自動評価を用いた機械翻訳規則のフィードバッククリーニング

今村賢治[†,††] 隅田英一郎[†] 松本裕治[††]

機械翻訳規則を対訳コーパスから自動獲得すると，自動獲得エラーや，コーパスに含まれる翻訳の多様性に起因して，不適切な規則の混入が避けられない．これらは誤訳や曖昧性増大の原因となる．この問題に対して本稿では，翻訳品質の自動評価を利用して最適な組合せを探索する，機械翻訳規則の取捨選択法を提案する．これをフィードバッククリーニングと呼ぶ．自動評価方法には BLEU を用い，組合せ最適化方法には，タスクの特徴を考慮した山登り法を用いた．本方式は，機械翻訳器の様々なパラメータや他の知識との干渉を考慮せずに，翻訳品質を向上させることができる．実験では，従来法に比べ，大幅に翻訳品質が向上することを確認した．

Feedback Cleaning of Machine Translation Rules Using Automatic Evaluation

Kenji Imamura,[†,††] Eiichiro Sumita[†] and Yuji Matsumoto[††]

When machine translation (MT) rules are automatically acquired from bilingual corpora, incorrect/redundant rules are generated due to acquisition errors or translation variety in the corpora. Such problematic rules cause implausible translations or increase ambiguity. As a new countermeasure to this problem, we propose a feedback cleaning method using automatic evaluation of MT quality, which removes incorrect/redundant rules as a way to increase the evaluation score. BLEU is utilized for the automatic evaluation. The hill-climbing algorithm, which involves features of this task, is applied to searching for the optimal combination of rules. Our method can improve MT quality without considering various parameters of an MT engine. Our experiments show that MT quality considerably improves than previous methods.

1. はじめに

対訳コーパスの充実にともない，コーパスベースの機械翻訳方式[3),13)]の研究がさかんになっている．従来人手によって知識が作成されていた構文トランスファ方式機械翻訳[5),6),14)]についても，その変換規則を対訳コーパスから自動獲得する手法が提案されてきている[2),7),9),23)]．しかし自動獲得した規則は，自動獲得エラー，あるいはコーパス中に含まれる翻訳の多様性によって，機械翻訳に不適切な，または競合する規則が獲得される．これらは誤訳や曖昧性増大の原因となる．このような不適切/競合規則（まとめて冗長規則と呼ぶ）を変換規則中から削除すれば，機械翻訳品質を向上させることができる．

従来，冗長規則への対処法として，以下の 2 つのアプローチがあった．

- 曖昧性解消の一環として，翻訳時に適切な規則を選択する方法（オンライン法[12)]）
- 自動獲得の後処理として，冗長規則をクリーニングする方法（オフライン法[8),11)]）

本稿では，第 2 のアプローチについて提案する．オフライン法としては，頻度による足切り[11)]，仮説検定によるクリーニング法[8)]が提案されている．頻度による足切り法は，コーパス中の各規則の出現頻度をカウントし，閾値以下の低頻度規則を一律に削除する．この方法は若干の翻訳品質向上が確認されているが，冗長な規則数に比べ，大幅な品質向上は達成できていない．また，仮説検定法は，ある規則の原言語側と目的言語側の対応が正しいかどうか，統計的に検定を行っている．しかし，統計的に信頼できる規則がコーパスサイズに比べて少ないため，翻訳に十分な規則数を得るためには，非常に大規模なコーパスを必要とするという問題点がある．

一方，翻訳品質の自動評価法も提案されてきている[1),4),18),24)]．これらは，機械翻訳システムの開発時，

† ATR 音声言語コミュニケーション研究所
ATR Spoken Language Translation Research Laboratories
†† 奈良先端科学技術大学院大学
Nara Institute of Science and Technology

規則番号	構文カテゴリ	原言語パターン		目的言語パターン	用例
1	VP	X_{VP} *at* Y_{NP}	⇒	Y' で X'	((*present*, *conference*) ...)
2	VP	X_{VP} *at* Y_{NP}	⇒	Y' に X'	((*stay*, *hotel*), (*arrive*, *p.m*) ...)
3	VP	X_{VP} *at* Y_{NP}	⇒	Y' を X'	((*look*, *it*) ...)
4	NP	X_{NP} *at* Y_{NP}	⇒	Y' の X'	((*man*, *front desk*) ...)

図 2 HPAT の変換規則例（前置詞 '*at*' を含むもの）
Fig. 2 Example of HPAT transfer rules that include prepositon '*at*'.

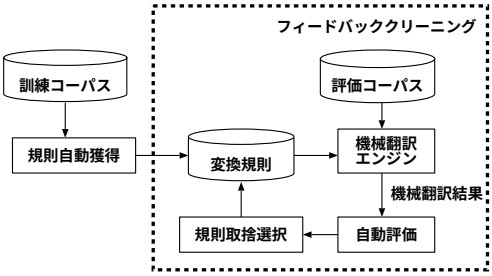


図 1 フィードバッククリーニングの基本構成
Fig. 1 Structure of feedback cleaning.

