Post-processing of Handwritten Pitman's Shorthand Using Unigram and Heuristic Approaches

Swe Myo Htwe¹, Colin Higgins¹, Graham Leedham², and Ma Yang²

¹ The University of Nottingham, School of Computer Science and IT,
Wallaton Road, Nottingham, NG8 1BB, UK
{smh, cah}@cs.nott.ac.uk

² Nanyang Technology University, School of Computer Engineering,
N4-#2C-77 Nanyang Avenue, Singapore 639798
{asgleedham, pg01874827}@ntu.edu.sg

Abstract. The computer transcription of handwritten Pitman's shorthand has enormous potential as a means of rapid text entry to today's handheld devices. Recognition errors caused in pattern segmentation and classification raises the incidence of ambiguous interpretation in existing systems and the paper proposes a well-established unigram technique and an efficient heuristic method to reduce ambiguity in a linguistic post processor. Heuristics applied in our transcription system are: - firstly, incorporating visual stimulus as used by human readers; secondly, applying knowledge of the most common words of Pitman shorthand; and finally, adding knowledge of collocation. An experiment using a phonetic Lexicon of 5000 entries shows the distribution of ambiguity in a shorthand lexicon due to the similarity of outlines' and estimates the transcription accuracy of 94%.

1 Introduction

Pitman shorthand is a speech-recording medium practiced at a practical rate of about 120-180 words per minute. It records speech phonetically as shown in Figure 1, and transliteration of phoneme primitives into orthographic target script relates to techniques applied to speech recognition systems such as Hidden Markov Model (HMM) and statistical language modeling techniques, which are used to convert a given utterance to the most likely word.

Fig. 1. Samples of Pitman shorthand notation and an illustration of ambiguous interpretations for a single input outline.

The potential of Pitman shorthand [1] as a means of rapid pen-driven text entry to a computer was firstly reported in the 1980's. Later works by Qiao & Leedham [2][3]

used statistical phonetic models to validate all legal pairs of primitive combination and reduce ambiguous candidates in the final text interpretation. The most recent work by Nagabhushan et al [4] introduced a shorthand specific dictionary-lookup method and concluded that further work is required in homophones (*outlines which are written similarly, but have different interpretations*) resolution area. This is a critical bottleneck in the recognition of human handwritten shorthand and a novel approach of homophone deduction by the use of unigram and contextual knowledge is proposed in this paper.

2 Unigram Approach

In general, homophones are caused by two factors which start in the recognition stage: - firstly, inability to differentiate between thick and thin pen strokes and secondly inability to detect an exact location of a vowel notation against a consonant kernel. These cause a script to be ambiguous and raise the occurrence of homophones.

“*First and Last Segment Algorithm (F&L)*” is a unigram approach used to filter ambiguous outlines in our system. The motive behind this approach is human readers rely on the legibility of the first and last segments of an outline in real text transcription. This finding is based on the transcription of 10 samples of shorthand notes by three professional Pitman writers where 55% of their transcription is done by the legibility of the first and last pattern primitives. The algorithm can be defined as

$$P(x_i) = \frac{P(f_i)P(l_i)}{\sum_{i=1}^n P(f_i)P(l_i)}, \quad (1)$$

where $P(x_i)$ is the normalized probability of the first and last segment and $\{f_1, f_2, \dots, f_n\}$ and $\{l_1, l_2, \dots, l_n\}$ denotes the first and last pen strokes in terms of n ambiguous candidate outlines.

Another resolution to filter ambiguous outlines is the use of “*Anchor Node (AN) Algorithm*” in which a segment with the highest accuracy is considered as an anchor node and the search is based on local variables i.e., descendant primitives of the anchor segment. The algorithm can be defined as

$$P(y) = \max (P(a_i), \quad (2)$$

where $P(y)$ is the probability of an anchor segment and $(a_1, a_2, \dots, a_n)$ denotes primitives of a candidate outline.

