文書の重要箇所抽出及び重み付け要約課題の解決方策についての分析

藤田 彬　　新井 紀子

国立情報学研究所

E-mail: { a-fujita, arai }@nii.ac.jp

1 はじめに

文字コンテンツのデジタル化，検索技術の向上，ユビキタスネットワーク，情報端末の小型化などの ICT の急速な発達に伴い，社会活動上の人間の能力に関する概念が変容しつつある。University of Oxford のチームにより，「知的活動の自動化が進むことでいくつかの職業が代替され，雇用に大きな影響がもたらされる可能性」が示唆された(Frey and Osbone 2013)。このような知的活動を支援するリッチなツールが多く手に入る環境においては，「解決の方法が直ぐには分かからない問題状況を理解し，情報を集め，適切な答えとして要約する能力」，いわゆる「問題解決能力 (Rychen and Salganik 2003)」がより重要性を増すと考えられる (Goos 2013; Hämäläinen, Cincinnato, Malin and Wever 2014)。しかしながら，本来人間が持ちうる問題解決能力がどの程度であり，人間にとってどのような方策が問題解決行動において適切であるかは，明確でない。

問題解決課題は，情報を収集した後の要約過程における認知的な負荷の大きさの別に，1) 情報源から重要箇所を抽出する手法(重要箇所抽出), 2) 情報源内の概念の重みを考慮して要約する手法(重み付け要約), 3) 抽象的な概念の流れを特定のクエリに沿ってまとめる手法(query-biased summarization)(Tombros and Sanderson 1998)，の 3 種に分類することができる。このうち，重要箇所抽出と重み付け要約は，人工知能分野の技術により一定水準の性能を持った自動手法が実現されている(平尾，磯崎，前田，松本 2013; Morita, Sasano, Takamura and Okumura 2013)。一方，query-biased summarization を行う技能は，自動化が実現されておらず，人間がもつ技能として特にユニークでありうる。例えば，「8 世紀から 10 世紀前半に、日本の政府が動員する軍事力の構成や性格はどのように変化したか」のような問いに指定された長さの文で答える行為は，query-biased summarization に該当する。この種の高度な要約を行う技能の育成は，国語科教育においても，以前より目標とされてきた。

しかしながら，筆者らの行った調査によると，実際は query-biased summarization を完遂する人間の数は多くないと見積もられる。入試偏差値が 60 を超える国内高校の 3 年生 70 名に，前述の日本史に関する要約課題を提示した。被験者は日本史の教科書内を完全一致検索する機能を持った端末を通じて資料を読み，記述を抜粋することができる。専門家により定められた 9 項の評価基準(含意すべき事柄)のうち，1 項以上を解答に含意できた被験者は全体の約 20%であり，最も多く含意できた被験者(全体の約 10%)においても含意項は 4 項であった。この結果からは，床効果により問題解決に働く要因を識別することができない。

従来，学校教育においては，query-biased summarization の技能を習得する足場かけ(scaffolding)として，重要箇所抽出，重み付け要約の手法が教材に用いられてきた。これらの手法を用いる課題については，十分な数の人間が問題解決に成功し，要因の識別が可能であることが明らかになっている(Fujita, Suzuki and Arai 2014)。しかしながら，どのような要素が問題解決の適切性に作用するかについては未だ明らかではない。

本研究では，被験者調査の結果に基づき，問題解決課題のうち「重要箇所抽出」及び「重み付け要約」の正確性に有意に働く要因を分析する。被験者は，ある限られたドメインの特定の事柄について説明する記述問題を解く。被験者には，解答となる記述を含む資料を検索・参照しながら解答文を編集する機能をもったインタフェイスが提供される。この時の「解答行動の時系列ログ」,「解答の評価結果」を分析対象として取り上げる。

2 手法

2.1 調査参加者

高校 3 年生(2014 年度)303 名が調査に参加した。参加者が所属する学校は，男子高校 A(4 クラス 148 名)と男子高校 B(2 クラス 78 名)，女子高校 C(2 クラス 77 名)に分かれる。調査参加者の全員が大学進学を希望する。調査参加者らが高校を受験した 2011 年度の各校の入試偏差値は，学校 A が 70 以上，学校 B,C が 60〜69 であった。

