Statistically-based Model for Computer-Aided Transcription Application

Benjamin K. Tsou, Tom B. Y. Lai, Samuel W. K. Chan,
Lawrence Y. L. Cheung, K. T. Ko, Gary K. K. Chan

Language Information Sciences Research Centre,
City University of Hong Kong,
Tat Chee Avenue, Kowloon,
Hong Kong SAR, China
Fax: (852) 2788-9734

Email: rlbtsou@uxmail.cityu.edu.hk

Abstract

The recent implementation of bilingualism in the Common Law system in Hong Kong has brought about an urgent need to develop a Computer-Aided Transcription (CAT) system to efficiently produce verbatim records of court proceedings conducted in Cantonese Chinese. The Cantonese Chinese CAT system essentially converts phonologically-based shorthand code, or stenograph code, into orthographic representation in Chinese characters. One big challenge in our development of a Cantonese Chinese CAT system is the ambiguity resolution for homophonous Chinese characters that share identical stenograph code. To solve the problem, the bigram model is used as the language model. We implemented the Viterbi algorithm to efficiently compute the most likely Chinese character string for each sequence of stenograph code input. The CAT system is trained with a 0.85 million character corpus. By incorporating enhancement features such as earmarked treatment of numerals, special encoding and domain-specific transcription, the Cantonese Chinese CAT system achieves as much as 96% transcription accuracy.

Keywords: Computational Linguistics, Text Corpora and Text Encoding

1. Introduction

The British rule in Hong Kong made English the only official language in the legal domain for over a century. It is not until the reversion of sovereignty to China in 1997 that Chinese has also come to enjoy official status in the Judiciary of Hong Kong. Legal bilingualism in Hong Kong has brought on an urgent need to create a Computer-Aided Transcription (CAT) system for Chinese to be on a par with the existing English CAT system. (Lun et al., 1995) A research project is undertaken to develop a Chinese CAT system. The system will enable the efficient maintenance of legally tenable records of bilingual (Chinese and English) court proceedings. Similar to English stenography, our Chinese CAT system is phonologically-based.

The major challenge of the development is to resolve the ambiguity given rise by the homonymy in the conversion of phonological code to Chinese characters. Probabilistic models have been widely applied to resolving ambiguity in natural language processing.

They find applications in areas like part-of-speech tagging (Bahl and Mercer, 1976), speech recognition (Rabiner, 1989; Waibel and Lee, 1990), and word sense disambiguation (Charniak, 1993, DeRose, 1988). This paper will present the application of statistical method to the development of a Chinese CAT System. The Viterbi algorithm (Viterbi, 1967) is employed to find the best solution in disambiguating homophonous Chinese characters for phonologically-based stenograph codes. Supplemented with some special measures, the system can achieve about 96% accuracy in the conversion of the stenograph codes to Chinese characters. The transcription system demonstrates the use of text corpora and statistical models in solving computational linguistic problems.

In this paper, we will outline the design of Cantonese Chinese CAT system. The next section briefly introduces the major parts of a CAT system. Section 3 describes the bigram statistical model for code conversion in automatic transcription. In Section 4, we will discuss three measures that help improve the transcription accuracy. Lastly, Section 5 summarizes our findings.

2. Overview of Computer Aided Transcription (CAT)

Three major components can be identified in a CAT system, as shown in Figure 1. First, the stenographer encodes speech, i.e. a sequence of syllables[i], into a sequence of stenograph code, or shorthand code, *simultaneously* when the litigant is speaking. Each stenograph code basically stands for a syllable. Then the sequence of code $\{s_1, \ldots, s_n\}$ is fed into the Computer Transcription System (CTS) to recover the original text $\{c_1, \ldots, c_n\}$. Lastly, some post-editing is needed to correct errors originated from typing mistakes or mis-transcription.

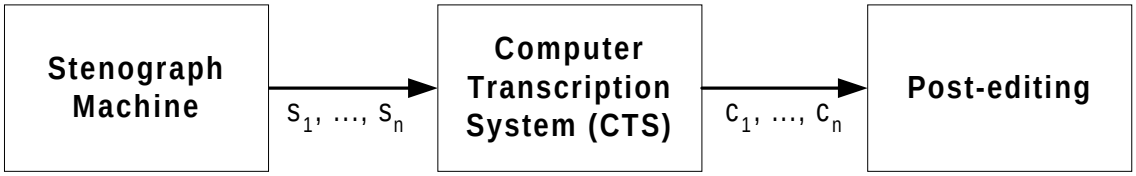

Figure 1 Three major components in CAT

The present paper will focus on CTS which concerns the conversion of stenograph code (representing a sequence of syllables) into Chinese characters. Our stenograph system is built on the basis of the existing CAT system for English so that the existing contingent of court stenographers can switch from one system to the other easily to produce the appropriate legal proceedings. However, the language differences mandate the use of more sophisticated techniques in the design of the Chinese CTS.

