音声認識結果と大規模コーパスに基づくユーザ意図に近い言語表現の検索

Retrieval of Expressions Similar to User Intention Based on Speech Recognition Results and a Large Corpus

竹澤 寿幸　　大熊 英男　　葦苅　豊　　清水　徹

Toshiyuki TAKEZAWA, Hideo OKUMA, Yutaka ASHIKARI, and Tohru SHIMIZU

独立行政法人 情報通信研究機構 知識創成コミュニケーション研究センター / ATR **音声言語コミュニケーション研究所**

National Institute of Information and Communications Technology / ATR

1　まえがき

音声言語コミュニケーションシステムで重要なのは，利用者のメッセージを相手に伝えることである[1]。異なる言語を話す人同士のコミュニケーションを支援する音声対話翻訳においてメッセージを相手に伝えるために必須となる機能は音声認識と機械翻訳である。音声認識，機械翻訳ともに近年の性能向上は著しいが，多数データを用いた平均値的な性能が向上したに過ぎず，システム利用者が自らのメッセージを込めた個々の発話の翻訳品質を保証するものではない。

音声認識については，結果をテキストとして画面に表示してフィードバックすればシステム利用者は自らのメッセージを機械がどの程度聞き取ったのか確認することができる。一方，機械翻訳については，結果をテキストとして画面に表示したところでシステム利用者は自らのメッセージがどのように相手に送られるのか判断することはできない。翻訳結果を逆方向にもう一度翻訳してその結果を提示することでシステム利用者に確認させる手法はあり得るが，多数データを用いた平均値的な傾向では役立つ可能性はあるものの，文や発話のような小さい単位で性能を保証するのは難しい[2]。

音声対話翻訳を音声言語コミュニケーションシステムとして実用化するためには，システム利用者が自らのメッセージを込めた発話の音声翻訳結果の性能を保証する技術が必要である。メッセージの言語表現に含まれるユーザ意図に近い表現がシステム開発用大規模コーパスにあれば，それを検索して提示し，システム利用者に選択させることはできるであろう。検索結果が選択された場合には，対訳そのもの，あるいは，対訳の訳語置換程度の範囲で翻訳することにすれば，その発話の音声翻訳結果の性能を保証することができると期待できる。

そこで，本稿では，大規模コーパスを用いて音声認識結果に類似した言語表現を検索する手法を検討する。対訳そのもの以外に訳語置換程度まで妥当な品質の音声翻訳結果が得られるものと想定し，音声認識結果の品詞を汎化する方針とする。システム利用者に即座に検索結果を提示するための高速処理手法が必要であるため，文献[3]で対訳用例検索に用いている編集距離と tf/idf による意味的な距離を併用する手法を利用する。そして，旅行会話文の読み上げ音声と実対話音声の複数のテストセットを用いた評価実験を行い，その効果を示す。

2　類似用例検索手法

2.1 音声認識結果の品詞の汎化

アナロジに基づく翻訳の考え方[4]に基づき，対訳そのものはもちろん対訳の一部を置換したものについても，文あるいは発話のような小さい単位で翻訳品質は十分であると仮定する。アナロジに基づく翻訳の一つの実現法である D3 [5]では，類似対訳用例の検索に意味辞書と編集距離を用いている。意味辞書は有用であるが，メンテナンスコストが必要であったり，研究限定などの制限があったりするため，品詞あるいは品詞の細分化情報を用いて汎化する方針とした。例えば，「東京」と「京都」は品詞の細部化情報である「地名」とすれば類似表現として検索できることになる。ここで，翻訳の品質を保証するために，日英方向であれば，日英の対訳で汎化された数に差があるものは置換利用するのに適切でない用例として検索対象から除く。もし過剰生成されてしまって意味的に不適切な連接となったとしても用例はシステム利用者に提示するのに使うため，利用者が意味的に不適当なものを選択することはないと仮定する。どの品詞あるいは品詞の細分化が良いかについては実験的に検証する。

2.2 対訳用例の検索

大規模コーパスに対して高速に対訳用例を検索する手法として文献[3]の手法を利用する。この手法は tf/idf による用例の予備選択と，編集距離と tf/idf を併用するスコア付けによる選択の 2 段階となっている。

予備選択は，日本語入力J_0に対して，次式でなされる。

$$P_{tf/idf}(J_k, J_0) = \sum_{i:J_{0,i} \in J_k} \frac{\log(N / df(J_{0,i})) / \log N}{|J_0|}$$

