The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

統計的機械翻訳における RIBES による最適化の効果

The effect of optimization by RIBES in statistical machine translation

河田　尚孝 *1　　磯崎　秀樹 *2　　菊井　玄一郎 *2
Naotaka Kawata　　Hideki Isozaki　　Genichiro Kikui

*1岡山県立大学大学院システム工学専攻
Okayama Prefectural University Graduate School of Computer Science and Systems Engineering

*2岡山県立大学情報工学部情報システム工学科
Okayama Prefectural University Faculty of Computer Science and Systems Engineering Department of Systems Engineering

Parameter optimization is indispensable for improving the performance of statistical machine translation. MERT is a standard optimization tool that uses BLEU as the typical objective function. However, for translation between very different language such as English - from/to - Japanese translation, BLEU has very low correlation with human judgement. Therefore, it is better to use more human-like evaluation function like RIBES instead of BLEU as the objective function in the optimization process. In this paper, we investigate effects of RIBES used as the objective function of MERT and found that RIBES achieves better than BLEU when SMT is forced to determine the word order which is very different from the source sentence.

1. はじめに

統計的機械翻訳（Statistical Machine Translation, SMT)[Koehn 09] においては，翻訳モデルや言語モデルなどを素性関数として対数線形モデルによって統合したモデルが広く利用されている．この方法では，個々の素性関数のパラメータを推定したあと，対数線形モデルにおける重みパラメータ（単に「重み」と呼ぶ）の推定（最適化と呼ぶ）が行われる．この最適化の手法として MERT(Minimum Error Rate Training)[Och 03] が知られている．MERT では，翻訳結果が参照訳（正解訳）とどの程度意味的に乖離しているかを評価するエラー関数を用意し，これを目的関数として検証用の対訳コーパスの翻訳結果に対して関数値が最小となるように対数線形モデルの重みを調整する．MERT では目的関数として翻訳評価関数 BLEU[Papineni 02]（の値を逆にしたような関数）がよく利用される．

しかしながら， BLEU は欧米言語間の翻訳のように語順が大きく変わらない言語間の翻訳では人手評価とある程度の相関あるが, 日英翻訳や英日翻訳では人手評価との相関が非常に低いことが報告されている [Echizen-ya 09]．当然ながら目的関数は人間の判断を反映していることが望ましいため，人手評価との相関が低い BLEU は目的関数として不適切である．この問題に対して，平尾ら [平尾 14] が提案した RIBES（の値を逆にしたような関数）を目的として用いることが考えられる．

ところが, Duh ら [Duh 12] によれば, RIBES は非常に「滑らかでない (unsmooth)」ため，翻訳最適化の目的関数として使うのは難しいことが指摘されている．この問題を回避するために，彼らは多目的最適化の枠組みを用いて BLEU と RIBES を混合したような評価関数を提案している．彼らは粗い近似として BLUE と RIBES を 3:1 で加重和した翻訳評価関数を推奨している．

本研究では統計的翻訳への入出力の語順の違いと MERT における評価関数の関係について再検討する．我々の仮説は次の通りである

> 「翻訳処理 (SMT) において語順を大きく変える必要がある言語対の場合は MERT の目的関数として BLUE より RIBES が適切であり，そうでない場合は BLEU の方が適切である」

本研究では Head Finalization を適用することによって統計翻訳において対応すべき語順の差を変更することにより，この仮説について検討する．以下，本稿の構成は次の通りである．2 章では我々が利用した統計的機械翻訳について，3 章では翻訳自動評価法のうち本研究で用いた BLEU と RIBES について説明する．4 章では実験について述べる．

連絡先: 河田　尚孝，岡山県立大学大学院, c226017c@cse.oka-pu.ac.jp

2. 対象とする統計的機械翻訳

本研究では以下の式で示される対数線形モデルに基づく統計翻訳を用いる．

$$\begin{aligned}\hat{e} &= \arg\max_{e} P(e|f) \\ &= \arg\max_{e} \sum_{m} w^{\mathrm{T}} h_m(f,e) \qquad (1)\end{aligned}$$

