タイトル	機械翻訳システムのための自動評価システムの提案とその性能評価
著者	越前谷, 博; 荒木, 荒木; Echizen ' ya, Hiroshi; Araki, Kenji
引用	工学研究 : 北海学園大学大学院工学研究科紀要(13): 53-59
発行日	2013-09-30

研究論文

機械翻訳システムのための自動評価システムの提案とその性能評価

越前谷　博* ・ 荒木健治**

Proposal of Automatic Evaluation System for Machine Translation Systems and its Effectiveness

Hiroshi Echizen'ya* and Kenji Araki**

概　要

近年，統計翻訳の研究が盛んに行われている．その際，円滑な開発サイクルの妨げとなっているのが評価である．人手評価が最も信頼の高い評価方法ではあるが，コストと時間がかかり，常に用いるのは困難である．このような背景のもと自動評価システムに対するニーズが高まり，様々な手法が提案されるようになった．しかし，これまでに提案されている自動評価システムには問題があり，不十分である．本報告では，従来手法に対して，人手評価との相関が高く，より高い精度で自動評価を行うことが可能な新たな自動評価手法を提案する．提案手法では，参照訳と翻訳文間に存在する共通部分を利用することで，語順を考慮し，かつ，全ての共通部分を評価値に反映した自動評価が可能である．更に，本報告では提案手法に基づく自動評価システムを用いて行った性能評価実験について述べる．

1　はじめに

機械翻訳の分野では統計翻訳[1]の研究が盛んに行われている．多くの研究者が統計翻訳に対して改良と実験を繰り返すことで，統計翻訳の発展に向けた研究を続けている．その際に問題となるのが評価である．機械翻訳システムが出力する翻訳文を評価する場合，人手による評価が最も信頼性の高い評価方法である．しかし，人手評価は時間とコストがかかるため，迅速な評価は困難である．このような人手評価の問題を解決するために自動評価の研究が近年，急速に進んでいる．

自動評価の研究は統計翻訳の研究が活発に行われるようになったことでそのニーズが一層高まった．それに応える自動評価手法として提案されたのがBLEU[2]である．BLEUが提案されたことで，統計翻訳の研究は更に加速した．このBLEUの普及により，現在の自動評価は機械翻訳システムが出力する翻訳文と人手により作成された訳文である参照訳との間の類似性をスコアとして計算し，得られたスコアを自動評価値とする方法が主流となった．BLEUは現在でも最も広く使用されている自動評価システムではあるが，その問題点も従来より指摘されており，BLEUよりも人手評価に近い精度で評価可能とされる手法も数多く提案されている．

このような状況において，我々は従来手法よりも高い精度で自動評価可能な新たな手法を提案する．BLEU，NIST[3]，PER[4]などの自動評価手法は翻訳文と参照訳間に存在する語順の違いに十分に対処できないという問題を抱えている．それに対して，提案手法は語順の違いをスコアに反映する際にパラメータを用いて制御することで対処する．また，WER[5]，METEOR[6]，GTM[7]，ROUGE-L[8]，ROUGE-W[8]，TER[9]などの自動評価手法は

* 北海学園大学大学院工学研究科
Graduate School of Engineering, Hokkai-Gakuen University
** 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

語順の違いには対応可能であるが全ての一致単語をスコアに反映できないという問題を抱えている．この問題を解決するために，提案手法では一致単語の抽出を再帰的に行うことで全ての一致単語をスコアに反映させることが可能である．提案手法に基づく自動評価システムを用いた性能評価実験の結果，人手評価との相関において提案手法が従来手法に比べ，高い相関係数を示すことを確認した．

2　提案手法

本章では，提案する自動評価手法におけるスコアの計算方法について述べる．提案手法では，従来手法と同様に機械翻訳システムの出力文である翻訳文と人手で作成された正しい訳文である参照訳を用いてスコアを計算する．

