医療通訳分野での音声翻訳機の有用性に関する研究

「HIV 検査と医療へのアクセス向上に資する多言語対応モデルの構築に関する研究」班

研究分担者　宮首　弘子　杏林大学外国語学部教授
沢田　貴志　神奈川県勤労者医療生活協同組合港町診療所所長
研究代表者　北島　勉　　杏林大学総合政策学部教授

研究要旨

経済のグローバル化や労働人口の減少に伴い、日本で暮らす在留外国人及び訪日外国人が年々増えている。この傾向は 2020 年の世界的新型コロナ禍によって、国際的な人的移動が大きく制限されたことで、予測できないものとなった。とは言え、外国人観光客や労働力としての外国人の受け入れに必要な医療通訳へのニーズが消えたわけではなく、コロナ感染拡大防止の対応策の一つとして電話や Zoom などによる遠隔医療通訳の需要が高まったことから、むしろ必要不可欠であることが再認識されたと言えるだろう。

日本政府（厚労省）は、医療通訳者の確保と養成を強化しているが、財源の確保と通訳人材の確保、とりわけ希少言語の人材確保は依然として困難である。また医療通訳に対する休日・夜間の対応、利用頻度が少ないことの非効率性が課題とされている。したがって、政府は、医療通訳者養成に力を入れる一方で、様々な IT ツールを使った電話通訳や遠隔通訳の対応が広がり、さらに、AI を活用した多言語「音声通訳・翻訳」（以下「音声翻訳」）機器で対応する方向性も整備している。総務省のグローバルコミュニケーション計画においても音声翻訳の重点整備分野の一つに医療通訳分野が挙げられている。

本研究は、音声翻訳機の一つとして医療現場で使用の広がりを見せているソースネクスト社の「POCKETALK(ポケトーク)」を用いて、その使い勝手の良し悪しや翻訳の信頼性について、医療現場での使用を想定した模擬実証研究を行って検証する。そこから POCKETALK の現状における通訳エラーのでやすいところを突き止め、コミュニケーションの成立を可能とする音声翻訳機の医療通訳としての有用性を考察したいと考える。

A. 研究目的

一般的に、文字テクストを他言語に変換することを「翻訳」、音声を他言語に変換することを「通訳」という。「音声翻訳」とは「音声認識→テクスト変換→他言語テクスト変換→他言語音声変換」という過程の総称であり、翻訳と通訳を複合した概念であると考えられる。

POCKETALK は現時点における最も汎用性のある音声翻訳機の一つであると認められる。変換エンジンには総務省情報通信研究機構（NICT）が開発しているクラウドサービス（VoiceTra）が使われている。

総務省のグローバルコミュニケーション計画 1)においても重点分野の一つに挙げられている医療通訳分野に用いた場合、現時点においてどのくらいの有用性が確認されるであろうか。このことを、医療通訳の模擬臨床現場におけるロールプレイ研修用シナリオを用いて検証してみたい。

B. 研究方法

1. AI 音声翻訳機の仕組みと検証目的

POCKETALK は、音声翻訳専用モバイル機器である。通信の高速化により、現場での通訳においてもタイムラグを感じない快適な使用感が実現していると謳われている。その音声翻訳の処理プロセスは､開発したソースネクスト社のホームページ[2)]の資料から次のような流れにまとめられる。

① 音声送信：ユーザーが喋った音声が 3G/4G 通信または Wi-Fi を通じて POCKETALK のクラウド・サーバにストリーミングで送られる。
② 音声認識：ストリーミング・データを音声認識エンジン（音声認識技術）がソース言語のテクストに起こす。
③ テクスト翻訳：ソース言語のテクストを翻訳エンジン（多言語翻訳技術）によってターゲット言語のテクストに翻訳する。
④ 音声変換：ターゲット言語のテクストを音声合成エンジン（音声合成技術）で音声に変換する。
⑤ 音声受信：ターゲット言語の音声がストリーミングで送り返されてくる。
⑥ テクスト受信：音声に併せて、ソース言語とターゲット言語の両方のテクストが画面に表示される。