$P(e|f)$ を表す式 (1) において w_m は素性関数 $h_m(f,e)$ の重みである．素性関数としては翻訳モデル, 言語モデル（に対応する関数）のほか任意の関数が利用可能である．また，重み w_m は後述する MERT により学習される．

2.1 誤り率最小化学習 (MERT)

MERT(Minimum Error Rate Training) は, 対訳データ $< F, E >$ に対して，以下の損失関数 $l_{error}(\cdot)$ を最小化する w_m を探索する．

$$l_{error}(F, E, C; \mathbf{w_m}) = error(E, argmax_{<e,d>\in c^{(i)}} \mathbf{w_m^{\mathrm{T}} h(f^{(i)} e, d)}$$

ここで，$< f, e >$ は $< F, E >$ の各対訳を表し, C は k-best リストである．そして, 対訳データに対して適切な前処理を施すことでより正確に学習することができると考えられる．例として本実験でも使用している前処理手法である Head Finalization について説明する．

2.2 Head Finalization

本研究では英日翻訳については，統計的翻訳の前処理として磯崎ら [Isozaki 10] により提案された Head Finalization（以下,「ヘッド・ファイナル化」）と呼ぶ処理を適用する．ヘッド・ファイナル化とは，フレーズ内の主辞を最後（一番右）に並び変える処理であり，英日翻訳のように 1 ）目的言語が主辞後置型の言語であり，2 ）原言語の語順がそれを大きく異なる言語であるような言語対において有効な統語的前処理手法である *1．この処理により, 言語間の語順の差が相当程度解消され, 翻訳精度が向上する．ヘッド・ファイナル化が行われた文章を図 1 に示す．

並び替え前　when the fluid pressure cylinder 31 is used , fluid is gradually applied .

並び替え後　the fluid pressure 31 cylinder used is when , fluid gradually applied is .

図 1: 英語データの事前並べ替え

3. 自動評価尺度

翻訳機を自動的に評価する方法は翻訳評価の効率化という観点から従来より様々な手法が提案されている．そこで用いられるのが, 出力された機械訳を用意された参照訳と比較して自動でスコアを与える自動評価尺度である．以下に, 今回の実験で比較対象とした自動評価尺度を示す．

3.1 BLEU

BLEU は n-gram 適合率 p_n に基づいてスコアの計算を行う．BLEU のスコアは以下の式で表される．

$$BLEU = BP * exp(\frac{1}{N}\sum_{n=1}^{N}\log p_n) \tag{2}$$

適合率 p_n については n-gram 精度の幾何平均を表す．また BP は機械訳が参照訳と比較して短い場合に課せられるペナルティ係数である．BP は機械訳の単語の総和を c, 参照訳の単語の総和を r とすると, 次の式 (3) で表される．

$$BP = \begin{cases} 1 & (c > r) \\ e^{1-\frac{r}{c}} & (c \leq r) \end{cases} \tag{3}$$

3.2 RIBES

RIBES[平尾 11] はグローバルな語順に注目し, 順位相関係数を用いて翻訳品質の評価を行う．RIBES スコアは次の式で表される．

$$RIBES = NKT \times P^{\alpha} \quad (0 \leq \alpha \leq 1) \tag{4}$$

RIBES は参照訳 R と機械訳 H の語順の順位相関係数を用いるので, まず参照訳 R の単語に出現した順に 0 から順位を与え, 整数のリストを作る．このリストを r とする．次に機械訳 H の単語の内, 参照訳に出現したものをリスト r の整数値に置き換えた整数リスト h をつくる．リスト r と h の順序列間の順位相関係数を計算することで参照訳と機械訳の間で一致して出現する単語の出現順の近さを測ることが出来る．本研究では順位相関係数として Kendall の τ を採用している．順位相関係数 τ は以下のように表す．

$$\tau = \frac{\text{昇順ペア数}}{\text{全ペア数}} \times 2 - 1 \tag{5}$$

ここで式 (7) 中の NKT は以下の式で表される．

$$NKT = \frac{\tau + 1}{2} \tag{6}$$

しかしこの NKT は機械訳の単語が短いと過剰に高い値となる可能性があるため, 平尾ら [平尾 14] より BLEU で用いられているペナルティ係数 BP を用いた RIBES の式が提案されている．

$$RIBES = NKT \times P^{\alpha} \times BP^{\beta} \quad (0 \leq \alpha \leq 1, 0 \leq \beta \leq 1) \tag{7}$$