2.1　共通部分列の決定

はじめに，最長共通部分列（**L**ongest **C**ommon **S**ubsequence：LCS）[10] を求める．その際には，以下の式(1)に基づきダイナミックプログラミング[11] の 2 次元配列を生成する．

$$D_{i,j}=\begin{cases}0, & i=0 \text{ or } j=0\\ max(D_{i-1,j}, D_{i,j-1}), & m_i \neq n_j\\ D_{i-1,j-1}+1, & m_i=n_j\end{cases} \quad (1)$$

例えば，参照訳として〝glass guide of the plastic mounting panel P〞，翻訳文として〝a glass guide molded in panel member P made of resin〞が得られた場合には，ダイナミックプログラミングの 2 次元配列は表 1 のようになる．

表 1 より LCS の値は 4 となる．しかし，この 4 になる過程（以後，LCS の値が得られるまでの過程を LCS 経路と呼ぶ）は 2 通り存在する．表 1 の場合は，下線が付与された値が LCS の値が増加する箇所を示しているが，3 と 4 が 2 箇所ずつ存在するため，LCS 経路が 2 つに分岐している．それぞれの LCS 経路を LCS 経路 No. 1，LCS 経路 No. 2 として以下に示す．

LCS 経路 No. 1

参照訳：[glass guide] of the plastic mounting [panel] [P]

翻訳文：a [glass guide] molded in [panel] member [P] made of the resin

LCS 経路 No. 2

参照訳：[glass guide] [of the] plastic mounting panel P

翻訳文：a [glass guide] molded in panel member P made [of the] resin

〝[〞と〝]〞の間の箇所は共通部分を示している．共通部分とは，一致単語が連続している部分で，かつ参照訳と翻訳文の間で同一に存在している部分である．LCS の値が 4 ということは一致単語の数が 4 であることを意味している．そして，LCS の値が同じであっても上述の例のように LCS 経路は複数存在する場合がある．提案手法では共通部分列を再帰的に決定するために，共通部分を一意に決定する必要がある．そこで，LCS 経路 No. 1 と LCS 経路 No. 2 から，より適切だと思われる LCS 経路を一つのみ決定する．

次いで，LCS 経路が複数存在した場合における一意の決定方法について述べる．上述の例においては，LCS 経路 No. 2 の共通部分〝of the〞は対応関係が成立していない．それに対して，LCS 経路 No. 1 の共通部分〝panel〞と〝P〞は対応関係

表 1：ダイナミックプログラミングの 2 次元配列の例

	i	1	2	3	4	5	6	7	8	9	10	11	12
	m_i	a	glass	guide	molded	in	panel	member	P	made	of	the	resin
j	n_j												
1	glass	0	0	<u>1</u>	1	1	1	1	1	1	2	2	2
2	guide	0	0	1	<u>2</u>	2	2	2	2	2	2	2	2
3	of	0	0	1	2	2	2	2	2	2	<u>3</u>	3	3
4	the	0	0	1	2	2	2	2	2	2	3	<u>4</u>	4
5	plastic	0	0	1	2	2	2	2	2	2	3	4	4
6	mounting	0	0	1	2	2	2	2	2	2	3	4	4
7	panel	0	0	1	2	2	2	<u>3</u>	3	3	3	4	4
8	P	0	0	1	2	2	3	3	<u>4</u>	4	4	4	4

にある．したがって，LCS 経路 No. 1 が選択されることが望ましい．そこで，提案手法では，全ての LCS 経路に対して，以下の式(2)と(3)を用いて score を計算し，score の値が最も大きい LCS 経路を一意に決定する．以下の score は共通部分の長さと位置の情報を用いている．

