MedSLT のユーザビリティの向上
--逆翻訳とヘルプ・システム--

中尾雪絵 [1,2]　Manny Rayner[3]　Nikos Chatzichrisafis[3]
神崎享子 [2]　Pierrette Bouillon[3]　Beth Ann Hockey[4]　井佐原均 [2]

1) ナント大学 2) 情報通信研究機構 3) University of Geneva 4) University of Santa Cruz

yukie-n@khn.nict.go.jp, mrayner@riacs.edu, Nikos.Chatzichrisafis@vozzup.com,
{kanzaki, isahara}@nict.go.jp,
pierrette.bouillon@issco.unige.ch,bahockey@email.arc.nasa.gov

1． はじめに

機械翻訳システムを用いた医療分野の翻訳においては、誤った情報の伝達を避けることが必要である。本稿が紹介するのは、医療診断用音声翻訳システム MedSLT が取り入れた逆翻訳とヘルプ・システムの二つの機能である。逆翻訳は、翻訳された文を中間言語を通して原言語に訳し戻すことにより、自分の発話がシステムに正しく理解されているか、ユーザが確認できる仕組みである。ヘルプ・システムは、認識可能な質問文をユーザ向けに表示し、ユーザがシステムに順応するのを助ける。これら二つの機能について詳述した後、現行での評価を提示し、両機能の意義や問題点を考察したい。

2． MedSLT の現在と課題

MedSLT (Bouillon et al 2005) の日本語版については、中尾　他 (2005)ですでに紹介されている。MedSLT が構想しているのは、例えば日本人の医師が日本語を理解しない患者に診察を行う際、簡単な質問を機械翻訳で補助するといったものである。現在、日本語版は英語・フランス語との相互の翻訳が可能で、スペイン語、カタロニア語、フィンランド語との翻訳も準備段階に入っている。

MedSLT のようなシステムを構築する際に重要となるのが信頼度であるということは、実際の医師の意見からも明らかである。誤訳は致命的な問題を起こすため、何としても避けなければならない。現場の声を踏まえた結果、我々は医師から患者への質問という限られた範囲に対象をしぼり、質問の内容も患者が「はい／いいえ」で答えられるものとし、誤訳を事前に防ぐ手段として、音声認識が正しく行われなかった場合、ユーザは認識をやり直すことができるようにした。音声認識は数百語程度に語彙を厳選した文法ベースのモデルを、翻訳は中間言語モデルを採用した。文法ベースの音声認識と中間言語ベースの翻訳は、オープンソースの Regulus プラットフォームで展開される (Rayner et al 2003, Rayner et al 2006)。特に、音声認識の言語モデルは、用例ベースの実行プロセスを使った高水準の feature grammar から得られる(Rayner et al 2003, Rayner et al 2006)。

このシステムの利点は、ユーザがシステムのカバーする文法の範囲内において、音声認識・翻訳ともに高い信頼度を持ちうることだろう (Bouillon et al 2005)。実際、我々が意見を聞いた医師達は、システムを介した会話が完全にコントロールされるという点で、原則的には信頼できると考えており、逆に翻訳の正確さという観点からは、統計的モデルの翻訳システムは、ブラックボックスが多すぎて、必ずしも信頼できるわけではないと考える。

とはいえ、システムを使った実験を行うと、課題もいくつか明らかになってきた。第一に「はい／いいえ」に限った質問は、答えを引き出すまでに手間がかかる。例えば「頭痛は二週間以上続いていますか」という質問に患者が「いいえ」と答えた場合、医師は「一週間？」「三日？」など患者が「はい」と答えるまで聞き続けなくてはならず、むしろ「どれくらい頭痛は続いていますか」といった、患者が具体的に答える質問が好ましい。第二に、ユーザにはシステムが自分の言ったとおり翻訳しているのかが不明瞭である。例えば、主語を省略した日本語文が、正しい主語を補って翻訳されているか、といった問題である。第三に、どのような文法がシステムに準備されているのか、ユーザには把握しにくい。例えば「鋭

