RIBES: 順位相関に基づく翻訳の自動評価法

平尾 努　　磯崎 秀樹　　Kevin Duh　　須藤 克仁　　塚田元
永田 昌明
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

1　はじめに

機械翻訳システムの性能を効率的に向上させていくためには，手間暇がかからない自動評価法が必須である．そして，当然それには人間の評価結果との間に高い相関があることが求められる．現在，機械翻訳の自動評価のデファクトスタンダードはBLEU [Papineni 02]であり，様々な研究で評価指標として利用されている．さらに，評価型ワークショップなどの公式評価指標として用いられることも多い．BLEUは一般的には人間による評価との間の相関が高いといわれており，その計算法も基本的にはリファレンスとシステム翻訳との間のNグラム (通常Nは1〜4) が一致した数を数えるだけで簡単であることがこうして広く用いられる理由であると考える．

しかし，BLEUは4単語までの連続した短い単語列しか評価しないため，システム翻訳の内容がリファレンスと大きく乖離していようともそれがリファレンスに含まれる単語列を局所的に保持しているだけで高いスコアを与える傾向にある．語順が似た言語対に対してはこうした問題は起こりにくいが，語順が大きく異なる言語対では大きな問題となる．

本稿では，語順が大きく異なる言語対を対象とし，こうした問題点を解決するため，システム翻訳とリファレンスとの間で共通して出現する単語の出現順序に着目した新たな自動評価法である RIBES (Rank-based Intuitive Bilingual Evaluation Score) を提案する．

2　Nグラムの一致率に基く自動評価法の問題点

いま，原文 (S) に対して，リファレンス (R) とシステム翻訳 (H1，H2) が以下の通り与えられたとしよう．

S 雨に濡れたので，彼は風邪をひいた．

R He caught a cold because he got soaked in the rain.

H1 He caught a cold because he had gotten wet in the rain.

H2 He got soaked in the rain because he caught a cold.

リファレンスは原文の直訳であり，H1もほぼそれに等しい．一方，H2は,「風邪をひいたので彼は雨に濡れた」という通常では考えることのできない意味であり，原文における因果関係が逆転している．こうした2つの機械翻訳に対し，言語としての流暢さ (fluency) に対するスコアは同等程度でも構わないが，内容としての適切性 (adequacy) は，H1がH2よりも高いスコアをとるべきである．

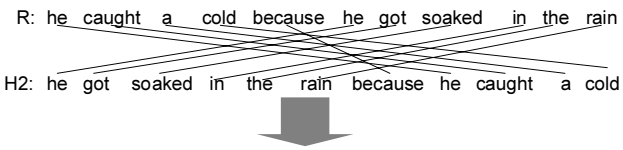
ここで，以下に定義する単一リファレンスの場合のBLEUスコアで2つの翻訳を評価してみよう．

$$\mathrm{BLEU} = \mathrm{BP} \cdot (p_1 p_2 p_3 p_4)^{1/4} \qquad (1)$$

ここで，p_n は N グラム適合率であり，BP は $\min(1, \exp(1 - r/h))$ である．r，h はそれぞれリファレンスとシステム翻訳の単語数を表す．

H1，H2 の BLEU スコアは，それぞれ，0.53，0.74であり，先に述べた直観を正しく反映していない．また，これらシステム翻訳の前半部分と後半部分を入れ替えた英訳として全く不適切な文を評価してもそのBLEUスコアは入れ替え前と大きく変化はしない．この原因は，Nグラム (Nは4以下) という局所的な単語列の一致率にしか着目していないことにある．よって，NISTスコア，METEOR [Banerjee 05] などにも同様の問題がある．

現在の統計的機械翻訳システムでは，計算量の観点から，大幅な語順の入れ替えが難しく，原文の語順を尊重した翻訳が出力される傾向にあるため，H2のような翻訳が出力される可能性は高い．よって，語順が大きく異なる言語対を対象とする場合，Nグラムの一致率で評価すると，局所的には正しい訳であったとしても文全体では正しくない訳に対し高いスコアを与える危険性がある．



$$r = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]$$
$$h = [6, 7, 8, 9, 10, 11, 5, 1, 2, 3, 4]$$

