Advances in Chinese Natural Language Processing and Language resources

Article · August 2009

DOI: 10.1109/ICSDA.2009.5278384

ADVANCES IN CHINESE NATURAL LANGUAGE PROCESSING AND LANGUAGE RESOURCES

[(1)]Jianhua Tao [(2)]Fang Zheng [(3)]Aijun Li [(4)]Ya Li
[(1)(4)] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing
[(2)] Tsinghua University, Beijing
[(3)] Institute of Linguistics, Chinese Academy of Social Sciences, Beijing
[(1)]jhtao@nlpr.ia.ac.cn [(2)]fzheng@tsinghua.edu.cn [(3)]liaj@cass.org.cn [(4)]yli@nlpr.ia.ac.cn

ABSTRACT

In the past few years, there have been a significant number of activities in the area of Chinese Natural Language Processing (CNLP) including the language resource construction and assessment. This paper summarized the major tasks and key technologies in Natural Language Processing (NLP), which encompasses both text processing and speech processing by extension. The Chinese Language resources, including linguistic data, speech data, evaluation data and language toolkits which are elaborately constructed for CNLP related fields and some language resource consortiums are also introduced in this paper. Aimed to promote the development of corpus-based technologies, many resource consortiums commit themselves to collect, create and distribute many kinds of resources. The goal of these organizations is to set up a universal and well accepted Chinese resources database so that to push forward the CNLP.

Index Terms— Chinese Natural Language Processing, Language Resource, Resource Consortium

1. INTRODUCTION

The research of language processing must be based on real language data and work on a large quantity of detailed materials. Only then more reliable works can be drawn. Most modern NLP are at least partly statistical. This approach allows the system to gather information about the frequency with which various constructions occur in specific contexts. During these years's development, CNLP has made great achievements and entered into a new fast developing era; though there are still many unresolved challenges in language processing and speech processing, i.e. deep structure analysis of syntax, phonology, speech, translation and Semantic etc.

To gain these achievements, the various well-constructed corpora are indispensable. Various kinds of high-quality corpora for speech recognition, speaker identification, voiceprint recognition, speech synthesis, machine translation and information retrieval, text classification, automatic summary etc. have been built during these ten years. Most of these resources are no longer for individual closely-held, but shared with other institutes or individuals who obey the authorization or license.

Resources sharing can save much duplication of effort. Considering these issues, many resource consortiums which provide corpora for CNLP have been set up. The goal of these consortiums is to set up a universal Chinese language database so that to enhance CNLP to an international level, by offering assistance in relevant fundamental research and the development of application, meanwhile to push forward the research on Chinese information processing.

The remainder of this paper is organized as follows. Section 2 summarized the basic processing technologies in NLP and the supporting resources, as well as CNLP toolkits. In section 3, some typical Chinese resource consortiums are introduced. Section 4 presents the conclusion and future work.

2. CHINESE NATURAL LANGUAGE PROCESSING

Techniques of automatic CNLP have been under development since the earliest computing machines, and in recent years these techniques have proven to be robust, reliable and efficient enough to lead to commercial products in many areas. To gain these achievements, the various well-constructed corpora are indispensable. There has been lots of progress in the fundamental resource construction during last ten years. By now the Chinese language recourse are as many as several hundred, covering linguistic data, speech data, evaluation data and language toolkits.

2.1. Language Processing

Language Processing is focused on developing efficient algorithms to process texts and to make their information accessible to computer applications. The processing scope can contain information at many different granularities, from simple word or token-based representations, to rich hierarchical syntactic representations, to high-level logical representations across document collections. The subproblems can divided into Word segmentation, POS tagging, Word sense disambiguation, Syntactic Parsing, Syntactic disambiguation, Text alignment, Phrase Extraction,

978-1-4244-4400-7/09/$25.00 ©2009 IEEE

Sentence Generation, Machine Translation, Information Extraction, Information retrieval, Question and Answering and Dialogue Systems etc. The fundamental processing, morphological analysis is relatively mature, 98% precision for word segmentation, and 95% for POS tagging in [1],whereas higher level processing related to syntactic and semantics parsing are still a great challenge. Natural language are not easily parsed by programs, as there is substantial ambiguity in the structure of human language, whose usage is to convey meaning (or semantics) amongst a potentially unlimited range of possibilities but only some of which are germane to the particular case.

