音声認識誤りが音声翻訳品質に与える影響の定量的評価

○大田健紘†‡　安田圭志†　菊井玄一郎†　柳田益造‡
†ATR 音声言語コミュニケーション研究所　‡同志社大学
{kenkoh.ohta, keiji.yasuda, genichiro.kikui}@atr.jp myanagid@mail.doshisha.ac.jp

1. はじめに

本研究は，音声認識処理と自動翻訳処理を結合することにより構築されている音声翻訳システムにおいて音声認識性能が最終的な翻訳結果に与える影響を明らかにすることを目的としている．従来より，音声認識処理と自動翻訳処理それぞれの性能の評価は行われているが，音声翻訳システムにおいては，これらを結合した音声翻訳システム全体としての品質が重要であり，音声認識等の要素技術に対しても翻訳結果への影響を考慮した評価指標が必要となる．たとえば同じ 1 語の認識誤りでも翻訳結果によりダメージを与える語については大きな重みにすべきであろう．

本稿では，ATR 音声言語コミュニケーション研究所で研究開発された，音声認識システム ATRASR[1]と自動翻訳システム SAT(Statistical ATR Translator)[2]を結合した音声翻訳システムを用いて上記の問題を検討する．具体的には，この音声翻訳システムの音声認識部の出力と，これを自動翻訳した結果の精度の関連を調べ，この結果から，どのような音声認識誤りが自動翻訳システムの出力に悪影響を及ぼすかを明らかにすると同時に，影響の程度を定量的に評価する．

本研究の知見を用いることにより，音声認識システムをそれと組み合わせる自動翻訳システムに応じて最適化することができ，最適な統合システムの実現が期待できる．

2. 実験システムの概要

2.1 音声認識部

大語彙連続音声認識システム ATRASR の探索は 2 パスで構成されており，1 パスでは音響モデルとして HMM，言語モデルとして 2-gram を用いた探索を行い，単語ラティスを出力する．そして 2 パスでは，3-gram を用いて探索結果の再評価を行い，*N*-best を出力する．ただし，分析には 1-best のみを用いる．

2.2 自動翻訳部

統計的自動翻訳システム SAT は，greedy というアルゴリズムを用いている．greedy とは，原言語と目的言語のパラレルコーパスを用いて，入力された原言語文に最も近い原言語文に対応する目的言語文を *N*-best で選択する．次に各目的言語文に対して，尤度を計算する．各目的言語文について単語の置き換えを行い，尤度を再計算する．これらの処理を尤度の変化が小さくなるまで繰り返し結果を出力する．

3. 分析対象データ

3.1 Machine-Aided Dialogues (MAD) Corpus

MAD コーパスとは，異なる言語を話す話者が音声翻訳システムを介して模擬対話を行い，人が音声翻訳システムに対してどのような言葉をどのように話すのかを調査するために収集されたコーパスである．MAD コーパスは日本語話者の 1370 発話と，英語話者の 1293 発話で構成されており，今回はこの中から抽出された日本語話者の 502 発話を用いた．このコーパスは音声データと，発話内容の人手による書き起こしテキストから構成されている．

3.2 主観評価による評価対象の絞込み

音声認識が正しく行われているにも拘わらず，自動翻訳品質がよくない発話は，本研究の分析対象データとして好ましくない．なぜなら，それらの発話は音声認識誤りが含まれたとしても，自動翻訳品質が悪くならないためである．そこで，評価対象を絞り込むために自動翻訳品質を評価する．評価方法[3-6]は大きく分けて主観評価と自動評価があるが，ここでは主観評価を行う．主観評価は，人手を要するためコストと時間がかかるという問題点があるが，各文に対する評価は利用者の直感を反映している．502 発話の人手による書き起こしテキストの自動翻訳出力を以下の基準に従って 4 段階で主観評価する．

A(完全訳)：訳文だけでまったく問題なし
B(部分訳)：訳文は少し情報が欠けている
C(可能訳)：訳文はかなり情報が欠けている
D(不可訳)：訳文からは，情報が想像もできない

主観評価結果が A もしくは B ランクの発話を分析に用いる．これにより，分析対象のデータは 502 発話の内 209 発話となる．209 発話に含まれる品詞の内訳を調査した．その結果を図 1 に示す．

