翻訳自動評価法
翻訳の質を推定する技術の進化

磯崎　秀樹
岡山県立大学

2016 年 11 月 25 日 第 4 回特許情報シンポジウム

翻訳ソフトの改良と人手評価

翻訳ソフトの作成では、出力を見て改善する作業をくりかえす。

しかし、よかれと思ってやった変更に副作用があり、全体的にはむしろ悪くなっていることがある。

そこで、新しい訳が古い訳よりも本当にいいか確認する必要がある。

色んな文を訳して、その訳を人間が見て採点すればよい。

これを「**人手評価**」といい、以下の 2 つの評価尺度が有名。

- **妥当性** (**adequacy**): 訳が原文にどれくらい忠実かの評価。
- **流暢性** (**fluency**): 訳がどれくらい流暢かの評価。

(これらは信頼性が低く、近年の人手評価では好まれない。)

自動評価の必要性

翻訳ソフトの出力する何千という文を、人手で評価するのは大変。
人件費も時間もかかる。

そこで、「**自動評価**」が考案された。

あらかじめ、人間が理想的な訳「**参照訳**」を作っておく。

翻訳ソフトの出力した訳「**機械訳**」と参照訳の**類似度を計算**。

世界で標準的に用いられている類似度は、IBM が提案した **BLEU** (BiLingual Evaluation Understudy))。

初期の自動評価法

BLEU より前には、音声認識分野で使われいる **WER** (Word Error Rate) が使われていた。

WER は、機械訳を書き換えて参照訳にするのに必要な、
単語の追加・削除・置換の操作の回数に基づく尺度。

機械訳が参照訳と同じなら WER $= 0.0$ で、
違うほど WER が大きいので、$1 -$ WER を類似度とする。

しかし、語順の近い欧米言語間では、逐語訳でもかなりわかり、
WER は語順の間違いに厳しすぎると批判された。

そこで、語順の間違いを大目に見る尺度が求められた。

PER と TER

これらは、WER をベースにして、語順の違いに甘くしたもの。

- **PER** (Position-independent word Error Rate):
  文を単語の集合とみなすことで、語順を完全に無視。

- **TER** (Translation Edit Rate): (Sover et al. 2006)
  複数語からなるフレーズを一度に動かすのを 1 操作とみなし、
  語順の違いを大目に見る。

磯崎が語順に関する論文を国際会議に投稿したとき、
「語順など重要な問題ではない」と書いた査読者がいる。

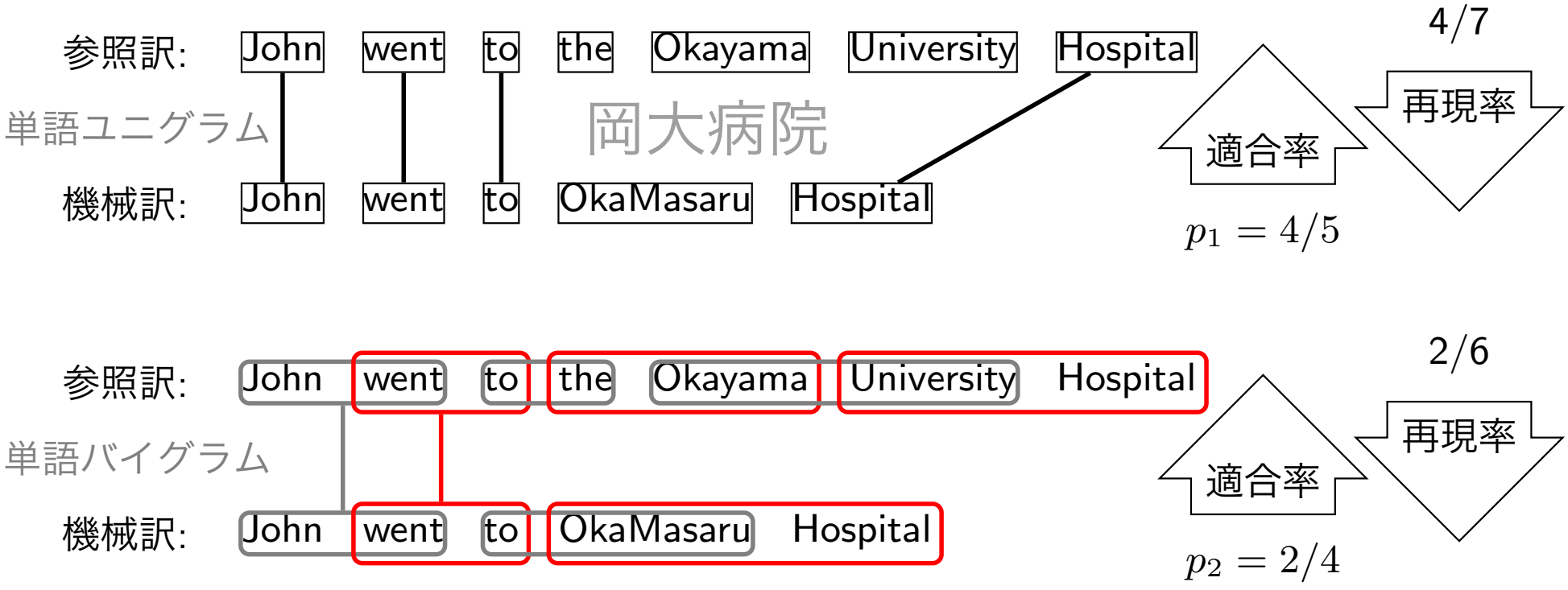
英語の語順で苦労している日本人には信じがたいが、
これが欧米の一部の研究者の認識らしい。

