J-039

手書きインタフェースにおけるテキスト抽出の一手法

A Text Data Extraction Method of Hand Writing User Interface

岩根 典之† 吉田 誠‡
Noriyuki Iwane Makoto Yoshida

1. はじめに

教材のデジタル化とともに教育学習の場でも１人１台のタブレット PC の利用が進むと考えられる．そのような環境において，スタイラスペンを用いた教材への書き込みは学習者の入力手段のひとつであり，自然なインタラクション形態となり得る．書き込みは読み手と書き手の対話とみなせるが，現状では読み手の一方的なインタラクションにとどまっている．読み手による教材への書き込みは，もともとの教材の記述とともに再び読み手により解釈される．そこでは教材から読み手である学習者に働きかけることはない．一方，エージェント技術を利用すれば教材はもっと自律的に学習者を支援できるはずである．エージェントは書き込み行為に対するフィードバックを読み手に返すこともできるようになる．

本研究では教材知識を利用して書き込みを解釈し，教育学習を促進する熟考型の質問エージェントの開発に取り組んでいる[1]．このエージェントを実現するためには読み手が書込んだ知識（書き込み知識）を抽出する必要がある．読み手はある暗黙的な意図をもって明示的に書き込む．すなわち書き込み知識が外在化される．しかし，手書きという自然なインタラクション形態を提供することは書き込みの曖昧性の問題を解決しなければないことも意味する．印刷資料に通常の筆記具で手書きされたアノテーションを自動抽出する研究[2]もあるが，本研究はスタイラスペンでデジタルテキストに手書きされたアノテーションを対象に知識レベルの抽出，すなわち質問生成のための知識獲得を目的としている．また，前者の研究とはデジタルとアナログの融合の考え方が異なり，テキスト抽出における課題も異なる．

本稿では，スタイラスペンで手書きしたタブレット PC 上のデジタルテキストから当該テキスト（書き込み知識）を抽出する一手法を提案する．本手法ではデジタルテキストの知識情報を利用するとともに手書き操作の種類を制限することにより曖昧性の解消を目指している．

2. 手書きによるテキスト抽出

2.1 テキスト抽出における考え方

本研究では質問生成のための知識獲得を目的としている．書き込み意図は，問い，答え，強調（重要）の三種類を基本とする．すなわち，デジタル教材を用いた初期段階の自主学習の支援を想定している．書き込みは，「これは何か」「それはなぜか」「それはどのようにしてか」など，デジタル教材を読みながら理解する過程で発現する問いや，読み進めながらそれら問いについて特定された答え，重要な個所などに対して行う．書き込みの基本操作は当該個所を囲むである．一方，問いと答えの対応はそれらを線で結ぶ．また，「これは何か」「それはなぜか」「それはどのようにしてか」などのタイプの違いも簡単な記号を書き込み意図の手掛かりとして手書きする．問いと答えの組は，ひとつの完結した学習知識であり，書き込みを行ったユーザの書き込み知識として獲得される．知識として完全な説明になっているのでその一部をブランク化することで穴埋め問題とその解答が生成できる．問いと答えの組からの完全な説明文は，タイプごとのテンプレート（フレーム）により生成する（完全説明フレーム）．手書きからのテキスト抽出では，書き込み意図を反映した問いや答え，強調に対する囲みから完全説明フレームの生成に必要な当該テキストを抽出する．

2.2 テキスト抽出問題

手書きという自然なインタラクション形態を提供することは書き込みの曖昧性の問題を解決しなければないことも意味する．書き込んだユーザが解釈する場合はその曖昧性は無意識にそのユーザにより解消される．しかし，本研究ではインタフェースエージェントが曖昧な書き込みから前節で述べた目的のためにテキストを抽出できる必要がある．また，書き込みはユーザの意図するテキストに対して，文字間スペースによっては囲まれた文字列が短かったり長かったり，すなわち不足する文字があったり，余分な文字が含まれたりする可能性がある．また，行間スペースによっては，余分な行の文字が含まれたりする可能性もある．このように手書きの囲みには，余分な文字が含まれたり，必要な文字が不足したりする曖昧性が存在する．テキスト抽出問題とは，インタフェースエージェントがユーザの曖昧な書き込みから質問生成に必要な知識としてテキストを抽出する問題である．

3. テキスト抽出法

テキストの抽出は，[1]着眼領域の決定，[2]着眼領域内の文字列抽出，[3]知識のタイプ記号の認識，[4]タイプ記号と抽出文字列から構文要素の比較，の４つのステップからなる（図 1）．各ステップでは以下の処理を行う．

[1] 着眼領域の決定

問いや答えなどとして着眼した部分が囲まれる．着眼領域はその囲みを含む矩形 FR として求める．着眼領域 FR は対角線上の最小点と最大点で表すが，手書きの性質を反映する．例えば，始点に対して終点は流れる．下から上への右回りの囲みでは図 1 の右側上の例のようにそれぞれの点の座標を決定する．

$$FR = [(\min X, \min Y), (\max X, \max Y)]$$

† 広島市立大学, Hiroshima City University
‡ 岡山理科大学, Okayama University of Science

[2] 着眼領域内の文字列抽出

着眼領域内 *FR* の文字列 *Str* は文字 *Chr* の集合として抽出する．ただし，GBc_i は文字 Chr_i の代表点でその文字のグリフの中心を表す．手書きの書き込みの曖昧性のため抽出された文字列に欠落や余分もあり得るので不完全な文字列の可能性がある．

$$Str = \{Chr_i \mid GBc_i \in FR\}, Strは文字列，\ Chrは文字$$

[3] 知識のタイプ記号の認識

手書きの囲みの周辺にその意図のタイプが 3 個前後の記号で書き込まれる．登録されたタイプ記号の書き込みとのテンプレートマッチングにより書き込まれたタイプ記号の種類を認識する．ただし，書き込みのタイプ記号はそれぞれのユーザごとに事前に登録しておく．

