Court Stenography-To-Text ("STT") in Hong Kong:

A Jurilinguistic Engineering Effort

Benjamin K. Tsou[1], Tom B.Y. Lai[2], K.K. Sin[2], Lawrence Y.L. Cheung[3]

[1] Language Information Sciences Research Center, City University of Hong Kong
[2] Department of Chinese, Translation and Linguistics, City University of Hong Kong
[3] Department of Linguistics, University of California, Los Angeles
{rlbtsou,cttomlai,ctsinkk}@cityu.edu.hk, yllc@ucla.edu

Abstract

Implementation of legal bilingualism in Hong Kong after 1997 has necessitated the production of voluminous and extensive court proceedings and judgments in both Chinese and English. For the former, Cantonese, a dialect of Chinese, is the home language of more than 90% of the population in Hong Kong and so used in the courts. To record speech in Cantonese verbatim, a Chinese Computer-Aided Transcription system has been developed. The transcription system converts stenographic codes into Chinese text, i.e. from phonetic to orthographic representation of the language. The main challenge lies in the resolution of the sever ambiguity resulting from homocode problems in the conversion process. Cantonese Chinese is typified by problematic homonymy, which presents serious challenges. The *N*-gram statistical model is employed to estimate the most probable character string of the input transcription codes. Domain-specific corpora have been compiled to support the statistical computation. To improve accuracy, scalable techniques such as domain-specific transcription and special encoding are used. Put together, these techniques deliver 96% transcription accuracy.

1. Introduction

Baum (1972) and Bahl et al. (1983) laid the groundwork of Markov-chain-based, hidden or otherwise, language modeling. With the availability of large corpora, language-model techniques have been used with success in various language engineering tasks, e.g. part-of-speech tagging, (DeRose, 1988; Dematas and Kokkinakis, 1995), error correction (Kukich, 1992; Mays et al., 1991) and speech-to-text conversion (Deroualt and Merialdo, 1986). We have worked with the Judiciary of Hong Kong on computer-aided transcription of stenographic records of court proceedings into orthographic text. We use trigram language modeling in the automatic conversion process. In the case of phonetic Chinese input methods, 96% accuracy has been reported (Lee, 1999). However, word boundaries, and tones, are not marked in stenography. Acute homonymy problems in colloquial Cantonese have made it difficult to achieve high accuracy in our automatic conversion module. We use domain-specific corpora, annotated but of modest sizes, and assign special stenographic codes to critical characters to overcome this difficulty. We have also studied how to identify and correct potential conversion errors in the post-editing process.

2. Court Stenography-To-Text ("STT") Jurilinguistic Engineering

2.1. Cantonese Stenograph Code to Character Conversion

Bristish rule in Hong Kong made English the only official language in the legal domain for over a century. After reversion of sovereignty to China in 1997, legal bilingualism brought on an urgent need to create a computer-aided transcription system for the Cantonese-speaking majority (Tsou, 1993; Tsou et al. 2000). With the support of the Judiciary, we have developed a transcription system for converting Cantonese stenographic code to Chinese characters.

2.2. Language Model of the Automatic Conversion Module

Speech to text conversion is ambiguous. We use the *N*-gram language model to determine the most likely character sequence $<c_1, \ldots, c_k>$ that corresponds to an input stenographic code sequence $<s_1, \ldots, s_k>$ by maximizing the conditional probability

(1) $P(<c_1, \ldots, c_k> | <s_1, \ldots, s_k>)$

Applying Bayes' rule and using using trigram approximation, we maximize

(2) $\prod_i P(c_i \mid c_{i-2}c_{i-1}) \cdot P(s_i \mid c_i)$

We assign weights to the monogram, bigram and trigram components of (2).

We use the Viterbi Algorithm to maximalize (2).

2.3. Domain-specific Training, Special Encoding and Post-editing

As explained below, we use training corpora in different domains like *traffic*, *assault* and *robbery* to obtain domain specific statistics. We assign special stenographic codes to critical characters that occur frequently in court proceedings to improve transcription accuracy.

Errors in the automatic conversion output are manually corrected in a post-editing phase.

