Construction of Chinese Segmented and POS-tagged Conversational Corpora and Their Evaluations on Spontaneous Speech Recognitions

Xinhui Hu, Ryosuke Isotani, Satoshi Nakamura
National Institute of Information and Communications Technology, Japan
{xinhui.hu, ryosuke.isotani, satoshi.nakamura}@nict.go.jp

Abstract

The performance of a corpus-based language and speech processing system depends heavily on the quantity and quality of the training corpora. Although several famous Chinese corpora have been developed, most of them are mainly written text. Even for some existing corpora that contain spoken data, the quantity is insufficient and the domain is limited. In this paper, we describe the development of Chinese conversational annotated textual corpora currently being used in the NICT/ATR speech-to-speech translation system. A total of 510K manually checked utterances provide 3.5M words of Chinese corpora. As far as we know, this is the largest conversational textual corpora in the domain of travel. A set of three parallel corpora is obtained with the corresponding pairs of Japanese and English words from which the Chinese words are translated. Evaluation experiments on these corpora were conducted by comparing the parameters of the language models, perplexities of test sets, and speech recognition performance with Japanese and English. The characteristics of the Chinese corpora, their limitations, and solutions to these limitations are analyzed and discussed.

1. Introduction

In corpus-based machine translation and speech recognition, the performance of the language model depends heavily on the size and quality of the corpora. Therefore, the corpora are indispensable for these studies and applications. In recent decades, corpus development has seen rapid growth for many languages such as English, Japanese, and Chinese. For Chinese, since there are no plain delimiters among the words, the creation of a segmented and part-of-speech (POS)-tagged corpus is the initial step for most statistical language processes. Several such Chinese corpora have been developed since the 1990s. The two most typical are People's Daily corpus (referred to as PKU), jointly developed by the Institute of Computational Linguistics of Peking University and the Fujitsu Research & Development Center [1], and the Sinica Corpus (referred to as Sinica) developed by the Institute of Information Science and the CKIP Group in Academia Sinica of Taiwan [2]. The former is based on the *People's Daily* newspaper in 1998. It uses standard articles of news reports. The latter is a balanced corpus collected from different areas and classified according to five criteria: genre, style, mode, topic, and source. Although conversational text is also contained in this corpus, it has only 75K of utterances and the domains are limited to a few fields, such as academia and economics, and the style is mostly in address and seldom in conversation.

Since the features of conversation differ from written text, especially in news articles, the development of a segmented and POS-tagged corpus of conversational language is promising work for spontaneous speech recognition and speech-to-speech translation.

In the Spoken Communication Group of NICT, in order to study corpus-based speech translation technologies for the real world, a set of corpora on travel conversation has been built for Japanese, English, and Chinese [3]. These corpora are elaborately designed and constructed on the basis of the concept of variety in samples, situations, and expressions. Now these corpora have been used in the NICT speech-to-speech translation (S2ST) system [8] and other services.

In this paper, we introduce our work on this Chinese corpora development and applications in S2ST speech recognition using these corpora. In Section 2, we provide a brief description of the contents of the NICT corpora, then describe how the Chinese data were obtained. In Section 3, we illustrate the specifications for the segmentation and POS tagging designed for these corpora. Here, we explain the guidelines of segmentation and POS tagging, placing particular emphasis on the features of conversation and speech recognition application. In Section 4, we outline the development procedures and explain our methods of how to get the segmented and POS-tagged data. Some statistical characteristics of the corpora will be shown here. In Section 5, evaluation experiments of speech recognition utilizing these corpora are reported by comparing the results using the same data sets of

Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 70–75,
Suntec, Singapore, 6-7 August 2009.

Japanese and English. Finally, in Section 6, we discuss the performance of the corpora, the problems that remain in the corpora, and give our ideas concerning future work.

2. Current NICT Chinese Corpora on Travel Dialog Domain

At NICT, in order to deal with various conversational cases of S2ST research, several kinds of corpora were elaborately designed and constructed [3]. Table 1 gives a brief description of the data sets related to the development of the Chinese corpora. Each corpus shown in this table was collected using different methods, for different application purposes, and was categorized into different domains.

