

## China's Machine Translation Technology

--keynote speech at The 1st Applied Linguistics  
Congress of China (May 15-21 , 2007, Beijing)

**FENG ZHIWEI**  
**Institute of Applied Linguistics**  
**Ministry of Education**  
zwfengde@bbn.cn

## Babel Tower

- Language barrier & communication



## First Machine Translation (1954)

Hurd, Dostert and Watson at the interface



IBM'S WATSON (right) AND FRIENDS: For a mathematical wizard . . .

## IBM-701 for first MT

IBM 701 at New York headquarters

- "filling a room as big as a tennis court" (New York Herald Tribune)



## Punched card input

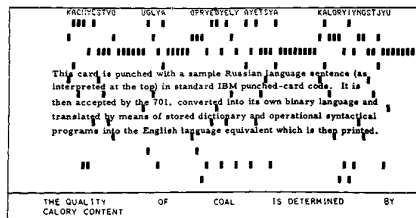
### Punched card input



Sentences in Russian are punched into standard cards for feeding into the electronic data processing machine for translation into English.

## Sentence input on punch card

### Sentence input on punch card



Specimen punched card and below a strip with translation, printed within a few seconds

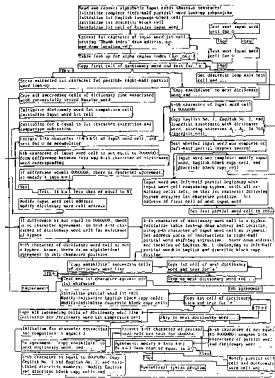
## Card reading unit

### Data input



## Flowchart for syntax analysis

### Sheridan's flowchart for syntax analysis (part of one rule)



## Dictionary for MT

### Dictionary output for example sentence

| • Russian input  | English equivalents |                  | 1st code<br>(PID) | 2nd code<br>(CDD <sub>1</sub> ) | 3rd code<br>(CDD <sub>2</sub> ) |
|------------------|---------------------|------------------|-------------------|---------------------------------|---------------------------------|
|                  | Eng <sub>1</sub>    | Eng <sub>2</sub> |                   |                                 |                                 |
| • vyelyichyina   | magnitude           | ---              | ***               | ***                             | **                              |
| • ugl-           | coal                | angle            | 121               | ***                             | 25                              |
| • -a             | of                  | ---              | 131               | 222                             | 25                              |
| • opryedyayetsya | is determined       | ---              | ***               | ***                             | **                              |
| • otmoshyeni-    | relation            | the relation     | 151               | ***                             | **                              |
| • -yem           | by                  | ---              | 131               | ***                             | **                              |
| • dlyin-         | length              | ---              | ***               | ***                             | **                              |
| • -i             | of                  | ---              | 131               | ***                             | 25                              |
| • dug-           | arc                 | ---              | ***               | ***                             | **                              |
| • -i             | of                  | ---              | 131               | ***                             | 25                              |
| • k              | to                  | for              | 121               | ***                             | 23                              |
| • radyius-       | radius              | ---              | ***               | 221                             | **                              |
| • -u             | to                  | ---              | 131               | ***                             | **                              |

## Output of English text

### Output on line printer

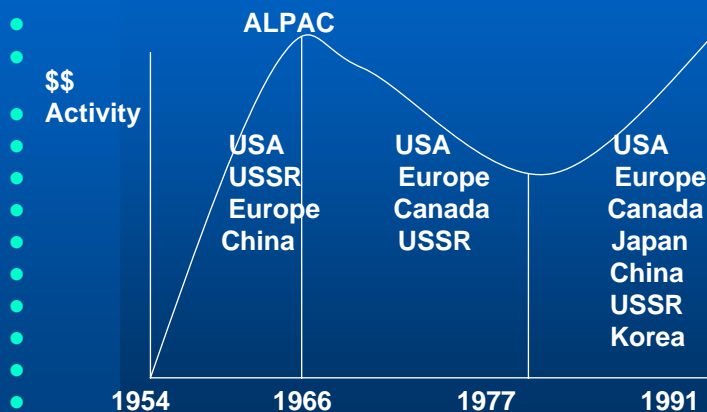


A continuous sheet of English-words sentences comes up on the printing unit seconds after sentences, in Russian, have been fed into the computer

