
Researches on MT and Language Information Processing of CCLIE & BIT

Prof. Huang Heyan

hhy63@bit.edu.cn



School of Computer Sci. & Tech.
Beijing Institute of Technology (BIT)



**Center of Computer & Language
Information Engineering (CCLIE)**
Chinese Academy of Sciences



Outline



Overview

II. Research Fields & Projects

III. Current Status of Research

IV. Accomplishments

V. Products & Applications

VI. Further Works



I. Overview

► History

1985-1992: IMT Project Team of ICT, CAS

1985-1990 Theoretical research & prototype system development

1990-1992 Development of the EC IMT system funded by National 863 plan

1992.6 The **practical IMT/EC-863 system** was developed. Licensed to HK Group Sense Ltd. Co. with USD 740K.

1993-1996: IMT Center of ICT, CAS

1993 EC IMT system won the first prize of CAS Award for Sci. and Tech. Progress

1993 Setup **Huajian Huizhi Co., Ltd** with registered capital of **USD 7.4 m** jointed with HK Group Sense Ltd. Co.

1995 IMT/EC-863 was awarded the **First Prize of National Advance in Sci. & Tech.**

I. Overview

1997-2008: Huajian Group Co. & CCLIE, CAS

1997.6 Setup Huajian Group Co. & *Center of Comp. & Lang. Info. Eng. (CCLIE)*, CAS

1998.6 Undertake the National Technological Innovation project
“Development of Network Information Translation Processing System”

1999.11 Setup **Huajian MT Co., Ltd.** with registered capital of **RMB100m.**

2001.6 Huajian “Multilingual MT System and Its Applications” won **National 9th Five-Year Technology Innovation Award** of China

2001.11 Huajian “Multilingual Large-scale Network Information Integrated Processing System” won the **2nd prize for Sci. and Tech. Progress, CAS**

2005.10 The “Multilingual Info. Processing Tech. and Its Application Industrialization” was listed in the special projects of National Deve. and Reform Commission of China

2006.6 Setup **Huajian E-tone Tech. (Beijing) Co., Ltd.** with registered capital of **RMB10m.**

2007.3 Setup **Ningbo Huajian Info. Tech. Co., Ltd.** with registered capital of **RMB2m.**

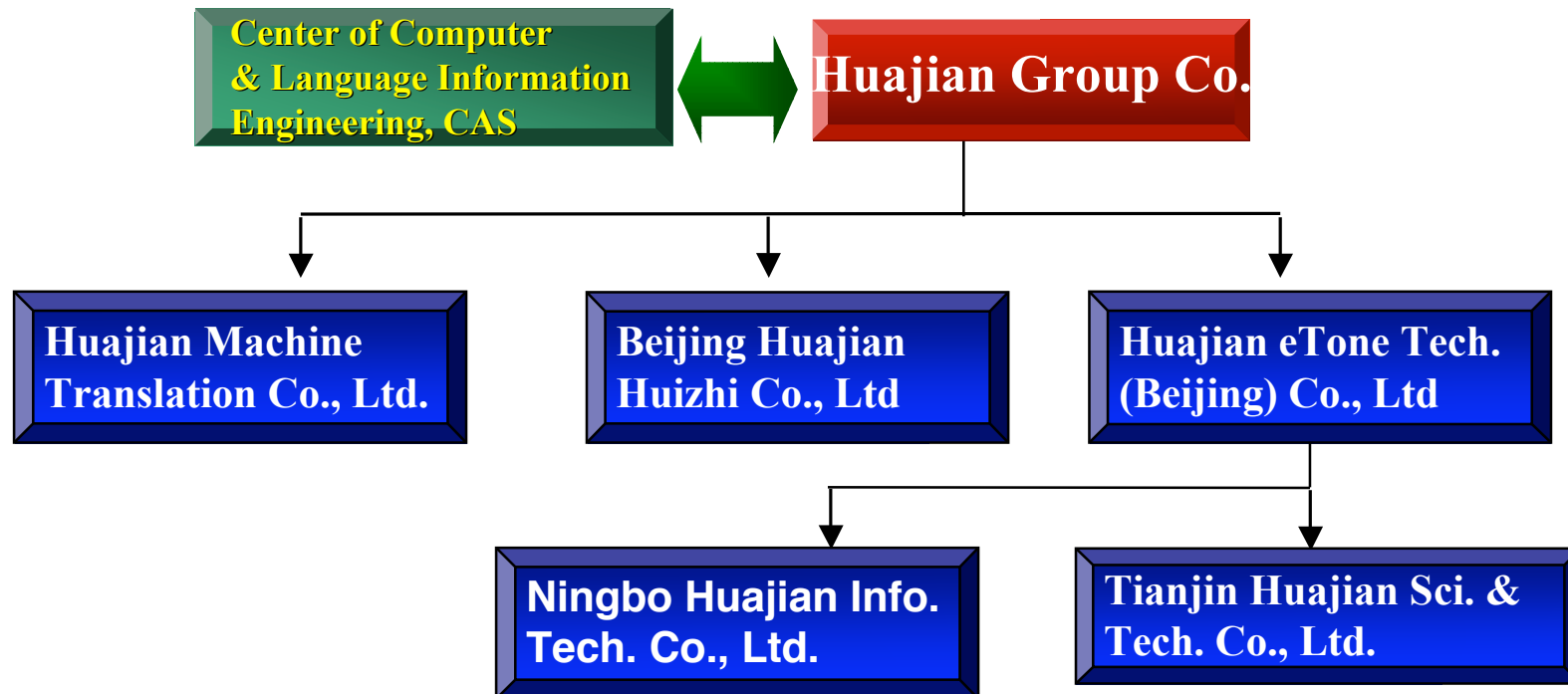
2008.6 Setup **Tianjin Huajian Sci. & Tech. Co., Ltd.** with registered capital of **RMB5m.**

