N-030

知識情報処理に基づく入試予測の可能性

A Possibility of Prediction about Entrance Examination Based on Knowledge Management

井出　明†
Akira Ide†

1. 研究の背景

試験問題を予測することは、古代より、人間にとって切実な欲求であった。特に、第二次世界大戦後、汎用のコンピュータが実用化してからは、「コンピュータ分析」と銘打った受験対策書が数多く出版され、たくさんの受験生たちを惹きつけ、また惑わせもしてきた。本稿ではフリーの言語解析ソフトを用い、知識情報処理の手法を用いて、ある一定期間の入試問題の分析を行った上で、その分析に基づく予測が的中しているのかという点の検証までを行う。

これまで多くの受験産業が「予測」を行っているものの、その予測がどの程度当たっていたのかという検証作業はほとんど行われていない。今回の実験と観察を通して、「コンピュータによる入試予測」なるものが、どの程度実用性のあるものかが推定できるようになる。

2. 実験内容

実験に当たっては、2007 年度から 2009 年度のセンター試験の「倫理」（本試験）を素材として用いた。手法としては、これを“KHCoder”にかけ、上位語が 2010 年にも出題されていたのかを検証するという方法をとった。

言語的な解析をするためには品詞を絞り込む必要があるが、副詞や形容詞については入試予測の対象としてなじまないため、品詞については、名詞（固有名詞、組織名、人名、地名を含む）、サ変名詞、形容動詞、動詞、可能の意味を持つ副詞、のみを抽出した。

語の分析に当たっては、段落単位で語の出現回数を数えた。これは、純粋に言葉の数を数えていった場合、同じ設問の中で人名等の固有名詞が多く出現するため、一問しか出されていない言葉であっても 5－6 回のカウントがなされてしまいかねないからである。

表 1　出力結果の一部

名詞	文書数(段落)				人名	文書数(段	地名	文書数(段	未知語	文書数(段
人間	89	悩み	6	民族	3	ベンサム	4	日本	25	アッラー	12
自分	42	文化	6	落語	3	荻生徂徠	4	仏	6	カント	7
他者	36	法則	6	利己	3	アリストテ	3	中国	4	ヘーゲル	6
世界	35	万物	6	お盆	2	ゴータマ	3	インド	3	ブッダ	5
理性	32	カルト	5	エゴイズム	2	ソクラテス	3	ユダヤ	3	NGO	4
思想	30	宇宙	5	ガス	2	王	3	江戸	3	NPO	4
近代	28	営み	5	クオリティ	2	内村鑑三	3	オタワ	2	キルケゴー	3
自己	27	患者	5	医療	2	孔子	3	ギリシア	2	ギリガン	3
個人	26	企業	5	怨み	2	イエス	3	パース	2	クルアーン	3
精神	21	権利	5	王国	2	コーラン	3	ヨーロッパ	2	アウグステ	2
責任	21	原因	5	王政	2	ニュートン	3	欧米	2	エピクロス	2
意志	20	古典	5	家事	2	パース	3	十七条	2	グロティウ	2
主義	19	国家	5	家庭	2	エリクソン	2	独	2	コギト	2
感情	18	国土	5	戒め	2	キャロル	2			サルトル	2
責務	17	所得	5	格差	2	ジェームス	2	その他名詞	文書数(段	デルフォイ	2
道徳	17	浄土	5	核兵器	2	パスカル	2	イスラーム	5	ホイジンガ	2
ケア	15	身体	5	学校	2	フロイト	2	キリスト	2	ムスリム	2
現代	15	人権	5	官僚	2	ベルクソン	2			ムハンマド	2
人生	14	世代	5	喜怒哀楽	2	山崎闇斎	2			メシア	2
法律	14	未来	5	基盤	2	井上哲次郎	2			ヤハウェ	2
環境	13	来世	5	機会	2	新渡戸稲造	2				
仏教	13	スト	4	教会	2	司馬	2				
ルール	12	旧約	4	軍国	2	植村正久	2				

†追手門学院大学　Otemon Gakuin University

3. 実験結果

出力結果には当然「一つ」や「記述」などの問題の指示部分に関する言葉も多く含まれるため、それは手作業で除いた。また人名についても、苗字と名前が別に出力されたため、これも手作業で結合した。さらに、「イエス」などは一般名詞として出力されるが、これを固有名詞に移動して表をリライトした。最後に、「解散」「生活」などといった一般的な語を排除し、考察段階に進むことにした。

今回は分析の対象期間として３年という期間を設定しているが、この中で２回以上登場している言葉は頻出語であることが推定できる。また一度しか出てきていない言葉はイレギュラーで出てきていることも考えられるため、次章の考察の対象とするのは２回以上登場した言葉に限定した。

出力された表の一部を表１として示す。

4. 考察

出力された単語は、「人間」という言葉が最も多く、この教科の特質を物語っているともいえる。出題の予測とは直接関係がないが、副詞としての「死後」や、動詞としての「生きる」なども上位にあり、出題者がこの科目を通じて高校生に何を学ばせたいかというメッセージも読み取ることができる。品詞別に分析していった場合、サ変名詞・可能の意味を持つ副詞・形容動詞は、この科目の特性をつかむうえでは確かに重要ではあるものの、「出題の予測」という観点からは情報を読み取ることが難しいため、最終的には除外して分析を行った。

出題予測の最も大きな手掛かりは人名地名等であるから、まずそこから次年度予測を試みた。具体的には、

①「アッラー」と「イスラム教」・・・頻度 5 と 12
②カントとヘーゲル（ドイツ観念論）・・・頻度 7 と 6
③ベンサム・・・・・頻度４
④荻生徂徠・・・・・頻度４

について、2010 年にどのように扱われているのかを検証した。

①については、出題があった。②は全く触れられていなかった。③は不要選択肢の中に登場した。④は直接の出題がなく、江戸時代の儒学という括りでの同一性しか確認できなかった。

以上から言えることは、ある言葉に着目して予測をするのではなく、出題されたドメイン（領域）を注視し、その領域からの出題があるのではないかという予測を行うほうが理に叶うことがわかる。語句単位で細かいコンピュータ分析をするよりも、俯瞰的な視野から出やすい領域を講じた方が対応として適切であると言えるかもしれない。

5. 今後の展望

今回は、初めての試みということで、「倫理」を素材として用いたが、この科目はあまり時事性がなく出題傾向が安定していることで知られている。今後は、時事性の強い「政治・経済」についても同様の分析を行うことで、この度得られた考察と矛盾がないか再度考えてみたい。

【参考文献・資料】

宮崎市定：『科挙―中国の試験地獄』岩波新書（1963）
センター試験倫理過去問（2007－2010）
樋口耕一：“KH　Coder” http://khc.sourceforge.net/