上記②、③で生成されるテクストデータはクラウド上の POCKETALK センターに保存される。

本研究では、②の音声認識、③のテクスト翻訳について、その有用性の検証を試みる。多言語音声翻訳のうち、「日本語→中国語」、「中国語→日本語」の音声翻訳を代表例として、その有用性を模擬実証して考察する。

２．模擬実証研究の設定

検証対象は、当該研究班の医療通訳研修で用いているロールプレイ実習用の次の２つのシナリオである[3)]。各シナリオには、原稿として日本語テクスト及び中国語テクストが用意されているので、他言語からの通訳の基準となる「参照訳」として利用できる。

・シナリオ１（S1）：医師 D（日本人）が患者 P（外国人）に HIV に感染していることを初めて告知する場面における両者の対話（詳細略）
・シナリオ２（S2）：保健師 H（日本人）が結核に感染した患者 P（外国人）に初回の面接を行い今後の治療について説明する場面における両者の対話（詳細略）

両シナリオにおいて、患者（外国人）として中国人を設定して、POCKETALK による中国語と日本語の間の翻訳を試みた。検証項目の区分は次のとおりである。

(1)語彙レベルの翻訳

各シナリオで用いられる医療専門用語及び医療者（日本語）がよく使うフレーズについて、日本語の音声認識及び多言語翻訳の精度を確認する。一単語あるいは一フレーズごとに音声翻訳して POCKETALK センターにテクストデータとして記録する。日本語音声は日本語ネイティブ、中国語音声は中国語ネイティブが担当した。

(2)対話レベル（シナリオの翻訳）

各シナリオ全体について医療者（日本語）・患者（中国語）それぞれの音声認識及びテクスト翻訳の精度を確認する。一文ごとに音声翻訳して POCKETALK センターにテクストデータとして記録する。医療者役（日本語）を日本語ネイティブ、患者役（中国語）を中国語ネイティブが担当した。

３．評価方法

(1)語彙レベル

この検証は医療専門用語に対応する能力の確認であり、正確に対応している割合をもって正確度として測定する。またエラー箇所を確認して発生の傾向を分析した。

(2)対話レベル

①BLEU スコア

各シナリオの各言語に対して、音声認識とテク

スト翻訳に分けて、機械翻訳の自動評価尺度として定着している BLEU スコアを用いて精度の評価を行った。BLEU(Bilingual Evaluating Understudy)とは、翻訳文に対し、基準となる参照訳を比較して、共通する語・フレーズの数を計測してその割合の高低で評価する方法である[4]。またテクスト翻訳の精度の比較のために、Google 翻訳を使用して、言語ごとに翻訳文を作成し、BLEU スコアによる評価測定を行い、POCKETALK の BLEU スコアと比較した[5]。

表１　BLEU スコアの解釈基準[6]

BLEU スコア	解釈
< 10	ほとんど役に立たない
10〜19	主旨を理解するのが困難である
20〜29	主旨は明白であるが、文法上の重大なエラーがある
30〜40	理解できる、適度な品質の翻訳
40〜50	高品質な翻訳
50〜60	非常に高品質で、適切かつ流暢な翻訳
> 60	人が翻訳した場合よりも高品質であることが多い

②エラー分析

BLEU スコアとは別に、当研究班員（宮首）は、各シナリオの音声認識（聞き取り）におけるエラーとテクスト翻訳におけるエラー箇所を、それぞれの変換テクストから洗い出し、一箇所ずつエラーの原因を分析した。そこから全体のエラーの傾向を考察した。

C. 研究成果

１．語彙レベルの音声翻訳

ここでは、各シナリオ中の医療者の発話（日本語）に含まれる医療専門語彙が POCKETALK によってどの程度正確に中国語に翻訳されるかを検証した。検証データは POCKETALK センターに保存したテクストデータであり、参照データは各シナリオにある参照訳中のデータである。

実証結果は表２のとおりとなった。

表２　医療専門語彙の正確率

シナリオ	語彙数	音声認識 正確数(率)	テクスト翻訳 正確数(率)
S１：HIV	28	27(96%)	27(96%)
S２：結核	31	29(94%)	29(94%)