3. Acute Homonymy in Cantonese Chinese

Over 90% of the population in Hong Kong are native speakers of Cantonese, a dialect of Chinese. Cantonese and Mandarin Chinese differ considerably in terms of phonology, phonotactics, word morphology, vocabulary and orthography. Mutual intelligibility between these two dialects of Chinese is low. This situation necessitates the Jurilinguistic Engjneering undertaking to develop an independent Cantonse computer-aided transcription (CAT) system for the local language environment.

A major challenge in developing a Cantonese CAT system is the problem of acute homonymy (homophony) in Cantonese, especially colloquial Cantonese. Chinese is a logographic language. Each Chinese character (logograph) represents a syllable. While the total inventory of Cantonese syllable types is about 720, there are at least 14,000 Chinese character types in use. There are thus many homophones in the language (Tsou, 1976).

We (Tsou et al. 2000) have reported that, in a one million character corpus of court proceedings, there are 565 distinct syllable types, 470 of which correspond to multiple homophonous characters. These 2,810 homophonous character types make up 94.7% of the 2,922 character types attested in the corpus.

Homophony in Cantonese Chinese means that Cantonese syllable to character conversion is a one-to-many process. Stenographic code to text transcription is particularly hard hit in this aspect as word boundaries are not recorded.

Measures to compensate for this obstacle are discussed in the next section.

4. Domain-specific Training and Special Encoding

As reported in Tsou et al. (2000), we found increasing the size of the training corpus unfruitful when we reach the 92~93% region:

Training Corpus	0.20	0.35	0.50	0.63	0.73	0.85
Accurcy	89.9	91.2	91.8	92.1	92.3	92.4

Table 1: Effect of training corpus size (million characters) to accuracy (%)

Because of this, we do not try to improve conversion accuracy by using enormous training corpora. Instead, we try to improve conversion accuracy by using different training corpora for different sub-domains and by means of other measure domain-specific measures.

4.1. Domain-specific Training Corpora

As reported in Tsou et al. (2000), using domain-specific training corpora significantly improves stenography to text accuracy:

Domain-specific Training	Not Applied		Applied	
Language Model	Bigram	Trigram	Bigram	Trigram
Training Data	General	General	Specific	Specific
Testing Data	Specific	Specific	Specific	Specific
Accuracy	92.6%	92.8%	94.7%	94.8%

Table 2: Effect of domain-specific training (training corpus sizes 085 mill., testing corpus sizes 0.2 mill.)

In accordance with the findings shown above, we have compiled training corpuses in different sub-domains of sizes 0.85 million ~ 1 million characters on the reckoning that this gives us a 2% improvement over using a general, undifferentiated corpus for training. This approach is practicable for our target text domain of court proceedings.

4.2. Special Encoding for Critical Characters

Examination of errors in automatic stenographic code to character conversion reveals that a small number of critical characters that occur frequently have great effects on conversion accuracy. We have identified 32 such characters and assigned special stenographic codes to them. Tsou et al. (2000) report a 2% improvement in accuracy when this measure is applied:

Special Encoding	Not Applied		Applied	
Language Model	Bigram	Trigram	Bigram	Trigram
Accuracy	92.4%	93.6%	94.7%	95.6%

Table 3: Effect of special encoding (training corpus 0.85 mill. characters, testing corpus 0.20 mill. characters, both not domain-specific)

Professional stenographers with whom we tested the use of specially-coded characters have found this practice acceptable.

4.3. The Combined Effect

When domain specific training and special encoding for critical characters are applied at the same time, their effects may not be additive. Tsou et al. (2000) report the results of an experiment:

Domain-specific Training	Not Applied		Applied	
Special Encoding	Not applied		Applied	
Language Model	Bigram	Trigram	Bigram	Trigram
Training Data	General	General	Specific	Specific
Testing Data	Specific	Specific	Specific	Specific
Accuracy	92.6%	92.8%	95.4%	96.2%

Table 4: Effect of applying domain-specific training and special encoding at the same time (training corpora 0.85 mill. characters, testing corpora 0.20 mill. characters)

We can thus reckon that applying both measures in our task brings us an improvement of 3% in accuracy. With a trigram language model, an accuracy of 96% is in general achieved.

5. Error Identification and Correction

While we have been able to achieve an automatic conversion accuracy of 96%, there is still work left for the human post-editor before the court proceedings transcriptions can be archived for later use. Professional stenographers who work with us have found this acceptable. Nevertheless, we have also giving thought to providing them with further help.