3 Heuristic Approach

“*Word Frequency (WF) Algorithm*” is a word level filter used to select the most probable word from a potential list of homophones. In order to cover a general domain, our transcription system uses the first 5000 of the most frequently used English words and each word is tagged with its corresponding frequency value. If an isolated outline is related to a list of homophones, a candidate with the highest frequency value is chosen as a potential successor. The algorithm can be defined as

$$P(f_i) = \frac{P(h_i)}{\sum_{i=1}^n P(h_i)}, \quad (3)$$

where $P(f_i)$ is the normalized probability of the frequency of a word in general English and $\{h_1, h_2, \dots, h_n\}$ denotes homophones of a single outline.

To clarify the “F&L”, “AN” and “WF” algorithms, consider the simple example shown in Figure 2. As shown in the example, the final word selection is done by the maximum likelihood probability of $P(WF) \cap P(F\&L) \cap P(AN)$.

Fig. 2. Illustration of transcription based on “F&L”, “AN” and “WF” algorithms.

4 Implementation and Experimental Results

The current goal of our experiment is to analyse the increase in similar outlines’ with the growth of a shorthand lexicon so that we can estimate which vocabulary-level contains the most similar outlines (*i.e.*, *homophones*) and which level has the least. Input to our system is a lexicon of up to the 5000 most frequently used English words, containing Pitman shorthand as indexes.

Experiment One in Figure 3 simulates perfect segmentation and recognition of an outline and demonstrates how a given outline can be uniquely identified in different sizes of shorthand lexicon. The experimental results show that about 94% of the 5000 most frequently used English words have unique shorthand notations. The maximum

ambiguity is 5 potential words per outline and an average ambiguity is 3 potential words per outline for the 6% without unique notations. Therefore a transcription accuracy of at least 94% can be estimated if there is no error in the prior outline segmentation and classification.

Experiment Two in *Figure 3* simulates recognition under line thickness ambiguity. This is the most likely case in real-time recognition as most digitisers cannot detect line thickness, whereas Pitman defines similar sounding consonants by the same strokes and differentiates between voiced and unvoiced sounds by thick and thin lines. The experimental result shows that outlines' uniqueness drops by about 4% given line thickness ambiguity and, therefore, the transcription accuracy at best can be expected to be 90% in the actual recognition of Pitman shorthand.

Experiment Three in *Figure 3* simulates recognition under vowel ambiguity. This is an essential consideration in the recognition of Pitman shorthand as it is a common practice of stenographers to omit vowels in an outline, and omitted positions are unpredictable as they vary widely from writers' experience or individual inclination. If a solution to the unpredicted omission of vowels in an outline is by the exclusion of vowel notations from the shorthand lexicon and matching up segmented primitives without vowel components, the new version of the lexicon is expected to have about 71% unique outlines. A strange peak i.e., 85% unique outlines occurred around the lexicon size of 4000 shows that outlines are more easily distinguishable without vowel components in a vocabulary level of 4000 words, but 4000 words is not enough for general vocabulary level.

Fig. 3. Analysis on shorthand Lexicon with the 5000 most frequently used words.

6 Discussion

Current findings are based on a shorthand dictionary and further experiment needs to be done on real handwritten data. Algorithms mentioned in the paper such as “*F&L*”, “*AN*” and “*WF*” need to be put into practice and further evaluation needs to be done on the whole of a word level and sentence level transcription.

Reference

1. Leedham C.G. and Downton A.C., ‘On-line recognition of Pitman’s shorthand: an evaluation of potential’, *Int. J. Man-Machine Studies*, Vol.24, pp.375-393, 1986

2. Qiao Y., Leedham C.G., 'Segmentation of handwritten Pitman shorthand outlines using an interactive heuristic search', Proc. 5th International Graphonomics Conference (ISBN 0-9630246-0-4), pp. 157-162, Tempe, Arizona, USA, 27-30 October 1991.
3. Qiao Y. and Leedham C.G., 'Segmentation and recognition of handwritten Pitman Shorthand outlines using an Interactive Heuristic Search', Pattern Recognition, Vol. 26 No. 3, pp. 433-441, March 1993.
4. P. Nagabhushan and Basavaraj. Anami, 'Dictionary Supported Generation of English Text from Pitman Shorthand Scripted Phonetic Text', Language engineering conference, Hyderabad, India, pp.33 December 13-15, 2002