2.2 課題

【資料】：資料には，検定済の日本史教科書『日本史Ｂ』(東京書籍株式会社, 2010)を用いた。資料は，歴史上特筆されるイベントや時代を特徴づける社会背景を取り扱う「トピック」という単位の文書にわかれる。1 トピック辺りの文書の平均分量は 1000 字である。調査参加者は当該教科書を授業等で参照したことはない。

【問題】：日本史に関する 2 種の問題(以下，問 1，問 2)を出題した。両問の内容および出題設定は，日本史教育の専

門家により監修された。個人の歴史観や意見を問う問題ではなく，検定済の教科書において史実とされる事柄について叙述的な説明を求める問題である。両問とも解答に字数制限及び時間制限を設けた。問題を以下に示す。

問 1：平安時代の政治体制では、藤原詮子や藤原彰子など、天皇の生母に当たる女性が重要な役割を占めていた。その理由を、当時の政治体制の在り方とともに，110 字から 130 字でまとめよ。(解答時間：10 分)

問 2：鎌倉時代、幕府により諸国に地頭が置かれたが、この地頭がどのような名目で設置され、実際にはどのような職務を担っていたか。またその後、承久の乱を経てどのように変化していったのかを、140 字から 160 字でまとめよ。(解答時間：15 分)

問 1 は，教科書上のある 1 つのトピックを探し出し，連続した箇所を抜き出すことで，過不足なく答えられる。

問 2 は，2 つの離れた箇所に記述されたトピックを探し出し，それぞれからある連続した箇所を 1 か所ずつ抜き出し，内容を要約することで，過不足なく答えられる。

問 1 では，該当箇所を抽出するのみで字数制限をクリアできる。一方，問 2 では字数制限を超過するため，要約を行う必要がある。

【評価基準】：大学受験予備校における日本史論述問題の評価指針に沿って，専門家が各問の評価基準を策定した。問 1 の評価基準は 2 項，問 2 は 7 項ある。評価基準の各項で示される事柄が答案に含意されるか否かを，日常的に記述式解答の評価を行う専門家が判定した。1 件の答案の含意状況を，異なる 2 名の専門家が判定した。判定結果が異なった際は合議をし，判定を確定するという形式をとった。評価基準を策定した専門家と評価を実施した専門家は異なる。

2.3 手続き

調査の冒頭に，調査参加者に「資料を検索しながら，解答を編集する機能」を有する図 1 のようなインタフェイスを提供した。インタフェイスは，「問題提示窓」，「検索フォーム」，「検索結果表示窓」，「下書きフォーム」，「解答フォーム」，「解答完了ボタン」から構成される。

問題提示窓：調査開始直前に監督者から提示される ID を入力すると，解答対象となる問題および解答制限時間，字数制限が表示される。提示された問題文字列はコピーが可能である。

検索フォーム及び検索結果表示窓：資料内のトピックを対象として，完全一致検索ができる(AND 検索可)。検索を実行すると，スニペット(検索語を含むトピックへのハイパーリンクとトピック内で検索語がヒットした付近の文字列の組み合わせのリスト)が表示される。検索クエリにヒットするトピックが存在しない場合は，その旨を示す文が提示される。スニペット上のハイパーリンクをクリックするとトピックが表示される。表示されたトピックからはコピーが可能である。トピックの上部には，スニペットに戻るハイパーリンクがある。

下書きフォーム及び解答フォーム：文章の編集及び各種フォーム・窓からのコピー&ペーストが可能である。それぞれのフォームの上部には，入力文字列の字数が逐次表示される。下書きフォームは「記述する内容を暫時的に整理するフォーム」であり，解答フォームは「最終的な解答をまとめるフォーム」である。この旨は，事前に監督者より調査参加者に提示される。

