機械翻訳可能性の自動評価

内元 清貴 † 林田 尚子 ‡ 石田 亨 ‡ 井佐原 均 †

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

‡ 京都大学
hysd@kuis.kyoto-u.ac.jp,
ishida@i.kyoto-u.ac.jp

1 はじめに

近年、インターネットの無料翻訳サービスが普及したこともあり、機械翻訳 (MT) システムは、プロの翻訳家から一般のユーザまで幅広く利用されるようになってきた。このような状況に鑑み、MT システムを使って国境を越えた人間同士のコミュニケーションを支援するプロジェクトが立ち上げられた。これは異文化コラボレーション実験プロジェクト (ICE プロジェクト) [1] と呼ばれており、アジア各国の大学、研究機関等によって共同で進められている。このプロジェクトの目標は異文化間かつ多言語間で行なわれる共同作業 (コラボレーション) を MT システムを使って支援することである。この目標を達成するための第一歩として、アジア各地の母国語を異にする人々がチームを組んでオープンソースソフトウェアの共同開発に取り組むという実験が行なわれた。この実験では、発信者は母国語でメッセージを作成しそれを MT システムで他の言語に翻訳して同じチームのメンバーに送信し、受信者は受け取ったメッセージを母国語で読んで理解する、という取り決めがあった。しかし、MT システムの翻訳には理解不能なあるいは誤解を生じるような誤りが多くあり、その誤りを改善して意図が伝わるようにするために、発信者は何度も元のメッセージを、より機械翻訳可能なもの、つまり、MT システムが適切に翻訳できるものに書き換える必要があった。ところが、一部の誤りが他にも影響し、文あるいは句全体が理解不能となることがよくあるため、受信者がメッセージのどこが理解不能な表現あるいは誤解を生じる表現で書き換えるべき個所であるかを送信者に伝えることは難しい。したがって、送信者は試行錯誤によって書き換えるべき個所を模索する必要があった。

そこで、本稿では、MT システムと入力文が与えられたときに、その MT システムによって入力文の機械翻訳がどの程度うまく行なえるかを機械翻訳可能性と定義し、これを自動評価する方法を提案する。入力文の機械翻訳可能性が高いと評価されるのはその翻訳が良い翻訳であった場合であり、良い翻訳であるかどうかを自動評価しようとすると一般に参照文が必要となる。しかし、提案手法では、MT システムには原言語から対象言語への翻訳と対象言語から原言語への翻訳がともに行なえることが要求されるが、MT(以下、自動翻訳と同義) の自動評価に必要とされるような参照文がなくても入力文を自動評価できる。このように機械翻訳可能性が自動評価できるようになることにより、MT システムを介したコミュニケーションがスムーズに行なえるようになることが期待できる。

2 確信度 (C-measure)

本研究では、確信度 (Confidence measure: 以下、C-measure) を、原文とその**折り返し翻訳文**との類似度と定義する。ここで、折り返し翻訳文は、原言語文を対象言語へ翻訳したものをさらに原言語に翻訳することによって得られる文と定義する。類似度の計算方法については、2.1 節で述べる。類似度の持つべき性質は、与えられた二文の意味が同じなら値が大きくなることである。本研究では、C-measure が高いほど翻訳が安定しており翻訳の信頼性が高いと仮定する。もちろん、この仮定はいつも真とは限らない。例外については、3 節で述べる。以下では、C-measure を計算するために、日英・英日の MT システムを用いる。実験に用いたのは商用の MT システムのひとつである。

2.1 C-measure の特徴について

本研究では、原文とその折り返し翻訳文の類似度が高いほど機械翻訳可能性が高いと仮定する。ここで、類似度は、MT 自動評価手法としてよく用いられる BLEU [2] を拡張したものを用いる。C-measure は下記の式により計算する。

$$\mathrm{CM} = \frac{2 \times \mathrm{CM_{bleu}(B|S)} \times \mathrm{CM_{bleu}(S|B)}}{\mathrm{CM_{bleu}(B|S)} + \mathrm{CM_{bleu}(S|B)}} \quad (1)$$

