Microsoft®
微软中国研究院

Overview of Natural Language Computing Group

Ming Zhou （周明）
Group Manager
Microsoft Research Asia
微软亚洲研究院

Microsoft®
微软中国研究院

Outline

- Introduction
- Research Projects
- Future Work

Microsoft Research Asia

http://www.microsoft.com/china/research

Microsoft® 微软中国研究院

- Microsoft Research China
  - Established in Nov. 1998
  - Corporate Research Arm in Asia-Pacific Region
  - Renamed to Microsoft Research Asia in Oct.2001
- Performing research in
  - Digital Media
  - Multimodal User Interface
  - Wireless and Networking
  - Information Processing
- People
  - 9 research groups
  - 100 researchers & 100 visitors

Natural Language Computing Group

Microsoft®
微软中国研究院

- One of the three key NLP groups in Microsoft
  - NLC group (Beijing), 20 researchers/visiting researchers
  - NLP group (Redmond), 50 researchers
  - NLG group (Redmond), 150 researchers/developers
- Co-work with other groups in Microsoft
  - Speech.NET group (Redmond), 150 researchers/developers
  - Speech group (Beijing), 20 researchers
  - User interface group (Beijing), 20 researchers/visiting researchers
  - Machine learning group (Redmond), 10 researchers
  - Information Retrieval group (Cambridge), 10 researchers
  - IME, Office, Windows, MSN, UI…
- Close collaboration with universities
  - Machine translation joint lab at Harbin Institute of Technology
  - Natural language processing joint lab at Tianjin University
  - ……

