2I3-4

大学入試センター試験歴史科目の自動解答

Solving History Problems of the National Center Test for University Admissions

狩野 芳伸[*1]
Yoshinobu Kano

[*1] 科学技術振興機構 さきがけ
JST PRESTO

We suggest a generic system that determines true or false for an input text from given knowledge source. This system does not depend on any specific domain nor language, works without training. We applied this system to History problems of the National Center Test for University Admissions (Center Test). Our system obtained the best result in the Mock Center Test challenge held by the Todai Robot project.

1. はじめに

大学入試問題に解答するには、人間のさまざまな知的処理が必要である。そのため大学入試問題の自動解答に挑戦することは、機械が人間の知的処理に迫るための重要なチャレンジであると考えられる。もし現在の技術で十分に正答できないのであれば、人間には可能な基礎的な処理が機械には未だできないということであり、知的処理のシステム一般に必要な技術が不足している可能性が高い。また、大学入試問題、特にセンター試験は採点基準が明確であり、皆が合意できる評価基準があるという点で特徴的である。すなわち、大学入試問題に自動解答するタスクは、現在の人工知能技術の達成度を測るベンチマークになりうると同時に、どのような技術が不足しているかを洗い出す非常に有用なタスクであると考えられる。

国立情報学研究所を中心とする「ロボットは東大に入れるか（東ロボ）」プロジェクト [1]では、大学入試試験の自動解答に挑戦している。本稿ではそのうち、我々の実装した大学入試センター試験の歴史科目の自動解答器について述べる。

歴史科目は暗記すれば解けるから、計算機による自動解答は容易だと思われがちだが、必ずしもそうではない。たとえばNTCIR-10 RITE2 タスク[Watanabe 12]では、センター試験社会科の問題を素材に含意関係認識のタスクとして精度を競ったが、試験の成績評価で最良の結果は 34.29 と、ほとんどが 25%がベースラインの四択問題であることを考えるとそれほど良い結果は得られていない。また、センター試験の問題文を加工して利用している分、元のセンター試験に直接解答するほうが難しい。

本稿で対象にする問題は、入力の正誤を判定する問題として一般化できる。正誤を判定する問題は、質問応答や検索のタスクに近いものがあるが、誤りを検出できなければならないという点で大きく異なる。我々はこの点に着眼し、入力内の誤りを引き起こすキーワードを特異的に検出するペナルティつきの知識源検索手法を提案する。我々の自動解答システムは、(1)入力からのキーワード抽出、(2)知識源内のキーワード分布によるペナルティつきスコアリング、(3)スコアからの解答生成で構成される。

正誤の判定を行うシステムとして、機械学習による手法が考えられる。高性能な質問応答システムの多くは、内部で機械学習を用いている [Shima 08]。しかし、機械学習が高性能を発揮するためには十分な学習データが必要である。これに対し我々は、知識源の記述中には正誤判定の証拠となる部分が一度しか現れないことが多いと考えた。たとえば教科書中に、同じ歴史的なイベントの記述が重複して現れることは非常に稀である。無償で利用できる Wikipedia の場合でも、やはり重複した記述は少ないと考えられる。我々のシステムは、証拠の記述が一度しか出現しない場合でも高い精度で正答することができる。

このシステムを用いて、東ロボプロジェクトで行われた代々木ゼミナール模擬試験（代ゼミ模試）チャレンジタスクで、世界史 58 点(偏差値 55)・日本史 56 点（偏差値 56）と参加者中最高の得点を達成した。また、RITE2 タスクのデータによる評価でも、RITE2 の参加者のうち最高のものを正答率で 9 ポイント上回る性能を達成し、現在のところ世界最高の性能である。

本稿で提案する手法は、センター試験の解答に限ったものではなく、与えられた知識源を用いてその範囲で入力の正誤を判定できる汎用のシステムである。特定のドメインにも言語にも依存せず、訓練も不要な手法であるため、応用の可能性は広いと考えられる。

