6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A COMPLEMENTARY APPROACH TO COMPUTER-AIDED TRANSCRIPTION: SYNERGY OF STATISTICAL-BASED AND KNOWLEDGE DISCOVERY PARADIGMS

Benjamin K. T'sou and Tom B. Y. Lai

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

ABSTRACT

The recent implementation of legal bilingualism necessitates the development of a Chinese Computer-Aided Transcription (CAT) system to produce Chinese court proceedings conducted in Cantonese. The transcription system converts transcription shorthand codes into Chinese text, i.e., from phonetic to textual representation of the language. Cantonese and Mandarin Chinese have many homophonous characters. The main challenge lies in the resolution of the severe ambiguity of the conversion. *N*-gram statistical model is incorporated to estimate the most probable character string during conversion. Domain-specific corpora have been compiled to support the statistical computation. With additional enhancement features, the CAT system delivers a transcription accuracy of 96%. An intelligent error detection tool is built into the system to facilitate the manual correction of the remaining errors. Using decision tree algorithm and a range of text and linguistic attributes, the system can effectively alert the users to possible errors.

Key Words: Speech to Text, Statistical Modelling, Cantonese, Chinese

1. INTRODUCTION

The Common Law system has been retained after the reversion of Hong Kong's sovereignty to China in 1997. English used to be the exclusive language in courtroom. Yet the recent introduction and implementation of bilingualism in its Judiciary has made Chinese an official language in the legal domain. The change has brought about an urgent need for Computer-Aided Transcription (CAT) system to efficiently produce and maintain legally tenable verbatim records of court proceedings conducted in Cantonese (T'sou 1993; Sin and T'sou 1994; Lun et al. 1995). Although Mandarin Chinese CAT system is available, Cantonese is the predominant Chinese dialect used by over 95% of the population in Hong Kong. The substantial linguistic differences between Cantonese and Mandarin (e.g. phonology, morphology, vocabulary and orthography) necessitate the Jurilinguistic Engineering undertaking to develop an independent Cantonese CAT system for the local language environment.

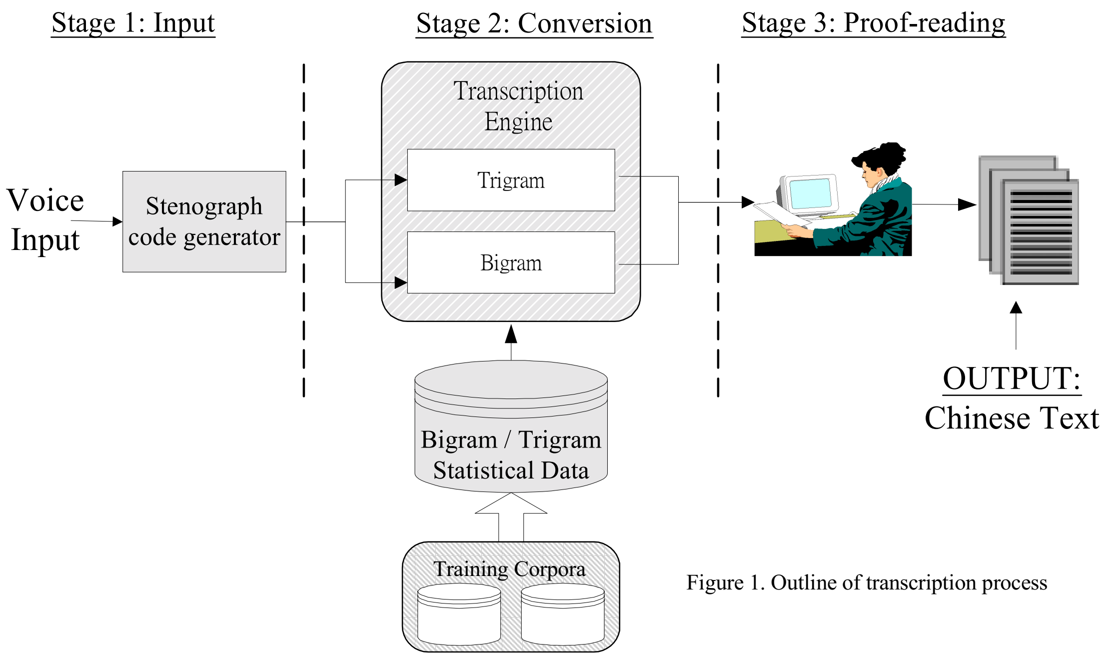

Figure 1. Outline of transcription process

2. Computer-Aided Transcription

The machine-aided transcription process can be divided into three stages, as shown in Figure 1. First the stenographer encodes speech as a string of stenograph codes via a stenograph machine. A transcription software converts the stenograph codes into Chinese character strings. Finally, the output text is proof-read to eliminate transcription errors. To support the CAT system, a Cantonese shorthand input scheme and a transcription software were developed. The **input scheme** is used to encode Cantonese syllables using stenograph codes. It enables stenographers to record Cantonese speech. To capitalize on the existing stenographers' skills in English and to meet the added requirement that both Cantonese and English may be used in court, the Cantonese CAT keyboard is based on the English CAT input scheme with minimal extension. The **transcription software** is the critical component of the system which converts the phonologically-based stenograph codes into Chinese characters.

