Jurilinguistic Engineering in Cantonese Chinese: An *N*-gram-based Speech to Text Transcription System

B K T'sou, K K Sin, S W K Chan, T B Y Lai, C Lun, K T Ko, G K K Chan, L Y L Cheung

Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon
Hong Kong SAR, China
Email: rlbtsou@uxmail.cityu.edu.hk

Abstract

A Cantonese Chinese transcription system to automatically convert stenograph code to Chinese characters is reported. The major challenge in developing such a system is the critical homocode problem because of homonymy. The statistical *N*-gram model is used to compute the best combination of characters. Supplemented with a 0.85 million character corpus of domain-specific training data and enhancement measures, the bigram and trigram implementations achieve 95% and 96% accuracy respectively, as compared with 78% accuracy in the baseline model. The system performance is comparable with other advanced Chinese Speech-to-Text input applications under development. The system meets an urgent need of the Judiciary of post-1997 Hong Kong.

Keyword: Speech to Text, Statistical Modelling, Cantonese, Chinese, Language Engineering

1. Introduction

British rule in Hong Kong made English the only official language in the legal domain for over a Century. After the reversion of Hong Kong sovereignty to China in 1997, legal bilingualism has brought on an urgent need to create a Computer-Aided Transcription (CAT) system for Cantonese Chinese to produce and maintain the massive legally tenable records of court proceedings conducted in the local majority language (T'sou, 1993, Sin and T'sou, 1994, Lun et al., 1995). With the support from the Hong Kong Judiciary, we have developed a transcription system for converting stenograph code to Chinese characters.

CAT has been widely used for English for many years and available for Mandarin Chinese, but none has existed for Cantonese. Although Cantonese is a Chinese dialect, Cantonese and Mandarin differ considerably in terms of phonological structure, phonotactics, word morphology, vocabulary and orthography. Mutual intelligibility between the two dialects is generally very low. For example, while Cantonese has more than 700 distinct syllables, Mandarin has only about 400. Cantonese has 6 tone contours and Mandarin only 4. As for vocabulary, 16.5% of the words in a 1 million character corpus of court proceedings in Cantonese cannot be found in a corpus consisting of 30 million character newspaper texts in Modern Written Chinese (T'sou et al, 1997). For orthography, Mainland China uses the Simplified Chinese character set, and Hong Kong uses the Traditional set plus 4,702 special local Cantonese Chinese characters (Hong Kong Government, 1999). Such differences between Cantonese and Mandarin necessitate the Jurilinguistic Engineering undertaking to develop an independent Cantonese CAT system for the local language environment.

The major challenge in developing a Cantonese CAT system lies in the conversion of phonologically-based stenograph code into Chinese text. Chinese is a logographic language. Each character or logograph represents a syllable. While the total inventory of Cantonese syllable types is about 720, there are at least 14,000 Chinese character types. The limited syllabary creates many homophones in the language (T'sou, 1976). In a one million character corpus of court proceedings, 565 distinct syllable types were found, representing 2,922 distinct character types. Of the 565 syllable types, 470 have 2 or more homophonous characters. In the extreme case, *zi* represents 35 homophonous character types.

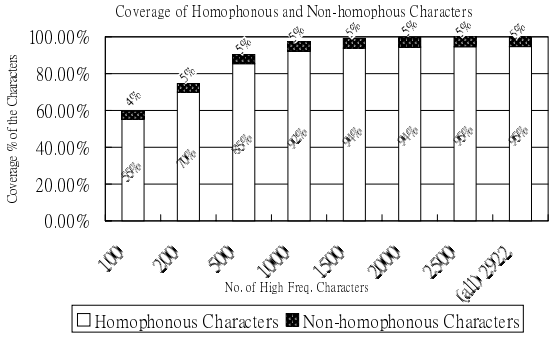

Figure 1. Coverage of Homonymous Characters

These 470 syllables represent 2,810 homophonous character types which account for **94.7%** of the text, as shown in Figure 1. The homocode problem must be properly resolved to ensure successful conversion.

