TR-S-0024

正解文追加類似度計算法の英日翻訳評価への適用

藤川 晶子
Akiko Fujikawa
安田 圭志
Keiji Yasuda
菅谷 史昭
Fumiaki Sugaya
竹澤 寿幸
Toshiyuki Takezawa

2001.3.30

翻訳自動評価技術の改善手段として、正解文追加類似度計算法が提案され、日英方向で有効性が検証されている。今回、英日方向の翻訳評価への適用により、その有効性を更に明らかにする。また、これまでの研究で得られている日英方向の翻訳文評価結果との比較について考察する。

1. 本研究の目的

翻訳能力自動評価法を改善する一つの試みとして、正解文追加類似度計算法が提案されている[1]。正解文追加類似度計算法の有効性に関しては、これまでの研究で、日英方向の翻訳評価における有効性は明らかにされてきた。しかしながら、英日方向の翻訳評価への適用はなされておらず、英日方向からの有効性は確認されていなかった。また、日英方向への適用結果との比較により更なる知見が得られるのではないかとの議論があった。そこで、正解文追加類似度計算法を、英日方向の翻訳評価に適用し、日英方向での結果との比較を行った。本稿では、この実験結果について考察する。

2. 翻訳評価法

本節では、従来の翻訳評価法と、正解文追加類似度計算法について説明する。

2.1 主観評価

従来の主観評価においては、評価者が以下の 4 基準で、翻訳文を主観により評価する方法である。

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

主観による翻訳能力の評価は、多大なコストを要し、評価者によって評価結果に差がみられることが、これまでの研究で分かっている。

2.2 従来の DP マッチングによる評価法

一つの原言語テスト文に対し、あらかじめ登録済みの目的言語の一文だけを翻訳正解文とする評価手法が提案されている[2]。この方法では、翻訳文と翻訳正解文とで DP マッチングにより、以下に定義する類似度を計算し、これを評価尺度とする方法である。

$$Similarity = \frac{Total - Sub - Ins - Del}{Total}$$

Total は、正解文の総語数、 *Sub* は、正解文と翻訳正解文を DP マッチングにより比較した場合の置換語数、*Ins* は、同様に比較した場合の挿入語数、*Del* は、同様に比較した場合の脱落語数を示す。これにより得られた値を類似度（*Similarity*）と呼ぶ。

従来 DP マッチングによる評価法では、類似した意味を持つが、異なる言い回しの翻訳文は、正しく翻訳されている場合でも、類似度が小さくなる。このことが、自動評価性能を低くする要因となっている。

2.3 正解文追加類似度計算法による自動評価法

2.1 節および 2.2 節で述べたように、主観評価と従来の DP マッチングによる評価手法には、それぞれの問題点がある。これらの問題点を改善するため、正解文追加類似度計算法が提案されている。この手法では、登録済みの正解文に加え、未登録の類似文を、正解候補として対訳コーパスから収集し、それらの中から翻訳文に対し最も類似度の大きい文を、翻訳文の評価尺度にする方法である。従来の DP マッチングによる評価手法と同様に、DP マッチングをベースとしているが、正解文を 1 つとせず、複数の中から選択することにより、評価能力を改善している。

具体的な処理の手順を、図 1 に示す。図中の原言語コーパス (Source language corpus) と目的言語コーパス (Target language corpus)、及び、原言語テスト文 (Source language test sentence)と目的言語テスト文 (Target language test sentence) は、対訳関係にある。

DP マッチングにより、原言語コーパスの中から、原言語テスト文の類似文を抽出する。その結果得られた類似文の中から、類似度がある一定の閾値以上となるものを選出し、それらに相当する文を目的言語コーパスから選択する。ここでの閾値を類似文検索閾値と呼ぶ。検索された目的言語文と、目的言語テスト文とを合わせて正解群とする。TDMT による翻訳文と正解群の中の各文とで、DP マッチングを行い、類似度を求める。その中から最も類似度の大きい文を正解群類似度(Answer set similarity)とし、これを翻訳文の評価尺度とする。

