Analysis on Characteristics of Chinese Spoken Language[1]

Chengqing Zong, Hua Wu, Taiyi Huang, Bo Xu

National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, China, 100080
{cqzong, huang, wh, xubo}@nlpr.ia.ac.cn

Abstract

For studying and developing human-computer dialog system and spoken language translation system oriented to the restricted domain, data collecting and analysis of the characteristics of spoken language are very important. How to establish proper strategies to process new linguistic phenomena and to enhance the expandability and transplantation of system is also important. This paper presents a method to deal with corpus from different domains. By using this method, the characteristics of Chinese spoken language in hotel reservation are studied and the results are presented in this paper.

Key Words: Corpus Analysis, Corpus Collection, Spoken Language Translation, Human-Computer Dialog, Spoken Language Parsing

1. Introduction

Collecting and analysis of corpus are very important tasks in research of human-computer dialog system and spoken language translation system. Especially, when the restricted domain is changed or expanded, how to deal with new linguistic phenomena and have the analysis algorithms not modified as much as possible are very important. So in corpus processing, how to establish proper strategies to enhance the expandability and transplantation of system is an important aspect addressed by this paper.

Recently, more and more research results on discourse processing of English or other languages are published.[1,2,3] Unfortunately, in Chinese information processing almost all research work in the past decades focussed on the text processing, such as statistic analysis of articles on newspaper or homepages on INTERNET etc. It is just very beginning to research on Chinese spoken language. Although some literatures involve Chinese spoken language.[4,5,6] The analysis on characteristics of Chinese spoken language is only qualitative. However, what is the difference between formal language and spoken language in Chinese? How about the informal linguistic phenomena in the Chinese spoken language? There is still no quantitative analysis and explanation.

In this paper, section 2 presents strategies to deal with corpus from different domains. Section 3 proposes a method to count the characteristics of Chinese spoken language, and the statistical results on characteristics of Chinese spoken language in hotel reservation domain are also presented in this section. Section 4 is concluding remarks.

2. Strategies for Processing Corpus

In this section, a new method is introduced which is designed by us to straighten out and analyze corpus in domain of hotel reservation.

2.1 Collection of Corpus

We collect corpus by using of an automatic record telephone. The dialog in Chinese between "guest" and "hotel service desk" is carried out freely, and the dialog content is recorded automatically.

[1] The research work described in this paper was supported by the National Natural Science Foundation of China under the grant No. 69835030, and supported by the National '863' Hi-Tech Program under the grant No. 863-306-ZT03-02-2, the China Post-doctoral Science Foundation and also the CAS K.C.Wong Postdoctoral Foundation.

Presently, we have already collected 112 dialogs, about 90K Chinese text. The topics are limited in hotel reservation including reservation time, room condition, price and traffic etc.

2.2 Pre-processing of Corpus

The purposes of pre-processing corpus mainly include tasks listed as follows:

* To convert acoustic signals on tapes into characters;

* To make word segmentation in the corpus;

* To make key marks for each dialog paragraph.

In our system, the corpus is automatically pre-processed under the help of human. The acoustic signals recorded on tapes are input into computer firstly, and then converted into Chinese characters by a speech recognition system. Finally the conversion results are checked and corrected by human. As the same way, character corpus is segmented automatically by a word segmentation software, and then the segmentation results are proofread by human.

2.3 Design of Universal Spoken Language Dictionary

For purpose to deal with corpus conveniently in different domains and to create dictionary easily for a spoken language processing system oriented to a new domain or task, we propose a strategy to establish an universal spoken language dictionary. The universal dictionary in our system consists of two parts: static dictionary (SD) and dynamic dictionary (DD).

Definition 1 Static Dictionary. If the number of words in dictionary is relative stable and the meaning of each word is generally fixed, the dictionary is called static dictionary. Signed as SD.

Definition 2 Dynamic Dictionary. If the number of words in dictionary may be increased or reduced, and the meanings of some words may be changed with different application domain, the dictionary is called dynamic dictionary. Signed as DD.

The SD and DD are comparative to each other. In our system SD mainly contains all Chinese functional words, pronoun and basic numeral including ordinal number and cardinal number etc. The DD mainly contains some noun, verb and adjective words etc. in common use. The basis to select noun, verb, adjective words and other content words is word frequencies which are counted based on large scale real corpus without any limitation. All words in SD and DD are tagged, and each entry contains part-of-speech, semantic information and corresponding English word etc. SD and DD together make up the system dictionary.