## New York Times, 1954-01-08

- 《701 translator》 published in New York Times, 1954-01-08
- In the demonstration, a girl operator typed out on a keyboard the following Russian text in English characters: "Mi pyeryedayem mislyi posryedstvom ryechi" (Мы передаем мысли посредством речи). The machine printed a translation almost simultaneously: "We transmit thoughts by means of speech." The operator did not know Russian. Again she types out the meaningless (to her) Russian words: "Vyelyichyina ugla opryedyelayatsya otnoshenyiyem dlyini dugi k radiusu." (величина угла определяется отношением длины дуги к радиусу) And the machine translated it as: "Magnitude of angle is determined by the relation of length of arc to radius." (New York Times, January 8, 1954)

## Zigzag of MT development



## MT in China

- Past
- Present
- Future

## Past

- the early experimental period (1956–1966)
- the stagnant period (1966–1975)
- the recovery period (1975–1987)
- the blossom period (since 1987)

## The early experimental period

- The National Plan for Developing the Science and Technology – Project “machine translation” (1956)
  - establishing of the translation rules of natural language
  - mathematical theory for natural languages
- Russian-Chinese MT experiment (RC-59)
  - vocabulary of 2,030 Russian words
  - algorithm of 29 flowcharts

## The recovery period

- TITLE-1 system (1976–1986, ISTIC)
- ECMT-1 system (1978, Liu Zhuo)
- JFY system (1976–1984, Liu Zhou)
- FAJRA system (1981, Feng Zhiwei, CETA)
- INSPEC system (1985)
- HT-83 system (1983, Uni-Heilongjiang)
- RI-84 system (1984, Uni-Heilongjiang)
- GCAT system (1985, Feng Zhiwei)
- FCAT system (1985, Feng Zhiwei)

## FAJRA system

- 汉语输入:

4 November 1981

语言和通讯在人类的生活中起着重要的作用。学者们正在积极地研究语言的自动翻译。他们提出了自动机的理论和形式语言的理论。自动机是能够识别语言的一种抽象机器。语言的规则能够把一个树形图自动地转换成另一个树形图。这个自动翻译系统已经被他们使用了五年。如果我们把这个系统同其他一些系统相比较,就可以看出它的优点。

## English output

### English output

英语输出

4 Novembre 1981

Language and communication play important role in human life. Scholars are actively studying automatic translation of language. They have proposed the theory of automaton and the theory of formal grammar. An automaton is an abstract machine that can recognize language. Grammar rule can transform a tree graph to another tree graph automatically. This system for automatic translation had been used by them for five years. If we compare this system with other systems, then we can see its advantage.

## French output

### French output

法语输出

4 Novembre 1981

Le langage et la communication jouent un rôle important dans la vie humaine. Les savants sont en train d'étudier activement la traduction automatique des langages. Ils ont proposé la théorie des automates et la théorie des grammaires formelles. un automate est une machine abstraite qui peut reconnaître le langage. La règle de la grammaire peut transformer automatiquement une arborescence en une autre arborescence. Ce système de la traduction automatique a été utilisé par eux pendant cinq ans. Si nous comparons ce système à d'autres systèmes, alors nous pouvons voir ses avantages.

## German output

### German output

德语输出

4 Novembre 1981

Die Sprache und die Kommunikation spielen im Leben der Menschen wichtige Rolle. Die Wissenschaftler sind gerade dabei, die automatische Übersetzung der Sprachen aktiv zu studieren. Sie haben die Theorie der Automaten und die Theorie der formalen Grammatiken vorgeschlagen. Ein Automat ist eine abstrakte Maschine, die die Sprache erkennen kann. Die Grammatikregel kann einen Baumgraphen in einen anderen Baumgraphen automatisch umwandeln. Dieses System zur automatischen Übersetzung war von ihnen seit fünf Jahren verwendet worden. Wenn wir dieses System mit anderen Systemen vergleichen, dann können wir seinen Vorzug sehen.

## Japanese output

Japanese output

日語輸出 4 Novembre 1981

言語と通信は人間の生活に重要な役割を果している。研究者は言語の自動翻訳を熱心に研究している。彼等はオートマトンの理論と形式文法の理論を提案している。オートマトンは言語を認識することができる一種の抽象機械である。文法の規則はひとつの木のグラフを別の木のグラフに自動的に変換することができる。自動翻訳のためのこのシステムは彼等によつて五年間使われていた。このシステムを他のシステムと比べると、われわれはその利点をわることができる。

## Russian output

Russian output

俄语输出 4 Novembre 1981

Язык и коммуникация играют важную роль в жизни человечества, ученые активно изучают автоматический перевод языков. Они выдвигают теорию автомата и теорию формальной грамматики. Автомат есть одна абстрактная машина, которая может распознавать язык. Правило грамматики может автоматически превращать одну древовидную схему на другую древовидную схему. Эта система автоматического перевода использована ими уже пять лет. Если мы сравниваем эту систему с другими системами, то мы можем видеть её преимущества.