Multiple Language Information Processing on Internet

Microsoft®
微软中国研究院

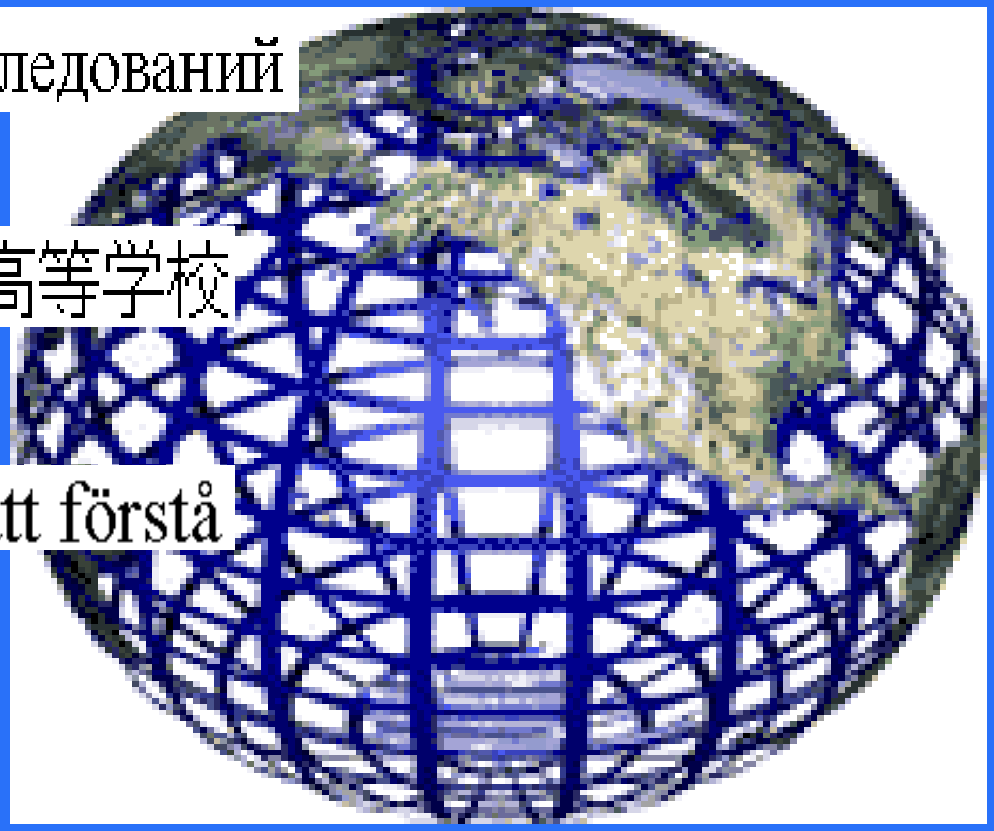
يا ليلي يا عيني

których można

Исследований

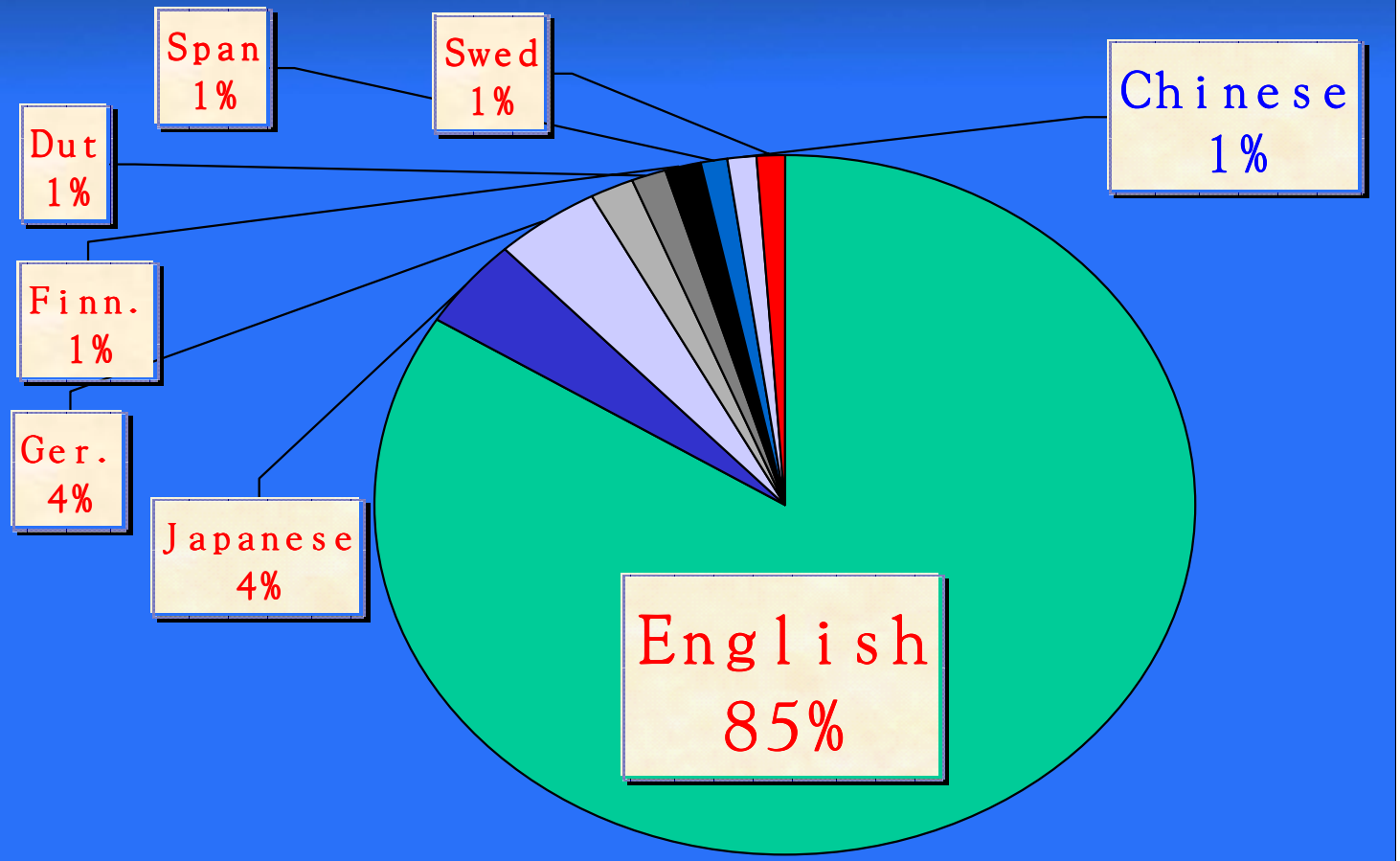
高等学校

att förstå



Language & Information Barrier

Microsoft®
微软中国研究院

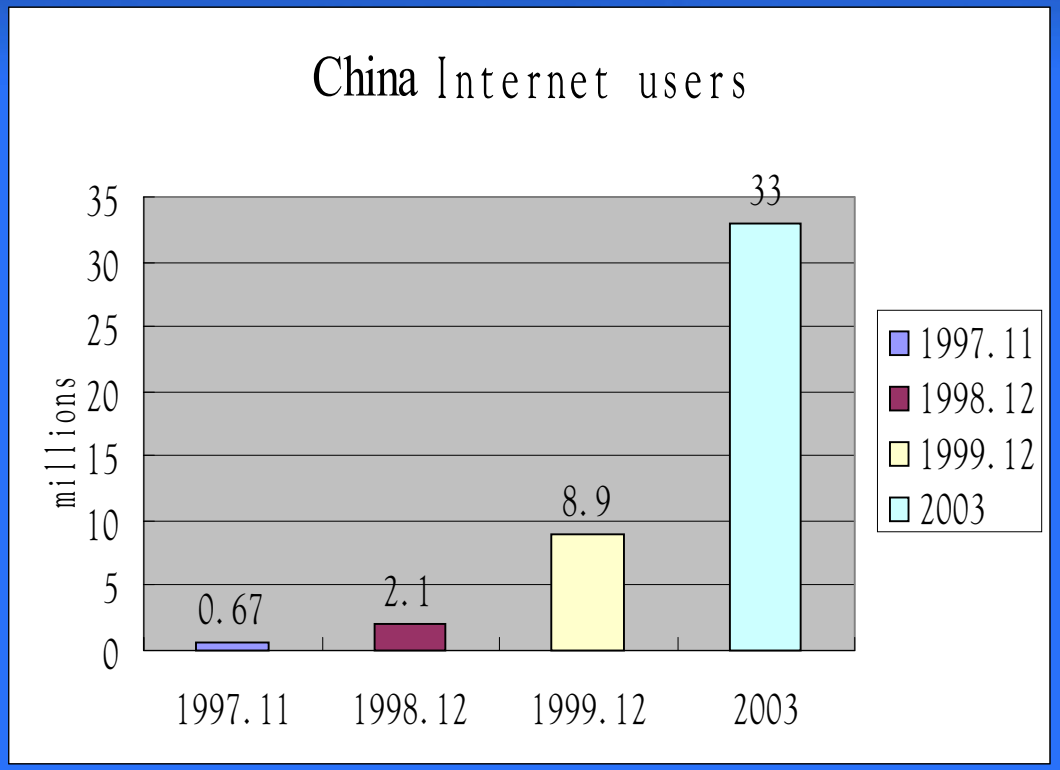


Source: Network Wizards Jan 99 Internet Domain Survey