Chinese is an ideographic language. Each character represents one syllable. Very often, many Chinese characters may share the same pronunciation, and thus be denoted by the same stenograph code. While the total inventory of Cantonese syllabic types is about 720 (based on the Jyutping Romanization system of the Linguistic Society of Hong Kong (1997)), there are

[i] In our Cantonese Chinese CAT system, the syllables are represented using Jyutping romanization scheme. See Linguistic Society of Hong Kong (1997).

at least 10,000 Chinese character types. Naturally the magnitude of the homocode problem can vary across different domains of application. We tried to estimate this for the legal domain with reference to a 0.85-million character corpus comprising of court proceedings. About 600 syllabic types were found to be used. However, these types have already been responsible for the pronunciations of over 2,500 character types found in the same corpus. Therefore each syllabic type on the average can be shared by up to 4 homophonous character types. In the extreme cases, 27 characters are found to be homophonous. The homocode problem thus is indeed very serious. If ambiguity is not properly resolved, the stenograph code will easily be mis-interpreted and the conversion will generate a large amount of errors. The statistical method is employed to intelligently select the most probable character out of a homophonous set for each shorthand code in context.

3. Statistical Approach to Ambiguity Resolution

3.1 Bigram Model

We apply the well-known bigram model to determine the best character sequence $\{c_1, \ldots, c_k\}$ given the input stenograph code sequence $\{s_1, \ldots, s_k\}$. In statistical terms, (1) should be maximized.

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

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

As huge amount of data is needed to generate reliable statistical estimates for (1), we may find approximation of (1) when certain assumptions are made. First, rewrite (1) as (2) using Bayes' rule.

(2) $$\frac{PROB(c_1,...,c_k) * PROB(s_1,...,s_k \mid c_1,...,c_k)}{PROB(s_1,...,s_k)}$$

As the value of $PROB\ (s_1, \ldots, s_k)$ is the same for any $\{ c_1, \ldots, c_k \}$, the issue now is to maximize the numerator in (2), i.e. (3).

(3) $$PROB\ (c_1,\ldots, c_k) * PROB\ (s_i ,\ldots,s_k| c_1,\ldots,c_k)$$

By making two more assumptions[ii], (3) can be approximated by (4).

(4) $$\prod_{i=1,\ldots,k} (PROB(c_i| c_{i-1}) * PROB(s_i \mid c_i))$$

The advantage of this approximation is that $PROB(s_i \mid c_i)$ and the bigram probability $PROB(c_i| c_{i-1})$ can be readily computed from a training corpus. The Viterbi algorithm is implemented to efficiently compute the maximum value of (4) for different choices of character sequence.

[ii] **Assumption 1: (Bi-gram model)** Using the bigram model, the expression $PROB(c_1, \ldots, c_k)$ can be approximated by the product of all conditional probabilities of a character and the previous character, i.e. c_i and c_{i-1} $\prod_{i=1,\ldots,k} PROB(c_i|c_{i-1})$.

Assumption 2: (Independence of Pronunciation) The pronunciation of c_i (as represented by s_i) is independent of that of the preceding or succeeding members in $c_1, \ldots, c_k$. Accordingly, the expression $PROB\ (s_1, \ldots, s_k| c_1,\ldots, c_k)$ can be approximated by $\prod_{i=1,\ldots,k} PROB(s_i|c_i)$.

3.2 Evaluation Prototypes

To evalute the system, we conducted some experiments to simulate the actual transcription processing using two prototypes. The first prototype, **CAT_{VA}**, implements the Viterbi algorithm for converting stenograph code into the Chinese text. A second prototype, **CAT_0**, (for control purpose) does not implement the Viterbi algorithm. Instead, it uses the crude method of conversion simply by selecting the character, within the homophonous set, that has the highest occurrence frequency. In this way, the improvement gained from the Viterbi algorithm can be estimated.

