Vol.2014-NL-219 No.2
2014/12/16

スクランブリングを考慮した和訳の自動評価法のNTCIR-9データによる検証

高地 なつめ[1,a)]　磯崎 秀樹[1,b)]

概要：語順を重視した翻訳自動評価手法 RIBES は，BLEU などの従来手法と比べ，日英・英日の機械翻訳の自動評価に適している．しかし，日本語にはスクランブリングがあるので，よい和訳が不当な評価を受けることがある．そこで我々は，参照訳の係り受け木から別の語順を自動生成することによって，この問題を改良する方法を提案し，NTCIR-7 データで文レベルの相関係数を向上させる効果があることを確認した [1]．なお，係り受け解析の文正解率はまだ 60%程度なので，人手で修正を行った．本発表では，この手法の有効性を NTCIR-9 のデータで検証し，やはり文レベルの相関を改良する効果があることを確認した．さらに，係り受け木を修正しなくても，同程度の効果が得られることを確認した．

1. はじめに

統計的機械翻訳 [2] の改善には，翻訳の品質評価が重要な役割を担っている．しかし，人手による品質評価には，時間がかかる．そこで，翻訳の品質評価を自動で行う「翻訳自動評価手法」がいくつも提案されている．中でも，広く用いられるのは，BLEU[3] と呼ばれる翻訳自動評価手法である．BLEU は n-gram によりローカルな語順を評価するが，大局的な語順は無視する傾向があり，日英・英日のような語順が大きく異なる言語間の翻訳では，人手評価との相関が低い [4]．

日英・英日の翻訳評価に適した翻訳自動評価手法として，RIBES[5][6] がある．RIBES は，順位相関係数 Kendall の τ によって，グローバルな語順を評価する．日英・英日特許翻訳での実験結果によれば，**表 1** に示すように，RIBES は，BLEU[3] や NIST などの従来手法に比べて人手評価との相関が高い．

これを受け，日英や英日の翻訳の評価には RIBES が使われるようになっている．しかし，現在の RIBES は「**スクランブリング**」を考慮していない．スクランブリングとは，文の意味は変わらず，語順のみが異なる文を指す．日本語の語順は比較的自由であり，語順を変えても意味は変わらない文が多い．英日翻訳の場合，正解である参照訳の他に，語順の異なる正解が複数存在する可能性が非常に高い．語順の近さを評価する RIBES では，1 つの参照訳のみを評価に利用した場合，本来は高い評価を得るはずの和訳が低い評価を受ける可能性がある．

表 1　RIBES と人手評価のシステムレベルの相関 (Spearman's ρ)

	BLEU	NIST	**RIBES**	備考 $\alpha = 0.25$
NTCIR-7 JE	0.515	0.343	**0.940**	[5], $\beta = 0$
NTCIR-9 JE	-0.042	-0.114	**0.632**	[4], $\beta = 0.1$
NTCIR-9 EJ	-0.029	-0.074	**0.716**	
NTCIR-10 JE	0.31	0.36	**0.88**	[7], $\beta = 0.1$
NTCIR-10 EJ	0.36	0.22	**0.79**	

この問題の解決策として，我々は語順の異なる参照訳を自動生成し，翻訳評価の際に，その複数の参照訳を用いる方法を提案した [1]．**表 2** は，NTCIR-7 データでの実験結果を示す．

2. スクランブリングの自動生成

原文 1 文に対する参照訳を複数用意するために，その参照訳からスクランブリングを自動生成する．スクランブリングの生成には，以下の 3 つの手法を用いる [1]．

- ポストオーダー法
- 格助詞交換法
- 単純格助詞制約を用いた格助詞交換法（提案手法）

いずれも参照訳の係り受け木に基づいて，異なる語順の文を自動生成する．係り受け木に基づく手法を用いる理由は，日本語の語順が係り受け関係に大きく依存しているためである．以下では，この 3 手法を説明する．

2.1 ポストオーダー法

日本語の場合，係り受け関係は語順に大きく影響する．

1 岡山県立大学 (Okayama Prefectural University)
a) cd25017y@cse.oka-pu.ac.jp
b) isozaki@cse.oka-pu.ac.jp

© 2014 Information Processing Society of Japan

表 2 NTCIR-7 英日翻訳データにおける妥当性と RIBES の文レベル相関のスクランブリングによる改善 [1]

Spearman's ρ	単一参照訳	→	提案手法
tsbmt	0.439	→	0.452
Moses	0.607	→	0.670
NTT	0.692	→	0.727
NICT-ATR	0.582	→	0.608
kuro	0.515	→	0.550

