Vol.2012-NL-206 No.14
Vol.2012-SLP-91 No.14
2012/5/11

音声情報案内システムのための統計的機械翻訳を利用した質問応答

西村一馬[†1]　川波弘道[†1]
猿渡洋[†1]　鹿野清宏[†1]

音声情報案内システム「たけまるくん」では質問文例と対応する応答文のペアで構成される質問応答データベース (QADB) から単語マッチングにより入力質問文に最も近い質問文例を選択し，それに対応する応答文を出力する応答文生成方式を採っている．この手法によりユーザ発話の意味理解など高度な解析をする必要なく応答文を出力できる．しかし，この手法では応答文は定型文しか出力できない．ユーザとのインタラクションを改善するための方法として，本研究では統計的機械翻訳による応答文生成を提案する．提案手法は質問文と応答文を別の言語とみなした翻訳を行うことで質問文を応答文に変換しようとするものである．書き起こし文による実験では適切な応答の生成率が約 60%であったが，10-Best および 50-Best の認識仮説を学習と入力を用いることで，適切な応答の生成率が約 10 ポイント向上した．生成された応答文の分析から，音声認識誤りの傾向を含有する翻訳モデルと翻訳ス コアによる応答文候補の選択が効果的に機能していることが明らかになった．

Response Generation Using Statistical Machine Translation in a Speech-Oriented Guidance System

Kazuma Nishimura,[†1] Hiromichi Kawanami,[†1]
Hiroshi Saruwatari[†1] and Kiyohiro Shikano[†1]

Takemaru-kun, a speech-oriented guidance system, employs a response generation using a question and answer database (QADB). This method searches a QADB for the example question most similar to a user utterance, and outputs the answer tagged to it. With this method, a system can get response sentences without high-level semantic analysis of user utterances. However, a system can output only sentences fixed beforehand. In this paper, a response generation using a statistical machine translation method is proposed. In the experiment using utterance transcriptions, about 60% of appropriate response sentences were generated. By learning and inputting multiple speech recognition hypotheses (10-Best and 50-Best), the rate of appropriate responses gained about 10 percentage points. The effectiveness of the proposed method was suggested though further improvement is necessary for actual operation.

1. はじめに

音声対話システムは人と機械が音声による対話をしながら情報案内など何らかのタスクを達成するシステムであり，幅広く需要が見込まれる．筆者らは**図 1** に示す音声情報案内システム「たけまるくん」を開発し，2002 年よりコミュニティセンターで運用を継続している[1)]．このシステムは用例ベース方式を採用しており，で質問応答データベース (QADB) を用いて質問応答を行う．入力質問文と最も類似した質問例をデータベース中から選択し，その質問例に対応付けられた応答文を出力するというシンプルな構造のため，応答内容の拡張が容易であり，一問一答式の情報提供を頑健に行うことができる．

ユーザにとってより親しみやすいシステムとするために，ユーザ発話の言い回しや表現に応じて応答文を柔軟に生成することが期待されているが，現在の方式では応答文は事前に設定された定型の文しか返すことができない．この要求を満たす一つのアプローチとして，質問例の表現や言い回しに対応させて応答文の表現を多様化することが考えられる．しかしながらそのような QADB の構築には人手がかかり，コストが高いものとなってしまう．

そこで，筆者らは統計的機械翻訳の手法を応用した応答文生成を提案している[2),3)]．この手法においては，質問文と応答文とを別の言語して扱い，システムへの質問発話を適切な応答文に「翻訳」すると考える．これまで，ユーザ発話の書き起こしを用いた予備実験[2)]，音声認識結果を用いた翻訳モデルによる音声入力からの応答性能の改善を報告した[3)]．本報告では 10-Best および 50-Best の音声認識候補を用いて翻訳モデルの学習，応答生成手法の詳細を述べ，生成された応答文の情報伝達性，自然性の両観点から実験結果の分析を行う．

