部分目標の達成度に基づく機械翻訳自動評価 — 部分目標の自動生成 —

内元 清貴　　小谷 克則　　張 玉潔　　井佐原 均

独立行政法人 情報通信研究機構
{uchimoto,yujie,isahara}@nict.go.jp
kat@khn.nict.go.jp

1　はじめに

機械翻訳の研究において、機械翻訳の品質評価は重要な課題であると認識されてきた。近年、その品質評価を自動化しその性能を向上させようという試みが数多くなされている [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] 。しかし、従来の機械翻訳自動評価手法では、個々の文について各システムの翻訳の優劣を判別するのは難しい。その原因のひとつは、従来の手法では翻訳文と参照訳文との間で共通する単語列が多く現れているほど訳質を高く評価する傾向があるために、局所的かつ致命的な翻訳誤りがあってもその誤りが過少評価されやすいからである。

一般に、ある文を翻訳する際には、英日翻訳で言えば前置詞や不定詞の訳し分けのように、翻訳品質を良好に保つために満たすべき条件がひとつ以上存在する。我々は、それらの局所的な条件を部分目標として設定し、その部分目標に着目した評価指標を従来の大局的な評価指標と統合することにより、従来の手法に比べて飛躍的に良く個々の翻訳文の品質を自動評価できることを示した [11]。しかし、部分目標は人手で与える必要があった。本稿では、部分目標を自動生成する方法を提案し、人手で与えた部分目標と同定度の性能が得られることを実験的に示す。

2　機械翻訳品質評価用テストセット

2.1　JEIDA テストセット

機械翻訳品質評価用データは、様々なものが公開されているが、そのほとんどは対訳例文のみからなり、例文に翻訳評価のための様々な情報が付与されているのは珍しい。本稿では、機械翻訳の品質評価用に収集された対訳文例で、翻訳評価のための情報が付与されているものを、テストセットと呼び、その典型例として、JEIDA(日本電子工業振興協会) のテストセット [12] を取り上げる。

JEIDA のテストセットに付与されている特徴的な情報としては、翻訳結果を評価するための yes/no 設問があげられる。この設問は、例えば、「for が「〜で」のように原因・理由を表すように訳されていますか?」といったもので、この設問に対し人間が yes/no で回答することによって、翻訳結果を客観的に評価することができるようになっている。例文と訳出例、設問の例は表 1 の上欄の通りである。設問は主として文法的な観点からカテゴリ分けされており、番号のハイフンより左 (1.1.7.1.3) がカテゴリを表わす。例えば、表 1 の「オリジナル」の欄にある設問は連鎖動詞に関するものである。

JEIDA のテストセットには英日機械翻訳用と日英機械翻訳用のものがあるが、以下では英日機械翻訳用テストセットを対象とする。この英日機械翻訳用テストセットで設問が設定されているのは 769 例文である。テストセットの分析のために、その 769 例文を 5 つの商用の機械翻訳システムでそれぞれ翻訳した。そして、各翻訳文に対し、yes/no 設問への回答による主観評価と、fluency と adequacy による主観評価を行なった。このとき、fluency と adequacy の主観評価は文献 [13] に従った。一方、yes/no 設問への回答による主観評価は、各翻訳文に対しその翻訳元の例文に付加された設問に yes/no で回答することにより行なった。

次に、yes/no 設問への回答による主観評価と fluency、adequacy との関係を知るために、両者の相関を調べた。以降であげる相関はいずれも 1%未満の有意水準で統計的に有意である。yes/no 設問への回答による主観評価については、yes を 1、no を −1 と数値化した。結果は、5 システムによる翻訳文、3,845 文に対し、fluency に対する相関係数は 0.53、adequacy に対する相関係数は 0.67 であった。それぞれに比較的強い相関があることが分かる。特に adequacy との相関が強い。

2.2　JEIDA テストセットの拡張

JEIDA のテストセットでは、ひとつの例文に対してひとつの設問が付与されているため、設問の対象外の部分に翻訳誤りがあっても評価には影響しない。そのため、重要な誤りを含む翻訳文は過大評価されやすい。そこで、この問題を解消するために、重要な翻訳誤りに対して設問を追加することによって、テストセットを拡張した。