本稿では、2 節で我々の手法と実装について述べ、3 節で実験とその結果、4 節で既存研究との比較、5 節で今後の展望を述べ、6 節で締めくくる。

2. 手法と実装

2.1 正誤を判定する問題

本節では、「四択から正しいもの（あるいは誤ったもの）を選べ」、「正誤の組合せのうち正しいものを選べ」というような正誤の判定を行う問題の解答手法について述べる。

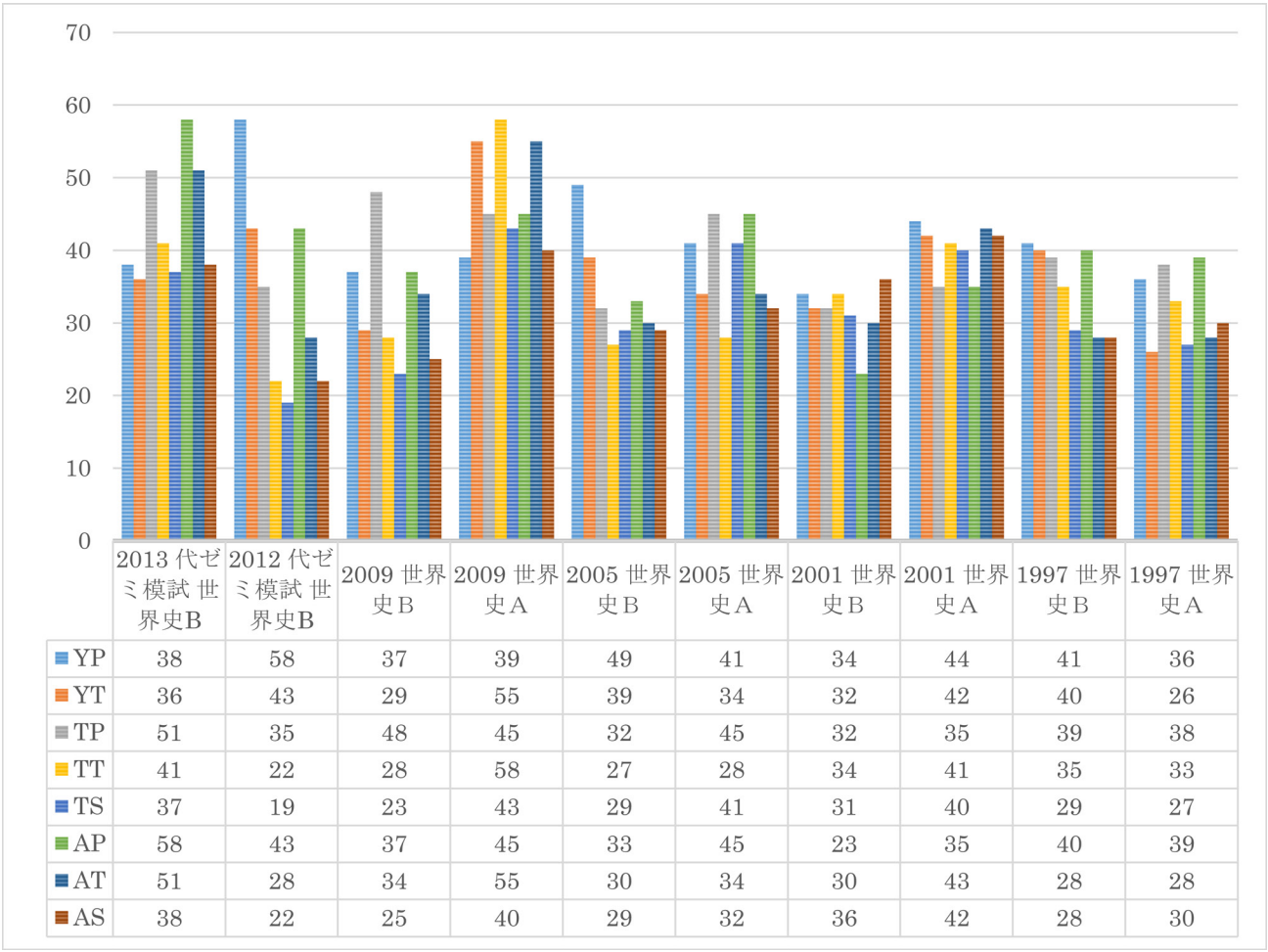
(1) 問題からのキーワード抽出

入力テキスト中で関連すると思われる部分から、キーワードを抽出した。RITE2 タスクのように、入力全体が関連すると思われる場合は、全体を対象として抽出する。センター試験の場合は、選択肢テキストに加え、小問などの共通部テキストがあり、さらに下線部などで別の箇所を参照していることがある。このような共通部や参照箇所も適宜対象としてキーワード抽出を行った。

キーワードの抽出は、日本語 Wikipedia の見出し語との完全マッチで行った。同一テキストにキーワードが重なる場合は、より長いキーワードを採用した。Wikipedia の見出し語には、たとえば「とら」など必ずしも適当でないものが含まれるため、そうした不適当な見出し語を前もって 100 個程度目視で除去した。

[1] http://21robot.org/

連絡先：狩野 芳伸 kano@nii.ac.jp http://kachako.org/kano/



	2013 代ゼミ模試 世界史B	2012 代ゼミ模試 世界史B	2009 世界史B	2009 世界史A	2005 世界史B	2005 世界史A	2001 世界史B	2001 世界史A	1997 世界史B	1997 世界史A
YP	38	58	37	39	49	41	34	44	41	36
YT	36	43	29	55	39	34	32	42	40	26
TP	51	35	48	45	32	45	32	35	39	38
TT	41	22	28	58	27	28	34	41	35	33
TS	37	19	23	43	29	41	31	40	29	27
AP	58	43	37	45	33	45	23	35	40	39
AT	51	28	34	55	30	34	30	43	28	28