2. 音声情報案内システム「たけまるくん」

2.1 システムと音声データベース

音声情報案内システム「たけまるくん」は生駒市の北コミュニティセンターに設置され，

†1 奈良先端科学技術大学院大学
Nara Institute of Science and Technology

Vol.2012-NL-206 No.14
Vol.2012-SLP-91 No.14
2012/5/11

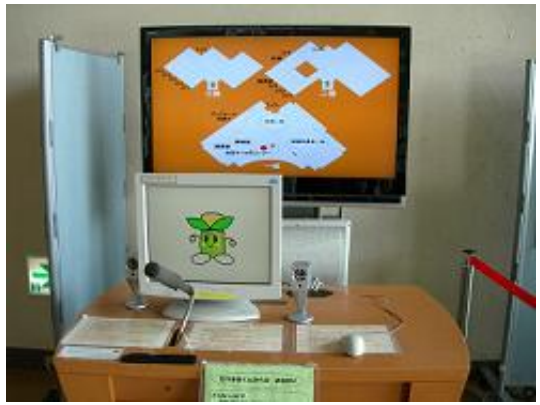

図 1 音声情報案内システム「たけまるくん」
Fig. 1 Speech-oriented information guidance system *Takemaru-kun*.

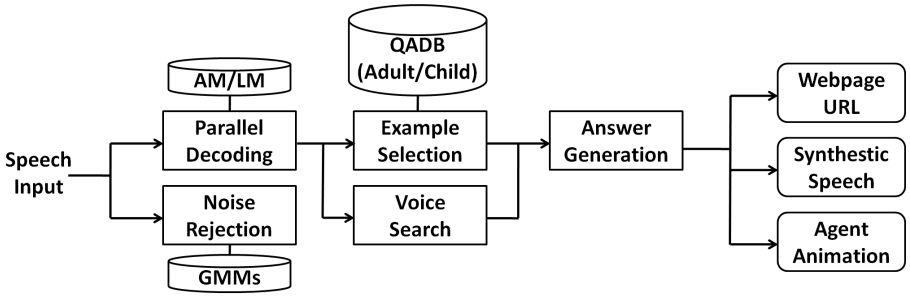

図 2 「たけまるくん」のシステム構成図
Fig. 2 Processing flow of *Takemaru-kun*.

2002 年 11 月から運用を開始し，現在まで約 10 年間に運用を継続している．「たけまるくん」の応答内容は，センターの施設案内，センターのサービス案内，周辺の観光案内などの他，ニュースや天気予報，日時などの一般的な情報である．また，情報案内を期待しない挨拶や，「たけまるくん」自身のプロフィールに関する雑談にも答えるなど，幅広いタスクに対応している．

システムの構成図を**図 2** に示す．システムは認識処理と平行して雑音と音声の 5 つの GMM の尤度計算を行い，雑音を応答処理にかけることなく棄却する．音声に対しては大

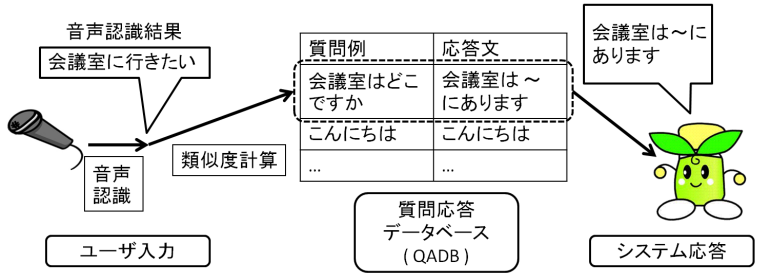

図 3 従来の「たけまるくん」の応答文生成
Fig. 3 Answer generation of *Takemaru-kun*.

人話者のモデル，子供話者のモデルをそれぞれ用いた並列デコーディングを継続し，モデルの尤度に基づい大人/子供を判別する．音声認識結果を用いて，大人・子供別に用意した QADB を参照し，形態素マッチング数に基づくスコアに基づいて質問用例を探索，応答文を生成する．システム出力にはウェブブラウザ，合成音声，エージェントアニメを用いる．ユーザとシステムは基本的に一問一答で対話を行う．

「たけまるくん」は長期にわたって運用されているが，雑音のみの入力も含めすべてのシステム入力音は記録されており，そのうち最初の 2 年 5 か月間の全システム入力については人手により書き起こし，雑音タグ，年齢層・性別ラベル，正しい応答ラベルが付与され，大規模音声データベースとして整備されている．