Table 1. NICT Corpora Used for Chinese Processing

Name	Collecting Method	Uttr.	Domain
SLDB	Bilingual conversation evolved by interpreters.	16K	Dialogues with the front desk clerk at a hotel
MAD	Bilingual conversation evolved by a machine translation system.	19K	General dialogues on travel
BTEC	Text in guidebooks for overseas travelers	475K	General dialogues on travel

The SLDB (Spoken Language Database) is a collection of transcriptions of spoken language between two people speaking different languages and mediated by a professional interpreter.

In comparison, the MAD (Machine Translation Aid Dialogue) is a similar collection, but it uses our S2ST system instead of an interpreter.

The BTEC (Basic Travel Expression Corpus) is a collection of Japanese sentences and their English translations written by bilingual travel experts. This corpus covers such topics related to travel as shopping, hotel or restaurant reservations, airports, lost and found, and so on.

The original data of the above corpora were developed in the form of English-to-Japanese translation pairs. The Chinese versions are mainly translated from the Japanese, but a small portion of BTEC (namely, BTEC4, about 70K of utterances) was translated from English. Every sentence in these corpora has an equivalent in the other two languages, and they share a common header (ID), except for the language mark. All the data in these three languages constitute a set of parallel corpora. The following shows examples of sentences in the three languages:

*Chn.: BTEC1\jpn067\03870\zh*我想喝浓咖啡。

Eng.: BTEC1\jpn067\03870\en\\\\I'd like to have some strong coffee.

*Jap.:BTEC1\jpn067\03870\ja*濃いコーヒーが飲みたい。

3. Specifications of Segmentation and Part-of-Speech Tagging

By mainly referring to the PKU and taking into account the characteristics of conversational data, we made our definitions for segmentation units and POS tags. Here, we explain the outlines of these definitions, then illustrate the segmentation and POS-tagging items relating to those considerations on conversations.

3.1. Guidelines of the Definitions

(1) Compatibility with the PKU and Taking into account the Demand of Speech Recognition of S2ST

Since the specification of segmentations and POS-tagging proposed by the PKU [4] has its palpability and maneuverability and is close to China's national standard [5] on segmentation and close to the specification on POS tags recommended by the National Program of China, namely, the 973-project [6], we mainly followed PKU's specification. We adopted the concept of "segmentation unit," i.e., words with disyllable, trisyllable, some compound words, and some phrases were regarded as segmentation units. The morpheme character (word) was also regarded as an independent unit.

However, we made some adjustments to these specifications. In the speech recognition phase of S2ST to deal with data sparseness, the word for "training" needed to be shortened. So a numeral was divided into syllabic units, while both the PKU and the Sinica took the whole number as a unit. For the same reason, the directional verbs (趋向动词), such as "上, 下，来，去，进去，and 出来," which generally follow another verb and express action directions, were divided from the preceding verb. The modal auxiliary verbs (能愿动词), such as "能，想，and 要," which often precede another verb were separated and tagged with an individual tag set. Because the numeral can be easily reunited as an integrated unit, such a processing method for numerals does not harm the translation phase of S2ST. Moreover, if the directional verb and the modal auxiliary verb can be identified, they will help the syntactic analysis and improve the translation phrase. These two kinds of verbs, together with "是 (be)" and "有 (have)" are more frequently used in

colloquial conversations than in written text, so we took them as an individual segmentation unit and assigned a POS tag to each. The special processes for these kinds of words aim at reflecting the features of spoken language and improve the performance of the S2ST system.

(2) Ability for Future Expansion

Although the corpora were developed for speech recognition in S2ST system, it is desirable that they can be used in other fields when necessary. This reflects in both segmentation and POS-tagging. In segmentation, the compound words with definitive suffix or prefix are divided, so they can be combined easily when necessary. In POS-tagging, the nouns and verbs are mainly further categorized into several sub-tags. We selected about 40 POS tags for our corpora, as shown in Table 1 in the Appendix. With such scale of tag sets, it is regarded to be suitable for language model of ASR. When necessary, it is also easy to choice an adequate tag set from it to meet the needs of other tasks.