At each processing level and processing purpose, specific corpora are needed to support the research.

Chinese Lexicon

Lexicon is the very basic resource in NLP. There are several lexicons for different use.

Chinese Lexicon for common use, i.e. word segmentation and POS tagging includes carefully selected Chinese word items, accompanied with information of POS, frequency and PINYIN, typically contain millions of items. Lexicon for word segmentation usually only contains word items, which is easy to construct and the scales can be great varied.

Chinese geographic name, corporation name and name lexicon is used in Name Entity Recognition. The Chinese Geographic Name Storehouse contains 470,000 city and town names in all the provinces of China. [2].

Chinese Web 5-gram Corpus contains all the Chinese words observed frequency data from unigram to 5-gram. The words it covered are about 883,000,000,000 in more than 102,000,000,000 sentences from accessible web pages before March, 2008 [3].

Chinese Pos tagged corpus

6-month news corpus with POS tagged from the People's Daily according to word segmentation and POS tagging norm is established by Peking University. It contains common Part-Of-Speech tags, special usage tags of verbs and adjectives, proper noun, place name of phrase type, organization name of phrase type and so on. The scale of the corpus is about 27 millions Chinese characters [4].

```
19970310-01-001-0020/m  新华社/nt  北京/ns
３月/t  ９日/t  电/n  [中共中央/nt  办公厅
/n]nt  近日/t  发出/v  通知/n  ，/w  要求/v
各级/r  党委/n  组织/v  干部/n  群众/n  认真
/ad  学习/v  悼念/v  邓/nr  小平/nr  同志/n
的/u  重要/a  文献/n  。/w
```

Figure 1 A sample in People's Daily corpus (CPC Central Committee General Office recently issued inform, which demands the masses of cadres and Party committees at all levels to seriously study the important literatures about mourning for Deng Xiaoping, reported by Xinhua News Agency, Beijing, 9th, March)

College of Computer and Information Technology, Shanxi University also constructs a POS tagged corpus with size of 5,000,000 Chinese Characters.

Multilingual Corpus

Chinese-English Olympic Dictionary consists of two parts: Sports Dictionary and Travel Dictionary, covering more than 60,000 bilingual terms. Each term is POS tagged [5].

Multilingual corpus is very useful for machine translation. There are bilingual and multilingual, paralleled and unparallel corpora. Several languages: Chinese, English, and Japanese are included. All the parallel sentences have been manual checked.

Syntactic tagged corpus

Tsinghua Chinese Treebank contains 44, 600 sentences, covering about 1,000,000 Chinese words from balanced literatures and each sentence is segmented and POS annotated. Complete parsing trees for each sentence are constructed and can be used to develop different Chinese parsers, i.e. Chinese multiword chunk parser, Chinese functional chunk parser, Chinese dependency parser, Chinese event parser and Chinese discourse relation parser [6].

Modern Chinese semantic Dictionary focus on semantic information based on intentional logical model, and several example sentences are given [7].

Linguistic resources for text classification, information retrieval and automatic summary etc. are also numerous. They contain several style of document, including argumentation, essay, fiction and narration.

2.2. Speech Processing

Speech processing technology is closely tied to natural language processing and digital signal processing. Speech processing can be divided into the following categories: Speech recognition, Speaker recognition, Speech coding, Voice analysis Speech synthesis, Speech enhancement etc. Speech technology is becoming more and more popular because speech is the most simplified and direct way to communication in daily life. While the speech technology application demands are numerous, there are some bottlenecks still need to be addressed in the following years. The first challenge is robustness in recognition part. Different environments, channels and speakers greatly impact the recognition results. In synthesis part, the naturalness is key problems, which refers to the prediction and generation of the language prosody. The latter challenge is closely related to NLP for the deficiency of deeper understanding of the language.

As a basic and important part in speech technology, more and more attention has been paid to the design of the speech corpus.