2009-now: Huajian Group Co. & CCLIE, CAS + CST, BIT

2009.3 Setup **Inst. Of Lang. Info. Proc. Of BIT**

I. Overview

► Structure of the Huajian Group Co. & CCLIE, CAS



I. Overview

► Research Team

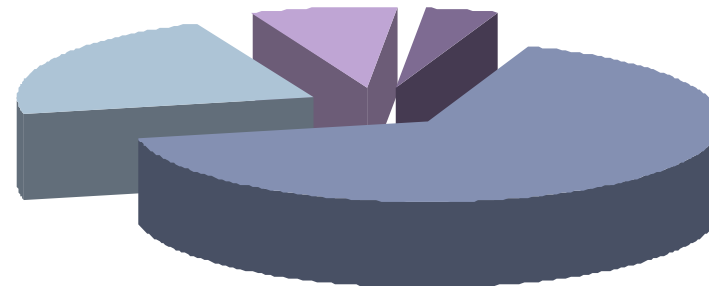
Two parts: CCLIE, about 60 persons; BIT, about 20 persons

Total : about 80

Average age: 30

Ph.D degree: 8%

Master degree: 22%



■ bachelor ■ master ■ doctor ■ others

II. Research Fields & Projects

- ▶ **Research Fields:**
 - ▶ **Machine Translation**
 - ▶ **Cross-Language Information Retrieval**
 - ▶ **Topic Detection**
 - ▶ **Adaptive Information Extraction**
 - ▶ **Linguistic sources Construction**

II. Research Fields & Projects

▶ Projects:

- ▶ 2007-2010 National High-tech R & D Program (863 Program):
Massive Information Processing IHSMTS Research
- ▶ 2007-2009 National Natural Science Funds: MT-oriented Text Identification
- ▶ 2007-2008 Ningbo Science and Technology Bureau: Ningbo Multilingual Information Platform for Public Service
- ▶ 2007-2008 National High-tech R & D Program (863 Program):
Multilingual Natural Spoken language Dialogue System Key Technology Research (in cooperation with Institute of Acoustics, CAS)
- ▶ 2006-2008 National High-tech R & D Program (863 Program):
Research on Internet-oriented Adaptive Information Extraction Technology

-
- ▶ **2005-2008** Shanghai Science and Technology Committee research projects (World Expo information specific project): Multilingual Information Integration Technology Research and Implementation
 - ▶ **2004-2006** National network and information security sustainable development project (confidential): Research on Emergencies and Topics Detection and Semantic Orientation Recognition Technology
 - ▶ **2004-2005** National Computer Network and Information Security Management Center: Cross-language IR system

III. Current Status of Researches-MT

► Intelligent Machine Translation

1986-1988 Theoretical pursue on IMT approach

1989-1990 Implementation of the prototype
system

1990-1992 Implementation of the practical
English-Chinese
IMT system

1992-2002 Implementation of the practical multi-
lingual

IMT system and specialty IMT system,
including:

English-Chinese

Chinese-English

Russian-Chinese

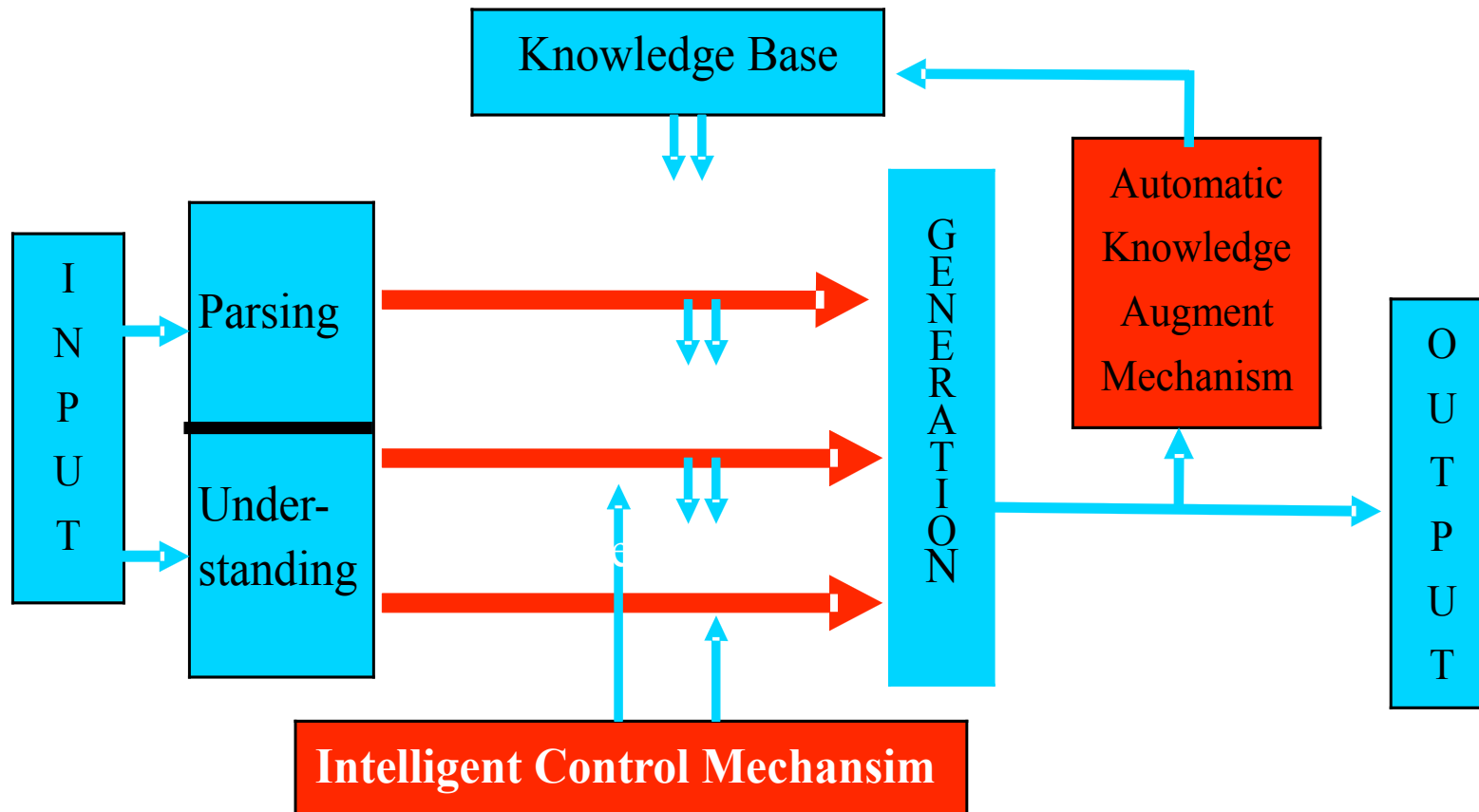
German-Chinese

Japanese-Chinese

Chinese

III. Current Status of Researches-MT

► Overall Structure of IMT:



III. Current Status of Researches-MT

► **Features:**

- **Uniform representation form for multiple knowledge – SC-grammer:**

<HeadCond> → <ContextCond> | <Reduction>,
<Trans>

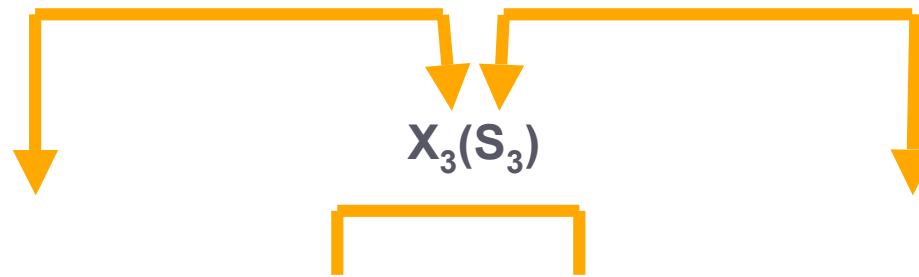
- **Adjustable classified hierarchical structure for description of component features pattern**
- **Context sensitive features pattern**

<S/NS> (L/R, <Ran>, <Comp>)

- **Analysis and transformation mechanism based on incomplete knowledge**
 - **Hierarchical adjustable consistent unification algorithm**



► **context sensitive Processing :**



$$X_1(S_1) X_2(S_2) \rightarrow [X_1(S_1) X_2(S_2)] S(L, <1, i>, X_i(S_i)) S(R, <1, j>, X_j(S_j)) | X_3(S_3)$$

► **Integrated parsing of multi-level features:**

$$X(X_1 (X_{11}, \dots, X_{1m}), X_2 \dots, \dots, X_n)$$

By changing the features' hierarchy attribute, sentence is incorporated analyzed with syntax, semantic and common sense knowledge.

