人間の能力との比較による音声翻訳システム 性能評価法の効率化に関する検討

安田圭志　菊井玄一郎
ATR 音声言語コミュニケーション研究所
{keiji.yasuda, genichiro.kikui}@atr.jp

1. はじめに

言語翻訳システムや音声翻訳システムの研究・開発において，システムの評価は不可欠であり，これまでに様々な評価手法が提案され，評価が実施されている[1-3]．人手を要する主観評価法として，評価者がシステムからの出力文ごとに，あらかじめ定義された訳質に対するランクを決定する「翻訳ランク評価法」[2]が用いられることが多い．

これらの評価手法は，あるテストセットに対してどの程度うまく訳せているかということのみを表しており，評価に用いたテストセットの翻訳の難易度については一切考慮されていない．しかしながら，音声翻訳システム実利用時におけるシステムの有用性は，ユーザにとって困難な翻訳をシステムが高品質に翻訳できる場合と，ユーザにとって容易な翻訳をシステムでも高品質に翻訳できる場合とでは大きく異なり、前者の方が有用性が高い．このような点を考慮するとシステム性能を正確に把握するためには，出力の訳質のみならず入力文の人間に対する翻訳の難易度を考慮した評価が必要である．

ここで，翻訳の難易度について注目すると，システムに対する翻訳の難易度は入力文のエントロピー等の指標で測定できることが明らかになっている[3]．その一方で，人間に対する翻訳の難易度についてはエントロピー等の指標では測定は不可能であり，現状においては，原言語文の特性等から直接的に測定する方法は存在しない．

そこで，能力が明らかな人間による翻訳結果とシステム出力とを訳質の観点から比較する相対評価を行うことにより，間接的に人間に対する翻訳の難易度を考慮できる評価法が提案されている[3]．この一対比較型 TOEIC 評価法では，TOEIC スコアが明らかな人間による翻訳結果とシステム出力とを一対比較することにより，システムの能力と均衡する TOEIC スコアを求める． TOEIC の点が高い人は難しい文でも高品質で翻訳でき，点の低い人は易しい文しかうまく翻訳できないことを考えると，この評価値が高い場合は（人間にとって）難しい文までうまく翻訳でき，低い場合には易しい文しかうまく翻訳できない，ということを表している．

このように, 一対比較型 TOEIC 評価法はテストセットの難易度を考慮に入れた評価を可能にするが，多くの評価コストがかかるという問題がある．そこで，本研究では，一対比較型 TOEIC 評価法と同等の評価性能を担保しつつ，評価コストを大幅に削減する「ソート型 TOEIC 評価法」を提案する．

2 では一対比較型 TOEIC 評価法について概説し，3 ではソート型 TOEIC 評価法について述べる．4 では実験結果を示し，最後に 5 で全体をまとめる．

2.一対比較型 TOEIC 評価法

図 1 に, 一対比較型 TOEIC 評価法の処理の流れを示す．ここでの翻訳方向は日英方向である．6 ヶ月以内に TOEIC を受験した複数の日本語ネイティブの被験者に日本語の問題文を音声で提示し，日本語を英語に翻訳させ，回答用紙に記入させる．被験者から回収された回答用紙は書き起され，人間の翻訳結果として，システムによる翻訳結果と比較する．これらの比較対象の翻訳結果から評価シートを作成し，日英バイリンガル評価者がシステムと各被験者の翻訳結果とを一対比較で主観評価する．

図 2 に，バイリンガル評価者による評価の流れを示す．まず評価者は，システムによる翻訳結果と，被験者による翻訳結果に対し，あらかじめ定義された以下の訳質ランクを決定する．

- S ランク：原文の情報が漏れ無く翻訳されており，訳出に文法的な間違いがない．使われている語彙もネイティブから見て自然である．
- A ランク：使われている語彙はネイティブから見て不自然であるが，原文の情報が漏れ無く翻訳されており，訳出に文法的な間違いがない．
- B ランク：原文のあまり重要でない情報が一部漏れていたり，訳出に文法的な間違いが若干あるが、容易に理解できる．
- C ランク:原文の重要な情報が漏れていたり，訳出に文法的な間違いが大分あって、かなり崩れた訳出であるが，良く考えれば理解出来きる．
- D ランク：重要な情報が誤訳されており，理解不能である．