正確率は二つのシナリオとも 90%以上であり、医療専門語彙についてほぼ正確な音声認識とテクスト翻訳が期待できることが確認された。

しかしながら、専門語彙にもかかわらず音声認識･テクスト翻訳において各３点のエラー（誤認、誤訳）が発生していることから、音声翻訳のリスクを確認するために、具体的にエラーを分析した。この結果、語彙レベルのエラーは音声誤認や同音異語の誤選択によって発生していることがわかった（表３）。

表３　語彙レベルのエラー

シナリオ	語彙（日本語）	音声認識（日→日）	テクスト翻訳（中→日）	エラー分析
S１：HIV	AR治療法	ALT治療法	ALT治疗	音声ＲとＬの誤認
S２：結核	排菌する	配筋する	分配	同音異義語の誤選択 意味不明なテクスト変換
	菌が外に出る	金が外に出る	出钱	同音異義語の誤選択

２．対話レベルの音声翻訳

ここでは、各シナリオ別に全体の対話を POCKETALK で翻訳し、日本語発話（医療者）と中国語発話（患者）に分けて集計して、それぞれの言語に対するテクスト翻訳の精度を測定した。集計する検証データは POCKETALK センターに保存したテクストデータであり、参照データは各シナリオにある参照訳である。

実証結果は表４のとおりとなった。

表４　対話レベルの日中翻訳の BLEU スコア

シナリオ	発話者	文数	音声認識 BLEU スコア	テクスト翻訳BLEU スコア	(参考) Google翻訳 BLEUスコア
S1	医師（日本語）	30	84.05	10.31	11.80
:HIV	患者（中国語）	18	60.09	17.21	10.64
S2	保健師（日本語）	55	77.60	7.84	5.72
:結核	患者（中国語）	24	54.41	17.81	14.18

(1)音声認識

①BLEU スコア

日中両言語とも BLEU スコアが 50 点超であり、POCKETALK が「非常に高品質」な音声認識の精度を有することが確認された。特に日本語の音声認識においては「人が翻訳した場合よりも高品質」であると解釈される。

②エラー分析

日本語音声認識については、シナリオ１（S1）では５箇所（４センテンス)、シナリオ２（S２）では 11 箇所（8 センテンス)、合計 16 箇所（12 センテンス）のエラーが確認された。

中国語音声認識については、シナリオ１（S1）では３箇所（３センテンス)、シナリオ２（S２）では 5 箇所（5 センテンス)、合計 8 箇所（8 センテンス）のエラーが確認された。

(2)テクスト翻訳

①BLEU スコア

２つのシナリオにおいて「日本語→中国語」「中国語→日本語」ともに、BLEU スコアが 20 点以下であった。このことから、POCKETALK の日本語・中国語のテクスト翻訳の精度は「趣旨を理解するのが困難なレベル」以下と判定される。

②エラー分析

「日本語→中国語」のテクスト翻訳についてはシナリオ１（S1）では 19 箇所（15 センテンス)、シナリオ２(S２)では 33 箇所(28 センテンス)、合計　52 箇所（43 センテンス）のエラーが確認された。

「中国語→日本語」テクスト翻訳については、シナリオ１（S1）では 9 箇所（8 センテンス)、シナリオ２（S２）では 10 箇所（9 センテンス)、合計 19 箇所(17 センテンス)のエラーが確認された。

D．考察

１．BLEU スコアの考察

POCKETALK のテクスト翻訳については、「日本語→中国語」の翻訳よりも「中国語→日本語」の翻訳のほうが、約２倍の高い BLEU スコアで評価された。このことは POCKETALK の翻訳能力の特性という以上に、日本語には主語が省略されるなど翻訳される言語で必須の要素が省略されることがあるため、一般的に日本語から他言語への翻訳が困難であることが裏付けられたものと考えられる。

また、BLEU スコアから見る限り、POCKETALK のテクスト翻訳は Google 翻訳によるテクスト翻訳より、「日本語→中国語」変換を除き、概ね優れていることが窺える。

BLEU は連接する語句の共通性で測定するスコアであることから、語順や意味は考慮されないため、はたして POCKETALK が「趣旨を理解するのが困難なレベル」であるかについては、具体的にエラー（誤認、誤訳）を分析・考察する必要がある。

２．音声認識エラーの考察[7)]