- Combination of transformation-generation process and parse process, which localizes ambiguity handling in the generation process, greatly optimizes translation process and improves efficiency and accuracy.
-

III. Current Status of Researches-MT

► Hybrid Strategy Machine Translation

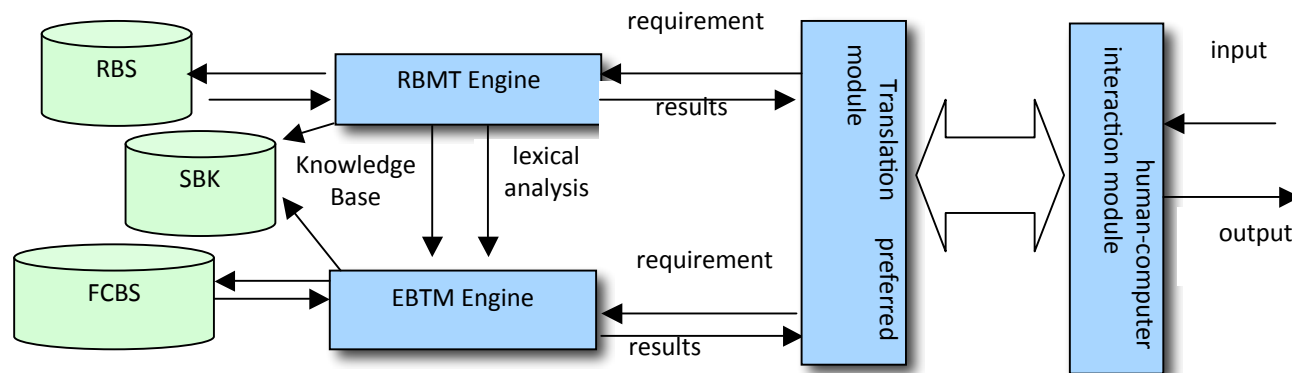
1998-1999	Theoretical pursue on hybrid strategy
	machine translation approach
2000-2001	Implementation of the prototype system
2001-2003	Implementation of the practical system
2004-now	Improvement



III. Current Status of Researches-MT

▶ Key Notes

- ▶ **Combination of rule-based method and corpus-based method**
- ▶ **Multi-level grammatical feature Architecture and Multi-knowledge unified SC (Sub-Category) Grammar**
- ▶ **Multi-level abstract knowledge representation of case model**
- ▶ **Context sensitive processing based on forward and feedback analysis**
- ▶ **Analysis and transformation mechanism based on incomplete knowledge with compatible features**
- ▶ **Sentence similarity computation based on multi-level features**



III. Current Status of Researches-MT

► Features

► Featured Case model:

Complex Logical Operation Based representation for Featured Case model: $M_1 M_2 \dots M_n$

M_i can be a word string, or a “<SynCate>(<FD>)”, in which SynCate indicates its grammar category, while FD indicate other properties: lexical, semantic & reduce length.

► Featured Case Analogy Matching Based on Feature Weight:

Feature weight: plays different role (key, subordinate)

Feature interval: similarity between two structure components, defined as:

$$D(f_1, f_2) = [0,1]$$

Threshold setting and heuristic strategy

► Combined hybrid strategies (example-based, rule-based, template-based) by featured case model matching

III. Current Status of Research-Cross-Language IR

- ▶ **Contents**

- ▶ Multilingual conversion of sentence and keyword search
- ▶ Text retrieval (including keyword retrieval, category retrieval, similarity retrieval)
- ▶ MT technology of Translated results contrasting Original texts

- ▶ **Features**

- ▶ Several information retrieval methods including cross-language keyword retrieval, cross-language category retrieval and cross-language similarity retrieval
- ▶ Multilingual support: Chinese, English, Japanese and Russian
- ▶ Fast response: keyword search response time of less than 1 second

