An Introduction to Corpora Resources of 863 Program for Chinese Language Processing and Human-Machine Interaction

QIAN YueLiang LIN ShouXun ZHANG YongDong

LIU Yang LIU Hong LIU Qun

Institute of Computing Technology of Chinese Academy of Sciences
No. 6 Kexueyuan Sourth Rd. Zhongguancun, Haidian, China 100080
{ylqian,sxlin,zhyd,yliu,hliu,liuqun}@ict.ac.cn

Abstract

Rapid progress has been made in many computer-based linguistic technologies in past several years. It is important to build large-scale linguistic resources to benefit most interested researchers. This paper introduces Corpora Resources of 863 Program for Chinese Language Processing and Human-Machine Interaction (i.e. Corpora Resources of 863 Program), which is now hosted by Institute of Computing Technology of Chinese Academy of Sciences (ICT). Corpora Resources of 863 Program is an extensive linguistic database, in the form of time-series data such as text, audio, image and video. We will also address issues such as its application, sharing mechanism and our future work.

1 Introduction

Rapid progress has been made in many computer-based linguistic technologies in past several years, including machine translation, text retrieval and understanding, text categorization, speech recognition, etc. However, because human language is so complex and information-rich, computer programs for processing it must be fed with enormous amounts of varied linguistic data, especially as corpus-based, stochastic, and learning approaches are introduced into natural language processing research. Such data are expensive to create and document, with maintenance and distribution adding additional costs. As a result, researchers at smaller companies and in universities may not be able to afford to make use of such valuable resources. Thus, most linguistic resources were not generally available for use by interested researchers.

In 1992, the Linguistic Data Consortium (LDC) was founded to provide a new mechanism for large-scale development and widespread sharing of resources for research in linguistic technologies. Based at the University of Pennsylvania, the LDC is a broadly-based consortium that now includes more than 100 companies, universities, and government agencies. Since its foundation, the LDC has delivered data to 197 member institutions and 458 non-member institutions (excluding those who have received data as a non-member and later joined).

However, LDC has not provided enough Chinese linguistic data. Several parallel or related efforts are underway in China.

Many universities and institutes have built an amount of Chinese linguistic corpora, but most of them have not been widely used partly due to not providing an effective sharing mechanism.

Corpora Resources of 863 Program is an extensive linguistic database. Under a grant from The National High Technology Research and Development Program (863 Program), ICT took part in constructing Corpora Resources of 863 Program with other universities, companies and institutes.

Section 2 is about the overview of Corpora Resources of 863 Program. Section 3 describes every corpus of Corpora Resources of 863 Program in detail. Section 4 is the application of Corpora Resources of 863 Program. Section 5 discusses the sharing mechanism. Section 6 is the conclusion and what future work we anticipate.

2 Overview

Human language resources, expensive to create and maintain, are in increasing demand among a growing number of research communities. To accelerate the development of Chinese language processing technology, under a grant from 863 Program, Institute of Computing Technology of Chinese Academy of Sciences took part in building Corpora Resources of 863 Program together with Institute of Automation of Chinese Academy of Sciences, Tsinghua University, Peking University, Beijing HanWang Technology Corporation, Anhui USTC iFLYTEK Corporation, Graduate School of the Chinese Academy of Sciences and Institute of Linguistics of Chinese Academy of Social Sciences.

Corpora Resources of 863 Program is an ongoing project, with its objective being to set up a linguistic database for Chinese Information Processing and Intelligent Human-Machine Interaction technology. It is an open and non-profitable database, aiming to benefit numerous researchers in this area, meanwhile to push forward the Chinese linguistic data processing technology.

In 1990, Chinese Online Handwriting Database was constructed, including 500 million Chinese characters. In 1996, Continuous Chinese Speech Corpus was completed. Now, Corpora Resource of 863 Program consists of 12 databases. It is divided into four major categories according to the type of data they contain, and then are further broken down into minor categories based on the source of data.

The major categories are text, audio, video and image.

In the following section, we will describe each corpus and database in detail.

3 Corpora

Corpora Resources of 863 Program is divided into four major categories according to the type of data they contain. These major categories are text, audio, video and image. The volume of Corpora Resources of 863 Program is over 200 GB.

3.1 Text

Currently, there is only one published corpus in the category of text. Some other text corpus, such as bilingual dictionaries, Chinese proper name dictionaries, and etc. are still under construction.

Chinese-English Parallel Corpus was constructed conforming to corresponding technique standards. The materials are in the form of dialogs and essays that sampled from several websites. The topics are mainly sports reports, weather forecasting, traffic, travel and living conditions. The corpus now

contains 200,000 Chinese-English aligned sentences, 50,000 Chinese-English aligned words and phrases.

Many corpora are still under construction, and one of them is Chinese-Japanese Parallel Corpus. Until now, we have completed 20,000 Chinese-Japanese aligned sentences.