AS	38	22	25	40	29	32	36	42	28	30

図 1 センター試験世界史の模試および過去問解答採点結果
教科書と snippet の組合せの結果を示す。
二文字のラベルは教科書（Y: 山川、T: 東京書籍、A: 両方）と snippet 単位(P:段落、T: 小節、S: 節）を示す。

Wikipedia でリダイレクト関係にある見出し語グループは、同一のキーワードとみなして処理を行った。

実際のキーワード抽出には、形態素解析器 MeCab[2]の Java 移植版である Kuromoji[3]のユーザ辞書に、前述の Wikipedia 見出し語から得られた単語セットを登録したものを利用した。

(2) 知識源の検索とスコア付け

正誤を判定するための知識源には、文章構造が明示されていることを前提にし、この文書構造を単位として検索とスコア付けを行った。たとえば教科書データの場合、段落・小節・節が明示的に記述されているので、これらのいずれかを snippet 単位として用いた。

各キーワードには知識源内の出現回数の逆数を重みとして割り振り、各入力内で重みの和が同一になるよう正規化した。

次に各 snippet 内における(1)で抽出したキーワードの分布を計算し、これを用いて各 snippet にスコア付けを行った。キーワードの分布は、(1)と同じ手法によりキーワード抽出を行い計算した。基本的なスコアは、snippet 内で出現したキーワードの重みの和である。これにより、最も関連があると思われる snippet を検出する。ここから、誤った入力に対応するペナルティを減算する。当該 snippet には出現しないが、ほかの snippet に出現するキーワードがあった場合、その重みをペナルティとして減算し、これを snippet のスコアとする。この計算を全 snippet について行う。

(3) 解答の生成

入力に対して(2)で計算された各 snippet スコアの最大値を用いて解答を生成した。

入力の正誤を判定する二値分類問題の場合は、閾値を定め、閾値と前述の最大スコアの比較により正誤を判定した。正誤の組合せを選ぶ場合、たとえば選択肢 a,b に対し「a 正-b 誤」のような場合は、それぞれ対応する選択肢について正誤判定をしたうえで、解答する組み合わせを判定した。

