統計的翻訳評価尺度に基づく日英翻訳文の訳質分析

鍔木　元[*1]　安田　圭志[*2]　山本　博史[*2]　匂坂　芳典[*1,2,3]

[*1] 早稲田大学 GITI　[*2] NiCT/ATR　[*3] 早稲田大学 ことばの科学研究所
hjm-tsubaki@asagi.waseda.jp　{keiji.yasuda, hirofumi.yamamoto, yoshinori.sagisaka}@atr.jp

１．はじめに

我々は、第二言語の生成、知覚能力の測定、自動評定についての研究を進めている。器械による音声言語能力の自動評定は、評定の省力化や評定者によらない評価の可能性等、有用性が期待できる。しかしながら、音声言語データの準備等には労力と時間がかかり、少ないデータを基に研究が緒に着いた段階である。

人間の翻訳、作文に関する言語能力の自動測定についてはこれまで、日本人英語の通じやすさの測定[1]、翻訳システムの自動評価手法の人間による翻訳の評価への適用[2]、英語コミュニケーション能力の自動測定[3]、自由英作文からの言語取得度の推定[4]等の研究がなされてきている。これらの研究結果からは、統計的な特徴量に基づく翻訳・作文能力の自動測定の可能性が伺える。

本研究では、日英翻訳を対象として、統計的自動翻訳で用いられる2つの統計的尺度についての利用可能性を検証した。日英翻訳文の英語らしさを示す単語Nグラム確率値、翻訳の妥当性を示す単語対応の翻訳確率値について、日英翻訳文に対して人が付与した主観評価値との相関関係を分析した。これらの尺度の翻訳評価への利用可能性、並びに、評定基準となる正解翻訳文が不要な翻訳評価の実現可能性を検討した。

２．翻訳評価のための統計的情報量分析

統計的言語翻訳では、次式に示すように2つの特徴量を用い、翻訳が行われる。

$$\hat{e} = \arg\max_{\text{all candidates}} p(e \mid j) = \arg\max_{\text{all candidates}} p(j \mid e)p(e)$$

p(e):対象言語文（英語）e の単語 N グラム確率
p(j|e):元文（英語）e が対象言語文（日本語）j に変換される翻訳確率

評価に用いる特徴量のうち、p(e)は、英語らしさを評価するための特徴量として理解できる。p(j|e)は、翻訳らしさを評価する特徴量と考えることができる。ここでは、これらの特徴量を用いて、客観評価尺度を考え、主観評価値との対応関係を分析した。

p(e)には、単語3グラム確率を使用した。学習者翻訳文のp(e)と正解翻訳のそれとの比を指標に用いた。p(j|e)には、単語対応の翻訳確率を用いた。同様に、学習者翻訳文のp(j|e)と、正解翻訳のそれとの比を求めた。

３．分析実験

2つの指標と主観評価値との対応関係を分析した。分析データには、ATR で収録された旅行会話基本表現データ（以下、BTEC）を用いた。内訳は次の通りである。

(1) BTEC 日本語会話文に対する、プロの翻訳者による日英翻訳文 162318 文。この対訳コーパスで言語モデル(英語)、翻訳モデルを構築。

(2) BTEC 日本語会話文の日本語文 510 文、それらに対するプロの翻訳者、日本人英語学習者の日英翻訳及び発話したものの書き下し文 11220 文を作成。決まり文句等を除いた日本語文 473 文、それらに対する日英翻訳文 10406 文を評価対象データに使用。翻訳文の構成は以下の通り。

- 日本語文１文あたり、プロの翻訳者が1つの日英翻訳文を生成（計 473 文）
- 日本語文１文あたり、21 名の日本人学習者が日英翻訳文を生成（計 9933 文）
- 日本人学習者の日英翻訳文には、1人のネイティブ評定者が表1の基準に基づき、5段階の主観評価値を付与