(3) Relation with the Corpora of Other Languages in NICT

The original data of the corpora are in Japanese or English. It is meaningful to build connections at the morphological level among these trilingual parallel corpora at least for "named entities." For example, we adopted the same segmentation units as in Japanese, and we subcategorized these words into personal names, organization names, location names, and drink and food names and assigned them each an individual tag. Personal names were further divided into family names and first names for Chinese, Japanese, and Western names. These subclasses are useful in language modeling, especially in the travel domain.

3.2. Some Explanations on Segmentation and POS-tagging

(1) About Segmentation

In our definition of a segmentation unit, words longer than 4 Hanzis (Chinese characters) were generally divided into their syntactic units. Idioms and some greeting phrases were also regarded as segmentation units. For example: "你好/，欢迎光临/，再见/，好的/." Semantic information was also used to judge segmentation unit. For example:

- 我/ 想/ 去/ **最/ 好/** 的/ 餐馆/ 。/ *(Tell me the best restaurant around here.)*
- **最好/** 是/ 价钱/ 不太/ 贵/ 的/ 宾馆/ 。/ *(I'd like a hotel that is not too expensive.)*

For segmenting compound words with different structures, we constituted detailed items to deal with them. These structures include "coordinated (并列), modifying (偏正), verb-object (动宾), subject-predicate (主谓), and verb-complement (述补)." The main consideration for these was to divide them without changing the original meaning. For those words that have a strong ability to combine with others, we generally separated them from the others. This was due to the consideration that if it were done in another way, it would result in too many words. For example, in the verb-object (动宾) structure, "买 (buy)" can combine with many nouns to get meaningful words or phrases, such as "买书 (buy book), 买肉 (buy meat), 买票 (buy ticket), and 买衣服 (buy clothes)." We prescribed separating such active characters or words, no matter how frequently they are used in the real world, to ensure that the meaning did not change and ambiguity did not arise. So the above phrases should be separated in following forms: "买/ 书/ (buy book), 买/ 肉/ (buy meat), 买/ 票/ (buy ticket), and 买／ 衣服 /buy clothes)."

For the directional verbs, we generally separated them from their preceding verbs. For example:

我/ 可以/ 换/ **到/** 别的/ 座位/ 吗/ ？/ *(Is it all right to move to another seat?)*

请/ 把/ 这/ 个/ 行李箱/ 保管/ **到/** 一点钟/ 。/ *(Please keep this suitcase until one o'clock.)*

Prefix and appendix were commonly separated from the root words. For example:

学生/ **们/** 都/ 去/ 京都/ 吗/ ？/ *(Are all students going to Kyoto?)*

我/ 是/ 自由/ 职业/ **者/.** *(I do free-lance work.)*

(2) About POS-Tagging

The POS tag sets are shown in Table 1 in the Appendix. The POS tagging was conducted by the grammar function based on how the segmentation unit behaves in a sentence.

4. Procedure of Developing the Chinese Corpora

The segmented and POS-tagged data were obtained in two steps. The first step was to get the raw segmented and POS-tagged data automatically by computer. The second was to check the raw segmented and POS-tagged data manually.

(1) Getting Raw Segmented and POS-Tagged Data

The text data were segmented and POS tagged by using the language model shown in formula (1).

$$P(L) = \alpha P(w_i \mid w_{i-1} w_{i-2}) + (1-\alpha) P(w_i \mid c_i) P(c_i \mid c_{i-1} c_{i-2}) \quad (1)$$

Here w_i denotes the word at the *ith* position of a sentence, and c_i stands for the class to which the word w_i belongs. The class we used here is a POS-tag set, and α is set 0.9.

