Construction and Evaluations of an Annotated Chinese Conversational Corpus in Travel Domain for the Language Model of Speech Recognition

Xinhui Hu, Ryosuke Isotani, Hisashi Kawai, Satoshi Nakamura

National Institute of Information and Communications Technology, Japan
(xinhui.hu, ryosuke.isotani, hisashi.kawai, satoshi.nakamura)@nict.go.jp

Abstract

In this paper we describe the development of an annotated Chinese conversational textual corpus for speech recognition in a speech-to-speech translation system in the travel domain. A total of 515,000 manually checked utterances were constructed, which provided a 3.5 million word Chinese corpus with word segmentation and part-of-speech tagging. The annotation is conducted with careful manual checking. The specifications on word segmentation and POS-tagging are designed to follow the main existing Chinese corpora that are widely accepted by researchers of Chinese natural language processing. Many particular features of conversational texts are also taken into account. With this corpus, parallel corpora are obtained together with the corresponding pairs of Japanese and English texts from which the Chinese was translated. To evaluate the corpus, the language models built by it are evaluated using perplexity and speech recognition accuracy as criteria. The perplexity of the Chinese language model is verified as having reached a reasonably low level. Recognition performance is also found to be comparable to the other two languages, even though the quantity of training data for Chinese is only half the other two languages.

Index Terms: corpus construction, language model, Chinese speech recognition

1. Introduction

In corpus-based speech recognition, machine translation, and information retrieval, the language model is an important aspect. A statistical language model assigns a probability to a sequence of $m$ words $p(w_1, ..., w_m)$ via a probability distribution. The language model is generally trained using annotated corpora. Because the model’s performance depends heavily on the size and quality of the training corpora, the corpora are essential 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, the most famous corpora used for natural language processing are the People’s Daily corpus, jointly developed by the Institute of Computational Linguistics of Peking University and the Fujitsu Research & Development Center (hereinafter “PKU”) [1], the Sinica corpus ("Sinica") developed by the Institute of Information Science and the CKIP Group in Academia Sinica of Taiwan [2], and the modern Chinese balanced corpus ("CNCL") developed by the Chinese National Commission on Language (国家语委) [3]. The PKU is based on the People’s Daily newspaper in 1998 and uses standard news report articles. The Sinica is the first balanced corpus in Chinese, consisting of texts mostly from Taiwan’s newspapers. These texts are collected from different areas and classified according to five criteria: genre, style, mode, topic, and source. The CNCL is a new corpus and the biggest among the present Chinese corpora. This corpus consists of three major categories (social sciences, natural sciences, and general) and about 40 subcategories, which contains more than 20 million Chinese characters. Though these corpora cover a very wide range, they are basically written texts, especially for the PKU corpus, and content is very formal. Although some transcripts of spoken data are contained in the above two balanced corpora, the quantities are relatively small and the domains are also limited. For example, texts tagged as “spoken” in Sinica are mostly excerpted from broadcast news, and have only 75,000 sentences. The domains of “spoken” data concern mainly academia and economics. The style is used mostly in formal communication and seldom in conversation. The CNCL contains about 30,000 sentences tagged as “spoken,” most of which are excerpted from drama scripts.

For practical, corpus-based natural language processing such as speech recognition, the language model should be trained using training data in which topics and expression styles are similar or close to the target of the task.

Since the characteristics of conversation differ from written text, especially from news articles, development of an annotated conversational corpus is promising work for speech recognition in the travel domain.

In the NICT Spoken Language Communication Group, 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 [4]. These corpora are elaborately designed and constructed around the concept of variety in samples, situations, and expressions. The corpora have been used in the NICT speech-to-speech translation (S2ST) system [5] and other services.

In this paper we introduce our work in the development of this Chinese corpus and report the results of an evaluation of it. In Section 2, we give a brief description of the contents of the NICT corpora and describe how the Chinese raw texts were obtained. Then, we illustrate in detail the specifications for the word segmentation and POS tagging designed for constructing this corpus. Here, we focus on discussing particular considerations for dealing with conversational text and automatic speech recognition (ASR) application. Next, we explain specific manual annotation methods for the corpus. In Section 3, we evaluate the corpus by using perplexity and speech recognition accuracy as criteria. Comparisons of these criteria are conducted among Chinese, Japanese, and English at the same time. Finally, in Section 4, we draw our conclusions about this corpus.

2. Construction of NICT Chinese Textual Corpus In Travel Domain

2.1. Acquisition of the Original Data

At NICT, to deal with various conversational cases of S2ST research and to meet the needs of practical applications, several kinds of corpora were elaborately designed and...
constructed [4]. Table 1 gives a brief description of the data sets related to the development of the Chinese corpus. Each set of data shown in this table was collected using different methods, and categorized into different tasks in the travel domain.