## The blossom period

- KEYI-1 English-Chinese system of Mars Institute (Beijing)
  - translation speed is 3,000 words/hour
  - the result of translation is readable
- TRANS-STAR system :
  - China National Software & Technology Service Co. (CS&S) bought the KEYI-1 copyright
  - KEYI-1 system was renamed as TRANS-STAR system.

## Chinese government funding

- The research work of all these MT systems was supported by Chinese government funds.
- The goal of the machine translation is just for translation of scientific documents in order to exchange the scientific information with developed countries.
- No any private company has interest for MT system in this period.

## Present

- GAOLI MT system (English-Chinese)
- 863-IMT/EC system (English-Chinese)
- SINO-TRANS system (Chinese-English)
- TONGYI system (English-Chinese)
- YIWANG system (English-Chinese)
- YIBA system (English-Chinese)
- E-to-J system (English-Japanese)

## GAOLI MT system (English-Chinese)

- Basic lexical dictionary: 60,000 entries
- Linguistic rules: more than 800 rules
- Background knowledge database: more than 150 entries
- Translation accuracy: 80%
- Readability of translated text: 80%-90%

## 863-IMT/EC system (English-Chinese)

- Basic English lexical base: 35,000 entries
- Basic Chinese lexical base: 25,000 entries
- Linguistic rules: 1500 rules
- Translation accuracy: 80%

## SINO-TRANS system (Chinese-English)

- Basic dictionary: 40,000 entries
- Two special subject technical dictionaries
  - Navel ships and boats (9312 entries)
  - rocket-gun (33,773 entries)
- Linguistic rules: 1,000 rules

## SUNSHINE YIWANG system

- Highest translation speed:  
100 sentences per second
- Can be used for browsing  
the text of INTERNET
- Web readworld:  
[www.readworld.com](http://www.readworld.com)
- Multi-windows display

## YAXIN-YIBA system

- Three translation models
  - on line translation model
  - automatic translation model
  - interface translation model
- Open to users: user can revise the  
dictionary and rules in MT system
- Rich special subject dictionaries:  
30 subjects (e.g. Computer,  
telecommunication, medicine)

## E-to-J system (English-Japanese)

- This system is developed by JEC company in Beijing.
- Technique of transformation from phrase tree (P-tree) to dependency tree (D-tree)
- Closely integrated with word processor

## Rule-based approaches abroad (1)

- Linguistic string analysis (Zellig Harris)
- Phrase structure Grammar (N. Chomsky)
  - Top-down parsing
  - Bottom-up parsing
  - Tomita Algorithm
  - Left-corner parsing
  - Cocke-Younger-Kasami algorithm (CYK algorithm)
- Augmented Transition Network (ATN, W. Woods)

## Rule-based approaches abroad (2)

- General Syntactic Processing (GSP, R. Kaplan)
- Chart Parser (Martin Key)
- Category Grammar (Y Bar-Hillel, J. Lambek)
- Link Grammar (D. Sleator, D. Temperley)
- Dependency Grammar or valency Grammar (L. Tesnière, G. Herbig)
- Government & Bounding Theory (GB, N. Chomsky)

## Rule-based approaches abroad (3)

- Lexical Functional Grammar (LFG, R. Kaplan, J. Bresnan)
- Functional Unification Grammar (FUG, Martin Kay)
- Montague Grammar (MG, R. Montague)
- Generalized Phrase Structure Grammar (GPSG, G. Gazdar, I. Sag)
- Head-driven Phrase Structure Grammar (HPSG, C. Pollard, I. Sag)
- Definite Clause Grammar (DCG, F. Pereira, D. Warren)
- Case Grammar (C. Fillmore)
- Preference Semantics (Y. A. Wilks)
- Conceptual Dependency Theory (R. Schank)

