2003－NL－155 (11)
2003／5／26

英語コミュニケーション能力の自動測定技術の提案

安田圭志[1,2]，隅田英一郎[1]，山本誠一[1]，柳田益造[2]，前川喜久雄[3]，菅谷史昭[4]

[1]ATR 音 声言語コミュニケーション研究所 [2]同志社大学 [3]国立国語研究所 [4]KDDI 研究所

本論文では，英語コミュニケーション能力を自動測定する手法を提案する．提案手法では，語彙や文法等に関する要素的能力ではなく，「英語文を構成する」総合的能力を測定する．提案手法は以下のステップからなる．(1) 受験者に日本語文を英語に翻訳してもらう．(2)受験者の訳文と正解の訳文との間の**一致度**を訳文間の *n* グラムの重なりや編集距離等を用いて機械的に測定する．(3)予め，様々な**能力値**と**一致度**との相関を学習しておき，未知受験者の**能力値**を**一致度**から推定する．能力値として TOEIC スコアを採用し，能力値が既知である 28 人が翻訳したデータを用いて能力値の推定実験を行い，良い相関を得た．TOEIC の部分スコアとの相関も確認できたので，提案手法は「読む・聴く」能力の測定にも利用可能である．

A Proposal for Automatically Gauging of English Language Proficiency

Keiji Yasuda[1,2], Eiichiro Sumita[1], Seiichi Yamamoto[1],
Masuzo Yanagida[2], Kikuo Maekawa[3], Fumiaki Sugaya[4]

[1]ATR Spoken Language Translation Research Labs. [2]Doshisha University
[3]The National Institute for Japanese Language [4]KDDI R&D Laboratories

It is vital to measure communication capability on demand for CALL. This paper proposes a computerized method of measuring communication capability using English. The proposed method does not measure elementary capabilities concerning vocabulary or grammar, but measures the integrated skills for structuring sentences, which is essential for writing and speaking. The proposal consists of three steps. Step one asks the subjects to translate Japanese sentences into English. Step two gauges the match between the translations of the subject and correct translations based on the *n*-gram overlap or the EDIT distance between translations. Step three learns the relationship between (1) capability and (2) match. By regression it finds the straight-line fitting a scatter plot of the capability and match of known subjects. Then, it estimates the capability of the unknown subject by using the line and the match. We conducted experiments that estimate the TOEIC score. We collected a set of English sentences translated from three-hundred thirty Japanese sentences by twenty-eight subjects with varied English capability. We found that the estimated score correlates with the actual score. Our method exhibited a correlation with the sub scores of TOEIC as well, so, it is applicable for measuring both listening and reading capability.

1 はじめに

国や人の交流が全地球的な規模で行われ，インターネットが爆発的に普及した現在，言わば国際共通語になっている英語によるコミュニケーション能力を高めることは，国家の戦略課題の一つであると言える[1]．しかし，これまでの英語教育は中学・高校・大学等や会話学校・通信教育等において膨大な時間と金とを費やしているにもかかわらず，日本人の

[1]例えば，文部科学省 (http://www.mext.go.jp/) の“英語指導方法等改善の推進に関する懇談会（2001/1/17）”や“「英語が使える日本人」の育成のための戦略構想（2002/7/12）”を参照.

英語力を向上することに成功していない．この停滞状況を打開するために，情報通信技術を活用して教育の質を高める e-Learning に期待が集まっている．また，自然言語処理などの情報処理技術の発展と大規模なコーパスの収集は外国語教育について新たな展開の方向を開くものと期待されている．

このような背景のもとに，本論文では，情報処理技術の外国語教育への利用の一つの新たな展開として，英語によるコミュニケーションの能力を自動的に測定する技術の研究開発を提案する．この技術を英語教育に導入し，学習者の実力を頻繁かつタイムリーに測定すれば，(1) 学習者は学習による進歩を敏速に把握でき，(2)指導者は，教材や指導方法の効果を把握し，また，問題点の発見・改良ができる．その結果，日本人の英語力を飛躍的に高められると考えられる．

