Chinese Input Method Based On Reduced Mandarin Phonetic Alphabet

Chun-Han Tseng, Chia-Ping Chen

Department of Computer Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan 800
M943040041@student.nsysu.edu.tw, cpchen@cse.nsysu.edu.tw

10.21437/Interspeech.2006-252

Abstract

In this paper we study the problem of simplifying Chinese input method and making it suitable for use with mobile devices. To see the feasibility of aggressively reducing the number of keystrokes per Chinese character, we compare three input modes: character-based, syllable-based and first-symbol-based. Specifically, we use these linguistic units as token types and compare the perplexities. With the language model trained by data based on the ASBC corpus, the perplexity of the data set we collect from on-line chat and instant messages is 102.6 for character-based model, 67.7 for syllable-based model and 16.3 for first-symbol-based model. Arguing from the relation between the perplexity and the number of "typical" sentences of a language model, our conclusion is that on average there are 6 to 7 characters per first-symbol in natural Chinese language.
Index Terms: speech synthesis, unit selection, join costs.

1. Introduction

With more powerful handsets and faster data communication speeds, mobile electronic devices appear to be the converging points for new information technologies, looming to replace the immobile counter-parts. However, for that to happen, the user interfaces on these devices do need significant overhauls.

Take the instant message (IM) service for example. Being used to run on desktops and laptops, it is now running on the mobile phones since the advent of 3G wireless network. However, in order to input a text message, the users can only use the key pads limited in size and the number of distinct keys. Since the set of potential text is large, this constraint in size posts a severe challenge for a convenient and healthy interface.

From the perspective of source coding, we can view the Chinese input problem as representing each Chinese sentence (source) by a codeword of input symbols. Ideally, a source code has a high probability of being decodable and a low expected code length. Here, in addition, we require that the number of code symbols (the size of the alphabet set) should be as low as possible.

The scenario of our input scheme is as follows. When a user wants to input a sentence, he inputs the sequence of first Mandarin phonetic symbols[1] of the characters in the sentence. Given the input sequence, the system outputs the most-likely candidate sentences for the user to choose from. Whether this is a feasible approach or not depends on the entropy of the text (source) and the entropy of symbol sequence. It is certainly feasible if these entropies are similar in magnitude. Otherwise, there will be many sentences (exponential in the input size) for given input symbol sequence. If this is the case, the system must be able to search efficiently for potential sentences and list the top candidates in the order of probability for the user to choose.

This paper is organized as follows. In Section 2, we review common Chinese input methods and researches on those methods related to Mandarin phonetic alphabet. We describe the principle and practice of our system in Section 3 and 4. We present our experiments and discuss the results in Section 5. In Section 6, we summarize our work.

2. Review

There are several common Chinese input methods: Pinyin (拼音輸入法), Pinzi (拼字), Complex (綜合), Hand-written (手寫), and Number (數字). The Pinyin is based on using the Mandarin phonetic symbols to represent a character, such as the Syllable (注音), the Microsoft New Syllable, and the Natural (自然) input methods. The Pinzi method is based on using parts of a character for representation, such as the Chang-Jie (倉頡) and Da-Yi (大易) input methods. The Complex is based on using the form, phoneme and morpheme of a character, such as the Liu (無蝦米) input method. The Hand-written is based on character recognition. In the Number input method, the basic strokes (筆劃) is coded by numbers and the user inputs the sequence of

This work was supported by the National Science Council of Taiwan ROC, grant number NSC94-2213-E-110-061.

[1] The first symbol in a Mandarin syllable is loosely known as the *head*, but they are not quite the same – sometimes the *tail* is the first-symbol if the syllable contains only one symbol.

strokes as numbers for a character.

On the Pinyin methods, there are several research works to improve the accuracy and efficiency. In [1], a statistical approach combining a trigram language model and a segmentation model is proposed to improve the conversion accuracy. In [2], an approach based on compression by partial match is implemented in the language model which outperforms modified Kneser-Ney smoothing methods. In [3], a scalar-quantized compact bigram is used on mobile phones to reduce computational resource.

