様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１６　日現在

機関番号：　８２６１６
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３５０３５７

研究課題名（和文）短答式記述テストにおける実用的な自動採点システム

研究課題名（英文）A Practical Japanese Short-answer Scoring　System for Writing Test

研究代表者

石岡　恒憲（Ishioka, Tsunenori）

独立行政法人大学入試センター・研究開発部・教授

研究者番号：８０３１１１６６

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：センター試験など大学入試レベルの短答式記述試験における自動採点および人間採点を支援する実用可能なシステムの試作および実装をした．自然言語における完全な意味理解はこの数年では不可能であるという判断のもと，採点は設問ごとに作題者が用意した「採点基準」に従った自動採点を基本とし，その結果を人間が確認・修正できるものとする．システムは「（予め用意された）模範解答」と「実際の記述解答」との意味的同一性や含意性を判定するほか，プロンプトと呼ばれる素材文と解答文との意味的近似性なども考慮する．また採点結果は多値分類であることから，サポートベクターマシンではなくランダムフォレストによる機械学習分類を使う．

研究成果の概要（英文）：We have developed an automated Japanese short-answer scoring and support machine for new National Center Test written exams. Our approach is based on the fact that recognizing textual entailments and/or synonymy has been almost impossible for scoring systems for several years. The system generates automated scores on the basis of evaluation criteria or rubrics, and human raters revise the scores. The system determines semantic similarity between the model answers and the actual written answers as well as a certain degree of semantic identity and implication. Owing to the need for the scoring results to be classified at multiple levels, we use random forests to effectively utilize many predictors rather than use support vector machines. An experimental prototype operates as a web system on a Linux computer.

研究分野：　情報数理

キーワード：　情報システム　自然言語処理　機械学習　CBT　Webアプリケーション　自動評価・自動採点

１．研究開始当初の背景

アメリカでは教育関係者のみならず一般においても、主観が入るエッセイ試験において、コンピュータによる評価採点の考えが受け入れられるようになってきた。実際、アメリカのビジネススクール入学のための共通テストである GMAT における作文（エッセイ）テストでは、1998 年より e-rater が、2006 年よりは IntelliMetric が採点を行なっている。他にも商用のシステムとして、PEG（Project Essay Grade）や IEA（Intelligent Essay Assessor）があり利用に供されている。我が国においても、我々のグループが日本語で始めての小論文自動採点システム Jess（Ishioka, 2006）を開発し、実用化の段階にある。成果については、朝日新聞夕刊の一面トップ（2005 年 2 月）に大きく掲載されたほか、Yahoo インターネットガイド（2006 年 6 月）、韓国 KBS テレビ（2007 年 2 月）等、多くのマスコミで紹介された。

しかし、本科研で実施しようとしている短答式（short-answer)テストをコンピュータで行なうことについては、その重要性は認められているものの技術的にさまざまな課題が未解決のままである。短答式テストとは、質問の解答を１文あるいは２文で返すものである。Vigilante(1999) は世界最大のテスト機関である ETS とニューヨーク大学とで、この分野における共同研究を行い、最初の報告をした。Leacock & Chodorow (2003)は、ETS が開発した c-rater の最新の仕様について報告している。Pulman & Sukkarieh (2005) は、情報抽出技術に隠れマルコフモデルなどの自然言語処理を用いて、システムが用意する正解文と同じ意味の文を幾つか自動生成する試みについて述べている。このように、1998 年ごろから短答式テストにおける自動採点の研究が進んでいるものの、本研究代表者の知る限り実用システムとしては ETS が唯一 c-rater を試作しているに過ぎない。しかも、その性能はエッセイの自動評価採点システム(e-rater)に比べ、人間による専門家との一致率は 10%以上も小さい。

一般に、短答式テストは多肢選択テストに比べ、以下の点が優れているとされる。
(1) 短答式テストの方がより真正(authentic)で信頼できる（と広く考えられている）。実際、現実世界における質問応答は、多肢選択ではなく、短答式テストに近い。
(2) 経済性。高品質な多肢選択問題を作成することは、通常、コストと手間がかかる。
(3) 多肢選択テストはテスト戦略を立てやすく、学生の問題についての理解を正しく評価することが難しい。また当て推量（まぐれ当たり）による効果も無視できない。