## Statistic-based Approaches

- N-gram Grammar
- Hidden Markov Model (HMM)
- Noisy Channel Model
- Parameter Estimation
  - Maximum Likelihood Estimation (MLE)
  - Sparse data problem
  - Parameter Smoothing approaches
    - Interpolated estimation
    - Adjusting frequency
- Preference-based Approaches
  - Collocation (strong [tee] / powerful [computer])
  - Word association (doctor / nurse)
- **Stochastic Context-Free Grammar (SCFG)**

## The rule-based approaches of machine translation of China

- MMT model (multi-labeled and multi-branched tree analysis model)
- IC analysis (Intermediate Constituent Analysis)
- LS method (Logic-Semantic method)
- ST method (String Transformation method)
- I-Tree method (Integrated Tree method)
- **How-Net**

## MMT model

- multi-label tree – algebraic values of the sentence
- multi-branched tree – geometric values of the sentence

## IC analysis (Intermediate Constituent Analysis)

- Logic-semantic principle
- Hierarchical principle
- Contradistinction principle

## LS method (Logic-Semantic method)

- The main logic-semantic features: agent, patient and action
- The subordinate logic-semantic features: space, color, role
- The action is the focus of the logic-semantic structure of a sentence

## ST method (String Transformation method)

- Holding position
- Justifying (changing) position
- Item addition
- Item deletion

## I-Tree method (Integrated Tree method)

- I-tree is the general expression of the structure of the sentence
- Analysis, transfer and generation are the operation of transformation for I-tree: addition and deletion of syntactical elements in the sentence

## How-Net

- Lexical knowledge description system (English & Chinese)
- Formal description for conceptions, the conceptions and their features are organized in a complete system
- Useful language resources in Internet
- Conceptual Design (1988-1993)
- Experiment (1993-1997)
- **Engineering implementation (1997-1999)**
- **Revision (1999-2003)**
- **Development of second resources as evaluation tools (2003-2005)**

## EBMT (example-based Machine Translation system)

- Since 1989, the corpus approach (ex. statistical approach, example-based approach) is introduced to machine translation
- The combination of machine translation with corpus approach will promote the development of the language translation technology in China.

## Japanese-Chinese EBMT system

- The corpus for Japanese and Chinese alignment sentences
- The example unit is sentence
- The similarity rate calculation based on word

## Different channels of Chinese government funding

- The national fund for social sciences (linguistic section)
- The national fund for natural sciences (information science section)
- The Hi-Tech 863 fund (863-IMT/EC system, SUNSHINE YIWANG)
- The 905 Chinese Language Processing Project

## Investment of private companies

- GAOLI MT system is supported by GAOLI private computer company
- YIBA MT system is supported by YAXINCHENG private software company (MINGTAI company)
- TONGYI MT system is supported by DATONG private software computer company

## Three types of users of MT software products

- Government
- Common people
- The state large-scale and medium-scale enterprises

## Common people users

- MT software steadily becomes the popular software that is necessary for common people
- The MT market was formed
- The private companies play more and more important role in driving the MT market

## The MT demand of the state large-scale and medium-scale enterprises

- In these enterprises, there are many technical documents need to be translated into Chinese
- The document is huge
- MT rough translation texts can be welcome by these enterprises

## The features of user region distribution for MT software

- The translation demand is concentrated in the big cities and developing regions
- The MT software must be oriented to big cities and developing regions

## New strategies in translation technology

- Combination with terminology data bank
- Combination with technique of language corpus processing
- Combination with speech technology
- Combination with Chinese characters recognition technology
- Developing the translation technology in INTERNET

## MT software combines with terminology data-bank

- The terminology is crystallization of scientific knowledge in language, it is an important language resource
- The terminology data bank is a very strong support to specialized machine translation

## The national standards for terminology data-bank

- GB/T 13725-92: General principles and methods for establishing terminology data bank, 1992
- GB/T 13725-92: Magnetic tape exchange format for terminological-lexicographical records, 1992
- GB/T 15387.1-94: Guideline for the development of terminology data banks, 1994
- GB/T 15387.2-94: Guideline for the documentation for developing terminology data bank, 1994
- GB/T 15625-95: Guideline for the evaluation of terminology data banks, 1995

## Terminology data banks (1)

- GLOT-C (data processing terminology), Chinese-English, 1988
- TAL (applied linguistics terminology), Chinese-English, 10,000 terms, 1990
- COL (computational linguistics terminology), Chinese-English, 10,000 terms, 1993
- Terminology data bank on machine-building industry: 250,000 terms, Chinese-English-French-German-Russian-Japanese, 1996