3. System Overview

The block diagram of our system is shown in Figure 1. First, a user inputs a symbol sequence into the system. With the input as the constraint condition, the system searches and generates a list of candidate sentences with significant probabilities. The list is redirected to the screen for the user to select.

Figure 2 illustrates the three modes of user input for "國立中山大學" (National Sun Yat-Sen University). In the character-based mode, a user has to type all characters correctly. This is virtually error-free as long as the user knows the correct characters. However, this is very time-consuming, and can be tedious on a small device such as a mobile phone. In the syllable-based mode, a user inputs the correct syllable sequence in the symbols of Mandarin phonetic alphabet. The system outputs the most likely character sequences as the input goes along. The user makes a selection when the input of a sentence, word or phrase is finished. This mode is currently the most commonly used mode for Chinese input with PCs or notebooks. In the first-symbol-based mode, a user inputs just the first Mandarin phonetic symbols of the intended characters. This is essentially the same idea of the syllable-based mode, but with a smaller alphabet and a smaller number of keystrokes per character. It relies on the "intelligence" of the system to do the rest of the job of outputting the intended text.

Since different characters can have the same syllable and different syllables can have the same first-symbol, it is expected that compared to character sequence, the ambiguity is higher with syllable sequence and even higher with first-symbol sequence. A higher ambiguity is reflected by a lower entropy. Let X be character sequence and Y be syllable sequence. The joint entropy is

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y). \quad (1)$$

Since $H(Y|X) = 0$ and $H(X|Y) \geq 0$, we have

$$H(X) \geq H(Y). \quad (2)$$

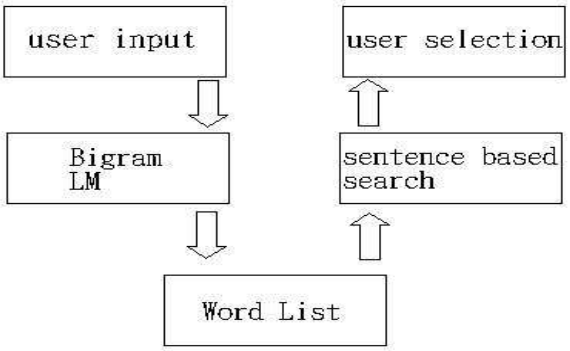

Figure 1: *The system block diagram.*

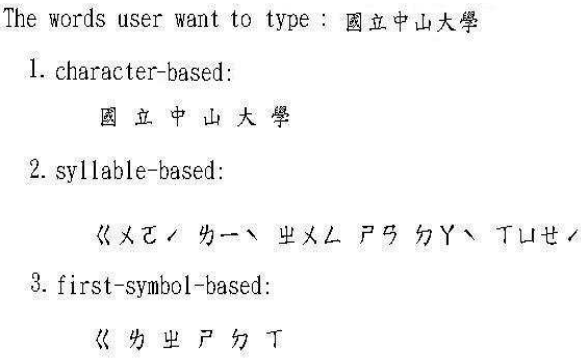

Figure 2: *The input sequences of three modes for "國立中山大學", the Chinese of "National Sun Yat-Sen University".*

4. Language Models

We use the bigram language model. In this model, the probability of a sentence $\mathbf{s}$ is

$$Pr(\mathbf{s}) = p(w_1| < s >) \prod_{j=2}^{l} p(w_j|w_{j-1})\, p(< /s > |w_l) \quad (3)$$

where $< s >$ and $< /s >$ are the symbols for start-of-sentence and end-of-sentence tokens. They are added artificially to each sentence in the corpus. With these tokens, the word unigram at the start of sentence can be replaced by a bigram and the probabilities of all sentences, not conditional on the sentence length, sum to 1.

Given the test set T and the language model P trained by the training set, we compute the perplexity

$$\text{PPL} = 2^{-\frac{1}{n}\log P(T)}, \quad (4)$$

where $P(T)$ is the probability of the test set T using the model of P, and n is the number of word tokens in the test

set. From (4), the number of typical sentences is approximately [5]

$$\frac{1}{P(T)} \sim (\text{PPL})^n. \tag{5}$$

Using the bigram model, we have

$$\log P(T) = \sum_{i=1}^{N} \left[\sum_{j=1}^{l_i} \log p(w_j^i | w_{j-1}^i) + \log p(< /s > | w_{l_i}^i) \right], \tag{6}$$

where N is the total number of sentences, l_i is the number of words in sentence i, and w_j^i is the jth word in sentence i.

