音声翻訳評価におけるパラフレーズデータの利用法

1R−3

菅谷 史昭† 安田 圭志† 竹沢 寿幸† 山本 誠一†

ATR音声言語通信研究所†

1. はじめに

著者等は音声翻訳システムの評価尺度として，翻訳一対比較法[1],[2],[3]により音声翻訳システムと人間能力を比較する方法を提案している．翻訳一対比較法では，主観によりシステムと人間の翻訳結果の優劣を判定していた．本比較法により，日英方向の音声翻訳システムの能力を日本人英語学習者のTOEIC能力に対応づけることができた．ATR-MATRIX音声翻訳システム[4]のTOEIC換算点は500点程度であることを評価実験により明らかにした．一方，システム開発を効率化するためには，自動評価が必要である．自動評価手法としては，予め準備した正解と翻訳結果のＤＰマッチングから計算される類似度を使った距離尺度を利用する手法[5],[6]がある．本距離尺度を利用するためには，翻訳結果の表現に沿った正解を準備する必要がある．そこで，正解をパラフレーズで収集した．本論では，翻訳一対比較法，翻訳一対比較法の自動化，類似度，パラフレーズデータの収集方法，そして自動評価法の実験結果について述べる．

2. 翻訳一対比較法

日英方向の翻訳一対評価法の処理の流れを図1に示す．翻訳一対比較法では，日本語のテスト問題をシステムと，異なる英語能力の日本語ネイティブに音声翻訳させた結果をネイティブのアメリカ人が比較し，勝率が同率となったところをシステムと人間との音声翻訳能力の均衡点とみなす．均衡点に対応した人間能力尺度をシステムの能力と定義する．人間能力尺度がTOEIC(Test of English for International Communication)の場合は，システム能力はTOEICスコアで表される．言語翻訳(TDMT)の翻訳結果と人間の音声翻訳能力の比較結果を図2に示す．縦軸は問題文330文に対する優劣判定結果である．

3. 自動翻訳一対比較法

図1の翻訳一対比較法では，システムと人間の翻訳結果を英語ネイティブが主観で判断していた．この比較判断を類似度尺度の大小で自動的に判定する．

類似度

文Siと文Sjの類似度sim(Si,Sj)は，文を構成する単語列間のＤＰマッチング結果である置換(Sub)，挿入（Ins)，削除（Del）を使い式（1）で計算する．単語間の距離は編集距離を使う．

$$sim(Si,Sj)=(N-Sub-Del-Ins)/N \quad (1)$$

ここで，Ｎは文Siの単語数である．

類似度はDPマッチングの単語比較を使っているので，表層の単語並びが異なれば正解でも，値が小さくなるという問題がある．そこで，表層表現の異なる正解を追加することにより，DPマッチングベースの類似度の改善を試みる．追加された正解を含む複数正解文 {Ai, i=1,...,L} と翻訳結果Tの類似度の最大値max{sim(Ai, T),i=1,...,L}を正解群類似度と呼ぶ．

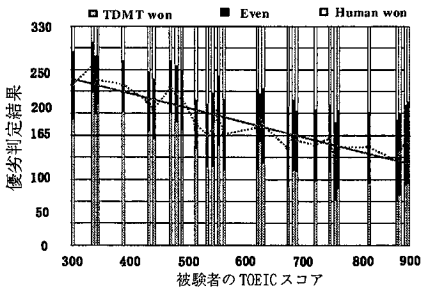

図2　言語翻訳部の評価結果

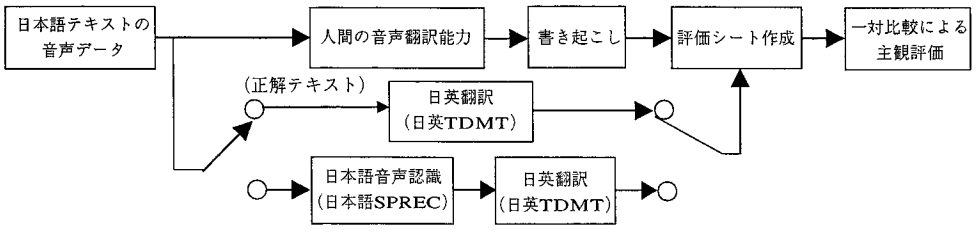

図1　翻訳一対比較法の処理の流れ

Usage of paraphrased data for evaluation of speech translation system

Fumiaki Sugaya, Keiji Yasuda, Toshiyuki Takezawa, and Seiichi Yamamoto

†ATR Spoken Language Translation Research Laboratories

4．パラフレーズデータ収集方法

ATRで独自に収集整理している日英バイリンガル旅行対話データベースから取られた23対話（330文）からなるSLTA1テストセットで日英自動評価実験をした．コーパスベースの言語翻訳で使われる一般的な対訳データベースには，原言語と目的言語が1対収められている．SLTA1テストセットも1対の（原言語，目的言語）ペアであるので，正解群類似度を計算するための翻訳正解を，原言語と同義な表現をパラフレーズにより集める．パラフレーズは5人の翻訳家に依頼し，翻訳家は各原言語文に対して異なる翻訳を対訳以外に3文ずつパラフレーズした．パラフレーズされる文の総数の最大値は15文であるが，重複が生じるので平均14.4文となった．5人の翻訳家で，各3文のパラフレーズでは，パラフレーズされる文の総数は飽和していない．