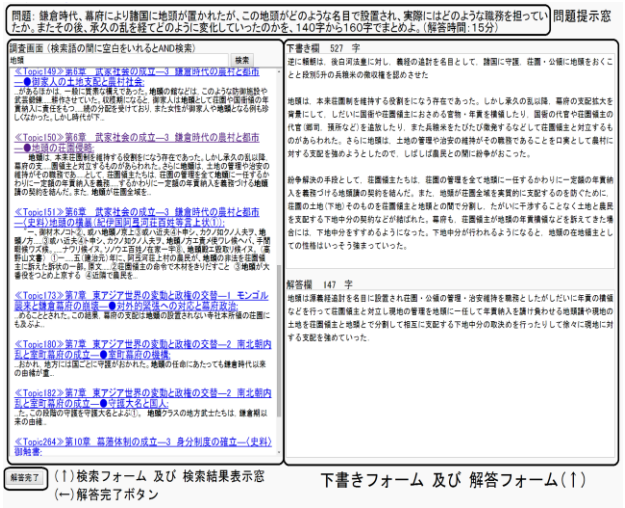

図 1： インタフェイスのイメージ

解答完了ボタン：調査参加者自身が解答の完了を判断した際に押す。解答制限時間が過ぎた場合は，監督者がボタンを押すよう指示する。

インタフェイスは，調査参加者が課題にあたる間の下記の行動に関する情報を，行動が起こった絶対時刻と共に記録する機能をもつ。(行動：取得される情報)

- 解答開始及び解答完了：(なし)
- 検索実行：検索語
- スニペット or トピックのロード：ロードされたページの URL
- 下書きの編集：下書きフォーム内の文字列
- 解答の編集：解答フォーム内の文字列

ここでいう「編集」は，文字(列)の追加，挿入，削除，書き換え行為を指す。

調査参加者がインタフェイス使用に慣れることを目的として，問 1，問 2 を提示する前に練習問題を 1 問出題した。調査参加者は，監督者の誘導を受けながら，「問題表示→検索語生成→検索実行→資料閲覧→解答編集→解答完了」という一連の操作方法を確認した。その後，問 1，問 2 の順に解答を行った。

調査はクラスごとに時間を隔てて行われた。1 クラスずつ，調査参加者を情報端末室に呼び，備え付けられた PC の Web ブラウザを通じてインタフェイス(Web アプリケーション)を使用し，調査を行った。

3 結果と考察

調査参加者の「解答行動中のログ」と「答案に含意される評価基準の項数」の間の関係を問題別に分析する。解答

中に調査参加者の使用端末にトラブルがあった場合，当該ログ及び答案を無効とした。ログ，答案共に有効であった調査参加者は，問 1 で 291 名，問 2 で 256 名であった。

3.1 答案の含意項数と関連する要素

問題別の含意項数の分布を表 1 に示す。両分布とも正規分布しない(Shapiro-Wilk test, 5%水準)。以下では，中央値を閾値として各分布を高得点群，低得点群の 2 群に分けることとする。含意項数の中央値は，問 1 が 1 項，問 2 が 3 項であった。問 1 については含意項数 1 項未満を"p1_low"(N=55)，1 項以上を"p1_high"(N=236)とした。同様に問 2 については，3 項未満を"p2_low"(N=80)，3 項以上を"p2_high"(N=176)とした。

各問には評価基準と共に，問 1 に 2 例，問 2 に 1 例の模範解答が用意されている。この模範解答は，評価基準を設定した専門家が，教科書の記述内容に沿って，教科書での出現順通りに評価基準の全項が記述される答案を作成したものである。この模範解答と各答案の間の編集距離を測り，群間で差を検定した。編集距離の測定には，Levenshtein distance の計算手法を用いた。問 1 の答案の編集距離は，2 つの模範解答のうち距離が短い方の模範解答との編集距離を測った。各群の編集距離の中央値は，p1_low: 98, p1_high:64, p2_low:135, p2_high:131 であった。Wilcoxon rank sum test を実施したところ，問 1，問 2 とも高得点群，低得点群の間で編集距離に有意な差が認められた($p < 0.01$)。このことから，模範解答と編集距離が近いものほど含意項数が多い傾向が確認できる。

編集距離を用いる場合，内容が同一の 2 文であっても，節の並び順や論理展開の順が異なることで，同一性を正しく反映した数値が得られない場合がある。そこで，出現順序を考慮しない単語単位での検討も行った。評価基準が言及する事柄が教科書中で記述される箇所を特定し，日本史ドメイン上の固有表現(named entity)と考えられる名詞(サ変名詞含む)を人手で抽出した(以下，キーワード)。問 1 については 24，問 2 については 46 のキーワードが抽出された。これらのうち，いくつのキーワードが各答案に出現するか異なり数を計数し，群間で差を検定した。出現キーワードの異なり数の中央値は，p1_low: 5, p1_high: 17, p2_low:10, p2_high:20 であった。Wilcoxon rank sum test によると，問 1，問 2 とも高得点群，低得点群の間で答案に含まれるキーワードの異なり数に有意な差が認められた($p < 0.01$)。