以下では，2 節で言語コミュニケーション能力について，3 節で提案手法の詳細について，4 節で，能力推定実験とその結果について，5 節で，関連研究との関係や今後の展望について説明し，6 節で論文を結ぶ．

2　言語コミュニケーション能力

2.1　「文を構成する」能力

コミュニケーション能力の捉え方は多様である．実際，質の高いコミュニケーションのためには文化的・社会的差異の理解やジェスチャーなどの非言語的な要素も重要であるが，本論文では，より基本となる「言語による」コミュニケーションに焦点をあてる．

また，「言語による」コミュニケーション能力もいわゆる四技能，すなわち，受信に関わる「読む・聴く」と発信に関わる「書く・話す」がある．また，この四技能のうち，「話す・聴く」は発音に関係するが，「書く・読む」はそうではない．

この四技能はすべて重要であるが，日本人は特に『発信型』の能力が弱いとされる．『発信型』のコミュニケーション能力に共通

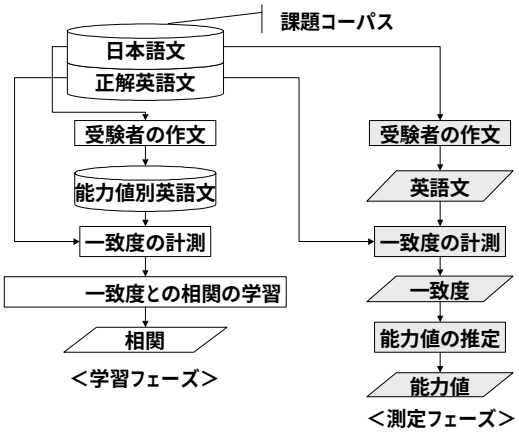

図 1 提案手法の処理の流れ

な「文を構成する」能力，すなわち，相手に伝えようとする内容を語順や訳語を適切に選んで外国語の文で表現する能力が必須であることは明白である．しかし，「文を構成する」能力の自動測定は従来扱われることが少なかった．本論文では，「文を構成する」能力を対象とする．

2.2　従来の能力測定

英語コミュニケーション能力の尺度として広く利用されているものには，実用英語技能検定[2]や国際連合公用語英語検定試験[3]や TOEIC[4]のペーパーテストや CASEC[5]のコンピュータテスト等がある．これらは，主に，「聴く・読む」能力に着目したテストであり，文法や語彙に関する知識を検査する多肢選択テストや聞き取りテストで構成される．

「話す・書く」能力に関しては，実用英語技能検定の面接や TOEFL の小論文評価のように，経験を積んだ専門家が採点している現状であり，採点の自動化の必要性は高い[6]．

---

[2] 財団法人日本英語検定協会（http://www.eiken.or.jp/）

[3] 財団法人日本国際連合協会(http://www.unate.or.jp/)

[4] http://www.toeic.or.jp/toeic/index.html

[5] http://casec.evidus.com/

[6] 近年，小論文評価の自動化の研究が進み，開発されたシステムe-ratorが実際の評価に取り入れられているが，使用方法は人間の評価を補助することに限定されている．

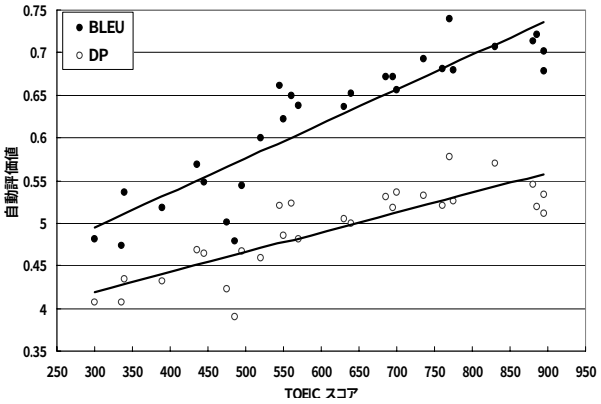

図 2 一致度と TOEIC スコアの相関

3 提案手法

提案手法は以下のステップからなる．(1) 受験者に日本語文を英語に翻訳してもらう．(2) 受験者の訳文と複数の正解の訳文との間の**一致度**を訳文間の n グラムの重なりや編集距離等を用いて機械的に測定する．(3)予め，様々な**能力値**と**一致度**との相関を学習しておき，未知受験者の**能力値**を**一致度**から推定する．図 1 に提案手法の処理の流れ，図 2 に一致度と能力値の相関のサンプルを示した．