POCKETALK による日本語の音声認識は、2 つのシナリオ合計 85 センテンスの日本語発話に対し、合計 16 エラー箇所及び 12 エラー・センテンスが確認された。それに対し、中国語の音声認識は、2 つのシナリオ合計 42 センテンスの中国語発話に対し合計 8 エラー箇所及び 8 エラー・センテンスが確認された（表５)。

表５　音声認識のエラー箇所のまとめ

	日本語 音声認識	中国語 音声認識
センテンス数(A)	85	42
エラー箇所	16	8
エラー・センテンス(B)	12	8
精度（A-B）/A	85.9%	81.0%
エラー箇所の分類		
同音異義語	4	5
音の聴き間違い	12	2
音の聴き漏れ		1

エラー箇所によりセンテンスの意味が不明になるものと想定して、POCKETALK の音声認識の精度を「意味の伝わるセンテンス（非エラー・センテンス）の全センテンスに対する割合」とするならば、日本語は 85.9%、中国語は 81.0%で、両言語の認識に遜色がないことがわかる。またこの数値は「非常に高品質」な音声認識機能の裏付けとなったものと考える。

エラー原因としては、同音異義語は文脈からの推定も難しいものであり、通訳者の限界にも類似している。それに対し、音の聴き間違いによるものが多く、通訳者ならば補ったであろう音声を聞き落としており、AI 翻訳の限界が窺える。

3．テクスト翻訳エラーの考察 [8)]

POCKETALK による日本語から中国語へのテクスト翻訳は、2 つのシナリオ合計 85 センテンスの日本語発話に対し合計 53 エラー箇所及び 43 エラー・センテンスが確認された。それに対し、中国語から日本語へのテクスト翻訳は、2 つのシナリオ合計 42 センテンスの中国語発話に対し合計 19 エラー箇所及び 17 エラー・センテンスが確認された（表６）。

表６　テクスト翻訳のエラー箇所のまとめ

	日→中 テクスト 翻訳	中→日 テクスト 翻訳
センテンス数(A)	85	42
エラー箇所	53	19
エラー・センテンス(B)	43	17
精度（A-B）/A	49.4%	59.5%
エラー箇所の分類		
音声誤認識(a)	14(26.4%)	5(26.3%)
明示化要す(b)	11(20.8%)	0
不要な重複(c)	4(7.5%)	0
不適な付加(d)	1(1.9%)	0
不適な省略(e)	4(7.5%)	5(26.3%)
不適な語彙選択(f)	13(24.5%)	7(36.8%)
不適な係り受け(g)	4(7.5%)	2(10.5%)
不適な語順(h)	2(3.8%)	0
エラー再分類		
音声認識エラー (a)	14(26.9%)	5(26.3%)
語用エラー (b)+(c)+(d)+(e)	20(37.7%)	5(26.3%)
意味エラー (f)	13(24.5%)	7(36.8%)
構文エラー (g)+(h)	6(11.3%)	2(10.5%)

音声認識と同様に、エラー箇所によりセンテンスの意味が不明になるものと想定して、POCKETALK のテクスト翻訳の精度を「意味の伝わるセンテンス（非エラー・センテンス）の全センテンスに対する割合」とするならば、「日→中」テクスト翻訳の精度は 49.4%、「中→日」テクスト翻訳は 59.5%である。POCKETALK のテクスト表示機能を使えば、誤った音声認識はキャンセルできるものの、一センテンスごとに意味の伝わる精度が 5 割あるいは 6 割であるとすると、連続した相互の対話は継続することが困難となるであろう。それゆえ、これらの数値は BLEU スコアによる「趣旨を理解するのが困難なレベル」判定の裏付けとして理解してよいと考える。

また、日本語から中国語への翻訳精度は中国語から日本語への翻訳精度に及ばない。このことも BLEU スコアの判定に合致するものと理解できる。

日中テクスト翻訳のエラー原因としては、日本語音声の誤認識に起因するもの 14 箇所（表５のエラー16 箇所中 2 箇所は翻訳において自動修正されている）、明示化が必要なもの 11 箇所、不適な語彙選択 13 箇所、不適な係り受け 4 箇所等である。それに対し、中日テクスト翻訳のエラー原

