海外トピックス「コンピュータによるエッセイの自動採点システム E-rater について」

大学入試センター 研究開発部
石岡 恒憲

昨年 (平成12年) の2月1日から11月30日まで、文部省長期在外研究員として、主にカーネギーメロン大学コンピュータサイエンス学科 (アメリカ/ピッツバーグ) にて「自然言語データベースからの統計的手法を用いた意味的検索・分類についての調査研究」を行った。その一環として、帰途、プリンストンにある教育テストサービス (ETS, Educational Testing Service: 以下ETSと略記) に立ち寄り、エッセイ (小論文) のための自動採点システムであるE-raterについて情報収集を行った。本稿はその調査報告の要約である。

1 E-rater とはなにか

E-raterはもともとはジル・バーシュテイン博士を始めとする幾人かの研究者によってETSで開発された。現在の開発および運用は、ETSの補助機関である教育テストサービステクノロジー (ETS Technologies) に移管されている。E-raterはビジネススクール (経営大学院) への進学希望者を対象とした適性試験であるGMAT(Graduate management Admission Test) におけるエッセイの採点に用いられている。1997年7月から、現在までに73万人以上の試験者に対して利用されてきているという。E-raterは得点結果だけでなく、作文能力が向上するよう、どの部分が良くないかを提示する機能を有する。

GMATのエッセイ問題は以下の2つの形式からなり、それぞれ1問ずつ出題される。試験時間は各30分である。

1. **論点 (issue) の分析 (analysis of an issue):** 論点に対して自分の意見を述べる問題である。効果的で説得力のあるエッセーであることが求められる。
2. **議論 (argument) の分析 (analysis of an argument):** 議論に対する批判と、どうすれば議論が良くなるかを述べる問題である。論理的批判能力と分析力が求められる。

ここで、E-raterのみでエッセイの最終評点が決められるのではない、ことに注意する必要がある。まず1つのエッセイは、専門家とE-raterによってそれぞれ独立に6点満点で採点される。次に、各評点に2点以上の差があった場合に限って、別の専門家が審査して最終点を決める。1点差の場合は、専門家の採点が優先するようである。専門家とE-raterによる採点の一致率（1点差以内）は、現時点でのホームページでの記載によれば97%を越える。1998年に書かれた研究レポートでは、その一致率は89%であり、運用開始から、かなり性能が向上していることがわかる。

2 E-raterによる採点基準

E-raterは以下の3つの基準によりエッセイを評価する:

構造: 文法の多様性、すなわちフレーズや文節および文の配列が多様な構造で表現されていること。

エッセイ中の文はすべてMSNLP(1997)などの適当な構文解析プログラムによって構文解析され、構成節、従属節、不定節、関係節が判別される。それら各節の数や、仮定法における助動詞(would, could, should, might, may)の出現回数などの情報を得ることができる。これにより、文ごとに構文構造タイプが決まり、それらの個数、あるいは出現比率を調べることで、1つのエッセイにおける構文多様性の尺度を得ることができる。

組織化: アイディアが理路整然と表現されていること。たとえば修辞的な表現、あるいは文や節の間の論理的な接続法が使われているか、議論がよく掘り下げられているかを判定する。

エッセイの議論を評価するために、まずエッセイを談話(discourse)と呼ばれる意味的な議論の構成単位に分割する。この談話は、形式上の段落とは必ずしも一致しないことに注意する。

談話単位に分割する方法には幾つかの方法があるが、E-raterでは（一般的ではあるが最も古典的な）キュー・ワード（cue word; きっかけ語）による方法が用いられている。(Quirk, et al, 1985)。たとえば、“In summary”や“In conclusion”は要約を示す形容詞句であるとか、“perhaps”や“possibily”は議論を掘り下げるときに信念や考えを示す語である、といったものである。“this”や“these”は、書き手が話題を変えずに関連をもたせるためにしばしば使われる。また新しい話題を始めることを示す不定詞句も同様に見つけることができる。

このようにして自動的に分割した談話単位に対し、注釈プログラム(APA, Annotation Program)によって「議論の始まりを示している」

あるいは「議論を掘り下げている」などのラベルを作成する。また「並列」と「対比」のような修辞的な関係を同定することも行う。

これにより、E-rater は、アイディアが理路整然と表現されているか、あるいは議論がよく掘り下げられているかを判定する。

内容: トピックに関連した語彙が用いられていること。

良いエッセイは、与えられたトピックに関連があって、内容の乏しいエッセイに比べて専門的で正確な語彙が用いられる傾向がある。したがって、良いエッセイは単語の選択において、他の別の良いエッセイ（模範エッセイ）と似ていると考えられる。そこで E-rater は、人間が評点してその結果、評点 1 から 6 までとなった各カテゴリーに含まれるトレーニング用の複数のサンプル・エッセイに含まれる単語と、採点するエッセイに含まれる単語とを比較することによって、字句と内容の評価を行う。