このため、我々は先の科研（挑戦的萌芽研究；H22-H24）で、短答式テストの自動採点という挑戦的な課題について取り組んだ。ここでは、我々が小論文の自動採点システムで培った自然言語処理技術を用い、「(テストの作成者が予め用意した）模範解答」と「(被験者の実際の」記述解答」とが十分近いかを、新聞記事などのコーパスから予め得ていた学習データを用いて判定することができるようにした。つまり、単純に単語の一致（パターンマッチ）だけで一致度をはかるのではなく、「模範解答」への意味的な近さや、表記上の揺らぎを吸収し、また係り受けの明瞭な正しい、わかりやすい文になっているかを、新聞の社説やコラムでの書き方と比較し、判定をおこなうことを可能とした。また、一つの「模範解答」から、それと同義な正解文の複数パターンを自動的に生成することを行った。これにより、非常に限定的な「模範解答」を一つないし少数を用意するだけで、意味的に一致する文を正しく判定できるようになった。

ところが近年になって言語処理の分野で、より具体的には 2005 年から 3 年間続いた評価型ワークショップ Pascal RTE(Recognizing Textual Entailment） Challenge をきっかけに、一方が他方の記述から含意（あるいは推論）されるか否かを判別する含意関係認識の必要性が注目され、従来の「浅い」処理から一歩意味に踏み込んだ解析技術の重要性が明らかになってきた。

２．研究の目的

本研究では、自然言語処理や情報アクセス研究に広く共通する課題である、テキスト間の含意（推論）・換言（同じ意味）・矛盾の認識を可能とする技術を取得し、実際の大学入試試験問題における記述試験採点で、実用に耐えるシステムを試作する。採点は、「(テストの作成者が予め用意した）模範解答」と「(被験者の実際の）記述解答」との意味的同一性や含意性を判定することによって行う。システムは Web にて公開することを目指す。日本語を処理する短答式テストの実用的な自動採点システムはまだ世の中に存在しないため、これができれば我が国で最初のシステムとなる。

３．研究の方法

大学試験の実際の問題を分析し、含意関係認識における主な研究課題について取り組む。入力が自然言語処理であるために自然言語処理は必須であるが、解決すべき課題は自然言語処理の枠組みに収まらない。知識処理・推論、意図や比喩など認知に基づく意味理解などさまざまな種類の知的情報処理が必要である。

このうち現時点で最も取り組みやすい「知識を問う」問題からアプローチを行ない、ある程度の成果を上げる。次により深い知的処理が必要な問題や暗黙の知識・推論を利用する技術について検討をおこなう。研究課題に対する大よその分担は以下の通り：

石岡（研究代表者）：総括、自然言語処理、含意関係認識

峯：自然言語処理、質問・応答メッセージの分析

初年度である平成 26 年度は、まず「知識を問う」問題に対して、含意関係認識を行うプログラムの実装を行う。ごく粗くいえば、日本語構文解析による語句の係り受けと、キーとなる語句の上位語・下位語・同義語などの置き換え、関連知識の検索技術の組み合わせ、および最終判断としての機械学習で達成できるのではないかと考えている。

また日本語構文解析については既にある CaboCha の利用を、機械学習器としては Random Forests の利用を考えており、これ以外の要素技術の補完とプログラムの自作をおこなう。

具体的は項目は以下の通り：

(１) 言語資源の利用に向けた調査・検討（石岡）

含意関係認識に有効と思われる各種の資源およびツールについて調査する。これらは全て CMU(カーネギーメロン大学)で開発・提供されている。また一般に公開されている利用可能な言語として、奈良先端大(NIST)の作成した Japanese WordNet および Hyponymy extraction tool を用いた Wikipedia 上位語・下位語ペア、また京都大学・黒橋研究室の作成した京都大学格フレーム、NTT の提供する日本語語彙体系（有料）についての利用を予定しており、その利用について検討を行う。

(2) 自然言語処理要素技術（峯・石岡）

試験問題のテキストは誤解が生じないようにわかりやすく記述されているが、それでもその解析は容易ではない。ほぼすべての科目を通じて、広い意味での参照関係の解析が必要である。たとえばテキストの表層に明示的に表れない先行詞や表層・構文関係に現れない参照表現の認識が必要である。

研究分担者の峯は就労マッチングに向けた相互推薦の研究をしており、その中で質問・応答メッセージ履歴の取得と，その分析を行っている。参照関係理解はそこでも重要なキー技術でありそこでの成果の活用が期待される。

(3) 含意関係認識プログラムの作成（石岡）

実際に上記の要素技術に基づいて、含意関係認識プログラムを試作する。本体となるスクリプト言語に加えて、日本語文字列を操作可能な sed や awk, perl, python などを用い、プログラムを行う。ただし、特に凝ったユーザインターフェイスを作ることはしない。

平成 27 年度（2 年目）以降は以下を実施する。

(4) 文章や状況の一貫性や自然さの認識（石岡）

国語（英語も）の会話・文章穴埋め問題では、文章や状況の一貫性や自然さを認識する必要がある。文章の自然さのレベルには、単語の選択制限や熟語の知識から判断できる比較的容易なものに加え、談話構造の認識が必要なもの、および世界知識・理解が必要なものがある。談話構造の認識が必要な問題としては、接続詞の選択や因果関係の認識が挙げられる。つまり１文あるいは２文で解答を返す場合に、それらの文が日本語として意味が通じていて、２文の場合は互いの関係が論理的に破綻のない一貫性や自然さをもっていることが必要となる。