ここで、$\mathrm{CM_{bleu}(B|S)}$ の S と B はそれぞれ、原文とその折り返し翻訳文を表わす。log をとると、次の式により表わされる。

$$\log(\mathrm{CM_{bleu}(B|S)}) = min\left(1 - \frac{s}{b}, 0\right) + \sum_{n=1}^{N} \frac{1}{N} \log p_n(B|S) \quad (2)$$

この式において、s は原文の単語長、b は折り返し翻訳文の単語長、N は考慮する単語 n-gram の最大の n の値を表わす。$p_n(B|S)$ は次の式で表わされる。

$$p_n(B|S) = \frac{\sum_{wn \in B} Count_{clip}(wn)}{\sum_{wn' \in B} Count(wn')} \quad (3)$$

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

$$Count_{clip}(wn) = min(Count(wn), Max\_Count(wn|S)) \quad (4)$$

ここで、$Max\_Count(wn|S)$ は S における単語 n-gram wn の出現頻度を表わす。式 (1) と BLEU score の計算式との違いは次の通りである。

- 依存構造木に基づく単語 n-gram

  日本語や韓国語などいくつかの言語においては、語順が比較的自由である。例えば、「太郎と花子はテニスをした」という日本語文の文節を単位とする依存構造は「((太郎と 花子は)(テニスを した))」のように表わされ、この依存構造からは、「太郎と花子はテニスをした」と「テニスを太郎と花子はした」の二種類の語順の日本語文が生成可能である。BLEU score はフラットな単語列における単語 n-gram に基づいて計算されるため、この二文の BLEU score は 1 とはならない。しかし、この二文は同じ意味

で、同じ訳文になると考えられるため、類似度は 1 にしたい。そこで、依存構造木に基づく単語 n-gram を採用することにした。単語の単位は形態素解析システム JUMAN [3] で定義されている形態素とし、文節内の単語はすべて隣に係り、係り文節における末尾の形態素は受け文節の先頭の形態素に係ると仮定する。予備実験を基に、式 (2) で N =3 とした。

- 調和平均

  BLEU score は MT 文における単語 n-gram の精度に基づいて計算されるため、 MT 文と参照文を入れ替えたときに計算される BLEU score は元のものとは異なる。しかし、類似度としては、入れ替えても同じ値になるようにしたい。そこで、式 (1) で表わされるように、参照文に対する MT 文の BLEU score だけでなく、 MT 文に対する参照文の BLEU score も考慮することにした。

- 汎化

  BLEU score は表層単語に基づいて計算される。したがって、類義語は別の単語として扱われる。しかし、二つの文の違いが類義語の関係にある単語のみであった場合には、類似度は 1 になるようにしたい。そこで、単語を単語クラスに置き換えて汎化することにした。ひとつの単語が複数の単語クラスに属する場合は、原文と折り返し翻訳文との間で一致する単語クラスの数が最も多くなるように山登り的に準最適な単語クラスの集合を探索する。単語クラスとしては「分類語彙表」 [4] の上位から 5 レベル目の階層を用いる。分類語彙表に収録されている単語の異なり数は 101,070 である。さらに、接続表現や数量表現をひとつのクラスに汎化するため、品詞カテゴリが「接続助詞」あるいは「数詞」である単語は品詞に汎化し、連続する数詞はひとつの数詞に置き換えた。そして、敬体と常体を区別しないために「接尾辞」の「ます」は無視するようにし、句読点の有無によって類似度が異なるようなことがないように句読点も無視するようにした。将来的には、句・節・文レベルの汎化も考慮したい。

2.2 MT 評価指標との関係

本節では、 C-measure と MT 自動評価指標である BLEU や NIST [5] および、人間による主観評価との関係について述べる。 C-measure は入力文とその折り返し翻訳文を基に計算され、 BLEU score および NIST score は、入力文を MT システムにより翻訳した英訳文と英語参照文を基に計算される。