四択の場合で正しいものを選ぶ場合は、各選択肢のスコアの最も大きいものを解答とした。誤ったものを選ぶ場合は、スコアの最も小さいものを解答とした。選択するものが正しいものなのか誤ったものなのかは、文字列パターンで判断した。現在まで試みたデータの範囲では、全問題についてどちらなのか正しく判断できている。

[2] http://mecab.sourceforge.net/

[3] http://www.atilika.org/

2.2 その他の問題

センター試験には、前節で述べた正誤の判定以外に「年代順に並べよ」という年代を問う問題がある。これについては、前節で述べた手法と同様に、最も関連すると思われる snippet を取得したうえで、その snippet 内の年代表記を抽出し、年代順に並べることで解答とした。

テキスト以外に画像や図表を解釈する必要のある問題については、解答を行わなかった。

2.3 解答器の実装フレームワーク

解答器は、統合言語処理システム Kachako [Kano 12][狩野 12a]互換の UIMA [Ferrucci 06] コンポーネント群として実装した。

Kachako[4]は UIMA 準拠の統合プラットフォームと互換 UIMA コンポーネント群を提供している。センター試験の解答および質問応答システム一般のコンポーネント化も行っている[狩野 12b]。Kachako は自然言語処理におけるユーザタスクを徹底した自動化によりサポートすることを目指したシステムで、プラットフォームおよびコンポーネントのインストール、ワークフロー生成、ウェブサービス展開、大規模処理、結果の視覚化、汎用比較評価などを全自動でサポートする機能を提供している。UIMA(Unstructured Information Management Architecture) [Ferrucci 06]は非構造化データ処理の相互運用性のための国際標準フレームワークである。

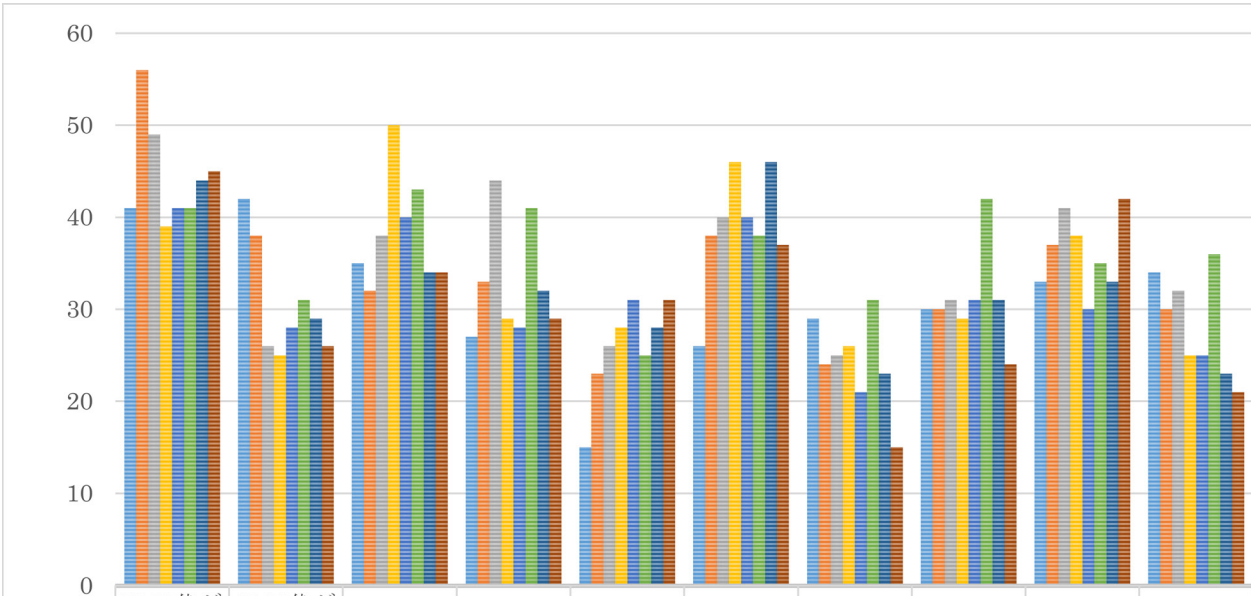
3. 実験と結果

3.1 知識源

知識源として、東京書籍および山川出版の電子化された教科書データを利用した。東京書籍のみ、山川出版のみ、および双方を同時に用いた場合を試みた。同時に用いる場合は、全体を連続した一つの知識源とみなして処理した。