Table 1. Data Sets of NICT Corpus for Chinese Processing

<table>
<thead>
<tr>
<th>Name</th>
<th>Collecting Method</th>
<th>Uttrs.</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTEC</td>
<td>Text in guidebooks for overseas travelers</td>
<td>465K</td>
<td>General dialogue on travel</td>
</tr>
<tr>
<td></td>
<td>Bilingual conversation advanced by interpreters</td>
<td>16K</td>
<td>Dialogue with hotel clerk</td>
</tr>
<tr>
<td>SLDB</td>
<td></td>
<td></td>
<td>General dialogue on travel</td>
</tr>
<tr>
<td>MAD</td>
<td>The Proper Nouns Expression Database (PNED)</td>
<td>19K</td>
<td>General dialogue on travel</td>
</tr>
<tr>
<td></td>
<td>Expressions with proper nouns about several areas of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Japan</td>
<td>15K</td>
<td></td>
</tr>
</tbody>
</table>

(*Temporary name)

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

2) The Spoken Language Database (SLDB) is a collection of transcriptions of two people speaking different languages with the mediation of a professional interpreter.

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

4) The Proper Nouns Expression Database (PNED) is a collection of expressions of proper nouns on five areas of Japan. These proper nouns include the names of famous places, souvenirs, and festivals of these areas.

The original data of the above corpus, one million utterances in total, were developed in the form of Japanese-English translation pairs. About half (515,000) of the utterances are translated into Chinese.

As shown in above, the Chinese data are acquired by translation of Japanese and English texts. Although the ideal data should be collected from real speech, so taking Chinese texts from Chinese speech for instance, this is impractical for reasons of cost. On the other hand, with the help of existing foreign corpora, we can efficiently build a Chinese corpus through translation. The effectiveness of the corpus has also been proven in practical applications.

For the peculiarities of conversational texts we made some adjustments to the specifications. These adjustments are reflected in two aspects: in considering the special characteristics of conversational texts and in dealing with the problem of sparse data when training the language model of speech recognition.

1) Consideration of characteristics of conversational texts: Directional verbs (趋向动词) such as 上, 下, 来, 去, 到, 进, and 出 are frequently used in daily conversation. These verbs generally follow another verb and express action directions, such as 看出来, 放进去.

Modal auxiliary verbs (能愿动词), such as 能, 想, and 要, which often precede another verb, also occur with high frequency in travel conversations.

Identifying directional and modal auxiliary verbs helps with syntactic analysis and improves the translation phrase. In our specifications, these two kinds of words are regarded as individual categories and are separated from their neighboring verbs.

For example:

1) 我/可以/换/ 别的/ 座位/ 吗/ ? (Is it alright to move to another seat?)
2) 请/把/这个/行李箱/保管/ 别一点钟/. (Please keep this suitcase until one o'clock.)

Besides the above two verbs, 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.

In our definition of a segmentation unit, words longer than 两汉字 (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/.

2) Data sparseness considerations for ASR.

To deal with the problem of data sparseness in training the language model, some considerations are taken into account.

Because the numeral can easily be reunited as an integrated unit, this processing method for numerals does not harm the S2ST translation phase; so we divided all numerals into individual characters.

The prefix and suffix were commonly separated from the root words. For example:

学生/们/去/京都/吗/？(Are all students going to Kyoto?)
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 candidates that have differences between the manually checked data and automatically segmented and POS-tagged data.
- Pick out the candidates that have 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 those sentences that have the same text but different segmentation and POS tags.
- List all words that have multiple POS tags and their frequencies. Designate infrequent ones as unreliable candidates and add them to the inconsistency tables. Sentences containing these words are again checked.

Refining the Transliteration
We found serious inconsistencies in the transliteration of proper nouns in the original raw texts. This is due to translation from Japanese or English from different sources. For example, “Visa card” is translated as "维萨卡", "Visa卡", "VISA卡", etc, “Los Angeles” is translated as "洛杉矶", "洛杉矶", "洛杉矶", "洛杉矶". To improve consistency, all proper nouns tagged as foreign names, location names, and organizations are picked out, and are again refined. This work is aided by referring to their equivalent utterances in the Japanese and English versions.