3.1 一致度

訳文と複数の正解訳の一致度を求める手法は機械翻訳品質の客観評価手法として広く利用されている．以下で代表的な 2 通りの一致度の求め方について説明する．

(1) n グラムを用いた一致度（BLEU スコア）：BLEU スコアは，評価対象の訳文と正解の訳文に関して n グラムの重なりに基づいて，次式で計算する (Papineni et al., 2002). スコアは良い方から悪い方へ 1.0 から 0.0 の値を取る（図 3 に計算例を示した）．

$$S_{BLEU} = \exp\left\{\sum_{n=1}^{N} w_n \log(p_n) - \max\left(\frac{L^*_{ref}}{L_{sys}} - 1, 0\right)\right\}$$

$$P_n = \frac{\sum_i \left(\begin{array}{l} \text{the number of } n\text{-gram in segment } i \\ \text{in the translation being evaluated,} \\ \text{with a matching reference coocurence} \\ \text{in segment } i \end{array} \right)}{\sum_i \left(\begin{array}{l} \text{the number of } n\text{-gram in segment } i \\ \text{in the translation being evaluated} \end{array} \right)}$$

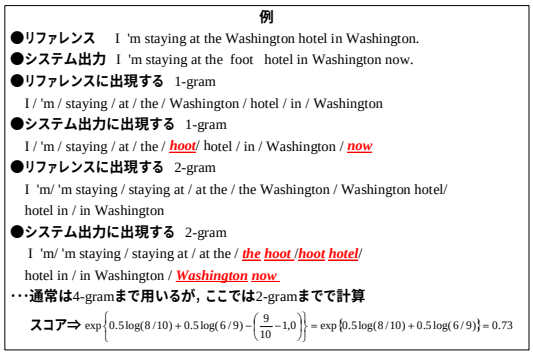

図 3 BLEUスコアの計算例

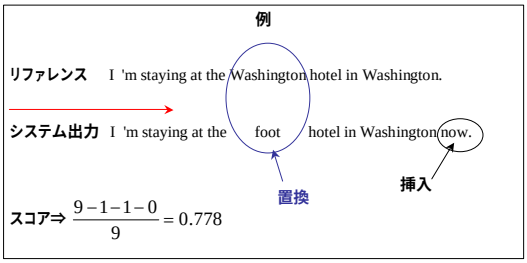

図 4 DP スコアの計算例

(2) DP を用いた一致度(DP スコア): DP は訳文と正解を比較し，置換，挿入，削除のペナルティの総和を最小化するアラインメントを算出するものであり，DP スコアは次式で計算する．

$$S_{DP} = \max_{i=1 \text{ to all reference}} \left\{ \frac{T_i - S_i - I_i - D_i}{T_i} \right\}$$

ただし，T_i は正解訳 i の総語数，S_i は正解訳 i と評価対象の訳文を DP マッチングにより比較した時の置換語数，I_i は挿入語数，D_i は脱落語数である（図 4 に計算例を示した）．

一つの原文に対する正解訳は様々なヴァリエーションが考えられる．上記の一致度はいずれもこの正解訳との相違を測るものであり，正解訳を多数用意しておくことが信頼性の観点から肝要である（4.1 及び 5.4 節参照）．また，機械翻訳の評価に関する知見であるが，正解訳を増加させると上記の 2 つのスコアが同一の値に近づいていくことが知られている (Yasuda 他 2003).