拡張の対象とした例文は、5 システムの翻訳結果に対する fluency および adequacy の平均値が 3 以下であった 94 例文、および、fluency および adequacy の平均値が 3 以上かつ 5 システムの翻訳結果に対する yes/no 設問への回答が 3 システム以上に対して no であった 56 例文の合計 150 例文とした。設問の追加は次の手順により行なった。

1. 翻訳結果から重要な翻訳誤りを抽出する。
2. 抽出した翻訳誤りに対し、テストセット中から、同様の誤りに関する設問を探す。あれば、対象文に対して同一内容の設問を生成し、同じ設問番号を付与する。なければ、新たに設問と番号を設ける。
3. 追加した設問を用いて各システムの翻訳結果を評価する。

この結果、ひとつの例文に対し、ひとつないし複数の設問が設けられた。新たに設問が設けられたのは、拡張対象とした 150 例文のうち 103 例文であった。追加された設問の合計は 148 問、ひとつの例文に付与された設問の最大数は 5 問であった。設問の追加により、fluency に対する相関係数は 0.53 から 0.57 に、adequacy に対する相関係数は 0.67 から 0.70 に向上した。この拡張の前後で得られた相関係数の差はいずれも 5%未満の有意水準で統計的に有意である。この結果は上記の手順でテストセットを拡張することにより yes/no 設問への回答による主観評価と fluency と adequacy による主観評価の相関をより強くすることができることを示している。ひとつの例文に対し設問が複数ある場合、yes/no 設問への回答による主観評価の評価値は yes と no の多数決により決定し、yes が多ければ 1、no が多ければ -1、同数なら 0 とした。拡張の結果、追加された設問の例を表 1 に、yes/no 設問への回答による主観評価の例を表 2 にあげる。

3 部分目標の達成度に基づく機械翻訳自動評価

3.1 機械翻訳品質自動評価指標

JEIDA のテストセットは、人手による主観評価向けに作成されたものであるため、自動評価には向いていない。しかし、2 節で述べたように、yes/no 設問への回答による主観評価結果は fluency や adequacy と比較的強い相関があるため、設問への回答を自動推定することができれば、fluency や adequacy と相関の強い指標ができる可能性が高い。本節では、yes/no 設問への回答を自動推定し、推定した結果を用いることによって機械翻訳の品質を自動評価する手法について述べる。

以下で、テストセットの各 yes/no 設問を部分目標とし、翻訳文が与えられたとき、yes/no 設問への回答が yes の場合に部分目標が達成され、no の場合に部分目標は達成されなかったもの (非達成) とする。ある翻訳文に対する評価値 A は、部分目標の達成度および非達成度と、翻訳文と訳出例 (参照文) との類似度を用いて、線形重回帰モデルにより次のように定義する。観測データに対する回帰直線は最小二乗法により求められる [14]。

$$A = \sum_{i=1}^{m} \lambda_{S_i} \times S_i + \sum_{j=1}^{n} (\lambda_{Q_j} \times Q_j + \lambda_{Q_j'} \times Q_j') + \lambda_\epsilon \quad (1)$$

$$Q_j = \begin{cases} 1 & : \text{部分目標が達成された場合} \\ 0 & : \text{それ以外} \end{cases} \quad (2)$$

$$Q_j' = \begin{cases} 1 & : \text{部分目標が非達成の場合} \\ 0 & : \text{それ以外} \end{cases} \quad (3)$$

ここで、Q_j は i 番目の ID を持つ部分目標の達成度を、λ_{Q_j} はその重みを表わす。Q_j' は i 番目の ID を持つ部分目標の非達成度を、$\lambda_{Q_j'}$ はその重みを表わす。n は部分目標の種類数を表わす。λ_ϵ は定数である。