図 1: 単語間の対応付け

3 RIBES (Rank-based Intuitive Bilingual Evaluation Score)

前節で指摘した問題点を解決するため，本稿では，リファレンスとシステム翻訳との間で共通に出現する単語の順序を順位相関係数で評価する RIBES (Rank-based Intuitive Bilingual Score) を提案する．

3.1 単語の対応付け

まず，リファンレスとシステム翻訳の間で共通する単語[1] のみを抽出する．次にリファレンスに含まれる単語に対して，出現した順に 1 から順位を与え，リスト $\boldsymbol{r}$ を得る．システム翻訳に対し，$\boldsymbol{r}$ 中の各要素に対応する単語が何番目に出てきたかを表すリスト $\boldsymbol{h}$ を得る．図 1 の例では，R と H2 では全ての単語 (11 単語) が共通しているため，R からはリスト $\boldsymbol{r} = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]$ を得る．次に $\boldsymbol{r}$ 中の要素に対応する単語が H2 で何番目に出てきたかを表すリスト $\boldsymbol{h} = [6, 7, 8, 9, 10, 11, 5, 1, 2, 3, 4]$ を得る．たとえば，$\boldsymbol{r}$ の 5 番目の要素に対応する単語は「because」であり，H2 において，それは 7 番目に出現する．よって，$\boldsymbol{h}$ の 7 番目の要素は 5 となる．

3.2 順位相関係数

$\boldsymbol{r}, \boldsymbol{h}$ のようなリストが与えられた場合，それらの間の順位相関係数としては，スピアマンの ρ とケンドールの τ を用いることができる．

[1] 1 文に複数回出現する単語については，対応付けに曖昧性が生じる．たとえば，図 1 における「he」については，どの「he」を対応を付けるか曖昧である．このような場合，ユニグラムではなくバイグラムで対応先を絞ってから対応付けを行う．図 1 の例では，「he caught」,「he got」というバイグラムでの対応関係を考えてから単語単位での対応付けを行う．

スピアマンの順位相関係数は以下の式となる．なお，本稿では単語の出現順位に着目しているため，同順位を考慮する必要がないことに注意されたい．

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{2}$$

ここで，n はリストの要素数，$d_i = r_i - h_i$ である．たとえば，$d_5 = 5 - 7 = -2$ である．d の絶対値が大きい場合は，リファレンスとシステム翻訳で語順の変動が大きいことを表し，小さいことは語順の変動が小さいことを表す．図 1 の例におけるスピアマンの順位相関係数は以下の式となる．

$$\rho = 1 - \frac{6(6 \times 5^2 + 2^2 + 4 \times 7^2)}{11^3 - 11} = -0.59 \tag{3}$$

一方，ケンドールの順位相関係数は以下の式となる．

$$\tau = \frac{\sum_{i=1}^{n-1} K_i - \sum_{i=1}^{n-1} L_i}{\frac{n(n-1)}{2}} \tag{4}$$

ここで，K_i は，h_i について $h_i < h_j$ となる場合の数，L_i は $h_i > h_j$ となる場合の数を表す．ただし，$j = i + 1, \ldots, n$ である．たとえば，$h_3 = 8$ なので，$K_3 = 3$，$L_3 = 5$ である．図 1 の例におけるケンドールの順位相関係数は以下の式となる．

$$\tau = \frac{21 - 34}{\frac{11 \times 10}{2}} = -0.23 \tag{5}$$

ρ, τ とも $\boldsymbol{r}$ と $\boldsymbol{h}$ の間の順序が完全一致の場合に $+1$，逆の場合に -1 をとる．このように，R と H2 では語順がマイナス相関にある．一方，同様にして，R と H1 との間の語順の相関係数を計算すると，それらの間で共通に出現する単語の出現順は完全に一致するのでその値は+1 となる．BLEU では H2 が H1 よりも高いスコアであったが，順位相関係数ではその逆となり，人手の評価に近くなる．

このようにリファレンスとシステム翻訳との間で共通に出現する単語の出現順を順位相関係数で評価すると文全体での語順に着目するため， N グラムという局所的な語順にしか着目しない手法よりも良い評価ができる．