(BTEC 日本語会話文 1 文に対し、21 名の日本人学習者が訳出した日英翻訳文の組を以後、テストセットとする。実験では、473 セットを使用）

表１ 学習者英文の評定基準

訳質レベル	評定基準
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。

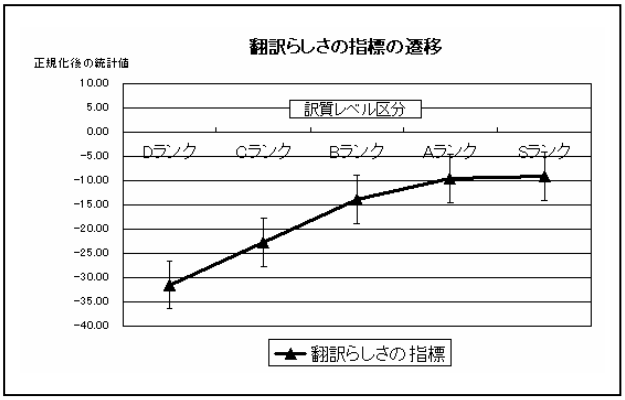

図1 翻訳らしさの指標と訳質レベルとの対応

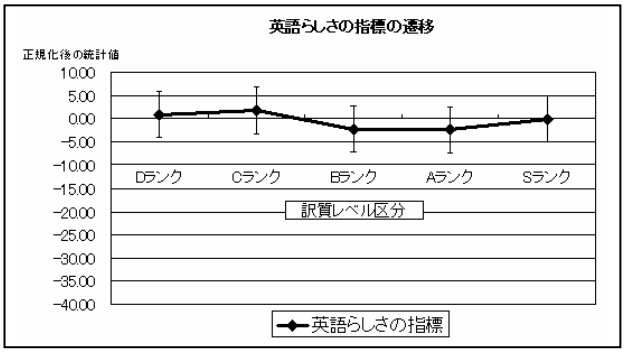

図2 英語らしさの指標と訳質レベルとの対応

日本人学習者が日英翻訳文 9933 文について、2つの指標を算出し、表1の訳質レベル毎にそれぞれ集計し、平均値を算出する。これら2つの指標の平均値と訳質レベルとの対応関係を図1、2に示す。

４．分析結果と考察

4.1 訳質レベルと統計的情報量の関係

翻訳らしさの指標は、図1のように訳質レベルが上がるに従い、平均値は増加傾向を示した。この対応関係から、この指標の翻訳評価への利用可能性が示された。

分析実験では、評価基準に基づいた絶対評価を行った。そのため、その基準に沿った正解翻訳が必要になった。一方、学習者の翻訳文同士の評価、すなわち相対評価の場合、評価対象翻訳文の統計的情報量のみの比較で評価できると考えられる。この指標の対応関係から、正解翻訳を必要としない翻訳評価の可能性も示めされた。

英語らしさの指標は、図2のように訳質レベルにかかわらず、横ばいの傾向を示した。この傾向は、人間が日英翻訳文を作る場合、英語の語順を意識して作成しているため、単語Nグラム確率において差異が生じないことを反映していると考えられる。

4.2 正解翻訳を用いる評価手法との比較

評価基準としての正解翻訳が翻訳評価に及ぼす影響を調べるために、正解翻訳文を用いる翻訳評価手法との比較を行った。評価手法には、BLEU を用いた。BLEU は機械翻訳による翻訳文の自動評価手法で、人間による正解翻訳を評価基準とし、単語 N グラムの重なりに基づき、翻訳システムの生成した翻訳文を評価する。スコアは次式で求められる。

$$S_{BLEU} = BP \times \exp\left(\sum_{n=1}^{N} w_n \log P_n\right)$$

$$\mathrm{BP} = \begin{cases} 1 & \text{if} \quad c > r \\ \exp(1-r/c) & \text{if} \quad c < r \end{cases}$$