2. Computer-Aided Transcription (CAT)

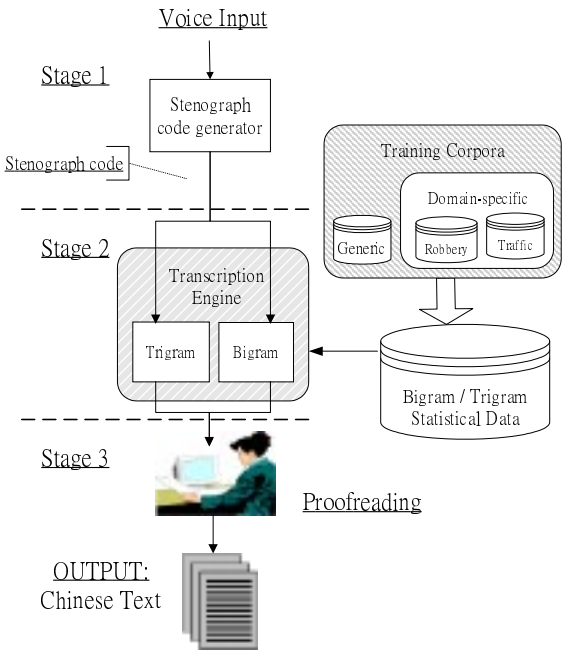

Figure 2. Automatic Transcription Process

Figure 2 outlines the transcription process in the Cantonese CAT system. Following typical courtroom CAT systems, our process is divided into three major stages. In **Stage 1**, *simultaneous* to a litigant speaking, a stenographer inputs speech, i.e. a sequence of transcribed syllables or stenograph codes, via a stenograph code generator. Each stenograph code basically stands for a syllable. In **Stage 2**, the transcription software converts the sequence of stenograph codes $\{s_1, \ldots, s_n\}$ into the original character text $\{c_1, \ldots, c_n\}$. This procedure requires the conversion component to be tightly bound to the phonology and orthography of a specific language. To specifically address homonymy in Cantonese, the conversion procedure in our system is supported by bigram and trigram statistical data derived from domain-specific training. In **Stage 3**, manual editing of the transcribed texts corrects errors from typing mistakes or mis-transcription.

3. System Architecture

3.1 Statistical Formulation

To resolve massive ambiguity in speech to text conversion, the *N-gram model* is used to determine the most probable character sequence $\{c_1, \ldots, c_k\}$ given the input stenograph code sequence $\{s_1, \ldots, s_k\}$. The conditional probability (1) is to be maximized.

(1) $\quad P(c_1, \ldots, c_k \mid s_1, \ldots, s_k)$

where $\{c_1, \ldots, c_k\}$ stands for a sequence of N characters, and $\{s_1, \ldots, s_k\}$ for a sequence of k input stenograph codes.

The co-occurrence frequencies necessary for computation are acquired through training. However, a huge amount of data is needed to generate reliable statistical estimates for (1) if $N > 3$. Consequently, N-gram probability is approximated by bigram or trigram estimates. First, rewrite (1) as (2) using Bayes' rule.

(2) $$\frac{P(c_1,...,c_k) \times P(s_1,...,s_k \mid c_1,...,c_k)}{P(s_1,...,s_k)}$$

As the value of $P(s_1, \ldots, s_k)$ remains unchanged for any choice of $\{c_1, \ldots, c_k\}$, one needs only to maximize the numerator in (2), i.e. (3).

(3) $\quad P(c_1,\ldots, c_k) \times P(s_1 ,\ldots, s_k|c_1,\ldots,c_k)$

(3) can then be approximated by (4) or (5) using bigram and trigram models respectively.

(4) $\quad \prod_{i=1,\ldots,k} (P(c_i|c_{i-1}) \times P(s_i|c_i))$

(5) $\quad \prod_{i=1,\ldots,k} (P(c_i|c_{i-2}c_{i-1}) \times P(s_i|c_i))$

The transcription program is to compute the best sequence of $\{c_1, \ldots, c_k\}$ so as to maximize (4) or (5). The advantage of the approximations in (4) and (5) is that $P(s_i|c_i)$, $P(c_i|c_{i-1})$ and $P(c_i|c_{i-2}c_{i-1})$ can be readily estimated using a training corpus of manageable size.