このような問題は一般に難しい問題とされているが、研究代表者は過去に小論文自動採点システム Jess を試作し、その要素技術の開発の中で、自然な文章や談話の接続の関係を評価する技術について研究成果を上げている。具体的には、接続詞を含む広い意味での接続表現、前の段落を示す指示語、話者の態度や考えを示すモダリティの検出とそれによる接続（付加、解説、論証、例示、転換、制限、譲歩、対比）の判定を、多くの新聞記事の社説やコラムを学習することで解決している。これと同様の、あるいは類似の技術が利用できるのではないかと考えており、そのプログラムの実装について検討を加える。

(5) 高度な判断に基づくアプローチ（石岡）

2009 年センター試験「政治・経済」の安保理に関する設問では、正解にたどり着くには選択肢②において「安全保障理事会の認定」と「総会の承認」は異なる概念であり、前者（「安全保障理事会の認定」）があれば「軍事的措置が行われる」から、「議会の承認が必要」という記述が誤りであることがわかる必要がある。このような問題は、教科書などの知識を前提にした含意関係認識とみなすこともできるが、さまざまな推論が複雑に絡み合っていて、現在の技術で解くことはきわめて難しい。このような複数の知識を組み合わせて答えを導く方法についても検討を行う。

４．研究成果

我々は自動採点に向けた要素技術の習得やシステム試作に努めてきたが、完全な含意関係認識技術すなわち正しい意味理解は現時点では困難であることがわかってきた。またシステムの限界も見えてきた。実際、国立情報学研究所の「ロボットは東大に入れるか」プロジェクトでも、受験ロボットがセンター試験の選択肢の正誤を、教科書からの知識源との含意関係認識技術を用いて解くということを試みてきたが、その結果わかってきたことは、完全な含意関係認識技術の困難性であり、いくつかの手法を組み合わせて半ばアドホックに解くことで、あたかも人間が解答したかのようにみせかけることがせいぜいであるという事実である。

そこで本科研では短答式記述テストの完全な自動採点は諦め、コンピュータで自動採点（仮採点）の結果と、その自動採点に至った根拠を示すことで人間（採点者）が自動採点を修正することのできる採点支援システムを実装することとした。採点者が自動採点

の結果に同意するならば、採点者はディフォルトで表示される採点結果を承認するだけで採点をすすめることができる。このような採点支援システムであっても、日本語を処理する我が国で最初の実用システムに向けた試作となった。

本システムは Web アプリケーションとして動作するが、下に示す図１は採点者が用いるユーザインターフェイス画面である。テキストで書かれた採点基準ファイルからこの画面が自動的に作成されるところに特徴がある。

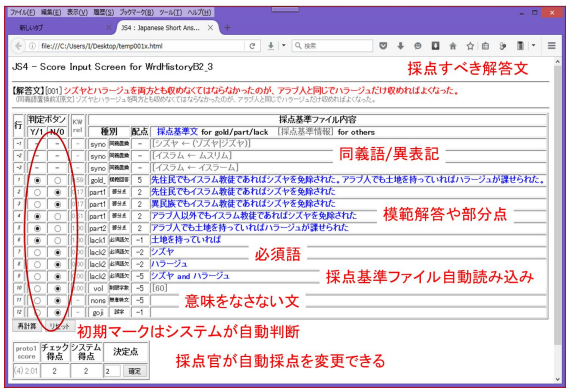

図１：短答式記述採点支援システムのユーザインターフェイス（採点基準の各項目に対し採点すべき解答文が合致しているか否かを自動判定するほか、機械学習に基づく予測値を出力する）

システムの性能を，2015 年度の「学研全国総合模試」で出題された地理・歴史 4 科目において出題された短答式記述解答部分（1 科目あたり各 2 問,計 8 問)を用いて評価した。模範解答の字数は 20－60 字程度である。8 つの設問に対し、予測モデルによる予測値と正解値を比較した。交差行列の詳細については紙面の都合で割愛するが、完全一致しない確率（誤分類率）は少なくない。ただ、正解と予測の違いが 1 点差内で収まる確率を求めると表１のようになり、正確な意味理解が必要な世界史 B〔②〕3 を除けば 71%から 95%とかなり使えるレベルであることがわかる。サンプルサイズは設問の種類によるが、概ね 70－130 程度である。