Natural Broadcasting Speech corpus

CASIA Mandarin corpus has been carefully recorded by a professional female speaker under studio conditions. The corpus contains 5000 phonetic context balanced sentences with about 7 hours. The text transcription with word boundaries, POS tags and pronunciation are also involved [8].

```
4. |法国 人民|深深 铭记着|将军$对 法兰西 民族的|
丰功 伟绩|.|
  fa3 guo2 ren2 min2 shen1 shen1 ming2 ji4
zhe5 jiang1 jun1 dui4 fa3 lan2 xi1 min2
zu2 de5 feng1 gong1 wei3 ji4
```

Figure 2 A sample in CASIA Mandarin corpus (The French people treasure up the great achievements toward France nation contributed by the general deeply)

ASCCD [9] is a read discourses speech, 18 texts with 300-500 syllables each. The speech was uttered by 5 female and 5 male speakers and recorded in two channels: speech waveform and the glottal impedance waveform through Laryngograph Segmental and prosodic annotations including canonical Pinyin and tone tier, initial / final tier of real pronunciation, sentence mode tier, and stress tier.

Multi-language corpus

Special scene and Special domain dialogue corpus can be separated as four domains: catering, transport, sports, weather. Each domain contains 100 sentences. Male and female voices are both recorded [10].

Chinese Hotel Reservation Dialogue is a multi-lingual hotel reservation dialogue corpus, which presently consists of Japanese, English and Chinese speech. The corpus includes speech by 50 speakers, 25 males and 25 females. Each speaker reads four dialogue sides, and a number of common language-parallel sentences. The total number of utterances is about 4,500, and the speech lasts about 4 hours [11].

Voiceprint Recognition

Corpus for Voiceprint Recognition contains speech from 10,000 male speakers aged 18-23. All utterances were required to be made twice, speaking clearly and naturally without any attempt to disguise the voice. For each speaker, the first time the utterance was recorded through a GSM mobile phone and the second time it was recorded through a landline telephone [12].

Dialect Speech and Emotion Speech

Regional accent is much different from Chinese mandarin. Several typical regional accent speeches of Changsha, Luoyang, Nanchang, Nanjing, Taiyuan, Wenzhou, Chongqing, Shanghai, Guangzhou and Xiamen, Tianjin and Sichuan accented Mandarin are presently recorded and other dialects are under construction.

Emotion corpus was designed for affect computing related tasks. Chinese Affective Speech Recognition corpus contains 50 records from 50 people (25 male, 25 female). Every emotional speech is consist of the affective reading of paragraphs and commands, where affects are listed as neutral, happiness, fear, angry and sad [13].

Articulatory databases

In order to record the signal more precisely, EMA (AG500 system) and the EPG are used to construct the speech database. 20 speakers are being recorded with the speaker’s tongue (three pints), jaw, lower lip and upper lip kinematics. Four speakers were recorded wearing customized artificial palates. The speech material is composed of segments, tonal syllables, phonetically balanced disyllabic words, phonetically [14].

Despite the corpora above, Telephone Speech Corpus, Dialogue and Spontaneous Conversation Corpus etc. are also available. Generally the corpus is uttered by different speakers of different ages and education background, recorded over fixed telephone network or in professional recording studio with professional recording equipments.

2.3. Evaluation Data

The goal of NLP evaluation is to measure one or more qualities of an algorithm or a system, in order to determine whether (or to what extent) the system answers the goals of its designers, or meets the needs of its users. Research in NLP evaluation has received considerable attention, because the definition of proper evaluation criteria is one way to specify precisely an NLP problem, going thus beyond the vagueness of tasks.

Evaluation data is elaborately designed for certain evaluation points. All the evaluation data contain three parts: test data, reference and evaluation tool. National 863 program (Chinese Hi-Tech Program), 973 program (National Key Foundation Research Program) and Chinese Information Processing Society of China (CIPSC) have carried out various types of evaluations.

Speech synthesis evaluation data

863 assessment for speech synthesis covers two evaluation aspects: understandability and naturalness. SUS (Semantic

Unpredictable Sentence) sentences are designed for understandability. Short papers are collected for naturalness test.

Speech recognition evaluation data

863 assessment for speech recognition evaluation data is usually composed of three parts: Desktop speech, Telephone speech and PDA speech. It was recorded in real environment with noise and covers single sentences and command words.

Part-of-speech evaluation data

Part-of-speech evaluation data contains 400,000 characters, includes the political, economic, sports, transportation, tourism, education and other aspects of the theme from books, newspapers, magazines and web pages after 1980s. All the segmentation and postaggting ambiguous are manual checked. Name Entity is also tagged in the corpus.

Name Entity recognition evaluation data

The corpus is composed of texts in both Chinese simplified and traditional types, of which 241 simplified documents (about 40 million words), traditional 126 documents (about 40 million words).