3.2 Viterbi Algorithm

The Viterbi algorithm (Viterbi, 1967) is implemented to efficiently compute the maximum value of (4) and (5) for different choices of

character sequences. Instead of exhaustively computing the values for all possible character sequences, the algorithm only keeps track of the probability of the *best* character sequence terminating in each possible character candidate for a stenograph code.

In the trigram implementation, size limitation in the training corpus makes it impossible to estimate all possible $P(c_i|c_{i-2}c_{i-1})$ because some $\{c_{i-2}, c_{i-1}, c_i\}$ may never occur there. Following Jelinek (1990), $P(c_i|c_{i-2}c_{i-1})$ is approximated by the summation of weighted trigram, bigram and unigram estimates in (6).

(6) $P(c_i \mid c_{i-2}\, c_{i-1})$

$$= w_3 \times \frac{f(c_{i-2}\, c_{i-1}\, c_i)}{f(c_{i-2}\, c_{i-1})} + w_2 \times \frac{f(c_{i-1}\, c_i)}{f(c_{i-1})} + w_1 \times \frac{f(c_i)}{\Sigma f(c_j)}$$

where (i) w_1, w_2, $w_3 \geq 0$ are weights, (ii) $w_1+w_2+w_3 = 1$, and (iii) $\Sigma f(c_j)$ is the sum of frequencies of all characters. Typically the best results can be obtained if w_3, the weight for trigram, is significantly greater than the other two weights so that the trigram probability has dominant effect in the probability expression. In our tests, we set w_1=0.01, w_2=0.09, and w_3=0.9.

The Viterbi algorithm substantially reduces the computational complexity from $O(m^n)$ to $O(m^2n)$ and $O(m^3n)$ using bigram and trigram estimation respectively where n is the number of stenograph code tokens in a sentence, and m is the upper bound of the number of homophonous characters for a stenograph code.

To maximize the transcription accuracy, we also refine the training corpus to ensure that the bigram and trigram statistical models reflect the courtroom language closely. This is done by enlarging the size of the training corpus and by compiling domain-specific text corpora.

3.3 Special Encoding

After some initial trial tests, error analysis was conducted to investigate the causes of the mis-transcribed characters. It showed that a noticeable amount of errors were due to high failure rate in the retrieval of some characters in the transcription. The main reason is that high frequency characters are more likely to interfere with the correct retrieval of other relatively lower frequency homophonous characters. For example, Cantonese, *hai* ('to be') and *hai* ('at') are homophonous in terms of segmental makeup. Their absolute frequencies in our training corpus are 8,695 and 1,614 respectively. Because of the large frequency discrepancy, the latter was mis-transcribed as the former 44% of the times in a trial test. 32 such high frequency characters were found to contribute to about 25% of all transcription errors. To minimize the interference, special encoding, which resulted from shallow linguistic processing, is applied to the 32 characters so that each of them is assigned a unique stenograph code. This was readily accepted by the court stenographers.

4. Implementation and Results

4.1 Compilation of Corpora

In our expreriments, authentic Chinese court proceedings from the Hong Kong Judiciary were used for the compilation of the training and testing corpora for the CAT prototypes. To ensure that the training data is comparable with the data to be transcribed, the training corpus should be large enough to obtain reliable estimates for $P(s_i|c_i)$, $P(c_i|c_{i-1})$ and $P(c_i|c_{i-2}c_{i-1})$. In our trials, we quickly approached the point of diminishing return when the size of the training corpus reaches about 0.85 million characters. (See Section 4.2.2.) To further enhance training, the system also exploited stylistic and lexical variations across different legal domains, e.g. *traffic*, *assault*, and *fraud offences*. Since different case types show distinct domain-specific legal vocabulary or usage, simply integrating all texts in a single training corpus may obscure the characteristics of specific language domains, thus degrading the modelling. Hence domain-specific training corpora were also compiled to enhance performance.

Two sets of data were created for testing and comparison: *Generic Corpus* (GC) and *Domain-specific Corpus* (DC). Whereas GC consists of texts representing various legal case types, DC is restricted to traffic offence cases. Each set consists of a training corpus of 0.85 million characters and a testing corpus of 0.2 million characters. The training corpus consists of Chinese characters along with the corresponding stenograph codes, and the testing corpus consists solely of stenograph codes of the Chinese texts.

