Abstract

In this paper we present the first public Japanese speech corpus for large vocabulary continuous speech recognition (LVCSR) technology, which we have titled JNAS (Japanese Newspaper Article Sentences). We designed it to be comparable to the corpora used in the American and European LVCSR projects. The corpus contains speech recordings (60 hrs.) and their orthographic transcriptions for 306 speakers (153 males and 153 females) reading excerpts from the newspaper’s articles and phonetically balanced (PB) sentences. This corpus contains utterances of about 45,000 sentences—a whole with each speaker reading about 150 sentences. JNAS is being distributed on 16 CD-ROMs.

1 Introduction

In the USA and Europe, effort such as ARPA (NAB)[1] and SQALE [2] have resulted in a large technology push in speaker independent, continuous speech recognition.

In Japan, the Acoustical Society of Japan (ASJ) Continuous Speech Corpora (ASJ-PB)[3] which contain about 10,000 PB sentences, have been widely used as a public resource for LVCSR research. However, we do not have been a large text database; the main reason is that Japanese texts are written without spacing between words, and we do not have an adequate automatic word segmentation tool. For this reason, Japanese LVCSR systems are not well developed. Recently, however, progress with morphological analysis systems has enabled automatic segmentation, and thus some LVCSR systems have begun to develop[4].

In Japan, we have been unable to compare different recognition methods and systems, because we did not have any common Japanese speech corpus for LVCSR research. To stimulate Japanese LVCSR research, we designed a Japanese speech corpus for LVCSR technology that is comparable to the corpora used for NAB and SQALE.

In developing the text database, we still have some language-dependent problems with training the language model. The main problem is that we do not have a general rule to separate text into words. One of the other problems is that, because Japanese text consists three character systems (hiragana, katakana, and kanji (Chinese characters)) there are a lot of variations of spelling.

These problems cause variations between morphological systems (grammar, lexicon, and so on) in Japanese natural language processing (NLP) research. Differences between morphological systems affect the word-frequency lists. For referential comparison, we need to normalize the morphological system or prepare a sharable referential tool as a public standard. We designed and developed the corpus after careful consideration of these points.

In constructing JNAS, the Large Vocabulary Continuous Speech Database Working Group (LVCSWG) of the Special Interest Group of Spoken Language Processing (SIG-SLP) of the Information Processing Society of Japan (IPSJ) designed and developed text sets for recording from 1995 to 1997, and the Speech Database Committee of the ASJ developed speech corpus in 1997. We do not plan any formal project for competitive evaluation such as ARPA or SQALE in our community, but we intended the

Oriental COCOSDA Workshop '98
corpus to be used as common reference data. In this paper, we describe the specification and development of JNAS.

2 Design and Development of the Corpus

2.1 Corpus Structure and Capabilities

The corpus is designed to be comparable to the ‘Wall Street Journal’ (WSJ0) database or the corpora used in the SQALE project.

The corpus is scalable and built to accommodate variable-sized large vocabularies (5K, 20K), variable perplexities (from 0 to 400), and variable sentence length (from 5 to 40 words). The training material for the acoustic models amounts to 100 sentences per speaker and over 100 speakers. The speaker adaptation material amounts to 50 PB sentences per speaker. All materials are recorded simultaneously with a standard close-talk microphone and a secondary desktop microphone. The speech data is collected in a “read” speech mode, and an equal number of male and female speakers are chosen. The corpus permits evaluation both with and without “out-of-vocabulary” lexical items and vocabulary open tests.

2.2 Japanese Language-Dependent Problems in Statistical Language Model Training

Japanese text is not divided by white space at word boundaries. In Japanese text, we use kanji (Chinese orthography) and kana (phonetic characters). There are two types of kana: hiragana and katakana. Kanji consists of ideograms. When writing with ideograms, since the objects of linguistic expression are innumerable, there are also an extremely large number of characters. For example, the four years of newspaper text we used, contained more than 5,000 different characters. Many kanji have multiple readings: one reading is derived from the Chinese pronunciation, and the other is the “Japanese” reading of the Japanese word that corresponds to the meaning of the Chinese character—what is called waajo. Many kanji have the same Chinese pronunciation. Therefore, it is very hard to disambiguate readings in the “mixed kana-kanji” style of writing. Moreover, in the Japanese language, we do not have general rules for what constitutes a single word, and verbs, adjectives, and other inflected words have many inflections.