2.2 用例ベース応答文生成

「たけまるくん」は質問例と応答文のペア（QA ペア）を集めた質問応答データベース (QADB) をシステム内に保持している．ユーザ発話が音声認識されると，その認識結果 I の N-best を用いた QADB 内の質問例 E との類似度計算が行われ，類似度 $s(I,E)$ が一番大きい質問例が QADB 中から最近傍法によって選択される[?]．

$$s(I,E) = w(I,E)/\max(g_I, g_E) \tag{1}$$
$$w(I,E) = \sum_{k \in I \cup E} \min(w_I(k), w_E(k))$$

ここで，$w_I(k), w_E(k)$ はそれぞれ単語 k の文 I, E の一文あたりの平均出現数であり，g_I, g_E はそれぞれ I, E の一文あたりの単語数である．

QADB による応答生成は，コンテンツの拡張，応答内容の制御が比較的容易なフレーム

Vol.2012-NL-206 No.14
Vol.2012-SLP-91 No.14
2012/5/11

ワークであるが，質問文に応じた多様な応答文を整備するには，人手による開発コストが問題となる．提案する統計的機械翻訳による応答生成はそれを解消することを目指すものである．

3. 統計的機械翻訳を用いた応答文生成

3.1 統計的機械翻訳の概要

統計的機械翻訳は対訳コーパスを分析して翻訳規則や対訳辞書にあたる統計モデルを自動学習し，ある言語の文を異なる言語の文へ変換する技術である．**図 4** に一般的な統計的機械翻訳の構成を示す．

原言語の文 $\mathbf{f}$ を目的言語の文 $\mathbf{e}$ に翻訳したいとする．このとき，翻訳結果 $\mathbf{e}$ の候補は無数に存在する．翻訳器は全てのペア $(\mathbf{e}, \mathbf{f})$ に対して $\mathbf{f}$ が $\mathbf{e}$ に翻訳される確率 $P(\mathbf{e}|\mathbf{f})$ を計算し，$P(\mathbf{e}|\mathbf{f})$ を最大化する $\hat{\mathbf{e}}$ を探索する．この問題は対数線形モデルにより，次のように表される．

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} P(\mathbf{e}|\mathbf{f}) = \arg\max_{\mathbf{e}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}, \mathbf{f}) \tag{2}$$

ここで，$h_m(\mathbf{e}, \mathbf{f})$ は素性関数であり，M は用いる素性の数である．それぞれの素性の重みを λ_m で表す．用いる素性には翻訳モデル，言語モデルなどがある．翻訳モデルは翻訳の可能性を表し，言語モデルはその文の言語としての流暢さを示す．言語モデルは目的言語のコーパスから学習される．元々は単語アライメントを学習して作成される IBM 翻訳モデル[4] が翻訳モデルとして使用されていたが，後に，次に示すフレーズベースの翻訳モデルが提案された[5]．フレーズベース翻訳モデルにおいては，単語ではなく句がアライメントの単位として用いられる．ここで,「句」とは任意の単語列を指し，名詞句や動詞句といった言語学的なまとまりを指すものではない．

フレーズベース翻訳モデルでは，翻訳モデルは以下のように定式化される．

$$P(\mathbf{f}|\mathbf{e}) = \prod_{i=1}^{I} \phi(\overline{f}_i|\overline{e}_i) d(a_i - b_{i-1}) \tag{3}$$

まず原言語文 $\mathbf{f}$ を I 個の句 $\overline{f}_1 \overline{f}_2 \ldots \overline{f}_I$ に分割し，$\mathbf{f}$ 中のそれぞれの句 $\overline{f}_i$ を目的言語の句 $\overline{e}_i$ に翻訳する．そして句 $\overline{e}_i$ の順序を入れ替える．$\phi(\overline{f}_i|\overline{e}_i)$ は句翻訳確率であり，$d(a_i - b_{i-1})$ は相対的な句歪み確率である．a_i は目的言語の i 番目の句に訳される原言語句の開始位置である．b_{i-1} は目的言語の $(i-1)$ 番目の句に訳される原言語句の終端位置である．