前述のように模範解答との編集距離が群間で異なることも考慮すると，高得点群の答案は低得点群の答案に比べ，「記述内容の順序関係」,「使用される固有表現」の両者が，教科書の記述に類似する傾向があることがわかる。

3.2 問題解決に必要な情報の符号化を試みた後に取る行動方策が答案の含意状況に与える効果

資料から適切な箇所を抜き出す(コピー&ペーストを行う)ことで，多くの評価基準を含意できることは自明である。しかしながら，多くの評価基準を含意する上で，資料を抜き出す方策のみが適するとは限らない。以下では，問題で要求される情報が記載されたトピック(以下，正解トピック)を符号化した後の調査参加者の行動を分析する。

表 1：調査参加者による答案の含意項数の分布

	0 項	1 項	2 項					
問 1	55	141	95					
	0 項	1 項	2 項	3 項	4 項	5 項	6 項	7 項
問 2	7	34	39	75	78	22	1	0

表 2：正解トピック初回表示後行動の得点群別調査参加者数(カッコ内は群中における百分率)

	p1_low	p1_high	p2_low	p2_high
正解部抜出	8 (15)	105(44)	10 (13)	56 32)
不適切箇所抜出	4 (7)	36 (15)	6 (8)	7 (4)
文字入力	17 (31)	61 (26)	19 (24)	22 (13)
スニペット再表示	3 (5)	13 (6)	29 (36)	68 (39)
再検索	4 (7)	21 (9)	9 (11)	21 (12)
正解トピック表示無	19 (35)	0 (0)	7 (9)	2 (1)

調査参加者の行動ログから 1 回目に正解トピックを表示した箇所を特定し，その直後に行った行動を 5 種類, a) 正解部抜出，b)不適切箇所抜出, c)文字入力, d)スニペット再表示, e)再検索に分類した。「正解部抜出」は評価基準のいずれかを含意する記述を包含する箇所を抜き出す行為を指す。それ以外の箇所を抜き出す行為を「不適切箇所抜出」とする。下書きフォームもしくは解答フォームにキーボードから文字を一文字ずつ追記・削除する行為を「文字入力」とする。問 2 については，2 ページある正解トピックのうちどちらか 1 ページを初めて表示した際を分析対象とした。これらの分類は，人手の判断により行った。

表 2 に，両問の高得点群，低得点群のそれぞれについて，正解トピック初回表示後の行動毎に各行動をとった調査参加者の人数及び群中における割合を示す。このうち，「正解トピック初回表示後の方策が正解部抜出または文字入力であること」と「最終的な答案の含意項数」の間の関係を分析した。問 1，問 2 とも，正解トピック初回表示後に正解部抜出を行った調査参加者の割合は高得点群の方が多く，文字入力を行った者の割合は低得点群の方が多い。これを踏まえ，各問について，高得点群と低得点群の調査参加者を，さらに正解部抜出の方策を取った群とその他の方策を取った群(正解トピック表示無も含む)に分けた。この 4 群の各群に属する参加者の人数について Fisher's exact test を実施したところ，問 1，問 2 の両問について，「正解部抜出の方策を取るか否か」と「どちらの得点群に属するか」という 2 つの要因の独立性が成立しないことが認められた(両側検定, $p < 0.05$)。文字入力についても同様に

Fisher's exact test を実施したところ，問 2 について「文字入力の方策を取るか否か」と「どちらの得点群に属するか」という 2 つの要因の独立性が成立しないことが認められた(両側検定, $p < 0.05$)。

正解トピック初回表示後に正解部抜出を行った群と，文字入力を行った群の間で, 答案の「模範解答との編集距離」と「含まれるキーワードの異なり数」の差を検定した。答案と模範解答との編集距離について Wilcoxon rank sum test を実施したところ，問 1，問 2 とも，正解部抜出群と文字入力群の間で有意な差が認められた($p < 0.01$)。また，含まれるキーワードの異なり数について, 問 1 に Wilcoxon rank sum test，問 2 に Welch's t test を実施したところ，問 1，問 2 とも 2 群間で有意な差が認められた($p < 0.01$)。模範解答との編集距離は, 両問とも正解部抜出群の方が文字入力群より短い。また，含まれるキーワードの異なり数は，両問とも正解部抜出群の方が文字入力群より多い。

このことから, 正解トピック初回表示後に文字入力を行った調査参加者に比べ, 正解部抜出を行った者は最終的に含意すべき意味内容をより多く含意し, かつ資料に記載された記述に表層的にも近い答案を作成する傾向がわかる。