China: The World's Fastest Growing Internet Market

Microsoft 微软中国研究院

- 1997-1999, China Internet users 300% increases each year
- China will become the largest Internet country



Soucce: China Internet Network Information Center

Research Overview

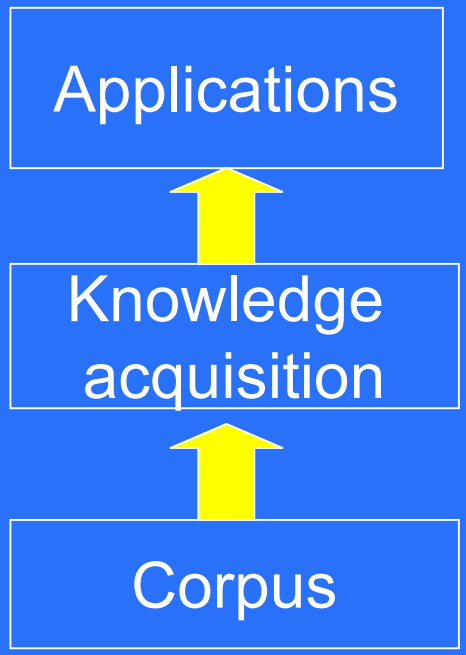
Microsoft®
微软中国研究院

Mission:

Overcoming Language & Information Barrier in Internet Era for Asian Users

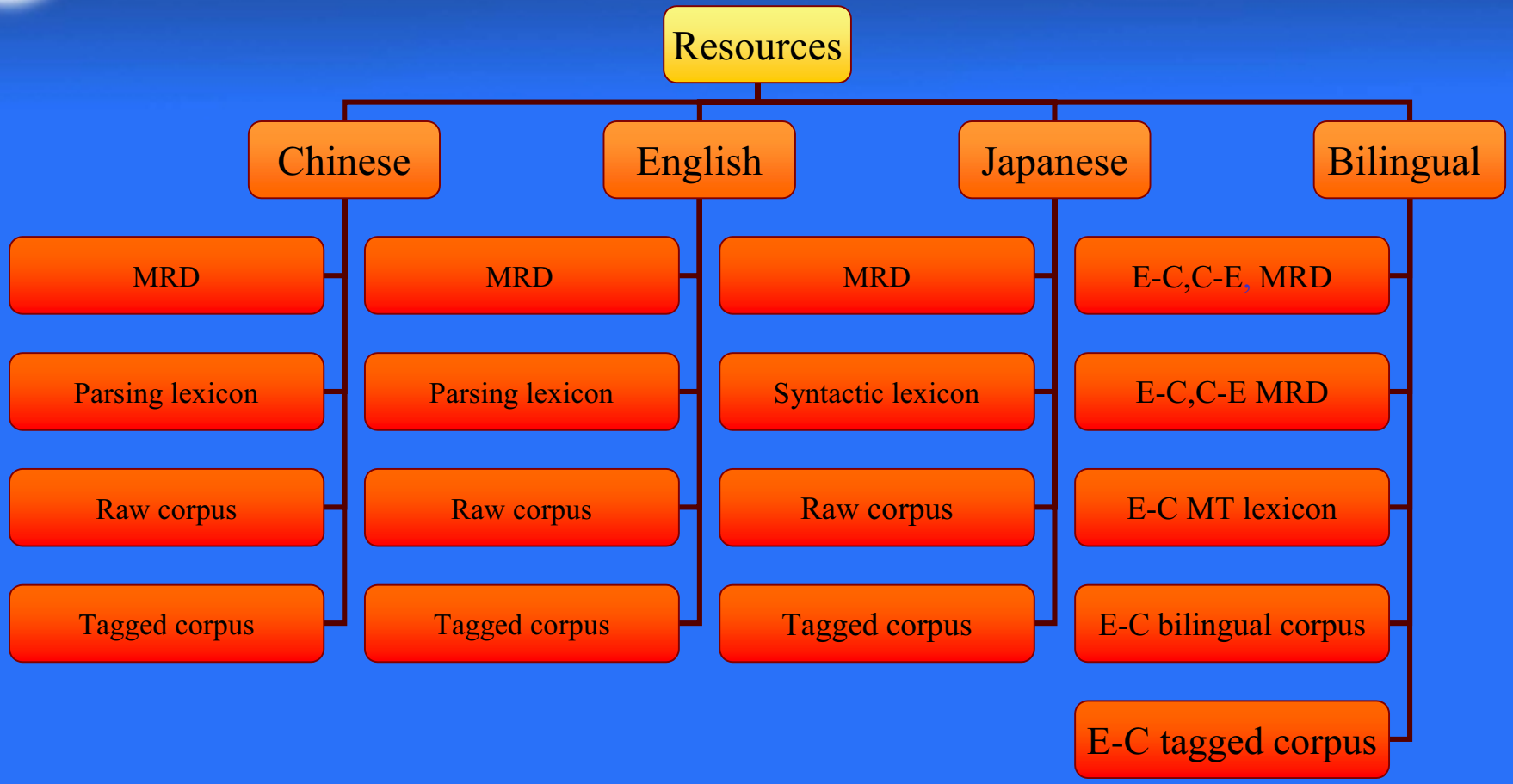
Research Directions

1. Asian language processing
2. Statistical language learning
3. Machine assisted translation
4. Information retrieval



NLP Resources

Microsoft®
微软中国研究院



Word Alignment

Microsoft 微软中国研究院

- {I}1 {will}2 {check}3 {this}4 {letter}5 {for}6 {you}7 {,}8 {if}9 {you}10 {want}11 {.}12
- {你}10 {要}9 {想}11 {让} {我} {给} {你} {查} {查} {这}4 封 {信}5 {，}8 {我}1 {就} {给}6 {你}7 {查 查}3 {。}12

- {But}1 {,}2 {in contrast}3 {,}4 {it}5 {is}6 {only}7 {in modern times}8 {that}9 {Galileo}10 {has}11 {become}12 {a}13 {problem}14 {child}15 {for}16 {historians of science}19 {.}20
- {但是}1 {相比之下}3 {，} {对于} {科学史家}19 {来说} {，} {伽利略}10 {只是}7 {在 现代}8 {才} {变成}12 {了} {一个}13 {成 问题}14 {的} {孩子}15 {了} {。}20