本実験では式 7 を用いて実験を行う．また式 7 中の α, β については参考文献 [平尾 11][平尾 14] より, $\alpha = 0.25$, $\beta = 0.1$ で評価している．

4. 実験

4.1 対訳データの準備

統計翻訳のトレーニングデータとして NTCIR-7 特許翻訳タスクの英日翻訳データ約 180 万文を使用する．最適化には同タスク内の開発データ 2655 文のうち無作為に抽出した 2155 文を用いる．残りの 500 文を自動評価尺度によるスコアを計算する際のテストデータとして使用する．以下に使用した対訳データの内訳を示す．

表 1: 対訳データの内訳

データの種類	文章の数
トレーニングデータ	約 180 万文
チューニング用データ	2155 文
テストデータ	500 文

全ての対訳データに対して, 以下のような処理を行った．

処理 1 日本語データを半角に変換

処理 2 特定の使用不可記号を文字列に変換

処理 3 括弧で囲まれている部分の削除

処理 4 データの分かち書き

処理 5 英語データの事前並び替え

処理 1 では, 日本語のデータに対して全角数字, アルファベット, 記号, カタカナをすべて半角に変換する処理を行う．

処理 2 では, Moses で使用が禁止されている記号を実体参照表現に置き換える．(例. & → &)

処理 3 では, 単語を補足する目的で括弧を使用している部分を括弧ごと削除した．　ただし ’ (a) ’ や ’ 1) ’ などの図や表を表しているものは中身を残して括弧のみを取り除く．理由は NTCIR-7 のデータには, ある単語を補足する目的で括弧を用いている部分が存在する．しかし, 括弧の位置が日本語と英語で対応していないものや, 片方の言語でしか括弧が現れていな

*1　日英翻訳のように逆方向の翻訳において，後処理として語順を並び変える方法 [Sudoh 11] が提案されているが本研究では利用しない

いもの, 括弧内の単語の対応が取れていないものが多数存在するため, 単語アライメントの計算や構文解析がより正確に行えるようにするためである.

処理 4 では, 日本語データの単語分割にはオープンソースの形態素解析エンジンである MeCab を使用した. 英語はカンマやピリオドなどの記号を分割する必要があったため, Moses に含まれる tokenizer.perl を用いて分割した.

処理 5 では, ヘッド・ファイナル化により, 英語の語順をあらかじめ単語を並び替えて日本語の語順に近づけておく. 単語の並び替えを行うためには英語の係り受け解析を行う必要があるため, 英語構文解析器である Enju[Yusuke 08] を用いた.

4.2 実験環境

前述した処理を行った対訳データと機械翻訳システム Moses を用いて英日, 日英翻訳機を作成した. 単語アライメントツールには GIZA++[*2] を使用した. 言語モデルの学習には KenLM を用い, トレーニングデータに含まれる文のみを使用した. 今回は最大 5 グラムまで学習させている. MERT による重み最適化には Moses に含まれる. mert.cpp を使用している. mert.cpp 内の目的関数を RIBES に変更することで RIBES による最適化を可能にしている.

4.3 実験項目

本研究では最適化における RIBES の効果を調べるために RIBES と BLEU を目的関数とした最適化によるスコアの推移を調べた. なお, 今回は英日, 日英翻訳機を使用しており, ヘッド・ファイナル化を行ったものを「HF あり」, ヘッド・ファイナル化を行っていないものを「HF なし」としている. また, RIBES と BLEU を線形結合した目的関数を用いることで, 語順と単語の適合率のどちらを重視した方がより最適化の効果が大きいのかを確かめた. RIBES と BLEU を線形結合した目的関数を以下に示す.

$$\text{目的関数} = \alpha BLEU + (1-\alpha) RIBES$$

二つの関数の割合について, α を以下のように設定して実験を行った.

$$\alpha = (1, 0.7, 0.5, 0.3, 0)$$

なお, $\alpha = 1$ は BLEU のみでの最適化を表している.

