





## Language, Industry and Language Industry

Alex Chengyu Fang  
Director, The Dialogue Systems Group  
Department of Chinese, Translation and Linguistics  
City University of Hong Kong  
acfang@cityu.edu.hk



## Language and Industry

- Language refers to the complex system of symbols, either written or spoken, acquired and used by human for communication.
- Industry refers to the production of an economic good or service within an economy.
- Language represents one of the least obvious associations with the industry.




## Language Industry

- Language industry refers to the sector of activity dedicated to designing, producing and marketing tools, products or services that benefit from machine understanding of human language.
- It is part of information technology that draws on linguistics and software engineering.




## Language Industry

- This presentation intends to show
  - that linguistics, the scientific study of language, has great potentials to demonstrate its social and economic accountability and
  - that, when carefully and appropriately targeted, effective language industry can be fostered to yield a direct bearing on the well-being of the local, regional and international communities.




## At CityU

- The Department of Chinese, Translation and Linguistics represents a critical mass of expertise in the area of applicable language technology.
- In 2008, The Dialogue Systems Group was founded by Dr Alex Fang as a task force devoted to applied research in human speech and language technology.



## The Dialogue Systems Group


- Over the past three years, the DSG has achieved a critical mass of expertise including
  - 4 faculty members
  - 4 full-time researchers
  - 12 affiliated members
    - UK, US, Holland, Korea, Austria
  - 4 PhD students
    - Linguistic research (Chinese, Japanese)
    - Commercial terminology
    - Natural language engineering



## Conventionally


- Access to the Internet or digital databases requires input via the keyboard.
- There are compelling situations where such access is not possible:
  - The visually impaired
  - In-car communication
  - Trading on the move
  - Workplace instruction

7




## Recently


- Internationally concerted efforts are being made to research into voice-enabled access to information:
  - Telephony banking systems (UK)
  - In-vehicle communication (Germany)
  - Speech-to-speech translation (US)



**Siri**  
Use your voice to send messages, set reminders, search for information, and more.




Images from <http://www.apple.com/iphone/>



## Capacity Building

- Language resource construction and annotation
- Knowledge discovery and representation from natural language texts
- Implementation of software systems to disseminate knowledge automatically mined from natural language texts
  - Via keyboard
  - Via voice




## Research Grants

- General Research Fund
  - Intelligent Terminology Management System
- UGC Knowledge Transfer
  - Interactive Delivery of Public Health Information
- CityU
  - Linguistic Foundations of Terminology
  - Linguistic Features of Text Types
  - Communicative Functions and Dialogue Acts



## Proactive Engagement

- To build successful business and industry partnerships, we have proactively engaged with the following:
  - Governmental and non-governmental professional service bodies
  - Local community
  - Businesses and industries



## Government Engagement

- We work closely with governmental and non-governmental professional service bodies, especially in the area of standards setting for language industries.
- This is a vital step to ensure fullest understanding of both policy decisions and needs from businesses.




## Professional Service Bodies

- International Organization for Standardization (ISO)
  - Technical Committee 37: Terminology and other language and content resources
  - Subcommittee 4: Language resource management
- Participation in standards authoring for
  - Linguistic Annotation Framework
  - Semantic Annotation including dialogue act analysis




## Professional Service Bodies

- China National Technical Committee on Terminology for Standardization (全國術語標準化技術委員會)
  - Regional discussions on word segmentation for Chinese, Japanese and Korean
  - Evaluation of national standard for language codes
  - Evaluation of national standard for bibliographical references




## Community Engagement

- We have actively engaged with the local community to identify practical needs and have targeted public health as a potential area of application.
- Support from non-governmental health and anti-drug abusing organizations.



## Prof Shiu Hung Lee visiting the DSG



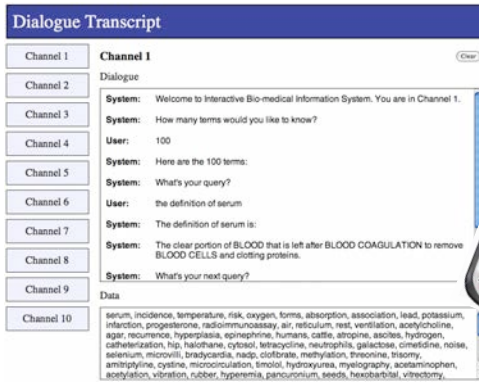


## Information Access

- A man-machine interactive system for delivery of public health information in Hong Kong
  - Access via keyboard
  - Access via voice
  - Intelligent retrieval of health-related information based on terminology as the basic conceptual units



## Demo



**Dialogue Transcript**

Channel 1  
Channel 2  
Channel 3  
Channel 4  
Channel 5  
Channel 6  
Channel 7  
Channel 8  
Channel 9  
Channel 10

**Channel 1**

Dialogue

**System:** Welcome to Interactive Bio-medical Information System. You are in Channel 1.

**System:** How many terms would you like to know?

**User:** 100

**System:** Here are the 100 terms:

**System:** What's your query?

**User:** the definition of serum


**System:** The definition of serum is:

**System:** The clear portion of BLOOD that is left after BLOOD COAGULATION to remove BLOOD CELLS and clotting proteins.

**System:** What's your next query?

Data

serum, incidence, temperature, risk, oxygen, forms, absorption, association, lead, potassium, infarction, progesterone, radioimmunoassay, air, reticulum, nest, ventilation, acetylcholine, agar, recurrence, hyperplasia, epinephrine, humans, cattle, atropine, ascites, hydrogen, catheterization, hip, isotopane, cytosol, tetracycline, neutrophils, galactose, cimetidine, noise, selenium, microvilli, bradycardia, nadp, clotbrate, methylation, threonine, trisomy, amphetamine, cystine, microcirculation, timolol, hydroxyurea, myelography, acetaminophen, acetylation, vibration, rubber, hyperemia, pancuronium, seash, hexobarbital, vitrectomy.




## Industry Engagement

- We have targeted those industrial entities in need of language processing technologies for knowledge and document management.
  - IBM
  - Huawei
  - World Bank
  - International Monetary Fund



## Conclusion

- Applicable research in natural language processing technologies at CityU
- Active engagement with governmental and non-governmental agencies as well as businesses and industries
- Synergy for productive language industry
- Demonstration of social and communal accountability



The banner features a stylized logo on the left consisting of a green and blue swirl. To the right of the logo, the text "KNOWLEDGE EXCHANGE CONFERENCE" is written in a small, sans-serif font. The background of the banner shows a close-up of hands holding a globe, with a blue and white grid pattern visible on the globe's surface.

- Thank you!