However, no matter how change the domain, the words in SD is generally fixed. As shown in figure 2-1, when domain is expanded and new corpus is collected, after pre-processing, the corpus will be counted comparing to the original dictionary, and all new words will be picked out. For expansion of the system dictionary, the only work that human will do are to decide which new word should be appended to dynamic dictionary and then to tag it. Similarly, it is easy to create a new dictionary based on the system universal dictionary and corpus collected from a new specific domain.

3. Statistic and Analysis on Chinese Spoken Language

Based on the corpus we collected from hotel reservation domain, the characteristics of Chinese spoken language are studied and analyzed quantitatively. The statistic results are presented in this section.

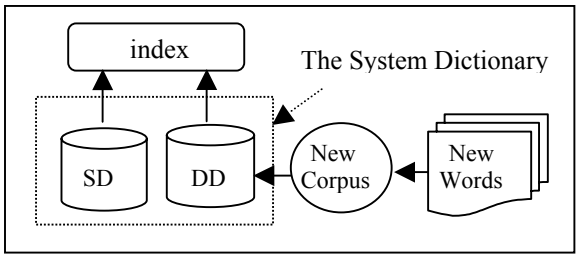

Figure 2-1 The Constitution of Dictionary

3.1 On Corpus Tagging

According to the strategies presented in section 2, we firstly design and construct a system universal dictionary of Chinese spoken language, and then create the domain-dependant dictionary (DDD). The corpus is tagged by using of DDD, and the part-of-speech of each word in dialog sentence is tagged. Some informal sentences are also recognized and marked automatically by system. The tagged corpus is finally checked and corrected by humans. The method for recognizing informal sentences is not described here due to the limitation of paper length, and it will be presented in another paper.

3.2 Statistic Results

The distribution of word length, dialog sentence length, part-of-speech and the proportion of each kind of informal sentences are all counted in basis of corpus that we collect in domain of hotel reservation.

(1) Distribution of Word Length. Comparing to word segmentation of text, the word segmentation of Chinese spoken language has its own characteristics. In spoken language some oral phrases or pet phrases appear more frequently and their meanings are generally fixed. They are consequently considered as words in our system although they are not real words according to the standards of word segmentation of Chinese language, such as "hao ma (it means IS IT OK ?)", "shi de (it means YES)" etc.

In our corpus the longest words contain 4 Chinese characters. The distribution of word length from 1 to 4 is shown in table 3-1.

Length	1	2	3	4
Rates(%)	28.50	57.20	12.99	1.31

Table 3-1 The Distribution of Word Length

In average the word length in spoken language is about 1.87. It is much shorter than the average length of words in Chinese text.[6]

(2) The Length of Dialog Sentence. In our experiment, we define the dialog sentence as follows:

Definition 3 Dialog Sentence From the beginning of speaker's talk to the end, the whole character sequence is considered as a dialog sentence, and the number of Chinese characters is called length of the dialog sentence.

According to definition 3, the lengths of dialog sentences in our corpus distribute from 1 to 67. The results are shown in table 3-2.

Length	1	2	3	4	5	6
Ratio(%)	15.12	8.34	9.28	8.54	7.68	6.78
Length	7	8	9	10	11-67	
Ratio(%)	5.27	5.27	4.78	4.09	24.84	

Table 3-2 The Distribution of Dialog Sentence Length

The average length of dialog sentence in our corpus is about 7.8. It is also much shorter than the average length of sentences in text.

(3) Distribution of Part-of-speech. In literatures regarding to part-of-speech of Chinese words, the division method and the number of part-of-speech are different. However, the authors think that how to divide the part-of-speech and the number of

part-of-speech are all not important. The key problem is how to use the part-of-speech(POS) in analysis of sentences. Here we divide the part-of-speech of Chinese words into 18 kinds as follows: noun(N), verb(V), judgement verb(J), auxiliary verb(X), adjective(A), place-name(W), conjunction(C), adverb(D), direction word(F), auxiliary word(H), classifier(L), pronoun(P), numeral(Q), preposition(R), mood auxiliary word(M), sound imitation word(Y), time word(T), idiom(I). The Idiom here mainly includes all respect word, insert phrases and interjection or response words used in spoken language. The results of distribution of these 18 part-of-speeches are listed in table 3-3.