3.3 調査参加者の属性による効果

調査にあたり, 事前に学校 A の調査参加者のクラスを担当する国語科教諭が，「読解して要約する力」について取り立てて優秀であると印象付けられる調査参加者を, 各クラス 10 名程度を目安に抽出した。この取り立てて優秀である群(問 1: N=31, 問 2: N=30)と, その他の群(問 1: N=108, 問 2: N=111)に調査参加者を分け，答案の含意項数の差を検定した。Wilcoxon rank sum test を実施したところ，問 1 においては有意な差は認められず($p > 0.05$), 問 2 においては有意な差が認められた($p < 0.05$)。問 2 においては，取り立てて優秀な群の含意項数の平均値が 2.77 であるのに対し，その他の群は 3.26 であった(中央値はともに 3)。このことから，重要箇所抽出タスク，重み付け要約タスクの両者については，国語科で扱われる「読解して要約する力」と異なる要素が主要因として働く可能性が示唆される。

4 おわりに

本稿では，資料検索・解答編集の機能を持ったインタフェイスを用いた被験者調査に基づき, 重要箇所抽出及び重み付け要約の正確性に有意に働く要因を分析した結果を述べた。解答内容に含意されるべき事項を多く含意した解答は，「記述内容の順序関係」，「使用される固有表現」の両者が資料の記述に類似する傾向が明らかになった。同時に，含意すべき事柄が記載された資料から情報の読解(符号化)を試みた後に取る方策が，「自ら情報を再構成して記述する方策」であるよりも，「含意すべき事柄を含んだ情報を抜き出し，編集を加える方策」である方が適切な解答生成につながるという結果が得られた。また，これらの問題解決課題を適切に処理する能力の主要因が, 従来扱われてきた読解して要約する力とは異なるものである可能性が示唆された。

本稿で扱った課題については，情報検索ツールを直接的に利用して問題解決を図る方が正確であるように報酬系が働き, その通りに行動した者が十分な解答を得たと考えられる。符号化した情報を再構成する能力がメリットとして働くタスクは query-biased summarization と考えられるが，前述の通り，現時点で要因を分析するだけの環境が実現できるとは言い難い。しかしながら，少なくとも，「重要箇所抽出, 重み付け要約」を「query-biased summarization」の scaffolding とすることの妥当性は再考の余地があると考えられる。

今後は，資料のドメイン及び調査課題の拡充，インタフェイスの最適化により, 調査結果を汎化することが求められる。

謝辞

本研究に際して, 埼玉県立総合教育センター及び国立大学法人筑波大学附属駒場高等学校より, 研究機会をご提供いただいた。また, 学校法人高宮学園代々木ゼミナールより, 日本史に関する専門知識をご提供いただいた。ここに感謝の意を表する。

参考文献

Frey, B. C. and Osborne, A. M. (2013). The Future of Employment: How susceptible are jobs to computerisation? Oxford Martin.

Fujita, A., Suzuki, M., & Arai, H. N. (2014). Cognitive Model of Generic Skill: Cognitive Processes in Search and Editing. Proceedings of the 36th annual meeting of the Cognitive Science Society, 2234-2239.

Goos, M. (2013). How the World of Work is Changing: A review of the evidence. ILO Research Paper. http://feb.kuleuven.be/public/n06022/ILO-20131205.pdf

Hämäläinen, R., Cincinnato, S., Malin, A. and Wever, D. B. (2014). VET workers' problem-solving skills in technology-rich environments: European approach. International Journal for Research in Vocational Education and Training, Vol.1, No.1, pp.57-80.

Morita, H., Sasano, R., Takamura, H. and Okumura, M. (2013). Subtree Extractive Summarization via Submodular Maxmization. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp.1023–1032.

Rychen, S. D. and Salganik, H. L. (2003). Key Competencies for a Successful Life and a Well-Functioning Society. Hogrefe & Huber Publishing.

Tombros, A. and Sanderson M. (1998). Advantages of Query Biased Summaries in Information Retrieval. SIGIR'98.

平尾努, 磯崎秀樹, 前田英作, 松本裕治. (2003). Support Vector Machine を用いた重要文抽出法. 情報処理学会論文誌, Vol.44, No.8., pp.2230-2243.

東京書籍株式会社. (2010). 日本史 B.