BLEU (単語 n グラム適合率)

Papineni et al. 2002

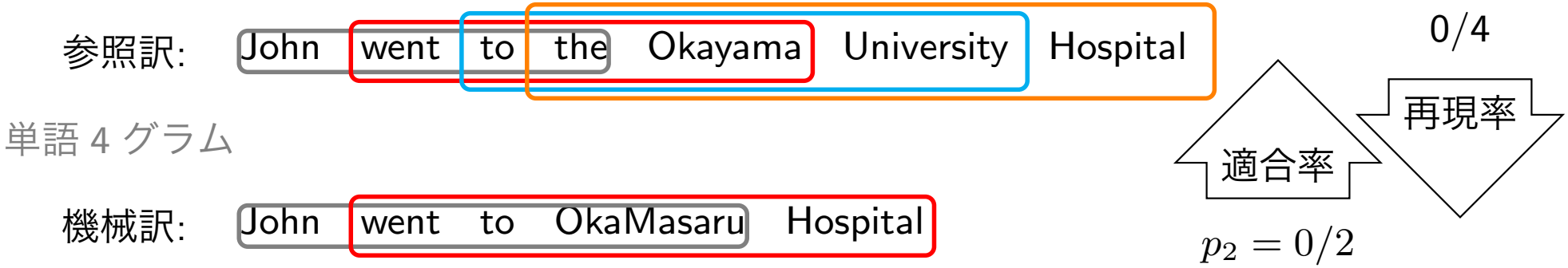
機械訳と参照訳の間で、共通している単語やフレーズが多いほど、よい訳である、という考え方にもとづく。

単語 n グラムの適合率を p_n で表し、p_1 から p_4 の相乗平均 $\sqrt[4]{p_1 p_2 p_3 p_4}$ をベースにしている。



BLEU (複数参照訳)

p_4 は、共通している単語 4 グラムが存在しないと 0 点なので、
BLEU $= \sqrt[4]{p_1 p_2 p_3 p_4}$ も 0 点になってしまう。