ここで，$J_{0,i}$はJ_0のi番目の形態素，$df(J_{0,i})$は形態素$J_{0,i}$の文書頻度(document frequency)，Nは対訳コーパスの用例数である。J_kに対象形態素があればその語句頻度(term frequency)は 1, なければ 0 とする。このスコアは入力長で正規化している。このスコアを用いて上位N_r ($\leq N$)個を選ぶ。

次に，予備選択された候補J_kに対してJ_0との編集距離$dis(J_k, J_0)$を求める。

$$dis(J_k, J_0) = I(J_k, J_0) + D(J_k, J_0) + S(J_k, J_0)$$

ここで，$k \le N_r$, $I(J_k, J_0)$, $D(J_k, J_0)$, $S(J_k, J_0)$はそれぞれ挿入，脱落，置換誤りの数である。そして，最終的に次式でスコア付けを行う。

$$score = \begin{cases} (1.0-\alpha)\left(1.0 - \dfrac{dis(J_k, J_0)}{|J_k|+|J_0|}\right) + \alpha P_{tf/idf}(J_k, J_0) & dis(J_k, J_0) > 0 \\ 1.0 & otherwise \end{cases}$$

$dis(J_k, J_0)$を正規化するにあたり，文献[3]では入力長$|J_0|$に対して行っているが，負にならないように$|J_k+J_0|$に修正した。重みαの値は実験により定める。

3 評価実験

3.1 音声認識およびテストセット

日本語音声認識は ATRASR [6]を用いた。音響モデルは MDL-SSS [7], 言語モデルはマルチクラス複合バイグラム[8]を用いた。

テストセットは表 1 に示すような 3 種類のものを用いた。BTEC (Basic Travel Expression Corpus) [9, 10]は旅行会話基本表現コーパスの読み上げ音声である。MAD (Machine-Aided Dialogs) [11]は音声翻訳システムを介して日本語話者と英語話者が実施した課題遂行型対話であり，音声認識システムの代わりにタイピストが発話を書き起こし，機械翻訳システムに入力する形態で集めたものである。FED (Field Experiment Data) [12]は関西国際空港で実施したモニタ実験から選んだデータである。平均発話長は BTEC と FED が同程度で，MAD が長い。パープレキシティは BTEC, MAD, FED の順に大きくなっている。

表 1　テストセット

	BTEC	MAD	FED
話者数	20	12	6
発話数	510	502	155
形態素数	4,035	5,682	1,108
平均発話長	7.91	11.32	7.15
パープレキシティ	18.9	23.2	36.2

利用者に使ってもらう状況を前提としているため，話し終えたらすぐに結果が表示されるような設定とした。具体的には，処理時間をリアルタイムファクタ(RTF)で表現したときに，RTF=1 とした。日本語音声認識結果の情報を表 2 に示す。

表 2　日本語音声認識結果

	BTEC	MAD	FED
単語認識率	94.8%	91.4%	89.4%
発話正解率	75.7%	53.8%	65.8%

3.2 実験条件および結果

品詞の汎化については，次の二つの条件を設定し，実験を試みた。条件 A, 条件 B と名付ける。

- 条件 A: 「人名」「地名」のみ汎化
- 条件 B: 「普通名詞」「形式名詞」「サ変名詞」「形容名詞」「サ変形容名詞」「形容詞」「形容動詞」「数詞」「副詞」を汎化

品詞を汎化してマッチングする手順は次のとおりである。

(1) 汎化する対象の表層表現を削除する。別途その情報を残す。
(2) 対訳用例の検索を行う。
(3) 対訳ペアで汎化する対象の数が異なる場合は候補から削除する。
(4) 汎化されたものを含む候補に対して，元の表層表現を追加する。

大規模コーパスとしては，旅行会話基本表現コーパス BTECを用いた。表 3 にその概要を示す。収集の時期とどの言語を起点に作成されたかによりサブセットに分けられており，そのうちのBTEC1, BTEC2, BTEC3, BTEC4 を用いた。BTEC1, BTEC2, BTEC3 は日本人が主に欧米へ行く場面の表現である。BTEC4 はアメリカやオーストラリアから旅行者が日本へ来る場面の表現である。検索対象となる対訳形式の発話表現数はあわせて約 49 万である。日英中の三言語パラレルとなっているが，そのうちの日英方向について実験を実施した。予備選択の数N_rは 30 に設定した。