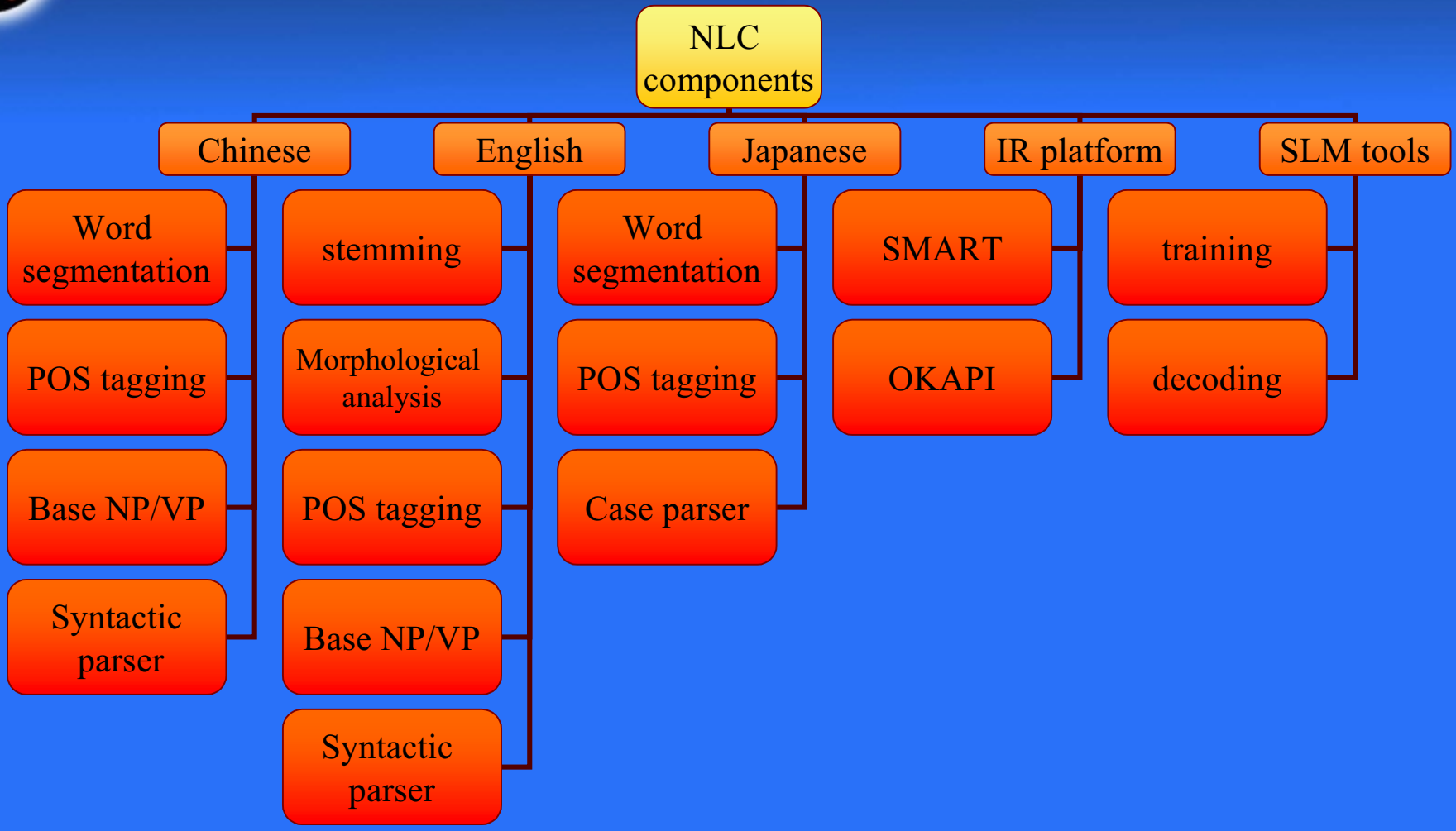
Corpus Tools

Microsoft® 微软中国研究院

- ✓ Corpus sampling
- ✓ Word segmentation
- ✓ POS tagging
- ✓ Bilingual text alignment(text, sentence, word level)
- ✓ Monolingual lexicon extraction
- ✓ Bilingual lexicon extraction
  - – from bilingual corpus, parallel or comparable
  - – from two mono-lingual corpus