3.2 センター試験形式の問題による評価実験

東ロボプロジェクトで作成された大学入試センター試験問題アノテーション済みデータ(問題構造・問題分類)を用いて実験



	2013代ゼミ模試 日本史B	2012代ゼミ模試 日本史B	2009 日本史B	2009 日本史A	2005 日本史B	2005 日本史A	2001 日本史B	2001 日本史A	1997 日本史B	1997 日本史A
YP	41	42	35	27	15	26	29	30	33	34
YT	56	38	32	33	23	38	24	30	37	30
TP	49	26	38	44	26	40	25	31	41	32
TT	39	25	50	29	28	46	26	29	38	25
TS	41	28	40	28	31	40	21	31	30	25
AP	41	31	43	41	25	38	31	42	35	36
AT	44	29	34	32	28	46	23	31	33	23
AS	45	26	34	29	31	37	15	24	42	21

図 2 センター試験日本史の模試および過去問解答採点結果

教科書と snippet の組合せの結果を示す。

二文字のラベルは教科書（Y: 山川、T: 東京書籍、A: 両方）と snippet 単位(P:段落、T: 小節、S: 節）を示す。

[4] Kachako 公式ウェブサイト http://kachako.org/ を参照。

を行った。センター試験の過去問と、代ゼミ模試の問題が同じXML フォーマットおよび DTD に基づいてアノテーションされている。図 1 および図 2 に結果を示す。これは実際の試験と同じ配点による採点であり、各年度 100 点満点である。年度・科目や知識源パターンによってばらつきがあるが、各科目各年度での最高得点はおおむね 40 点から 60 点の範囲で分布している。

3.3 NTCIR-10 RITE2 データセットによる評価実験

NTCIR-10 RITE2 Exam Search タスクのデータでも評価を行った。RITE2 Exam Search では、与えられた文章が正しいかどうかを、Yes/No の二値で判定する。RITE2 Exam Search はセンター試験の社会科目からデータセットを作成しており、評価値としてもともとの 4 択問題等の形式で評価した正答率を算出している。表 1 に結果を示す。我々のシステムは RITE2 タスク開催時に最良の結果であった参加者よりもよい結果を得ることができた。

RITE2	参加者最高値	我々の結果
正答率	34.26	43.51

表 1. NTCIR-10 RITE2 タスクデータでの評価結果

4. 既存研究との比較

4.1 検索エンジンによるベースラインシステムとの比較

本稿で提案した手法は、誤った選択肢に対しより適切にスコア付けをするためのペナルティを用いている。一方、我々は一般的な TF/IDF ベースの検索エンジンの確信度出力順ランキングによる自動解答を試みた[狩野 13]。検索エンジンには Indri[5] を用い、教科書テキストをインデックスして用いた。これは本稿のペナルティ手法と比較するベースラインシステムとして適当なものである。結果は 30 点台から 40 点台と、本稿の手法のほうが常におおむね 10 ポイント程度良い結果であった。

4.2 Factoid 型質問応答システムとの比較

[石下 14]は問題文を質問形式に変換し、既存の Factoid 型質問応答システムに入力してその結果から問題解答を試みている。結果は 30 点台から 40 点台であり、本稿の手法のほうが良い結果となっている。Factoid 型質問応答システムでは正誤解答には不要なモジュールが多く、固有表現のカテゴリも対応が不十分であったことが主な要因と考えられる。

5. 今後の課題

今後の課題としては、文脈を考慮してキーワードの選択を行うことが挙げられる。キーワードの適切な選択には、科目・設問・選択肢それぞれのレベルでのトピックが影響していると考えられるので、複雑なトピック解析が必要である。また、もう一つの課題として、文章構造以外の snippet 利用が挙げられる。構文や意味的な関係を考慮した snippet の生成などを検討したい。