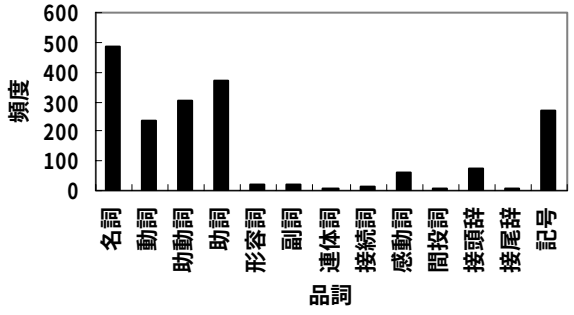

図 1 テストセットの品詞の内訳

4. 音声翻訳出力品質の分析

4.1 分析手順

まず 209 発話の原言語の音声データに対して音声認識を行う．音声認識の探索ビーム幅をさまざまな値に変更し，実験を行うことで大量の音声認識結果を生成する．音声認識結果については，正解系列との DP マッチングを行うことで，含まれている音声認識誤りを抽出する．そして，自動翻訳品質の評価には客観評価の 1 つである単語誤り率(WER)と主観評価を用いる．ただし，客観評価は一文ごとの評価の精度はあまりよくないが，大量のデータに対する評価では主観評価と高い相関が見られる．WER は，音声認識誤りを含む系列に対して自動翻訳を行い，正解系列を自動翻訳した系列を参照文として式(1)により計算する．

$$WER = \frac{\text{挿入語数＋置換語数＋脱落語数}}{\text{参照文の単語数}} \quad (1)$$

以降音声認識結果の WER は WER_{rec} と記す．そして，認識誤りの種類と自動翻訳結果の WER(以降 WER_{trans})を用いて重回帰分析を行い，音声認識誤りと WER_{trans} の関係を定量的に評価する．

4.2 音声認識結果と自動翻訳結果の評価

分析に用いるデータは，MAD コーパスのテストセット 209 発話を音声認識システム ATRASR の探索ビーム幅を 5 種類に変更し，のべ 1045 発話を作成した．これらの音声認識結果に含まれる音声認識誤りの種類を調査した．表 1 に，置換，削除，3 種類の認識誤りの割合を示す．

表 1　誤挿入，削除，置換誤りの割合(%)

誤挿入	削除	置換
28	14	58

各認識誤りについて，品詞ごとの頻度を調査した(表 2)．置換誤りについては，最も頻度の多かった名詞についてどういう品詞に置換されたか，またどういう品詞から置換されたかを図 2 に示す．音声認識誤りに関しては，削除誤りは記号が多く，誤挿入は助詞の誤りが多い．また，置換誤りは，名詞から名詞に置換されることが最も多く次いで，名詞から助詞への置換が多い．

4.3 重回帰モデルを用いた分析

1)WER_{rec} と WER_{trans} の関係を調べるために単回帰分析を行った．独立変数は WER_{rec} とし，従属変数は WER_{trans} とする．ここで実際に分析に用いることができるデータは，のべ 2510 発話のうち，音声認識の正解系列の自動翻訳出力の主観評価結果が A または B ランク(209×5 発話)である．重回帰分析を行った結果，式(2)の関係を得た．

$$\textit{自動翻訳の}WER = 1.34 \times \textit{音声認識の}WER + 0.046 \quad (2)$$

横軸に WER_{rec}，縦軸に WER_{trans} をとり 1045 発話のデータをプロットし，回帰直線を引いた(図 3)．

表 2 各認識誤りの品詞別の割合(%)

	削除	誤挿入	置換	
			前	後
名詞	4.14	8.28	27.10	25.63
動詞	0.13	2.54	7.61	7.21
助動詞	0.27	1.34	4.81	5.21
助詞	3.07	8.54	10.68	13.48
形容詞	0.53	0.00	1.07	0.67
副詞	0.00	0.80	1.20	0.27
連体詞	0.00	0.80	0.00	0.67
接続詞	0.00	0.27	0.67	1.47
感動詞	0.00	0.13	0.00	0.00
間投詞	0.67	0.53	2.00	1.20
接頭辞	0.53	2.54	0.27	0.67
接尾辞	0.00	0.00	0.67	0.80
記号	5.07	2.27	1.47	0.27

表 3 置換，誤挿入，削除誤りの標準回帰係数 (**：p<0.01)