3.2 能力値

能力値の候補はいろいろ考えられるが，本論文では受験者数が多く，受験者の能力の分布が幅広い

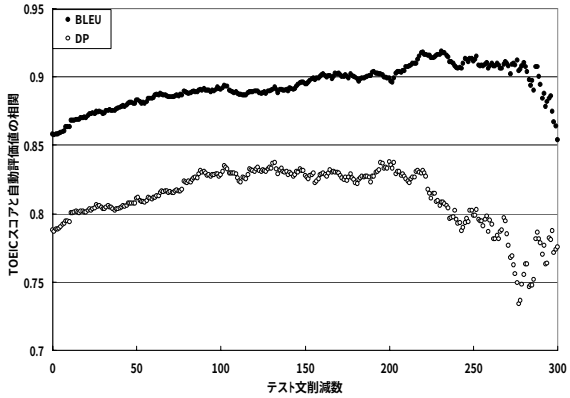

図 5 テスト文削減の影響
（TOEIC スコア）

TOEIC スコア′を採用した．同スコアは最低 10 ポイントで最高 990 ポイントである．

4 能力推定実験

実験では TOEIC スコアが既知である 28 人が翻訳したデータを用いて実験し，能力値の推定に関して良い結果を得た．TOEIC スコアは，2 つの部分スコア（「聴く」能力を測定するためのリスニングセクションのスコアと，「読む」能力を測定するためのリーディングセクションのスコア）の合計からなるが，これらの部分スコアとの相関も確認できたので，提案手法は「読む・聴く」能力の測定にも利用可能である．

以下で詳細を述べる．実験条件

- テストセットとして，ATR バイリンガル旅行対話データベースからとられた 23 会話(330 発話) からなる SLTA1 セットを用いた．
- TOEIC スコア 300 点台～800 点台の被験者 28 名に同一の日本語文の音声を聞かせ翻訳させ，書き起こした．
- 正解訳は 1 文につき 16 を準備した (翻訳者 5 名が 1 文につき 3 翻訳を作成し，元の対訳コーパスの 1 文と合わせて 16 文).

[7] TOEIC は ‘Test of English for International Communcation’ のアクロニムであり，英語の熟達度を測定する (http://www.chauncey.com/).

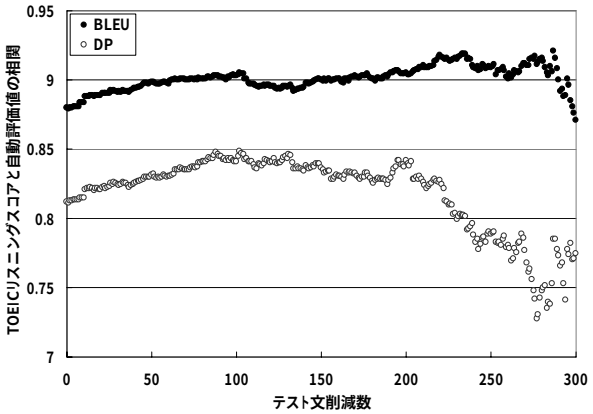

図 6 テスト文削減の影響
（TOEIC リスニングスコア）

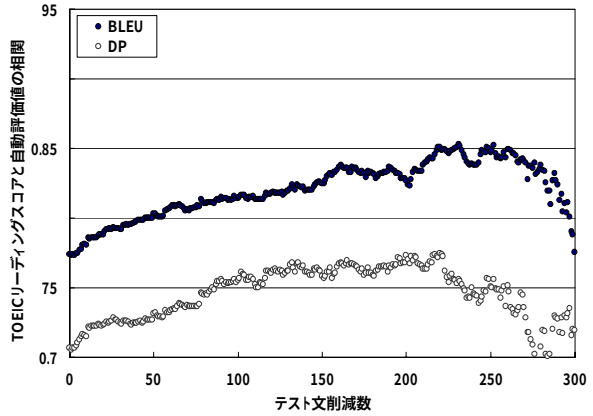

図 7 テスト文削減の影響
（TOEIC リーディングスコア）

4.1 実験結果

実験では，BLEU スコアの場合，推定スコアと実際の TOIEC スコアとで 0.9 程度の相関が得られ，DP スコアでは，推定スコアと実際の TOIEC スコアとで 0.8 程度の相関が得られ，自動推定法が有望であることが確認できた[8].