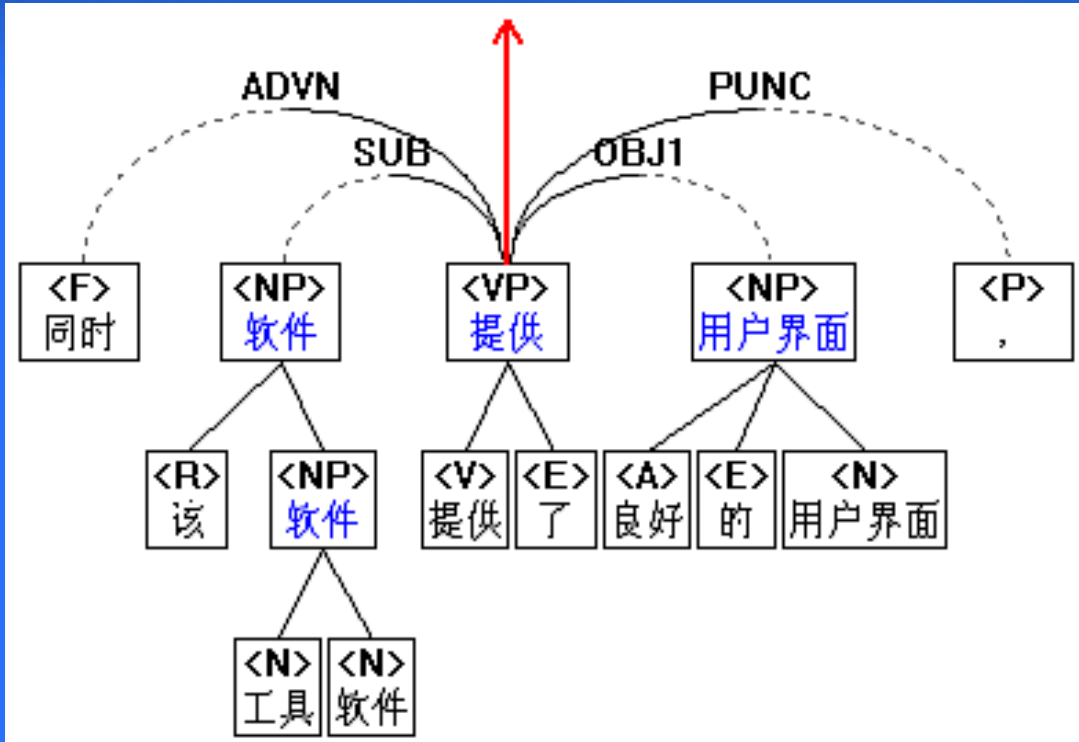
NLP components

Microsoft®
微软中国研究院



Chinese Block Based Dependency Parser

Microsoft
微软中国研究院



- Input a text, output a dependency structure at high speed
- Block identification plus dependency analysis
- Rule-based robust parser
- 200,000 words
- Practically used

Knowledge Acquisition

Microsoft 微软中国研究院

- Statistical Language Modeling (SLM)
  - Based on character, word and class;
  - Long distance dependency
- Lexical knowledge
  - New terms
  - Word similarity/collocation, thesaurus, association;
- Syntactic knowledge
  - Phrase and dependency rules
  - Resolving ambiguous structure
- Translation knowledge
  - Translation selection
  - Translation template

Research Projects

Microsoft®
微软中国研究院

- Chinese IME
- Chinese spelling check (Apr. 2000~)
- Information Retrieval (Mar. 2000~)
- English reading wizard (Nov. 1999~)
- English writing wizard (Oct. 1999~)
- Example based machine translation for MS Localization Tool (Apr. 2001~)
- Web dictionary builder (July, 2001~)

Microsoft®
微软中国研究院

Chinese IME

Powerful Language Model
(Presented in ACL-2000)

Microsoft 微软中国研究院

- Role of language model in conversion
  - For all possible word* strings that match typed pinyin:
  - Select W with the highest language model probability.
- How to estimate Pr (W):
  - $\Pr(W) = \prod^{n} \Pr(W_n \mid W_{n-1}, W_{n-2}, W_{n-3}\ldots, W_1)$
  - Trigram Approximation : $\prod^{n} \Pr(W_n \mid W_{n-1}, W_{n-2})$

Microsoft®
微软中国研究院

Demo: Chinese Input

Microsoft®
微软中国研究院

Information Retrieval

Highlights

Microsoft®
微软中国研究院

- Participated in CLIR track in TREC-9
  - Finding the best indexing units for Chinese IR
  - BaseNP identification and translation (ACL-2000, SIGIR'2001)
  - Enhancing bilingual lexicon with parallel corpus
  - Among top performance evaluation
- Participated in Web track in TREC-10
  - Collaboration with Stephen Robertson (Cambridge）
  - Enhance conventional IR techniques for web retrieval
  - Investigate the use of link information
- NLP Enabled IR technologies
  - Extended indexing
  - Automatic word clustering/word thesaurus building
  - Avoiding noisy query expansion
  - New query translation model & translation selection