置換	削除	誤挿入
0.404**	0.286**	0.264**

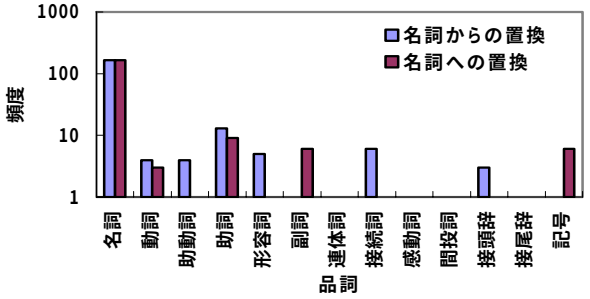

図 2 置換誤りのうち名詞について詳細に分類

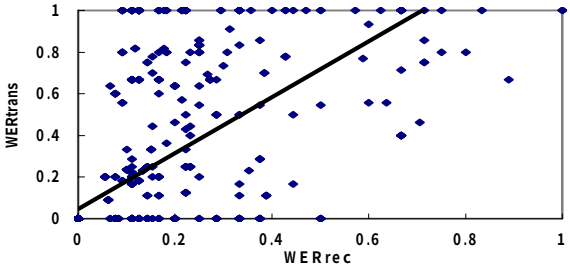

図 3 WER_{rec} と WER_{trans}

単回帰分析の結果求まったモデルの当てはまりのよさを表す指標に寄与率がある．式(2)のように求まったモデルの場合，寄与率は 0.451 である．寄与率は 1 に近いほど，当てはまりがよいということを表している．

2)次に，置換，誤挿入および削除の各認識誤りが自動翻訳品質に与える影響を調べた．独立変数は置換数，誤挿入数および削除数を発話の単語数で割ったものとし，従属変数を WER_{trans} として重回帰分析を行った．その結果，WER_{trans} に最も大きな影響を与えるのは，置換誤りであり，次いで削除誤り，そして誤挿入である．表 3 に重回帰分析の標準回帰係数を示す．重回帰分析の標準回帰係数は従属変数へ

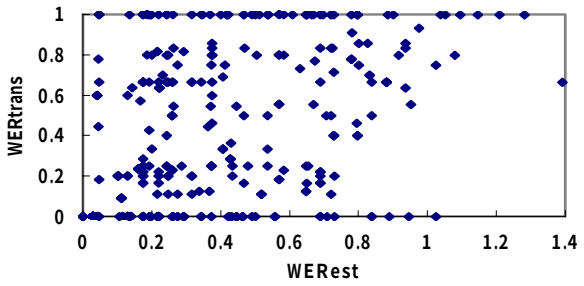

図 4 詳細に分類した音声認識誤りによる重回帰

の影響の大きさを表す指標である．

表 3 の結果より，各係数は有意水準 1%で有意であり，信頼できる係数であることがわかる．この場合の寄与率は 0.459 である．

3)各音声認識誤りをさらに品詞別に分類し，それぞれが自動翻訳に与える影響を調査した．独立変数は，削除誤りおよび誤挿入の名詞，動詞，助動詞，助詞，形容詞，副詞，連体詞，接続詞，感動詞，間投詞，接頭辞，接尾辞および記号，置換に関しては頻度の高かった名詞から名詞への置換，名詞から助詞への置換，助詞から名詞への置換，その他の置換を用いた．独立変数は WER_{trans} とした．ただし，出現頻度が 0~2 の品詞は独立変数から取り除いた．表 4 に各音声認識誤りの標準回帰係数を示す．

このモデルに対する寄与率は 0.601 である．重回帰により求めた WER_{est} と分析に用いた 1045 発話の WER_{trans} をプロットした(図 4).

5. 考察

5.1 音声認識誤りの詳細分類による分析

表 3 に示したように置換，誤挿入および削除誤りを比較すると，置換誤りが，WER_{trans} に対して最も影響が大きいことがわかった．

ここでは，さらに詳細に各音声認識誤りの品詞ごとの影響を調べた．表 4 より，記号の削除誤りが最も影響が強いという結果を得た．そしてほぼ同程度で名詞から名詞への置換誤りという結果を得た．その他の置換誤りに関しては，複数品詞を足し合わせているので，各品詞が置換されることによる影響は小さいと考えられる．記号の削除誤りが最も影響が強い理由は，自動翻訳の翻訳モデルを学習したコーパスの影響と考えられる．記号が削除されると，文の区切れが判別できず，自動翻訳に失敗している．

重回帰分析は，各変数間に独立性を仮定しているため, 相関があると正しく分析できない. そのため, 各変数間の相関係数を調べた．その結果，名詞の削除と名詞から名詞への置換(-0.492)，名詞の誤挿入と名詞から名詞への置換(-0.346)，そして名詞の誤挿入と助詞の誤挿入(-0.298)の 3 組において他と比較して高い相関があった．これは，長い名詞が短い 2 つ以上の名詞に分割された場合，置換と誤挿入が同時に起こるためと考えられる．逆も同様に起こりえる．