3.2 Audio

The category of audio is composed of six corpora:

- Chinese Speech Synthesis Corpus
- Prosody Analysis Corpus
- Dialectical Mandarin-Chinese Corpus
- Chinese and English Mixed Reading Corpus
- Chinese Speech Recognition Corpus
- Chinese Telephone Speech Recognition Corpus

Chinese Speech Synthesis Corpus

Chinese Speech Synthesis Corpus is composed of four databases.

- **Mandarin-Chinese TTS System Database**

We invited a man and a woman[1] to read prepared materials in Mandarin-Chinese and recorded the whole process. The database may benefit text-to-speech (TTS) research. The detail is shown in Table 1.

	Man	Woman
Sentences	4,491	6,046
Recorded	1,007MB	1,350MB
Annotated	20MB	42MB

Table1: Detail of Mandarin-Chinese TTS System Database

- **Mandarin-Chinese TTS System Test Database**

The database is somewhat alike the above database, except that it is for the purpose of test. The detail is shown in Table 2.

	Man	Woman
Sentences	1,000	1,000
Recorded	250MB	272MB
Annotated	14MB	14MB

Table2: Detail of Mandarin-Chinese TTS Test System Database

- **Mandarin-Chinese Intonation Analysis Database**

Only one female volunteer was asked to read prepared texts in Mandarin-Chinese, varying in mood and intonation. The text is composed of 1,000 sentences. The volume of recorded data is 286MB, and annotated data is 15MB.

- **Continuous Prosody Essay Database**

A man and a woman read prepared materials in both Mandarin-Chinese and dialect. The database contains about 4 hours of news releases in consideration of various topics. The volume of recorded data is 3.2GB, and annotated data is 2MB.

Dialectical Mandarin-Chinese Corpus

We chose Shanghai, Xiamen, Guangzhou and Chongqing as the first sampling area. In each city we invited 200 volunteers, totally 800. Considering the correlation between age and education background, the age distribution is shown in Table 3.

Age	Percentage
[18, 25]	15%
(25, 45]	70%
(45, 55]	15%

Table3: Age distribution in Dialectical Mandarin-Chinese Corpus

Each volunteer spent 1 or 2 hours in reading prepared materials: 220 short paragraphs (each paragraph has less than 50 Chinese characters), 160 colloquial dialogs and some popular dialectical phrases. Until

[1] All of the volunteers are ordinary people, who are not professional narrators.

now, we have annotated 40 hours of data. Its volume is about 100 GB. The Corpus is still under construction.

Prosody Analysis Database

The database contains about 17 hours of data: 549 paragraphs and 4,693 sentences, in which 596 sentences are interrogative, exclamatory and imperative ones. Its volume is 10GB.

Otherwise, the database also includes about 30 minutes of CCTV news releases read by 12 male and 7 female newscasters. The volume of this part of data is 52MB.

Chinese and English Mixed Reading Corpus

The reading materials are mainly in Chinese, mixed with English abbreviations, single English word and phrases. We have designed 3,000 sentences, including 1,000 English sentences, 600 Chinese sentences and 1,400 Chinese and English mixed sentences.

The data was sampled in 16 KHz and saved as 16 bit PCM.

Until Now, we have completed man-reading data construction and annotated some of the data. Its volume is about 450MB. Woman-reading data is still under construction.

Chinese Speech Recognition Corpus

The volume of Chinese Speech Recognition Corpus is 17GB. It consists of eight corpora:

- Southern Dialect Speech Corpus
- Northern Dialect Speech Corpus
- Chinese Dialog Speech Corpus
- English Dialog Speech Corpus
- Embedded Circumstance Speech Corpus
- Kiosk Channel Self-Adaptive Speech Corpus
- Olympic-related English Speech Corpus
- Olympic-related Japanese Speech Corpus

The statistics of the eight corpora is shown in Table 4. The first column is the sequence number of each corpus above. For instance, NO 2 refers to Northern Dialect Speech Corpus.

NO	materials	volume	length
1	100,000 sentences	8.2GB	70 hours
2	75,000 sentences	5.8GB	59 hours
3	10,000 sentences	1.1GB	8.4 hours
4	10,000 sentences	1GB	8 hours
5	2,331 words	71.1 MB	1.3 hours
6	1,000 words	46MB	40 minutes
7	15,000 sentences	900MB	7 hours
8	2,000 sentences	400MB	3 hours

Table4: The statistics of the eight corpora

Telephone Speech Recognition Corpus

The corpus is a collection of two-sided telephone conversations among 700 Chinese speakers sampled legally from Beijing local Call Records and long distance call to 8 other cities. The data were sampled in 8 KHz and saved as 16 bit PCM. The volunteers were from 28 cities, and their age ranged from 14 years old to 84. The telephone conversations are about 10 numbers, 5 number strings, 20 popular phrases, 5 expressions about currency format, 6 expressions about time, 8 questions and 20 sentences.

The English and Japanese counterparts are being developed.

3.3 Video

The category of video consists of three

corpora and databases:

- Chinese Hearing and Vision Bi-Model Corpus
- Human Motion Video Database
- Continuous Sign Language Database

3.4 Image

The category of image is composed of two databases:

- Face Recognition Database
- Chinese Online Handwriting Database

Chinese Online Handwriting Corpus has been completed, collecting all Chinese characters in GB18030 Character Set (27484 Chinese characters) excluding those also in GB2312 Character Set (6763 Chinese characters). The corpus contains over 2,000,000 handwriting Chinese characters written by 100 persons. Figure 1 shows samples in the corpus.