Microsoft®
微软中国研究院

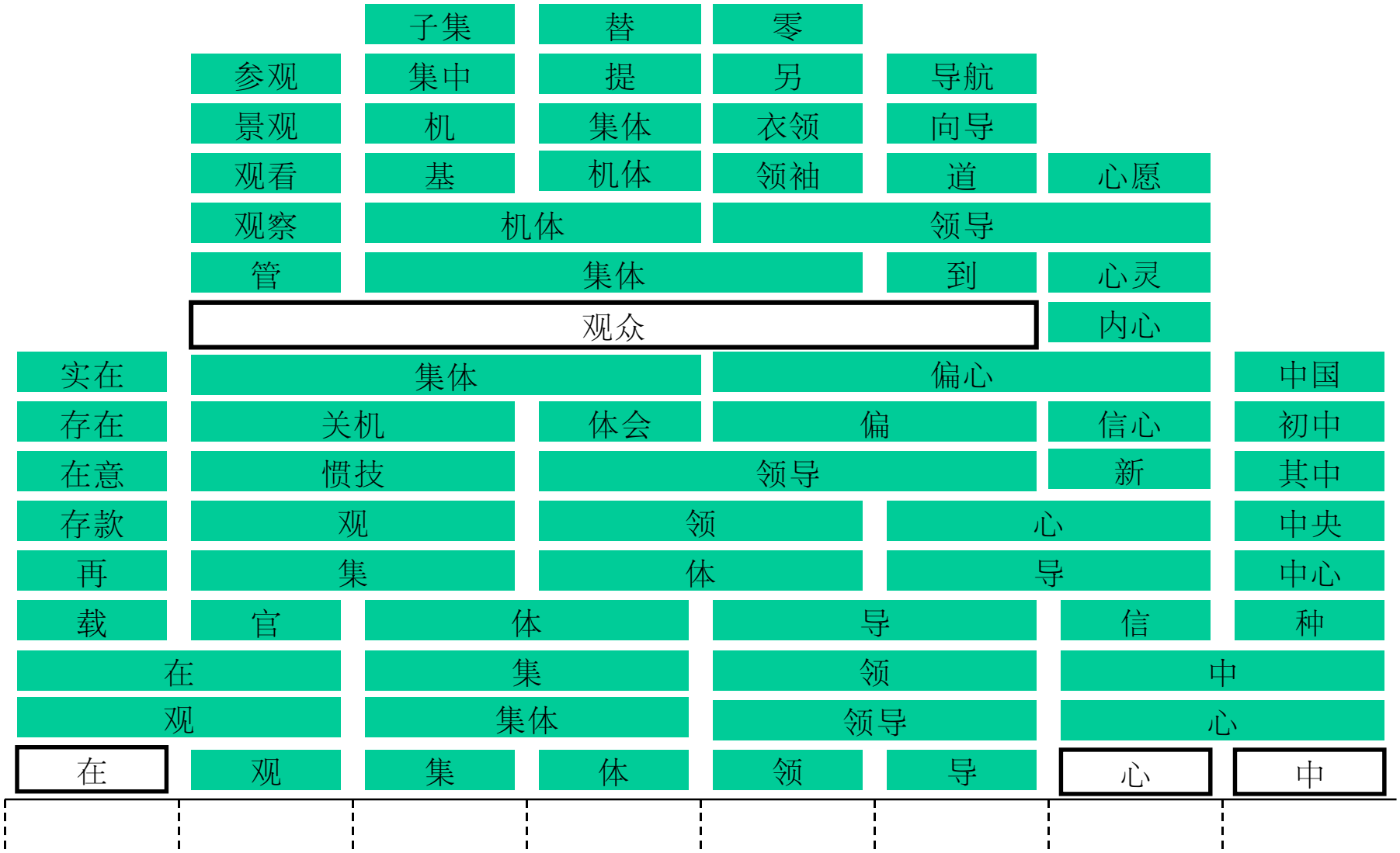
Chinese Spelling Check

Error types in Chinese Text

Microsoft 微软中国研究院

- Non-word error
  - Character substitution “一鸣惊人” 一鸣惊人
  - String substitution “实施求是” 实事求是
  - Character insertion “惊天天动地” 惊天动地
  - Character deletion “忠耿耿” 忠心耿耿
- Real-word error
  - Word substitution “统治” 通知
  - Word insertion “基于基于” 基于
  - Word deletion “一龙” 一条龙

Error Detection/Correction by Fuzzy Matching(*ACL-2000*)



Demo: Chinese Spelling Check

Microsoft®
微软中国研究院

English Reading Wizard

Key Features

Microsoft®
微软中国研究院

- Quick Gisting
  - Automatic abstract
  - Keyword extraction
- Reading Assistant
  - Word/Phrase level translation
  - Optimal translation selection based on context

Microsoft®
微软中国研究院

English Writing Wizard

Motivation

Microsoft®
微软中国研究院

- Biggest difficulty for non-natives
  - Spelling?
  - Grammar?
  - Polish!
- Effective way to overcome the difficulty
  - Spelling checker & grammar checker?
  - Dictionary?
  - Others?
- Research effective method to support English polish

Microsoft®
微软中国研究院

Localization Tool

Microsoft®
微软中国研究院

Localization Tool

- Over $300 million/year localization cost, $100 million/year for translation
- Improve efficiency of MS product localization, expect to realize cost reduction
- Research on new MT methodology
  - Translation knowledge acquisition
  - EBMT & SBMT

