学習効果を最大とするための英文誤り検出の性能評価

永田　亮†　中谷　和秀††
† 甲南大学知能情報学部　†† 甲南大学理工学部
E-mail: †rnagata@ konan-u.ac.jp.

1. はじめに

自由記述形式の英作文（以下，単に英作文と表記する）の添削には多くの時間と労力を要するため，英文中の文法誤りを検出する様々な手法が提案されている（例えば，文献 [2]〜[4], [7], [9], [10] など）．これらの手法は，ライティング学習支援としての応用が期待される．すなわち，誤り検出結果をフィードバックとして学習者へ与え，そのフィードバックに基づき学習が英作文を書き直すという学習支援である．実際，多くの研究（例えば，文献 [7] など）は，研究目的として学習支援への応用を挙げている．現状では，冠詞の誤り，単数／複数に関する誤り，一致の誤り，前置詞の誤りなどが検出可能であり，検出性能だけでなく検出できる誤りの種類という観点からも関連研究は進歩している．

しかしながら，語学学習支援への応用を考えた場合，従来研究には大きく欠落した部分がある．学習支援という観点からは，誤り検出の利用による学習効果を評価することが最も重要である．すなわち，誤り検出の結果をフィードバックとして学習者に与えることにより，どの程度ライティング能力が向上するかということを評価する必要がある．現状では，従来手法の評価は検出性能（多くの場合，検出率，検出精度，F 値）のみに基づいて行われている．

このことを深く追求すると次のような疑問が起こる — 現状の誤り検出手法は学習支援として利用可能であるのか？．この疑問に答えることは非常に重要である．なぜなら，検出性能を完全にすることはほぼ不可能であるからである．したがって，どれほど検出性能が向上したとしても，不完全な性能の誤り検出を学習支援に使用しなければならない．実用を考えた場合，検出性能がどこまで低くても学習効果があるのかを明らかにしなければならない．

関連して次のようなことも議論すべきである．多くの誤り検出手法では，閾値やパラメータの設定により，検出率と検出精度のバランスを調整することが可能である．それにも関わらず，学習効果が高くなるのは検出率重視であるのか検出精度であるのかの詳細な議論はなされていない（以後，この問題を検出率／検出精度問題と呼ぶ）．文献 [2], [3] は，検出精度重視がより良いと主張しているが，明確な理由は挙げられていない．この問題が解決されると，学習効果がより高くなるよう，検出率と検出精度のバランスを調整することが可能である．

このような背景を受けて，本研究では誤り検出性能と学習効果の関係を調査する．そのため次の二つの仮説を立てる：

仮説 1　学習効果が最大となるのは検出率，検出精度ともに 100%より低いときである

仮説 2　検出率重視よりも検出精度重視の誤り検出のほうが学習効果は高い

仮説 1 は，検出性能が高いほうが学習効果は高いという直感に反する仮説である．しかしながら，我々は，次のような理由を根拠として**仮説 1** を立てる．仮に，検出率も検出精度も 100%になったとする．すなわち，誤検出なしに全ての誤りを検出できる手法があるとする．この手法で学習者にフィードバックを与えると，学習者は，「どこが間違っているのか」，「なぜ間違っているのか」を考えずとも全ての誤りを見つけ出すことができる．一方，検出性能が完全でない誤り検出手法では，検出漏れや誤検出があるため，与えられたフィードバックを手掛りにして，「どこが間違っているのか」と「なぜ間違っているのか」を自ら考えて判断しないといけない．この時に学習者にかかる負荷が，より高い学習効果をもたらすと予想できる．更に，情報量が多いフィードバックが必ずしも高い学習効果をもたらすとは限らないことを示唆する実験結果がいくつか報告されている．例えば，Robb ら [11] は，4 種類のフィードバック ((1) 誤り箇所と正解，(2) 誤り箇所と誤りの種類，(3) 誤り箇所，(4) 一行あたりの誤り数) を比較し，必ずしも情報量の多いフィードバック (1) が学習効果が高いわけではないと報告している．

