部分目標の達成度に基づく機械翻訳自動評価

内元 清貴 [*1]　小谷克則 [*1]　小倉健太郎 [*2]　島津美和子 [*3]
張玉潔 [*1]　介弘達哉 [*4]　富士秀 [*5]　松川淑子 [*6]　井佐原 均 [*1]

[*1] 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

[*2] 日本電信電話株式会社 NTT
サイバースペース研究所
ogura.kentaro@lab.ntt.co.jp

[*3] 東芝ソリューション株式会社
shimazu.miwako@toshiba-sol.co.jp

[*4] 沖電気工業株式会社 研究開発本部
ユビキタスシステムラボラトリ
sukehiro564@oki.com

[*5] 株式会社富士通研究所
fuji.masaru@jp.fujitsu.com

[*6] 日本電気株式会社 メディア情報研究所
y-matsukawa@ah.jp.nec.com

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] 。自動評価の性能が向上することにより、機械翻訳システムの利用や改良が効率良くできるようになることが期待されるからである。例えば、自動評価の指標をシステムパラメータのチューニングに利用することにより、翻訳性能が向上したという報告がある [4] 。この報告は、自動評価の性能が向上すれば翻訳性能が自動的に向上することを示している。

しかし、これまでに提案されている自動評価手法では、300 文程度のまとまったデータがあれば翻訳システムの優劣を判別することができるが、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。本稿では、それを目指した評価用テストセットの作成方法および機械翻訳自動評価手法を提案する。

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。本稿では、それらの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得るシステムを作成することによって、従来の手法に比べて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す。

2　機械翻訳品質評価用テストセット

2.1　テストセットの現状

機械翻訳品質評価用データには、無作為に例文を収集することによって作成されたものと、意図的に機械翻訳が困難な言語現象を含む例文を収集することにより作成されたものの大きく二種類が存在する。近年、機械翻訳技術の向上を支援する目的で、IWSLT (International Workshop on Spoken Language Translation) [*] などの機械翻訳品質評価のための評価型ワークショップが開催されており、MT-05 (NIST 2005 Machine Translation Evaluation) [†] や HTRDP Evaluation [‡] などのベンチマークテストも行なわれている。これらのワークショップ、ベンチマークテストで用いられているデータは、公平さを保つために新聞記事や旅行対話データなどから無作為に集められたもので、上記の分類では前者に相当する。一方、後者の分類に相当するものとしては、NTT から公開されているもの [6]、JEIDA(日本電子工業振興協会) により作成されたもの [6, 7] などが挙げられる。このうちほとんどのデータは対訳例文のみからなり、JEIDA のデータのように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。このように、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、以降で、テストセットと呼ぶことにする。

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 設問が挙げられる。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、設問の例は次の通りである。

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？

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、上の設問は連鎖動詞に関するものである。

[*] http://www.is.cs.cmu.edu/iwslt2005/

[†] http://www.nist.gov/speech/tests/mt/index.htm

[‡] http://www.863data.org.cn/

JEIDA のテストセットには英日用と日英用のものがあるが、以下では英日用テストセットを対象とする。この英日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、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。一方、fluency と adequacy の主観評価は文献 [8] に従った。

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、3845 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 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でも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 の平均が 3 以下であった 94 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. 抽出した翻訳誤りのそれぞれに対し、テストセット中から、類似した設問を探す。あれば、対象文に対しそれと類似した設問を生成し、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、必要に応じて複数の設問が設けられた。設問の追加により、Fluency に対する相関係数は 0.53 から 0.55 に、Adequacy に対する相関係数は 0.67 から 0.69 に向上した。設問が複数ある場合の yes/no 設問への回答による主観評価については、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、翻訳評価の例を表 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、システム開発者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たものとする。この部分目標の達成度を Q、翻訳文と訳出例 (参照文) との類似度を S とし、ある翻訳文に対する評価値 A を次の式で定義する。

$$A = S + \lambda \times Q \tag{1}$$

$$Q = \begin{cases} 1 : \sum_{i=1}^{n} f_i \times Q_i > 0 \text{ のとき} \\ 0 : \sum_{i=1}^{n} f_i \times Q_i = 0 \text{ のとき} \\ -1 : \sum_{i=1}^{n} f_i \times Q_i < 0 \text{ のとき} \end{cases} \tag{2}$$

ここで、Q は n 個の部分目標が与えられた場合の部分目標達成度を表わす。λ は部分目標達成度の重みである。Q の値は、部分目標の合否による多数決により決まり、達成されている部分目標の方が多ければ 1、少なければ-1、同数であれば 0 となる。$Q_i(0 \geq i \geq n)$ は i 番目のカテゴリを持つ部分目標に対する達成度を表わす。この Q_i は部分目標が達成されていれば 1、達成されていなければ-1 の値を返す。f_i は、評価対象の翻訳文に依存して 0 か 1 のいずれかの値をとる。

S は翻訳文と参照文との類似度である。類似度としては様々なものが考えられる。ここでは、機械翻訳自動評価手法としてよく用いられる BLEU [2] を用いた。BLEU score の計算式は次の式により表わされる。

$$\log \text{BLEU} = min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} \frac{1}{N} \log p_n \tag{3}$$

この式において、r は参照文の単語長、c はテスト文の単語長、N は考慮する単語 n-gram の最大の n の値を表わす。p_n は次の式で表わされる.