To prepare for the simulation, authentic Chinese court proceedings were obtained from the Hong Kong Judiciary. Since both prototypes require training, a training corpus of about 0.85 million characters was compiled. The corpus consists of Chinese characters along with the corresponding stenograph codes. In CAT_{VA} training, the bigram co-occurrence probabilities between characters in the texts were calculated. In CAT_0, the character with the highest frequency was calculated for each syllable type/stenograph code. A testing corpus of about 0.15 million characters was also set up. The Chinese characters were converted into the corresponding stenograph code. The test data thus consists solely of stenograph codes, simulating the input by stenographers. In our expreriments, the trained prototypes were used to convert the testing corpus which consists solely of stenograph code into the Chinese text.

To measure the transcription accuracy, the transcribed text from each prototype and the original text were compared. Each transcribed character was checked against the character in the original text. The number of identical character pairs were calculated. CAT_0 and CAT_{VA} achieve an accuaracy of about **78.0%** and **92.4%** respectively. The application of the statistical method in CAT_{VA} offers about 14% gain in accuracy over CAT_0.

4. Improving the Transcription Accuracy

The statistical approach to ambiguity resolution itself does not guarantee 100% accuracy. To boost the accuracy further, three measures are taken, namely, special coding scheme derived from error analysis, ear-marked treatment of numerals, and domain-specific transcription. In the subsequent tests, the same transcription engine, CAT_{VA}, is used. However, different measures are adopted to improve the accuracy rate.

4.1 Special Encoding Scheme

An error analysis was conducted to investigate the possible causes of the mis-transcribed characters. The study revealed that a noticeable amount of errors were due to high failure rate in retrieving some characters in the transcription. It was found that the ambiguity resolution was poor even when the statistical method was employed. The reason is that some characters have an exceptionally high frequency. Their stenograph codes are shared by a number of homophonous characters. Whenever these stenograph codes are encountered, the program will usually output the character with an exceptionally high frequency because of their high rate of occurrence in the training data. The exceptionally high frequency character interferes the correct retrieval of other characters sharing the same stenograph code. For example, in

Cantonese, *hai* ("to be") and *hai* ("at") are homophonous in terms of segmental makeup and thus have identical stenograph code. Their absolute frequencies in our training corpus are 8,695 and 1,614 respectively. Due to the large discrepancy in frequency, the latter is mis-transcribed as the former 44% of the times.

To minimize the interference, we encoded 32 such high frequency characters that consistently produce interference with separate unique stenograph codes. Although this kind of exceptional encoding deviates from the phonologically-based scheme, court stenographers can generally handle the small set with ease. Applying the special encoding scheme to CAT_{VA}, we set up a third prototype, $\mathbf{CAT_{VA,SE}}$. $CAT_{VA,SE}$ offers about 2% increase in accuracy over CAT_{VA}, resulting in **95.03%** accuracy.

4.2 Domain-specific Transcription

The second measure to raise the accuracy is domain-specific transcription. In English Stenograph CAT system, automatic transcription is supported by special "Job dictionaries." These domain-specific dictionaries contain professional vocabularies that are used in cases of similar types. At transcription time, they can be dynamically activated depending on the type of the case recorded. In our analysis of the authentic legal transcripts, we also note that different case types have specific legal terms or usage that may not be frequently found in other categories. For instance, chemical vocabulary that is frequently found in drug-trafficking cases will certainly not be as frequent in transcripts related to fraud or traffic offences. A similar measure is built into our statistically-based Chinese CAT system. The advantage of capitalizing on the vocabulary differences across various domains is even more apparent in our system. Integrating all vocabularies in a single training corpus may obscure the co-occurrence probabilities of some characters that frequently occur in some domains but not the others. Limiting the training to a confined set of vocabulary helps lower the ambiguity of the stenograph code. The statistical data obtained will model the language of the domain better.

To model "Job dictionaries" in our Chinese CAT system, we have exploited this domain-specificity of the lexical items found in the legal transcripts. A test has been conducted. In the previous tests, the training and testing corpora are drawn from court transcripts of various case types, e.g. traffic, assault, and robbery. In the new test, we compiled another set of the training and testing corpora which consist solely of transcripts related to the "Traffic" case type. The sizes of these corpora are the same as those in the previous prototypes for comparison. The fourth prototype, $\mathbf{CAT_{VA,\,TR}}$, is identical to CAT_{VA}. The prototype gives an accuracy of 94.54%. At present the other categories including *Traffic*, *Assault*, and *Robbery* are being compiled. More domains will be added in the near future.

