対訳コーパスを用いた翻訳自動評価法の性能評価

5Y−1　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)

Performance evaluation of an automatic method for evaluation of translation quality using parallel corpus

Keiji 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，これまでに対訳コーパスを用いた翻訳自動評価手法[1]を提案してきた．本手法は，対訳コーパスから検索した複数の翻訳正解と，システムによる翻訳結果とをDPマッチングにより比較する方法である．これまでの研究では，本手法と，人手によりパラフレーズして得られた複数正解を翻訳自動評価に用いた場合[2]との比較は行われていなかった．以降，対訳コーパスを用いた翻訳自動評価法を「検索型自動評価法」，人手でパラフレーズして得られた翻訳正解を用いた自動評価法を「パラフレーズ利用型自動評価法」と呼ぶ．本論文では検索型自動評価法と，パラフレーズ利用型自動評価法とを，評価性能の観点から比較する．

2　検索型自動評価法

本章では，検索型自動評価法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．

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

ここで，Sは類似度であり，tは翻訳正解文の総語数，sは翻訳正解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．

検索型自動評価法では，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．

パラフレーズ利用型自動評価法においては，正解群の代わりに，人手で作成した複数正解を用いて同様の処理を行う．

本論文で検索型自動評価法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．また，翻訳品質の評価対象は，ATR-MATRIXの言語翻訳サブシステムによる日英方向の翻訳結果とした．

3　複数正解収集法

パラフレーズ利用型自動評価法に用いる複数正解は，日本語が分かる5名の英語ネイティブが，各テスト文につき，3文のパラフレーズをすることにより収集した．5名が個別にパラフレーズを行うため．パラフレーズの結果が，同じ文になる場合もあるが，その数は少なく，1文の原言語文につき，平均で14.4文の異なり目的言語文が収集されている．複数正解DPマッチングにおいては，ここで収集したデータに加え，対訳テストセットの英語側をも用いるため，正解数が最大で16文となる．

4　評価結果

本章では，評価性能の観点から，検索型自動評価法とパラフレーズ利用型自動評価法の比較，検討を行う．

両自動評価手法の評価性能の比較は，人手により以下の基準でランク付けされた翻訳ランクと，各自動評価法による評価結果との相関により行う．

(A) 完全訳: 訳文だけで全く問題なし．
(B) 部分訳: 訳文は少し情報が欠けている．
(C) 可能訳: 訳文はかなり情報が欠けている．
(D) 不可訳: 訳文からは，情報が想像もできない．

相関を求めるにあたって，各翻訳ランク(A)〜(D)を3〜0に数値化し，各文について5評価者の平均（MOS値）を求め，これを用いる．

図1に両手法と，MOS値との相関を示す．図1において，縦軸はMOS値との相関を表しており，横軸はパラフレーズ利用型自動評価法で用いた正解数及び，検索型自動評価法を表している．

図1より，複数正解の数が増えるほどMOS値との相関が高くなっていることが分かる．また，検索型自動評価法においては，パラフレーズ利用型自動評価法で正解数を10文とした場合に相当する相関が得られている．

図2は検索型自動評価法とMOS値の相関と，パラフレーズ利用型自動評価法とMOS値の相関との差分を，エントロピー毎に求めた結果である．ここで用いたエントロピーは，マルチクラス複合2グラムの言語モデルから求めた原言語文の文毎の平均単語エントロピーである．

図2において，縦軸は相関の差分を表しており，横軸はエントロピーを表している．また，図2の白の棒グラフは，パラフレーズ利用型自動評価法における正解数が1文の場合の結果を表しており，黒の棒グラフは正解数が16文の場合の結果を表している．図2より，エントロピーが低い部分では検索型自動評価法が有効であり，エントロピーが高い部分では，パラフレーズ利用型自動評価法の方が有効であることが示されている．これは表1に示すように，検索型自動評価法ではエントロピーが低いテスト文において，人手で作成するより十分な数の正解が集められており，一方，エントロピーが高い文において，人手で作成するほど十分な文が集められていないためであると考えられる．

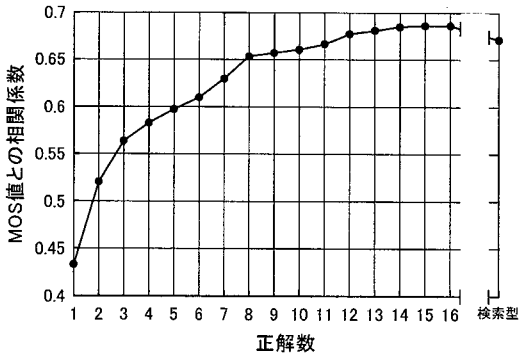

図1 正解数とMOS値との関係

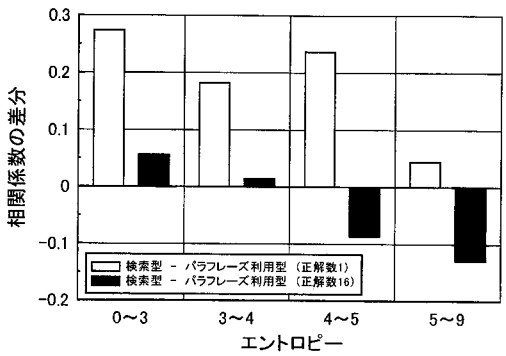

図2 エントロピー毎の相関の差分

表1 エントロピーと
検索型自動評価法で追加される正解数の関係

エントロピー	正解群に含まれる異なり文数の平均
0〜3	30.47
3〜4	10.72
4〜5	7.54
5〜9	2.42
0〜9	11.72

5 むすび

評価性能という観点から，翻訳ランクとの相関に着目し，検索型自動評価法とパラフレーズ利用型自動評価法との比較，検討を行った．この結果，検索型自動評価法は，パラフレーズ利用型自動評価法で正解数を10文用いた場合に相当する評価性能であることが示された．また，エントロピー毎に評価性能を比較すると，エントロピーが低い文では検索型自動評価法が有効であり，エントロピーが高い文ではパラフレーズ利用型自動評価法が有効であるとの結果が得られた．

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]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.

[2] H.S., Thompson “Automatic Evaluation of Translation Quality: Outline of Methodology and Report on Pilot Experiment”, Proc. Evaluators' Forum, pp.215-223, 1991.