表１: 予測が正解の 1 点差内に収まる確率

設問		確率	設問		確率
世界史 B [2]	1	0.75	日本史 B [2]	1	0.86
世界史 B [2]	3	0.48	日本史 B [2]	3	0.71
日本史 B [1]	2	0.76	地理 B	1	0.91
日本史 B [1]	4	0.88	地理 B	4	0.95

模範解答（正解文）と記述解答文の意味的同一性チェックは、構文的・意味的レベルの複雑な照合が必要でまだまだ技術的に困難であるが，現在の双方のキーワード群同士の照合は、「表層的」な側面だけでなく「意味的」な側面も予測変量として加えており、その第一歩のアプローチとしては、十分に現実的だと判断できる。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　16 件）

①Shaymaa E. Sorour1 and Kazumasa Goda and Tsunenori Mine, Comment Data Mining to Estimate Student Performance Considering Consecutive Lessons, Journal of Educational Technology & Society, 2017, 20(1), 73-86.(査読あり)

②石岡恒憲，コンピュータ上で実施する記述式試験 ―エッセイタイプ，短答式，マルチメディア利用について―，電子情報通信学会誌， 99 (10), 2016, 1005-1011.（招待論文）

③石岡恒憲,再設計される（redesigned)SAT について ― 改訂の 意図と背景 ―， 大学入試研究ジャーナル，26，2015，163－170.(査読あり)

④石岡恒憲，狩野芳伸，橋本貴充，大津起夫，全文検索による試験問題検索システム ― 新規作成問題の類似文書検索を中心として―，大学入試研究ジャーナル，24，2015，129-135.（査読あり）

⑤Tsunenori Ishioka, Investigations into Missing Values Imputation Using Random Forests for Semi-supervised Data, Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services, 2014, 296-301, DOI: 10.1145/2684200.2684288(査読あり)

〔学会発表〕（計　22 件）

①石岡 恒憲，コンピュータ上で実施する記述式試験について［招待講演］, 第 133 回 コンピュータと教育研究会，情報処理学会，情報処理学会電子図書館 CE-133-19．2016/2/13-14.（東京都小金井市）

②石岡恒憲，コンピュータ上で実施する記述式試験―エッセイタイプ，短答式，マルチメディア利用について― 平成 28 年度全国大学入学者選抜研究連絡協議会大会（第 11 回）研究発表予稿集， 2016，108-116. 2016/6/1-3，立命館大学 茨木キャンパス(大阪府茨木市)

③石岡恒憲，コンピュータ上で実施する記述式試験, 2016 年度 統計関連学会連合大会，2016，報告集 403．2016/9/5-9/7 金沢大学角間キャンパス(石川県金沢市)

④石岡恒憲，コンピュータ上で実施する記述式試験 ～人工知能活用の観点から～ [特別講演]，電子情報通信学会，信学技報 AI，2016，23-25，2016/12/9 クアージュ湯布院（大分県由布院市）

⑤石岡恒憲，亀田雅之，劉東岳．人工知能を利用した短答式記述採点支援システムの

開発，電子情報通信学会，信学技報 NLC，2016，87-92，2016/12/21-12/22 NTT 武蔵野研究開発センタ（東京都武蔵野市）
⑥亀田雅之，石岡恒憲，劉東岳．短答記述式問題解答文の採点支援システム JS4 の試作，言語処理学会第 23 回年次大会(NLP2017)，2017，1137－1140．2017/3/13-3/17　筑波大学春日キャンパス（茨城県つくば市）
⑦石岡恒憲，亀田雅之，劉東岳．人工知能を利用した短答式記述採点支援システムの開発，計測自動制御学会，第 44 回知能システムシンポジウム，2017，SY0004/17/B2-3．2017/3/13-3/14 東海大学 高輪キャンパス（東京都品川区）

〔図書〕（計２件）
①石岡恒憲，米国における SAT の改革と入試研究，「大学入試における共通試験」第Ⅲ部「海外における共通試験」第１章，高等教育ライブラリー12，東北大学出版会，2017，153－164，総ページ数 221．
②石岡恒憲，テストの現代化と大学入試（繁桝算男 編「新しい時代の大学入試」，第 2 章），金子書房，2015，57-78，総ページ数 205．

〔産業財産権〕

○出願状況（計　０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
Ishioka's Home Page:
http://www.rd.dnc.ac.jp/~tunenori/

６．研究組織
(1) 研究代表者
石岡 恒憲（ISHIOKA Tsunenori）
大学入試センター・研究開発部・教授
研究者番号：８０３１１１６６

(2)研究分担者
峯 恒憲（MINE Tsunenori）
九州大学・システム情報科学研究院・准教授
研究者番号： ３０２４３８５１

(3)連携研究者
なし

(4)研究協力者
亀田 雅之（KAMEDA Masayuki）
大学入試センター・研究開発部付・学術支援員