また，問題数を変化させる[9]と様々な観測が得られた（図 5，6，7）．

問題数が多いと相関が乱れるのは，識別力の低い簡単な問題の割合が増加するためと想定できる．また，問題数が少なすぎる（図の

[8] CASECスコアとTOEICスコアの相関は0.86である．

[9] (Sugaya他, 2001)で提案された方法を用いた．

右側）と非常に乱雑な結果となる．本実験では，200 文ほど削り 100 文ほど残ったセットが良い性能を示している．高い精度を得るには適切な問題の数と種類の選択が重要と考えられるが，適切な選択法は今後の検討課題である．

TOEIC の部分スコアとの相関（図 6，7）も確認できたので，提案手法は「読む・聴く」能力の測定にも利用可能である．但し，「聴く」能力との相関は常に強く，「読む」能力との相関は弱い．TOEIC で「聴く」能力を測るときの問題と本実験のテストセットの親和性が高かったと考えられる．

5 関連研究及び議論

5.1 大規模対訳コーパスを使ったテスト問題自動作成

従来，テスト問題は，文法事項や語彙やイディオムなどの知識を調べたり，長文の大意要約を求めたりする，多肢選択方式や穴埋め方式が主流であった．問題の作成・評価には専門家が年単位の時間を費やしてあたる必要がある．また，テスト問題の件数も数千件にとどまり，繰り返し受験する場合には，同じ問題を同一受験者に提示することを避けることが困難となり，評価の信頼性が揺らぐ可能性がある．

これに対して，提案手法は，作文課題であり，適切な原文を提示するだけである．大規模な対訳コーパスがあれば，単にそこから選択するだけになる．たとえば，ATR は旅行関係表現を集めた 70 万件の日英対訳コーパス（Takezawa 他 2002）をもっており，これは，通常の問題プールに比べ 2 桁大きく，同じ問題を同一受験者が受ける可能性を非常に小さくできる．

テスト問題はこのように対訳コーパスから自動選択するので，貿易，経理，技術など専門分野毎の対訳コーパス作成をすれば，各専門分野で必要とされる英語力を測るテストも自動的に作成できる．また，容易に多言語展開でき，今後，重要度が増加すると言われている中国語などに，すばやく，低いコストで適用できる．

作文課題の難易度と文の特徴（エントロピー[10]，文長，単語の頻度・重要度など）との関係を利用して，受験者の回答に動的に合わせた適応型テストに向けて研究を進めている．

5.2 誤り診断情報の提示

本提案手法は翻訳力の高低を測定しているが，能力が低かった場合に，その原因と対策を示すことが出来ていない．各作文課題とそれに直接関係する学習項目（語彙，文法事項など）との関連を付けたコーパス[11]を作成し，課題文と学習項目の関連付けの自動化を行い，受験者にフィードバックする手法を確立することは重要な課題の一つである．

5.3 自動推定方法の改良

本論文では BLEU スコアと DP スコアを紹介したが，主観評価と相関が高い客観評価手法については機械翻訳を対象として幾つか提案されており，また，構文情報の利用する手法もある．機械翻訳を離れて，英語コミュニケーション能力を測定する高精度の手法を追求することは今後の重要な課題である．

5.4 多重正解の自動作成

これらのアルゴリズムで用いる複数の等価訳は今のところ人手で作成される．我々は既に，これを効率的に行う方法を提案している(Sugaya 他 2002).

また，編集距離を使って類似の対訳文を対訳データから検索し，複数の等価訳を得る方法(Yasuda 他 2001)や単一の正解訳をプログラムで言い換えて複数の等価訳を得る手法も提案している(Shimohata 他 2002; Finch 他, 2002).

