Vol.2016-NL-227 No.8
2016/7/30

和訳の自動評価のための係り受け木の比較

門田 悠一郎[1,a)]　磯崎 秀樹[1,b)]

概要：翻訳の質の良し悪しを自動的に評価する翻訳自動評価手法は、翻訳ソフトの改良に不可欠である。語順の近い欧米言語間では IBM の BLEU [1] が採用されているが、英語と日本語のように語順が大きく入れ替わる言語対では、BLEU と人手評価の相関が低い。磯崎ら [2], [3] は、語順の類似度を重視する **RIBES** を考案した。NTCIR の日英・英日の特許翻訳タスクにより、RIBES は人手評価とのシステムレベル相関が BLEU などの他の翻訳自動評価法に比べて高いことが確認されてる。しかし、RIBES は語順の類似度を重視しているので、日本語の語順の自由度（**スクランブリング**）と相性が悪いことが予想される。そこで Isozaki and Kouchi [4] は、参照訳の単語を並べ替えて、誤解を招かない他の語順を自動生成する「**compDep（係り受け比較法）**」という手法を考案し、生成された他の語順も参照訳として採用することで、RIBES と人手評価の文レベルの相関係数が向上することを示した。本稿では、この手法を日本語で紹介するとともに、同じ手法を WER と IMPACT に適用し、文レベル相関がどう変化するかを示す。

1. はじめに

翻訳ソフトの改良には、翻訳の評価（採点）が必要であるが、人手評価には時間や人件費がかかるので、自動評価が使われる。欧米言語間の翻訳では、IBM の BLEU [1] が使われている。BLEU は、翻訳ソフトが出力した「機械訳」と人手で作成した理想的な訳である「参照訳」の類似度を単語 N グラムの適合率に基づいて採点する手法である。

しかし、日本語と英語のように、語順がまったく異なる言語間の翻訳では、人手評価と BLEU の点数の相関が低い。たとえば、「彼は雨に濡れたので、風邪をひいた。」という日本語を英語に翻訳した場合に、以下のような結果が得られたとしよう。[5]

- 参照訳: He caught a cold because he got soaked in the rain.
- 機械訳 1: He caught a cold becuse he had gotten wet in the rain.
- 機械訳 2: He got soaked in the rain because he caught a cold.

この場合、機械訳 1 は参照訳にそっくりであり、「濡れた」の表現が若干違うだけである。一方、機械訳 2 は参照訳と因果関係が逆転しており、とても忠実な訳とはいえない。

ところが、BLEU は機械訳 1 に 0.53 点、機械訳 2 に 0.74 点を与える。つまり、因果関係が逆転している機械訳 2 の方を高く評価する。これは、BLEU が単語 4 グラムまでしか見ていないため、それより広い範囲で問題が起きていても、無視されるせいである。このような問題のある BLEU を評価に使うと、翻訳ソフトを間違った方向に導くので、よりよい評価手法が望まれる。

	RIBES		BLUE	
	JE	EJ	JE	EJ
NTCIR-7	0.926	0.835	0.588	0.676
NTCIR-9	0.614	0.895	−0.026	−0.032
NTCIR-10	0.88	0.93	0.69	0.76

表 1　人手評価と自動評価のシステムレベル相関（Spearman's ρ））
RIBES は BLEU より高い相関を示す。[3], [6]

そこで磯崎ら [2], [3] は、よりよい評価法として、機械訳と参照訳の語順の類似度に基づく自動評価法 RIBES を提案した。

RIBES は、英日・日英の特許翻訳で、人手評価と高い「システムレベル相関」を示す。「システムレベル相関」とは、同じテストセットに含まれる何百という文を各翻訳システムが翻訳して得られる訳文の、人手評価と自動評価の点数をそれぞれ平均した点数の間の相関である。

NTCIR の特許翻訳タスクで、単一参照訳を用いた場合の BLEU と RIBES の「システムレベル相関」を表 1 に示す。なお、Pearson の積率相関係数は外れ値の影響を受けやすいので、ここでは外れ値の影響を受けにくい Spearman's ρ だけを示している。

人手評価とある自動評価法のシステムレベル相関が高け

[1] 岡山県立大学 (Okayama Prefectural University)
[a)] cd28045y@cse.oka-pu.ac.jp
[b)] isozaki@cse.oka-pu.ac.jp

れば、その自動評価法を用いて、どの翻訳システムがよいかを決めることができる。なお、NTCIR-9 の人手評価には、adequacy と acceptability があるが、ここでは NTCIR-7 でも利用された adequacy を使う。