参照訳が一つだけだと、BLEU が 0 点の文が多くなる。

そこで BLEU を使うときは、なるべく 0 点にならないように、
参照訳をいくつも用意しなければならない。

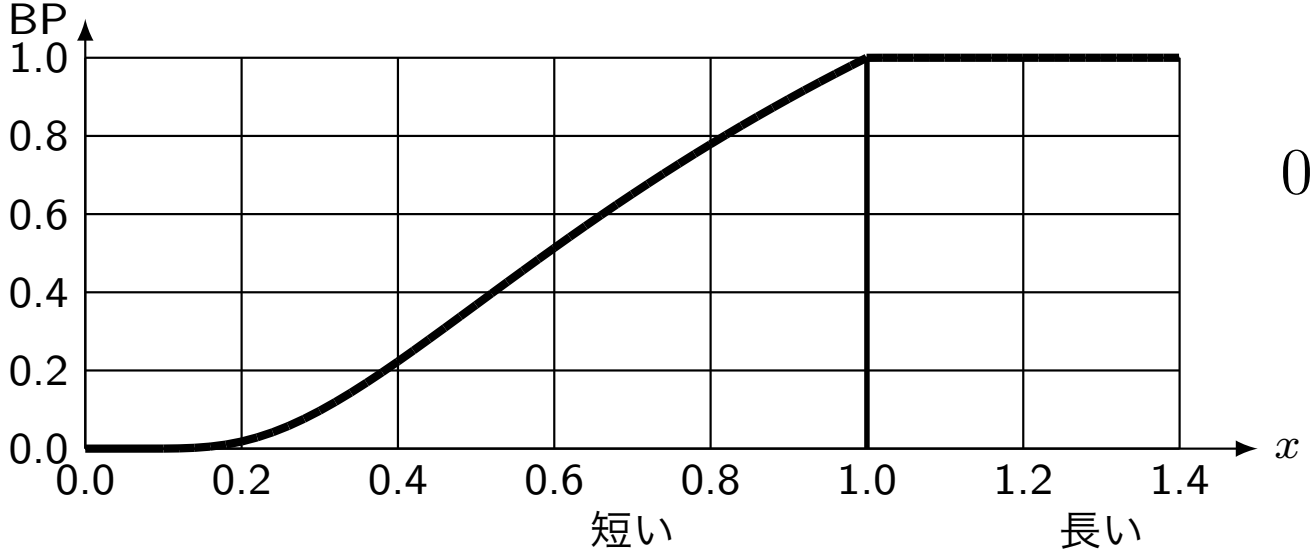
BLEU (Brevity Penalty)

複数参照訳のせいで、BLEU は再現率を使ってないので、
自信のない部分を出力しなければ適合率が上がる。

これを防ぐため、BLEU では、短い機械訳にペナルティを与える。

文の長さの比「機械訳の長さ/参照訳の長さ」を x とすると、以下の **BP** (Brevity Penalty) を掛けることで、短すぎる訳の点数を下げる。

$$\mathrm{BP}(x) \stackrel{\mathrm{def}}{=} \min\left(1, \exp\left(1 - 1/x\right)\right), \qquad \mathrm{BLEU} = \mathrm{BP} \times \sqrt[4]{p_1 p_2 p_3 p_4}$$



$$0.0 \leq \mathrm{BLEU} \leq 1.0$$

BLEU (人手評価との相関の低さ)

NTCIR 特許翻訳タスクの実験結果によると、
英日翻訳や日英翻訳では、BLEU と人手評価の相関が低い。
NTCIR-7 の日英翻訳タスクの場合、
BLEU と人手評価の順位相関 (Spearman's ρ) は 0.5 程度しかない。

BLEU が大局的な語順を考慮していないのが原因。

BLEU は、因果関係が逆の訳などに高い点数を与えることがある。

原文	彼は雨に濡れたので、風邪を引いた。	妥当性	BLEU
参照訳	He caught a cold because he got soaked in the rain.	○	1.00
機械訳1	He caught a cold because he had gotten wet in the rain.	○	**0.53**
機械訳2	He got soaked in the rain because he caught a cold.	×	**0.74**

RIBES(人手評価との相関の高さ)

Isozaki et al. EMNLP-2010, 平尾ら 2011

そこで磯崎は、大局的な語順を考慮した **RIBES** を提案。
RIBES は人手評価と相関が高い。

NTCIR-7 日英翻訳における人手評価と自動評価の順位相関
(Spearman's ρ)