4.2 Experimental Results

For evaluation, several prototypes were set up to

test how different factors affected transcription accuracy. They included (i) use of bigram vs. trigram models, (ii) the size of the training corpora, (iii) domain-specific training, and (iv) special encoding. To measure conversion accuracy, the output text was compared with the original Chinese text in each test on a character by character basis, and the percentage of correctly transcribed characters was computed. Five sets of experiments are reported below.

4.2.1 Bigram vs. Trigram

Three prototypes were developed: the *Bigram Prototype, CAT_{VA2},* the *Trigram Prototype, CAT_{VA3},* and the *Baseline Prototype, CAT_0. CAT_{VA2}* and *CAT_{VA3}* implement the conversion engines using the bigram and trigram Viterbi algorithm respectively. *CAT_0,* was set up to serve as an experimental control. Instead of implementing the *N*-gram model, conversion is accomplished by selecting the highest frequency item out of the homophonous character set for each stenograph code. GC was used throughout the three experiments. The training and testing data sets are 0.85 and 0.20 million characters respectively. The results are summarized in Table 1.

Prototypes	CAT_0	CAT_{VA2}	CAT_{VA3}
Corpus	GC	GC	GC
Accuracy	**78.0%**	**92.4%**	**93.6%**

Table 1. Different *N*-gram Models

The application of the bigram and trigram models offers about 14% and 15% improvement in accuracy over Control Prototype, CAT_0.

4.2.2 Size of Training Corpora

In this set of tests, the size of the training corpora was varied to determine the impact of the training corpus size on accuracy. The sizes tested are 0.20, 0.35, 0.50, 0.63, 0.73 and 0.85 million characters. Each corpus is a proper subset of the immediately larger corpus so as to ensure the comparability of the training texts. CAT_{VA2} was used in the tests.

Training Size	0.20	0.35	0.50
Training Corpus	GC	GC	GC
Accuracy	**89.5%**	**91.2%**	**91.8%**
Training Size	0.63	0.73	0.85
Training Corpus	GC	GC	GC
Accuracy	**92.1%**	**92.3%**	**92.4%**

Table 2. Variable Training Data Size

The results in Table 2 show that increasing the size of the training corpus enhances the accuracy incrementally. However, the point of diminishing return is reached when the size reaches 0.85 million characters. We also tried doubling the corpus size to 1.50 million characters. It only yields 0.8% gain over the 0.85 million character corpus.

4.2.3 Use of Domain-specific Training

This set of tests evaluates the effectiveness of domain-specific training. Data from the two corpora, GC and DC, are utilized in the training of the bigram and trigram prototypes. The size of each training set is 0.85 million characters. The same set of 0.2 million character testing data from DC is used in all four conversion tests. Without increasing the size of the training data, setups with domain-specific training consistently yield about 2% improvement. A more comprehensive set of corpora including *Traffic*, *Assault*, and *Robbery* is being compiled and will be reported in future.

Domain-specific	Not Applied		Applied	
Prototypes	CAT_{VA2}	CAT_{VA3}	CAT_{VA2}	CAT_{VA3}
Training Data	GC	GC	DC	DC
Testing Data	DC	DC	DC	DC
Accuracy	**92.6%**	**92.8%**	**94.7%**	**94.8%**

Table 3. Application of Domain-Specificity

4.2.4 Special Encoding

Following shallow linguistic processing, special encoding assigns unique codes to 32 characters to reduce confusion with other characters. Another round of tests was repeated, identical to the CAT_{VA2} and CAT_{VA3} tests in Section 4.2.1, except for the use of special encoding. The use of training and testing corpora have 0.85 and 0.20 million characters respectively.

S. Encoding	Not Applied		Applied	
Prototypes	CAT_{VA2}	CAT_{VA3}	CAT_{VA2}	CAT_{VA3}
Corpus	GC	GC	GC	GC
Accuracy	92.4%	93.6%	**94.7%**	**95.6%**

Table 4. Application of Special Encoding

Table 4 shows that the addition of special encoding consistently offers about 2% increase in accuracy. Special encoding and hence shallow linguistic processing provide the most significant improvement in accuracy.