4.4 評価指標

翻訳機の品質を人手で評価することはコストがかかるため, 自動で評価できることが望ましい. そこで自動評価尺度を翻訳機の評価指標として用いる. 本実験では評価指標として RIBES を使用する. 使用した RIBES は式 (7) である. また, $\alpha = 0.25$, $\beta = 0.1$ で実験を行っている.

4.5 実験結果と考察

評価関数を RIBES に変更し, RIBES と BLEU を目的関数として最適化を行った結果を図 2 に示す. 図 2 より, ヘッド・ファイナル化を行ったものについて, BLEU の方が最適化後のスコアが高いことが分かった. しかし, BLEU のみで最適化した場合と BLEU と RIBES を線形結合した場合のスコアは同程度であった. これは Duh らが述べているように RIBES での最適化がうまくいかない原因の一つとして考えられている RIBES が「滑らかではない」関数であるという問題が BLEU を加えることで緩和されるからだと考えられる. 次にヘッド・ファイナル化を行っていない場合について, RIBES で最適化した方が BLEU よりも最適化後のスコアが上がっていることが分かった. また, RIBES の割合を大きくするほどスコアが上がっている. この結果から, 語順が異なる言語対に対して RIBES は語順を正そうと働くため, 単語の適合率を重視する BLEU よりも最適化の目的関数として適していることが分かった.

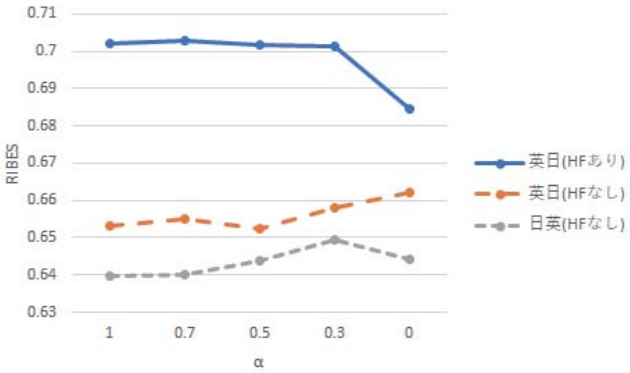


図 2: RIBES と BLEU の比率を変えたときの最適化

次に RIBES と BLEU を線形結合した目的関数を用いた最適化による RIBES スコアの推移について図 3(a)〜3(c) に示す. 図 3(a) より, ヘッド・ファイナル化を行った場合に効果がある BLEU を目的関数としたとき, スコアは単調に増加していることが分かる. 一方 RIBES での最適化では, 1 回の学習によるスコアの変化が大きく, 最適化が BLEU よりも早く終わっている.

続いて図 3(b) について, ヘッド・ファイナル化を行わない場合, RIBES での学習により単調にスコアが増加していった. 図 3(c) の場合は, 最適化のときに BLEU の方が大きくスコアが変化していた. また図 3(a) と図 3(b) の共通点として, 学習後のスコアがより高くなる目的関数を用いた学習ではスコアが単調増加している. それに対してスコアが低くなる目的関数を用いた場合, 学習の回数は少なく, 1 回の学習によるスコアの変化が大きくなる傾向が見られた.

5. おわりに

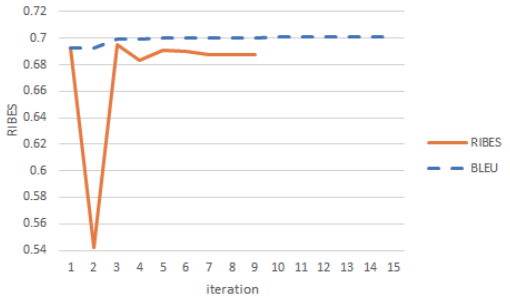
本研究では ngram 再現率による翻訳自動評価尺度である BLEU，および，大域的な語順を考慮した翻訳自動評価法である RIBES を目的関数として MERT の最適化の効果について日本語一英語間の特許文翻訳を対象として検証した. その結果, RIBES は日英翻訳やヘッド・ファイナル化なしの英日翻訳のように統計翻訳において語順を大きく並び替える必要がある言語対に対して BLEU より効果的であることが分かった. 一方，ヘッド・ファイナル化による事前処理を行って統計的翻訳における語順の並び替えが限定的である場合には BLEU のほうが効果的であり，さらに RIBES と BLEU を線形結合することで BLEU のみよりやや効果的であることが分かった.