仮説 2 は，検出率／検出精度問題に関する仮説である．我々は，検出精度重視のほうが学習効果が高いと考えている．その理由として，学習者にかかる負荷の量と質を挙げることができる．小数の誤りを高い精度で検出した場合，他の検出されていない多くの誤りを学習は自ら探し出す必要がある．その際には，学習者は，検出された誤りからの類推，自分自身の知識，教科書や辞書などの外部からの知識を併用し(注 1)，誤りを探し出さなければならない．この誤りを探し出すという負荷が学習効果につながると考えている．検出率重視の場合，この負荷は相対的に少なくなり，検出結果が正しいかどうかを判断することが学習の中心となる．更に，検出精度が低いため類推による学習がうまくいかないと予想できる．

（注 1）：ここでは，誤り検出結果と共に教科書や辞書を用いて英文を書き直すという学習環境を想定している．学習環境の詳細については 2.1 で述べる．

2. 検証方法

予備実験を行い，検証方法を検討した．予備実験には，被検者 10 名が参加した．学習回数は，一人最大 10 回，のべ 56 回である．

2.1 学習方法

本検証方法は，英作文の作成と書き直しという学習活動に基づく．**仮説 1** と**仮説 2** を検証し，一般的な結論を得るためには，様々な誤りと様々な誤り検出手法を対象として，被検者にこの学習活動を行ってもらう必要がある．しかしながら，始めから多くの誤りと多くの誤り検出手法を対象として検証を行うことは現実的でない．そこで，本研究では，冠詞の誤りと単数／複数に関する誤りを対象とした．この理由として，(i) 冠詞の用法と単数／複数の使い分けは学習者にとって難しい，(ii) そのため，学習者の英文に頻出する，(iii) 一方で，これらの誤りの検出手法は最もよく研究されており多くの手法がある（例えば，文献 [2], [4], [7], [9], [10] など）が挙げられる．誤り検出手法は，可算／不可算の判定に基づいた誤り検出手法 [10] を選択した．この手法は，英文中の名詞が可算であるか不可算であるかを判定し，その判定結果に基づき誤りの検出を行う．例えば，英文 “I have an information.” 中の名詞 “information” が不可算であると判定できると，不可算名詞に不定冠詞を使用していることから誤りと判断される．この手法は，冠詞の誤りと単数／複数に関する誤りの検出に対して有効であることが示されている [10]．また，可算／不可算の判定を行う際に得られる確率に対して閾値を設けることで検出率重視／検出精度重視の制御が可能であるという好ましい性質も有する．

本検証方法における一回の学習活動の詳細を表 1 に示す．1. で，“University life” のなどトピックが被検者に提示される．2. では，それを受けて，被検者は準備として書く内容を考える．被検者には白紙が与えられ，下書きなどに自由に使えることとした．3. で，被検者は実際に英文を書く．その際，辞書や教材などの使用は認めず，被検者は自力で書く．10 文以上を書くことを条件とした．英作文の作成は，ブログを基に開発した英文ライティングシステム上で行う．一般的なブログシステムに誤り検出手法 [10] を実装し，英作文，誤り検出，書き直しを一つのシステム上で行えるようにした（ただし，他者の英文や過去に書いた英文は閲覧できないようにした）．このシステムにより，多人数同時に学習活動が行え，効率的に検証実験が進められる．また，実験データの管理も容易になる．4. では，システム上の「添削ボタン」を押すことで，誤り検出が行われる．検出結果として，入力英文全文を提示し，検出した誤りを赤字で表示する．検出処理自体は数秒で終了するが，被検者の足並を揃えることと休憩をかねて 5 分のインターバルを設けた．最後に，5. で，システムからのフィードバックに基づき被検者は英作文の修正を行う．補助として，文法書 [5], [6] を参考として独自に作成した冠詞と単数／複数に関する説明書き（A4 一枚もの）と英和辞書 [8] を使える環境とした．また，被検者に対して，検出結果には検出漏れと誤検出が含まれる可能性があることを説明した．なお，学習の流れと実験の時間配分は予備実験の結果に基づき調整した．

表 1： 学習の流れ

手順	時間
1. トピックを被検者に提示	–
2. 被検者はトピックを踏まえ書く内容を考える	5 分
3. 考えた内容に基づき実際に英文を書く	35 分
4. システムが誤りを検出	5 分
5. 検出結果を受けて誤りを修正	15 分

以上が，本検証方法における学習活動である．この学習活動を被検者が繰り返し行い，その際の学習効果を測定する．被検者間で，誤り検出の傾向（検出率重視／検出精度重視）を変えることで仮説の検証を行う．