い痛みですか」という文はシステムにカバーされているのに、「頭の痛みは鋭いですか」はカバーされていないというケースもある。

第一の問題に関して、我々は省略処理（ellipsis processing）を試みている段階である。本稿が扱うのは、第二の問題に対処した逆翻訳並びに第三の問題に対処したヘルプ・システムについてである。

3． 逆翻訳

ユーザが、翻訳ボタンを押す前に自分の発話を確認できれば、誤訳の可能性は自ずと減るが、問題は、認識結果から最終的なアウトプットの翻訳がどのようになるかを推測するのが簡単であるとは限らない、ということである。「逆翻訳」という考えは、こうした問題を改善しようとしたものであり（Frederking et al 1997）、MedSLT ではユーザの発話を中間言語を介して日本語に訳し戻す仕組みとなっている。

例えば、原言語を日本語とすると、日本語から中間言語への翻訳プロセスは、省略処理などを伴う混み入ったものだが、中間言語から目標言語への翻訳はより簡素である。この目標言語への翻訳を仲介する中間言語を、原言語の日本語にも訳し戻してユーザに提示すると、原言語から目標言語への翻訳過程における不確かさも解消される。

なお、中間言語からの逆翻訳は、中間言語から原言語への翻訳ルールを適用している（英語→日本語、仏語→日本語でも同ルールを採用）。

4． ヘルプ・システム

ヘルプ・システムは、二つの発想から成り立っている（Starlander et al 2005, Rayner et al 2005）。

一つは統計的音声認識の使用である。というのも、統計的音声認識を文法ベースのモデルを用いた認識と比べると、システムの文法がカバーする想定内の発話の場合は精度が劣るものの、文法をカバーしていない想定外の発話については、文法ベースの音声認識よりも、よい結果になる。

もう一つは、MedSLT が扱う質問文の範囲の狭さである。質問内容が絞り込まれているので、数百の発話から成るコーパスで、ユーザの質問文はほぼ網羅されているといってもよい。

先述のように、MedSLT は音声認識に文法ベースのモデルを使用している。従って、ヘルプ・システムを使用するとき、ユーザの発話は文法ベースと、統計の二つのモデルの認識システムで認識されることになる。統計的モデルは Nuance SayAnything ツールによる標準型 N グラム言語モデルで実装し、文法ベースの学習に用いたのと同じコーパスで学習させる。語のクラス特定には、クラスごとに二つから三つの例を記述した宣言セットを使う。続いて、プログラムを用いて語彙セットからすべての語が探し出され、クラスごとに類似する語が抽出される。

ヘルプ・システムは、文字化した日本語の例文を常にシステムの最新版に準拠して備蓄しており（ヘルプ・コーパス）、コーパス内の例文がいずれも正しい翻訳文を生成する文であることもあらかじめ確認されている。統計ベースの認識システムが作動すると、ヘルプ・システムは出力文をヘルプ・コーパス内のすべての日本語文と照合し、もっとも近い候補文の群をユーザに提示する。照合の際、ストップワードは取り除かれ、N グラムモデルが同クラスの語を抜き出す。例えば「コーヒー」と「チョコレート」は飲食物を指す名詞であることから同じクラスになっている。すなわち、「コーヒー」を含む発話からは「チョコレート」を含む認識候補文が抽出されるといった具合である。同じ N グラムを多く有するほど、発話に近い候補文と見なされ、N グラムの多いものから優先的にヘルプ画面に表示される（Starlander et al 2005）。

ヘルプ・システムは、文法ベースモデルの出力ではなく、N グラム言語モデルの出力を用いる。というのも、N グラムモデルのほうが、システムの想定外の文に対し、性能が良いからである。これにより、ヘルプ機能の頑健性は高いものとなっている。

英語版を用いた実験では、ヘルプ・システムを使う被験者の学習率は、使わない被験者の二倍となることが分かった（Starlander et al 2005, Rayner et al 2005）。この学習率は、実験の前々半と後々半の習熟度差を被験者一人ずつ算出したものである。