4.3 Ear-marked Treatment of Numerals

In the original English stenography, numerals, e.g. 1998, 250000, 0.2, are encoded using the numerals themselves instead of phonologically-based stenograph code. Special numeric keys can be found on the stenograph keyboard for input. As a result, each numeral type represents a distinct stenograph code. However, this representation prevents us from capturing some regularities of numerals. For example, in Chinese, the numeral is often followed by a set of classifiers or units of measurement, e.g. dollar, metre. If each numeral type is represented by

a distinct code, this regularity can hardly be captured by the bigram probabilities as the occurrences of individual numeral types, say, 651, 23, etc, can be pretty low even in a comparatively large corpus. The negative impact of the problem is that any numerals that do not appear in the training data will become a new code.

To improve the transcription of the code adjacent to previously unseen numeral code, we treat all numerals as one single category. Instead of representing each numeral types using a distinct code, a special symbol, "NUM", is introduced to serve as the stenograph code for all numeral types. Each time when the training procedure encounters a numeral type, the frequency and the co-occurrence probability of the NUM entries and its adjacent code will be updated. In this way, we can track the probability of the co-occurrence of the numerals (instead of individual numeral) with other Chinese characters. An experiment was done by incorporating the feature on top of the Viterbi algorithm. We call this prototype $\mathbf{CAT_{VA,NUM}}$. In the initial testing, there was noticeable improvement of the accuracy. However, as we scaled up the training size as in the previous test, the accuracy is raised only by 0.04% to **92.42%**. The net improvement is small. There are two reasons for this. First the occurrence of numerals is relatively small in the testing corpus (0.6% in 0.2 million character). Second, we found that the majority of the numeral types are the ten one-digit numerals (e.g. 1, 2, 3, etc.) used for enumeration in both the training and testing data. Their frequencies are quite high. As a result, their co-occurrence probabilities with other codes are taken care of by the training set. However, the treatment will still be useful for preventing the potential negative impact on transcription we mentioned for unanticipated novel numerals found in new test data. We are still conducting experiments in this connection.

4.4 Combing All the Measures

It is evident from the tests mentioned that of the three measures, special encoding and domain-specific processing offer larger improvement. To push the accuracy even further, the three measures discussed in Section 4.1 to 4.3 are employed simultaneously to form the fifth prototype, $\mathbf{CAT_{ALL}}$. The "Traffic" training and testing corpora are processed with special encoding and numeral treatment. CAT_{ALL} achieves **96.73%** accuracy.

5. Conclusion

To summarize, we created a unique Chinese CAT system that adapts the phonologically-based stenograph machine and its encoding scheme originally designed for Indo-European languages. The system has managed to produce accurate transcription in a language which has many homophonous characters. The success makes it possible for the court stenographer to operate in Chinese and English without re-creating a different stenograph keyboard and input scheme, facilitating the implementation of legal bilingualism in the Judiciary of Hong Kong.

To realize all these, the key issue lies in the resolution of ambiguity given rise during the conversion from phonologically-based code to the Chinese characters. The bigram statistical model has been applied to select, using the Viterbi algorithm, the most probable sequence of characters out of the homophonous character sets, a critical issue in the design of CAT for a basically monosyllabic language. With additional measures such as special encoding and domain-specific processing, the Chinese CAT system has attained about 96% transcription

accuracy. The results also have significant implications to finding a good solution for inputting Chinese characters using phonologically-based romanization on computers.

References

Bahl, L. R. and R. L. Mercer. (1976) "Part of Speech Assignment by a Statistical Algorithm." IEEE International Symposium on Information Theory, Ronneby, Sweden, June 1976.

Charniak, E. (1993). Statistical Language Learning. Cambridge, MA: MIT Press.

DeRose, S. (1988). Grammatical Category Disambiguation by Statistical Optimization. Computational Linguistics, 14: 31-39.

Linguistic Society of Hong Kong. (1997) Yueyu Pinyin Zibiao (Cantonese Jyutping Transliteration Word List). Hong Kong: Linguistic Society of Hong Kong.

Lun, Suen, K. K. Sin, Benjamin K. Tsou and T. A. Cheng. (1995) Diannao Fuzhu Yueyu Suji Fangan. (The Cantonese Shorthand System for Computer-Aided Transcription)" (in Chinese) Paper presented at the 7th International Conference on Cantonese and Other Yue Dialects.

Rabiner, L. R. (1990) "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." Proceedings of IEEE. Reprinted in Waibel and Lee (1990).

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

Waibel, A., and K. F. Lee (eds.). (1990) Readings in Speech Recognition. San Mateo, CA: Morgan Kaufmann.