III. Current Status of Research-Cross-Language IR

▶ Progress

- ▶ Realization of Word segment technology of Chinese, English, Japanese, Russian in the same MT engine, and construction of stop word lists corresponding above four languages for information retrieval
- ▶ Supporting large-scale corpus above 100GB
- ▶ Providing functions of keyword retrieval, category retrieval and similarity retrieval
- ▶ Realization of multilingual coding conversion and automatic language identification
- ▶ Multi-format processing and presentation of original text and translated results (Fully compatible with WORD, RTF, HTML, etc.)
- ▶ Fast response: keyword search response time of less than 1 second
- ▶ High precision and recall:
10p=92.8%;@100p=88.9%;Recall=93.2%

III. Current Status of Researches - Topic Detection

▶ **Contents**

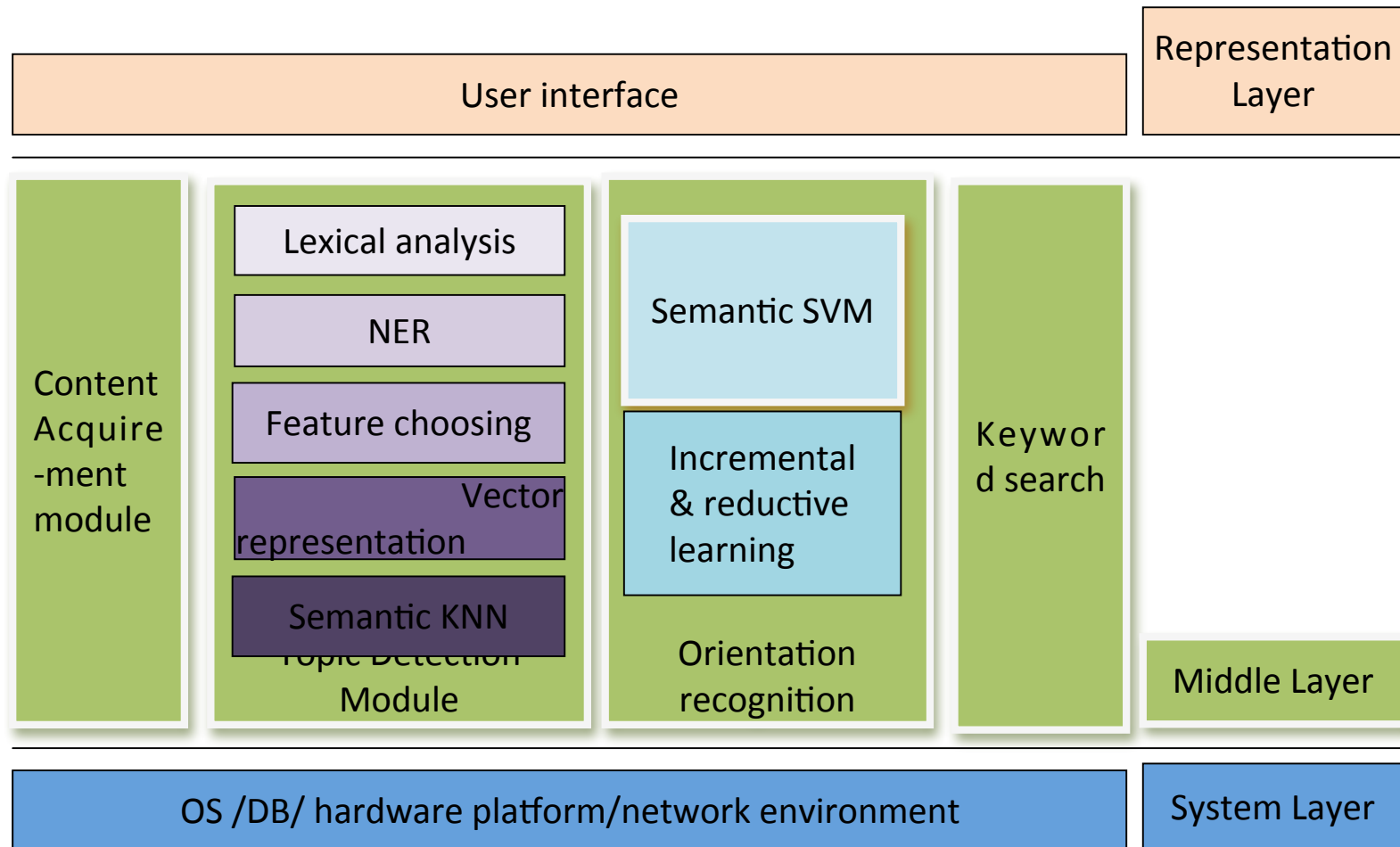
- ▶ **Technology of real-time acquirement to Internet information**
- ▶ **New topics detection technology**
 - ▶ **Clustering and data mining technology based on semantic analysis**
 - ▶ **Keyword extraction**
 - ▶ **Abstract generation**
- ▶ **Technology of sensitive contents semantic orientation recognition**
 - ▶ **Clustering technology based on semantic orientation**
 - ▶ **Automatic semantic feature extraction**
 - ▶ **Text semantic orientation recognition**
 - ▶ **Recognition model reduction and incremental learning**
 - ▶ **Recognition model fast training technology**
- ▶ **Technology of Local storage management**