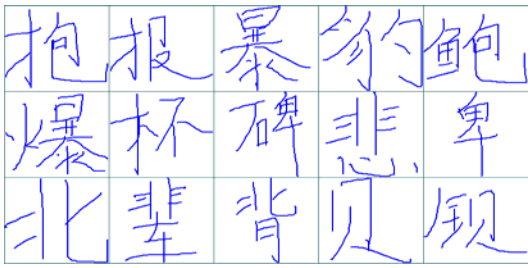

Figure1: Samples in Chinese Online Handwriting Corpus

4 Application

Corpora Resources of 863 Program is an extensive linguistic database, so researchers interested in various areas will benefit from it. Actually, the audio resources in Corpora Resources of 863 Program have been purchased by many companies and individual researchers.

It is the diversity of Corpora Resources of 863 Program that satisfy the research and development needs of researchers in various areas including machine translation, text retrieval and understanding, text categorization, speech recognition, speech synthesis, text-to-speech, face detection, handwriting characters recognition, etc.

Another important application is that we can provide evaluation based on such abundant linguistic resources. In 2003, launched by Expert Committee of Software and Hardware Technology of 863 Program, ICT organized the 863 evaluation '2003 on Chinese Information Processing and Intelligent Human-Machine Interface. The evaluated technologies include Chinese Word Segment and POS Tagging, Information Retrieval, Text Classification, Text Summarization, Machine Translation, Speech Recognition, Speech Synthesis and Chinese Online Handwriting Recognition. 46 systems have been evaluated. [2] The evaluation was successful and had an impact on the development of mentioned technologies.

Corpora Resources of 863 Program has also been used to support Information Infrastructure Construction for Olympic Games that will be held in Beijing in 2008.

5 Sharing Mechanism

One of the aims of Corpora Resources of 863 Program is to unite numerous researchers in various areas and then to establish a universally accepted Chinese linguistic database including not only linguistic data.

We believe that shared resources provide benefits that closely-held or proprietary resources do not.

Although formal sharing mechanism of 863 Corpora Resources has not been worked out, here are some points. First, we prefer to divide users into members and non-members. Members are entitled to one copy of each corpus released in their years

[2] More details will be found at http://www.863.org.cn

of membership only after paying an amount of fees. Non-members can get access to non-free corpus only after purchasing it. Second, we are welcome to corrections and additions provided by individual users. We would like negotiate to exchange corpus with other corpus providers. Actually, we maintain relations with other groups around the world who gather and/or distribute linguistic data. One of our collaborators is Chinese Linguistic Data Consortium (ChineseLDC), which was founded in 2002. Third, every corpus or database in 863 Corpora Resources is copyrighted by original constructors and ICT only manages and distributes all the resources.

6 Conclusion

Rapid progress has been made in many computer-based linguistic technologies in past several years, including speech recognition and understanding, optical and pen-based character recognition, text retrieval and understanding, machine translation, and the use of these methodologies in computer assisted language acquisition. As a result, it is important to build and share large-scale linguistic resources.

Under a grant from The National High Technology Research and Development Program (863 Program), ICT took part in constructing Corpora Resources of 863 Program with other universities, companies and institutes.

Corpora Resources of 863 Program is an extensive linguistic resources database, and may benefit many computer-based linguistic technologies.

In the future, we plan to complete the resources that still in progress and add new resources to the corpora. Existed resources will be refined, making it more robust and effective. We will also seek international cooperation and negotiate to exchange resources with other corpus providers. Corpora Resources of 863 Program will keep on developing to adapt to the needs of that research community and international developments in technology research.

Acknowledgement

The ICT authors are supported by National High Technology Research and Development Program contract “Generally Technical Research and Basic Database Establishment of Chinese Platform” (Subject No. 2001AA114010). We are grateful to Expert Committee of Software and Hardware Technology of 863 Program, our colleagues in Digital Technology Laboratory of ICT and all of our collaborators.

References

Christopher Cieri and Mark Liberman. 2000. *Issues in Corpus Creation and Distribution: the Evolution of Linguistic Data Consortium.* Proceedings of the Second International Language Resources and Evaluation Conference. Athens, Greece.

David Graff and Steven Bird 2000 *Many uses, many annotations for large speech corpora: Switchboard and TDT as case studies.* 2nd Language Resources and Evaluation Conference (LREC 2000) Athens, Greece, May 2000

Mark Liberman and Christopher Cieri. 1998. *The Creation, Distribution and Use of Linguistic Data.* Proceedings of the First International Conference on Language Resources and evaluation. Granada, Spain.

Steve Cassidy and Steven Bird. 2000. *Querying Databases of Annotated Speech.*

Proceedings of the Eleventh Australasian Database Conference.

http://www.ldc.upenn.edu

http://www.863data.org.cn

http://www.chinesldc.org