Machine translation evaluation data

The machine translation evaluation data covers several languages, including Chinese, English, Japanese and French. A certain amount reference translations which are manual checked are available for both conversation and discourse.

Despite the evaluation data above, evaluation data for information retrieval, automatic index, text classification, full text retrieval etc. are also included. They contain several styles of documents, including argumentation, essay, fiction and narration. Some references are manually checked and the others are automatic generated based pooling.

2.4. NLP Toolkit

In the field of NLP, linguistic resources and some fundamental processing are both important, such as Word Segmentation, POS tagging and syntactic parsing etc. Considering the extensive usage of these basic technologies, some research institutes have opened their research achievements in the hope of reaching a win-win outcome, as users can freely use their work, and they can get bug report or other valuable feedbacks from users.

ICTCLAS and LTP are two widely used such NLP toolkits.

ICTCLAS

The most famous resource in CNLPP [15] is Chinese Lexical Analyzer ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System). The main components are word segmentation, POS tagging, Named Entity Recognition and New Word Recognition. More than 30,000 people or institutes have downloaded ICTCLAS and go on with their works based on ICTCLAS [16].

ICTCLAS uses Cascaded Hidden Markov Model as a unified framework, which combine the entire processing step to achieve an excellent overall performance.

It is reported that the in ICTCLAS 3.0 segmentation speed is 996KB/s and precision is over 98%, whereas the overall Application Programming Interface (API) is no more than 200KB and the compressed lexicons is less than 3M [1].

LTP

Language Technology Platform (LTP) [17] is shared by Information Retrieval Laboratory of Computer Science and Technology School, Harbin Institute of Technology. It is a uniformed language processing system based on XML presentation. All operations are done in DOM. The processing technologies cover Lexical Analysis, Part-of-speech Tagging, Named Entity Recognition, Dependency Parsing, Graph-based Dependency Parsing, Word Sense Disambiguation, Shallow Semantics Labeling, Semantic role labeling and Language Technology Markup Language etc. Until December, 2008, more than 260 institutes shared LTP for research freely in two years.

Other NLP related open toolkits are quite a few. These valuable works greatly cut down the duplication of effort and can be used for further development, which can greatly promote the research level of NLP.

3. LANGUAGE RESOURCE CONSORTIUM

The Linguistic Data Consortium (LDC) [18] is an open consortium which creates, collects and distributes language resources for research and development purposes. It provides a new mechanism in terms of language resources construction and sharing, which greatly promotes the development of related fields. Resources sharing can save much duplication of effort and permit replication of published results, support fair comparison of alternative algorithms or systems, even permit the research community to benefit from corrections and additions provided by individual users. Some Chinese language resource consortiums have been set up in recent years and have already made remarkable progress. The recourses they provided can be speech- or text-based; read or spontaneous; wideband or narrowband; standard or dialectal Chinese;

clean or with noise; or of any other kinds which are deemed helpful for the related research.

3.1. Chinese Corpus Consortium (CCC)

Chinese Corpus Consortium (CCC) is a non-profit, academic consortium sponsored by Dr. Thomas Fang Zheng and Co-Founded by 8 international companies and scientific research institutes in 2002 [19].

Now CCC has more than 40 resources, which covers corpus for ASR, TTS, VPR, and Emotion Computing etc. All corpora contain rich information for data description, such as:

Table 1 Standard Data Description in CCC

Name of the corpus
IPR Holder
Corpus type: Speech or Text
If it is a speech corpus:
Purpose
Language
Style
Channel
Sampling rate
Sampling precision
Corpus size
SNR level
Transcriptions
If it is a text corpus
Language
Domain
Corpus size
Tag information
A brief description of the corpus

3.2. Chinese Linguistic Data Consortium (CLDC)

Chinese Linguistic Data Consortium (CLDC) is another academic and non-profitable resource consortium in China, initialed by Chinese Information Processing Society of China (CIPSC) in 2004 [20].

The resources collected by CLDC cover various research fields in Chinese information processing, such as word segmentation, POS tagging, syntactic/semantic parsing, translation, speech synthesis, speech recognition, dialog speech, etc. Until now, CLDC has collected more than 100 kinds of resources which include more than 40 kinds of data supported by National 863 program and 973 program, and hundreds of universities, institutes and companies among Asia, Europe, America are using data from CLDC.

