Microsoft Research IME Corpus

December 5, 2005
MSR-TR-2005-168

Hisami Suzuki and Jianfeng Gao
Microsoft Research
One Microsoft Way, Redmond WA 98052

Abstract. This document describes the Microsoft Research IME (MSR-IME) Corpus, a test corpus for language modeling research based on the task of Japanese character conversion for text input. The aim of the corpus is to facilitate research in language modeling by providing a data set on a realistic, non-trivial task that is yet easy to use. The corpus consists of 6,000 sentences, and includes the gold-standard output of conversion, the corresponding input phonetic stream in kana, and the 100-best list of conversion candidates, specifically prepared for the re-ranking formulation of the task. This report includes the description of the task of Japanese character conversion for text input, the details of the corpus as well as the guidelines used to assign the readings to the reference file.

1 Introduction

This report describes the Microsoft Research IME (MSR-IME) Corpus, a test corpus for language modeling research, using the application of Japanese Input Method Editor (IME). IME is a software technology that provides a standard method for inputting text in Asian language such as Chinese and Japanese, and has been the subject of language modeling research in the context of *Pinyin-to-Character conversion* in Chinese (Gao et al., 2002a) and *Kana-Kanji conversion* in Japanese (Gao et al., 2002b).

The task of IME consists of converting the input phonetic strings provided by the user into the appropriate word string using ideographic characters. This is a similar task to speech recognition in many ways: the most obvious similarity is that the problem can also be viewed as a Bayesian decision problem. Let A be the input phonetic string, which corresponds to the acoustic signal in speech. The task of a text input system is to choose the most likely word string W^* among those candidates in **GEN**(A) that could have been converted from A:

$$W^* = \arg\max_{W \in \mathbf{GEN}(A)} P(W \mid A) = \arg\max_{W \in \mathbf{GEN}(A)} P(W)P(A \mid W) \quad (1)$$

Unlike speech recognition, however, there is no acoustic ambiguity in text input, because the phonetic string is provided directly by users. Moreover, in most cases we can assume a unique mapping from *W* to *A* in text input, i.e., $P(A|W) = 1$, so the decision of Equation (1) depends solely on $P(W)$, i.e., the language model probability. A unique mapping from *W* to *A* also means that it is relatively simple to convert *W* to *A*, which facilitates the creation of training data. The mapping from *A* to *W*, however, is highly ambiguous, subject to both word segmentation and character conversion ambiguities. This will constitute a challenging task for language modeling, yet the cost of developing the training and test data is relatively low, which makes it an ideal application for evaluating language modeling techniques. The MSR-IME corpus provides a test data set for this purpose.

2 The Corpus

The MSR-IME Corpus consists of three types of data:

Reference files. They consist of 6,000 sentences randomly extracted from two Japanese news sources, and serve as the gold-standard reference conversion for the task of IME.

Reading files. These files consist of corresponding kana readings for the sentences in the reference files, and serve as the input files for IME. The readings are assigned by two native speakers of Japanese manually inspecting and correcting the results of an automatic conversion. The specification for the reading assignment is found in Section 4.

N-best files. These files contain 100-best conversion candidates for each sentence in the reading files, produced by our word-trigram-based baseline language model (Gao et al., 2002a). These files can be used specifically for evaluating re-ranking approaches.

3 Baseline results

We ran our baseline word trigram language model mentioned above on the MSR-IME Corpus. The model was trained on a 36-million-word Nikkei newspaper corpus, and used a lexicon consisting of 167,107 entries, which are the same settings under which we have run our previous experiments (Gao et al, 2002b; Gao and Suzuki, 2003; Gao et al., 2005; Suzuki and Gao, 2005; Yuan et al., 2005). Table 1 summarizes the results: CER refers to *character error rate*, which is the number of characters wrongly converted from the phonetic string divided by the number of characters in the reference transcript. Oracle CER is the best CER obtained by choosing the best analysis in the 100-best candidate list.

source	# sentences	# characters	CER (%)	oracle CER (%)
news1	3,000	158,796	3.73	1.50
news2	3,000	190,474	2.29	0.72

Table 1: CER of the baseline trigram model

4 Annotation guidelines for assigning reading information

This section contains the specification for creating the reading files mentioned in Section 2.