$$score = \sum_{c \in c\text{-}num} (length(c)^{\beta} \times pos) \tag{2}$$

$$pos = \left(1.0 - \left|\frac{c_i}{m} - \frac{c_j}{n}\right|\right) \tag{3}$$

式(2)の c は共通部分，β は共通部分の長さに基づく重みパラメータであり，1.0 以上の値をとる．式(3)の pos は参照訳と翻訳文の間の共通部分の相対的な位置のずれを意味する．m と n はそれぞれ翻訳文と参照訳の構成単語数である．c_i と c_j は翻訳文と参照訳におけるそれぞれの位置である．式(2)と(3)を用いて score を求めると，パラメータ β の値を 1.2 とした場合，LCS 経路 No. 1 の score は $3.4933 (= 2^{1.2} \times (1.0 - |\frac{2}{12} - \frac{1}{8}|) + 1^{1.2} \times (1.0 - |\frac{6}{12} - \frac{7}{8}|) + 1^{1.2} \times (1.0 - |\frac{8}{12} - \frac{8}{8}|))$，LCS 経路 No. 2 の score は $3.4461 (= 2^{1.2} \times (1.0 - |\frac{2}{12} - \frac{1}{8}|) + 2^{1.2} \times (1.0 - |\frac{10}{12} - \frac{3}{8}|))$ となる．したがって，score の値がより高い LCS 経路 No. 1 が選択され，最適な共通部分列の決定が可能となる．

2.2　スコアの計算方法

2.1 節で述べたように，提案手法では複数の LCS 経路が存在する場合には，式(2)と(3)に基づき一意に LCS 経路を決定する．そして，その LCS 経路より自動評価としての評価値を算出する．その計算式を以下の式(4)と(5)，そして，式(6)に示す．

$$R = \left(\frac{\sum_{i=0}^{RN-1} (\alpha^i \sum_{c \in c\text{-}num} length(c)^{\beta})}{m^{\beta}}\right)^{\frac{1}{\beta}} \tag{4}$$

$$P = \left(\frac{\sum_{i=0}^{RN-1} (\alpha^i \sum_{c \in c\text{-}num} length(c)^{\beta})}{n^{\beta}}\right)^{\frac{1}{\beta}} \tag{5}$$

$$AE\ score = \frac{(1+\gamma^2)RP}{R + \gamma^2 P}. \tag{6}$$

式(4)と(5)の $\sum_{c \in c\text{-}num} length(c)^{\beta}$ は個々の共通部分ごとに得られる値の総和である．したがって，上述した例の場合，共通部分〝glass guide〟，〝panel〟，〝P〟が対象となる．これらの共通部分の値としては，パラメータ β が 2.0 の場合，$6 (= 2^{2.0} + 1^{2.0} + 1^{2.0})$ となる．更に，提案手法では，決定された共通部分を除き，更に共通部分が存在する場合には，改めて LCS 経路を一意に決定し，決定された共通部分を用いて同様の計算を行う．すなわち，共通部分列の決定の再帰処理を行う．上述した例の場合，〝of the〟が改めて共通部分として存在するため，$4 (= 2^{2.0})$ が得られる．このような新たに決定される共通部分は語順が異なる共通部分である．そして，このような語順の異なる共通部分をどの程度，スコアに反映させるかを制御するためにパラメータ α を用いている．式(4)と(5)の α^i の i は再帰処理の回数を示すカウンタである．上述した例では，LCS 経路に基づく共通部分列の決定処理は 2 回行われるため，カウンタ i は 0，1 となる．パラメータ α は 1.0 以下の値を有する．1.0 の場合には，カウンタ i の値が増加しても α^i は 1.0 のままとなる．これは，語順が異なっていても共通部分の重みが変わらないことを意味する．逆に，0.0 に近くなるほど，カウンタ i の値の増加に伴って，α^i の値は小さくなるため，語順が異なる共通部分についてはその値が小さくなることを意味する．このようにパラメータ α は語順の異なる共通部分列に対する重みづけパラメータとして用いられる．また，式(4)，(5)の $RN-1$ は再帰処理の回数を意味している．上述の例では，カウンタ i が 1 になるまで再帰処理が行われるため，$RN-1$ の値は $1 (= 2-1)$ となる．