今回は最適化の効果を RIBES で評価したが，今後の課題として, 人手評価による検証があげられる.

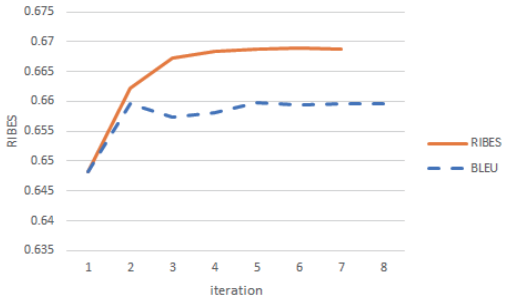
参考文献

[Duh 12] Duh, K., Sudoh, K., Wu, X., Tsukada, H., and Nagata, M.: Learning to Translate with Multiple Objectives, *Proceedings of the 50th Annual Meeting of the Associateion for Computational Linguistics*, Vol. 1, pp. 1–10 (2012)

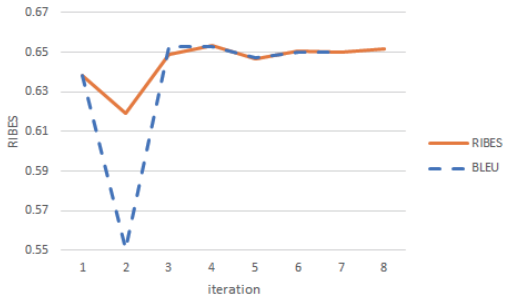
*2 http://www.statmt.org/moses/giza/GIZA++.html



(a) 英日 (HF あり) の最適化の推移



(b) 英日 (HF なし) の最適化の推移



(c) 日英 (HF なし) の最適化の推移

図 3: 最適化の推移

[Echizen-ya 09] Echizen-ya, H., Ehara, T., Shimohata, S., Fuji, A., Utiyama, M., Yamamoto, M., Utsuro, T., and Kando, N.: Meta-Evaluation of Automatic Methods for Machine Translation using Patent Translation Data in NTCIR-7, *Proceedings of the 3rd Workshop on Patent Translation*, pp. 9–16 (2009)

[Isozaki 10] Isozaki, H., Sudoh, K., Tsukada, H., and Duh, K.: Head Finalization: A Simple Reordering Rule for SOV Langages, *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMERT*, pp. 244–251 (2010)

[Koehn 09] Koehn, P.: Statistical Machine Translation, *Cambridge University Press* (2009)

[Och 03] Och, F. J.: Minimum Error Rate Training in Statistical Machine Translation, *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 160–167 (2003)

[Papineni 02] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J.: 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 (2002)

[Sudoh 11] Sudoh, K., Wu, X., Duh, K., Tsukada, H., and Nagata, M.: Post-ordering in statistical machine translation, *Proceedings of the 3rd Workshop on Patent Translation* (2011)

[Yusuke 08] Miyao, Y., Jun'ichi Tsujii:Feature Forest Models for Probabilistic HPSG Parsing, *Computational Linguistics*, Vol. 34, No. 1, pp. 35–80 (2008)

[平尾 11] 平尾努, 磯崎秀樹, 須藤克仁, Duh, K., 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 第 17 回年次大会 発表論文集, pp. 1115–1118 (2011)

[平尾 14] 平尾努, 磯崎秀樹, 須藤克仁, Kevin, D., 塚田元, 永田昌明：語順の相関に相関に基づく機械翻訳の自動評価法, 自然言語処理, Vol. 21, No. 3, pp. 421–444 (2014)