2.2 学習効果の測定方法

学習効果の測定方法を議論するためには，まず学習者の能力を定義する必要がある．本研究で扱うのは誤り検出であるので，英作文中の誤りの少さを能力とするのが一つの妥当な手段である．そこで，本研究では定量的に扱うことができる誤り率を学習者の能力と定義する．誤り率は，

$$e = \frac{\text{英作文中の誤りの数} + 1}{\text{英作文中の名詞句の総数} + 1} \tag{1}$$

で定義する．ただし，表 1 の「3. 考えた内容に基づき実際に英文を書く」が終了した時点での英文中の名詞句数と誤り数を用いる．分母，分子に +1 するのは，後に数学的な取扱がよくなるためである．また，誤りが全くない英文の場合，より長いほうが能力が高いと評価できる利点もある．

能力を誤り率で定義すると，学習効果は学習活動による誤り率の減少率と定義するのが自然である．言い換えると，学習活動を重ねることで，どの程度誤り率を減少させられるかである．直感的には，学習回数と誤り率の関係をプロットし，誤りの減少率を見ることで学習効果が測定できることがわかる．単純な方法としては，実験データに対して線形回帰分析を行うことで誤りの減少率，すなわち学習効果が求められる．

しかしながら，本研究では，(a) 誤り率が低くなればなるほど，誤り率を減少させるのは難しい（能力が高くなればなるほど，能力を向上させるのは難しい），(b) 誤り率の最小値は 0 であり，学習により漸近的に 0 に近づくと予想される，という二つの理由から指数回帰を用いる．指数回帰として，

$$e = \exp\{a(t + b)\} \tag{2}$$

を用いる：ここで，t，a，b は，それぞれ，学習回数，誤り

率の減少率（学習効果），学習開始時の能力を表す．学習効果 a と学習開始時の能力 b は，実験データに対して最小二乗法を適用することで推定できる．以上より，実験を行うことで学習効果を測定できることがわかる．

2.3 比較条件

仮説 1 と**仮説 2** を検証するためには，複数の誤り検出性能に対して学習効果 a を求め，各 a の値を比較しなければならない．本研究では，次の 4 種類の検出性能を比較する．

第一の条件として，全く誤り検出を行わないという条件を選んだ．この条件では，被検者はシステムからのフィードバックなしで学習活動を行うことになる．もし，この条件で得られる学習効果よりも学習効果が低い場合，誤り検出を行うこと自体が学習支援として意味を持たないことになる．したがって，この条件はベースラインの役割を果たす．以後，この条件を **No-feedback** と表記する．

第二と第三の条件は，検出率重視と検出精度重視である．予備実験により得られた英文に対して，誤り検出手法の閾値を区間 $[0, 1]$ で 0.05 刻みで変化させ，F 値が最大となる閾値を求めた．その結果，閾値は 0.6 となった．この条件を検出率重視とした（以後，**Recall-oriented** と表記）．この条件より検出精度が高くなるよう閾値を 0.9 に設定し，検出精度重視とした（以後，**Precision-oriented** と表記）．

最後の条件は，完全な検出性能となるようにした．すなわち，検出率も検出精度も 100%である．ただし，現状ではそのような誤り検出手法を実現することは不可能であるので，英文添削の経験がある英語母語話者による採点により仮想的に実現した．以後，この条件を **Human** と表記する．

3. 実　　験

3.1 実験条件

被験者は大学 1 年生〜4 年生の 26 人とした．この 26 人の被験者を **Human** 6 人，**Recall-oriented** 7 人，**Precision-oriented** 7 人，**No-feedback** 6 人と各条件に割り当てた．

学習回数は 10 回とした．各回のトピックは，順に “University life”，“Summer vacation”，“Gardening”，“My hobby”，‘My frightening experience”，“Reading”，“My home town”，“Traveling”，“My favorite thing”，“Cooking” とした．実験は，2008 年 10 月〜12 月に実施し，週二回の学習活動を基本とした．ただし，この期間に全ての学習活動を終えることができない被験者がいたため，最終的に 8 回以上の学習を行っていない被験者は実験から除外した．その結果，**Human** 4 人，**Recall-oriented** 7 人，**Precision-oriented** 6 人，**No-feedback** 5 人の計 22 人の被験者となった．