WER, METEOR, GTM, ROUGE-L, ROUGE-W，TER などの語順を考慮した自動評価手法では，語順が大きく異なる共通部分（例えば，〝of the〟がこれに該当する）は無視され，スコアに反映されない．語順の制約が強い英語などの言語においては，語順の異なる共通部分を無視しても大きな問題とはならないが，日本語などの語順の制約が緩い言語においては，完全に無視することは最適な自動評価の妨げとなることがある．この問題に対して，提案手法では，語順の違いをどこまでスコアに反映させるかをパラメータ α を用いて制御することで対処している．

上述の例の場合，α の値を 0.5 とすると，$\sum_{i=0}^{RN-1} (\alpha^i \sum_{c \in c\text{-}num} length(c)^{\beta})$ の値は $8 (= 0.5^0 \times 6 + 0.5^1 \times 4)$ となる．更に，式(4)と式(5)の値はそれぞれ 0.2357（$= \sqrt{\frac{8}{12^{2.0}}}$）と 0.3536（$= \sqrt{\frac{8}{8^{2.0}}}$）になる．

式(6)の γ は P/R より得られる．上述の例では，γ の値は 0.6666 となる．その結果，式(6)の AE score の値は

$$0.3065\left(=\frac{(1+0.6666^2)\times 0.2357\times 0.3536}{0.2357+0.6666^2\times 0.3536}\right)$$

となる．

3 性能評価実験

3.1 実験データ

実験データには，NTCIR-7[12] の翻訳タスクデータ，WMT10[13]，WMT11[14] の自動評価タスクデータを用いた．NTCIR-7 データでは，14 の機械翻訳システムが日本語文 100 文を英語文 100 文に翻訳した，計 1,400 の翻訳文を用いた．また，参照訳の数は 1 である．人手評価は Adequacy と Fluency の観点より 3 名の評価者が全翻訳文に対して 5 段階で評価したものを用いた． 3 名の評価結果に対してはメジアン値を最終的な人手評価とした．

WMT10 と WMT11 においてはチェコ語，ドイツ語，スペイン語，フランス語から英語に翻訳した文を翻訳文とした．参照訳の数は NTCIR-7 と同様に 1 である．

3.2 実験方法

実験は，全ての翻訳文と参照訳に対して，〝tokenizer.perl〞[14] と 〝lowercase.perl〞[14] を用いて前編集を行った．そして，提案手法に基づく自動評価システムを用いて，人手評価との相関係数を求めた．自動評価システムに対する評価結果としては，システムレベルと文レベルの相関係数を求めた．その際，システムレベルはスピアマンの順位相関係数，文レベルはケンドールの順位相関係数を用いた．

更に，提案手法の有効性を確認するために，比較実験を行った．比較に使用した自動評価手法は BLEU（ver.12），METEOR（ver.1.4），RIBES（ver.1.02.3）[15]，TER（tercom ver 0.7.25）である．

なお，提案手法における自動評価システムのパラメータ α と β の値には，予備実験に基づき 0.1 と 1.2 をそれぞれ用いた．

3.3 実験結果と考察

表 2 に NTCIR-7 を用いたシステムレベルにおけるスピアマンの順位相関係数を示す．表 3 には NTCIR-7 を用いた文レベルにおけるケンドールの順位相関係数を示す．また，表 4 には WMT10 を用いたシステムレベルにおけるスピアマンの順位相関係数，表 5 には WMT10 を用いた文レベルにおけるケンドールの順位相関係数を示す．そして，表 6 には WMT11 を用いたシステムレベルにおけるスピアマンの順位相関係数，表 7 には WMT11 を用いた文レベルにおけるケンドールの順位相関係数を示す．

表 6 と表 7 における〝indiv〞は 1 つの機械翻訳システムより得られた翻訳文が評価対象となっている．〝comb〞は 2 つの機械翻訳システムの組み合わせより得られた翻訳文が評価対象となっている．なお，表 5 と表 7 では，BLEU が存在しないが，BLEU はシステムレベルでの自動評価を前提として提案されている自動評価手法であり，文レベルには適さないことが広く知られていることから，WMT10 と WMT11 においては文レベルの相関係数を求めていない．