S_i は翻訳文と参照文との類似度で、λ_{S_i} はその重みである。類似度を計算する方法としてはこれまで様々なものが提案されてきた [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] 。実験では、類似度 S_i の計算に 23 の評価値、つまり、BLEU [3] を 4 種類 (単語 n-gram の最大長=$\{1, 2, 3, 4\}$)、NIST [4] を 5 種類 (単語 n-gram の最大長=$\{1, 2, 3, 4, 5\}$)、GTM [6] を 3 種類 (べき指数 e=$\{1, 2, 3\}$)、METEOR (exact) [9]、WER [1]、PER [5]、ROUGE [15] を 8 種類 (単語 n-gram の最大長=$\{1, 2, 3, 4\}$、および、4 つの異形 (LCS, S$*$, SU$*$, W-1.2)) 利用した。したがって、式 (1) の m の値は 23 である。実験では、単語分割には JUMAN [*] を用いた。

3.2 部分目標達成度の自動推定

部分目標の達成度は、各 yes/no 設問に対し回答を自動推定することにより決定する。本節では、単純なパターンに基づく回答自動推定手法について述べる。

設問に対する回答の自動推定は、翻訳文にパターンが含まれるか否かをチェックすることにより行なう。パターンは設問ごとに作成する。パターンの表記は仮名で統一し、パターンの適用は、翻訳文と訳出例を句読点なしの仮名文に変換した後に行なう。実験では、仮名文への変換には JUMAN を用いた。

表 3 は例文と設問および作成したパターンの例である。ここで、記号「｜」は「OR」を表わす.

表 3: パターンの例

例文	She lived there by herself.
設問	"by herself" が「独りで」のように訳されていますか？
パターン	【ひとり（だけ｜きり）で｜たんどくで｜たんしんで】が訳中に含まれていれば回答は yes、そうでなければ、回答は no
例文	They speak English in New Zealand.
設問	「ニュージーランドでは英語を話す」のように、人称代名詞 "they" の訳語が省略されて訳されていますか？
パターン	【かれらは｜それらは】が訳中に含まれていれば回答は no、そうでなければ、回答は yes

3.3 部分目標の自動生成と部分目標達成度の自動推定

yes/no 設問と参照文の関係を観察することにより、翻訳品質を良好に保つ鍵と考えられる表現は、例文に付与された yes/no 設問に関係があり、かつ、各例文の参照文に共通して含まれていることが多いことが分かった。したがって、本稿では、yes/no 設問を次の手順で自動生成する。

[*] http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html

表 1: 拡張例

オリジナル	番号	1.1.7.1.3-1
	例文	The percentage of stomach cancer among the workers appears to be the highest for any asbestos workers.
	訳出例	労働者の胃癌の割合は、アスベスト労働者のために最高となるようだ。
	設問	appear to が「ようだ」のように助動詞として訳されていますか？
追加	番号	1.1.6.1.3-5
	翻訳誤り	for が正しく訳出されていない。
	設問 1	for が「〜で」のように原因・理由を表すように訳されていますか？
追加	番号	追加 1
	翻訳誤り	長文のため、訳抜けがある。
	設問 2	英語文が全て日本語文に訳されていますか？

表 2: 評価例

システム	翻訳結果	回答 設問	設問 1	設問 2	fluency	adequacy
1	労働者の間の胃癌のパーセンテージは、どのような石綿労働者のためにでも最も大きいようである。	Yes	No	Yes	2	3
2	労働者の間の胃癌のパーセンテージは、あらゆるアスベスト労働者のために最も高いように思われます。	Yes	Yes	Yes	2	3
3	労働者の間の胃癌のパーセンテージはどんなアスベストのためにも最も高いように見えます	Yes	No	No	1	2
4	労働者の間の胃癌のパーセンテージは任意の石綿には最も高く見えます。	Yes	No	No	1	2
5	労働者の中の胃癌の割合はどんなアスベストにも最も高いように見える。	Yes	No	No	1	2

1. 各例文に対し、参照文に共通に現れる語の集合を抽出する。
2. 抽出した語集合の要素から構成される任意の n 語の組み合わせのうち、参照文に同じ語順で共通して現れるスキップ単語 n-gram を共通スキップ単語 n-gram の集合とする。
3. 部分目標を、「すべての共通スキップ単語 n-gram が翻訳文に含まれていますか？」という yes/no 設問として定義する。