Therefore, it is difficult to define a word unit, and there are a lot of variations between morphological systems. In making referential comparisons, we need to considered these points carefully.

2.3 Text Preprocessing

As the text for training and evaluation material, we are using the Mainichi Newspaper, one of the major Japanese dailies, because its copyright permission is most suited to our purpose of releasing the resultant corpus to the public.

Ideally, the text preprocessing should divide the sentence into words and resolve the ambiguities with all of the readings of the words. This preprocessing is similar to the type that might be used in a text-to-speech system. A text-to-speech systems' preprocessor, however, can only give the readings, and cannot give any grammatical and/or morphological information, such as segmentation of words or the part of speech of the word, which is useful for constructing a language model for speech recognition. In the research community for Japanese natural language processing, a system called “morphological analysis system” is widely used, and the system estimates word segmentation, part of speech, and inflection.

However, estimation of the reading of the word is beyond the ability of the current morphological analysis system, which is developed for text processing. Moreover, in Japanese we don't have any standard general rule for the definition of vocabulary, morphological grammar, or a system for parts of speech. Therefore, we fixed the goal of text preprocessing as analysis of a sentence by a morphological analysis system.

There were no public morphological analysis systems which had the ability to construct a language model, and so we decided to use the morphological tagged corpus of the Mainichi Newspaper which is distributed as the RWCP-Text-Corpus (RWC-DB-TEXT-95-1) by the RWCP (Real World Computing Partnership) as a standard text database for training of the language model.

An example of the RWCP-Text-Corpus is shown in 1. In the example, each line stands for a morpheme. The first column contains the notation for the morpheme, the second contains the basic form (dictionary form) of the morpheme, and the third contains the ID number of a part of speech (POS).

The RWCP proposed a POS system called THiMCO (Tagset of High quality for Integrated Multi-usage Corpus Openly available to public). In the RWCP-Text-Corpus, THiMCO95 was used.
Figure 1: An example of the RWCP-Text-Corpus

Table 1: text corpus

<table>
<thead>
<tr>
<th>Size</th>
<th>Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5K</td>
<td>85.8</td>
</tr>
<tr>
<td>8.1K</td>
<td>90.0</td>
</tr>
<tr>
<td>20K</td>
<td>95.7</td>
</tr>
<tr>
<td>27.6K</td>
<td>97.0</td>
</tr>
<tr>
<td>291K</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table 2: Frequency-weighted word coverage (from the word-frequency list)

Next, a word bigram language model was constructed to calculate the test-set perplexity of the sentence. The word bigram language model was generated using the CMU SLP Toolkit [5]. The language model was an open vocabulary backoff word bigram which was constructed with the cutoff set to 2, the discount strategy specified as “Good Turing discounting,” and a vocabulary size of 20K words.

Sentences in articles covering three months were classified into 30 categories based on the bigram model. Each category is characterized by the sentence length (2 levels), the vocabulary size (5 levels) and perplexity (3 levels).

A statistically-controlled text set consists of 90 sentences (SC-sentences) collected from the categories according to 3 and about 10 connected sentences taken from a few paragraphs which consisted of only the three or more sentences which were satisfied in any category in 3.

Five other text sets were “article” sets. An “article” set consisted of three articles. Each article included 10 or more paragraphs that consisted only of sentences classified into any class in 3. Each paragraph contained 3-10 sentences. We didn’t check for duplication of sentences between “sentence” sets and “article” sets.
<table>
<thead>
<tr>
<th>VOC</th>
<th>LENGTH = NORMAL</th>
<th>LENGTH = LONG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PERP=P_L</td>
<td>PERP=P_M</td>
</tr>
<tr>
<td>MID</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>MID+</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>LAR</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>LAR+</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>LAR++</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

VOC=MID: 5k voc. without an unknown word
VOC=MID+: 5k voc. with one unknown word

LENGTH=NORMAL: 5-19 morphemes
LENGTH=LONG: 20-39 morphemes
PERP=P_L: 0 < perplexity < 40
PERP=P_M: 40 ≤ perplexity < 85
PERP=P_H: 85 ≤ perplexity < 400

VOC=LAR: 20k voc. without an unknown word
VOC=LAR+: 20k voc. with one unknown word
VOC=LAR++: 20k voc. with two unknown words

LENGTH=NORMAL: 5-29 morphemes
LENGTH=LONG: 30-39 morphemes
PERP=P_L: 0 < perplexity < 70
PERP=P_M: 70 ≤ perplexity < 130
PERP=P_H: 130 ≤ perplexity < 400

Table 3: The number of sentences collected from each category
2.5 Recording

The speech data were recorded in collaboration with 39 sites, so the recording conditions and AD conversion characteristics, including low-pass filter characteristics, were not unified. Each recording site collected data sets for 4-10 speakers (equal numbers of male and female speakers chosen). Each speaker read one set (about 100 sentences) from SC sentence sets, and one subset (about 50 sentences) from the ATR PB sentence sets.