c は、翻訳システム結果の文長
r は、人間による正解翻訳の文長
P_n は、修正 N グラム精度

翻訳システムの結果と人間による正解翻訳を、学習者の翻訳文とプロの翻訳者による正解翻訳文に置き換え、スコアを算出し、訳質レベルとの対応関係を見た。

図3のように、訳質レベルが上がるに従い、BLEU スコアの平均値は増加傾向にあった。図1の翻訳らしさの指標では誤差の重なりが大きい、AランクとSランク

が、BLEU スコアでは明確に区別された。この結果により、正解翻訳の語彙が翻訳評価に深く影響することを示した。

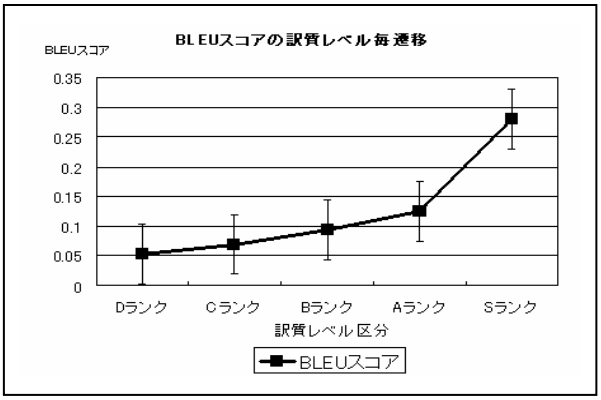

図3 BLEU スコアと訳質レベルとの対応

4.3 テストセット数と統計的情報量の関係

翻訳らしさの指標について、訳質レベルとの明確な対応関係が確認でき、翻訳評価に利用できる可能性が見出せた。しかし、誤差が大きく、訳文1文の指標のみで、その訳文を評価することは、現時点では難しい。そこで、 どのぐらいの量のテストセットを用意すれば、評価が可能になるのかを見るために、テストセット数を増加させ、翻訳らしさの指標と訳質レベルとの対応関係、分散を調べた。

図4のように、テストセット数が少ない場合、翻訳らしさの指標と訳質レベルとの対応関係が安定しない。だが、テストセット数の増加に伴い、対応関係は安定していくことが分かる。さらに表2からは、ランク内分散が小さくなっていることがわかる。このことは、一定量のテストセットがあれば、この指標で学習者の翻訳能力を評価する可能性を示している。

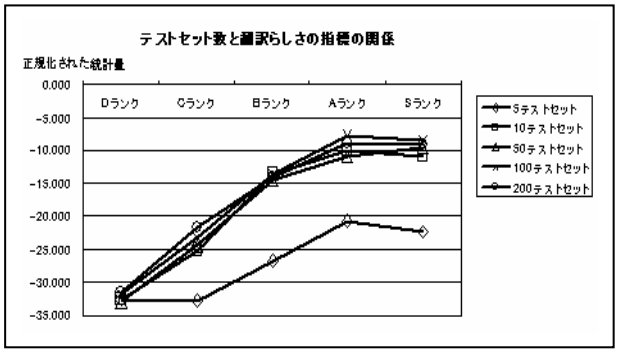

図4 テストセット数と訳質レベルの関係

表2 テストセット数とランク内分散

テストセット数	ランク内分散
10	14.34
50	10.74
100	9.66
200	9.61

5．まとめ

統計的自動翻訳で用いられる統計量を学習者の訳文評価に用いる可能性について検討を行った。英語らしさの指標、翻訳らしさの指標、それぞれについて翻訳文の客観評価への適用可能性について調べた。翻訳らしさの指標について、訳質レベルに対し規則性を示した。この結果、この指標は、翻訳評価に有用性を示したが、そのままでは、直接、用いられないことが判明した。

分析では、正解翻訳を用いて尺度化を計ったが、学習者間の比較は、正解文がなくとも行える。従来の客観尺度化の議論では、テスト文に対する正解は必須であるとされていた。本研究で考案した評価尺度の利用は、そのような正解を必ずしも必要としない評価法としての可能性に道を開いた。

謝辞

日頃統計的翻訳に関するご指導をいただく隅田室長をはじめとする ATR 音声言語コミュニケーション研究所の皆様に感謝致します。

参考文献

[1] 和泉絵美，内元清貴，井佐原均“日本人英語の通じやすさに関する研究”，言語処理学会 第12回年次大会, S1-4 pp16-19, 2006

[2] 山本誠一，菅谷史昭，安田圭志，隅田英一郎，“音声翻訳技術開発の経験に基づく外国語能力評価法の提案”，電子情報通信学会技術報告書, pp30-31, 2003

[3] 安田圭志，隅田英一郎，山本誠一，柳田益造，前川喜久雄，菅谷史昭“英語コミュニケーション能力の自動測定技術の提案”，情報処理学会研究報告 pp65-70, 2003

[4] 坂田浩亮，新保仁，松本裕治，“コーパスを用いた言語習得度の推定”，情報処理学会研究報告 pp113-119, 2007