3.3. Chinese Natural Language Processing Platform

Chinese Natural Language Processing Platform (CNLPP) [15] is aimed to create an entirely open environment. The recourses in CNLPP are free to download and not limited in text or speech corpora, but also include related source code, demos, books, papers etc.

Apart from the consortiums above, Institute of Computational Linguistics, Peking University (ICLPKU) [21], Institute of Linguistics, Chinese Academic of Social Sciences (ILCASS), The Institute of Ethnology and Anthropology, Chinese Academic of Social Sciences (IEACASS), Tsinghua University, Harbin Institute of Technology, Xiamen University [22] etc. have carried out various resource sharing activities. Meanwhile, because of the great concerned of Chinese all over world, some data sharing consortiums overseas have increased the Chinese language corpus construction and sharing, such as LDC, The CJK Dictionary Institute [23] etc.

4. CONCLUSION AND FUTURE

In the past few years, great achievements have been made in CNLP, which can be proved by many reliable and efficient commercial products. These remarkable achievements in CNLP are due to the progress in Chinese language resources collection and data sharing. By now the Chinese language recourse are as many as several hundred, covering linguistic data, speech data, evaluation data and language toolkits.

On the other hand, Considering the importance of corpus in information processing, many consortiums endeavored to collect corpus for sharing in order to set up a universal and well-accepted Chinese linguistic database so that to push forward the Chinese NLP. They offer a good service for data users and try to promote the researches for both of them. With the support of linguistic resource, related evaluation programs have also gained rapid development.

Future works should be focus on the followings:

Create resources more related to Chinese characteristics, such as prosodic labeled corpus.

Create resources with more reference to linguistics.

Create more large-scale and well-accepted corpus, such as HowNet etc.

5. REFERENCES

[1] http://ictclas.org/

[2] http://192.168.88.2/resourse.asp?name=中国地名机构名库&alone=1

[3] http://192.168.88.2/resourse.asp?name=中文互联网 5-gram 语料库&alone=1

[4] Shiwen YU, "Foundational Language Resource Data on Web Sites", *Terminology Standardization and Information Technology* 2001 (4).

[5] http://192.168.88.2/resourse.asp?name=奥运领域汉英词库&alone=1

[6] qiang ZHOU, “Annotation Scheme for Chinese Treebank”, *Journal of Chinese Information Processing* 18(004) pp 1-8, 2004.

[7] http://192.168.88.2/resourse.asp?name=现代汉语内涵逻辑语义词典&alone=1

[8] Jianhua Tao, Fangzhou Liu, Meng Zhang, Huibin Jia, “Design of speech corpus for mandarin text to speech”, *The Blizzard Challenge 2008 workshop*, Oct, 2008

[9] Li Aijun, Lin Maocan, ChenXiaoxia, et.al., “Speech corpus of Chinese discourse and the phonetic research”, *ICSLP*’2000.

[10] QIAN Yueliang et al. “Design and Construction of HTRDP Corpora Resources for Chinese Language Processing and Intelligent Human-Machine Interaction”, *Chinese High Technology Letters* 15(001): 107-110, 2005.

[11] http://www.d-ear.com/CCC/corpora/2003-CHRD.pdf

[12] Thomas Fang Zheng, “The Voiceprint Recognition Activities over China - Standardization and Resources,” *Oriental COCOSDA* 2005, pp. 54-58, December 6-8, 2005, Jakarta, Indonesia

[13] http://www.d-ear.com/CCC/corpora/2006-CCC_AC2006_ASR.pdf

[14] Aijun LI and Fang ZHENG,, “O-COCOSD Activities in China, Country Report”, *COCOSDA*, Hanoi, Vietnam, 2007.

[15] http://www.nlp.org.cn/

[16] Zhang, H., H. Yu, et al.. “HHMM-based Chinese lexical analyzer ICTCLAS”, *Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pp 184–187, 2003

[17] Jun, L., L. Ting, et al.. "LTP: an XML-based open language technology platform." *Proceedings of the 25th Anniversary of the Chinese Information Processing Society of China*, Beijing: Tsinghua University Press: 561-572, 2006

[18] http://www.ldc.upenn.edu/

[19] http://www.CCCForum.org.

[20] www.chineseldc.org/

[21] http://icl.pku.edu.cn

[22] http://ncl.xmu.edu.cn

[23] http://www.cjk.org/