To estimate the parameters in the bigram language model, we use a maximum-likelihood-based estimator modified by smoothing and backing-off. The maximum-likelihood estimate (MLE) is simply the relative frequency

$$p(u|v) = \frac{n(u,v)}{n(v)}, \tag{7}$$

where $n(u,v)$ is the count that the bigram $(w_j = u, w_{j-1} = v)$ appears in the train set. To cope with bigrams unseen in the train set, we use the add-one smoothing scheme, adjusting the counts to be

$$\tilde{n}(u,v) = (n(u,v)+1)\frac{n(v)}{n(v)+V}, \tag{8}$$

where V is the size of vocabulary, and use the MLE for the adjusted counts

$$\tilde{p}(u|v) = \frac{\tilde{n}(u,v)}{\sum \tilde{n}(u,v)} = \frac{n(u,v)+1}{n(v)+V}. \tag{9}$$

On top of smoothing, we also incorporate backoff scheme into our bigram language model,

$$\tilde{p}^*(u|v) = \begin{cases} \tilde{p}(u|v), & \text{if } n(u,v) > 0, \\ \alpha(v)\tilde{p}(u), & \text{if } n(u,v) = 0, \end{cases} \tag{10}$$

where $\alpha(v)$ is chosen so that the total probability is 1.

5. Experiments

5.1. Data

5.1.1. Dictionary and Vocabulary

We extract a dictionary from the open-source **xcin**[2] and related library source. An entry in the dictionary is a Chinese character (similar to orthography in English) followed by all its pronunciation variations (similar to homographs). We call this dictionary the xcin dictionary.

[2]xcin is a server for Chinese input under X Window system. See http://xcin.linux.org.tw/

For the character-based mode, a character in the text is labelled by itself. We use all characters in the xcin dictionary, for a total number of 13065 characters. The vocabulary (of a task) is a subset of the dictionary, containing those characters appearing in the train set.

For the syllable-based mode, a character in the text is labelled by the first syllable of the character's entry in the xcin dictionary. For the label set, we use all syllables that appear in the xcin dictionary as the first (or the sole) syllable for some characters, resulting in a total number of 1256 syllables. Note that *toned* syllables are used.

For the first-symbol-based mode, a character is labelled by the first phonetic symbol of the first syllable in the xcin dictionary. It is straightforward to use the set of Mandarin phonetic alphabet, which contains a total number of 37 (first-) symbols.

5.1.2. Text Sets

Two text sets are used in this study. The first, called the ASBC set, is extracted from the Academia Sinica Balanced Corpus [4]. The number of characters in this set is approximately 7.7 million. After adding the start and end tokens, the number of tokens in ASBC is approximately 8.3 million. The content in ASBC is of seven different subjects: literature, life, society, science, philosophy, art, and none.

We collect the second, called the "CHAT" set, from on-line chat messages. As the name indicates, the content of this set is essentially "chats" between friends or classmates. The number of characters collected in CHAT is approximately 130 thousands. Examples of sentences in CHAT are "問妳一下唷" (Let me ask you something) or "那有什麼問題呢" (No problems), the kind of utterances commonly used in on-line conversations or instant messages to communicate with other people.

These two sets are of quite different natures. The ASBC set is of various genres and is quite formal (well-written). The CHAT set is more informal and interactive, imitating the spoken language to a large extent.

5.2. Results

For each mode (character-, syllable- and first-symbol-based), we compute the perplexities of test set using language model trained by train set, of the 4 cases listed in Table 1. Since there are 3 modes, a total number of 12 runs of experiments are conducted in this evaluation, as shown in Table 2. The results on perplexities are summarized in Table 3.