3. System Architecture

3.1. *N*-gram Statistical Model

The challenge of developing the transcription system lies in ambiguity resolution due to the problematical homonymy in the conversion.

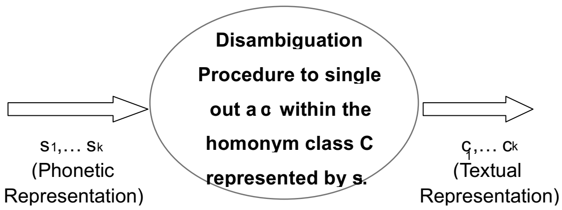

Figure 2. Disambiguation

The transcription module has integrated an *N*-gram statistical language model to compute the best sequence of characters. Here is the statistical formulation for the conversion process. Let $\{s_1, \dots, s_N\}$ be the input stenograph code sequence, and $\{c_1, \dots, c_N\}$ be any character sequence corresponding to $\{s_1, \dots, s_N\}$. The best transcription is the character string that maximizes the conditional probability in (1).

(1) $\mathrm{P}(c_1, \dots, c_N \mid s_1, \dots, s_N)$

By Bayes's theorem, (1) can be rewritten as (2)

(2) $$\frac{\mathrm{P}(c_1,\dots,c_N)\times \mathrm{P}(s_1,\dots,s_N \mid c_1,\dots,c_N)}{\mathrm{P}(s_1,\dots,s_N)}$$

Since $\mathrm{P}(s_1,\dots,s_N)$ remains unchanged for any character sequence, it is necessary to find the string that maximizes (3).

(3) $\mathrm{P}(c_1,\dots, c_k) \times \mathrm{P}(s_1 ,\dots, s_k|c_1,\dots,c_k)$

N-gram probability is rather difficult to estimate when *N* is larger than 3. So (3) is approximated by bigram or trigram in (4) and (5) respectively.

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

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

Trigram statistical model is computationally more expensive than bigram counterpart but is slightly more accurate. Viterbi algorithm (Viterbi 1976) is used for efficient computation. The estimation of bigram and trigram probability is derived from a 0.85 million character training corpus which is drawn from authentic Chinese court proceedings and judgement scripts.

3.2. Enhancement Measures

Some additional enhancement techniques are implemented to raise the overall accuracy. They include special encoding and domain-specific transcription. **Special encoding** refers to the exceptional encoding for some characters that are often mis-transcribed. In our trial tests, error analysis was conducted to identify the set of characters that are often mis-transcribed. 32 characters that contribute to about 25% of all errors are identified and assigned a unique stenograph code instead of phonologically-based stenograph code. **Domain-specificity** also improves the transcription accuracy. The system exploits the vocabulary difference across various domains, e.g. *traffic*, *assault*, and *fraud offences*. Five domain-specific training corpora were prepared for generate the bigram and trigram estimation. Texts of different domains can be modelled with greater accuracy. The transcription engine can activate different sets of domain-specific data in transcription.

4. Results

For evaluation, the design features of the transcription engine have been systematically compared. Here are the results. Three prototypes were produced, namely, CAT_0, CAT_{VA2} and CAT_{VA3}. CAT_{VA2} and CAT_{VA3} are implemented using the bigram and trigram transcription algorithm respectively. Instead of using the *N*-gram model, CAT_0 only selects the highest frequency character out of the homophonous character set for each stenograph code. It serves as a baseline prototype.

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

Table 1. Different *N*-gram Models

Table 1 shows that transcription engines equipped with the N-gram language modelling consistently outperform the baseline prototype.

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

With the help of enhancement features, the bigram and trigram implementations of the transcription engine achieve 95.4% and 96.2% accuracy respectively. The system performance is comparable with, if not better than, other advanced Romanized Chinese Speech-to-Text input applications under development.

5. Intelligent Error Detection Aid

An intelligent error detection aid is designed to help pick out the mis-transcribed characters in the proof-reading stage. Various techniques have been described previously to raise the transcription accuracy to 96%. However, there remain ~4% persistent errors in the output of the automatic transcription. Because of the stringent requirement for the accuracy of recording, these errors are corrected manually in post-editing stage of court proceedings production. The majority of the errors are erroneous selection of homophonous characters in automatic transcription. Spotting the errors in the text manually is a painstaking task. The error detection tool facilitates the identification of potential errors.