因は､不適な語彙選択 7 箇所､不適な省略 5 箇所、中国語音声の誤認識に起因するもの 5 箇所（表 5 のエラー8 箇所中 3 箇所は翻訳において自動修正されている)、等である。

このテクスト翻訳エラーを通常の通訳プロセスで考察すると、音声翻訳（通訳）エラーは音声認識エラー、語用エラー（明示化要す、不要な重複、不適な付加、不適な省略)、意味エラー（語彙の誤選択)、構文エラー（不適な係り受け、不適な語順）に再分類することができる（表６)。

エラーの分類で興味深い点は、日中テクスト翻訳と中日翻訳テクストのエラーの比率の相違である。音声認識エラーおよび構文エラーは日中・中日で同じ程度の割合であるが、日中テクスト翻訳が語用エラーの割合が高く、また内容は「明示化が必要」に偏っている。翻訳の語用（対話レベル）的等価は文脈からの高度の推論を必要とすることから、日本語から中国語への対話の変換が難しいことを反映しているものと推測される。またこの特徴は、多くの日中通訳者の感覚とも符合していると考える。現在の AI 翻訳はまだ語用的推論機能を十分に組み込んでいないので、語用的推論能力は現時点で人間の通訳者の優位なポイントとなるものとも考えられる。

E. 結論

結論として、POCKETALK は医療専門用語・フレーズへの対応はほぼ申し分なくカバーしている。また音声認識においても高い精度である。しかしながら、テクスト翻訳については、自動評価でも人による評価においても対話レベルに十分に対応しているとは言えないとの結論になる。その理由の一つとしては、対話であっても、翻訳機は一つ一つの発話を単体として処理し、とりわけ省略された意味のつながりを認知できないのではないかと考えられる。その点は、通訳者であれば対話の流れから自然に感じ取ることができ、スムーズなコミュニケーションにつなげることができる。

ただし、エラー発生の原因は限られていて、現状の大規模コーパスがさらに整備されるならば、かなりの改善が可能であろう。またエラー発生は通訳者と共通している点があり、特に語用エラーの克服には経験を踏んだ通訳者が有するノウハウを活用できるのではないか。たとえば医師:「これがエイズです」に対し、POCKETALK 訳；“这是艾滋病”、通訳経験者訳：“这就是艾滋病” と言外のニュアンスを表現できる。この点は研究課題として今後再考したい。

本研究では検証しなかったが、POCKETALK のクリアな音声、レスポンスの速さ、文章の滑らかさには驚きを感じる。これらの優れた点はAI翻訳技術の賜物であり、それこそが商品化の前提であろう。それに対し、POCKETALK は小型軽量化され携帯に便利な点からみても、主に観光旅行上の場面での使用を想定していることは否めない。医療現場での使用には医療者側によるハンズ・フリーの工夫や一センテンスごとの発話など、かなりの制約を伴うと言わざるを得ない。医療現場に特化した大規模コーパスの整備及び使用上のインターフェイスの工夫が求められると考える。

参考文献

1) 総務省(2020)「グローバルコミュニケーション計画 2025」
https://www.soumu.go.jp/main_content/000678485.pdf

2) (株)ソースネクスト「POCKETALK 」
https://pocketalk.jp/

3) 北島勉、他(2017)『外国人に対する HIV 検査と医療サービスへのアクセス向上に関する研究』平成 28 年度総括・分担研究報告書（厚生労働省・科学研究費補助金エイズ対策研究事業）

4) 内山将夫(2008)「自動評価尺度 BLEU」
https://www2.nict.go.jp/astrecatt/member/mutiyama/corpmt/4.pdf

5) 西野竜太郎「シンプル MT スコア」
https://www.nishinos.com/simple-mt-score

6) GoogleCloud「モデルの評価」

https://cloud.google.com/translate/automl/docs/evaluate?hl=ja

7) 具体的内容は G．研究発表を参照されたい。

8) 同上

Ｆ．健康危険情報

なし

Ｇ．研究発表

張弘(宮首弘子)(2021)「音声翻訳機の医療通訳における有用性」『杏林大学外国語学部紀要』第 33 号

Ｈ．知的財産権の出願・登録状況

1. 特許取得

　なし

2. 実用新案登録

　なし

3. その他

　なし