[4] タイプ記号と抽出文字列から構文要素の比較

着眼領域内の文字列 *Str* とデジタルテキストの知識情報，すなわち文法的に可能な文字列の候補 *Cand* から完全文字列 *CStr* を求める．着眼領域内の文字列 *Str* を包含する候補文字列 *Cand* のうち，文字列の長さが最小の候補を曖昧性のない完全文字列 *CStr* とする．

$$CStr_k = \min_j |Cand_j|,\ \ Str \subset Cand_j, |\cdot|は文字列の長さ$$

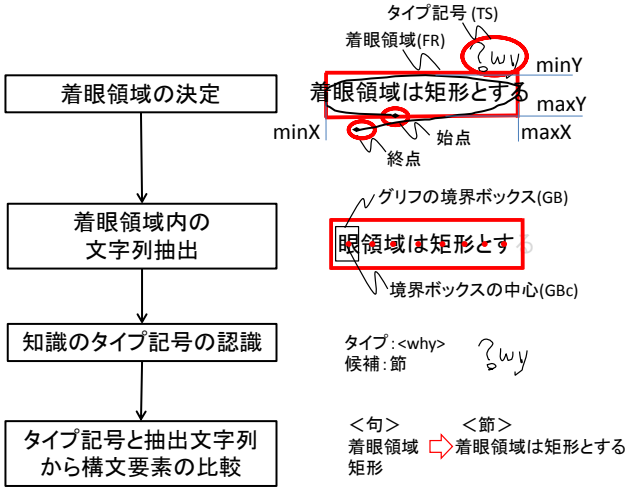

図 1　テキスト抽出処理

4. 予備実験

4.1 目的と方法

提案手法は表示される文字情報以外にもデジタル教材に格納する知識を利用して曖昧性を解消する．ここでは，まず，曖昧性がどのような具合なのか手書きデータを収集して予備実験を試みる．6 名の被験者の大学生に 200 文字前後の説明的文章のテキストに対して指定文字列 8 個のリストに従って手書きで囲んでもらった．テキストの内容 8 種類に対して，文字間（0 ピクセル／5 ピクセル），行間（5 ピクセル／10 ピクセル），フォントサイズ（15 ポイント／30 ポイント）をそれぞれ 2 種類，すなわち各テキストに対して 8 通りのバリエーションで文字列の長さや字種が異なる指定リストに従って 8 か所の書き込みを行ってもらった．8 種類のテキストは被験者にとって初見の内容とした．処理を簡単にするためテキストは画像データの pdf ファイルで文字は等幅に設定した．

4.2 結果と考察

全体では，指定通り抽出した正抽出率約 60%，余分な文字のある誤抽出率約 40%（誤抽出 1），不足文字のある誤抽出率約 10%（誤抽出 2）であった．さらに誤抽出 1 について，同じ行での誤りは約 30%，複数行に渡る誤りは約 80%であった．フォントサイズ文字間の違いによる結果の差異はほとんどなかった．しかし，行間の違いは，行間が狭い方が誤抽出率は倍近かった．また，行間の違いによる誤抽出は，行間が狭い方が複数行に渡る誤りにおいて 4 倍近かった．この予備実験を通じて以下が明確になった．

囲みが複数行にまたがった場合，着眼領域内の文字列抽出で余分の文字が抽出されることがあるが，この曖昧性は文字列候補との比較により解消できる．しかし，同一行でも候補となる文字列の長さが等しく，かつ着眼領域内の文字列を包含する場合は完全文字列を特定できないので，そのようなケースにも対応できるよう手法の改良が必要である．また，テキストは画像データではなく通常の pdf 形式で確認する必要がある．一方，書き込みスタイルの個人差は簡単な処理で吸収することにしているが，提案手法の有効性は被験者を増やして確認する必要がある．たとえば，タイプ記号の認識は事前に登録した各自の書き込み事例とのテンプレートマッチで十分なのか，囲みの始点位置，右回りや左回りなどスタイルの違いを吸収できるのかなど確認する必要がある（異なるスタイルの被験者も含めて人数を増やして確認するなど）．提案手法ではタイプ記号など書き込みに対して制限を設けたが，個人的に確立された書き込みスタイルを持っている場合，処理の簡単さを優先した本手法は自然な書き込みを阻害しないのかなど確認する必要がある．

5. おわりに

タブレット PC とスタイラスペンによる手書きインタフェースのためのテキスト抽出法を提案した．予備実験で書き込みの曖昧性がどの程度かを調べ，提案手法の課題を考察した．今後，タブレット PC のためのシステムの試作を通じて手法を評価改良する予定である．

謝辞

本研究は JSPS 科研費 24501142 の助成を受けたものです．

参考文献

[1]岩根典之,吉田誠，“デジタル教材から質問を生成するためのエージェントフレームワーク”，情報処理学会第 75 回全国大会講演論文集,4H-1 (2013).

[2]A. Mazzei, F. Kaplan, and P. Dillenbourg, “Extraction and Classification of Handwritten Annotations”, Proceedings of the 1st International Workshop on Paper Computing (2010).