テストセットとしては、 NTT より配布されている MT テストセット [6] *を用いた。このテストセットは、日英 MT システムの評価用に作成されたもので、 3,718 文の日本語文とその英訳からなる。このテストセットには、 ICE プロジェクトの実験においても実際に出てきそうな問題が含まれていると考える。このテストセットにおける各日本語文に対し、ひとつずつ英語参照文を選択し、自動評価に用いた。 BLEU score と NIST score の計算には、 mteval (version v11a) †. を用いた。

図 1 から図 4 に、それぞれ、 C-measure と BLEU score および NIST score との関係、 C-measure と人間による主観評価 (fluency,adequacy) との関係を示す。これらのグラフは、 C-measure に閾値を設けて 0 から 1 の間で変化させ、各閾値に対しその閾値を越える入力文を抽出し、その英訳の自動評価値を計算することによって描いたものである。 fluency と adequacy の主観評価は文献 [7] に従った。主観評価にはテストセットの内、先頭から奇数番号の例文を 950 文抽出して用いた。 C-measure として、 2.1 節に述べた各特徴を採用した場合と採用しなかった場合との違いが比較できるように、各図の各グラフには、 C-measure として、それぞれ、 BLEU(”bleu”)、 NIST(”nist”) ‡、調和平均 (”harmonic”)、依存構造木に基づく単語 n-gram(”tree-ngram”)、汎化 (”generalization”) およびこれらの組み合わせを採用した場合の結果を示した。

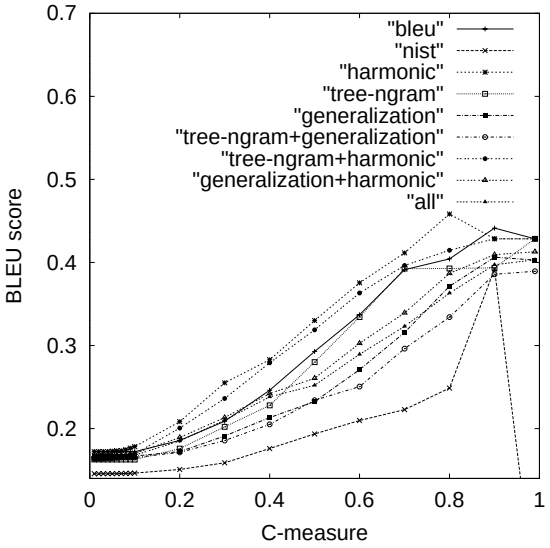

図 1: C-measure と BLEU score との関係

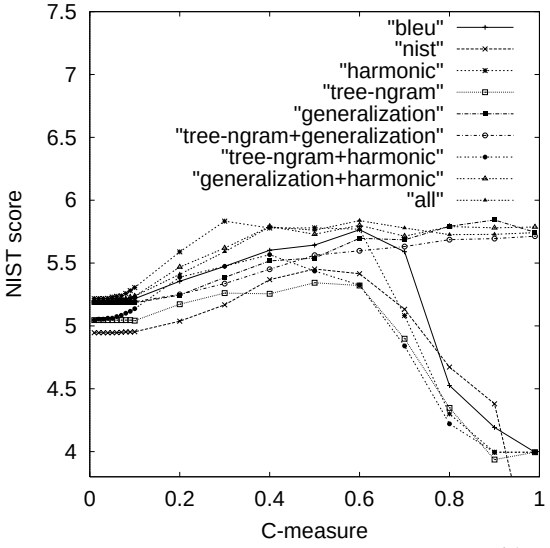

図 2: C-measure と NIST score との関係

C-measure として、 2.1 節に述べた全特徴を採用した場合、図 1 から図 4 における C-measure と BLEU score、 NIST score、 fluency、 adequacy との相関係数は、それぞれ、 0.9911、 0.8948、 0.9758、 0.9536 と高かった。これは、 C-measure によって平均的に評価値の高い翻訳を選別することができることを示している。また、参照文を用意しなくても、 C-measure が低い文を集めることによって、 MT システムにとって翻訳が難しい文の集合を自動的に収集できることも示している。 C-measure が高く、かつ、原文と折り返し翻訳文との差異が小さいものを信頼するようにすれば、より信頼性の高い翻訳が得られると考えている。