5．言語翻訳部のTOEIC換算点

TOEIC受験者のSLTA1テストセットに対する翻訳結果の正解群類似度のテストセット平均を図3に示す．正解群類似度の複数正解文はパラフレーズされた英文である．TOEICスコアの増加に伴い正解群類似度は増加している．直線は回帰直線である．言語翻訳部（TDMT）の正解群類似度は0.48であり，音声認識（SPREC）部を介した言語翻訳部の正解群類似度は0.45である．これらのシステムの正解群類似度と回帰直線から，言語翻訳部単体のTOEICスコアは682.9，そして音声認識部を介した場合は547.3と求まる．主観の優劣判定による翻訳一対比較法では，それぞれ707.6，548.1となり，最大でも差分は25点程度である．

6．パラフレーズ正解文の効果

パラフレーズにより正解文を追加することにより，翻訳に沿った正解を準備できる確率が増加すると考えられるので，客観尺度である正解群類似度と翻訳の質の相関が高くなること期待できる．一方，優劣判定による翻訳一対比較のシステムの勝率は，評価者が主観により翻訳結果を判定した結果である．これらの主観値と客観尺度の関係を調べるために，図4に言語翻訳部のシステムの勝率と正解群類似度の関係を示す．X印が1つの正解の場合である．○印は図3のデータに対応していて，15文のパラフレーズデータを付け加えた場合である．パラフレーズをすることにより，正解群類似度の値が大きくなっている．また，システム勝率と正解群類似度の相関をみると，正解が一つの場合は0.89，正解群サイズを増やした場合は0.91となり相関が大きくなることがわかる．

7．むすび

主観の優劣判定が必要な翻訳一対比較法の自動化法として，複数正解をパラフレーズにより追加するDPマッチングベースの類似度尺度を用いた自動化手法を提案した．本手法が，主観の翻訳一対比較法から求まるTOEICスコアと良い一致を示すことを確認した．また，翻訳一対比較法のシステム勝率と類似度が強い相関があること，またパラフレーズにより相関が大きくなることを示した．パラフレーズされた文のサイズとその効果については今後の検討課題である．

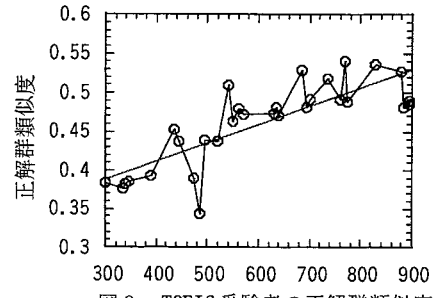

図3　TOEIC受験者の正解群類似度

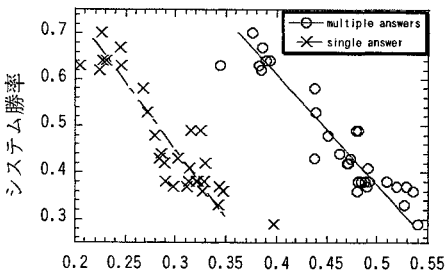

図4　言語翻訳部のシステム勝率と正解群類似度の関係

文献

[1] 菅谷史昭，竹澤寿幸，横尾昭男，山本誠一，"ATR-MATRIXと人間との音声翻訳能力比較実験"，日本音響学会2000年春季研究発表会講演論文集I，March 2000.

[2] 菅谷史昭，竹澤寿幸，横尾昭男，山本誠一，"音声翻訳システム（ATR-MATRIX）の評価"，信学技報，SP2000-23，pp. 39-45, June 2000.

[3] F. Sugaya, T. Takezawa, A. Yokoo, Y. Sagisaka, S. Yamamoto,"Evaluation of the ATR-MATRIX speech translation system with paired comparison method between the system and humans", Proc. ICSLP2000, pp. 1105-1108.

[4] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida. F. Sugaya, A. Yokoo, S. Yamamoto, "A Japanese-to-English speech translation system: ATR-MATRIX", Proc. ICSLP 1998, pp. 2779-2782.

[5] H. S. Thompson, "Automatic evaluation of translation quality: outline of methodology and report on pilot experiment", Proc. of the evaluators' forum, pp. 215-223,1991.

[6] 安田圭志，菅谷史昭，竹澤寿幸，横尾昭男，山本誠一，柳田益造，"対訳コーパスを用いた表層的類似度に基づく翻訳能力自動評価法"，信学技報，SP2000-111，pp. 97-102, Dec. 2000.