From table 3-3 we can see that numeral, verb and noun are most frequently used in analyzed corpus. It is consistent with Chinese language that noun and verb

POS	A	C	D	F	H
Rate(%)	4.00	1.52	6.84	0.52	3.98
POS	I	J	L	M	N
Rate(%)	10.77	2.63	2.87	5.37	14.69
POS	P	Q	R	T	V
Rate(%)	10.88	15.61	0.66	3.10	15.31
POS	W	X	Y		
Rate(%)	0.47	1.63	0.00		

Table 3-3 The Distribution of Part-of-speech

are widely used. The reason why numeral ratio is so high is due to the specific domain. In procedure of hotel reservation, the digits are often spoken out in forms as telephone number, price, date and room number etc. So the high ratio of numeral is dependent on the specific domain.

(4) Appearance Ratio of Informal Sentences. In spoken language, generally there are various of informal sentences. These informal sentences are major obstacles for parsing speaker's sentences syntactically, but how many ratio the informal sentences take in spoken language, there is still not quantitative result. In this paper we divide informal sentences into 4 types mainly: a) redundant sentences(RdS); b) repetition sentences(RpS); c) word-order confusion(WoC) and d) incomplete sentences(IcS). What is so called redundant sentence means that one word at least is redundant in a sentence. Similarly, word-order confusion means that one word at least is at wrong position in a sentence, and so on. The one-word-only sentence(OwS) is also counted as a special linguistic phenomenon, and the results are also listed in tables 3-4.

Linguistic Phen.	RdS	RpS	WoC
Ratio (%)	4.70	3.56	1.23
Linguistic Phen.	IcS	Ows	TpC
Ratio (%)	32.61	44.59	5.68

Table 3-4 Appearance Ratio of Informal Sentences

Where TpC in table 3-4 means two or more than two informal linguistic phenomena coexist in a same sentence.

From the results shown in table 3-4 we can see that informal linguistic phenomena widely exist in Chinese spoken language. Especially the sum of omission sentences and one-word-only sentences takes more than 50% in total sentences. So it brings parsing algorithm much trouble in Chinese language understanding. On the other hand, it is a good thing for speech-to-speech translation that one-word-only sentences appear so many, because it is not difficult to translate a word or phrase into another language as long as the word or phrase exists in system dictionary.

4. Conclusion

Spoken language parsing is one of key issues in research of spoken language processing , and collection and analysis of corpus are basis for designing parsing algorithm. Although the method and results presented in this paper are based on the corpus restricted in specific domain, the results show the common law of modern Chinese spoken language, and the processing method is of general meanings. The authors believe that it will provide beneficial reference for research of Chinese discourse processing. However, more key techniques and strategies in corpus collecting and analyzing are still remained to study in further. In next step of our work, the following issues will be addressed:

- Automatic detecting of domain-dependant words;
- Automatic detecting of various ill-formed sentences;
- Statistic analysis on sentence type of Chinese spoken language.

5. Acknowledgement

The authors are grateful to Mr. Zhao Hongjian for his helpful work. The authors also would like to say a very big thank to the anonymous reviewers for their beneficial comments.

References

[1] Rebecca J. Passonneau, Diane J. Litman. Discourse Segmentation by Human and Automated Means. Computational Linguistics. Vol. 23, No. 1, 1997. Pages 103~139.

[2] Marilyn A. Walker, Johanna D. Moore. Empirical Studies in Discourse. Computational Linguistics. Vol. 23, No. 1, 1997. Pages 1~12.

[3] Alexandra Georgakopoulou, Dionysis Goutsos. Discourse Analysis. Edinburgh University Press, 1997.

[4] Chen Jianmin. Modern Chinese Spoken Language. Beijing Press 1984.

[5] Zong Chengqing, Zhang Xin, Huang Taiyi and Zhao Shubin. The Chinese Spoken Language Understanding Based on the Dialog Knowledge (in Chinese). In Proceedings of 1998 International Conference on Chinese Information Processing (ICCIP'98). Nov. 18 - 20, Tsinghua University, China. pp. 143-148.

[6] Huang C., Xu P., Zhang X., Zhao S.B., Huang T.Y., Xu B., "Lodestar: A Mandarin Spoken Dialogue System For Travel Information Retrieval", To Appeared in EuroSpeech ' 99, Sept.5-9, 1999, BUDAPEST, HUNGARY.

[7] Liu Yuan, Liang Nanyuan and Shen Xukun. The Standards of Chinese Word Segmentation for Information Processing and the Methods of Chinese Word Segmentation (in Chinese). Tsinghua University Press 1994.