Microsoft®
微软中国研究院

Web Dictionary Builder

Web Dictionary

Microsoft 微软中国研究院

- Terms, particularly new terms (English)
  - Words
  - Base noun phrases

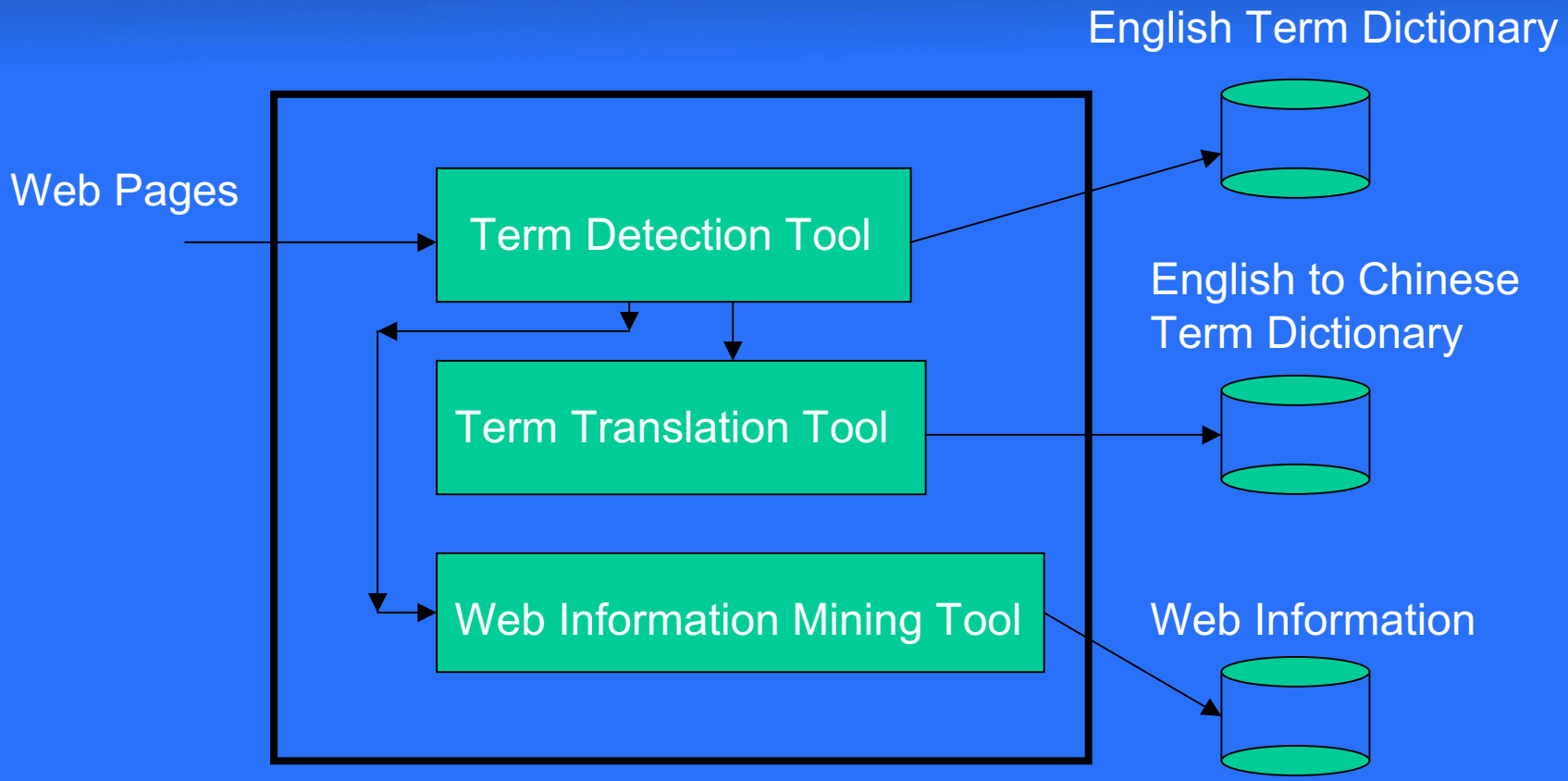
    Eg., information age, information asymmetry
  - Proper nouns

    Eg., WTO, George W Bush
- Translations of terms (English to Chinese)

  Eg., information age --- 信息时代
- Web information on terms
  - Eg., URLs of descriptive web pages

Web Dictionary Builder

Microsoft®
微软中国研究院



Research Achievements

Microsoft®
微软中国研究院

- Tech. Transfer
  - Chinese IME ->Window & Office XP, WinCE
  - English Writing Assistant, Chinese Spelling Check ->Office.NET
- Papers: 23 in FY2001
  - 6 in ACL-2000, 1 in SIGIR'2001, 1 in TREC-9
  - ACM Transaction on Asian Language Information Processing
  - International Journal of Computational Linguistics & Chinese Language Processing
- Patents
  - 10 filed, 10 disclosures

Vision – Overcoming Language/Information Barrier for Asian Users

Microsoft®
微软中国研究院

- Chinese IME
- Spelling check
- English Reading
- English Writing

本文研究的是在复杂背景
面人脸，使用基于样本学习
集人脸图像和非人脸图像作

- Japanese IME
- Spelling check
- English Reading
- English Writing

勝で授与されたばかりの
。しかしそれ以上に目を
我前の肖像と微妙に異な

- Asian IME
- Spelling check
- English Reading
- English Writing

유란시아- 이는 너희 세
는 하나님, 신성(神性), 신
칭으로 부르는 신성한 성격

- Same core technology
  - SLM modeling
  - Powerful NLP/IR/MT toolkits
- Same code base
- Adapt to all Asian language applications

Microsoft®
微软中国研究院

Thanks!

For more information:

mingzhou@microsoft.com