*http://www.kecl.ntt.co.jp/icl/mtg/resources/index.php

†http://www.nist.gov/speech/tests/mt/resources/scoring.htm

‡最大値が 1 になるように正規化している。

図 1 と図 2 において、BLEU socre と NIST score それぞれに対する相関係数の平均値が最も高かったのは、「依存構造木に基づく単語 n-gram ＋汎化」を採用した場合で、相関係数はそれぞれ、0.9848 と 0.9701 であった。このとき、図 3 と図 4 において、人間の主観評価 (fluency, adequacy) に対する相関係数はそれぞれ、0.9344、0.8999 と高かった。この結果は、機械翻訳可能性の自動評価においては、折り返し翻訳文における単語 n-gram の調和平均よりも精度を重視する方が良いことを示している。

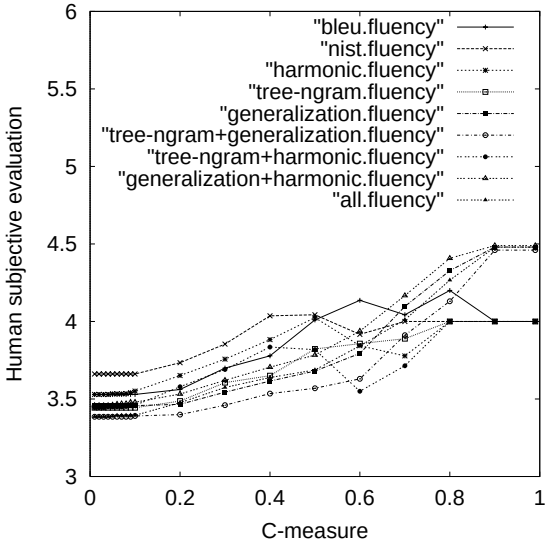

図 3: C-measure と fluency との関係

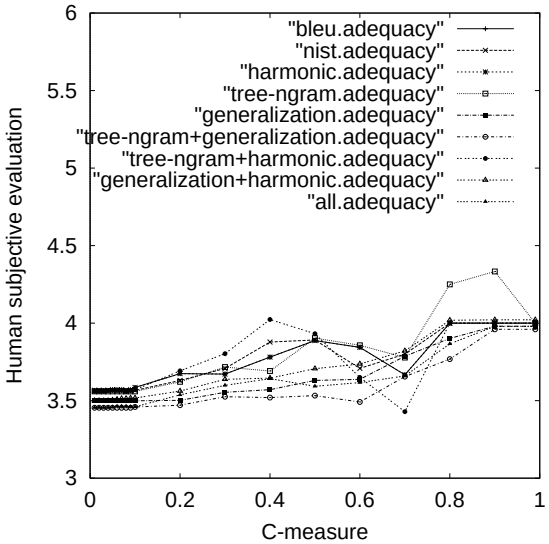

図 4: C-measure と adequacy との関係

今回の実験では、一種類の MT システムを対象としたが、今後、他の MT システムに対しても調べたい。

3 半自動翻訳による翻訳支援

3.1 機械翻訳不適個所の推定

前節で示したように、与えられた文の機械翻訳可能性は C-measure によって評価できそうなことが分かった。したがって、与えられた文の各断片についても同様に C-measure を求められれば機械翻訳可能な部分とそうでない部分を特定できる可能性が高い。そこで、そのような断片として、与えられた文の各部分木を利用する。つまり、与えられた文におけるすべての部分木に対し C-measure を計算することを考える。ここで、与えられた文におけるすべての部分木の集合を SST とし、与えられた文そのものも SST に含まれるものとする。与えられた文の依存構造木は JUMAN [3] と KNP [8] で解析することによって得られ、部分木はその依存構造木から得られる。