In our work on less-than perfect real-world voice recognition outputs of Cantonese speech, we have developed statistical techniques to identify potential speech-to-text transcription errors and to correct them (Cheung et al., 2003). We tested two speech recognition products available on the market and found that they could only achieve accuracy rates of 60~70% when working on colloquial Cantonese court proceedings. Conceiving of a post-processing module as shown in Figure 1, we studied statistical techniques to generate potential error candidates and to correct errors.

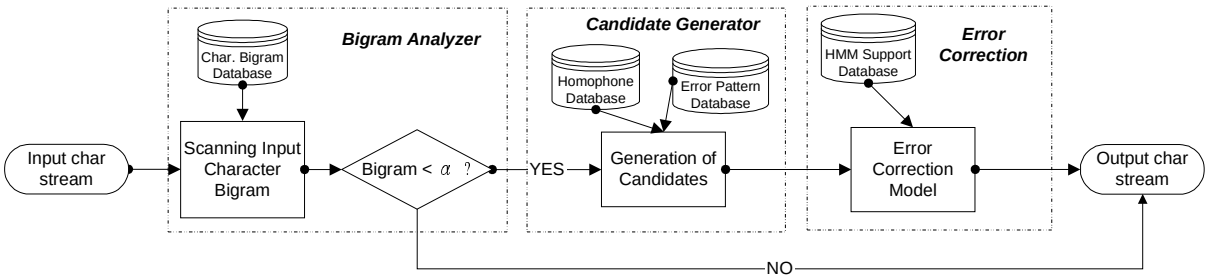

Figure 1: Architecture of speech recognizer post-processing

5.1. Potential Error Identification

We (Cheung et al., 2003) propose to use a "bigram measure" to identify potential errors. The "bigram measure" value of the juncture between characters c_{n-1} and c_n is defined as:

$$(3)\ BG(c_{n-1}, c_n) = \frac{count(c_{n-1}, c_n)}{count(c_{n-1}) \times count(c_n)}$$

If $BG(c_{n-1}, c_n)$ is smaller than a threshold value, c_n will be considered to be a potential error.

5.2. The Bigram Viterbi Disambiguator

5.2.1. Replacement Candidates

First, we need to form a set of possible replacements for each of the potential errors.

C*onfusion* (Kukich, 1992) is defined as follows:

$$(4)\ pr((c_i',t_i')|(c_i,t_i)) \times \left[1 - \frac{1}{\sqrt{\text{count}((c_i,t_i),(c_i',t_i'))}}\right]$$

$pr((c_i',t_i')|(c_i,t_i))$ is the probability that character c_i' with transcription t_i' is incorrectly recognized as character c_i with transcription t_i. It gives us a measure of how likely a character seen in the speech recognition output should indeed be another character. The square-bracketed multiplier in (4) is to discount infrequent confused patterns.

Characters that give high enough confusion probability with a potential error are put into a replacement candidate set. We (Cheung et al., 2003) form a candidate set of homophonesand a candidate set of sound-alikes for each potential error and then pass the union of the two sets to the bigram disambiguator.

5.2.2. Making Decisions

Our (Cheung et al. 2003) Viterbi disambiguator uses the *confusion* probabilities in its language model.

5.3. Error Identification for STT Post-editing

The error identification and correction techniques described above are for use with speech recognizers. Stenography to text conversion has a much higher accuracy rate than speech recognizers. Our professional stenographer partners actually find it acceptable to have 5% errors left for post-editing.

Nevertheless, we are giving thought to adapting the potential error identification technique described above for our stenograph to text system.

While we (Cheung et al., 2003) report achieving an accuracy improvement from 69.9% to 73.7% for speech recognizers, it should be noted that our percentage of correct replacement decisions is only about 40~45%. On the other hand, the recall and precision rates of potential error identification are both around 65%. This suggests that it may be worthwhile to add a potential error identification capability to our STT system.

Our study of speech recognizer errors (Cheung et al., 2003) shows that only 30% of the errors involve homophones:

Type	Homonyms	Sound-alikes	Others
%	30.8	30.5	34.2

Table 5: Error types of speech recognizers

While stenographers do make errors of a "typo" kind, which can be compared to sound-alike errors in speech recognition, we reckon that most of the errors of our STT system should be caused by homonymy. The usefulness, or otherwise, of a module for identifying potential conversion errors for the attention of human post-editors must be established independently.