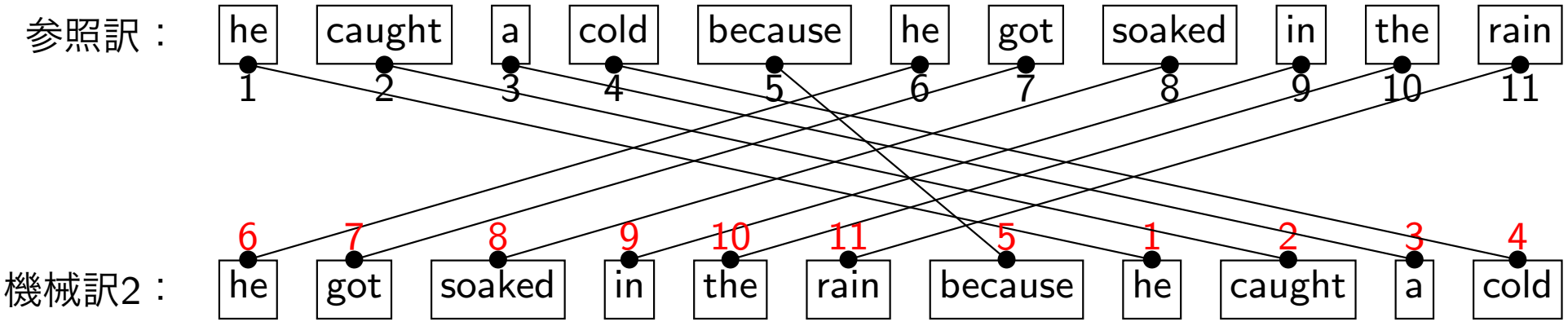
自動評価法	妥当性	流暢性
RIBES	**0.947**	**0.879**
BLEU	0.515	0.500

原文	彼は雨に濡れたので、風邪を引いた。	妥当性	BLEU	**RIBES**
参照訳	He caught a cold because he got soaked in the rain.	○	1.00	1.00
機械訳1	He caught a cold because he had gotten wet in the rain.	○	0.53	**0.93**
機械訳2	He got soaked in the rain because he caught a cold.	×	0.74	**0.38**

RIBES(大局的語順の考慮)