表 2 から表 7 において，表 4 の WMT10 を用いたシステムレベルにおけるスピアマンの順位相関係数のみ提案手法が従来手法に比べ低い値となっ

表 2 ：NTCIR-7 を用いたシステムレベルにおけるスピアマンの順位相関係数

Metrics	Adequacy (14 systems)	Fluency (14 systems)	Avg.
提案手法	0.9912	0.9253	**0.9582**
BLEU	0.8505	0.8242	0.8374
METEOR	0.8022	0.7538	0.7780
RIBES	0.9121	0.8374	0.8747
TER	−0.9473	−0.8769	−0.9121

表 3 ：NTCIR-7 を用いた文レベルにおけるケンドールの順位相関係数

Metrics	Adequacy (1,400 sentences)	Fluency (1,400 sentences)	Avg.
提案手法	0.4138	0.3503	**0.3820**
BLEU	0.1146	0.1491	0.1319
METEOR	0.1838	0.2060	0.1949
RIBES	0.3558	0.2950	0.3254
TER	−0.2664	−0.2605	−0.2635

表 4 ：WMT 10 を用いたシステムレベルにおけるスピアマンの順位相関係数

Metrics	cz-en (12 systems)	de-en (25 systems)	es-en (14 systems)	fr-en (24 systems)	Avg.
提案手法	0.6643	0.7115	0.6381	0.5635	0.6443
BLEU	0.7203	0.7885	0.3890	0.6862	**0.6460**
METEOR	0.5594	0.8538	0.4330	0.4957	0.5855
RIBES	0.4895	0.5423	0.6615	0.5200	0.5533
TER	−0.8042	−0.3700	−0.5429	−0.3983	−0.5288

表 5 ：WMT 10 を用いた文レベルにおけるケンドールの順位相関係数

Metrics	cz-en (2,481 sentences)	de-en (5,031 sentences)	es-en (5,289 sentences)	fr-en (3,852 sentences)	Avg.
提案手法	0.0610	0.0553	0.0194	0.0384	**0.0435**
METEOR	0.0711	0.0703	−0.0024	0.0299	0.0422
RIBES	0.0415	0.0394	0.0205	0.0411	0.0356
TER	−0.0700	−0.0209	−0.0036	−0.0412	−0.0339

表 6 ：WMT 11 を用いたシステムレベルにおけるスピアマンの順位相関係数

Metrics	cz-en indiv (8 systems)	de-en indiv (20 systems)	es-en indiv (15 systems)	es-en comb (6 systems)
提案手法	0.9048	0.1722	0.7857	−0.3714
BLEU	0.8333	0.2309	0.8204	−0.1739
METEOR	0.9286	0.5308	0.8321	−0.6000
RIBES	0.8333	0.0406	0.5393	−0.0667
TER	−0.9524	−0.1985	−0.7250	0.8286

Metrics	fr-en indiv (18 systems)	fr-en comb (6 systems)	Avg.
提案手法	0.7750	0.6377	**0.4840**
BLEU	0.7730	−0.1449	0.3898
METEOR	0.7998	0.0857	0.4295
RIBES	0.7337	−0.0857	0.3324
TER	−0.7564	0.0286	−0.2959

たが，その他の全ての相関係数で提案手法が最も高い値を示した．提案手法は表 4 においてのみ最大の相関係数を示さなかったが，BLEU に次いで 2 番目に高い相関係数を示している．このことから今回使用した実験データのほぼ全てにおいて提案手法は従来手法に比べて，高い相関係数を示していることが確認できた．このように提案手法が高い相関係数を示すことができた要因としては，先にも述べたように，語順を考慮し，かつ，全ての共通部分をスコアに反映させることができているためと考えられる．しかし，表 5 と表 7 より提案手法は従来手法との比較では最も高い値を示したが人手評価と相関は非常に弱いものであった．したがって，自動評価システムとしては改良の余地が多分に残されている．文レベルでも高い相関係数を得るためには，語彙，構文などの様々な観点からの類似性をスコアに反映させる必要があると考えられる．