2 つの翻訳の優劣は，ここで決定されたランクに基づいて決定するが，2 つの翻訳のランクが同じであった場合，各翻訳の自然性を考慮して優劣を決定する．自然性まで考慮しても優劣を決めることができない場合については，同等 (Even)と評価する．被験者による翻訳結果にはスペルミスが含まれる場合が

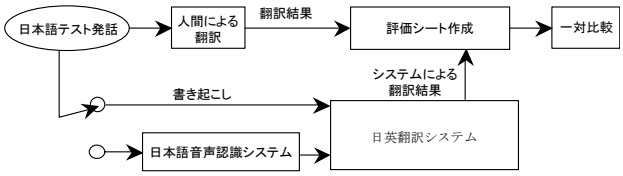

図 1. 一対比較型 TOEIC 評価法の処理の流れ

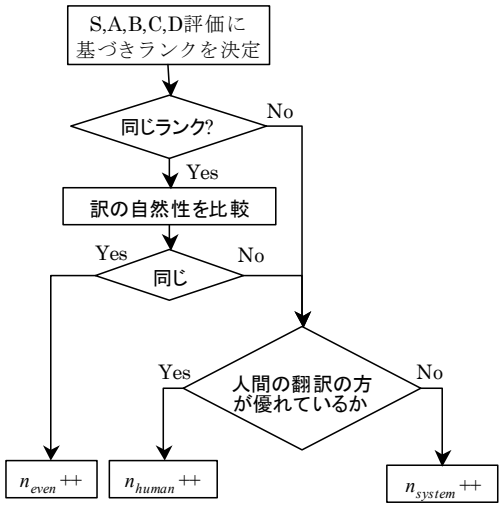

図 2. バイリンガル評価者による評価の流れ
(一対比型 TOEIC 評価法)

あるが，本手法の目的は，英語の語彙能力を評価することではなく，音声翻訳能力を評価することであるため，評価者はスペルミスが含まれる場合でも，“center”とすべきところを“senter”とした誤りのように，バイリンガルの評価者が音にして単語を推定できると判断した場合については，誤りとしない方針で評価を行う．

全ての被験者とシステムとの一対比較が完了した段階で，回帰分析によりシステム性能に相当する TOEIC スコア(システム TOEIC 換算点)を計算する．

回帰分析[1]を行うにあたり，システム勝率(W_S)を次式により計算する．

$$W_S = (n_{system}+0.5\times n_{even})/n_{total} \quad (1)$$

ただし，n_{total} はテストセットに含まれる文数を表す．式(1)では，両者の能力が同等とみなせる文数（n_{even}）を二分し，システム優位の文数（n_{system}）に加える修正を行っている．

回帰分析において，各被験者の TOEIC スコアを独立変数(X_i)，各被験者の W_S を従属変数(Y_i)とし，以下の関係を満たすものとする．

$$Y_i = \begin{bmatrix}1 & X_i\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} + \varepsilon_i \qquad (i = 1,2,\Lambda\ ,m) \quad (2)$$

$$= \boldsymbol{X_i}\boldsymbol{\beta} + \varepsilon_i \qquad (i = 1,2,\Lambda\ ,m)$$

[1]実例に沿った説明は 4 において後述する

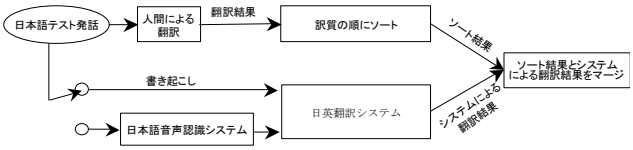

図 3. ソート型 TOEIC 評価の処理の流れ

表 1. TOEIC 評価の実験条件

	BTEC	MAD	FED
被験者数 (m)	21	35	21
テスト発話数(n_{total})	510	502	155
パラメータのバリエーション(N_{param})	3	3	3

ただし，β_1 と β_2 は母回帰係数を表し，mは被験者数を表す．また，ε_i は誤差項である．ここで

$$\boldsymbol{Y} = \begin{bmatrix}Y_1\\ Y_2\\ \mathrm{M}\\ Y_m\end{bmatrix} \quad (3), \qquad \boldsymbol{\chi} = \begin{bmatrix}\boldsymbol{X_1}\\ \boldsymbol{X_2}\\ \mathrm{M}\\ \boldsymbol{X_m}\end{bmatrix} \quad (4)$$