表 4 各音声認識誤りの標準回帰係数
(*：p<0.05，**：p< 0.01)

誤りの種類	標準回帰係数	誤りの種類	標準回帰係数
削除名詞	0.066**	誤挿入名詞	0.142**
削除助詞	0.169**	誤挿入動詞	0.207**
削除形容詞	0.023	誤挿入助動詞	-0.051*
削除間投詞	-0.007	誤挿入助詞	0.015
削除接頭辞	0.077**	誤挿入副詞	0.103**
削除記号	0.352**	誤挿入連体詞	0.013
置換名詞→名詞	0.291**	誤挿入間投詞	0.151**
置換名詞→助詞	0.035	誤挿入接頭辞	-0.044*
置換助詞→名詞	0.109**	誤挿入記号	0.101**
置換その他	0.329**		

表 5 各音声認識誤りの標準回帰係数
(*：p<0.05，**：p< 0.01)

誤りの種類	標準回帰係数	誤りの種類	標準回帰係数
削除名詞	0.123**	誤挿入名詞	0.028
削除助詞	0.105**	誤挿入助詞	0.249**
削除形容詞	0.006	誤挿入副詞	0.082**
削除間投詞	-0.001	誤挿入間投詞	0.251**
削除接頭辞	0.087**	誤挿入接頭辞	0.012
削除記号	0.44**	誤挿入記号	0.072**
置換名詞→名詞	0.337**	置換助詞→名詞	0.132**
置換名詞→助詞	-0.012	置換その他	0.337**

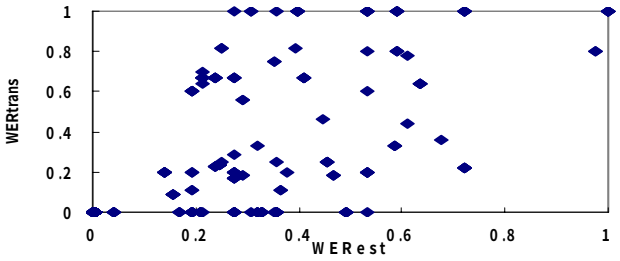

図 5 WER_{rec} が低い場合の重回帰

5.2 WER_{rec} が低い場合の分析

現状の音声認識システムの性能は WER で 0~0.2 程度である．また，図 3 より WER_{rec} がある一定以上の値になると，自動翻訳品質は非常に悪くなることがわかる．そのため，WER_{rec} がある値以上になると WER_{trans} を評価する必要が無いと考えられる．そこで，WER_{rec} が 0.2 以下のデータに対してのみ，3)の分析を行った. データ数は 894 発話である. ただし，従属変数および独立変数は 3)と同様であり，その中から出現頻度が 0~2 の品詞は分析から取り除いた. 表 5 に各音声認識誤りの標準回帰係数を示す.

表 5 より, WERtrans に最も影響が大きい認識誤りは，記号の削除誤り，名詞から名詞の置換である．この結果は全データを使って分析した表 4 の結果と同じである．そして，分析により得られたモデルに対する寄与率は 0.605 であり全データを使った場合より若干当てはまりがよくなる．図 5 に WERest と WER_{trans} をプロットした.

表 7 音声認識誤りにより主観評価結果が変化する場合の WER_{trans} の平均と WER_{rec} の平均（()は標準偏差）

		音声認識出力の自動翻訳							
		A		B		C		D	
		翻訳	認識	翻訳	認識	翻訳	認識	翻訳	認識
書き起こしテキストの自動翻訳	A	0.035(0.14)	0.026(0.076)	0.62(0.41)	0.17(0.098)	0.64(0.26)	0.26(0.15)	0.72(0.28)	0.45(0.25)
	B	0(0)	0(0)	0.028(0.12)	0.038(0.085)	0.59(0.32)	0.16(0.14)	0.73(0.25)	0.28(0.14)