[10] Sugaya他 (2000) 参照.

[11] 機械翻訳の評価では池原他(1994)，井佐原他(1996）等の先行研究がある．

5.5 多様な分野での検証とコーパス開発

本実験は ATR で開発した旅行会話コーパスを用いて行われ，ある指標が TOEIC スコアと高い相関も持つことが確認できた．本手法は分野に依存しないと考えられるが，ビジネス分野や技術に関する打ち合わせ等他の分野への拡張は今後検証が必要な点である．この検証には多様な分野でのコーパスの開発が必要となる．

5.6 「話す」能力評価への展開

様々なレベルの日本人学習者の英語翻訳データを大量に集めることが出来たとすると，この学習者データを活用して，レベル毎に学習者の言語モデルを作ることが可能となる．これによって，学習者のレベルに適応した音声認識装置が開発でき，間違った英語も「正確に」認識できるような新機能を生み出せる．すなわち，音声入力も受け付けて能力を評価するシステムや，学習者の誤用を検出し，矯正を支援するような総合的な英語教育システムを構築できる．

6 おわりに

本論文では，英語コミュニケーション能力を自動測定する手法を提案する．提案手法では，語彙や文法等に関する要素的能力ではなく，「英語文を書く」総合的能力を測定する．提案手法は以下のステップからなる．(1) 受験者に日本語文を英語に翻訳してもらう．(2)受験者の訳文と正解の訳文との間の**一致度**を訳文間の *n* グラムの重なりや編集距離等を用いて機械的に測定する．(3)予め，様々な**能力値**と**一致度**との相関を学習しておき，未知受験者の**能力値**を**一致度**から推定する．

能力値として TOEIC スコアを採用し，能力値が既知である 28 人が翻訳したデータを用いて実験し，能力値の推定に関して良い結果を得た．TOEIC の部分スコアとの相関も確認できたので，提案手法は「読む・聴く」能力の測定にも利用可能である．

参考文献

Finch, A, Watanabe, T., and Sumita, E., 2002 Paraphrasing by Statistical Machine Translation, FIT-2002, E-53, pp.187—188

池原他 1994，言語表現体系の違いに着目した日英機械翻訳機能試験項目の構成，人工知能学会誌, 9 (4)

井佐原他 1996，開発者の視点からの機械翻訳システムの技術的評価，自然言語処理，3 (3)

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. 2002 Bleu: A Method for Automatic Evaluation of Machine Translation, Proc. of the 40th Annual Meeting of ACL, pp. 311—318.

Shimohata, M. and Sumita, E. (2002a) Automatic paraphrasing based on parallel corpus for normalization, Proc. of LREC.

Sugaya, F., Takezawa, T., Yokoo, A., Sagisaka, Y., and Yamamoto, S. 2000 Evaluation of the ATR-MATRIX Speech Translation System with a Pair Comparison Method Between the System and Humans, Proc. of ICSLP, pp. 1105—1108.

Sugaya, F., Takezawa, T., Kikui, G. and Yamamoto, S., 2002. Proposal of a very-large-corpus acquisition method by cell-formed registration, Proceedings of the LREC.

Sugaya, F., Yasuda, K., Takezawa, T. and Yamamoto, S. , 2001 Quality-Sensitive Test Set Selection for a Speech Translation System Proc. of ACL 2001 Workshop on S2S, pp. 109—116.

Takezawa, T. et al. 2002 Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World, Proc. of LREC.

Yasuda, K., Sugaya, F., Takezawa, T., Yamamoto, S. and Yanagida, M. 2001. An automatic evaluation method of translation quality using translation answer candidates queried from a parallel corpus, Proc. of MT-SUMMIT-VIII

Yasuda, K., Sugaya, F., Takezawa, T., Yamamoto, S. and Yanagida, M. 2003. Applications of Automatic Evaluation Methods to Measuring a Capability of Speech Translation System, Proc. EACL.