III. Current Status of Researches-Topic Detection

► **Features**

- Detection of emergencies by semantic kNN-based text clustering algorithm, which can improve efficiency while effectively recognizes topics-specific in network information stream.
- Recognition of semantic orientation using semantic SVM. Parsing information on semantic level to determine semantic orientation of texts with sensitive contents.
- Training algorithms of semantic orientation reorganization not only has high efficiency but also can well adapts to changes in number of training base through incremental and reduction learning.
- Building Semantic feature space based on Huajian MT lexical analysis engine, and introduction of weight computing with separated named entities and general features

III. Current Status of Researches-Topic Detection



III. Current Status of Researches- Topic Detection

► **Progress:**

- Classify the newly detected topic as a new information category, and automatically classify the following relevant articles into this category
- Semantic orientation recognition algorithm in support of online reduction or incremental learning
- An open system that can support cluster computing and load balancing functions
- Information acquirement from HTTP-based forum or BBS
- Average accuracy rate of detection required to be over 45% (relying on the amount of information and features of topics)
- Average accuracy rate of recognition with given topics required to be over 90% (relying on features of topics)

III. Current Status of Research- Adaptive IE

- ▶ Contents:
 - ▶ Language independent Information Extraction technology to develop a set of fundamental structure with adaptive capacity using statistical machine learning method
 - ▶ Named Entity Recognition (including person name, location name, and organization name)
 - ▶ Domain-specific Terminology extraction
 - ▶ Complex NP recognition (length \geq 5 words)
 - ▶ Anaphora and Co reference Resolution

III. Current Status of Research- Adaptive IE

► **Features:**

- Little reliance on specific linguistic rules, minimal dependence on annotated corpus, with adaptive capacity of information extraction in different languages and fields.
- Use of semantic information contained in radicals of Chinese characters based on maximum entropy framework, to study the value of radicals as language sources.
- Adaptive NER method based active learning technology, adopting pool-based sample selection with weighed density, to evaluate raining value through computing weighed credibility and sample density.
- Coreferential resolution framework based on domain ontology sources, to identify formalized examples from texts with ontology as support platform, the essence of which is that the core is ontology rather than dictionaries or gazetteers.

III. Current Status of Research- Adaptive IE

▶ **Progress:**

▶ **Adaptive Named Entity Recognition**

- ▶ F-measure : date, expressions of time, percentage and amount up to more than 90%, location and person name basically 80%, and organization name basically 70%

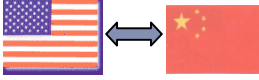
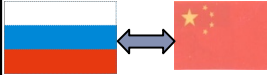
▶ **Complex maximal noun phrase (xMNP) recognition**

- ▶ Research on key problem of current MNP recognition: recognition of xMNPs with no less than 5 words
- ▶ Basically up to 60%











▶ **Anaphora and Coreference Resolution in domain**

- ▶ Coreference resolution framework based on domain ontology, introducing domain semantic feature, precision of domain coreferential resolution basically up to 85%