BLEU [1] は、個々の文の採点をするのではなく、ある翻訳システムが出力した数百の訳文全体の採点を行う。単語 4 グラムの適合率が利用されているので、個々の文の点数を計算しても、多くの場合 0 点になってしまうという問題がある。

しかし、翻訳ソフトの開発時には、文単位で成績の良し悪しを知りたい。RIBES [2], WER [7], IMPACT [8] などの翻訳自動評価法は、文単位で点数を出力するようになっている。文単位で人手評価と自動評価の相関を求めるのが「文レベル相関」である。

しかし、英日翻訳の場合、日本語の語順の自由度のせいで、問題のない和訳が低い点数になってしまうことがある。たとえば、参照訳が「出力部を図 2 に示す。」で、機械訳が「図 2 に出力部を示す。」の場合、RIBES は語順が違うので低い点をつけるが、人手評価では高い点がつく。このように、語順が変わっても、問題がないことがあり、これを「**スクランブリング**」という。

しかし、どんな語順でもいいわけではない。例えば、「中野は聖護院で八つ橋を食べた」を以下のように並べてはいけない。

(1) 「をで食べたは中野八つ橋聖護院」
(2) 「は中野で聖護院を八つ橋食べた」
(3) 「中野は八つ橋で聖護院を食べた」

(1) や (2) は日本語の語順になっていない。(3) は日本語の語順だが、原文と意味が変わっている。つまり、日本語の語順を入れ替える場合、意味が変わらないように注意しなければならない。

Isozaki ら [4], [10] は、参照訳にスクランブリングを適用して、問題のない文を自動生成し、それらを参照訳として追加することによって、RIBES と人手評価の文レベル相関が向上することを示した。

本稿では、新しい「係り受け比較法」を WER や IMPACT に適用して、文レベル相関が向上するか否かを実験によって確認する。RIBES, WER, IMPACT はいずれも文単位でスコアが出せる評価指標であるのに対して、BLEU [1] はそのままの定義では文レベルのスコアを出すのに向かないため、ここでは対象としない。

2. スクランブリングの自動生成

2.1 POSTORDER 法

Isozaki ら [10] は、参照訳を係り受け解析して得られた文節の係り受け木から、日本語の語順を守りながら語順を入れ替えるため、「**POSTORDER 法**」という手法を提案した。

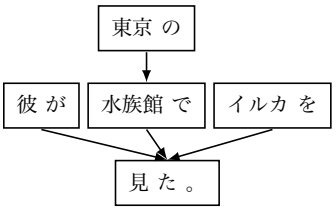

図 1 「彼が東京の水族館でイルカを見た。」の係り受け木

日本語では、修飾語が被修飾語に先行する傾向があり、この傾向は「主辞後置型（head final）」[11] と呼ばれる。POSTORDER 法は、この主語後置性を守りながら文節を並べかえるので、(1) (2) (3) のような文は生成されない。

たとえば、「彼が東京の水族館でイルカを見た。」は、図 1 の係り受け木で表される。POSTORDER 法は、このような係り受け木を post-order で出力する。つまり、木の各ノードは、そのノードの子ノードがすべて出力されたあとにでなければ出力されない。なお、子ノードが複数ある場合、それらをどんな順序で出力してもよい。このようにして生成される文のことを、本稿では一般に「PO 文」と呼ぶ。図 3 は、たったひとつの参照訳を係り受け解析して得られる「木 0」から PO 文が多数生成され、その中で、係り受け解析を行った結果得られる「木 i」が、元の「木 0」と係り受け関係が同じ文だけを採用することを示す。

今の場合、以下の 3!=6 通りの「PO 文」が生成される。

(1) 彼が 東京の 水族館で イルカを 見た。
(2) 彼が イルカを 東京の 水族館で 見た。
(3) 東京の 水族館で 彼が イルカを 見た。
(4) 東京の 水族館で イルカを 彼が 見た。
(5) イルカを 彼が 東京の 水族館で 見た。
(6) イルカを 東京の 水族館で 彼が 見た。

しかし、POSTORDER 法では、以下のような誤解を招く語順が生成されることがある [12]。

- 彼が本を買った後に、友人から電話があった。

→ 友人から彼が本を買った後に、電話があった。

上の文では、「友人から」は「あった。」を修飾しているが、下の文では、「買った」を修飾しているように聞こえるので、誤解を招く。

上の文は図 2 の係り受け木を持っている。「あった。」の 3 つの子ノード「後に、」「友人から」「電話が」を「友人から」「後に、」「電話が」の順で post-oreder 出力すると、→ の文が得られる。