## Terminology data banks (2)

- Thesaurus data bank on agriculture: Chinese-English, 25,000 terms, 1991
- Thesaurus bank on chemical industry: Chinese-English, 25,000 terms, 1989
- Encyclopedia terminology data bank: Chinese-English, 800,000 terms, 1997
- Terminology data bank for standardization: Chinese-English

## Chinese language corpus

- Comprehensive Chinese corpus (1983), 20 millions Chinese characters, Beijing Aviation & Space-flight University
- Corpus on Chinese language teaching materials for middle school (1983), 1.068 millions Chinese characters, Beijing Normal University
- Corpus on Chinese Newspapers (1988), 2.5 millions Chinese characters, SHANXI University
- Corpus of People Dairy, Peking University.

## Chinese National Corpus Project

- 70 million Chinese characters
- The selection of this Corpus has three restrictions:
  - Diachronic restriction
  - Cultural restriction
  - Usage restriction

## Corpus Processing

- Automatic segmentation of Chinese writing text in corpus
- Automatic POS (Part of Speech) tagging for Chinese Corpus
- Automatic phrase bracketing and syntactic annotation for Chinese Corpus

## From text corpus to tree bank

- 1 [zj 纱笼/n 。 /w ]
- 2 [zj [fj [dj 纱笼/n [vp 是/v [np [np 马来/n 民族/n ]的/u [np 传统/n 服装/n ]]]] , /w [vp [vbar 富/a 有/v ] [np 浓厚/a 的/u [np 热带/n 情调/n ]]]] 。 /w ]
- 3 [zj [fj [dj [np 纱笼/n 的/u 用途/n ] [ap 很/d 广/a ]] , /w [dj [pp 除了/p [vp [tp 出外/v 时/n ]穿/v ]] , /w [vp 也/d [vbar 被/p [vp 当做/v [np 浴衣/n 、 /w 睡衣/n 和/c [np 婴孩/n 的/u 摇篮/n ]]]]]] ] 。 /w ]

## The speech technology and MT

- Text to Speech software
  - TINGWANG: XUNFEI Company, Anhui Province.
- Chinese speech recognition is relatively easy
  - Chinese: 420 syllables
  - English: 4,030 syllables
  - Russian: 2,960 syllables

## Chinese characters recognition technology (OCR) and MT

- To recognize 6763 Chinese characters in GB 2312-80
- Recognition rate: 99.9%
- Recognition rapidity: real time

## The translation technology in INTERNET

- Many MT software can be used in Internet.
- The advantage for MT software in Internet:
  - Higher translation speed
  - Real-time translation
  - Large machine dictionary
  - Cheap price
  - Possibility to add the new words

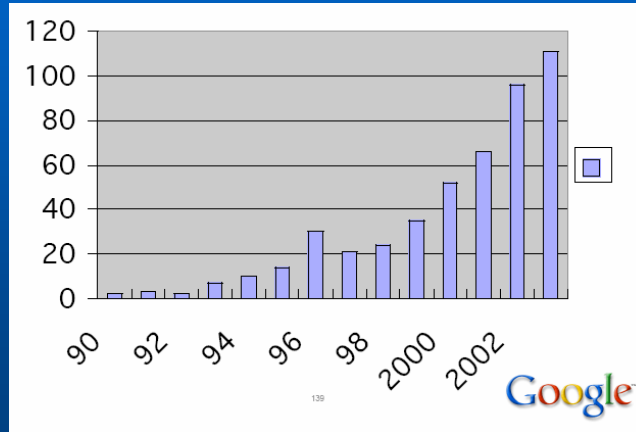
## New Project supported by Chinese government – 973 Project

- 973 will support the innovation research in NLP including MT
- Speech-to-speech MT:
  - NLPR, Institute of Automation, Academia Sinica
  - Kern Member of C-STAR (Consortium for Speech Translation Advanced Research)
- MT system based on HNC (Hierarchical Network of Concepts) theory, Institute of acoustics, Academia Sinica

