A Country Report - COCOSDA Activities in China

O-COCOSDA'2009

Beijing, China

Aijun LI, Thomas Fang ZHENG

Activities – ChineseLDC

- **ChineseLDC sponsored one workshop in 2008**
  - 4th China Workshop on Machine Translation (CWMT2008) in Nov 27-28,2008.

- **New Resources**

□ CLDC-2009-001

- One of the CWMT2008 Machine Translation Evaluation Data, --- Chinese-English MT evaluation data in news domain

1,006 Chinese sentences in discourse style, 41,042 Chinese characters, English is translated, by Institute of Computation, CAS

□ CLDC-2009-002

One of the CWMT2008 Machine Translation Evaluation Data

--- English-Chinese MT evaluation data in news domain, 1000 English sentences, 21,767 words, Chinese is translated, by Institute of Computation, CAS

□ CLDC-2009-003

One of the CWMT2008 Machine Translation Evaluation Data

--- the candidate input data for the “System Combination” task

Activities in Chinese Academy of Social Sciences(CASS)

- Institute of Linguistics, CASS
  - Telephone speech corpus with detailed linguistic and phonetic transcription is being jointly developed by CASS and the Institute of Automation, CAS. 8000 dialogues covering 5 service fields: travel, hotel, airport, restaurant and taxi.
  - Infant and parent speech database with linguistic and phonetic transcription
  - CASS_NOKIA EFL speech corpus extended for more sentence types
  - Physiological database with EMA and Nasal/oral airstream.

- Institute of Ethnology and Anthropology (IEA), CASS (joint by Tibetan Language & Character Committee, Xin Jiang University and Southwest University For Nationalities)
  - Tibetan, Uigur and Yi Speech Acoustic Databases
    - fulfill purposes for single language phonetic research and multi-lingual comparative phonetic research.
    - three Standard (broadcasting) language acoustic parameter databases of Tibetan, Uigur and Yi languages
    - index and software platform

Activities in Broadcast Media Language Branch of National Language Resources Monitoring & Research Center

- A joint institution of Ministry of Education (MOE), the State Administration of Radio Film and Television (SARFT), and Communication University of China (CUC), China.
- Main function is to monitor the language resources in time continuously and dynamically, and to submit the investigating result in order to supply governments with data for establishing language policy and feasible language standard.

  - Broadcast Media Language Monitoring Corpus
    - Research activities base on text data
      - Annual research report: Language Situation in China ,Annual dictionary of Chinese New Words ,New Chinese Lettered Words Dictionary base on the monitoring corpus (in press)
    - Some special activities base on audio or video data: Broadcasting news read speech corpus for Chinese speech synthesis oriented semantic computing model
  - Chinese Mandarin Digital Multi-modal Corpus (CMDMC)
    - It's a dynamic miniature model (or speech museum) with diachronic, opened, cross-media and sharable features. The materials generated by the authoritative speakers (e.g. announcers in radio or TV, actors/actresses in movie or drama) with normality are required samples. The finished corpus has about 2000 hours video or audio files since 1930s with metadata tagging. Furthermore, a management and inquiry system is also available.

Activities in Language Institute of the Central University for Nationalities

Supported by national projects of ‘211‘ and ‘985‘, a several Speech Corpora of Chinese Minority Languages are being collected, which are benefit for intangible culture preservation, information sciences, minority language teaching, Standard Chinese teaching and linguistic research. The corpora include:

- Speech Corpus of Endangered Ethnic Languages
- The Interlanguage Corpus of Tibetan Speaking Mandarin Chinese
- Phonetic File of Ethnic Languages in China
- Multimedia Archive in International Phonetic Alphabet of Ethnic Languages in China

Activities in the Institute of Psychology,CAS

- Discourse corpus
  - with prosody, syntax and information structure annotation.
  - More than 100 texts have been annotated.