句歪み確率は句（もしくは単語）の翻訳前後の位置の違いで与えられるペナルティである．句翻訳確率は以下のような相対頻度で与えられる．

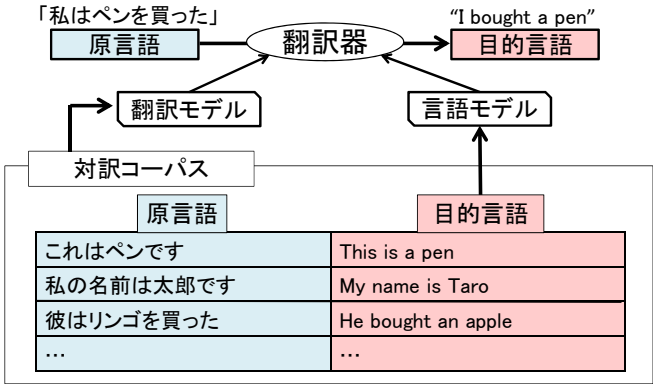

図 4 統計的機械翻訳の構成
Fig. 4 Flow of statistical machine translation.

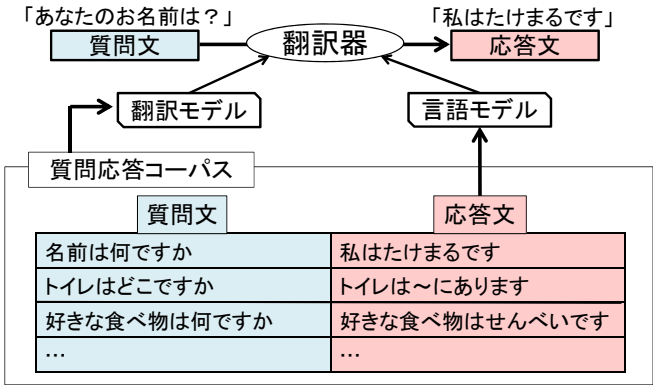

図 5 統計的機械翻訳を用いた質問応答の構成
Fig. 5 Flow of responce generation using statistical machine translation.

Vol.2012-NL-206 No.14
Vol.2012-SLP-91 No.14
2012/5/11

$$\phi(\overline{f}|\overline{e}) = \frac{\mathrm{count}(\overline{f},\overline{e})}{\sum_{\overline{f}'}\mathrm{count}(\overline{f}',\overline{e})} \quad (4)$$

フレーズベースの統計的機械翻訳ツールキットとして Moses[⋆1]がある．Moses では対訳コーパスから IBM 翻訳モデルの単語アライメントを基にしたヒューリスティックを用いてフレーズ抽出を行う．

3.2 N-Best 認識結果を用いた統計的機械翻訳による応答文生成

統計的機械翻訳はある言語の文を他言語の文へ翻訳する技術であるが，質問文と応答文を別言語とみなすことで，質問文から応答文に「翻訳」する．**図 5** が統計的機械翻訳の手法を用いた応答文生成手法の構成となる．言語間翻訳においては，翻訳モデルは例えば日本語と英語のように互いに異なる言語の対訳コーパスから学習されるが，応答文生成においては，翻訳モデルは質問応答ペアの集合から学習される．

提案手法の有効性を検証するため，すでに書き起こし文を質問例とした評価実験を行っている[2)]．実際のシステム運用ではユーザ発話の音声認識結果が入力となる．これらは書き起こし文と異なり認識誤りを含む．そのため書き起こし文の質問例から作成した翻訳モデルに対して音声認識結果を入力とすると，認識誤りが応答文生成の性能低下を起こす．そこで，音声認識結果を用いて翻訳モデルを作成する．翻訳モデル学習データ，入力データとしてそれぞれ N-Best 認識結果を用いると，応答性能が改善することが示唆されている[3)]．

翻訳モデルの学習フェーズにおいては，**図 6** に示すように N-best の認識仮説それぞれに対して応答文を複製してペアを作り，それを学習データセットに用いる．これにより認識誤りの多様性を含む大量の学習データが獲得できる．

応答文生成のフェーズにおいては，**図 7** に示すように N–best の認識仮説それぞれを翻訳器にかけて応答文の候補を生成する．生成された応答文候補の中から最も翻訳スコアの高いものを最終的な出力とする．これにより，より優れた応答文が選ばれると期待できる．

4. 実　　験

4.1 実験条件

翻訳モデルの学習データに用いる認識仮説数と，応答文生成の際の入力に用いる認識仮説数によって，数種類の組み合わせで実験を行った．音声認識エンジンは Julius4.2 である．