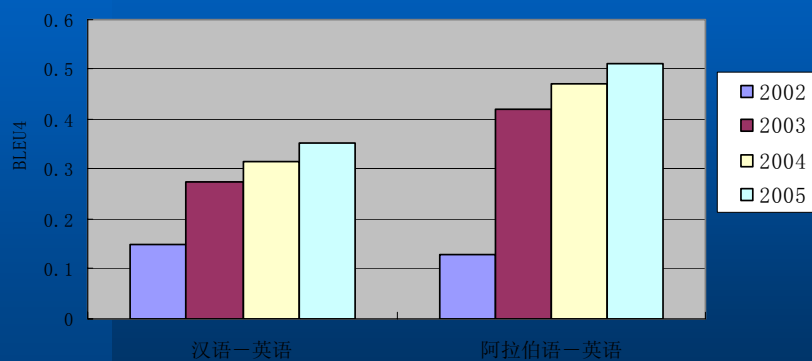
## SMT

- **Machine Translation based on corpus and statistics – Statistical Machine Translation (SMT)**
- Hidden Markov Model (HMM)
- Noisy Channel Model (NCM)
- Parameter Estimation (PE)
- Maximum Entropy (ME)

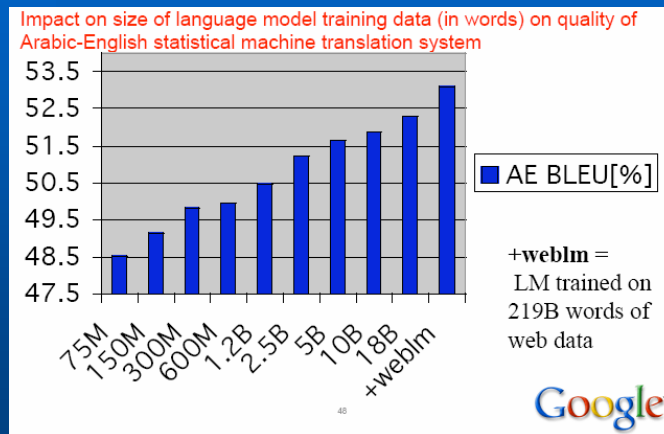
## Paper publication on statistical MT incrementally increased



## BLEU index of SMT



## Impact of size of language model training data on quality of SMT

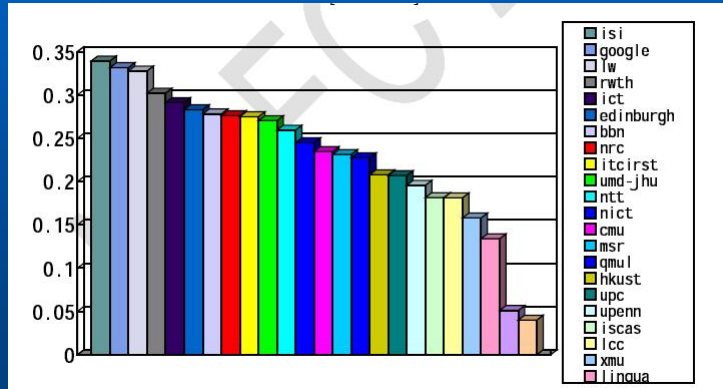


## Example of SMT in Chinese Academy of Sciences

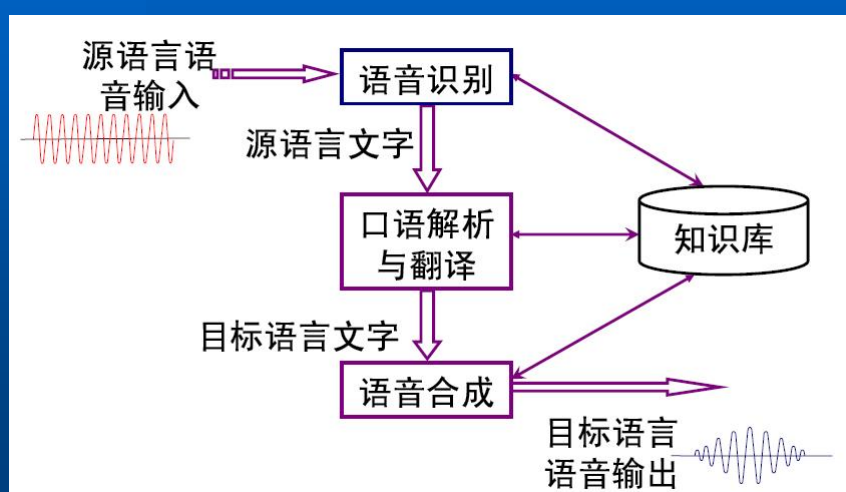
- 新华网拉萨7月2日电，这是举世瞩目的历史时刻：7月2日零时31分，首趟进藏旅客列车鸣响汽笛，稳稳停靠在拉萨火车站1号站台。
- Xinhuanet, Lhasa July 2 (Xinhua), this is the world's historical moment: 0:31 on July 2, the first trip into Tibet, passenger trains rung first, its docked in Lhasa Station No.1 of the **campaign**.