5． 評価

我々は、544 文の日本語発話文データによる日本語版 MedSLT の評価を行った。データは４人の日本人話者の

発話を集めたもので、翻訳先言語は英語である。評価は、1）文法ベースの認識システムと統計的認識システムの比較、2）逆翻訳の性能、3）ヘルプ・システムの性能の三点について行った。

〈文法ベースと統計的認識システムの比較〉

英語版 MedSLT の認識システムの評価実験では、文法ベースの認識はシステムの想定範囲の発話に関しては非常に良く、想定外の発話については悪かった。まず、我々は日本語版MedSLTでも同じ結果が出るのかどうかを確かめることにした。

544 の発話文のうち、324 文がシステムの想定範囲、220 文が想定外である。この両グループを上記二種のリコグナイザーで認識し、ワードエラー（WER）と文エラー（SER）の割合を算出した。意味エラー率（SemER）については、文法ベースの認識システムは意味論的な表示ができるのに対し、統計的認識システムでは不可能である。このため、両者の比較は人手により、認識結果の出力文を発話の書起し文と照らし合わせ、直感的に同義と取れるかどうかという基準で行った。そのうち、認識結果が書起し文と同じ意味ではないと判断された発話文を対象にSemERを算出した。結果は表１の通りである。

想定内での WER の差は、文法ベースの 3.7%に対し、統計的 4.8%とわずかだが（相対率 23%）、SER を見ると、その差は文法ベース 12.7%に対し、統計的 17.6%と広がりを見せ（同28%）、SemERになると文法ベース3.7%、統計的 9.9%（同 63%）と大きく違いが出る。一方、想定外の発話に関しては、英語版認識システムの実験結果と同様、N グラム準拠の統計的認識システムが、文法ベースよりやや高い性能を示した。

	文法ベース		統計的	
WER	想定内	想定外	想定内	想定外
	3.7%	41.1%	4.8%	32.6%
SER	12.7%	99.5%	17.6%	81.8%
SemER	3.7%	77.7%	9.9%	74.1%

表１　文法ベースと・統計的な認識システムの性能比較

(想定内の発話は 324 発話、想定外は 220 発話)

〈逆翻訳〉

逆翻訳システムの目標は、ユーザの発話を中間言語を用いて逆翻訳し、ユーザに提示することである。逆翻訳がうまくいかないときには、二通りの理由が考えられる。一つは、中間言語が正しくても、誤った逆翻訳を行ったり、逆翻訳がまったく生成されない場合、また中間言語に主語の省略が補われていないなどの問題があるにも関わらず、逆翻訳が生成されてしまう場合である。

逆翻訳の評価について、我々は試験コーパス内の発話を処理したときに、どれくらい誤った中間言語表現があるのかを調べてみた。人手の評価は時間がかかるうえ、見落としなどが起きやすいので、対象となる発話すべてを中間言語を介して目標言語（英語・仏語）に訳してから、日本語へ訳し戻した。また、日本語の認識結果の評価も行い、他の四つの評価結果と（英語訳、仏語訳、日本語の逆翻訳、認識文）と比べた。その結果、四つの評価がすべて「正しい」と認定された文は中間言語も正しく、四つすべてが「誤り」と認定されたものは中間言語も誤りであった。四つすべてが正しい、あるいは誤りとなる発話は、全体の 90%近くにのぼった。残り 10%について、我々著者のうち二名が中間言語を一つずつ検討した。その結果、中間言語が正しいときには逆翻訳も正しく行われていることが、ここでも裏付けられた。まとめると表２のようになる。

	逆翻訳の結果			
中間言語	**Good**	**OK**	**Bad**	**訳なし**
正しい	310	18	0	0
誤り・なし	2	0	58	156

表２　逆翻訳の評価の結果

中間言語が正しく生成されると、94.9%（328 文のうち 310 文）の逆翻訳がほぼ正しく作られ、残る 5.1%も許容範囲（「は」と「が」の違いなど）の訳文であった。中間言語の欠落や、誤った中間言語が含まれている場合（通常、認識誤りと判断）、妥当な逆翻訳はわずか 0.9%（216 文のうち 2 文）であった。

