

統計的自動翻訳に基づく日本人学習者の英文訳質分析

鏑木 元*¹ 安田 圭志*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, 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。

言語翻訳に関する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]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能力の自動測定の可能性が伺える。本稿では、日英翻訳を対象として、自動翻訳で用いられる統計値について、自動評定への利用可能性を調べた。

2. 翻訳評価のための統計的情報量分析

翻訳訳質の自動評価については、自動翻訳器の性能評価を目的として、自動翻訳文の評価がなされている[4]。これらの研究では NIST や BLUE といった訳文の統計的言語特徴量を使用することの有用性が示されている。ここでは、器械が生成した自動翻訳文に代えて学習者が生成した翻訳文を対象とし、それらが有する統計的言語特徴量と訳文評価の関係について分析した。

2.1 翻訳評価分析に用いた言語データ

分析には、ATR で収録された旅行会話基本表現データを用いた。このデータは 510 文からなる旅行会話文（日本語文）を、日本人英語学習者が翻訳し、発話したものを書き下した文全 10710 サンプルとなっている。これらの各訳文は、表 1 に示す基準に基づいて、評定者により、5 段階の主観評価値が与えられている。

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。

2.2 統計的言語特徴量と訳質の対応

分析の手始めとして、各訳文の N グラム統計値を求め、各評定基準毎に求めた値を表 2 に示す。

表 2 評定基準に対応する N グラム確率統計

訳質レベル	サンプル数	N グラム統計値		
		1-グラム	2-グラム	3-グラム
S ランク	3288	0.00017	0.0080	0.0557
A ランク	1445	0.00012	0.0068	0.0221
B ランク	2683	0.00015	0.0040	0.0125
C ランク	1372	0.00001	0.0076	0.0056
D ランク	1901	0.00020	0.0077	0.0480

この表に見られるように、各訳質レベルに対する全データを用いた場合、訳質レベルが上がるに連れてNグラム統計値は増加している。1グラムに比べ、2グラム、3グラムの順に対応が良くなっていることが判る。訳質が最低のDランクに上昇が見られるが、これは、Dランクに特徴的に見られる翻訳傾向に依存するものと考えられる。Dランクの訳文では、訳出が難しい箇所に対して訳出を行わないなど、訳文としては不完全さが目立つが、その分訳出される箇所そのものの間違いは少なくなっている。このことから、文書情報検索における再現率(recall)と適合率(precision)のように、単に翻訳英文の単語列出現確率のみならず、元の日本語文に対応する確率をも考慮した統計値が必要と考えられる。このため、以下に示すような統計的翻訳確率を加味した統計値により比較した。

$$\begin{aligned} & \text{統計的翻訳を考慮した特徴量} \\ & = \text{単語列}_1 \text{出現確率} \times \text{単語列}_1 \text{翻訳確率} \\ & \quad \times \text{単語列}_2 \text{出現確率} \times \text{単語列}_2 \text{翻訳確率} \\ & \quad \times \dots \\ & \quad \times \text{単語列}_n \text{出現確率} \times \text{単語列}_n \text{翻訳確率} \end{aligned}$$

出現確率、翻訳確率が高ければ、この特徴量は大きくなる。つまり、使用頻度の高い英語の単語列(英語らしい英語)で、元の日本語文を翻訳していることになる。

よく知られているように。統計的翻訳では、対訳コーパスを用いて以下に示す統計的翻訳確率の最大値を与える単語列候補を翻訳結果として用いる。

$$e = \operatorname{argmax}_{\text{all candidates}} p(e|f) = \operatorname{argmax}_{\text{all candidates}} p(f|e) p(e)$$

ここで
 $p(e|f)$:元文(日本語) fが対象言語文(英語) eに変換される条件付確率
 $p(f|e)$:元文(英語) eが対象言語文(日本語) fに変換される条件付確率
 $p(e)$:対象言語文(英語)のN-gram生起確率
 統計的翻訳確率を求める上で、ここでは簡便に、被験者の翻訳文が、正解翻訳に対応する単語列をどの程度含んでいるかを示す尺度として特徴量を用いている。結果を表3に示す。

表3 統計的翻訳を考慮した特徴量と評定基準の対応

訳質レベル	統計的翻訳を考慮した特徴量
S ランク	0.11172
A ランク	0.01622
B ランク	0.00042
C ランク	0.00004
D ランク	0.03358

この表に見られるように、統計的翻訳を考慮した特徴量は、Nグラムのみによる場合よりランクの差違を明確に示しており、訳質が最低のDランクに上昇も少なくなっている。

2.3 統計的言語特徴量を用いた評価

前節の統計値が示すように、統計的翻訳で用いられる種々の特徴値を訳文全体の評価として使用する可能性はあるが、個々の訳文に対してそのまま利用することはできない。ちなみに、上記の統計的翻訳を考慮した特徴量でもランク内分散(標準偏差)は0.00007~0.21となっており、ランク間の重なりは大きい。評定に前述した特徴量を用いるためには、どのような対象にどのように使用するかをさらに検討する必要がある。ここでは、まず、複数の文を用いた評定としての可能性調べるため、文数を50、100、150とした場合の統計的翻訳を考慮した特徴量の遷移を求めてみた。この結果、表4に示すように文数が増加するに連れて、評定値とクラスの対応が良くなってゆくことが判った。

表4 統計的翻訳を考慮した特徴量の文数における遷移

訳質レベル	50文	100文	150文
S ランク	0.0351	0.0458	0.0447
A ランク	0.0079	0.0211	0.0198
B ランク	0.0001	0.0002	0.0002
C ランク	0.00001	0.00001	0.00001
D ランク	0.0305	0.0225	0.0233

また、きちんとした統計的翻訳確率を求めた比較を行った上で調べるべきであるが、各文にこのような統計値を適用する場合、統計量を用いる対象の小ささによる推定誤差は避けることができない。確率値とランクの相違が大きいサンプルを調べた結果、以下のようなものが問題点として考えられた。

主観評価値の最低のランク D であっても、統計的翻訳を考慮した特徴量が高い値を示す。D ランクの上昇がここでも観察される理由でもある。

例：

元文：もっと大きな金庫はありませんか。

翻訳文：do you have

上記例の翻訳文には、元文の“もっと大きな金庫”が翻訳されていない。訳質基準に基けば、ランク D である。しかし、“”はありませんか”が、do you have と翻訳されており、生起確率、翻訳確率において高い確率が求まる。

3. まとめ

統計的翻訳で用いられる種々の特徴値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。この結果、訳文全体の評価として、単語 N グラム統計量、翻訳確率等の有用性が確認できた。また、単なる出現確率特徴量で測定が難しいと思われる要因の存在が認められた。今後は、精確な統計的翻訳確率の利用を初めとする統計的翻訳特徴の利用法に関する精緻化と、語彙情報を初めとする他の言語知識の併用について検討を進める予定である。

謝辞

日頃統計的翻訳に関するご指導をいただき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へ感謝致します。

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] 安田圭志, 菊井玄一郎 “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”, 言語処理学会 第 12 回年次大会, P3-3 pp252-255