These PB-sentences were chosen by ATR Interpreting Telephony Research Laboratories. Entropy was calculated based on the clusters of two phonemes (120 CV's, 227 VC's and 55 VV's, making 402 clusters in all) and three phonemes (69 CVC's where C is an unvoiced consonant, 18 CVC's where C is a nasal consonant and 136 VCV's where C is a semivowel, making 223 clusters in all) on the assumption that they occur independently. 10,196 original sample sentences were extracted at random from newspapers, magazines, novels, letters, textbooks, etc. Of these, 503 PB sentences were chosen to maximize the entropy. They were sorted so that each set of 50 sentences was also phonetically balanced.

From the 150 SC sentence sets, 138 sets were read by both one male speaker and one female speaker, 4 sets were read by both of two male speakers, 4 sets were read by both of two female speakers, 2 sets were read by one male speaker, and 2 sets were read by one female speaker. At each of the recording sites, all of the speakers read the same PB sentence subset.

The utterances were recorded with two microphones simultaneously: a standard close-talk microphone (Sennheiser HMD410/HMD25-1 or equivalent) and a desktop microphone which was selected independently at each site (Sanken, Sony, and similar). The two versions of the data were stored in separate files.

Each utterance was checked at each recording site. In the prompting text, each word was given a single reading. However, in Japanese, there are words which have several readings (i.e., “Japan” has two readings: nihon, and nippon). An orthographic transcription was created of any changes to readings made at each recording site. No changes to the content of the newspaper articles were permitted under the copyright permission, and the orthographic transcription was not modified for any other errors or variations. A list of these errors was collected in the check list file at each recording site.

3 Speech Corpus: JNAS

The data described here was compiled into 16 CD-ROMs and titled JNAS (Japanese Newspaper Article Sentences). 8 discs (Vol. 1 through Vol. 8) contain the close-talk microphone data, and the others (Vol. 9 through Vol. 16) contain the desktop microphone data. Additional components are the prompting text, the modified transcription in three styles: kanji-text with ruby (readings), katakana, and romaji, original recorded text, check report files, and the bigram language model. The speech waves were digitized at a 16kHz sampling frequency and quantized at 16 bits. They were stored with the NIST SPHERE Headers in the compressed format, using the “shorten” compression technique implemented in the NIST SPHERE PACKAGE.

The CD-ROMs have been released to the public.

4 Conclusion and Future Works

The JNAS corpus and its components have been designed and developed for LVCSR research by the joint efforts of the LVCSD-WG IPSJ and the Speech Database Committee of ASJ.

The Speech Database Committee of ASJ are now selecting text sets for referential evaluation. It plans a training set that contains 100 speakers x 100 sentences and an evaluation set that contains 25 speakers x 4 sentences.

To promote both research of component technologies and development of systems for LVCSR, we have recognized the necessity of a sharable software repository which includes recognition engines, acoustic models and language models. Thus, we are developing a Japanese Dictation Tool Kit [], sponsored by the Information-Technology Promotion Agency (IPA), in Japan. In the project, we will also develop a tool to normalize the differences between morphological analysis systems.

Acknowledgments

The prompting texts and the bigram language models for the Mainichi Newspaper article sentences were prepared by Akinori Ito (Yamagata Univ.), Takehito Utsuro (NAIST), Tatsuya Kawahara (Kyoto Univ.), Toru Shimizu (KDD), Masafumi Tamoto, Kazuhiro Arai (NTT), and Nobuaki Minematsu (TUT). We used the NIST SPHERE package to attach headers.
to the wave files and for the "shorten" compression technique used to reduce the number of CD-ROMs. The NIST SPHERE package was implemented by the Spoken Natural Language processing group, National Institute of Standards and Technology, U.S.A. The 'shorten' compression technique was developed by Tony Robinson at Cambridge University and SoftSound Limited, UK. The speech data was collected by the efforts of many volunteers at the 39 research institutes.

We would like to thank all of the above groups and individuals.

References