RIBES は語順の近さをベースとした類似度。

語順の近さは、Kendall's τ という順位相関係数で測定。

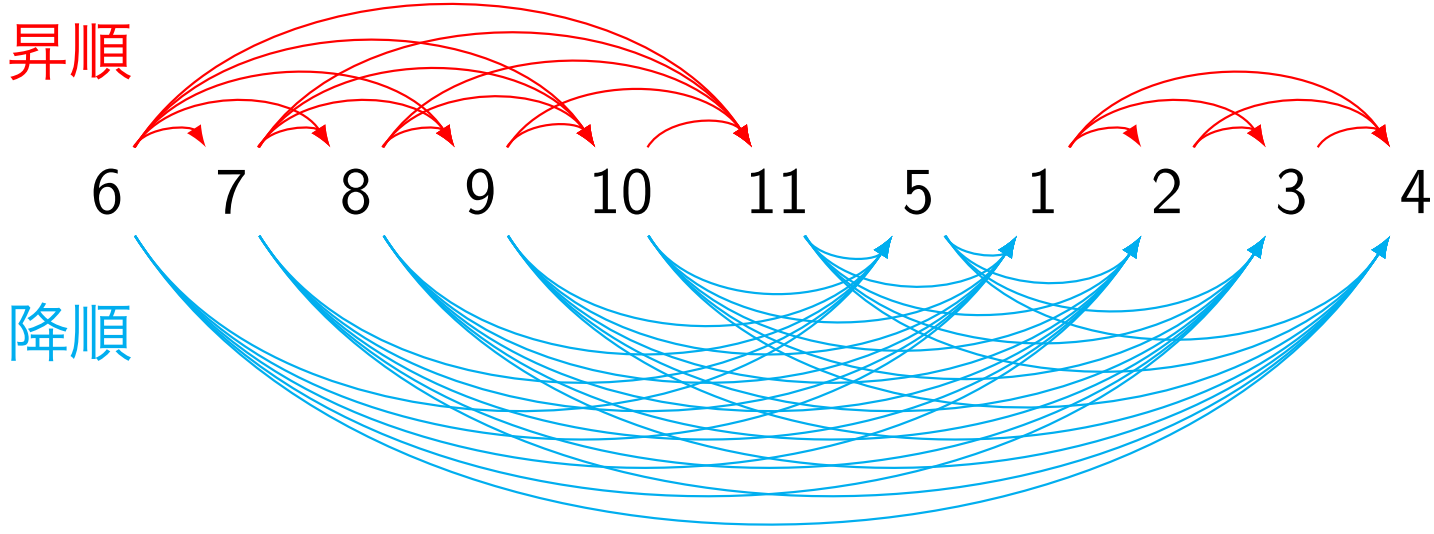


この機械訳の語順は、[6,7,8,9,10,11,5,1,2,3,4] という整数のリストで表せる。

RIBES(NKT)

整数リストから、整数を 2 つ取り出してできるペアのうち、昇順ペアの割合を NKT (Normalized Kendall's τ) と呼ぶ。

この場合、要素が 11 個あるので、${}_{11}C_2 = 55$ ペアあるはず。



昇順なのは、6〜11 の部分の ${}_6C_2 = 15$ ペアと、1〜4 の部分の ${}_4C_2 = 6$ ペアの合計 21 ペアなので、NKT $= 21/55 = 0.38$。

NKT を日英翻訳の自動評価に使ってみると、
人手評価と高い相関があることが判明。

RIBES(ペナルティー)

しかし、参照訳と機械訳の間で共通する単語 (i.e., p_1) が少ないと、NKT は過大 (過少) 評価する、という弱点がある。

そこで、単語適合率 p_1 の α 乗をペナルティとして掛ける。

しかし、適合率は自信のあるところだけを出すという方法で上げられる。そこで、BP の β 乗を掛ける。

以上により、RIBES は以下の式により定義される。

$$\text{RIBES} \stackrel{\text{def}}{=} \text{NKT} \times P^{\alpha} \times \text{BP}^{\beta}$$

デフォルトでは $\alpha = 0.25$, $\beta = 0.1$。

RIBES(スクランブリング)

RIBES は語順の類似度によって訳を評価。

しかし、日本語の語順は、比較的自由。
以下の 2 つの文は同じ意味で、どちらでもよい。

❶ 彼が東京の水族館でイルカを見た。

❷ 東京の水族館で彼がイルカを見た。

RIBES は語順を重視するので、❶を参照訳、❷を機械訳として採点すると、悪い点になり、人手評価とずれる。

これは係り受け解析で解決できそうだが、機械訳が係り受け解析できるほど質が高いとは限らない。

また、どんな語順でもいいわけではない。次の文は意味が変わる。

❸ イルカは水族館で太郎を見た。

RIBES(スクランブリング)