3.2 実験結果

図 1 に実験結果を示す．図 1 は，実験で得られた学習効果 a の値を条件ごとに平均して式 (2) をプロットしたものである．ただし，条件間の差異が比較し易いように，$b = 0$ とした．具体的な a の平均値は，**Human**：$a = -0.046$，**Recall-oriented**：$a = -0.015$，**Precision-oriented**：$a = -0.038$，**No-feedback**：$a = -0.024$ となった．

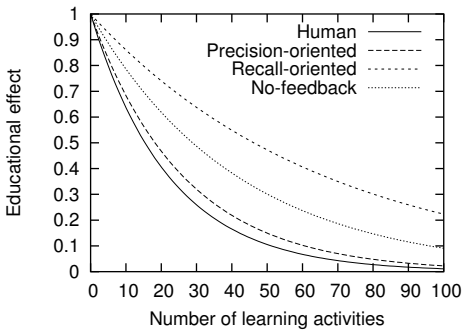

図 1：　実験結果

4. 考　　察

仮説 1 には反するが，実験結果は **Precision-oriented** には **Human** に近い学習効果があることを示している．このことは具体例で比較するとより明確になる．学習開始前の平均の誤り率 32%(注 2)を半分の 16%に減らすためには，**Human** では 16 回，**Precision-oriented** では 18 回の学習活動が必要となり，両者にはほとんど差がない．一方，**No-feedback** では 29 回であり，**Human** の倍近くかかる．また，**Recall-oriented** では 47 回と更なる学習活動が必要となる．したがって，人手の添削（もしくは検出性能が完全な誤り検出手法）に比べると学習効果は劣るものの，添削情報が全くない自学自習よりも検出精度重視の誤り検出のほうが高い学習効果が得られるといえる．

以上の結果から，何らかの理由で人手の添削が十分に利用できない状況では，積極的に検出精度重視の誤り検出を使用すべきであるといえる．そのような状況として，ライティングの授業などが挙げられる．Burstein ら [1] によると，ライティング能力を向上させるためには，(1) 英文を書く，(2) 添削してもらう，(3) 書き直す，の 3 ステップを繰り返すことが重要である．しかしながら，多くの人数（例えば 30 人）から成るライティングの授業では，これを実践することは非常に困難である．このような状況では，誤り検出を利用して学習者自身が英文を何度か書き直すことでこの 3 ステップの繰り返しが可能となる．最後に，教師の添削を受けることで最終的な添削の質も保証され，添削にかかる労力も大幅に減らせる．実験結果からは，このような誤り検出の利用法が理想的であると結論付けられる．ただし，今回対象とした誤り

(注 2)：第一回目の実験時に被験者が書いた（修正前の）英文から得られた誤り率の平均値．言い換えると，平均的な能力．

は限定的である，かつ，実験規模が小さいため，この結論が一般的に成り立つとは限らないことに注意する必要がある．また，別の条件（例えば，より高い閾値を設定する）では，**Precision-oriented** の学習効果が **Human** の学習効果を上回る可能性も残されていることに注意しなければならない．すなわち，**仮説 1** が完全に否定されたわけではない．

仮説 2 については，予想通りの実験結果が得られた．**Recall-oriented** は，**No-feedback** よりも学習効果が低い結果となった．**Recall-oriented** では誤検出が多く，学習者を混乱させたと予想できる．一方，**Precision-oriented** では精度良く少数の誤りを検出したため高い学習効果が得られたと考えられる．実際，**Recall-oriented** では，検出率 0.31，検出精度 0.60，**Precision-oriented** では検出率 0.25，検出精度 0.72 であった．更に，**Recall-oriented** では，**Precision-oriented** に比べ，より多くの検出を行う一方で検出精度が低いため，実際よりも検出精度が低く感じられる可能性がある（すなわち，誤検出の数が多い）．そのため，学習意欲が低下したとも予想できる．以上をまとめると，実験結果は，検出精度重視は検出率重視よりもよいという従来からの考え方をサポートする．

上述の検出率と検出精度に関する結果は，別の観点からも興味深い．従来，誤り検出性能は，F 値で総合的に評価することが多く行われてきた．上述の結果からは，**Recall-oriented** は $F = 0.41$，**Precision-oriented** は $F = 0.37$ となり，**Recall-oriented** のほうが性能が高いと判断される．しかしながら，学習効果という観点からは **Precision-oriented** のほうが優れている．このことは，F 値が常に最適な評価尺度とは限らないことを示唆している．