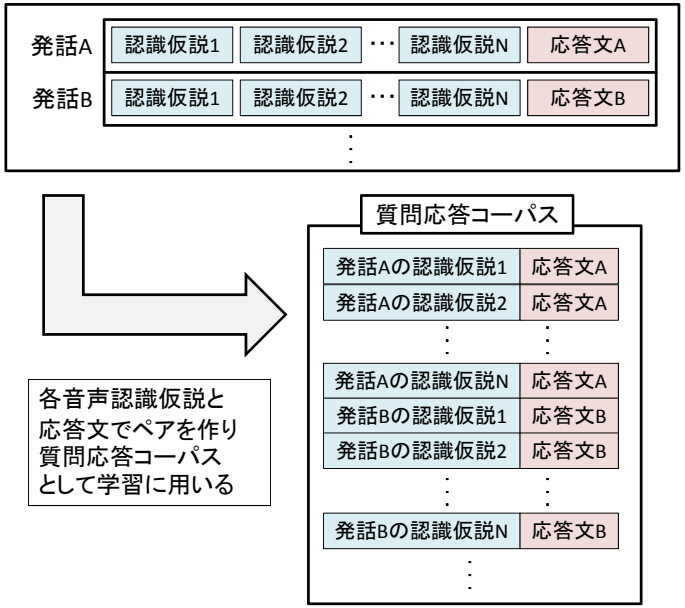

図 6 複数の認識仮説を用いた学習データセット
Fig. 6 Training data using N-Best ASR candidates.

実験には「たけまるくん」で収集された大人ユーザによる有効発話のデータセットを用いた．それぞれの発話には 276 種類の応答文のうち一つが付与されている．2002 年 11 月から 2004 年 10 月の 2 年間に収集されたデータから，2003 年 7 月と 8 月の 2ヶ月を除いて作成した質問応答ペアを学習データとし，翻訳ツール Moses[⋆2]により翻訳モデルを作成した．質問応答ペアのうち，応答文から SRILM により 3-gram の言語モデルを作成した．

2003 年 7 月のデータ 872 件は開発データとして式 (2) における素性重みのチューニングのための開発データとして用いた．開発データは音声認識仮説 1-best と応答文から作成した質問応答ペアである．テストデータとして，2003 年 8 月に収集された有効発話の中から，意図が不明瞭である発話を除外したあとの 959 件とその応答文のペアを用いた．

⋆1 http://www.statmt.org/moses/

⋆2 http://www.statmt.org/moses/

Vol.2012-NL-206 No.14
Vol.2012-SLP-91 No.14
2012/5/11

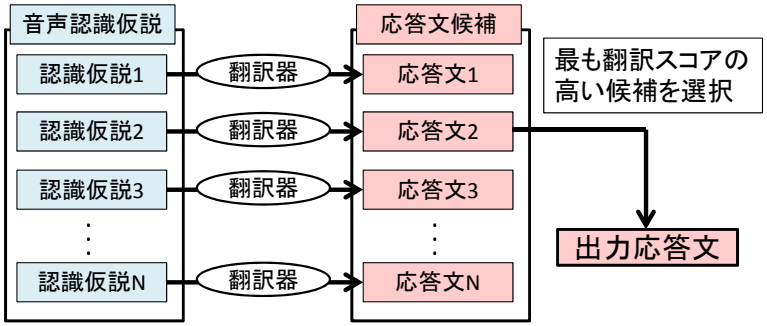

図 7 複数の認識仮説を入力とする応答文生成
Fig. 7 Response generation using N-Best ASR candidates.

学習データの認識仮説数，入力データの認識仮説数として 1-best，10-best，50-best の 3 種類のデータを作成した．

4.2 評価尺度

一般的に統計的機械翻訳の評価尺度として，BLEU スコアが用いられている．BLEU スコアは翻訳結果文と正解参照文を比較して単語 n-gram の一致度を測るものである．BLEU スコアは主観による翻訳結果の評価と相関があることが知られている[6]．よって計算機による高速な評価を可能とする BLEU スコアは統計的機械翻訳の研究分野では広く用いられている．