4.2.5 Incorporation of Domain–Specificity and Special Encoding

As discussed above, both domain-specific training and special encoding raise the accuracy of transcription. The last set of tests deals with the integration of the two features. Special encoding

is utilized in the training and testing data of DC which have 0.85 and 0.20 million characters respectively.

Prototypes	CAT_{VA2}	CAT_{VA3}
Training/Testing Data	DC	DC
S. Encoding	Applied	Applied
Accuracy	**95.4%**	**96.2%**

Table 5. Integration of D. Specificity and S. Encoding

Recall that Domain-Specificity and Special Encoding each offers 2% improvement. Table 5 shows that combining BOTH features offer about 3% improvement over tests without them. (See non-domain-specific tests in Section 4.2.3)

The 96.2% accuracy achieved by CAT_{VA3} represents the best performance of our system. The result is comparable with other relevant advanced systems for speech to text conversion. For example, Lee (1999) reported 94% accuracy in a Chinese speech to text transcription system under development with very large training corpus.

5. Conclusion

We have created a Cantonese Chinese CAT system which uses the phonologically-based stenograph machine. The system delivers encouragingly accurate transcription in a language which has many homophonous characters. To resolve problematic ambiguity in the conversion from a phonologically-based code to the logographic Chinese characters, we made use of the *N*-gram statistical model. The Viterbi algorithm has enabled us to identify the most probable sequence of characters from the sets of possible homophonous characters. With the additional use of special encoding and domain-specific training, the Cantonese CAT system has attained 96% transcription accuracy. The success of the Jurilinguistic Engineering project can further enhance the efforts by the Hong Kong Judiciary to conduct trials in the language of the majority population. Further improvement to the system will include (i) more domain-specific training and testing across different case types, (2) fine-tuning for the optimal weights in the trigram formula, and (3) optimizing the balance between training corpus size and shallow linguistic processing.

Acknowledgement

Support for the research reported here is provided mainly through the Research Grants Council of Hong Kong under Competitive Earmarked Research Grant (CERG) No. 9040326.

References

Hong Kong Government. 1999. *Hong Kong Supplementary Character Set*. Information Technology Services Department & Official Languages Agency.

Jelinek, F. 1990. "Self-organized Language Modeling for Speech Recognition." In A. Waibel and K.F. Lee, (eds.). *Readings in Speech Recognition*. San Mateo: CA: Morgan Kaufmann Publishers.

Lee, K. F. 1999. "Towards a Multimedia, Multimodal, Multilingual Computer." Paper presented on behalf of Microsoft Research Institute, China in the 5th Natural Language Processing Pacific Rim Symposium held in Beijing, China, November 5-7, 1999.

Lun, S., K. K. Sin, B. K. T'sou and T. A. Cheng. 1997. "Diannao Fuzhu Yueyu Suji Fangan." (The Cantonese Shorthand System for Computer-Aided Transcription) (in Chinese) *Proceedings of the 5th International Conference on Cantonese and Other Yue Dialects*. B. H. Zhan (ed). Guangzhou: Jinan University Press. pp. 217—227.

Sin, K. K. and B. K. T'sou. 1994. "Hong Kong Courtroom Language: Some Issues on Linguistics and Language Technology." Paper presented at the Third International Conference on Chinese Linguistics. Hong Kong.

T'sou, B. K. 1976. "Homophony and Internal Change in Chinese." *Computational Analyses of Asian and African Languages* 3, 67—86.

T'sou, B. K. 1993. "Some Issues on Law and Language in the Hong Kong Special Administrative Region (HKSAR) of China." *Language, Law and Equality: Proceedings of the 3rd International Conference of the International Academy of Language Law (IALL)*. K. Prinsloo et al. Pretoria (eds.): University of South Africa. pp. 314-331.

T'sou, B. K., H. L. Lin, G. Liu, T. Chan, J. Hu, C. H. Chew, and J. K. P. Tse. 1997. "A Synchronous Chinese Language Corpus from Different Speech Communities: Construction and Applications." *Computational Linguistics and Chinese Language Processing* 2: 91—104.

Viterbi, A. J. 1967. "Error Bounds for Convolution Codes and an Asymptotically Optimal Decoding Algorithm." *IEEE Transactions on Information Theory* 13: 260—269.