III. Current Status of Researches— Linguistic sources

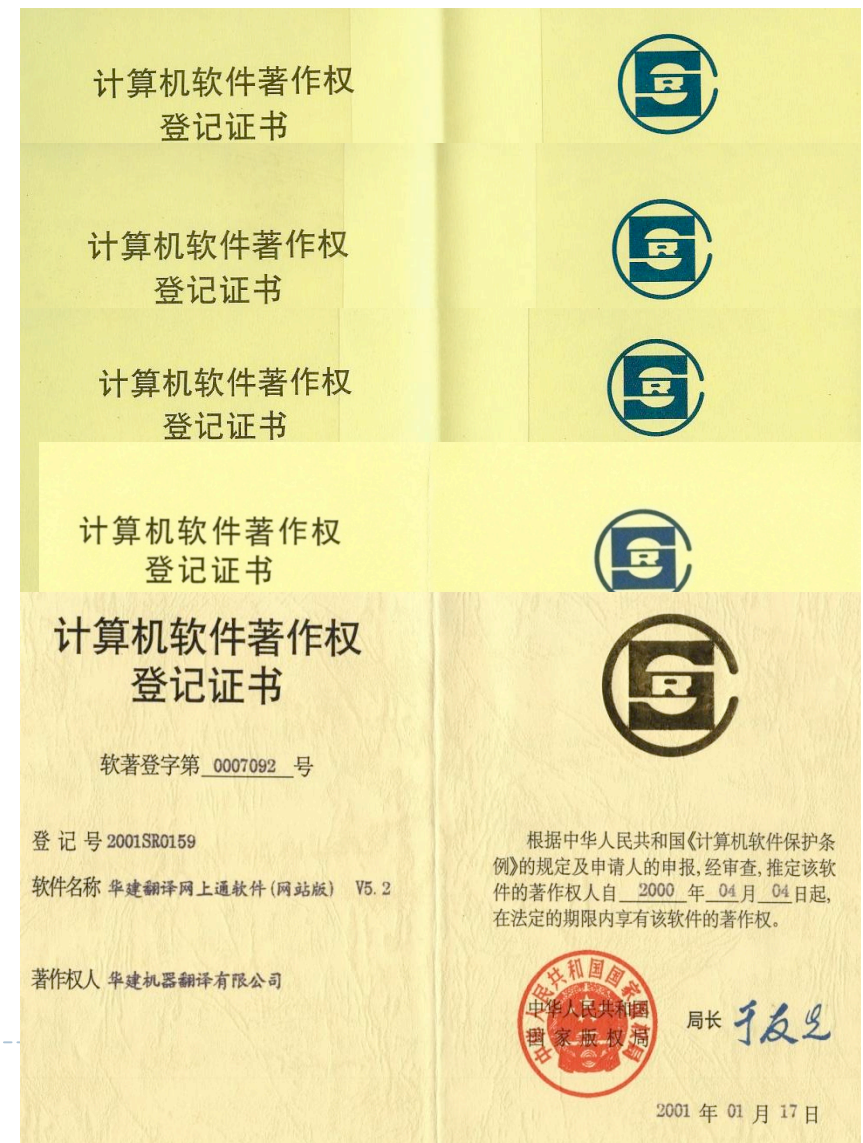
System	Description	Data	Specialty Bases
Eng→Chn 	Having the longest development history, stable and mature, having relatively complete specialty bases, mature speech MT system	Vocabulary: 280,000 rules: 6000	29 professional databases
Chn→Eng	developed Having been tested and tuned with Internet corpus, good translation quality and high readability rate of basic sentences and dialogues.	Vocabulary: 250,000 rules: 4928	24 professional databases
Rus→ Chn 	After many years' development, the system is relatively stable, because it has been trained with a great deal of Internet corpus, the translation result is good.	Vocabulary: 190,000 rules: 3920	Professional database in astronomy, mechanics, etc
Chn→ Rus	Being developed from early 2003, not mature and need to be tuned and improved with a great deal of Internet corpus.	Vocabulary: 180,000 rules: 2070	

III. Current Status of Researches— Linguistic sources

System	Description	Data	Specialty Bases
Jap→ Chn  ↔ 	Being developed from 2000. The speech MT system is of better readability for basic sentences and dialogues.	Vocabulary:230,000 rules: 2736	
Chn→ Jap	Being developed from 1999, relatively mature system; speech MT system developed.	Vocabulary:240,000 rules:2306	Professional database in sports and traveling
Ger→Chn  → 	Being developed from 1993	Vocabulary:112, 000 rules:2400	Professional database in automobile
Chn →French  → 	Being developed from 2000	Vocabulary:148,000 rules:3500	
Chn → Spn  → 	Being developed from 2000	Vocabulary:160,000 rules:3700	
Chn→Kor  → 	Starting from 07/2006; satisfactory translation of common sentences and daily conversations	Vocabulary:85,000 rules:2200	
Kor→Chn	Starting from 01/2007; spoken language system in stage of improving	Vocabulary:95,000 rules:2000	

IV. Accomplishments

- ❖ 8 Patents
- ❖ More than 60 Software Copyrights
 - ❖ Intelligent EC MT System IMT/EC V1.0
 - ❖ Pocket Intelligent EC MT System V1.0
 - ❖ Huajian Net Wizard V1.0—V6.0
 - ❖ Huajian Bilingual Browser V1.0
 - ❖ Huajian Net Wizard (Web Site Edition) V5.2
 - ❖



V. Products & Applications -1

❖ Embedded Intelligent MT System

❖ **GSL(InstantDic), HK:**

- ❖ Instant 863A/B/C
- ❖ Instant Longman 8688 Bilingual Translator

❖ **Besta, InvenTec, TaiWan:**

- ❖ Learning Machine, Electronic Dictionary, PDA
- ❖ With an annual sales volume of more than 200,000 sets