ただし本研究においては，応答文としての適切さと BLEU スコアとが相関があるかどうかが不明であるため，主観による評価を行った．評価基準は以下の通りである．

- 応答として必要な情報を含むこと (情報伝達性)
- 日本語の文として自然であること (自然性)

この 2 点を両方満たすものを「適切」な応答文とする．今回の評価は 1 名の評価者によって行った．なお，BLEU スコアについても参考のため算出を行った．

4.3 実験結果

図 8 に提案手法による適切な応答の生成率を示す．横軸は翻訳モデルの構築に用いた認識仮説数を表している．

書き起こし文を学習データ，評価データとして実験を行ったところ，60.0 % が適切な応答文であった．音声認識結果 1-Bes のみを用いて翻訳モデルを学習した場合，この値に及ばないが 10-Best, 50-Best で学習した場合は書き起こしによる性能を越える結果となった．学習データの増加と認識誤りの傾向が学習モデルに適切に反映された結果と考えられる．

入力データについて観察すると，10-best，50-best ともに 1-best を大きく上回った．特に翻訳モデル 10-best 及び 50-best を用いた場合には，1-best 入力時と比べて約 10 ポイント性能が上昇しており，翻訳スコアに基づく生成応答文の選択が有効に機能していることが分かる．

図 9 は適切さが満たすべき 2 つの評価尺度とした，「情報伝達性」と「自然性」を個別に評価した結果である．上の図が情報伝達性を，下の図が自然性をそれぞれ実現した文章の生

表 1 実験データの諸元
Table 1 Details of experimental data

学習データ	収集期間	2002 年 11 月-2004 年 10 月 (2003 年 7，8 月を除く)
	データ数	18509 件 (1-best) 184983 件 (10-best) 912289 件 (50-best)
開発データ	収集期間	2003 年 7 月
	データ数	872 件
テストデータ	収集期間	2003 年 8 月
	データ数	959 件
単語正解精度		86.88%

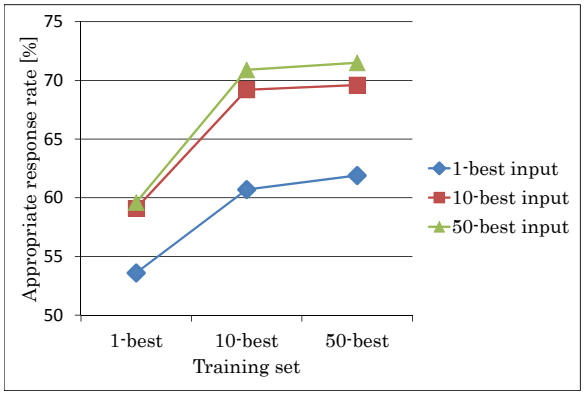

図 8 適切な応答の生成率
Fig. 8 Rate of appropriate responces.

Vol.2012-NL-206 No.14
Vol.2012-SLP-91 No.14
2012/5/11

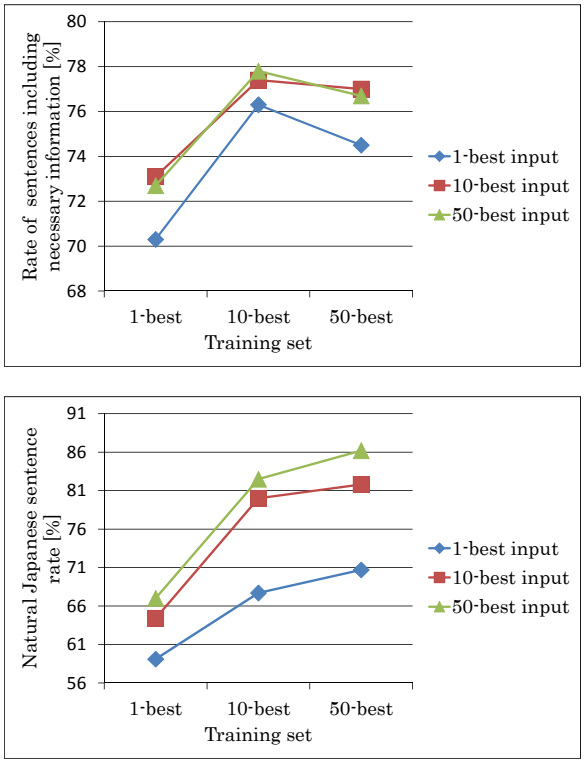

図 9 「適切さ」の条件の個別での評価：情報伝達性 (上) と自然性 (下) の実現率
Fig. 9 Rates of informative (above) and natural (below) responses.