とすると，標本回帰直線式(5)の回帰係数

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix}$$

は，式（6）で与えられる．

$$\boldsymbol{Y} \cong \boldsymbol{\chi\beta} \quad (5)$$

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\chi}^{+}\boldsymbol{Y} = (\boldsymbol{\chi}^{T}\boldsymbol{\chi})^{-1}\boldsymbol{\chi}^{T}\boldsymbol{Y} \quad (6)$$

ただし，$\boldsymbol{\chi}^{+}$ は，$\boldsymbol{\chi}$ の最小 2 乗型一般逆行列である．

システム TOEIC 換算点は，式（5）を用い，システムの能力と被験者の能力が均衡する点, 即ち Y=0.5 となる X により求めることができる．

3. ソート型 TOEIC 評価法

実際の音声翻訳システムの研究開発においては，システムパラメータや学習コーパスなどを様々に変更した上でのシステム評価が必要とされる．このような評価を一対比較型 TOEIC 評価法により実施する場合，最大で次式で計算される n_{pairs} 対の評価を行う必要がある．

$$n_{pairs} = m \times n_{total} \times N_{param}$$

ここで，N_{param} は，評価対象のシステムパラメータ設定のバリエーションを表す．

旅行会話基本表現集コーパス（BTEC），実験室において収集されたシステムを介した対話（MAD），関西空港等において収集されたシステムを介した対話（FED）の 3 種類のテストセット[4]を用いて評価を実施した．表 1 に各テストセットでの評価における実験条件を示す．

パラメータバリエーションについては，音声認識システム[5]では，リアルタイムファクターが 1 のシステム（ASR(rtf1)），リアルタイムファクターが 5 のシステム（ASR(rtf5)），テキスト入力を想定し認

表 2. テスト発話「大阪駅は大阪の中心部にある駅です」に対するソート結果の例

翻訳結果	ソート結果(訳質順位)	被験者のTOEICスコア
Osaka station is located at the center Osaka ⋮ Osaka station is in the senter part of Osaka	1 ⋮ 1	925 ⋮ 580
Osaka station is center of Osaka ⋮	2 ⋮	370 ⋮
A Osaka station is at the center of Osaka ⋮	3 ⋮	715 ⋮
Osaka station is center area of downtown	4	875
Osaka station is the center	5	480
There is Osaka station that where is middle of Osaka	6	385

識誤りが無い場合（IN(text)）の 3 パターンを用意した．また，機械翻訳システム[6, 7]については，処理時間がかかるが高い訳質を出力できるシステム(MT(high perform))と，訳質は落ちるが処理時間が早いシステム(MT(high speed))を用意した．実際の評価ではこれらの音声認識システムと翻訳システムの組合せの内，IN(text) MT(high perform)，ASR(rtf5) MT(high perform)，ASR(rtf1) MT(high speed)の 3 パタンの評価を実施した．また，これらの条件に加え，従来の一対比較型 TOEIC 評価法の研究と同様の実験条件での評価として，SLTA1 テストセットと，翻訳システム TDMT を用いた評価も実施した[3].

テストセット BTEC を用いた評価では，32130 対（=21×510×3）の一対比較が必要であるが，熟練した評価者においても 1 時間あたりに可能な評価量は 100 対程度であるため，BTEC を用いた評価だけでも 300 時間以上を要することになる．

このような評価コストの問題を解決するため，翻訳自動評価法を応用してシステム TOEIC スコアを自動計算する方法が提案されている[8]．このような方法により評価コストの問題は解決されるが，得られたシステム TOEIC スコアの信頼区間の幅が大きくなり，一対比較型 TOEIC 評価法ほどは高い評価性能が得られないという問題がある．