ここで使われている文書処理技術は、「同一文書で何度も出現する単語の重みを大きくする TF(Term Frequency) 法」と、「どの文書にも現われる（いわゆる一般的な）単語の重みを小さくする IDF(Inverse Document Frequency) 法」を組み合わせた TF・IDF 法である。これを用いて、採点エッセイの（評点 1 から 6 までの評点を有する）サンプル・エッセイとの類似度、すなわちコサイン相関を計算する。

このような方法は、一般には「ベクトル空間モデル」による方法と呼ばれる。基本となる考え方は、互いに似たベクトルをもったもの同士は、互いに近い関係にある、と判断するものである。

E-rater は最終的なエッセイの評点を、人間の採点を目的変数とする線形の重回帰モデルにより算出する。説明変数は、前述の「構造」、「組織化」、「内容」から得られる 57 の説明変量である。

ただ全部の説明変数がいつも使われているわけではなく、通常の場合は、このうち 8-12 変量が用いられている。エッセイの内容によって、8-12 変量の組み合わせは異なるのであるが、現在、（組み合わせの異なる）75 のモデルが使われている。

75 のモデルのうち、最も使われる変量は以下の通りである。

1. 単語の出現頻度ベクトルから得られるコサイン相関スコア
2. いわゆる一般的な単語の重みを低くした単語の重みベクトルから得られるコサイン相関スコア
3. 仮定を表す助動詞の数
4. 仮定を表す助動詞の数の全体の語数に対する割合
5. 議論を深めるためのきっかけ語の数

6. 議論の始まりに現われる代名詞の数
7. 議論の始まりに現われる補足句 (complement clause) の数
8. 議論の始まりに現われる要約語 (summary words) の数
9. 議論の始まりに現われる詳細語 (detail words) の数
10. 議論を深める修辞句 (rhetorical words) の数

これ以外の変数や、また重回帰係数については、部外秘であるらしく教えていただけなかった。なお重回帰係数は、新しいテストエッセイを評点するたび更新される。

3 おわりに

ETSの部外者であっても、http://www.etstechnologies.com/html/criteriondemo-all.htm にて デモ使用 を申請し、ID とパスワードを入手すれば、E-rater のデモ版を利用することができるようである。

また、E-rater についてより詳しく知りたい方は、Web から以下の資料を入手することができる:

- 巻末参考文献に示されている論文と 2 セットのスライド:
  E-rater で用いられている自然言語技術についてかなり詳細に知ることができる。
  http://www.ets.org/research/erater.html
- GMAT における実際のエッセイ問題
  http://www.gmat.org/
- GMAT の実際のエッセイ問題についての模範解答
  http://www.west.net/~stewart/awaissue.htm

参考文献

[1] Burstein, Jill and Daniel Marcu (2000). Towards Using Text Summarization for Essay-Based Feedback. Le 7e Conference Annuelle sur Le Traitement Automatique des Langues Naturelles TALN'2000, Lausanne, Switzerland, October 2000.

[2] Powers, D.E., Burstein, J.C., Chodorow, M., Fowles, M.E., & Kukich, K. (in press). Comparing the validity of automated and human essay scoring (GRE No. 98-08a). Princeton, NJ: Educational Testing Service.

[3] Burstein, Jill C., Susanne Wolff and Chi Lu. (1999). Using Lexical Semantic Techniques to Classify Free-Responses. The Depth and Breadth

of Semantic Lexicons. Edited by Nancy Ide and Jean Veronis. Kluwer Academic Press.

[4] Burstein, Jill and Martin Chodorow (1999). Automated Essay Scoring for Nonnative English Speakers. Joint Symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies, Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing, College Park , Maryland, June, 1999. Slides from the presentation are also available.

[5] Burstein, Jill, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris (1998). Automated Scoring Using A Hybrid Feature Identification Technique. In the Proceedings of the Annual Meeting of the Association of Computational Linguistics, August, 1998. Montreal, Canada.

[6] Burstein, Jill, Karen Kukich, Susanne Wolff, Chi Lu, & Martin Chodorow (1998). Enriching Automated Scoring Using Discourse Marking. In the Proceedings of the Workshop on Discourse Relations & Discourse Marking, Annual Meeting of the Association of Computational Linguistics, August, 1998. Montreal, Canada.

[7] Burstein, Jill, Karen Kukich, Susanne Wolff, Chi Lu, & Martin Chodorow (1998). Computer Analysis of Essays, NCME Symposium on Automated Scoring, April 1998, Presented by Karen Kukich

[8] Burstein, J., Randy Kaplan, Susanne Wolff, and Chi Lu. (1996). Using Lexical Semantic Techniques to Classify Free-Responses. In Proceedings from the SIGLEX 1996 Workshop, Annual Meeting of the Association of Computational Linguistics, University of California, Santa Cruz.

[9] MSNLP(1997) http://research.microsoft.com/nlp/