ただし、共通スキップ単語 n-gram がない場合、yes/no 設問は生成されないものとする。yes/no 設問に対する回答は、すべての共通スキップ単語 n-gram が翻訳文に含まれていれば yes、そうでなければ no とする。

4 実験と考察

実験では、5 つの機械翻訳システムの翻訳文とその主観評価結果のうち、3 システム分を 3.2 節に述べたパターンの作成や式 (1) のパラメータ λ_{S_i}、λ_{Q_j}、$\lambda_{Q'_j}$、λ_ϵ のチューニングのための学習セットとし、残りの 2 システム分をテストのための評価セットとして用いた。

学習セットでは、各例文には設問と参照文がそれぞれ少なくともひとつ以上設けられ、3 システムによる機械翻訳結果とその主観評価結果が与えられている。3.2 節に述べたパターンは人手で作成した。769 例文に設けられた yes/no 設問は 917 問であり、そのうち、767 例文に設けられた 898 問に対してパターンを作成することができた。残りの 19 設問は、例えば、「一文全体がひとつの文として訳されていますか？」のようなもので、3.2 節に述べたような単純なパターンを作成するのは困難であった。このパターンを用いたとき、yes/no 設問に対する回答の自動推定結果は表 4 の通りである。

表 4: 設問への回答の自動推定結果

セット	精度
学習	97.6%(2,629/2,694)
評価	82.8%(1,487/1,796)

表 5: 評価値 A と fluency/adequacy との相関係数 (参照文がひとつのとき)

手法	fluency		adequacy	
	学習	評価	学習	評価
従来手法	0.53	0.51	0.49	0.49
提案手法 (自動)	0.92	0.60	0.94	0.63
提案手法 (上限)	0.92	0.66	0.94	0.75

表 6: 評価値 A と fluency/adequacy との相関係数 (参照文が 5 つのとき)

手法	fluency		adequacy	
	学習	評価	学習	評価
従来手法	0.56	0.56	0.53	0.56
提案手法 (自動)	0.92	0.61	0.94	0.64
提案手法 (完全自動)	0.86	0.61	0.87	0.64
提案手法 (上限)	0.93	0.66	0.94	0.75

テストセットの 769 例文を対象に、式 (1) の評価値 A

と fluency および adequacy との相関を調べた。パターンを作成できなかった 19 設問に対しては Q_j の値は 0 とした。まず、線形重回帰分析 [14] により、学習セットを用いて、評価値 A と fluency および adequacy との相関係数が最大となるように式 (1) のパラメータ λ_{S_i}、λ_{Q_j}、$\lambda_{Q'_j}$、λ_ϵ の値を最適化した。次に、パラメータの最適値セットを用いて、評価セットでの相関係数を調べた。結果を表 5 と表 6 にあげる。表で、「従来手法」は評価値 A の計算に類似度 S_i のみを用いたときに得られた相関係数を表わす。「提案手法 (自動)」は式 (1) の Q_j および Q'_j の計算に yes/no 設問に対する自動推定結果を用いたときに得られた相関係数を示し、「提案手法 (完全自動)」は式 (1) の Q_j および Q'_j の計算に部分目標の自動生成および部分目標達成度の自動推定の結果を用いたときに得られた相関係数を示している。このとき、部分目標の生成にはスキップ単語 trigram、スキップ単語 bigram、スキップ単語 unigram を用いた。「提案手法 (上限)」は Q_j および Q'_j の計算に人間による判定結果を用いたときに得られた相関係数を示している。

提案手法と従来手法の相関係数の差は、参照文が増えても、fluency、adequacy ともに 5%未満の有意水準で有意である。つまり、提案手法により得られた相関係数は従来手法に比べて有意に高い。この結果は個々の翻訳文の品質を自動評価するのに部分目標の達成度を考慮するのが有効であることを示している。