❖ **OKWAP:**

- ❖ English Learner 885i

❖ **Aigo:**

- ❖ MID Product



鴻文8688



V. Products & Applications -2

- ▶ **Multilingual Value-added Service Platform**
 - ▶ **SMS Translation Service Platform for Telecommunication Value-added Service**
China Unicom: 938686
China Mobile : 501286
Four Network Union: 10662008
 - ▶ **WAP Intelligent Translation System**
 - ▶ **MSN Real-time Translation System— Xiao i Robot, 240,000 times per week**



Multilingual Mutual Translation Platform



Huajian Xiaofan MSN Xiao i Robot

V. Products & Applications -3

- ▶ **Intelligent Aided Translation Platform:**

Combining hybrid translation strategies, including: rule-based, corpus-based, case-based heuristic analogy matching, etc., and realizing the organic combination of aided translation, bilingual alignment, terminology management and translation process management to professional translation organizations /translators

- ▶ **Successful Cases:**

Beijing 2008 Olympic Organizing Committee

State Intellectual Property Office

China National Institute of Standardization

Navy Translation Team, Equipment Department

Institute of Air Force

First Institute of the Ministry of Aviation

China Academy of Transportation Science

Deloitte Law Office

...

V. Products & Applications -4

▶ PC Translation Software

- ▶ **Huajian Net Wizard**: With an annual sales of nearly 450,000 sets in overseas market (PCCW as the general agent)
- ▶ **Huajian Bilingual Browser**: Bounded with PC manufactures such as Lenovo, Hisense, Tongfang and TCL,), with the annual sales of more than 3 million sets (Hong Kong PCCW as the general agent , “Pacific CyberTrans Bilingual Browser”)
- ▶ **Huajian EasyTrans**: With Japanese Logo Vista as the agent, Authorizing Taiwan Invente SDK
- ▶ **Speed translation-family EICQ**: With Loto (USA) as the agent
- ▶ **Huajian Dictionary**



V. Products & Applications -5

- ▶ **Sector-oriented Multilingual Application and Integration:** Applied extension of multilingual information processing system with MT as the technological core, including:
 - ▶ **Multilingual Topics Detection**
 - ▶ **Multilingual Information Retrieval**
 - ▶ **Multilingual Sensitive Information Filtering**
 - ▶ **Multilingual Text Classification**
- ▶ **Application**
 - ▶ **State Intellectual Property Publishing House**
 - ▶ **Multilingual Mutual Translation Platform of Beijing Post**
 - ▶ **Safety Departments such as National Computer Network and Information Safety Management Center**
 - ▶ **Ningbo Multilingual Information Service Platform**

选项

SQL Server 登录

DB服务器(S): 210.72.14.156 文件服务器(S): 210.72.14.156

☐ 使用信任连接DB服务器(U)

登录 ID(L): hjs spider 密码(P): *****

主题设置

更新的相似度阈值: 0.45 归类的相似度阈值: 0.15



VI. Further Works

❖ Questions:

❖ How to be well situated for computer :

Huge conceptual knowledge base:

conceptual entities & relations: wordNet, HowNet,

common sense knowledge: Wiki, iAsk of Sina,

various background knowledge: e-science, digital lib.

Ability to sense & recognize environment:

computer vision, image processing, sensor network,

❖ How to be real intelligent for computer:

learning ability:

reasoning, mining,

selecting & judging ability based on value evaluation:

cognitive science, brain science, psychology,

❖ A long way to go for computer intelligence and for NLP

VI. Further Works

❖ Further researches:

- ❖ Multilingual ontology, such as for Euro languages and Chinese, task-oriented design of ontology for MT, Cross Language IR
- ❖ The uniform description method of multi-lingual concepts and related knowledge
- ❖ Meaning computing algorithms based on the multilingual ontology
- ❖ Combined the related works, such as computer vision, sensor network, machine learning and so on with NLP
- ❖ The architecture of multi-lingual processing & service platform based on cloud computing
- ❖ Various approaches & technologies to improve the performance of NLP, such as pre-parsing and pre-editing for documents, long sentence processing, contextual processing, etc.

VI. Further Works

❖ Industrial-oriented Development

❖ Aiming at different users

- ❖ Government, medium and small enterprises, professionals, and general users

❖ Focusing on different fields

- ❖ Practical MT tools
- ❖ High-performance IAT tools
- ❖ Multilingual dialogue and short sentence systems
- ❖ Multilingual information service of specific domains under discourse and full-text

❖ Supporting for different applications

- ❖ Multilingual real-time browsing and knowledge management
- ❖ Multilingual value-added service for Internet & Telecommunication
- ❖ Multilingual data mining and information parsing
- ❖ Multilingual Enterprise Information Portal (EIP)



Thanks !