与えられた文 s における任意の部分木 $st_i(\in S)$ の確信度スコア $Scr(st_i)$ を次のように定義する。

$$Scr(st_i) \quad = \quad (st_i\text{のCM}) \times \frac{st_i\text{の文節数}}{\text{与えれた文の文節数}} \qquad (5)$$

そして、下記のように確信度スコアが最大となる部分木集合 ST_{best} を求める。

$$ST_{best} \quad = \quad \underset{ST}{\text{argmax}} \sum_{s_i \in ST} Scr(s_i) \qquad (6)$$

ただし、ST は SST の部分集合であり、ST 中の部分木の文節は重ならないものとする。つまり、ST に含まれる部分木を単純に繋ぎ合わせると元の文 s が得られるものとする。複数の部分木が同じ確信度スコアを持つ場合は、最長のものを優先する。ここで、長さは部分木中の文節数と定義する。与えられた文の C-measure がすべての部分木の中で最大となる場合は、与えられた文そのものが ST_{best} として選ばれる。最適な部分木集合は山登り法により探索する。

最適な部分木集合がそれぞれの C-measure とともにユーザに提示される。械翻訳不適個所の候補は、次の手順で推定される。

1. 最適な部分木集合において、部分木の C-measure がすべて閾値よりも低い場合は、全部分木集合の中から C-measure が最低となる部分木を抽出して機械翻訳不適個所の候補としてユーザに提示する。このような場合は、文末の文節が機械翻訳不適個所であるか、主語が欠落しているなど翻訳に必要な情報が不足している場合が多い。複数の部分木が同じ C-measure を持つ場合、最長のものが優先される。すべての部分木の C-measure が同じ場合は、末尾の文節を優先する。
2. 最適な部分木集合に、C-measure が閾値を越える部分木がある場合、その部分木は機械翻訳可能であることが多く、残りの部分木のうち、C-measure が閾値より低いものが機械翻訳不適個所である場合が多い。このような場合、後者を機械翻訳不適個所の候補としてユーザに提示する。さらに、全部分木集合の中に、機械翻訳不適個所の候補を含み、かつ、C-measure が閾値を越える部分木がある場合、その部分木と機械翻訳不適個所の候補との差異の部分が機械翻訳不適個所である場合があるので、参考としてユーザに提示する。

出力例を図 5 に示す。「Partial translation」で示されているのが最適な部分木集合であり、「Check!」で示されているのが機械翻訳不適個所の候補である。実験では、閾値は 0.5 とした。

3.2 実験と考察

機械翻訳不適個所の推定が、原文の機械翻訳可能性を向上させるのに貢献するかどうかを調べる実験を行なった。2.2 節に示した実験で使ったテストセットの先頭の 100 文に対し、最適な部分木集合と機械翻訳不適個所候補を推定して被験者に提示した。その情報をもとに、書き換えるべきであり、かつ、書き換え可能だと判断したものについてのみ原文を書き換えてもらった。