表 6 音声認識誤りによる主観評価結果の変化

		音声認識出力の自動翻訳			
		A	B	C	D
書き起こしテキストの自動翻訳	A	643	28	34	60
	B	0	208	30	42

5.3 主観評価による分析

分析により得られた結果について主観評価との対応を調べるために，まず音声認識誤りにより主観評価がどのように変化するかを調べた(表 6)．そして，ランクの変化が起こる場合に対応する WER_{rec}，WER_{trans} がどの程度になっているのかを調べた(表 7)．表 7 より，書き起こしテキストの自動翻訳の主観評価が A ランクと B ランクの場合を比較すると，B ランクの方が WER_{rec} が低くても主観評価が悪くなっていることがわかる．つまり，重大な認識誤りが無い状態で自動翻訳しても品質の悪いデータは，少しの音声認識誤りの影響を受けやすいということである．この結果から，図 3 のように WER_{rec} が低い場合は，WER_{trans} の分散が大きくなると考えられる．そこで，書き起こしテキストの自動翻訳結果が A ランクの発話と B ランクの発話を分けて 1)の分析を行った(図 6)．図 6 より，A ランクの発話は式(2)の回帰直線より傾きが小さくなり，B ランクの発話は傾きが大きくなった．音声翻訳の出力品質は書き起こしテキストの翻訳品質にも依存していることがわかる．

また，A ランクから C，D ランクに評価が低下する場合の WER_{trans} と，B ランクから C，D ランクに評価結果が低下する場合の WER_{trans} は同程度である．このことから，音声認識出力の自動翻訳品質に関して，自動評価と主観評価に関連があると考えられる．

次に，主観評価結果を数値に対応させて，図 4 とあわせてプロットし，自動評価と主観評価の場合の対応を確認する．A ランクを 0，B ランクを 0.34，C ランクを 0.67，そして D ランクを 1 とした(図 7)．

そして WER_{trans}(観測値)と，WER_{est}(理論値)，主観評価を数値化した値の間の相関を調べた(表 8)．表 8 より，観測値，理論値，主観評価の間には比較的強い相関が見られる．そのため，重回帰により得られる理論値は音声翻訳システムの出力品質を評価できていると考えられる．

6. まとめ

重回帰分析を用いて，音声認識誤りと音声翻訳の出力品質の関係を調査した．その結果，記号の削除，名詞から名詞の置換，その他の置換が音声翻訳の出力品質への影響が大きいことがわかった．逆に助詞の誤挿入や連体詞の誤挿入はほとんど影響を与えないという結果になった．しかし，図 4 や図 7 のように分散が大きく必ずしも分析に用いた独立変数のみに WER_{trans} が依存していないという結果も得た．

今後はデータ量を増加させ，より詳細な分類を行い，音声認識誤りと音声翻訳品質の関係を導出する．

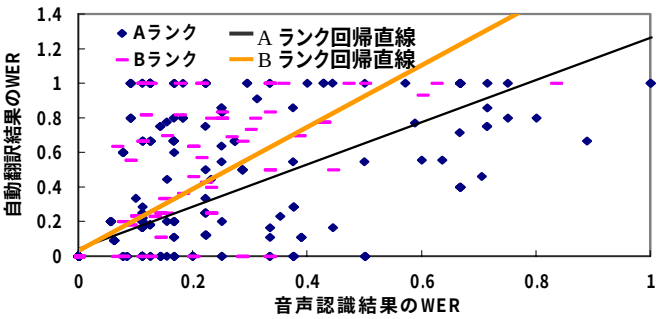

図 6 A ランクと B ランクの回帰直線

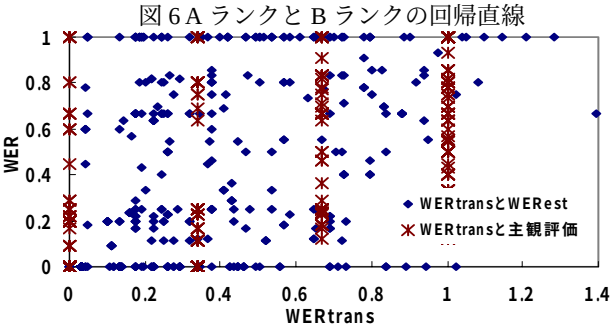

図 7 自動評価と主観評価の対応

表 8 観測値，理論値，主観評価の間の相関

	主観	観測値	理論値
主観	1		
観測値	0.712	1	
理論値	0.63	0.67	1

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] 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.

[2] Watanabe,T.et.al.,”Example-based Decording for Statistical Machine Translation”, MT Summit IX, pp.410-417, 2003.

[3] Papineni,K.et.al.,“Bleu: a Method for Automatic Evaluation of Machine Translation”, Proc. ACL, pp311-318, 2002.

[4] NIST, “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics”, http://www.nist.gov/speech/tests/mt/, 2002.

[5] 安田圭志他 “ 対訳コーパスを用いた翻訳品質自動評価法 ”，情報学論，Vol.43，No.7，pp.2108-2116，2002.

[6] 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.