ただし，順位相関係数は $[-1, +1]$ の間の値をとる．そこで，従来法と同様，$[0, 1]$ の値をとるように順位相関係数 corr を以下の式で $[0, 1]$ の値に正規化する．

$$\text{Normalized corr} = \frac{\text{corr} + 1}{2} \tag{6}$$

3.3 ペナルティ

先に述べたとおり，順位相関係数はリファレンスとシステム翻訳の間で共通に出現する単語にのみ着目して計算する．よって，それらの間で共通する単語が極端に少ない場合，過剰に高いスコアを与える可能性がある．たとえば，以下の例では，リファレンスとシステム翻訳に共通な単語は「John」,「yesterday」のみで，それらの間に語順の入れ替えはなため，内容が異なる訳であっても，順位相関係数が 1 になってしまうという問題が生じる．

R John went to a restaurant yesterday.

H John read a book yesterday.

これを避けるため，以下に示す，システム翻訳がリファレンス翻訳に含まれる単語を含む割合をペナルティとして導入する．

$$P = n/h \tag{7}$$

n はシステム翻訳とリファレンスとの間で共通な単語の数，n はシステム翻訳の単語の数である．さらに，このペナルティに対する重みパラメータ α $(0 \leq \alpha \leq 1)$ を導入し，RIBES を以下の式で定義する．

$$\text{RIBES(S)} = \text{NSR} \times P^{\alpha} \tag{8}$$

$$\text{RIBES(K)} = \text{NKT} \times P^{\alpha} \tag{9}$$

NSR は，式 (6) で正規化したスピアマンの ρ であり，NKT は式 (6) で正規化したケンドールの τ である．

4 評価実験

4.1 実験に利用したデータと比較した自動評価法

RIBES がどの程度人間の評価との間に相関があるのか，あるいは，従来手法と比較してどの程度相関が高いのかを調べるため，NTCIR-7 の特許翻訳タスク [Fujii 08] の英日翻訳データを用いて評価実験を行った．このタスクにはオーガナイザが提供したベースラインシステムを含む 15 システムが参加しており，うち 2 つがルールに基く翻訳システム，残りが統計翻訳システムである．翻訳課題は 1,381 文あり，リファレンス翻訳が 1 つ用意されている．このうち 100 文に対し，3 名の被験者が言語としての流暢さ (fluency)，内容としての適切さ (adequacy) という 2 つの観点で 5 段階のスコアを付与してある．

各文に対し，3 名が与えたスコアの平均値をシステム毎に平均した値と自動評価法が与えたスコアのシステム毎の平均との間の相関をスピアマンの順位相関係数で評価した．なお，自動評価法の評価指標として，ピアソンの積率相関係数を用いる場合もあるが，本実験ではサンプル数が 15 しかないため，順位相関にのみ着目した．

比較対象として用いた従来の自動評価指標は，ROUGE-L [Lin 04]，IMPACT [Echizen-ya 07]，METEOR [Banerjee 05]，BLEU [Papineni 02] である．

ROUGE-L はリファレンスとシステム翻訳との間の最大共通部分単語列 (LCS) に基く自動評価法であり，IMPACT はそれを改良したものである．これらは，LCS を用いているため，提案手法ほど直接的ではないが，語順を考慮した評価法である．一方，METEOR，BLEU は，先に述べた通り，N グラムの一致率に基く自動評価法である．

4.2 実験結果と考察

実験結果を表 1 に示す．表より，adequacy に関しては，RIBES(S)，RIBES(K) とも，従来法よりもより人間の評価結果に対し相関が高いが，fluency に関しては，ROUGE-L にやや劣る結果となった．一方，自動評価法として広く用いられている BLEU や METEOR は adequacy，fluency ともに相関は非常に低い．この結果は，N グラムという局所的な単語の並びに着目することが，日英のような語順の大きく異なる言語対を対象とした翻訳の評価には不向きであることを示している．一方，ROUGE-L，IMPACT はある程度語順を考慮した評価法であるため，これらより高い相関が得られている．