このように、POSTORDER 法は誤解を招く文を生成することがあり、これをいかにして排除するかが問題となる。

そこで Isozaki ら [10], [12] は、経験則による制限を加えた。しかし、厳しい制限を加えたせいで、多くの文で別の語順がまったく出力されない、という問題が発生した。

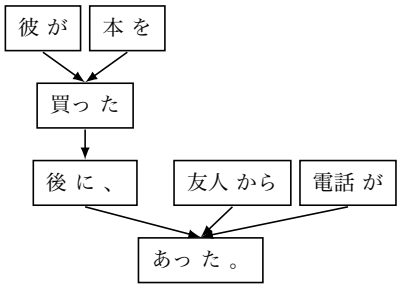

図 2 「彼が本を買った後に、友人から電話があった。」の係り受け木

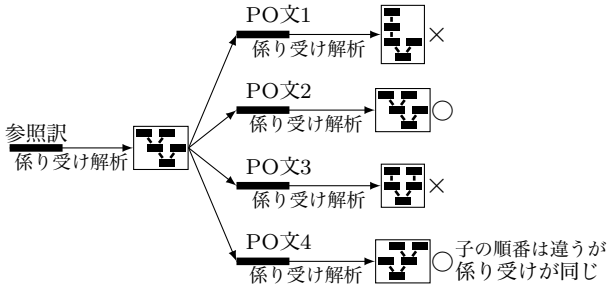

図 3 係り受け比較法によるスクランブリングの自動生成法。「PO文i」を係り受け解析して得られる「木i」の係り受け関係が元の「木0」と同じときだけ採用

2.2 係り受け比較法

POSTORDER 法で生成される、誤解を招く文を除去するため、Isozaki and Kouchi [4] は、**compDep** という手法を考案した。本稿ではこれを「**係り受け比較法**」と呼ぶ。

係り受け比較法は「人間が誤解するなら、係り受け解析器も誤解するだろう。」という単純なアイデアに基づく。

先ほどの電話の例の場合、原文では「友人から」が「あった。」を修飾しているが、生成された文では「買った」を修飾しているように見える。それは、「買った」が「あった。」より先に出現しているからである。これは人間に限らず、CaboCha のような係り受け解析器でも同じはずである。実際に `cabocha -f3` が出力した係り受け木を図 4 に示す。案の定 CaboCha も、「友人から」が「買った」を修飾していると誤解したことがわかる。

「係り受け比較法」はこのように、POSTORDER 法で自動生成された文を係り受け解析し、文節間の係り受け関係が元の文の係り受け木と同じになっているかどうかを確認する。係り受け関係が同じであれば、その文をスクランブリングとして認め、新しい参照訳として採用する。係り受け関係が違っていれば、誤解を招く文と見なし、採用しない。

3. 実験結果

前述の「係り受け比較法」によって生成されたスクランブリングをすべて参照訳として採点を行うと、人手評価 (HUMAN) と RIBES の文レベルの相関係数が向上することが明らかになった [4]。この実験では、NTCIR-9 英日特許翻訳 [13] に参加した 17 システムのデータを利用した。

