

表層的類似度に基づく翻訳一対比較法の自動化に関する検討

安田圭志^{††} 菅谷史昭[†] 竹澤寿幸[†] 山本誠一[†] 柳田益造[†] ([†]ATR 音声言語通信研究所, ^{††}同志社大)An automatic evaluation method of translation capability
and its application to a translation paired comparison methodKeiji YASUDA^{††}, Fumiaki SUGAYA[†], Toshiyuki TAKEZAWA[†], Seiichi YAMAMOTO[†], Masuzo YANAGIDA[†]([†]ATR Spoken Language Translation Research Laboratories, ^{††}Doshisha University)

1 はじめに

翻訳システムの性能改善の効率化や主観評価のコストを削減するためには、言語翻訳の自動評価技術が必要である。われわれは、これまでに ATR 音声翻訳通信研究所で研究開発された日英双方向音声翻訳システム ATR-MATRIX の評価を通じて翻訳システムの評価方法についての研究を進めてきた。翻訳の品質を 4 ランクに評価者が主観で割り当て、翻訳ランクを決定する「翻訳ランク評価法」や、翻訳システム能力と人間能力との比較を通じて翻訳システム能力を評価し TOEIC スコアに換算する「翻訳一対比較法」[1]を実施してきた。しかしながら、何れの方法も、主観による判定が必要であり、それに要するコストは少なくない。翻訳システムの性能改善の効率化には客観評価が必要となってきた。

客観評価手法としては、「DP マッチングによる評価手法」や「正解文追加類似度計算法」[2]が提案されている。正解文追加類似度計算法では、正解表現として登録されていない未登録の正解に対して類似度が小さくなるという DP マッチングによる評価手法の問題点を、対訳コーパスを用いることにより改善している。

本論文では、正解文追加類似度計算法を用いて、翻訳システム能力の TOEIC スコアへの換算を自動化する手法を検討し、その手法を ATR-MATRIX の評価に適用した結果と翻訳一対比較法での結果とを比較する。

2 正解文追加類似度計算法

本章では、正解文追加類似度計算法[2]について簡単に説明する。

翻訳評価のための類似度(Similarity)を、以下に定義する。

$$S = \frac{t - s - i - d}{t} \quad (1)$$

ここで、 S は類似度であり、 t は正解翻訳文の総語数、 s は正解翻訳文と翻訳システムからの翻訳出力を DP マッチングで比較した時の置換語数、 i は同様に比較した時の挿入語数、 d は同様に比較した場合の脱落語数である。

正解文追加類似度計算法では、対訳コーパスから式(1)により、原言語側で類似した表現に対する目的言語による表現を追加することで得られた複数の翻訳結果を用いる。以降これを「正解群」と呼ぶ。

言語翻訳結果と正解群の中の各文とで、DP マッチングにより言語翻訳結果のスコアリングを行ない、類似度を求める。この結果として、正解群に含まれる文の数だけ類似度が求まるが、その最大類似度を正解群類似度(answer set similarity)とし、これを翻訳文の評価尺度としている。

3 TOEIC スコア評価の自動換算法

本章では、翻訳システム能力の TOEIC スコアへの自動換算法について説明し、ATR-MATRIX に適用した結果について述べる。以降、この手法を「TOEIC スコア自動換算法」、翻訳システム能力の TOEIC スコアへの換算結果を「TOEIC 換算点」と呼ぶ。

正解文追加類似度計算に用いたコーパスは、ATR で構築された 618 会話 (16110 文) からなるバイリンガル旅行対話データベースである。評価用のテストセットはこの内の 23 会話 (330 文) である。この 23 会話は、言語翻訳部に対してオープンである。