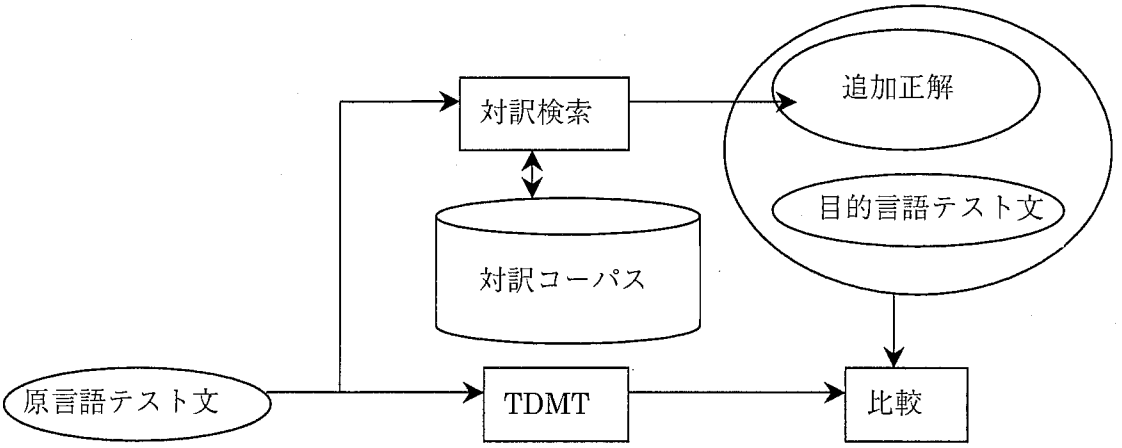

図 1　正解文追加類似度計算法の処理の流れ

3. 正解群追加類似度計算法の TDMT への適用

3.1 実験条件

今回の実験では、原言語は英語、目的言語は日本語としている。テスト文は原言語、目的言語それぞれ 344 文で、対訳コーパスは 16110 文の対である。評価対象は、TDMT にテキスト入力して得られた翻訳文、すなわち、音声認識による誤りを含まない翻訳文（以下、TDMT 翻訳文と呼ぶ）を対象としている。

また、正解文追加類似度計算法における原言語の類似文検索については単語単位での DP マッチングを行い、目的言語でのスコアリングでは形態素単位での DP マッチングを行っている。

3.2 実験結果

3.2.1 類似文検索閾値と追加される正解文数の関係

類似度がある一定の閾値以上の文を類似文、閾値を類似文検索閾値と呼ぶ。類似文検索閾値は、0.1 毎に、0.1 から 1.0 までとする。表 1 で、類似文検索閾値別に、追加される正解文の例を示す。

重複を許す場合ののべ文数を比較した結果を図 2 に、重複を許さない場合の異なり文数を比較した結果を図 3 に示す。図 2 および図 3 において、横軸は類似文検索閾値を、縦軸は追加される文の数の平均値を表している。

重複を許す場合と許さない場合共に、英日方向の追加文数が、日英方向に比べ少なくなることが分かった。また、類似文検索閾値の増加に伴い追加文数が対数的に減少するという同一の傾向が見られた。

表 1 類似文検索閾値別にみる追加文の例

テスト文
I have a twin room with a partial ocean view at two hundred and sixty dollars room per night.
TDMT 翻訳文
一泊につき２６０ドルの一部分海洋眺めのツインルーム御部屋がございます。
正解文
１つが一泊２６０ドルで海が少し見える御部屋になります。

閾値別にみる追加される正解文の例

類似文検索閾値=0.4
ツインの御部屋で一泊１６０ドルぐらいまでの部屋があるといいんですが。
１つが一泊２６０ドルで海が少し見える御部屋になります。
はい海が少しご覧になれるツインの御部屋は一泊２６０ドルです。
そしてツインでエキストラベッドをお入れした場合には一泊１７０ドルでございます。
ツインの料金は一泊２５０ドルからでございます。
ツインは一泊２５０ドルからになっております。