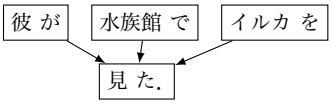

図 1 「彼が水族館でイルカを見た.」の係り受け木

日本語の係り受けでは，係り先となる文節は必ず係り元の文節の後に位置する．つまり，日本語は被修飾語が修飾語より後に現れる語順である．これはヘッド・ファイナルと呼ばれる．ヘッドとは，修飾される語（被修飾語）を指す．

係り受け木から，ヘッド・ファイナルな語順を実現するためには，係り受け木の子ノードを全て出力してから，その親ノードを出力すればよい．これはつまり，係り受け木をポストオーダーで出力することに他ならない．例えば，「彼が水族館でイルカを見た.」という文について考えると，係り受け木は**図 1** のようになる．これをポストオーダーで出力した場合，以下のような 6 文が得られる．

1. 彼が/水族館で/イルカを/見た.
2. 彼が/イルカを/水族館で/見た.
3. 水族館で/彼が/イルカを/見た.
4. 水族館で/イルカを/彼が/見た.
5. イルカを/彼が/水族館で/見た.
6. イルカを/水族館で/彼が/見た.

全て語順は異なるが，同じ意味の文である．つまり，この 6 文全てが，スクランブリングとなる．このように係り受け木をポストオーダーで出力し，その全順列をスクランブリングとする方法を「ポストオーダー法」と呼ぶ．

2.2 格助詞交換法

生成されたスクランブリングは参照訳つまり正解の和文となる．そのため，スクランブリングの生成では，可能な限り，元の参照訳と意味が異なっていたり，日本語として意味が通じない文を生成しないようにすべきである．ポストオーダー法では，このような問題のある文が数多く生成されてしまう．たとえば，「カラスと青い鳥が空を飛んでいた.」という文について考えると，係り受け木は**図 2** のようになる．この文のスクランブリングをポストオーダー法で生成した場合，以下のような 4 文が得られる．

1. カラスと/青い/鳥が/空を/飛んでいた.
2. 空を/カラスと/青い/鳥が/飛んでいた.
3. 空を/青い/カラスと/鳥が/飛んでいた.

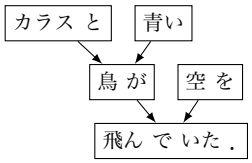

図 2 「カラスと青い鳥が空を飛んでいた.」の係り受け木

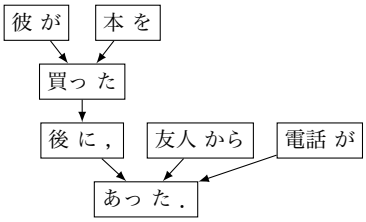

図 3 「彼が本を買った後に，友人から電話があった.」の係り受け木

4. 青い/カラスと/鳥が/空を/飛んでいた.

上の 2 文は意味が同じだが，下の 2 文は「青い」のが「鳥」ではなく「カラス」に変わっている．このような意味の違う文は生成すべきでない．

この場合，名詞を修飾する形容詞を，修飾先ではない名詞の直前に置くから誤解を呼ぶので，形容詞を動かすのは慎重になるべきである．同様に，用言を修飾する副詞を別の用言の前に置くと，やはり誤解を呼ぶであろう．

そこで，「**連続する** 格助詞句の並び替えのみを許可する」という制約を加える．格助詞句とは，格助詞で終わる文節を指す．この制約により，先ほどの「青い」のような形容詞は動かせなくなり，別の名詞を修飾してしまうことがなくなる．同様に用言を修飾する副詞も動かせなくなり，他の用言を修飾してしまうことがなくなるのではないか．

先ほどの「カラスと青い鳥が空を飛んでいた.」の場合，「と」は並列助詞で，格助詞は「が」と「を」だけなので，以下の 2 文だけが得られる．

1. カラスと/青い/鳥**が**/空**を**/飛んでいた.
2. 空**を**/カラスと/青い/鳥**が**/飛んでいた.