Table 2 shows some of the statistics for the 510,000 utterances in Table 1 (the PNED is not contained here) for different languages. Among the trilingual parallel languages, Chinese has a minimum total of words while it has a maximum vocabulary.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>510K</td>
<td>6.95</td>
<td>3.50M</td>
<td>47.3K</td>
</tr>
<tr>
<td>Japanese</td>
<td>510K</td>
<td>8.60</td>
<td>4.30M</td>
<td>45.5K</td>
</tr>
<tr>
<td>English</td>
<td>510K</td>
<td>7.74</td>
<td>3.80M</td>
<td>32.9K</td>
</tr>
</tbody>
</table>

3. Evaluations of the Corpus
To verify the performance of the developed Chinese corpus, we used it to build language models of speech recognition, and then evaluated the perplexity and speech recognition accuracy. For comparisons with the other two languages, Japanese and English, the same evaluations are also conducted. But the training data for these two languages uses all training data, including those utterances that are not translated into Chinese. The total utterances of Japanese and English are respectively 1,000K, double that of the Chinese data.

3.1. Data Sets for Training and Tests
All utterances (515,000) except for the test sets are used for training language models. A total of 1,524 utterances from the BTEC set are used to evaluate perplexity. To evaluate speech recognition evaluation, we selected two test sets. One is a test set containing 510 utterances from the above 1,524; denoted as BTEC. The other is a test set containing 502 utterances from MAD; denoted as MAD. To prevent noise in the language model caused by punctuation, all sentences are normalized before being sent to the training process. All commas, periods, and question marks in a sentence are replaced with a special
tag, “SENT-START-END.” The other punctuation is removed from the training data.

3.2. Language Models

In speech recognition of our S2ST system, a multiclass 3-gram is used for rescoring and a multiclass composite 2-gram for lattice decoding.

The multiclass n-gram is the extension of the conventional word class n-gram. The idea of multidimensional word classes was introduced to separately represent left- and right-context Markovian dependence [7].

The multiclass composite n-gram is a language model further extended from the above multiclass n-gram. In this model, higher-order word n-grams are partially introduced by regarding frequent variable-length word sequences as new word succession entries. Word sequences with high frequency are regarded as independent word entries and their probabilities are approximated by a word n-gram, but the other words are processed by the class-based n-gram [7].

3.3. Perplexity and Speech Recognition

Figure 1 shows the perplexities of the test set (1,524 utterances) respective to the multiclass 3-grams for different class numbers and for three different languages. Among the three languages, Chinese has the highest perplexity. This can be regarded as a reflection of its word distribution – with the maximum vocabulary and minimum total words. We can see that the perplexities of all languages tend to decline in line with the increase in the class number, but become saturated when they approach 8,000. For Chinese, we verified that when the class number is larger than 8,000, its perplexity will rise. We therefore selected 8,000 as the optimized class number.

Table 3 shows the results of speech recognition for two test sets. ChrAcc refers to character accuracy and is specifically used for Chinese (Chn) recognition evaluation. Among the three languages, the performance of Japanese (Jap) is the best for both test sets, but it is a little difficult to judge Chinese and English (Eng) because for BTEC the word accuracy (WrdAcc) of Chinese is higher than it is for English, but for MAD the word accuracy of Chinese is inferior to English. However, it must be pointed out that the sentence accuracy (SenAcc) of Chinese is in both cases higher than that of English.

4. Conclusions

This paper described the development of an annotated Chinese conversational corpus used for speech recognition in an S2ST system. The corpus covers a wide range of travel-related tasks. The data styles contain sentences excerpted from standard texts and sentences collected from several speech environments. The annotation is conducted with careful manual checking. The specifications on word segmentation and POS-tagging are designed to follow the main Chinese corpora that are widely accepted by most researchers of Chinese natural language processing. Many particular features of conversational texts are also taken into account. Finally, an annotated corpus of about 515,000 utterances, or about 3.5 million words, is constructed. Moreover, a set of parallel corpora was constructed using these 515,000 pairs of utterances of Chinese, Japanese, and English. These corpora now play a substantial role in conversational language speech processing, and are used in the NICT Chinese-Japanese-English Speech-to-Speech Translation System and other communication services [5]. In evaluating this corpus, we found that the perplexity of the Chinese language model has reached a good level and its recognition accuracy is as high as we expected, although there are still some differences with Japanese. The main reasons for these differences caused by the factor of the language model are first assumed to be an insufficient quantity, since it is just half of the other two languages, and second to arise from the data translated from the other two languages, mainly from Japanese. There are few expressions containing Chinese proper nouns.

To sum up, the Chinese corpus, mainly built by translation, is verified as effective for building a conversational language model of ASR. These corpora could be extended to other languages using similar methods when necessary.

For future work, we will expand the Chinese corpus from external data resources, including real field experiments, websites, and a huge text database to extract more useful utterances. In addition, the adaptation of a language model will be researched so that it can be made more robust for different topics.

5. Reference