The initial data for training the model were from the Sinica due to their balanced characteristics. The annotated data were added to the training data when producing new data. When the annotated data reached a given quantity (here, the BTEC1 was finished, and the total words in the corpora exceeded 1M), the Sinica data were not used for training. We have conducted an experiment with this model for an open test text of 510 utterances from BTEC, and the segmentation and POS-tagging accuracy was more than 95%. Furthermore, proper noun information was extracted from Japanese corpora and marked in the corresponding lines of the Chinese segmented and POS-tagged data.

(2) Manual Annotation

The manual annotations were divided into two phases. The first was a line-by-line check of the raw segmented and POS-tagged data. The second was to check the consistency. The consistency check was conducted in the following manner:

- Find the candidates having differences between the manually checked data and the automatically segmented and POS-tagged data.
- Pick up the candidates having a high frequency of updating in the above step, and build an inconsistency table. The candidates in this table are the main objects of the later checks.
- Check the same sentences with different segmentations and POS tags.
- List all words having multiple POS tags and their frequencies. Determine the infrequent ones as distrustful candidates and add them into the inconsistency tables.

The released annotated data were appended with a header ID for each token (pair of word entry and POS tag) in an utterance including a start marker and end marker, shown as follows:

BTEC1\jpn067\03870\zh\\\00010\||||UTT-START||||
BTEC1\jpn067\03870\zh\\\00020\我|我||我|r||||
BTEC1\jpn067\03870\zh\\\00030\想|想||想|vw||||
BTEC1\jpn067\03870\zh\\\00040\喝|喝||喝|v||||
BTEC1\jpn067\03870\zh\\\00050\浓|浓||浓|a||||
BTEC1\jpn067\03870\zh\\\00060\咖啡|咖啡||咖啡|n||||
BTEC1\jpn067\03870\zh\\\00070\。||||UTT-END||||

Table 2 shows some of the statistics for the 510K utterances in Table 1 for different languages.

Table 2. Some Statistics of Each Corpora in NICT

	Utter.	Ave. words /Uttr.	Words	Vocab.
Chinese	510K	6.95	3.50M	47.3K
Japanese	510K	8.60	4.30M	45.5K
English	510K	7.74	3.80M	32.9K

Figure 1 shows the distributions of utterance length (words in an utterance) for 3 languages among the 510K annotated data. From Figure 1, we know that the Chinese has the fewest words in an utterance, followed by English, with the Japanese having the most.

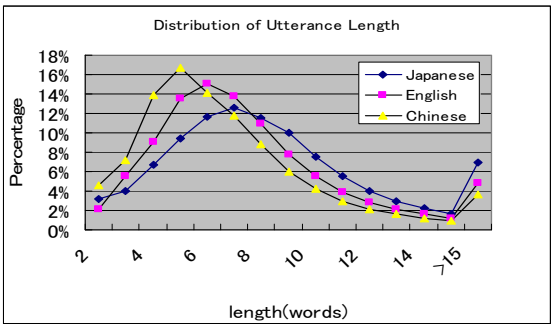

Figure 1. Distribution of utterance length

5. Evaluation Experiments

To verify the effectiveness of the developed Chinese textual corpora, we built a language model for speech recognition using these corpora. For comparisons with other languages, including Japanese and English, we also built language models for these two languages using the same training sets. Meanwhile, the same test set of each language was selected for speech recognition.

5.1. Data Sets for Language Models and Speech Recognitions

For simplicity, we adopted word 2-gram and word 3-gram for evaluating perplexities and speech recognition performance. The training data were selected from the 510K utterances in Table 1, while the test sets were also extracted from them, but they are guaranteed not to exist in the training sets. In evaluations of perplexity, 1524 utterances (a total of three sets) were chosen as the test set. In evaluation of recognition, 510 utterances were chosen as test set. For Japanese and English, the same data sets were also chosen for comparisons.