表 3　大規模コーパス

	BTEC1	BTEC2	BTEC3	BTEC4
発話表現数（$\times 10^3$）	172	46	198	74
日本語延べ形態素数（$\times 10^3$）	1,174	342	1,434	548
日本語異なり形態素数（$\times 10^3$）	28	20	43	22
言語方向(起点:対訳)	J:EC	J:EC	J:EC	E:JC

検索の重みパラメータ α の値を 0 から 1 まで 0.1 きざみとし，表 1 の 3 種類のテストセットを用いて実験を行った。音声認識結果を入力として第 1 位候補で完全一致した正解率(Top 1 Accuracy), 同じく第 30 位までの中に完全一致したものが含まれている正解率(Top 30 Accuracy), 正解形態素列を入力とした場合の正解率(Oracle Accuracy)を条件 A (Condition A), 条件 B (Condition B)に対して求めた結果を発話正解率(Correct Utterance)とともに図に示す。図 1 が BTEC, 図 2 が MAD, 図 3 が FED の結果である。

BTEC の実験はクローズドのものである。MAD, FED の実験については，挨拶のような定型的表現は含まれているであろうが，オープンの結果である。BTEC については，汎化条件 A が汎化条件 B よりも良い結果となっているが，それ以外の MAD, FED は汎化条件 B が汎化条件 A より良い結果となっている。BTEC で汎化条件 A の場合のみ，音声認識結果を入力として 30 位までの候補を出すと 1 位のみの場合に比べて正解率の向上が見られるが，それ以外のテストセット，条件は 1 位のみと 30 位までの累積の値の間にあまり差はない。

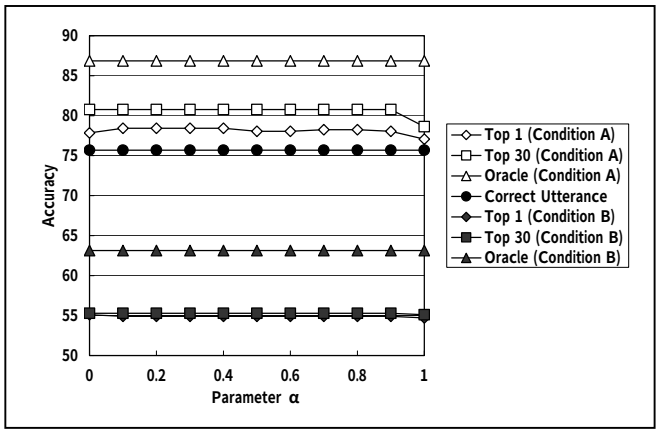

図 1　BTEC の実験結果

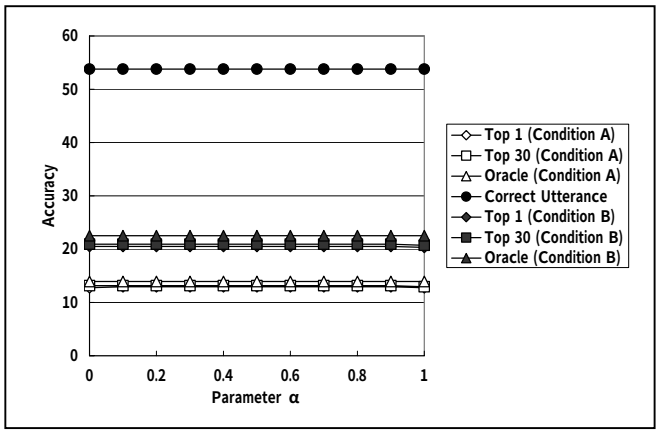

図 2　MAD の実験結果

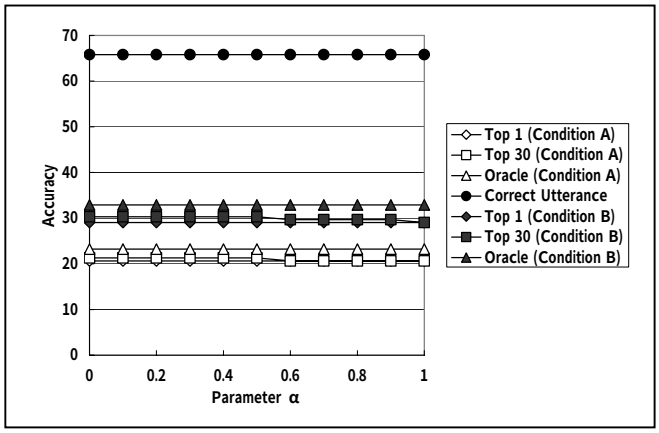

図 3　FED の実験結果

3.3 人手評価および考察

図 1，図 2，図 3 の結果は得られた類似用例が元の言語表現と形態素列として完全一致している割合を求めたものである。実際には表現が若干異なっていても同じ意味内容を伝達するとみなせる場合がある。そこで，実対話音声のテストセットである FED に対して人手でその内容を評価してみた。類似用例として選ばれた第 1 位の候補のみを次の ABCD ランクに分類した。