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。	○	×	○	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。	○	○	○	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	○	×	×	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。	○	×	×	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。	○	×	×	1	2

$$p_n = \frac{\sum_{C\in \text{参照文集合}} \sum_{n-gram\in C} Count_{clip}(n-gram)}{\sum_{C'\in \text{参照文集合}} \sum_{n-gram'\in C'} Count(n-gram')} \quad (4)$$

ここで、$Count(n-gram')$ はテスト文における単語 n-gram の出現頻度を表わす。$Count_{clip}(n-gram)$ は、次の式で表わされる。

$$Count_{clip}(n-gram) = min(Count, Max\_ref\_Count) \quad (5)$$

ここで、Max_Ref_Count はテスト文における単語 n-gram の出現頻度を表わす。

3.2 部分目標達成度自動推定システム

部分目標の達成度の判定には、yes/no 設問に対し回答を自動推定するシステムを作成して用いる。本節では、単純なパターンに基づくルールベースシステムについて述べる。

設問に対する回答の自動推定は、パターンを含むか否かの判定によって行なう。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。仮名文への変換には JUMAN [§] を用いた。

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

以下は例文と設問および作成したパターンの例である。

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る
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ない

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のうち、3 システム分をパターンの作成や式 (1) の λ 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。

まず、式 (1) の類似度 S と adeqaucy との相関が弱い 185 設問を対象に部分目標達成度自動推定システムを作成した。自動推定の精度を表 3 にあげる。

上記の 185 設問が付与されている 185 例文を対象に、式 (1) の部分目標達成度 Q と fluency および adequacy との相関を調べた。式 (3) の N は 3 とした。結果を表 4 にあげる。表で、「提案手法」は Q の計算に自動推定結果を用いた場合の結果を示し、「提案手法 (上限)」は Q の計算に人間による判定結果を用いた場合の結果を示し

表 3: 設問への回答の自動推定結果

	精度
closed test	95.9%(532/552)
open test	87.8%(325/370)

表 4: 部分目標達成度と fluency, adequacy との相関係数

手法	fluency		adequacy	
	closed	open	closed	open
BLEU	0.14	0.24	0.13	0.18
提案手法	0.49	0.44	0.66	0.53
提案手法 (上限)	0.56	0.47	0.76	0.68

表 5: 式 (1) の評価値 A と fluency, adequacy との相関係数

手法	fluency			adequacy		
	λ	closed	open	λ	closed	open
BLEU	0	0.38	0.37	0	0.36	0.37
提案手法 (自動)	0.4	0.44	0.41	0.6	0.47	0.43
提案手法 (人間)	0.5	0.46	0.42	0.7	0.50	0.45

ている。提案手法では BLEU に比べてかなり fluency, adequacy との相関が強くなっている。両者の相関係数の差は 0.5%未満の有意水準で有意である。この結果は、部分目標達成度を考慮することで特に adequacy と強い相関がある自動評価が可能となることを示している。

次に、769 例文を用いて、学習セットで λ を 0 から 0.1 刻みで変化させることによって、式 (1) の A と fluency, adequacy との相関が最大となる λ を調べ、その λ を用いて、評価セットでの相関係数を調べた。ただし、部分目標達成度の結果を用いたのは、上記で自動推定が可能となった 185 例文についてのみである。結果は表 5 の通りである。この表で、「提案手法 (自動)」は、評価値 A の計算の際に、185 例文についてのみ Q の自動推定結果を用い、残りについては BLEU score を用いた場合の結果を示しており、「提案手法 (自動)」は、評価値 A の計算の際に、185 例文についてのみ Q の人間による判定結果を用い、残りについては BLEU score を用いた場合の結果を示している。相関係数の差は、adequacy に対しては 5%未満の有意水準で有意である。この結果は、たとえテストセットの一部であっても、部分目標を設定し、その達成度を推定して利用することにより、従来の手法よりも adequacy との相関が強くなることを示している。

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得るシステムを作成することによって、従来手法に比べて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。

今後の課題としては、まず、テストセットのさらなる拡張、部分目標達成度自動推定システムの拡張と精度向上があげられる。次に、今回は機械翻訳自動評価において設問の重みを一定としたが、これを設問の重要性に応じて変えることがあげられる。

将来的には次のような発展が考えられる。

- 部分目標の自動生成

  現在は部分目標を人間が与えているが、例文に応じて部分目標である設問を自動生成し、その回答を自動推定できるようにしたい。より適切に自動生成するためには、原文の難しさ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が可能になれば、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。さらに自動生成の手法が発展すれば、機械翻訳自動評価に参照文が必要ではなくなる可能性もある。

- テストセットの自動生成

  現在はテストセットの例文も人間が収集あるいは作文しているが、将来的には翻訳が難しい文を自動的に収集できるようにしたい。

- テストセット、システムの多言語化

  現在は日英・英日のみが対象であるが、他の言語にも拡張してきたい。

以上が可能になれば、本手法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。

参考文献

[1] Sonja Niessen, Franz Josef Och, Gregor Leusch, and Hermann Ney. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the LREC 2000*, pp. 39–45, 2000.

[2] Kishore Papineni, Salim Roukos, Todd Ward, and Weiing. Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, 2002.

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

[4] Franz Josef Och. Minimum Error Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 160–167, 2003.

[5] Joseph P. Turian, Luke Shen, and I. Dan Melamed. Evaluation of Machine Translation and its Evaluation. In *Proceedings of the MT Summit IX*, pp. 386–393, 2003.

[6] 池原悟, 白井諭, 小倉健太郎. 言語表現体系の違いに着目した日英機械翻訳機能試験項目の構成. 人工知能学会誌, Vol. 9, No. 4, 7 1994.

[7] 井佐原均, 内野一, 荻野紫穂, 奥西稔幸, 木下聡, 柴田昇吾, 杉尾俊之, 高山泰博, 土井伸一, 永野正, 成田真澄, 野村浩郷. 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.

[8] 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.