類似文検索閾値=0.6
はい海が少しご覧になれるツインの御部屋は一泊２６０ドルです。
１つが一泊２６０ドルで海が少し見える御部屋になります。

類似文検索閾>0.9
１つが一泊２６０ドルで海が少し見える御部屋になります。

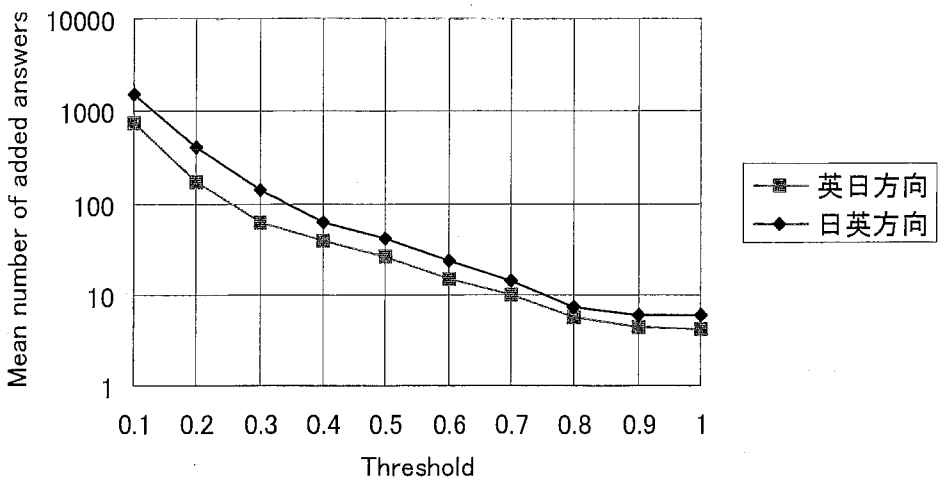

図 2　類似文検索閾値と追加文のべ数の関係

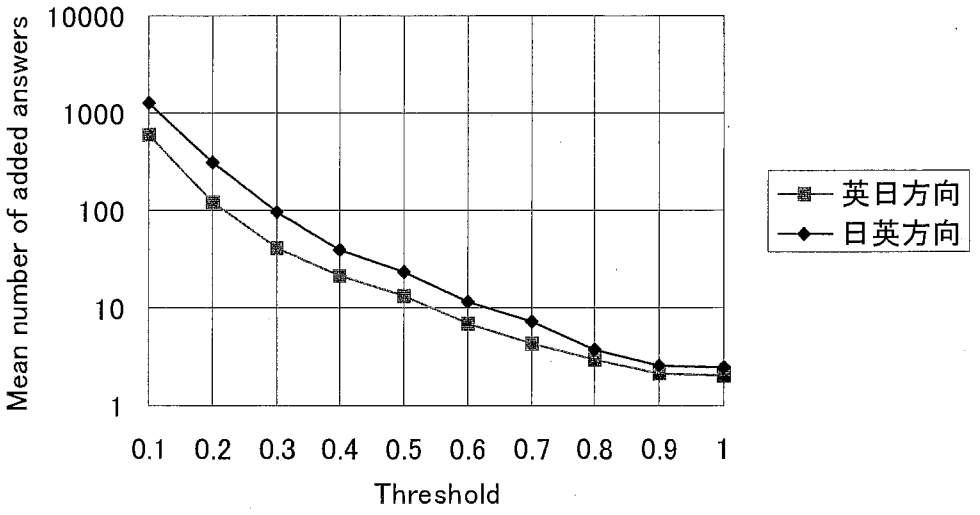

図 3 類似文検索閾値と追加文異なり数の関係

3.2.2　主観評価による翻訳ランクと正解群類似度の関係

図 4 に、閾値 0.6 の場合の英日方向に適用した結果得られた主観評価のランクと正解群類似度の関係を示す。図 5 は、日英方向の場合の結果である。図 4 および図 5 において、横軸は翻訳ランク、縦軸は正解群類似度を表しており、図中の点は各ランクでの平均正解群類似度を表しており、エラーバーは各ラン

ク内での正解群類似度の1σの区間を表している。図4では、翻訳ランクが高いほど、平均正解群類似度も大きくなる傾向がみられている。これは、図4と図5の比較から、日英方向の翻訳文評価結果と同一の傾向であることが分かる。

日英方向と英日方向との相違点としては、日英方向の適用結果に比べ、類似度の平均値はほぼ0.1下がる。Aランクの差が最大であり、日英方向での平均値が約0.7であるのに対し、英日方向では約0.4となり、0.3程度低くなった。