この 2 文は意味が変わらない．この制約により，意味の異なる文の生成を大幅に減らすことができると期待できる．このように，ポストオーダ法へ「**連続する** 格助詞句の並び替えのみを許可する」という制約を加えた方法を「格助詞交換法」と呼ぶ．係助詞や接続助詞などは，格助詞の実験結果を見て考える．

2.3 単純格助詞制約を加えた格助詞交換法 (提案手法)

ポストオーダー法によって生成される文に比べ，格助詞交換法により生成される文には，問題のある文が少なくなると期待できる．しかし，格助詞交換法を用いても，問題のある文を生成してしまう例がある．たとえば，「彼が本を買った後に，友人から電話があった.」という文の係り受け木は**図 3** である．この文のスクランブリングを格助詞交換法で生成すると，以下の 12 文が得られる．

1. 彼が/本を/買った/後に, /友人から/電話が/あった.
2. 彼が/本を/買った/後に, /電話が/友人から/あった.
3. 本を/彼が/買った/後に, /友人から/電話が/あった.
4. 本を/彼が/買った/後に, /電話が/友人から/あった.
5. 友人から/電話が/彼が/本を/買った/後に, /あった.
6. 友人から/電話が/本を/彼が/買った/後に, /あった.
7. 電話が/彼が/本を/買った/後に, /友人から/あった.
8. 電話が/本を/彼が/買った/後に, /友人から/あった.
9. 電話が/友人から/彼が/本を/買った/後に, /あった.
10. 電話が/友人から/本を/彼が/買った/後に, /あった.
11. 友人から/彼が/本を/買った/後に, /電話が/あった.
12. 友人から/本を/彼が/買った/後に, /電話が/あった.

1~4 の文は正しいスクランブリングであると言える．それ以降の 5~12 の文は，意味を理解するのに苦労する．

また，9~12 の文は，「本を友人から買った」というような誤解が生まれる文となっている．このような理由から，5~12 の文は正しいスクランブリングではないと判断する．

5~12 の文では，「あった」という動詞に係る「友人から」「電話が」という 2 つの文節が，「買った」という動詞より前に並び替えられている．これにより，「友人から」「電話が」という文節が，「買った」に係るという誤解が生まれる．

このような誤解を生む文を生成しないために，格助詞交換法に「用言を修飾する格助詞句は，先行する用言の前に置いてはならない」という制約を加える．

先ほどの「彼が本を買った後に，友人から電話があった.」という文では，用言「あった」を修飾する格助詞句は，先行する用言「買った」の前に置くべきではない．

この制約の元でスクランブリングを生成すると，以下の 4 文が得られる．

1. 彼が/本を/買った/後に, /友人から/電話が/あった.
2. 彼が/本を/買った/後に, /電話が/友人から/あった.
3. 本を/彼が/買った/後に, /友人から/電話が/あった.
4. 本を/彼が/買った/後に, /電話が/友人から/あった.

この 4 文は同じ意味であると言える．格助詞交換法にこの制約を加えることにより，紛らわしい文の生成を阻止することができた．

この制約は改良の余地がある．この制約では「用言を修飾する格助詞句」が対象だが，形容詞は「を」格を取らない．そのため，形容詞の前に「を」格の格助詞句を並び替えても問題はないと考えられる．たとえば，「美しい花に水をあげた.」という文について考えると，係り受け木は **図 4** のようになる．上記の制約に「を格の格助詞句は，先行する形容詞の前に並び替えても良い」という条件を設けて，この文のスクランブリングを生成した場合，以下のような 2 文が得られる．

1. 美しい/花に/水を/あげた.
2. 水を/美しい/花に/あげた.

この 2 文は正しいスクランブリングであると言える．

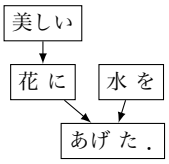

図 4 「美しい花に水をあげた.」の係り受け木

表 3 参照訳の増加による各システムの RIBES スコアの上昇 (NTCIR-7 のデータ)[1]

	妥当性	RIBES			
		単一参照訳	**提案手法**	格助詞交換法	ポストオーダー法
tsbmt	3.5	0.715	0.7188	0.719	0.7569
moses	2.9	0.706	0.7192	0.722	0.781
NTT	2.7	0.671	0.683	0.686	0.7565
NICT-ATR	2.6	0.655	0.664	0.670	0.749
kuro	2.4	0.629	0.638	0.647	0.752

形容詞のように，動詞の中にも特定の格を取らないものがある．たとえば，「咲く」は「を」格を取らない．このような，動詞がどの格を取るかという情報は，日本語語彙大系 [8] の構文体系のような辞書に記載されているが，今回は辞書を利用せず，ある動詞がある格を取らないというような条件を制約に加えることはしない．

そこで，格助詞交換法に以下の「**単純格助詞制約**」を加え，これを「提案手法」という．

- 用言を修飾する格助詞句は，先行する用言の前に置いてはならない．
- ただし，「を」格の格助詞句は，先行する形容詞の前に並び替えても良い．

表 3 に示すように，これらの参照訳自動生成法を用いると，一般に各システムの RIBES スコアが上昇する．しかし，これは改良とは言えない．とくにポストオーダー法は，すべてのシステムのスコアが一様に上がってしまい，人手評価の妥当性との相関が失われてしまう．

3. 実験

本論文では，NTCIR-9[4] のデータにより，提案手法の有用性を検証する．NTCIR-7[9] のデータでは，提案手法を利用した場合，妥当性と RIBES の間の文レベルの相関が，表 2 の通り，人手評価のあるすべての翻訳機で改善された [1].