C-meausre としては、2.2 節の図 1 から図 4において相関が高かった「依存構造木に基づく単語 n-gram ＋汎

```
#ORIGINAL: 鉛筆は、２ＢかＨＢを使ってください。
#-------------------------------------------------------------------------
#  原文(部分木)                    折り返し翻訳                    確信度スコア
#-------------------------------------------------------------------------
#---Subtrees-->
２ＢかＨＢを使ってください。        ２ＢまたはＨＢを使ってください。  0.58
ＨＢを使ってください。              ＨＢを使ってください。            0.5
鉛筆は、２Ｂか使ってください。      ２Ｂまたは鉛筆を使います。        0.26
使ってください。                    使ってください。                  0.25
２Ｂか使ってください。              ２Ｂまたは使用。                  0.23
鉛筆は、２ＢかＨＢを使ってください。 鉛筆使用２ＢまたはＨＢ。        0.22
鉛筆は、ＨＢを使ってください。      鉛筆使用ＨＢ。                    0
鉛筆は、使ってください。            鉛筆を使ってください。            0
鉛筆は、                            鉛筆                              0
２Ｂか                              それは２Ｂですか？                0
ＨＢを                              ＨＢの                            0
#<--Subtrees---
#-------------------------------------------------------------------------
#  原文(部分木)                    折り返し翻訳                    C-measure
#-------------------------------------------------------------------------
#---Partial translation-->
２ＢかＨＢを使ってください。        ２ＢまたはＨＢを使ってください。  0.77
［鉛筆は、                          鉛筆                              0 ］
（鉛筆は、２Ｂか使ってください。    ２Ｂまたは鉛筆を使います。        0.26）
#<--Partial translation---
#---Check!-->
［鉛筆は、                          鉛筆                              0 ］
#<--Check!---
EOD
```

図 5: 機械翻訳不適個所の推定の例

化」を採用した。被験者に提示した情報の例を図 5 に示す。図 5 の例では、「鉛筆は、」が候補として示されている。ここで、原文の「鉛筆は、2B か HB を使ってください。」を例えば「2B か HB の鉛筆を使ってください。」に書き換えることができれば、「Use the pencil of 2B or HB.」といったよりよい翻訳を得ることができる。上記 100 文に対する書き換えの後、MT システムで再度翻訳し、翻訳文を評価したところ、BLEU score、NIST score がそれぞれ、0.1739 と 3.3162 から、0.2161 と 3.6674 に向上した。書き換えた文は 43 文であった。書き換え前後で翻訳文の質を下記に示す主観評価により調べたところ、29 文について前より良くなっており、悪くなったものはなかった。この 29 文中、18 文 (62%) については、被験者が修正した個所とシステムが提示した機械翻訳不適個所が重なっていた。他の 11 文については、原文と折り返し翻訳文との差異から主語を追加するべきと判断したものが 2 文あり、残りは、「Partial translation」の折り返し翻訳が不自然であることから修正したと考えられる。書き換えの際には「Partial translation」の折り返し翻訳が参考になった。この結果は、システムの提示した情報が有効に働いていることを示している。主観評価は、同じ 100 文に対し、5 段階で行なった。この主観評価では、各文に対し 1 点 (とても悪い) から 5 点 (とても良い) の点数が評価点として与えられた。3 点以上が理解可能であるとした。主観評価の点数は、書き換え前後で平均値 2.73 点から 3.52 点へと向上した。書き換えた 43 文については、平均値 1.63 点 (合計 70 点) から 3.47 点 (合計 149 点) へと大幅に良くなった。

図 6 は対象言語における出力の例である。翻訳とともに、C-measure や他の翻訳候補が示されている。他の翻訳候補としては、各部分木の翻訳だけでなく、機械翻訳不適個所の推定により得られた部分木集合の各翻訳を単純に組み合わせて生成したものも提示している。例えば、図 6 の [Use 2B or HB.][The pencil] が後者の生成例である。これは、元の翻訳文「The pencil use 2B or HB.」よりは理解しやすくなっている。