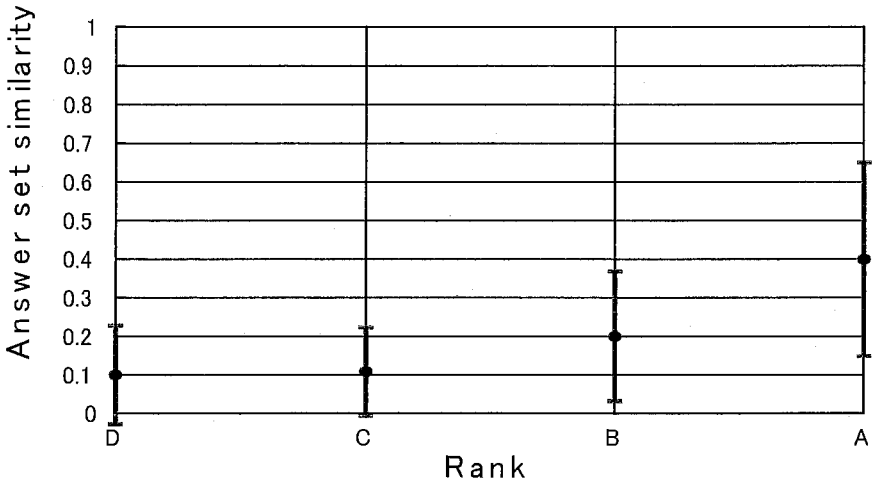

図4 英日方向における主観評価による翻訳ランクと正解群類似度の関係

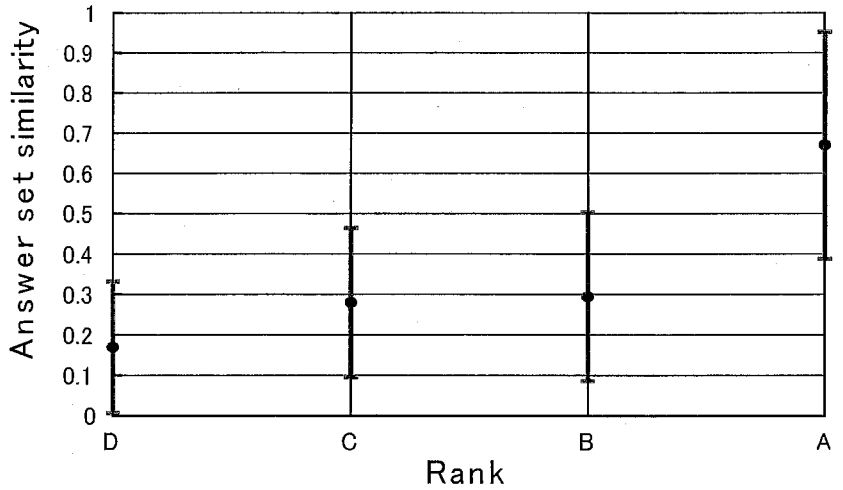

図5 日英方向における主観評価による翻訳ランクと正解群類似度の関係

3.3 類似文検索閾値と判別率の関係

正解群類似度から翻訳ランクの自動判定を行うため、主観評価で決定されたランクを用いて判別分析を行った。

2クラス分けと4クラス分けのそれぞれの場合について、以下に定義する判別率を求める。2クラス分けでは、Aランクと B,C,D ランク、 A,B ランクと C,D ランク、 A,B,C ランクと D ランクの3通りとする。4クラス分けについては、主観評価で得られる4ランクを対応させた4クラスとする。判別方法としては、最近傍則判別を用いている。

判別率 ＝ 正しく判別された文の数／全体の文の数

判別率が高いほど、類似度による自動評価が主観評価に近いことを意味する。

図6に正解文追加類似度計算法を英日方向の翻訳評価に適用した結果を示す。図7は日英方向での結果である。図6及び図7において、横軸は類似文検索閾値または、従来の DP マッチングによる評価結果を用いた場合を表し、縦軸は判別率を表している。

日英方向の翻訳評価結果と比較してみると、英日方向の判別率が全体的に低いが、ほぼ同一の傾向が見られた。全てのクラス分けにおいて、従来の DP マッチングによる評価結果を用いた場合と比較し、評価結果が主観評価により近づくことが分かった。実際、正解を追加しない場合に比べ、提案手法による判別率が上昇している。この傾向は日英方向でも確認されており（図 7)、これまでに正解の対訳が1つとされていたTDMT翻訳文に複数の類似文を追加することによって、正解の幅を広げることが両方向ともに可能となったためであると考えられる。