年代順に並べる問題では、時代によってはそもそも年代がはっきりせず、教科書中の記述順をもって年代順を示唆する場合がいくつもみられたため、このような場合の対応が必要である。

知識源の選択による影響も大きい。さらに異なる知識源の場合にどう振る舞いが変化するか実験を試みたい。

6. おわりに

本稿では、我々の実装した大学入試センター試験の歴史系科目の自動解答システムについて述べた。我々のシステムを代ゼミ模試チャレンジタスクの世界史・日本史および NTCIR-10 RITE2 タスクに適用し、いずれも最高の性能を達成した。

歴史系の試験問題以外にも、さまざまなシステムの応用が考えられる。たとえば、技術マニュアルを知識源とする操作支援システムや、医学書を知識源とする医療診断支援などである。

入試偏差値 50 を超えたということは、その意味で人並み以上である、ということができる。一方で、得点はまだ半分程度であり、大いに改善する余地があると思われる。

謝辞

ご協力くださった東ロボプロジェクトメンバーの各氏に深謝申し上げる。大学入試センター試験の問題および解答データについては、株式会社ジェイシー教育研究所が販売する「大学入試センター試験問題データベース センターTen 2011 通常版 全教科セット」を利用した。また、東京書籍株式会社および株式会社山川出版社の教科書データを利用させていただいた。代々木ゼミナールの模擬試験のデータは、学校法人高宮学園に提供いただいたものを使用した。

参考文献

[Ferrucci 06] Ferrucci, D., Lally, A., Gruhl, D., Epstein, E., Schor, M., Murdock, J. W., Frenkiel, A., Brown, E. W., Hampp, T., et al. (2006) Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report.

[Ferrucci 12] Ferrucci, D. (2012). Introduction to "This is Watson". *IBM Journal of Research and Development*, 56(3.4), 1:1–1:15.

[Kano 12] Kano, Y. (2012) Kachako: a Hybrid-Cloud Unstructured Information Platform for Full Automation of Service Composition, Scalable Deployment and Evaluation. *In the 1st International Workshop on Analytics Services on the Cloud (ASC), the 10th International Conference on Services Oriented Computing (ICSOC 2012).*

[Shima 08] Shima, H., Lao, N., Nyberg, E., & Mitamura, T. (2008). Complex Cross-lingual Question Answering as Sequential Classification and Multi-Document Summarization Task. In NTCIR-7 Workshop.

[Watanabe 12] Y. Watanabe, Y. Miyao, J. Mizuno, T. Shibata, H. Kanayama, C.-W. Lee, C.-J. Lin, S. Shi, T. Mitamura, N.Kando, H. Shima, and K. Takeda, (2012) "Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10," *in the 10th Conference of NII Testbeds and Community for Information access Research (NTCIR-10),* 2013, pp. 385–404.

[狩野 12a] 狩野 芳伸. Kachako: 誰でも使える全自動自然言語処理プラットホーム. 2012 年度人工知能学会全国大会（第 26 回）. 山口県教育会館, 2012 年 6 月 12 日.

[狩野 12b] 狩野 芳伸. 統合研究基盤：質問応答システムの互換コンポーネント化による再利用性向上と開発自動化支援. 人工知能学会誌,27(5) 特集号「ロボットは東大に入れるか」, pp.492-495. 2012 年 9 月.

[狩野 13] 狩野 芳伸, 神門 典子. 大学入試センター試験は教科書の肯定的表現密度のみで解けるか. 情報知識学会第 21 回（2013 年度）年次大会. お茶の水女子大学. 情報知識学会誌 Vol. 23 (2013) No. 2, pp. 179-184. 2013 年 5 月 25 日.

[石下 14] 石下円香, 狩野芳伸, 神門典子. 質問応答システムを用いた多岐選択式問題の解答器の作成に関する研究. 情報処理学会第 215 回自然言語処理研究会. 国立情報学研究所. 2014 年 2 月 6 日.

[5] http://www.lemurproject.org/indri/