ここで，一対比較型 TOEIC 評価法と同等の評価性能を担保しつつ，評価コストを大幅に削減するソート型 TOEIC 評価法について説明する．ソート型 TOEIC 評価法では，あらかじめ TOEIC スコアが既知の被験者による翻訳を，訳質の順に並べ替えたデータを作成することにより，評価コストを大幅に削減することが可能である．図 3 に日英翻訳方向にけるソート型 TOEIC 評価法の処理の流れを示す．まず，バイリンガルの評価者は，複数の被験者による翻訳結果を，図 4 に示した評価用ツールを用いて並べ替える．この評価ツールは，表示された翻訳をドラッグアンドドロップで順番を入れかえたり，同じ訳質の文を一つにまとめ上げる機能などを持っている．表 2 は，ソート結果の例である．ここで示すように，

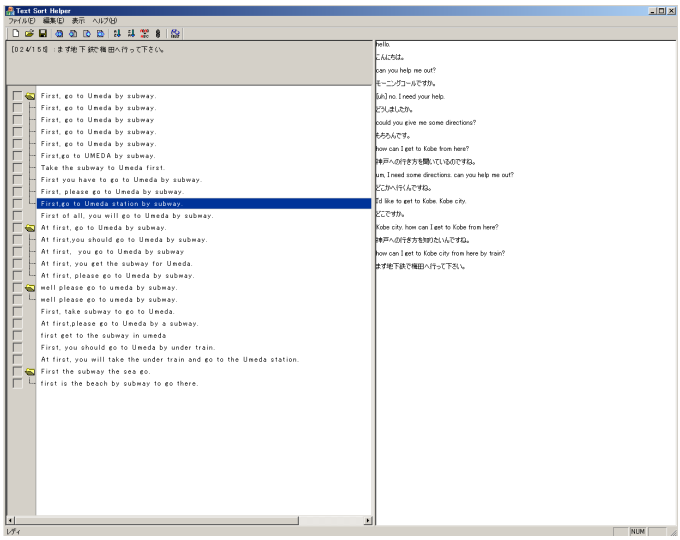

図 4. ソート型 TOEIC 評価に用いた GUI

テスト発話単位で見た場合，訳質の順と TOEIC スコアの順が必ずしも一致するわけでは無く，評価者によるソート作業が必要となる．

次に，先の作業で得られた「ソート結果」と，評価対象であるシステムによる翻訳結果とを見比べ，どの被験者による翻訳の訳質に相当するかを決定する．

これらの作業により，テスト発話単位で，システムによる翻訳と被験者による翻訳との優劣関係が明らかになることから，システム勝率（W_S）を計算することができる．最後に，このシステム勝率を用いて一対比較型 TOEIC 評価の場合と同様の手順でシステム TOEIC スコアを計算する．

4. 評価結果

図 5 は，ソート型 TOEIC 評価法で，テストセット BTEC を用いた評価における被験者の TOEIC スコアとシステム勝率の関係を表している．図 5 において，縦軸はシステム勝率を，横軸は被験者の TOEIC スコアをそれぞれ表す．また，●は IN(text) MT(high perform)の設定のシステムを評価した結果であり，