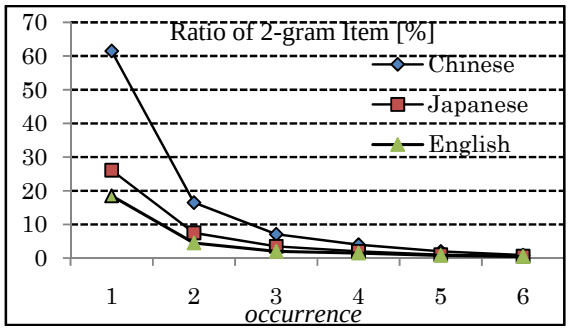

Figure 2. Ratio of 2-gram items with low occurrence

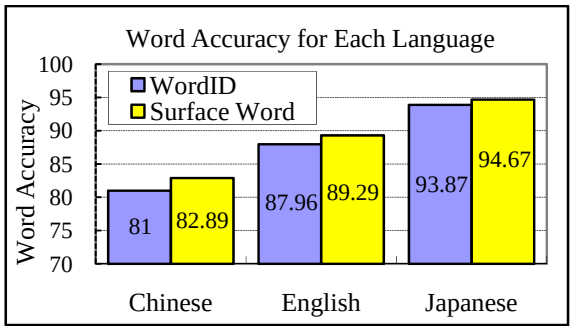

Figure 3. Word recognition accuracies of 3 languages

5.2. Comparisons of Language Models

Using the above utterances in the training sets, a word 2-gram and a 3-gram were built respectively for each language. The distributions of items inside these models were investigated. Figure 2 shows the ratios of 2-gram's items which have low occurrences (from 1 to 6) in the 2-gram model.

Compared with the other two languages, the Chinese has the biggest vocabulary. Moreover, it also has a large amount of low-frequency 1-gram, 2-gram, and 3-gram items. For example, more than 60% of its 2-gram entries appear only once. This can be regarded that the Chinese has more varieties when expressing a same meaning than the other two languages. It is also partly due to bias occurred in the translation process, compared to the original languages. So the probability computations in 2 or 3-gram related to these entries were estimated by using a smoothing approach, so the accuracy is not high.

Table 3 shows average sentence entropies (ASE) of the test sets to the 3-gram models. The ASE is obtained as follows: (1) first to get the product of average word entropy and the total word count in test set. (2) then divide the product by the total sentences in the test set. From the table, we know the Chinese has the maximal sentence entropy (or maximal perplexity) among the three languages. This means that when predicate a sentence in the recognition process, Chinese requires a much bigger search space than the other two languages.

Table 3. Average Sentence Entropy of the Test Sets to 3-gram Models

	Chinese	Japanese	English
Vocab. of Test Set	10,030	12,344	10,840
Ave. Sen. Entropy	294.58	165.80	202.92
Word Perplexity	45.0	20.1	28.5

5.3. Comparison of Speech Recognition Performances

The 2-gram language model was used for decoding recognition lattice, while the 3-gram model was used for rescoring process. The recognition results are shown in Figure 3. Here, WordID refers to the word's outer layer (entry) together with its POS tag, other information like conjugation of verbs, declension of nouns, etc., while the surface word contains only its outer layer, no POS tag is contained in this case .

The difference in word accuracy of speech recognition between these two forms is about 2% for Chinese, and 1% for English and Japanese.

6. Summary

This paper described the development of Chinese conversational segmented and POS-tagged corpora that are used for spontaneous speech recognition in S2ST system. While referring mainly to the PKU's specifications, we defined ours by taking into account the needs of S2ST. About 510K utterances, or about 3.5M words of conversational Chinese data, are contained in these corpora. As far as we know, they are presently the biggest ones in the domain of travel, with a style of conversations. Moreover, a parallel corpus was obtained using these 510K pairs of utterances of Chinese, Japanese, and English. These corpora now play a big role in spontaneous language and speech processing, and are used in the NICT/ATR Chinese-Japanese-English Speech-to-Speech Translation System [8] and other communication services. However, according to our evaluations in this paper, there are still some difference in performance among Chinese and other languages, especially Japanese. There is still some room to improve the quality of these corpora mainly because the Chinese text data were translated from other languages, mainly Japanese, with a few words from English. There is some bias in expression, especially for the transliterations of proper nouns. For examples, "Los Angles" is translated as "洛杉矶, 洛杉机, 洛杉基, and 洛山矶." also, some utterances are not like those spoken by native speakers,

like sentence of “非常感谢你的热情” which corresponds to the original sentence of “ご親切に感謝します(I appreciate your kindness).”