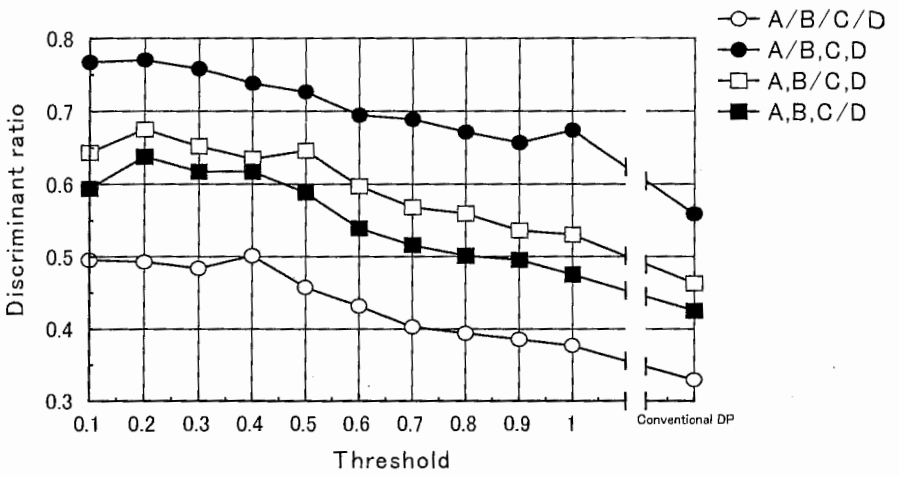

図６　英日方向における類似文検索閾値と判別率の関係

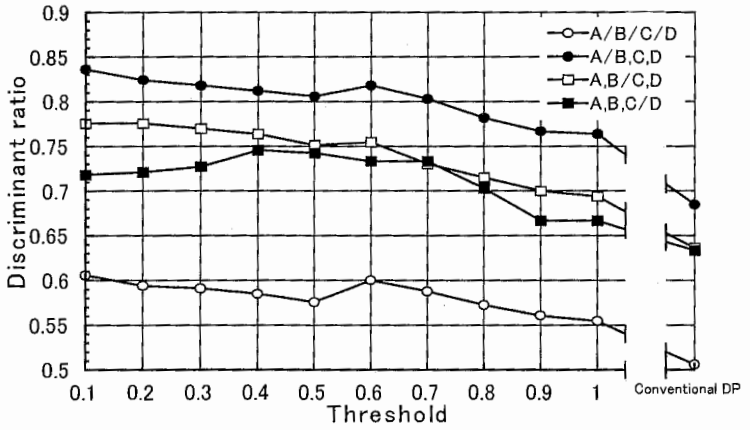

図７ 日英方向における類似文検索閾値と判別率の関係

3. まとめ

正解文追加類似度計算法を用いて英日方向の翻訳文の評価を行い、日英方向の翻訳文評価結果と比較した。その結果、類似文検索閾値と追加される文数の関係、主観評価による翻訳ランクと正解群類似度の関係、判別率と類似文検索閾値の関係において、ほぼ同一傾向を示した。正解文追加類似度計算法と従来の DP マッチングによる評価手法の評価能力を判別率により比較した場合、英日方向では、日英方向と同様に、判別率の改善がみられた。これにより、正解文追加類似度計算法は、日英、英日両方向の翻訳評価において有効であることが確認された。

但し、追加文数、正解群類似度、判別率が英日方向では、日英方向での適用結果よりも低くなっており、今後、この要因についての調査が必要である。

謝辞

ATR 音声言語通信研究所で、1ヶ月間の本実習の機会を与えてくださった匂坂芳典第2研究室室長をはじめ、研究を進めるにあたり、様々な御支援を頂いた山本博史主任研究員、中嶋秀治研究員、吉川徹氏、吉岡由紀氏、村上純子氏、TSG の皆様に厚く感謝致します。

最後に、音の学習と対話学習の研究に関するご鞭撻をくださったATR 先端情報科学研究部　山田玲子プロジェクトリーダーに心から御礼申し上げます。

参考文献

[1] 安田圭志, 菅谷史昭, 竹澤寿幸, 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 SP 2000-111, pp.97-102, 2000.

[2] K. -Y. Su, M. -W. Wu, and J. -S. Chang, “A new quantitative quality measure for machine translation systems”, Proc. COLING, pp. 433-439, 1992.