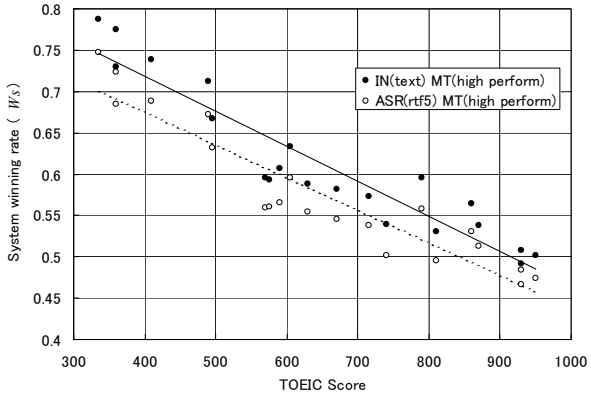

図 5. テストセット BTEC の評価におけるシステム勝率（W_S）と TOEIC スコアの関係

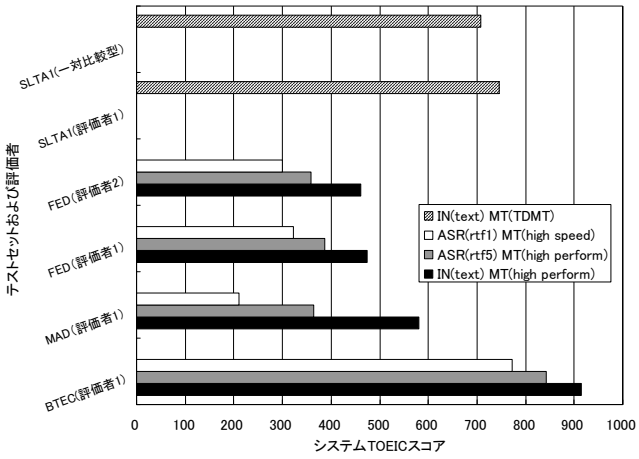

図 6. ソート型 TOEIC 評価法での評価結果

○は ASR(rtf5) MT(high perform)での結果である．ここで，式(2)～(6)との対応について述べると，縦軸は式(3)における $\boldsymbol{Y}$ に，横軸は式(4)の $\boldsymbol{\chi}$ に対応する．図中の直線は，式(6)により計算された回帰直線である．

図 6 に, 評価により得られたシステム TOEIC スコアを示す．横軸はシステム TOEIC スコアを表し，縦軸のラベルは，テストセットおよび評価者を表す．図 6 において最上部の棒グラフが従来の一対比較型 TOEIC 評価法での結果であり，それ以外はソート型 TOEIC 評価法での結果である．まず，テストセット SLTA1 における，一対比較型 TOEIC 評価法と，ソート型 TOEIC 評価法の結果を比較すると（図中の SLTA1（一対比較型）と SLTA1（評価者 1））, 多少の評価結果のズレがあるものの，ほぼ同様の結果が得られている. また, テストセット FED においては, 異なる 2 名の評価者がソート型 TOEIC 評価法により, 同様の評価を行っている（図中の FED（評価者 1）と FED（評価者 2））．これらの結果を比較すると，ほぼ同様の結果が得られており，評価者間の評価の揺れが小さいことが分かる．またテストセット BTEC では，最も高いシステムパフォーマンスが得られており，音声認識誤りが無い場合で 914 のシステム TOEIC スコアが得られている．

5.まとめと考察

一対比較型 TOEIC 評価法を効率化する方法として，ソート型 TOEIC 評価法を提案し，種々の実験条件における評価結果を示した．実験の結果，従来の一対比較型 TOEIC 評価法とほぼ同等の結果がえられた．また，ソート型 TOEIC 評価法においては，異なる評価者においても，評価結果のズレが少ないことが示された.

最後に，ソート型 TOEIC 評価法により，どの程度の効率化が可能であるかについて考察する．同一の評価者が, 一対比較型 TOEIC 評価とソート型 TOEIC 評価とを行った結果が無いため，厳密な比較が出来ないが，ソート型 TOEIC 評価法では，一対比較型 TOEIC 評価法に比べ, 評価時間を約 20%程度に短縮することが可能であった.

謝辞

本研究は情報通信機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである.

参考文献

[1]長尾 真他,“Mu プロジェクトにおける日英翻訳結果の評価”,情処 NL 研, 1984-NL-047，1984.

[2]Sumita E. et. al.,“Solutions to Problems Inherent in Spoken language Translation : The ATR-MATRIX Approach”, Proc. of MT Summit, pp.229-235, 1999.

[3]菅谷史昭他 “音声翻訳システムと人間との比較による音声翻訳能力評価手法の提案と比較実験”, 信学論, D-Ⅱ, Vol. J84-D-Ⅱ, No.11, pp.2362-2370, 2001.

[4]水島 昌英他“実環境における音声翻訳システムを介した対話実験-実験室環境との発話スタイルの比較”，音響講論，1-P-17，pp.185-186，2005-09

[5]伊藤玄他“音声認識統合環境 ATRASR の概要と評価報告”，音響講論，1-P-30，pp.221-222，2004.

[6]Akiba, Y. et. al.,“Using language and translation models to select the best among outputs from multiple systems”，Proc. of COLING, pp.8-14，2002.

[7]Imamura, K. et. al.,“ Practical approach to syntax-based statistical machine translation”，Proc. of MT summit, pp.267-274，2005.

[8]Yasuda, K. et. al.,“ An Objective Method for Evaluating Speech Translation System: Using a Second Language Learner's Corpus”，IEICE Trans. on Info & Sys, VOL.E88-D, NO.3, pp.569-577, 2005.