The error detection procedure accepts a stream of Chinese characters $\{c_1, \ldots, c_n\}$, and determines whether the occurrence of character c_i where $1 \leq i \leq n$ is considered well-formed or ill-formed in the context $\{c_1, \ldots, c_n\}$. In our study, $\{c_1, \ldots, c_n\}$ is delimited by punctuation marks. The detection tool alerts the user when potential errors are found. The user can then decide to make corrections.

5.1 Decision Tree Algorithm

The novel approach of using C4.5 decision tree algorithm (Quinlan 1993) is applied to the intelligent detection of errors. C4.5 learning algorithm is a common data mining technique that has been applied in various text processing applications such as finding associations in a collection of texts (Feldman and Hirsh 1997) and mining online text (Knight 1999).

A decision tree is a tree structure where at each node a test on a particular attribute of the data is performed, and where the leaves corresponds to a particular class. The path from the root node to a particular leaf is then a series of tests on the attributes that classifies the data to the class defined by the particular leaf. C4.5 is the most used algorithm for inducing decision trees. Using the information theory approach, an entropy criterion selects the most informative or discriminative feature through training. The best attribute is selected under a statistical property, called information gain, for the testing at any point in the tree. It measures how well a given attribute differentiates the training examples according to their target classificatory scheme and to select the most suitable candidate attribute at each step while expanding the tree.

The C4.5 machine learning algorithm has been employed to identify possible mis-transcribed characters in the output text. Twelve attributes such as word segmentation information, frequency and part of speech of words are used to provide contextual information for distinguishing between correct and incorrect characters. Through training, decision trees are built, and rules are acquired using the transcription output. With the rules, the error detection procedure can be applied to determine whether a given character is considered well-formed or ill-formed in the linguistic context.

In the experiment, the context window size of each string is restricted to 5 words. The words are labelled as follows: $\{W_{M2}, W_{M1}, W_0, W_{P1}, W_{P2}\}$. W_0 is the word containing the problematic character. W_{M2}, W_{M1}, W_{P1} and W_{P2} are the context of W_0. The training data consist of a set of instances (5-word strings) with the corresponding contextual attribute values. A set of attributes are defined for each string. The attributes can be classified into two categories: (1) Information Retrieval-based attributes and (2) Linguistically-based attributes.

Information Retrieval-based

- Bigram frequency between neighbouring words
- Character and word frequency
- Word segmentation property

Linguistically-based

- Presence of bound morpheme as a word (implies incorrect character selection)
- Part-Of-Speech (POS) of neighbouring words

The rules acquired by C4.5 decision tree on average can classify the training strings into well-formed or ill-formed classes with an accuracy of 93%. Among these attributes, we found that word bigram frequencies between the problematic word and the immediately preceding and following words are far more dominant factors affecting the classification. In other words, the word collocation information makes important contribution to determine the well-formedness of characters, and the results are encouraging. The capability of classification will significantly facilitate the identification of the remaining ~4% errors.

6. Conclusion

The creation of the Chinese CAT system has laid a cornerstone for efficient maintenance of Chinese court proceedings in Hong Kong. The success of the Jurilinguistic Engineering application can further enhance the efforts by the Hong Kong Judiciary to conduct trials in the language of the majority population, Cantonese Chinese. The system has brought in a phonetically-based input solution for Chinese CAT. The critical component in the system lies in the ambiguity resolution of homonyms during CAT code to text conversion. Using the *N*-gram statistical language model, the system enables the identification of the most probable sequence of characters from the sets of possible homophonous characters. With the use of trigram implementation, special encoding and domain-specific training, the Cantonese CAT system delivers up to 96% transcription accuracy. The unique feature of the system is the integration of an error detection aid for identifying potential errors in the transcription output. Using C4.5 decision tree algorithm and various statistical and linguistic attributes, the system can effectively locate the remaining 4% errors from the transcribed text for final proof-reading. The blending of transcription and text mining technology makes the CAT system an ideal partner for transcription task.

7. References

Feldman R. and H. Hirsh. 1997. “Finding associations in collections of text.” In R.S. Michalski, I. Bratko and M. Kubat. (eds.), *Machine Learning and Data Mining: Methods and Applications*, pp. 224-240. Wiley.

Knight K. 1999. “Mining online text.” *Communications of the ACM* 42(11): 58-61.

Lun, S., K. K. Sin, B. K. T'sou, and T. A. Cheng. 1995. “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.

Quinlan, J. R. 1993. *C4.5 Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann.

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. 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)*. (eds.) K. Prinsloo et al. Pretoria: University of South Africa. pp. 314—331.

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