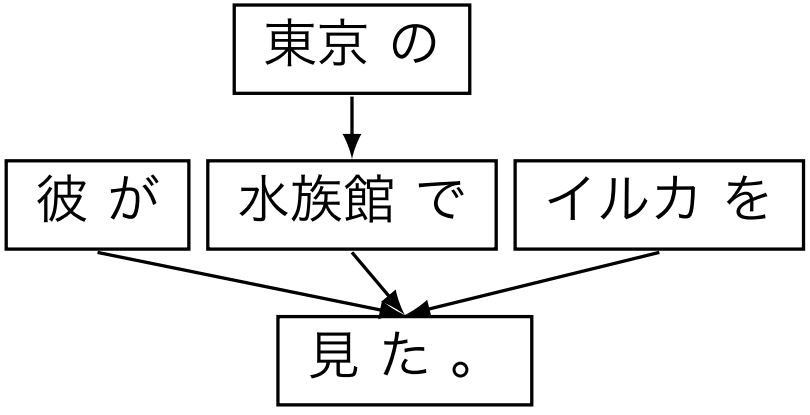
スクランブリングに対応するため、参照訳を自動で増やす。

日本語は head-final と呼ばれる語順。

修飾する表現を先に出して、修飾される表現 (head) を後で出す。

これは、係り受け木を postorder で出力することに他ならない。

兄弟ノードをどの順で出力するか、という自由度があり、語順は一意に定まらない。

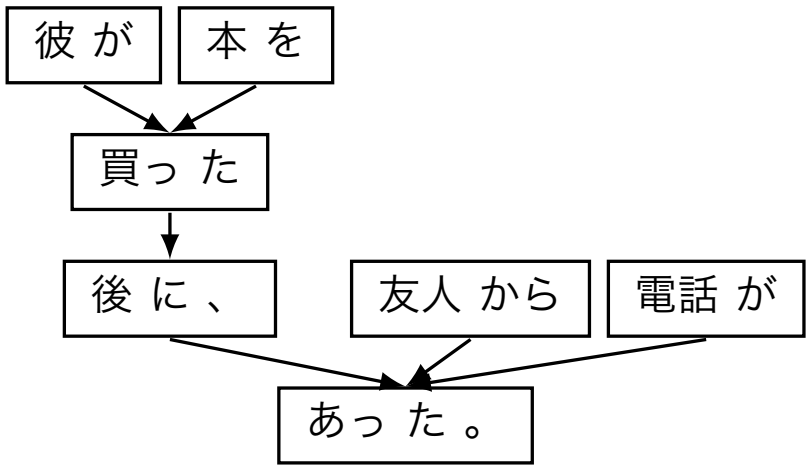


- 彼が東京の水族館でイルカを見た。
- 彼がイルカを東京の水族館で見た。
- 東京の水族館で彼がイルカを見た。
- 東京の水族館でイルカを彼が見た。
- イルカを彼が東京の水族館で見た。
- イルカを東京の水族館で彼が見た。

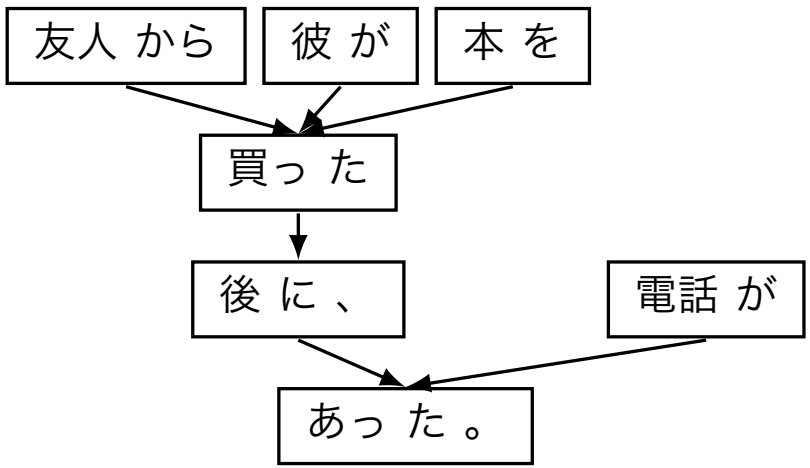
RIBES(スクランブリング)