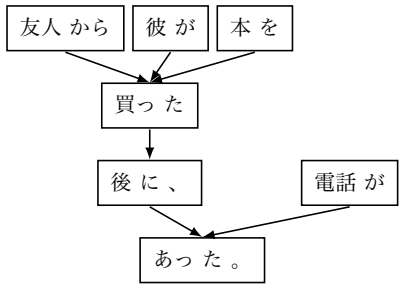

図 4 「友人から彼が本を買った後に、電話があった。」の係り受け木

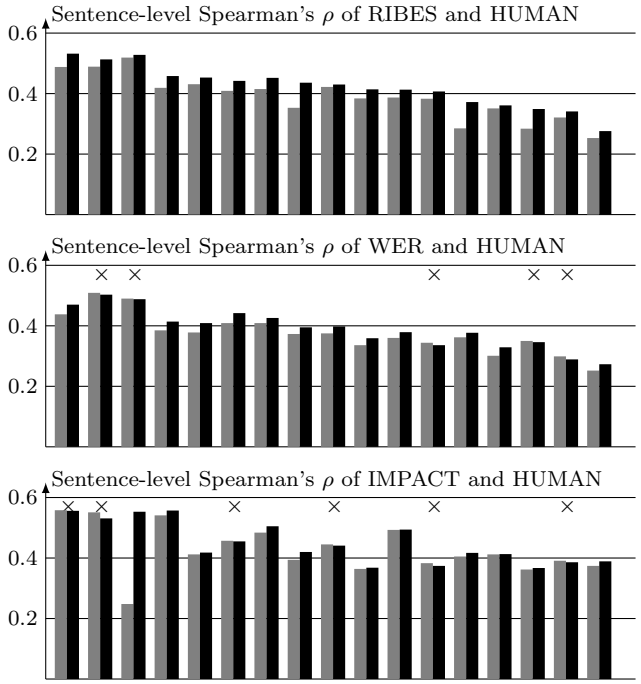

図 5 係り受け比較法によって生成されたスクランブリングを参照訳として採点した場合（黒い棒）の各翻訳自動評価法と人手評価の文レベル相関の変化。× は単一参照訳（灰色の棒）より悪くなったことを示す。(NTCIR-9 英日特許翻訳の 17 システム)

本稿では、同じ手法を WER [7] と IMPACT [8] にも適用して、これら 3 つの翻訳自動評価法の文レベルの相関係数がどう変化するかを調べた。その結果を図 5 に示す。

- RIBES は 17 システム中全 17 システムで文レベルの相関係数が向上した。符号検定によれば、$p = 0.00001526$ で、統計的有意差がある。
- WER は 17 システム中 12 システムで文レベルの相関係数が向上した。符号検定によれば、$p = 0.1435$ で、統計的有意差はない。
- IMPACT は 17 位ステム中 11 システムで文レベルの相関係数が向上した。符号検定によれば、$p = 0.3323$ で、統計的有意差はない。なお、IMPACT は RIBES や WER に比べて相関係数がもともと高い。

つまり、「係り受け比較法」で生成される参照訳を追加することによって得られる文レベル相関の向上が、統計的に有意な差となるのは RIBES だけであり、WER や IMPACT

でも改善するものの、今回用いた NTCIR-9 の英日翻訳のデータだけでは、統計的有意差がない。

これは、RIBES が語順を直接測定する評価指標であるのに対して、WER や IMPACT は直接語順を測定していないので、影響を受けにくいからだと考えられる。

しかし、いずれの自動評価法も、文レベルの相関係数は 0.6 未満であり、相関が強いとはいえない。まだ改良の余地がある。

4. おわりに

「係り受け比較法」は、多くの文に対して別の語順を生成でき、人手評価との文レベルの相関係数が向上することが明らかになった。

本稿では、係り受け比較法を WER や IMPACT に適用しても、人手評価との文レベル相関が RIBES ほど向上しないことを示し、その原因を解明した。

なお、本稿の係り受け木はすべて、`cabocha -f3` が出力する XML を `cabochavees` 環境 [*1] で自動描画したものである。

「係り受け比較法」を実装してくれた高地なつめ氏に感謝します。

参考文献

[1] 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).

[2] 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).

[3] 平尾　努，磯崎秀樹，須藤克仁，Kevin, D.，塚田　元，永田昌明：語順の相関に基づく機械翻訳の自動評価法，言語処理学会年次大会，Vol. 21, No. 3, pp. 421–444 (2014).

[4] Isozaki, H. and Kouchi, N.: Dependency Analysis of Scrambled References for Better Evaluation of Japanese Translation, *Proc. of the Workshop on Statistical Machine Translation*, pp. 450–456 (2015).

[5] 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).

[6] 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).

[7] Su, K.-Y., Wu, M.-W. and Chang, J.-S.: A New Quantitative Quality Measure for Machine Translation Systems, *Proc. of the International Conference on Computational Linguistics (COLING)*, pp. 433–439 (1992).

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

[9] Lin, C.-Y. and Och, F. J.: Automatic Evaluation of Translation Quality Using Longest Common Subsequences and Skip-Bigram Statistics, *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*, pp. 605–612 (2004).

[10] 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).

[11] Isozaki, H., Sudoh, K., Tsukada, H. and Duh, K.: Head Finalization: A Simple Reordering Rule for SOV Languages, *Proc. of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pp. 250–257 (2010).

[12] 高地なつめ，磯崎秀樹：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，情報処理学会研究報告 NL，2, Vol. 219, pp. 1–5 (2014).

[13] 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).

[*1] `http://softcream.oka-pu.ac.jp/isozaki/TeXmac/cabochatrees.pdf`