3.1 TOEIC スコア自動換算法

TOEIC スコア自動換算法には、テストセットを

ATR-MATRIXにより翻訳した結果と、TOEIC受験者により翻訳させた結果を用いる。TOEIC受験者のTOEICスコアは300点から900点までである。TOEIC受験者は100点台毎に5名、計30名であり、895点の被験者が2名である以外は、全て異なるスコアである。尚、これらのTOEIC受験者による翻訳結果は、翻訳一対比較法によりTOEIC換算点を求める際に用いられているものと同じである。

これらの各TOEIC受験者による翻訳結果について、正解文追加類似度計算法により各文の正解群類似度を求める。次に、TOEIC受験者ごとにテストセット330文についての平均正解群類似度を求める。この結果がFig.1である。Fig.1の縦軸がTOEIC受験者のTOEICスコアで、横軸が平均正解群類似度である。図中の●は、TOEIC受験者を表しており、図中の直線は回帰直線である。TOEICスコア自動換算には、この回帰直線にATR-MATRIXによる翻訳結果330文についての平均正解群類似度を代入し、翻訳システム能力に相当するTOEICスコアを求める。図中の▲がATR-MATRIXによる翻訳結果の平均正解群類似度であり、■は翻訳システム能力に相当するTOEICスコアである。

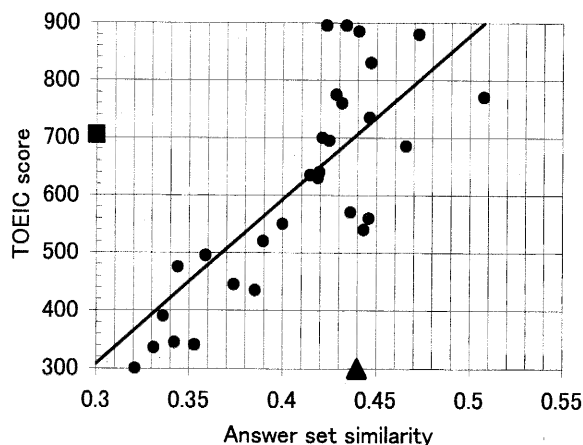


Fig.1 Relation between TOEIC score of human subjects and the average of answer set similarities

3.2 TOEICスコアの換算結果

文献[1]によれば、翻訳一対比較法によるTOEIC換算点は708点であり、Fig.1よりえられる提案手法によるTOEIC換算点は705点となっている。両手法によるTOEIC換算点の差は、わずか3点であり、提案

手法により翻訳一対比較法と近い結果が得られている。ただし、これら結果は音声認識の段階で誤りが含まれない条件のものである。音声認識誤りが含まれる場合においては、提案手法によるTOEIC換算点が618点、翻訳一対比較法によるTOEIC換算点が548点である。両手法による換算点の差は70点となり、音声認識誤りが含まれない場合よりも、両手法間のTOEIC換算点の差が大きくなっている。

Table1は、これらの結果をまとめたものである。括弧内が音声認識誤りが含まれる場合のTOEIC換算点である。

Table1 System's TOEIC scores measured by the proposed method and the translation paired comparison method

	System's TOEIC score
Translation paired comparison method	708 (548)
proposed method	705 (618)

4 むすび

正解文追加類似度計算法を用いた翻訳システム能力のTOEICスコアへの自動換算手法について検討した。この手法をATR-MATRIXの評価に適用した結果、人手による評価手法である翻訳一対比較法と近い結果が得られることが示唆された。音声認識誤りが含まれる場合については、本手法と翻訳一対比較法によるTOEIC換算点に70点の差があった。

今後は、まず音声認識誤りが含まれる場合の結果の分析を行う。更に、別の実験を追加し、本手法の信頼度や誤差の検討を進めたい。

謝辞 本研究の一部は、同志社大学学術フロンティア事業の援助を受けた。

文献

- [1] F. Sugaya et al., "Evaluation of the ATR-MATRIX Speech Translation System with Pair Comparison Method Between the System and Humans", Proc. ICSLP 2000, pp.1105-1108, 2000.
- [2] 安田, 菅谷, 竹澤, 山本, 柳田, "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, 2000.