今回の実験において，RIBES(S) と RIBES(K) を比較すると，adequacy，fluency ともにやや RBIES(K) のがややよい相関を示した．スピアマンの順位相関係数は，単語の入れ替わりを距離として評価し，ケンドールの順位相関係数では直接的な距離ではなく半順序関係で評価するという違いがある．RBIES(S) の方が大きな語順の入れ替わりにより敏感であることが相関係数の差に現れたと考える．ただし，2 つのうちどちらが翻訳の自動評価法として優れているかを調べるためには，今後，他のデータを用いるなどさらに実験を重ねる必要があると考える．

表 1: NTCR-7 日英特許翻訳データにおける人手評価と自動評価との間の相関

	adequacy	fluency
RIBES(K),$\alpha = 0.0$	0.894	0.844
RIBES(K),$\alpha = 0.1$	0.933	0.861
RIBES(K),$\alpha = 0.2$	**0.947**	0.879
RIBES(K),$\alpha = 0.3$	0.940	0.887
RIBES(K),$\alpha = 0.4$	0.929	0.861
RIBES(K),$\alpha = 0.5$	0.926	0.872
RIBES(K),$\alpha = 0.6$	0.922	0.858
RIBES(K),$\alpha = 0.7$	0.919	0.869
RIBES(K),$\alpha = 0.8$	0.919	0.869
RIBES(K),$\alpha = 0.9$	0.908	0.861
RIBES(K),$\alpha = 1.0$	0.879	0.833
RIBES(S),$\alpha = 0.0$	0.747	0.729
RIBES(S),$\alpha = 0.1$	0.854	0.815
RIBES(S),$\alpha = 0.2$	0.883	0.833
RIBES(S),$\alpha = 0.3$	0.887	0.826
RIBES(S),$\alpha = 0.4$	0.915	0.847
RIBES(S),$\alpha = 0.5$	0.926	0.858
RIBES(S),$\alpha = 0.6$	0.922	0.840
RIBES(S),$\alpha = 0.7$	0.894	0.836
RIBES(S),$\alpha = 0.8$	0.894	0.836
RIBES(S),$\alpha = 0.9$	0.829	0.765
RIBES(S),$\alpha = 1.0$	0.797	0.736
ROUGE-L	0.903	**0.889**
IMPACT	0.826	0.751
METEOR	0.490	0.508
BLEU	0.515	0.500

5 まとめと今後の課題

本稿では，日英のような語順が大きく異なる言語対でも人間の評価結果との間に高い相関を持つ新しい自動評価法である RIBES を提案した．RIBES はリファレンスに含まれる単語がシステム翻訳に出現する順に着目し，これを順位相関関係として評価する．さらに，リファレンスとシステム翻訳との間で共通する単語が少ない場合には順位相関係数を過剰に高く与える問題点を解決するため，システム翻訳がリファレンスに含まれる単語を含む割合をペナルティとして用いる．NTCIR-7 の日英特許翻訳タスクを用いて RIBES を評価したところ，人間の評価結果との間の相関は 0.947 であり，翻訳内容の適切性という観点からは，従来の自動評価法と比較して最も高い相関を示した．

今後の課題としては，複数リファレンスが与えられた時の RIBES の拡張がある．一般的には原文に対する正解訳は複数あることが多い．BLEU はもともと複数リファレンスを前提として設計された指標であり，RIBES でもこれに対応することは，より良い自動評価指標を実現するために必須であると考える．

参考文献

[Banerjee 05] Banerjee, S. and Lavie, A.: Meteor: An Automatic metric for MT evaluation with improved correlation with human judgements, in *Proc. of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and Summarization*, pp. 65–72 (2005)

[Echizen-ya 07] Echizen-ya, H. and Araki, K.: Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum, in *Proceedings of MT Summit XII Workshop on Patent Translation*, pp. 151–158 (2007)

[Fujii 08] Fujii, A., Utiyama, M., Yamamoto, M., and Utsuro, T.: Overview of the Patent Translation Task at the NTCIR-7 Workshop, in *Working Notes of the NTCIR Workshop Meeting (NTCIR)*, pp. 389–400 (2008)

[Lin 04] Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries, in *Proceedings of Workshop on Text Summarization Branches Out*, pp. 74–81 (2004)

[Papineni 02] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation, in *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*, pp. 311–318 (2002)