统计方法的思考

2、Verbmobil 口语翻译系统对两种方法的比较*

（1）Verbmobil 系统概况:

- 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

** Wolfgang Wahlster, Verbmobil Multilingual Processing of Spontaneous Speech, www.dfki.de/~wahlster/VM-final*

11.7 关于分析方法与统计方法的思考

(2) Verbmobil 口语翻译系统构成:

- 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，8.9 德词/Sent.，9.4英词/Sent.
- 基于格的翻译（Case-based）引擎：30,000模板(Template)
- 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：334个多语言FST (Finite State Transducers)
- 基于子串的（Substring-based）翻译器：24,680德英语句对训练

11.7 关于分析方法与统计方法的思考

(3) 系统评测方法：

- 基于网络的大规模测试：43,180 Translations
  (25,345(德)+17,835(英))
- 65评估人员
- 句子长度：1－60单词

11.7 关于分析方法与统计方法的思考

实验系统的测试结果：

翻译引擎与最终系统	翻译词正确率 ≥50%	翻译词正确率 ≥75%	翻译词正确率 ≥80%
Case-based Translation	37%	44%	46%
Statistical Translation	69%	79%	81%
Dialogue-act based Trans.	40%	45%	46%
Semantic Transfer (SeT)	40%	47%	49%
Substring-based Translation	65%	75%	79%
Automatic Selection	57% / 78%*	66% / 83%*	68% / 85%*

** After training with instance-based learning algorithm*

11.7 关于分析方法与统计方法的思考

3、问题与思考

- ATR的D^3与Verbmobil系统中的 SeT 比较孰优孰劣？意义？
- 两个系统翻译的语言对差异较大：日 ↔ 英，德 ↔ 英
- 两个统计翻译引擎都是基于词对位，基于短语、块或其它将如何？
- 训练语料的规模对统计翻译系统的影响？（D^3: 20.4 万; SeT: 5.8万）
- 统计翻译方法和基于分析的翻译方法都是孤军奋战？

11.7 关于分析方法与统计方法的思考

4、统计翻译方法的“石头汤 (Stone soup)”之说

(http://spanky.triumf.ca/www/fractint/stone_soup.html)

- in Eastern Europe, there was a great famine
- a peddler drove his wagon into a village
- iron cauldron, water, stone, fire
- most of the villagers had come to the square or watched
- cabbage, salt beef, potatoes, onions, carrots, mushrooms, … …
- a delicious meal

11.7 关于分析方法与统计方法的思考

5、观点与认识

1) 统计方法的崛起打破了分析方法在整个自然语言处理领域一统天下的僵局，但并不意味着分析方法的结束

2) 在很多方面，分析方法与统计方法并不是完全对立的

- 绝对化与模糊化
- 唯一化与多选化
- 主观性与客观性
- 定性与定量

11.7 关于分析方法与统计方法的思考

3) 统计方法与分析方法的结合是必然的结局

- 知识准备与资源建设
  - 词对 / 规则 / 模板 / 实例 的自动抽取与学习
- 处理方法实现过程
  - 句法分析与语义提取、Chunk / 句子边界检测、远距离相关分析

4) 多翻译引擎协同竞争机制是口语翻译系统首选的策略

- 基于模板（Template）或模式（Pattern）的方法
- 基于规则或中间语言（语义表示）的方法
- 基于统计的方法
- 基于实例的方法

小结

- 口语翻译的特点与难点
- 口语翻译的基本原理
- 口语翻译基本方法与技术现状
- 关于国际语音翻译联盟（C-STAR）
- 关于统计翻译方法与分析翻译方法的思考

习题

1. 认真查阅关于 IBM 统计翻译模型及其相关技术的文献，学习并掌握统计翻译方法。
2. 实现一个小型的基于统计翻译模型的汉英翻译系统，掌握实现技术。
3. 查阅有关口语理解方法论文，了解口语解析技术的基本思想。
4. 思考如何将基于规则的翻译方法和基于统计的翻译方法相结合，实现高质量的口语翻译系统？

Thanks

谢谢!