The cross entropy (CE) is an upper bound for the entropy rate of the stochastic process of natural languages. In other words, it is an approximation to the entropy. PPL and entropy are thus related via CE. Compare the perplexities using ASBC as the train set and CHAT as the test set

Table 1: *Usage of data sets for evaluation.*

	train set	test set
A1	ASBC	CHAT
A2	CHAT	ASBC
A3	CHAT	CHAT
A4	ASBC	ASBC

Table 2: *The list of task IDs for our experiments.*

	character	syllable	first-symbol
A1	X1	Y1	Z1
A2	X2	Y2	Z2
A3	X3	Y3	Z3
A4	X4	Y4	Z4

(X1, Y1 and Z1). The perplexities are $102.6, 67.7$ and 16.3 respectively for character-based, syllable-based and first-symbol based modes. On average, the ambiguity of input mode is 1.5 characters per syllable and 6.5 characters per first symbol.

For all three modes, using CHAT as the train set and ASBC as the test set (X2, Y2 and Z2) has the highest perplexity. This is due to the fact that CHAT is a small set with a small vocabulary, resulting in many OOV (out-of-vocabulary) tokens in the test set.

The fact that using CHAT outperforms using ASBC as the train set on CHAT as test set is not too surprising since most probability is distributed to the patterns that appear in the train set.

5.3. Discussion

The result on the syllable-based mode actually supports the fact that syllable-based approach is highly feasible. The search space of character sequences for a given syllable sequence is manageable and fast search can be implemented without significant computational resource.

For the feasibility of first-symbol-based input mode, further research work is required as the search space is enormous. It is necessary to structure the search space so that good candidates can be approached efficiently.

The current framework does not consider adapting the system to specific users: if a user frequently inputs certain patterns, the model parameters can be adjusted accordingly to reflect such idiosyncrasy for better performance.

The language model used here is a bigram model with smoothing and back-off. Although good for fast evaluation, there is a risk that this model is over simplified and unable to capture important dependencies between linguistic patterns.

The CHAT set is quite limited in size. The collection of such data is a difficult issue because text in on-line chat or instant message is quite personal. Instead of switching to other sets, we will continue to work on this domain, since the application in mind is IM with mobile devices.

6. Conclusion

In this paper, we evaluate the feasibility of a Chinese input method based on the first Mandarin phonetic symbols of the syllables of characters. We use the ASBC corpus and collect on-line chat messages. We compute perplexities using bigram language models with smoothing and backoff. We base our evaluation on the ambiguity of the input symbol sequence in specifying the output character sequence. The experimental results suggest that side information may be needed to reduce the ambiguity for the first-symbol-based mode, and justify the feasibility of the syllable-based mode.

Table 3: *Experimental results. OOV = out-of-vocabulary, rate = OOV rate, CE = cross entropy.*

ID	OOV	rate	CE	PPL
X1	15	0.01	6.7	**102.6**
X2	297k	3.6	9.6	782.2
X3	0	0	6.6	98.8
X4	0	0	6.7	103.1
Y1	0	0	6.1	**67.7**
Y2	45k	0.5	8.3	315.6
Y3	0	0	5.9	57.5
Y4	0	0	6.2	73.8
Z1	0	0	4.0	**16.3**
Z2	362	0.004	4.6	23.4
Z3	0	0	4.1	17.1
Z4	0	0	4.0	16.1

7. References

[1] Zheng Chen and Kai-Fu Lee, “A New Statistical Approach to Chinese Pinyin Input”, ACL-2000. The 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, 3-6 October 2000.

[2] Jin Hu Huang and David Powers, “Adaptive Compression-based Approach for Chinese Pinyin Input”, ACL SIGHAN Workshop, pp.24-27.

[3] Feng Zhang, Zheng Chen, Mingjing Li, Guozhong Dai, “Chinese Pinyin Input Method for Mobile Phone”, ISCSLP2000.

[4] 中央研究院漢語料庫的內容與說明, http://www.sinica.edu.tw/SinicaCorpus/98-04.pdf.

[5] T. Cover and J. Thomas, “Elements of Information Theory”, John Wiley and Sons, Inc., 1991, USA, ISBN: 0-471-06259-6.