postorder で出力された文の中には、誤解を招くものがある。

参照訳: **彼が本を買った後に、友人から電話があった。**



PO 文: **友人から彼が本を買った後に、電話があった。**

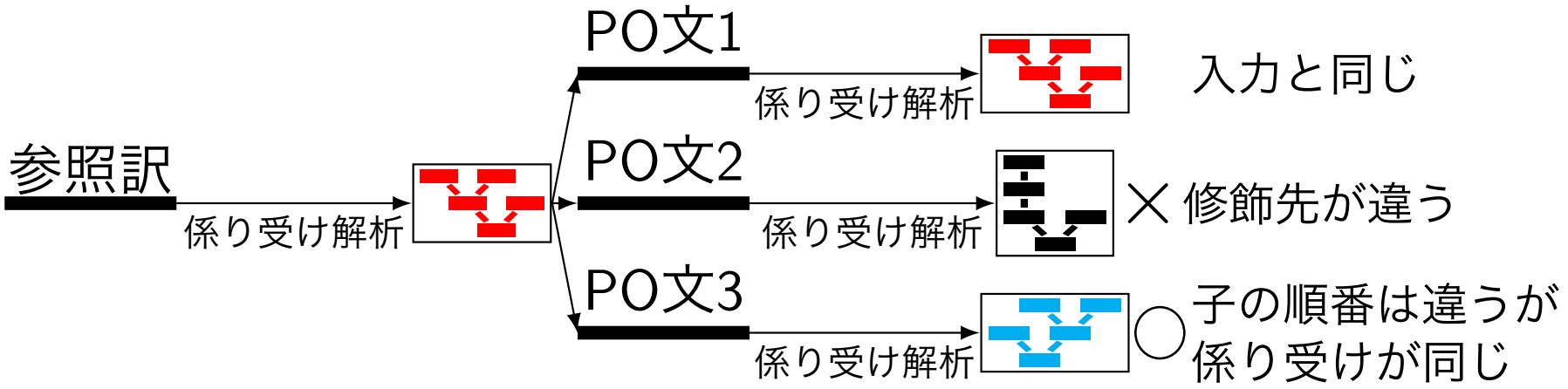


修飾先が変わるのは NG

RIBES(スクランブリング)

Isozaki and Kouchi WMT-2015 は機械訳を係り受け解析せず、
参照訳を自動的に増やしてスクランブリングに対応。

1. 係り受け解析器を使い、参照訳の係り受け木を作成。
2. この木を **postorder で出力**し、スクランブルされた文を生成。
3. 自動生成された文を係り受け解析して**修飾先の変わる文を除去。**
   残った文を参照訳に加える。



RIBESによる評価と人手評価の文レベル相関が向上。

最近の翻訳自動評価の動向

ここまで、英日・日英翻訳の翻訳自動評価の話をしてきた。

欧米にも BLEU に対する不満の声はあり、国際会議

WMT $\left\|\begin{array}{l}\text{Workshop on Statistical Machine Translation}\\ \longrightarrow \text{Conference on Machine Translation}\end{array}\right\|$

などで、新しい翻訳自動評価の手法が活発に研究されている。

RIBES と同じ 2010 年に、イギリスで LRscore という Kendall's τ をベースにした自動評価法が独立に提案されているが、語順の近い中英で実験しているので、RIBES ほどの効果はない。

2010〜2012 年の欧米での翻訳自動評価法は、以下の資料を参照。

磯崎：最近の自動評価法の研究動向と RIBES、AAMT/Japio 特許翻訳研究会、特許文書の機械翻訳結果評価方法検討会資料集, 2012.

`http://aamtjapio.com/kenkyu/files/discussion01/AAMT_Japio_discus(20120907)-02.pdf`

最近の翻訳自動評価の動向

他の研究分野同様、翻訳自動評価の分野でも、最近はニューラルネットが利用されている。

- Neubig et al. WAT-2015: ニューラルネットを使ったリランキングで、文法エラーが減少。
- Shah et al. WMT-2016: 参照訳なしで訳文を評価する WMT QE タスクで、ニューラルネットのバイリンガルな素性を利用。
- Kim et al. WMT-2016: 単語が正しく訳されているかを判定する RNN と、最終的な品質を判定する RNN で評価。

まとめと今後の課題

- 欧米で標準的に用いられている BLEU という翻訳自動評価法は、日英・英日翻訳では人手評価と相関が低い。
- そこで、語順の近さに注目した自動評価法 RIBES を提案した。RIBES は人手評価と相関が高い。
- 参照訳の係り受け木を用いて参照訳を増やすことで日本語のスクランブリングに対応させた。

RIBES は日英・英日翻訳にかかわる多くの翻訳研究者に利用されているが、以下の問題点が指摘されている。

- RIBES を目的関数としてチューニングすることが難しい。
- ニューラルネットを用いた NMT (Neural Machine Translation) は、SMT と間違いの傾向が異なり、NMT に合わせた改良が必要。