(A) 形態素列として完全に一致している。
(B) 表現は異なるが，同じ意味内容である。
(C) 一部の役に立つ情報を含んでいる。
(D) 意味が異なり，役に立たない。

類似用例が得られなかったものは次の二つに分類した。

(OK) 音声認識結果が発話を単位として正しい。
(NG) 音声認識結果が誤っている。

汎化条件 A に対する人手評価結果を図 4 に，汎化条件 B に対する人手評価結果を図 5 に示す。

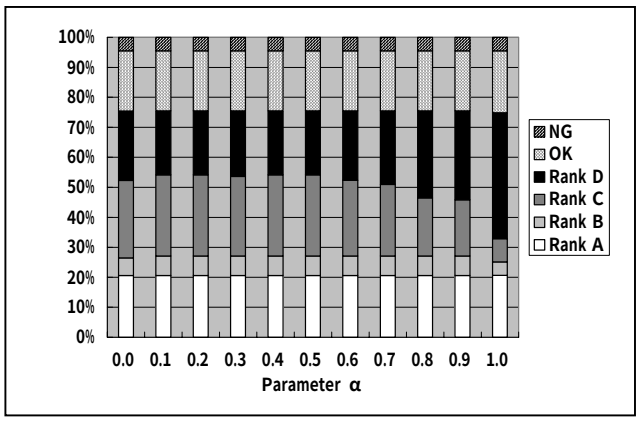

図 4　FED 汎化条件 A に対する人手評価結果

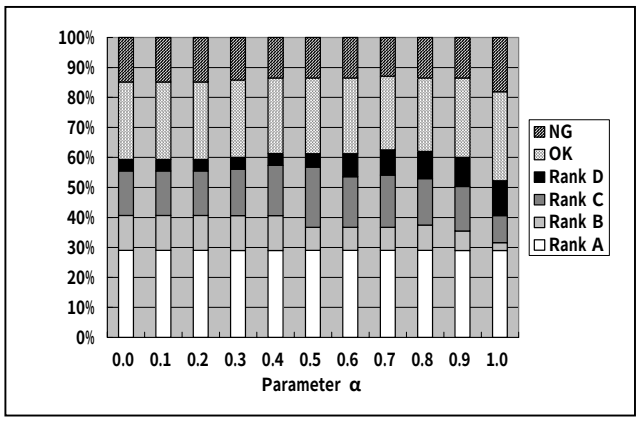

図 5　FED 汎化条件 B に対する人手評価結果

実対話音声テストセット FED では，汎化条件 A の場合に類似用例を出力する割合は約 75%であるが，役に立たないものが 3 分の 1 程度含まれている。汎化条件 B の場合には類似用例を出力する割合は約 60%になるものの，役に立たないものが少ない。汎化条件 A でランク D であったものが，汎化条件 B でランク B に向上した例を示す。

【入力音声】電車[ん]もしくはバスがありますが，どちらがよろしいですか。
【汎化条件 A】プリント紙は光沢と絹目がありますが，どちらがよろしいですか。
【汎化条件 B】電車とバスではどちらがよろしいですか。

類似用例が検索されなかったが，認識結果が正しいという OK に分類されるものの例を示す。

【OK の例】二百三十円です。

4 議論および関連研究

実対話音声テストセット FED は発話正解率が約 66%であった。汎化条件 B でパラメータαが 0.4 の時にランク AB の累積，ランク ABC の累積ともに良い。その場合，ランク AB の約 41%については翻訳品質が良いと期待できることになる。できる限り文脈に依存せず，かつ，情報の過不足のないように対訳コーパスが整備されていれば，対訳をそのまま使う手法の品質は保証されることになる。対訳の一部を置換する手法，および，それと組み合わせた評価については今後の課題とする。