いずれの場合も問題の原因は、日本語の主語を中間言語で補完するような文法ルールが欠落していたことである。一方、中間言語から日本語という方向では、正しい

ルールが存在しないにも関わらず、訳文が生成されてしまうという現象も見られた。今後、中間言語をさらに細かくチェックし、この問題を解決していきたいと考えている。

〈ヘルプ・システム〉

ヘルプ・システムについては、日本人被験者三名による評価を行った。対象としてシステムの想定外の発話から 215 発話を取り出し、先述した N グラム言語モデルを用いてヘルプ候補文を作成した。被験者は、発話と、発話ごとに選ばれたヘルプ候補文を見ながら、候補文が「使える」か「使えない」かを判断していった。「使えない」という評価の大半は、医師の質問とは関係のない発話（215 発話のうち 41 発話）であった。こうした無関係な発話に対しては、ヘルプ・システムはいかなる候補文も提示し得ない。一方、システム対象内の発話（「言い差し」なども含む）については、被験者三人の平均にして 67.6%のヘルプ候補文が「使える」と評価された。なお、システムの「対象内」（in domain）の発話とは、MedSLT の想定内・外（in/out of coverage）に関わらず、医師の質問としてふさわしい発話を差す。結果は表３のようになった。

	システムの対象内の発話に対する「使える」評価の数
被験者１	122
被験者２	158
被験者３	100

表３　ヘルプ・システムの動作評価の結果

この結果は、ヘルプ・システムが活用するに値するものであることを示唆するものだが、今後さらに改良する必要がある。４節で説明したように、ヘルプ・システムの N グラムのクラスは、統計的言語モデルのクラスと合致するようになっている。言語モデルのクラスは、対象となる入力言語に適用できるよう調整されている。ヘルプ・システムの目的は、発話された文と関連のある候補文を提供することであり、今後システムの動作を良くするためには、ヘルプ・システムの N グラムのクラスを構成する方法を改良することも必要だろう。

６． 終わりに

文法ベースの音声認識と中間言語の翻訳は、正確さと信頼度の高さという点で、医療診断という限られた分野での音声翻訳システムとして期待の持てる組み合わせである。しかし、ユーザにとってはシステムがカバーする文を使いこなすのに時間がかかったり、発話と翻訳文との間にずれがないかという不確かさがあったりするのも事実である。本稿では、これらの問題を解決するべく導入した二つの機能、すなわち中間言語を原言語に言い換える逆翻訳、並びに頑健な認識に基づく知的ヘルプ・システムを紹介した。５節で見た評価により、両機能の日本語版 MedSLT における性能も裏付けることができたかと思う。システムの対象内と見なせる発話のうち、たとえ認識に失敗しても 67%はヘルプ候補文を抽出することができる。また、逆翻訳機能のフィードバックで、データの 99%において正確である。これは両機能を搭載によって、システムがユーザにとってさらに使いやすいものとなることを明示している。

参考文献

P. Bouillon et al, 2005. “A Generic Multi-Lingual Open Source Platform for Limited-Domain Medical Speech Translation”. Proceedings of EAMT 2005, Budapest, Hungary.

R. Frederking et al, 1997. “Interactive Speech Translation in the DIPLOMAT Project”. Proceedings of the Spoken Language Translation Workshop at the 35th ACL Conference.

中尾　他, 2005 年. “医療診断用音声翻訳システム MedSLT における日本語規則の作成”, 第 11 回日本情報処理学会.

M. Rayner et al, 2006年刊行予定. *Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler.* CSLI Press.

M. Rayner et al, 2003. “An Open Source Environment for Compiling Typed Unification Grammars into Speech Recognisers”. Proceedings of EACL 2003 (interactive poster and demo track).

M. Rayner et al, 2005. “A Methodology for Comparing Grammar-Based and Robust Approaches to Speech Understanding” Proceedings of INTERSPEECH 2005.

M. Starlander et al, 2005. “Practising Controlled Language through a Help System integrated into the Medical Speech Translation System (MedSLT)”. Proceedings of MT Summit X.