4.1 Treatment of numbers

Alphanumerical characters are left intact in the reading file. For example:

・６，０００キロから１５，０００キロに延長
・６，０００きろから１５，０００きろにえんちょう

However, kanji numerals are converted into their reading, as in:

・午前十時二十分現在の円相場
・ごぜんじゅうじにじゅっぷんげんざいのえんそうば

Some numbers written in kanji characters have the reading that is not compositional of the readings of component kanji characters. For example, 九四 is pronounced as きゅうじゅうよん in the following context, where じゅう is not realized in writing:

・九四年十一月には二回目の会合を開く

In such a case, we chose to assign reading for each character rather than for the whole word:[1]

・きゅうよん[*きゅうじゅうよ]ねんじゅういちがつにはにかいめのかいごうをひらく

We designated the following readings in parentheses for each kanji numeral:

一(いち)二(に)三(さん)四(よん)五(ご)六(ろく)七(なな)八(はち)九(きゅう)〇(まる)

The above rule is applied *only when* two or more of kanji numerals are used consecutively. Otherwise, readings are assigned for the whole word as they are pronounced:

・一日: ついたち or いちにち (depending on the context)
・九月: くがつ[*きゅうがつ]
・六百万: ろっぴゃくまん[*ろくひゃくまん]
・四一六日: よんーろくにち

4.2 Treatment of alphabets and symbols

Roman alphabets are left intact in the reading file:

・ＥＵが会合を開く。
・ＥＵがかいごうをひらく。

[1] Expression in brackets with an asterisk means we did not choose this annotation.

Below are the symbols that remain intact in the reading file:

。、％＆＃＄＊（）「」｛｝＝＋－・ー

4.3 Ambiguous reading

Words with multiple possible readings are disambiguated as much as possible using contextual knowledge. For example, 今日 can either be read as こんにち or きょう, but it is unambiguous in the following examples:

・長い歴史を経て今日の形になった。
・ながいれきしをへてこんにちのかたちになった。

・今日の夕飯はパスタにしよう。
・きょうのゆうはんはぱすたにしよう。

Certain words have multiple valid readings in a single context (for example, 明日 can be あす or あした) – in these cases, our annotators chose a reading according to their preference.

4.4 Reading of proper names

For proper names, whose readings are difficult to estimate correctly, we resorted to as many resources as possible, both on the web and in books, including lists of Japanese addresses, technical terms, and who's who.

For Chinese- and Korean-origin names written in kanji, commonly used readings are assigned if there is one available (e.g., 司馬遷 as しばせん; 胡錦涛 as こきんとう; 上海 as しゃんはい; 金泳三 as きむよんさむ). Otherwise, we assigned our best guess reading based on the Japanese pronunciation of the kanji characters.

4.5 Other

We did not convert the katakana ヴ into hiragana, and used the character in the reading file.

・ヴェネツィアの歴史
・ヴぇねつぃあのれきし

5 Conclusion

This document described the MSR-IME corpus, a test corpus for language modeling research. The corpus is available for download from the following website: http://research.microsoft.com/research/downloads/. As this is the version 1.0 release, unintended errors may be found in the corpus. We appreciate your feedback – for question or feedback, please send us email at msrimeco@microsoft.com.

References

1. Gao, J., J. Goodman, M. Li and K.-F. Lee. 2002a. Toward a unified approach to statistical language modeling for Chinese. In *ACM Transactions on Asian Language Information Processing*, 1-1: 3-33.
2. Gao. J., H. Suzuki and Y. Wen. 2002b. Exploiting headword dependency and predictive clustering for language modeling. In *Proceedings of EMNLP*, pp.248-256.
3. Gao. J. and H. Suzuki. 2003. Unsupervised learning of dependency structure for language modeling. In Proceedings of *ACL*, pp.521-528.
4. Gao. J., H. Yu, W. Yuan and P. Xu. 2005. Minimum sample risk methods for language modeling. In *Proceedings of HLT/EMNLP*, pp.209-216.
5. H. Suzuki and J. Gao. 2005. A comparative study on language model adaptation using new evaluation metrics. In *Proceedings of HLT/EMNLP*, pp.265-272.
6. W. Yuan, J. Gao and H. Suzuki. 2005. An empirical study on language model adaptation using a metric of domain similarity. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP 05)*, pp.957-968.