# *Criterion* SM Online Essay Evaluation:

An Application for Automated Evaluation of Student Essays

## 1. Introduction

- 望ましいライティングの指導法:
  Write →feedback→ revise
- 教師の負担を軽減し、学習者に対してエッセイライティングの機会をもっと設定できるように開発された
- Criterionの構成
  自動採点機能＋診断的フィードバック

## 2. Application Description

- *E-rater* スコアリングエンジン
  TOEFL スコアリングガイドの評価規準として明示されている言語的特徴を反映させている
- 具体的な評価項目:
  syntax, discourse, topical content, lexical complexity

## 2.1. The *E-rater* scoring engine

- 統語解析器が統語構造を評価する
- 概念構成別にディスコースマーカーの辞書を備え、出現の有無により評価する
- 主題別の内容評価については、主題に応じて使用されることが予想される語彙に注目し、使用頻度による重みづけを行い、その量をベクトル化して分析・評価する

## (参考) Content Vector Analysis

- あるエッセイでFi回使用されているｉ という単語の重みづけWi:
  Wi = (Fi/MaxF) * log(N/Ni)
  MaxF(当該のエッセイで最も使用されている語の出現回数)
  N(エッセイの総語数)
  Ni(全エッセイのiという単語の出現回数)

## (参考) Content Vector Analysis

- Wi = (Fi/MaxF) * log(N/Ni)
- Wi1=(Fi1/MaxF1)*log(N/Ni)
  Wi2=(Fi2/MaxF2)*log(N/Ni)
  …
  Wi6=(Fi6/MaxF6)*log(N/Ni)
- Wi とW1～W6のcosine値を計算し、最大のものをscore point value(1-6)、最高ランクのエッセイ(通常はスコア6)W6でWiを割った値(cosine correlation value)

## 2.1. The *E-rater* scoring engine

- レトリックタイプ別に使用が予想される語の辞書を備え、エッセイで使用された全ての語との対応状況により評価する
- 使用語彙の難易度については、特徴的な語の数、1語あたりの平均文字数、5～6文字によって構成される長めの単語の数などを指標として評価する

## 2.1. The *E-rater* scoring engine

- スコアについては、まず270のエッセイを人間が6段階で採点し、それぞれの段階に評定されたエッセイについて50の言語的特徴をステップワイズ法によりカテゴリー化する。そして、カテゴリー化された特徴を採点対象となるエッセイの特徴に等価してスコアを与える

## 2.2 *Critique* Writing Analysis Tools

- Agreement errors, verb formation errors, wrong word use, missing punctuation, typographical errors を発見し、フィードバックを与えてくれる
- Bigram(隣りあう語と品詞の組み合わせ)単位でコーパス上の語彙と比較し、出現し得ないものを「誤り」と判定させる

## 2.2 *Critique* Writing Analysis Tools

- 「誤り」判定のcomplementary methodとて関連性のある語、関連性のない語をによる判定方法も備えている
- 隣り合う語による判定では誤りとなってしまうような用例については”filter”を設定する
- 同音異義語のようなConfusable wordsについては左右の隣り合う2語で判断する

## 2.2 *Critique* Writing Analysis Tools

- 「受動態」「長すぎる(短すぎる)文」「過度の同一語の繰り返し」などを望ましくないstyleとして指摘する
- レトリックに応じて、想定されるdiscourse elementsの有無や文章構成の適切さを判断して、フィードバックする

## 3. Evaluation Criteria

- *Criterion (E-rater, Critique)* が学習者に対して有益で信頼できるフィードバックを与えていると判断するための規準
- 2名の評定者による採点結果の一致度(golden standard)と評定者とシステムによる採点一致度(agreement)を比較する
- 約97%の一致度が見られたケースもあるが、*E-rater*の一致度評価の基本線は75%～80%

## 3.2. *Critique* performance evaluation

- *Critique*がフィードバックを与えるべきであると判断した箇所と人間がフィードバックを与えるべきであると判断した箇所の「数」を比較する(precision)
- *Critique*がフィードバックを与えるべきであると判断した箇所と人間がフィードバックを与えるべきであると判断した箇所の内、「一致した数」を比較する(recall)

## 3.2. *Critique* performance evaluation

- Grammar, Usage, Mechanicsのprecisionは90～100%
- Recallに関しては、bigram, confusable words ごとに結果が異なり、例えば「主語と動詞の一致」では人間とシステムが指摘した箇所の一致度は40%だったが、所有格のアポストロフィー70%, confusable words 71%であった

## 3.2. *Critique* performance evaluation

- 同一語の繰り返しに関する判定では、precision, recall に加え、F-measureを用いる
- 300のエッセイの内、2名の評定者によって繰り返しがあると判定されたエッセイについて総語数に対する繰り返し語の割合を計算したところ5%以上が両者に共通する基準となっていた。
- これに該当するエッセイについては Precision(0.27), recall(0.54) , F-measure(0.36)

## 3.2. *Critique* performance evaluation

- Discourse内の配置によってそれぞれの文に与えられるラベルに関しては、1462エッセイに対して評定者が行ったラベルづけの結果と比較して、Precision(0.71), recall(0.70) , F-measure(0.70)という結果になった

## 4. Application Use

- 2002年9月の提供を開始し、同年12月には利用者数が50,000人を越えた。アメリカ国内だけでなく、中国、台湾、日本などでも利用されている。高校3年生までの中等教育段階での利用が多く、1週間に約7,000のエッセイを処理している
- 利用者による評価は、エッセイに対する評価・フィードバックの速さ、使いやすさなど概ね良好である

## 5. Application Development and Deployment

- 約15名、予算100万ドルの開発チーム
- チームは、definition, analysis, development, implementationの各段階に沿って作業を進める
- ブラウザ画面を利用してどのようにフィードバックを与えるかが大きな課題だった

## 6. Maintenance

- 新年度の開始時(9月)の導入にあわせてバージョンアップを予定している
- 集中管理方式となっているので、アップデートなども自動的に行われ、開発チームの担当者が管理に当たっている

## 7. Conclusion

- ある特定の母語話者に多い誤りなどについて分析し、文法的特徴に取り入れたいと考えている
- 文の形式面だけでなく、意味・内容に関わるフィードバックを与えるような方法についても研究開発を行う