成率である．

必要な情報を含む文の生成率は 10-best の翻訳モデルを用いた場合に最大となった．また，入力の認識仮説数を 1-best から 10-best に増やした際に性能向上が見られたが，10-best から 50-best に増やした際には性能向上は見られなかった．

自然な文の生成率はモデル学習データの認識仮説数を増やせば増やすほど，また入力の認識仮説数を増やすほど性能が上昇するという結果になった．

なお，50-best の翻訳モデルを用いて応答文生成をした際の BLEU スコアは，1-best 入力で 0.590，10-best 入力で 0.633，50-best 入力で 0.639 であり，入力の認識仮説数とともに上昇が見られる．

5. 考　　察

5.1 モデル学習における N-Best の有効性

学習データとして 10-Best, 50-Best を用いたとき，書き起こし文による性能を凌駕した．音声認識誤りの傾向を含んだデータの利用，学習データ数増加が双方の効果が考えられる．

ただし，図 9 に見られたように，情報伝達性に関して最適な仮説数が存在する可能性がある．自然性は，今回の実験の範囲では仮説数の増加とともに改善されている．ただし，さらに実験を拡張することで，最適な仮説数の存在や情報伝達性とのトレードオフが観察される可能性もある．

5.2 入力における N-Best の有効性

認識結果の N-Best を入力とし，翻訳スコアによる選択を行うことが応答性能向上に有効であった．

1-best 入力では不適切な応答が生成されたものが，N-best 入力では適切な応答となった例を挙げる．学習，入力ともに 10-Best を用いた実験において，認識仮説第 4 位から生成された応答文が採用された例を示す．

・認識仮説第 1 位
　入力：えっと天気予報教えて下さい
　出力：和室天気予報のホームページにアクセスします
・認識仮説第 4 位
　入力：天気予報教えて下さい
　出力：天気予報のホームページにアクセスします

フレーズ対応を**図 10** に示す．認識仮説 1 位ではフィラー「えっと」に「和室」という無関係な単語が対応づけられているが，認識仮説 4 位ではフィラーが存在せず，適切な応答文が生成されている．

また，同一実験環境で，次のような例も観察された．

・認識仮説第 1 位
　入力：トイレはどこですか
　出力：はの奥かはばたきホール入り口の近くにございます

Vol.2012-NL-206 No.14
Vol.2012-SLP-91 No.14
2012/5/11

・認識仮説第 9 位

入力：トイレはどこですかあ

出力：トイレは左の奥かはばたきホール入り口の近くにございます

この例では認識仮説 9 位の語尾は不要なものに思えるが,「ですかあ」が「トイレは左」に変換されたことで，生成文の言語モデル尤度は高くなり，全体としてもスコアが高くなり，選択されたと考えられる.「ですかあ」-「トイレは左」のようなフレーズ対は学習データに複数音声認識仮説を用いた効果と考えられる．実際に，1-Best 学習で作成した翻訳モデルでは，認識仮説第 9 位からは適切な応答文は生成されなかった.

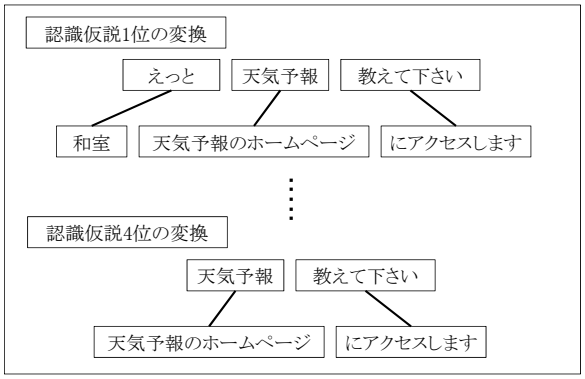

図 10 フィラーのない下位の認識仮説から生成された文が選ばれた例
Fig. 10 Example of generated correct response from lower ASR candidate.