6. Conclusion

Our experience with court stenography-to-text conversion shows that problematic homonymy in colloquial Cantonese Chinese is an obstacle in the development of highly accurate automatic stenography-to-text transcription. We have been able to achieve 96% accuracy by training trigams with modestly-sized domain-specific corpora with special stenographic codes assigned to critical elements. It would also be logical to apply statistical error-identification techniques that we have developed for use with less-than-perfect real-world speech recognition to identify potential conversion errors for the attention of human post-editors.

7. Acknowledgements

This study was supported by a Competitive Earmarked Research Grant (CERG) from the Research Grant Council of Hong Kong (CityU 1238/00H).

8. References

Damerau, F. (1964). A Technique for the Computer Detection and Correction of Spelling Errors. *Communications of the ACM*, 7(3), pp. 171-176.

Bahl, L.R., Jelinek, F. and Mercer, R.L. (1983). A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5, pp. 179-190.

Baum, L.E. 91972) An Inequality and Associated Maximization Technique in Statistical Estimation of a Markov Process. *Inequalities*, 391), pp. 1-8.

Cheung, L.Y.L, Tsou, B.K. and Lai, T.B.Y. (2003). Error Identification and Correction in Chinese Speech Input Post-processing. In Proceedings of Oriental COCOSDA 2003, Singapore: COLIPS Publications, pp. 233-240.

DeRose, S.J. (1988). Grammatical Category Disambiguation by Statistical Optimization. *Computational Linguistics*, 14(1), pp. 31-39.

Dematas, E. and Kokkinakis, G. (1995). Automatic Stochastic Tagging of Natural Language Texts. *Computational Linguistics*, 21(2), pp. 137-164.

Deronalt, A.M. and Merialdo, B. (1986). Natural Language Modeling for Phoneme-to-Text Transcription. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8, pp. 742-749.

Kukich, K. (1992). Techniques for Automatically Correcting Words in Text. *ACM Computing Surveys*, 24(4), pp. 377-439.

Lee, K.F. (1999). Towards a Multimedia, Multimodal, Multilingual Computer. Paper presented on behalf of Microsoft Research China at the 5th Natural Language Processing Pacific Rim Symposium, Beijing.

Mays, E., Demerau, F. and Mercer, R. (1991). Context Based Spelling Correction. *Information Processing and Management*, 27(5), pp. 517-522.

Tsou, B.K. (1976) Homophony and Internal Change in Chinese. Computational Analysis of Asian and African Languages, 3, pp. 67-86.

Tsou, B.K. (1993). Some Issues on Law and Language in the Hong Kong Special Adminstrative Region (HKSAR) of China. In K. Prinsloo at al. (Eds.), *Language, Law and Equality: Proceedings of the 3rd International Conference of the International Academy of Language Law (IALL*), University of South Africa, pp. 314-331.

Tsou, B.K., Sin, K.K. Chan, S.W.K., Lai, T.B.Y., Lun, C. Ko, K.T., Chan G.K.K. and Cheung, L.Y.L. (2000). Jurilinguistic Engineering in Cantonese Chinese: An *N*-gram-based Speech to Text Conversion System. In *Proceedings of COLING2000*, Universitaet des Saarlandes, Saarbruecken, Germany, pp. 1121-1125.

Tsou, B.K., Tsoi, W.F., Lai, T.B.Y., Hu, J. and Chan, S.W.K. (2000). LIVAC, A Chinese Synchronous Corpus, and Some Applications. In *Proceedings of the ICCLC International Conference on Chinese Language Computing*, Chicago, pp. 233-238, http://livac.org.

Weischedel, R.M., Meteer, M., Schartz, R., Ramshaw, L. and Palmucci, J. (1993). Coping with Ambiguity and Unknown Words through Probabilistic Models. *Computational Linguistics*, 19(2), pp. 359-382.

Xiang, T. and Evans, D. (1996). A Statistical Approach to Automatic OCR Error Correction in Context. In Proceedings of the Fourth Workshop on Very Large Corpora, Copenhagen, Denmark, pp. 88-100.