関連研究に下畑等のもの[13]がある。目的は，話し言葉に含まれる言い淀み，言い直し，助詞の脱落などの影響を減らすことにある。まず，D3 [5]で翻訳用例を検索する際に編集距離がある閾値より大きいものを翻訳不能文と分類する。そして，その翻訳不能文に対して単言語コーパスから類似文を検索する。元の文の代わりに検索された類似文を翻訳することで適切に翻訳できる文が増えることを，対話の書き起こしテキストを用いた実験で示している。これに対し，我々の目的はシステム利用者を含んだ系で音声翻訳結果の品質を保証したいというものである。そのため，認識結果に近い対訳用例を検索してシステム利用者に提示し選択させる。対訳用例そのもの，あるいは，対訳用例の一部を置換したものの翻訳品質は良いという考え方は共通である。実装にあたり，D3 [5]は意味辞書が必要であるが，メンテナンスコストの必要な意味辞書を使わなくとも，品詞の汎化操作と，汎化した数が対訳ペアで一致するかどうかによる適切性判定を行えばよいという可能性を示した。

5 むすび

音声対話翻訳を音声言語コミュニケーションシステムとして実用化するためには，システム利用者が自らのメッセージを込めた発話の音声翻訳結果の性能を保証する技術が必要である。メッセージの言語表現に含まれるユーザ意図に近い表現をシステム開発用大規模コーパスから検索して提示し，システム利用者に選択させることで対訳そのものに近い形で翻訳することにすれば，その発話の音声翻訳結果の性能を保証することができると期待できる。そこで，大規模コーパスを用いて音声認識結果に類似した言語表現を検索する手法を検討し，旅行会話文の読み上げ音声と実対話音声の複数のテストセットを用いた評価実験を行って，その効果を示した。今後は，利用者がさらにその一部を編集するような機能等について検討を行う予定である。

参考文献

[1] 古井貞熙, “話し言葉の音声理解へ―存在感のある研究を期待して―,” 情報処理学会研究報告, Vol. 98, No. 114, SLP-24-18, pp. 129-136 (1998).

[2] Uchimoto, K., Hayashida, N., Ishida, T., and Isahara, H., “Automatic rating of machine translatability,” Proc. of MT Summit X, pp. 235-242 (2005).

[3] Watanabe, T. and Sumita, E., “Example-based decoding for statistical machine translation,” Proc. of MT Summit IX, pp. 410-417 (2003).

[4] Nagao, M., “A framework of a mechanical translation between Japanese and English by analogy principle,” Elithorn and Banerji (Editors): Artificial and Human Intelligence, NATO Publications (1984).

[5] Sumita, E., “Example-based machine translation using DP-matching between word sequences,” Proc. of ACL 2001 Workshop on Data-Driven Machine Translation, pp. 9-16 (2001).

[6] 伊藤玄, 葦苅豊, 實廣貴敏, 中村哲, “音声認識統合環境 ATRASR の概要と評価報告,” 日本音響学会 2004 年秋季研究発表会講演論文集 I, 1-P-30, pp. 221-222 (2004).

[7] 實廣貴敏, 松田繁樹, 藤本雅清, Herbordt, W., 堀内俊治, 中村哲, “ATR における日本語音声認識の評価―日本語音響モデル―,” 日本音響学会 2006 年春季研究発表会講演論文集, 1-P-21, pp. 185-186 (2006).

[8] 山本博史, 菊井玄一郎, “ATR における音声認識の評価―コーパスと言語モデル―,” 日本音響学会 2006 年春季研究発表会講演論文集, 1-P-22, pp. 187-188 (2006).

[9] Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., and Yamamoto, S., “Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world,” Proc. of International Conference on Language Resources and Evaluation, pp. 147-152 (2002).

[10] Kikui, G., Sumita, E., Takezawa, T., and Yamamoto, S., “Creating corpora for speech-to-speech translation,” Proc. of 8th European Conference on Speech Communication and Technology, Vol. 1, pp. 381-384 (2003).

[11] Takezawa, T. and Kikui, G., “A comparative study on human communication behaviors and linguistic characteristics for speech-to-speech translation,” Proc. of International Conference on Language Resources and Evaluation, pp. 1589-1592, (2004).

[12] 菊井玄一郎, 竹澤寿幸, 水島昌英, 山本誠一, 佐々木裕, 河井恒, 中村哲, “音声対話翻訳システムの実環境におけるモニタ実験,” 日本音響学会 2005 年秋季研究発表会講演論文集, 1-7-10, pp. 11-12 (2006).

[13] 下畑光夫, 隅田英一郎, 松本裕治, “発話を対象とした類似文検索と機械翻訳への適用,” 自然言語処理, Vol. 11, No. 4, pp. 105-126 (2004).