また，入力を 10-best から 50-best としても適切な応答の生成率に大きな改善が見られなかった．**図 12** に 50-Best 入力の実験 (10-Best 学習) において，採用された認識仮説の順位の割合を示す．最終的な応答文のうち，約 75 % は 10 位までの認識仮説から生成されており，入力仮説数の拡大による応答性能の上昇にも上限がある可能性が示唆されている.

5.3 今後の課題

学習，入力について様々な仮説数を用いて実験を行い，最適な仮説数を検討する.

フレーズベースの翻訳モデルを導入したが，単語共起モデル等応答生成により適したモデルの調査を行う.

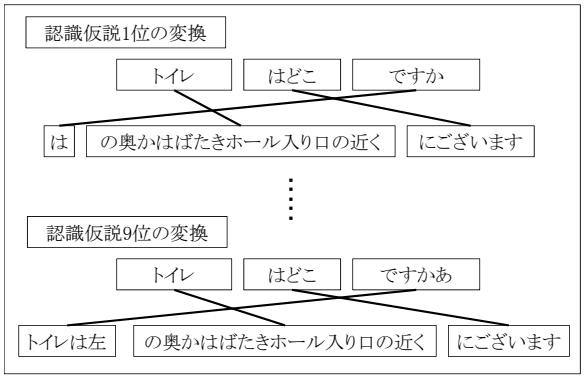

図 11 ノイズ単語が混入した下位の認識仮説から生成された文が選ばれた例
Fig. 11 Example of generated correct response from lower ASR candidate with noisy word.

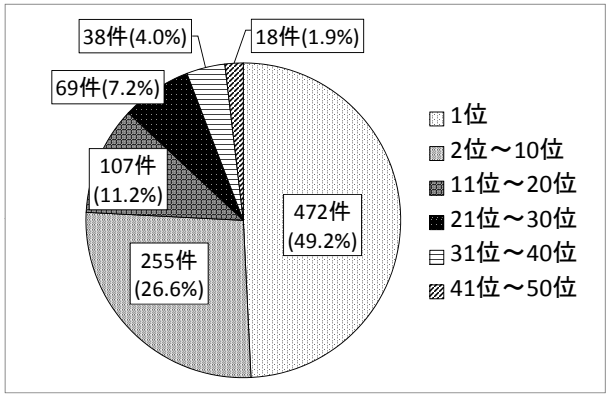

図 12 採用された認識仮説の順位の割合
Fig. 12 Rates of selected candidates' ranking.

また，本提案手法を従来の QADB 方式の応答文生成の代替となる手法と位置付けたが，代替手法としてではなく，QADB 方式で適切に応答できない部分を補う方法として捉え，両方の手法を併用して応答性能の向上を図るアプローチも検討する.

Vol.2012-NL-206 No.14
Vol.2012-SLP-91 No.14
2012/5/11

6. 結　　論

統計的機械翻訳の手法を応用した応答文生成手法を提案した．10-Best または 50-Best の音声認識結果を翻訳モデルの学習データとし，同様に入力データについても 10-Best または 50-Best の認識仮説を用いることで，書き起こしを用いた場合よりも 10 ポイントの性能向上が見られた．誤った情報を含んだ応答文，自然性にかける応答文の生成された原因の分析結果からも，音声認識誤りの傾向に対応した翻訳モデル，翻訳スコアを用いた生成文選択が効果的に機能していることが示唆された．

謝辞　本研究の一部は，科学技術振興事業団・戦略的基礎研究推進事業 (CREST) の支援を受けて実施された．

参 考 文 献

1) Ryuichi Nishimura *et al.*, Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1, 433–436, 2004.
2) Kazuma Nishimura *et al.*, Investigation of Statistical Machine Translation Applied to Answer Generation for a Speech-Oriented Guidance System, Proceedings of APSIPA Annual Summit and Conference 2011, 2011.
3) 西村一馬 *et al.*, 音声認識結果を用いた統計的機械翻訳による音声情報案内システム応答文の分析, 日本音響学会春季講演論文集. 3-7-13, 113-116, 2012.
4) Peter F. Brown *et al.*, The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, 19(2), 263–311, 1993.
5) Philipp Koehn *et al.*, Statistical Phrase-Based Translation, Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 1, 48–54, 2003.
6) Kishore Papineni *et al.*, BLEU: a method for automatic evaluation of machine translation, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 311–318, 2002.