NTCIR-9 で人手により妥当性が評価された文は 300 文ある．この 300 文を RIBES で評価した場合のスコアと人手による妥当性との文レベル相関が，提案手法によって改善するか否かを調べる．

単一参照訳は，オーガナイザから提供された参照訳であり，この単一参照訳を用いて求めた RIBES スコアは，一般的な RIBES スコアである．つまり，この参照訳を使った場合と，他 3 通りの自動生成された参照訳を使った場合

を比較することによって，上記の手法の有用性の検証が可能となる．

係り受け木は CaboCha [10] などのツールによって自動的に得られるが，文全体の係り受け関係がすべて正しい文の割合（文正解率）は，まだ 50%程度にとどまる [11], [12] ので，人手評価された 300 文を得られた係り受け木の間違いを人手で修正する．この修正により，結果がどのように変化したかも調査する．修正した係り受け木から上記の 3 通りの方法により，スクランブリングを生成する．

本論文では，単一参照訳と自動生成による複数参照訳を用いて，RIBES により評価を行う．評価対象は NTCIR-9 の英日翻訳に参加し，人手評価されたシステム訳である．複数参照訳を用いた場合，全ての訳の RIBES スコアをそれぞれ求め，その中で値が最大となるものをその文の RIBES スコアとした．

今回の実験では，以下の 3 点を検証する．

- 複数参照訳が RIBES に与える影響
- 提案手法により RIBES と妥当性の文レベル相関の改善
- 係り受け木の修正の影響

4. 実験結果

RIBES と妥当性の文レベルの相関を**図 5** に示す．左から，「単一参照訳」「提案手法」「提案手法（係り受け木の修正無し）」をそれぞれ利用した時の RIBES と妥当性の文単位の相関を示す．

係り受け木を修正した場合，単一参照訳と提案手法を比べると，NTT-UT-3，RBMT5，KYOTO-1 以外は提案手法を用いた方が妥当性との相関が高い．

NTT-UT-1 の場合，たとえば以下のような改善例がある．

- 機械訳：(妥当性 5)
  第 1 抵抗 変化 部 2 3 は， ドレイン 領域 2 1 と チャネル 形成 領域 1 2 と の 間 に 設け られ て いる .
- 与えられた参照訳：
  ドレイン 領域 2 1 と チャネル 形成 領域 1 2 と の 間 に 第 1 抵抗 変化 部 2 3 が 設け られ て いる .
  RIBES = 0.701
- 提案手法により生成された 4 つの語順のうちの一つ：
  第 1 抵抗 変化 部 2 3 が ドレイン 領域 2 1 と チャネル 形成 領域 1 2 と の 間 に 設け られ て いる .
  RIBES = 0.979

提案手法は 17 システム中 14 システムで Spearman の ρ が向上している．これは符号検定で $p = 0.013$ であり，5%有意水準で統計的有意差がある．

文レベルの相関の向上が小さいのは，おかしな文が生成されないように，提案手法が語順の入れ替えを極めて限定しているのが原因であろう．61%の文で別の語順が生成されていない．