次に、英語側の情報をもとに誤っていると思われるところを指摘し、その指摘にしたがって適切に原文の日本語文を書き換えることができるかを調べた。まず、上述と同じ 100 文から原文の C-measure が 0.5 以上のもの 32 文を抽出した。その中から、英語を母国語とする人が、英語としてはおかしいものの、機械翻訳不適個所の候補の英訳に問題がないものを選択し、「Subtrees」で示されている部分木の訳の情報をもとに、問題と思われる文節を指摘した。指摘できたのは、32 文中 7 文であった。その指摘を受けて、日本語を母国語とする被験者が日本語の部分木とその折り返し翻訳の情報のみを参照しながら原文を書き換えるという作業を行なった。その結果、7 文中、1 文については悪くなったが、5 文については評価点が良くなり、理解不能だった文は、7 文中 5 文から 1 文に減った。数は少ないが、指摘した部分についてはうまく修正できることが多い。

```
#ORIGINAL:  The pencil use 2B or HB.
#-------------------------------------------------------------------------
# 原文(部分木)                    翻訳                          C-measure
#-------------------------------------------------------------------------
#---Subtrees-->
1 2 3                            Use 2B or HB.                 0.58
2 3                              Use HB.                       0.5
0 1 3                            2B or use a pencil.           0.26
3                                Use.                          0.25
1 3                              2B or use.                    0.23
0 1 2 3                          The pencil use 2B or HB.      0.22
0 2 3                            The pencil use HB.            0
0 3                              Use a pencil.                 0
0                                The pencil                    0
1                                Is it 2B?                     0
2                                of HB                         0
#<--Subtrees---
#-------------------------------------------------------------------------
# 原文(部分木)                    翻訳                          確信度スコア
#-------------------------------------------------------------------------
#---Partial translation-->
1 2 3                            Use 2B or HB.                 0.77
[0                               The pencil                    0   ]
(0 1 3                           2B or use a pencil.           0.26)
                                 [Use 2B or HB.][The pencil]
#<--Partial translation---
#---Check!-->
[0      The pencil                                             0   ]
#<--Check!---
EOD
```

図 6: 対象言語における他の翻訳候補の出力例

本研究では、C-measure が高ければ高いほど翻訳の信頼性は高くなると仮定している。しかし、直訳を折り返し翻訳した場合、直訳は不自然であるにも関わらず、原文と折り返し翻訳文は類似していることが多い。このような場合、機械翻訳可能性は過大評価されることになる。このような過大評価を避けるために、対象言語における言語モデルを利用する予定である。

4 まとめと今後の課題

本稿では、MT システムと入力文が与えられたときに、その MT システムによって入力文の機械翻訳がどの程度うまく行なえるかを自動評価する方法を提案し、翻訳支援に役立つことを示した。本稿では、単独の MT システムを用いたが、複数の MT システムや翻訳メモリなどを利用することにより機械翻訳可能性自動評価の性能を向上させ、自動でより良い翻訳が得られるようにすることも検討している。また、自動で原文を修正する方法についても検討している。その際、様々な言い換えの技術 [9, 10] が適用可能であろう。

謝辞

ツール開発に協力してくださった豊橋技科大の土屋雅稔氏、京大の木田充洋氏、NICT の竹内和広氏、京大石田研究室の方々、コメントをくださった京大の宇津呂武仁氏に感謝いたします。

参考文献

[1] 野村早恵子, 石田亨, 船越要, 安岡美佳, 山下直美. アジアにおける異文化コラボレーション実験 2002: 機械翻訳を介したソフトウェア開発. 情報処理学会論文誌, Vol. 44, No. 5, 2003.
[2] Kishore Papineni, Salim Roukos, Todd Ward, and Weiing. Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th ACL*, pp. 311–318, 2002.
[3] 黒橋禎夫, 長尾眞. 日本語形態素解析システム JUMAN 使用説明書 Version 3.61. 京都大学大学院情報学研究科, 1999.
[4] 国立国語研究所（編）. 分類語彙表－増補改訂版－. 大日本図書, 2004.
[5] NIST. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Technical report, NIST, 2002.
[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] 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.
[8] 黒橋禎夫. 日本語構文解析システム KNP 使用説明書 Version 2.0b6. 京都大学大学院情報学研究科, 1998.
[9] 乾健太郎, 藤田篤. 言い換え技術に関する研究動向. 自然言語処理, Vol. 11, No. 5, 2004.
[10] 松吉俊, 佐藤理史, 宇津呂武仁. 機能表現「なら」の機械翻訳のための言い換え. 情報処理学会 自然言語処理研究会 NL159-28, pp. 201–208, 2004.