最後に，実験の結果明らかとなった本検証方法の課題について考察する．より一般的な結論を得るためには実験規模を拡大しなければならない．その一環として学習回数を増やすことが挙げられる．しかしながら，今回の実験の経験からは，一人の被験者に 10 回より多く英文を書いてもらうのは困難であるように感じられた．何度も同じ学習活動を繰り返すことで飽きを感じる学習者がいるように見受けられた．そのため，学習者が飽きを感じない検証方法を考案する必要がある．このことは，誤り検出を利用した実際の学習活動でも工夫しなければならない点である．別の課題として，実験中に学習者の能力が変化するため誤り検出性能も変化してしまうという課題がある．学習活動が進むにつれて学習者の能力は向上する．能力が向上するに従い，誤りは減り，誤り検出は難しくなる．そのため，同じ誤り検出手法でも相対的に検出性能は低下することになる．実際，**Precision-oriented** では前半は $F = 0.39$，後半は $F = 0.35$，**Recall-oriented** では前半は $F = 0.44$，後半は $F = 0.38$ であった．誤り検出性能と学習効果の関係をより詳細に調査するためには，このことを考慮して検証実験を行わなければならない．

5. おわりに

本研究では，誤り検出性能と学習効果との関係を調査した．二つの仮説，**仮説 1**：学習効果が最大となるのは検出率，検出精度ともに 100%より低いときである，**仮説 2**：検出率重視よりも検出精度重視の誤り検出のほうが学習効果は高い，を実験により検証した．**仮説 1** については成り立たないものの，検出精度重視では人手の添削に近い学習効果が得られるという興味深い結果が得られた．この結果，教師の添削が十分に利用できない学習環境では，検出精度重視の誤り検出を積極的に使うべきであると結論付けた．また，実験結果は**仮説 2** を支持することが明らかとなった．更に，必ずしも F 値が最適な性能評価指標とは限らないことも示唆した．

今後は実験規模を拡大し，より一般的な結論を導く予定である．また，対象とする誤りも前置詞や一致の誤りへ拡張する予定である．なお，本研究で収集した英作文は，文法誤りタグ付きの学習者コーパスとして公開している．利用希望者は，第一著者に連絡されたい．

謝　辞

本研究の一部は科研費（19700637）の助成により実施した．

参考文献

[1] J. Burstein, M. Chodorow, and C. Leacock, “Automated essay evaluation: The *Criterion* online writing service,” AI Magazine, vol.25, no.3, pp.27–36, 2004.

[2] M. Chodorow and C. Leacock, “An unsupervised method for detecting grammatical errors,” Proc. 1st Meeting of the North America Chapter of ACL, pp.140–147, 2000.

[3] M. Chodorow, J.R. Tetreault, and N.R. Han, “Detection of grammatical errors involving prepositions,” Proc. 4th ACL-SIGSEM Workshop on Prepositions, pp.25–30, 2007.

[4] N.R. Han, M. Chodorow, and C. Leacock, “Detecting errors in English article usage by non-native speakers,” Natural Language Engineering, vol.12, no.2, pp.115–129, 2006.

[5] 広田成章, 高校新基礎英語, 桐原書店, 東京, 1992.

[6] 飯塚茂，萩野敏, プレステージ綜合英語, 文英堂, 東京, 1997.

[7] E. Izumi, K. Uchimoto, T. Saiga, T. Supnithi, and H. Isahara, “Automatic error detection in the Japanese learners’ English spoken data,” Proc. 41st Annual Meeting of ACL, pp.145–148, 2003.

[8] 小西友七，南出康世, ジーニアス英和辞典第 4 版, 大修館書店, 東京, 2007.

[9] R. Nagata, A. Kawai, K. Morihiro, and N. Isu, “A feedback-augmented method for detecting errors in the writing of learners of English,” Proc. 44th Annual Meeting of ACL, pp.241–248, 2006.

[10] R. Nagata, T. Wakana, F. Masui, A. Kawai, and N. Isu, “Detecting article errors based on the mass count distinction,” Proc. 2nd IJCNLP, pp.815–826, 2005.

[11] T. Robb, S. Ross, and I. Shortreed, “Salience of feedback on error and its effect on EFL writing quality,” TESOL QUARTERY, vol.20, no.1, pp.83–93, 1986.