これより，NTCIR-7 のデータ同様に，NTCIR-9 のデータでも提案手法を利用した場合，妥当性と RIBES 間の文レベルの相関が改善された．

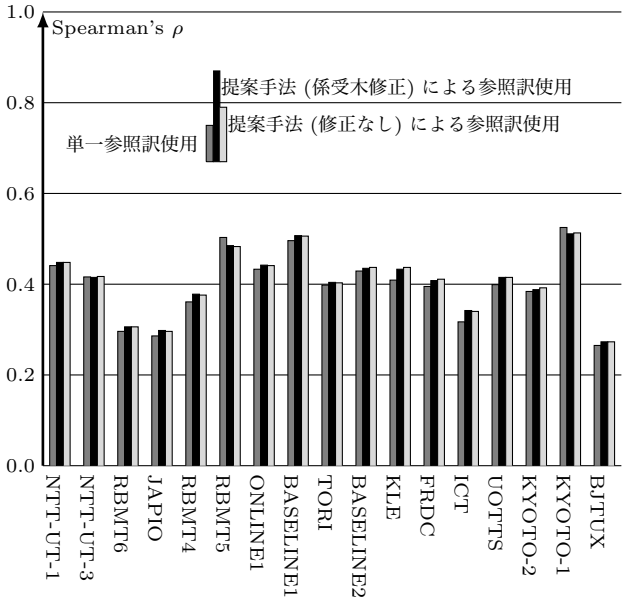

図 5 NTCIR-9 データでの提案手法による妥当性と RIBES の文レベル相関の改善

ポストオーダー法は数十万もの語順を生成する文があり，時間がかかるので，今回の結果には含めていない．

しかし，係り受け木の修正は，文正解率が低いので，かなりの手間になる．そこで，係り受け木を修正しない場合の結果も調べたが，係り受け木を修正した場合とほとんど同じ結果になった．つまり，今回の提案手法に関して言えば，係り受け木の修正には目立った効果はなく，文レベルの相関は同程度に改善されることがわかる．

5. おわりに

本論文では，語順を重視した翻訳自動評価法 RIBES で和訳を評価する場合に発生する，スクランブリングの問題を緩和するため，参照訳から同じ意味の別の語順の文を自動生成する方法について，NTCIR-9 のデータにより検証を行った．

その結果，提案手法により，人手評価による妥当性のスコアと RIBES のスコアの文レベルの相関が若干向上することが確認できた．

スクランブリングの生成には係り受け木を利用した．正しい係り受け関係を利用するために，CaboCha の出力を人手で修正した．しかし，今回の実験では，CaboCha の間違いを修正した効果は確認できなかった．

これにより，NTCIR-7，NTCIR-9 のデータでは，スクランブリングを自動生成し，参照訳として利用する手法が，RIBES と妥当性の文レベルでの相関を改善するということが確認できた．なお，NTCIR-10[7] の各翻訳機の出力は本原稿の執筆時点で未公開であり，NTCIR-10 データでの

検証は今後の課題である．

スクランブリングの生成では，格助詞句のみという狭い範囲での並び替えを行った．今後，不適切な文を生成しないように注意しながら，この制約を緩和していきたい．

本研究では，日本語のスクランブリングに対処するために係り受け解析を利用したが，つい最近，ヨーロッパの言語間の翻訳自動評価においても，係り受け解析の結果を利用して翻訳評価を行う RED, REDp という手法が提案 [13] されており，システムレベルで，人手評価と高い相関が確認されている．この手法との比較は今後の課題である．

参考文献

[1] Isozaki, H., Kouchi, N. and Hirao, T.: Dependency-based Automatic Enumeration of Semantically Equivalent Word Orders for Evaluating Japanese Translations, *Proc. of the Workshop on Statistical Machine Translation*, pp. 287–292 (2014).

[2] Koehn, P.: *Statistical Machine Translation*, Cambridge University Press (2009).

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

[4] Goto, I., Lu, B., Chow, K. P., Sumita, E. and Tsou, B. K.: Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop, *Working Notes of the NTCIR Workshop Meeting (NTCIR)* (2011).

[5] Isozaki, H., Hirao, T., Duh, K., Sudoh, K., Tsukada, H. and Nagata, M.: Automatic Evaluation of Translation Quality for Distant Language Pairs, *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 944–952 (2010).

[6] 平尾　努，磯崎秀樹，Duh, K.，須藤克仁，塚田　元，永田昌明：RIBES: 順位相関に基づく翻訳の自動評価法，言語処理学会年次大会，pp. 1115–1118 (2011).

[7] Goto, I., Chow, K. P., Lu, B., Sumita, E. and Tsou, B. K.: Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop, *Working Notes of the NTCIR Workshop Meeting (NTCIR)* (2013).

[8] 池原　悟，宮崎正弘，白井　諭，横尾昭男，中岩浩巳，大山芳史，林　良彦：日本語語彙大系，岩波書店 (1997). http://www.kecl.ntt.co.jp/mtg/resources/GoiTaikei/.

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

[10] Kudo, T. and Matsumoto, Y.: Japanese Dependency Analysis using Cascaded Chunking, *Proceedings of CoNLL*, pp. 63–69 (2002).

[11] Tamura, A., Takamura, H. and Okumura, M.: Japanese Dependency Analysis Using the Ancestor-Descendant Relation, *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 600–609 (2007).

[12] Sassano, M.: Linear-Time Dependency Analysis for Japanese, *Proc. of the International Conference on Computational Linguistics (COLING)*, pp. 8–14 (2004).

[13] Yu, H., Wu, X., Xie, J., Jiang, W., Liu, Q. and Lin, S.: RED: A Reference Dependency Based MT Evaluation Matric, *Proc. of the International Conference on Computational Linguistics (COLING)*, pp. 2042–2051 (2014).

© 2014 Information Processing Society of Japan