For future work, while continuing to improve the consistency of the corpora, we will expand the Chinese corpora from external data resource, such as Web sites and LDC databases, to extract original Chinese spontaneous text data.

7. References

[1] H.M. Duan, J. Song, G.W. Xu, G.X. Hu and S.W. Yu, “The Development of a Large-scale Tagged Chinese Corpus and Its Applications.” http://icl.pku.edu.cn/icl_tr

[2] C.R. Huang, and K.J. Chen, “Introduction to Sinica Corpus,” CKIP Technical Report 95-02/98-04, http://www.sinica.edu.tw/SinicaCorpus

[3] G. Kikui, E. Sumita, T. Takezawa, S. Yamamoto, “Creating Corpora for Speech-to-Speech Translation.” 8th European Conference on Speech Communication and Technology, Vol.1, pp.381-384, Sep., 2003

[4] S.W. Yu, X.F. Zhu, and H.M. Duan, “The Guideline for Segmentation and Part-Of-Speech Tagging on Very Large Scale Corpus of Contemporary Chinese.” http://icl.pku.edu.cn/icl_tr

[5] The National Standard of PRC, “Standardization of Segmentation for Contemporary Chinese.” GB13715, 1992.

[6] Institute of Applied Linguistics of the Ministry of Education, China, “Specification on Part-of-Speech Tagging of Contemporary Chinese for Information Processing (Draft).” 2002.

[7] H. Yamamoto, S. Isogai, and Y. Sagisaka, “Multi-class Composite N-gram Language Model,” Speech Communication, 2003, Vol.41, pp369-379.

[8] T. Shimizu, Y. Ashikari, E. Sumita, J.S. Zhang, S. Nakamura, “NICT/ATR Chinese-Japanese-English Speech-to-Speech Translation System.” Tsinghua Science and Technology, Vol.13, No.4, pp540-544, Aug. 2008.

Appendix Table 1. Chinese POS Tag Table

POS Tag		Description		POS Tag		Description	
		Chinese	English			Chinese	English
a		形容词	Adjective	n	nppx	人名中的姓	Chinese family name
b		区别词	Non-predicate adjective		nppm	人名中的名	Chinese first name
c		连词	Conjunction		nppxj	日本人的姓	Japanese family name
d		副词	Adverb		nppmj	日本人的名	Japanese first name
de		结构助词	Attributive		nppxw	欧美式人名的姓	Western family name
e		叹词	Interjection		nppmw	欧美式人名的名	Western first name
g		语素字	Morpheme Word		npl	地名	Place
h		前缀词	Prefix		npo	组织名	Organization
i		成语, 习用语	Idiom		npfd	饮食物名	Drink and food
j		简略语	Abbreviation	o		拟声词	Onomatopoeia
k		后缀词	Suffix	p		介词	Preposition
m	m	数词	Numeral	q		量词	Quantifier
	ma	数量定词	Numeral Classifier	r		代词	Pronoun
	mb	概数词	Approximate numeral	u		助词	Auxiliary
n	n	普通名词	Noun	v	v	普通动词	Verb
	nd	方位词	Directional locality		v1	系动词“是”等	Auxiliary verb
	ns	处所名	Space word		v2	动词“有“	Verb “Have”
	nt	时间词	Time word		vt	趋向动词	Directional verb
	nx	非汉字, 字符	Numeric, character string		vw	能愿动词	Modal verb
	np	专有名词	Proper noun	w		标点符号	Punctuation
	npp	人名	Personal name	y		语气助词	Modal particle