## Bleu4 (NIST 2006)

- CAS-ICT's Bleu4 is fifth in NIST test

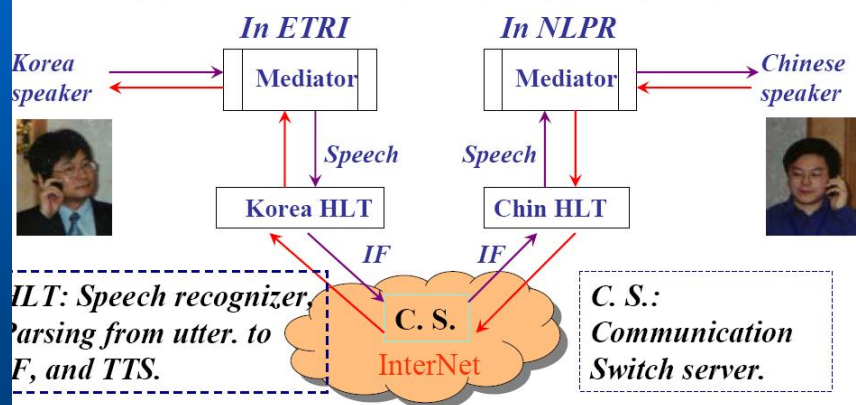


## Speech Translation



## Speech Translation in NLPR

### ❖ 基于普通手机的中韩双向语音翻译系统



## C-STAR (Consortium for Speech Translation Advanced Research)

- China is kern member of C-STAR



## C-STAR kern member delegates



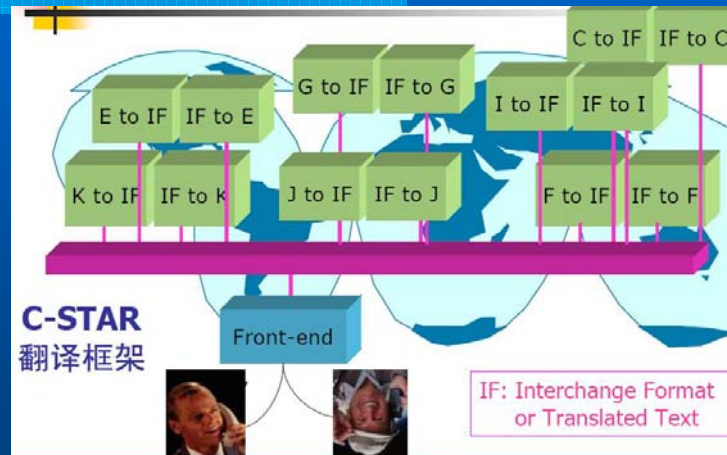
## C-STAR III Goal

### C-STAR III Goal

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



## Interchange Format (IF) of C-STAR

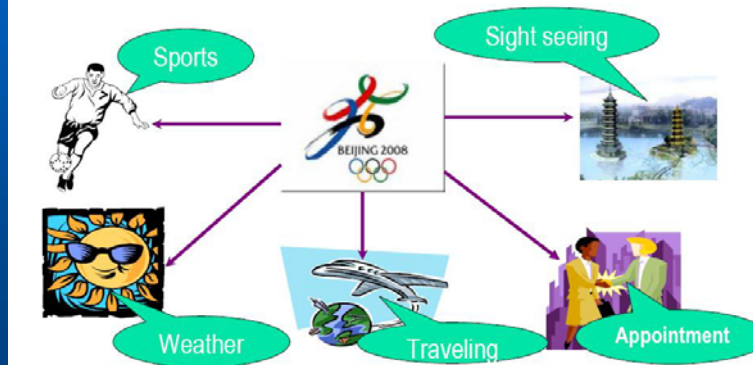


## Japanese-Chinese Speech MT in NLPR



## Multilingual Service system for Olympic Game 2008

□ 面向2008北京奥运会的多语言智能信息服务系统



End

● Thank you !