部分目標の自動生成および部分目標達成度の自動推定に基づく手法により得られた相関係数は、人手による部分目標の生成および部分目標達成度の自動推定に基づく手法により得られた相関係数と等しい。この結果は提案手法により、部分目標の設定に要する人手作業のコストを軽減できることを示している。しかし、「提案手法 (完全自動)」と「提案手法 (上限)」との相関係数にはまだ大きな差がある。この差を埋めるのは今後の課題である。

5 まとめと今後の課題

本稿では、翻訳品質を良好に保つために満たすべき条件を設問の形で各テスト文に付与したテストセットと、個々の設問に対する回答を自動推定するシステムを作成し、その設問と従来の評価指標を複数統合することにより、従来の手法に比べ個々の翻訳文の品質をより適切に自動評価することが可能となることを示した。また、yes/no 設問の形で部分目標を自動生成し、その達成度を自動推定する方法を提案し、その有効性を示した。

今後、テストセットをさらに拡張し、提案手法により得られる相関係数の上限値を向上させるとともに、提案手法も改良し、部分目標の生成および部分目標達成度の推定の性能を向上させたい。与えられた文の部分目標をより適切に自動生成するためには、文の複雑さや原文と翻訳文の間の対応情報も考慮する必要があると考えている。自動生成と自動推定の性能が向上すれば、翻訳の品質だけでなく、翻訳誤りの部分も分かるようになるだろう。研究がさらに発展すれば、将来、機械翻訳の自動評価指標を用いて機械翻訳システムのパラメータチューニングを行なうことにより、高性能な機械翻訳システムの実現も期待できる。

謝辞

テストセットの拡張指針は AAMT 技術動向調査委員会で作成されたものに基づいています。ご協力くださった委員の皆様、特に、小倉健太郎氏、島津美和子女氏、介弘達哉氏、富士秀氏、松川淑子女史に感謝いたします。

参考文献

[1] S. Niessen, F.J. Och, G. Leusch, and H. Ney. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the LREC 2000*, pp. 39–45, 2000.

[2] Y. Akiba, K. Imamura, and E. Sumita. Using Multiple Edit Distances to Automatically Rank Machine Translation Output. In *Proceedings of the MT Summit VIII*, pp. 15–20, 2001.

[3] K. Papineni, S. Roukos, T. Ward, and W. Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th ACL*, pp. 311–318, 2002.

[4] NIST. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Technical report, NIST, 2002.

[5] G. Leusch, N. Ueffing, and H. Ney. A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation. In *Proceedings of the MT Summit IX*, pp. 240–247, 2003.

[6] J.P. Turian, L. Shen, and I.D. Melamed. Evaluation of Machine Translation and its Evaluation. In *Proceedings of the MT Summit IX*, pp. 386–393, 2003.

[7] B. Babych and A. Hartley. Extending the BLEU MT Evaluation Method with Frequency Weightings. In *Proceedings of the 42nd ACL*, pp. 622–629, 2004.

[8] C. Lin and F.J. Och. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of the 20th COLING*, pp. 501–507, 2004.

[9] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pp. 65–72, 2005.

[10] J. Gimeńez, E. Amigo; and C. Hori. Machine translation evaluation inside qarla. In *Proceedings of the IWSLT'05*, 2005.

[11] 内元, 小谷, 小倉, 島津, 張, 介弘, 富士, 松川, 井佐原. 部分目標の達成度に基づく機械翻訳自動評価. 言語処理学会 第 12 回 年次大会 発表論文集, pp. 865–868, 2006.

[12] 井佐原, 内野, 荻野, 奥西, 木下, 柴田, 杉尾, 高山, 土井, 永野, 成田, 野村. 開発者の視点からの機械翻訳システムの技術的評価 — テストセットを用いた品質評価法 —. 自然言語処理, Vol. 3, No. 3, pp. 83–102, 1996.

[13] TIDES. Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Arabic-English and Chinese-English Translations. http://www.ldc.upenn.edu/Projects/TIDES/Translation/TransAssess02.pdf, 2002.

[14] N.R. Draper and H. Smith. *Applied regression analysis. 2nd edition.* Wiley, 1981.

[15] C. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pp. 74–81, 2004.