4　まとめ

本報告では，機械翻訳システムのための新たな自動評価手法を提案した．提案手法に基づく自動評価システムを用いた性能評価実験の結果，従来手法に比べて，人手評価との間でより高い相関を示した．これは提案手法の有効性を示すものである．今後は，文レベルにおいてより高い相関係数を得るための改良を行う予定である．

表 7 ：WMT 11 を用いた文レベルにおけるケンドールの順位相関係数

Metrics	cz-en indiv (2,205 sentences)	de-en indiv (4,350 sentences)	es-en indiv (2,687 sentences)	es-en comb (1,792 sentences)
提案手法	0.0199	0.0491	0.0421	−0.0297
METEOR	0.0342	0.0516	0.0844	−0.1006
RIBES	0.0002	−0.0306	0.0306	−0.0046
TER	−0.0289	−0.0435	−0.0412	−0.0270

Metrics	fr-en indiv (3,318 sentences)	fr-en comb (1,285 sentences)	Avg.
提案手法	0.0369	−0.0328	**0.0441**
METEOR	0.0320	−0.1396	−0.0054
RIBES	0.0236	−0.0512	−0.0046
TER	−0.0364	−0.0931	−0.0386

謝辞

性能評価実験で使用した NTCIR-7 データは日本特許翻訳機構（Japio）及び国立情報学研究所（NII）より提供された．ここに記して，感謝の意を表す．

参考文献

［1］P. Koehn. 2010. Statistical Machine Translation. *Cambridge University Press*.

［2］K. Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). pp.311-318.

［3］NIST. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Vo-Occurrence Statistics. http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf.

［4］Keh-Yih Su, Ming-Wen Wu and Jing-Shin Chang. 1992. A New Quantitative Quality Measure for Machine Translation Systems. Proceedings of the 17th International Conference on Computational Linguistics (COLING). pp.433-439.

［5］G. Leusch, N. Ueffing and H. Ney. 2003. A Novel String-to-String Distance Measure With Applications to Machine Translation Evaluation. Proceedings of the 9th Machine Translation Summit (MT Summit). pp.311-318.

［6］A. Lavie and A. Agarwal. 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. Proceedings of the Second Workshop on Statistical Machine Translation. pp.228-231.

［7］P. Turian, L. Shen and I. Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. Proceedings of the 11th Machine Translation Summit (MT Summit). pp.386-393.

［8］Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL). pp.606-613.

［9］M. Snover, B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas (AMTA). pp. 223-231.

[10] D. S. Hirschberg. 1975. A Linear Space Algorithm for Computing Maximal Common Subsequences. *Communications of the ACM*. Volume 10 Issue 6. pp. 341-343.

[11] T. Komori and S. Katagiri. 1992. GPD Training of Dynamic Programming-based Speech Recognizers. *Journal of the Acoustical Society of Japan (E) 13(6)*. pp.341-349.

[12] A. Fujii, M. Utiyama, M. Yamamoto and T. Utsuro. 2008. Overview of the Patent Translation Task at the NTCIR-7 Workshop. Proceedings of the Seventh NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access. pp.389-400.

[13] C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki and O. F. Zaidan. 2010. Findings of

the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. Proceedings of the Join Fifth Workshop on Statistical Machine Translation and Metrics MATR. pp.17–53.

［14］ C. Callison-Burch, P. Koehn, C. Monz and O. F. Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation, Proceedings of the Sixth Workshop on Statistical Machine Translation. Proceedings of the Sixth Workshop on Statistical Machine Translation. pp.22–64.

［15］ H. Isozaki, T. Hirao, K. Duh, K. Sudoh and H. Tsukada. 2010. Automatic Evaluation of Translation Quality for Distant Language Pairs. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp.944–952.