従来の人手による主観評価を自動評価に置き換えることにより，コストダウンおよび開発サイクルのスピードアップを狙ったものである．しかし，これらは開発支援ばかりでなく，翻訳システムそのものの自動チューニングにも利用できると考えられる[20]．

本稿では，そのような自動チューニング法の 1 つとして，自動評価を翻訳規則の取捨選択に利用する，フィードバッククリーニング方式を提案する（**図 1**）．具体的には，個々の規則の翻訳結果への寄与度を評価し，自動評価値を高めるように規則の取捨選択を行う．したがって，自動評価法が翻訳品質と十分に相関があるならば，クリーニング後の翻訳品質も向上する．

本方式の特徴は，翻訳結果だけを評価対象としているため，機械翻訳エンジンの様々な状態，たとえばパラメータ，他の知識との干渉，曖昧性解消方式等を意識する必要がない点である．たとえ機械翻訳エンジンが，オンライン法で冗長規則を避けていたとしても，解消しきれないエラーは存在する．本方式は，残ったエラーだけに焦点をあてて規則をクリーニングする．つまり本方式は，オンライン法を補完し，ルールセットを翻訳エンジンの状態に適応させる効果がある．

2. 使用した機械翻訳システムと自動獲得の問題点

2.1 翻訳エンジン

構文トランスファ方式の機械翻訳は，入力文を構文解析し，構文木（句構造や依存構造）を目的言語の構造に変換することにより翻訳を行う．変換時に，語順を大幅に変更することができるため，英語と日本語のように語順が著しく異なる言語間の翻訳にはよく用いられている方式である[5),6),14)]．

本稿では，構文トランスファ方式の機械翻訳システムとして，HPAT[8)] を用いる．HPAT の翻訳知識のうち，最も重要な知識は，原言語と目的言語の表現を対応づけた，変換規則である．英日翻訳における例を**図 2** に示す．これは基本的には同期文脈自由文法規則であり，図中の X_{VP} と X' 等は，対応する非終端記号を表す．

翻訳する際は，まず変換規則の原言語パターンを用いて入力文を構文解析する．次に，使用された原言語パターンを目的言語パターンにマッピングすることにより，目的言語の木構造を得る．目的言語構造の葉に非終端記号が残った場合は，対訳辞書を参照して，目的言語の単語を埋め込む．

解析または変換時に発生した曖昧性は，入力文と用例との意味距離（シソーラス[17)] 上でのノード間の距離）を用い，最も距離が近い規則が選択される[5)]．たとえば，入力が "*leave at 11 a.m.*" であった場合，図 2 の規則のうち，用例（*arrive*, *p.m.*）との距離が最も近くなるため，規則 2 が選択され，「11 時（*11 a.m.*）に 出る（*leave*）」という訳が作られる．

2.2 自動獲得の問題点

HPAT は，変換規則を対訳コーパスから階層的句アライメント[7)] を用いて自動的に獲得する．しかし，自動獲得された規則には，冗長な規則が多数含まれる．その原因を大別すると以下のとおりであり，コーパスから自動獲得した規則を用いる翻訳システムでは避けられないものである．

- 規則の自動獲得エラー
  完全な自動獲得方式はないため，自動獲得した規則には必ず誤りが含まれる．
- 対訳コーパスに起因するもの
  - 文脈や状況に依存する訳が含まれ，規則を一般化できない．
  - 同じ原言語でも複数の訳がコーパスに含まれている．

文献 8) の実験では，12 万対訳から変換規則を自動

獲得した結果，約 9 万 2 千規則が生成された☆．この規則はほとんどがコーパス中に 1 回，または 2 回しか出現しない低頻度規則である．低頻度規則を削除し，規則数を約 1/9 に減少させたにもかかわらず，翻訳品質は若干向上したと報告している．しかし，低頻度規則には，慣用表現の翻訳に必要な規則等も含んでいるため，これらを一律に削除しても翻訳品質を大幅に向上させることはできない．

3. 機械翻訳品質の自動評価

本稿では，自動評価方法として，BLEU[18] を用いる．文献 18) によると，BLEU は主観評価と十分に相関があるとされており，文献 15) ほか，近年の機械翻訳の評価に多く用いられていることを考慮した．

BLEU スコアは，機械翻訳結果と人が翻訳した結果（参照訳）を比較し，その類似度を数値化したものである．類似度は，両者の単語 N-gram 一致数で測定される．N は可変であるが，本稿では標準設定に従い，1-gram から 4-gram までを用いて測定する．1-gram は，単語レベルでの一致数を表すため，単語訳の正しさを表す指標となっている．また，高次の N-gram は，翻訳の流暢さを表す指標となっており，BLEU スコアは両者を組み合わせた指標となっている．

ここで注意したいのは，BLEU スコアを算出するためには，ある程度の大きさを持った評価用の対訳文集合を用いる必要がある点である．BLEU スコアを 1 文ごとに算出することも可能ではあるが，主観評価とのずれが大きい．BLEU は各文の類似度を翻訳結果集合全体について総和をとることにより，個々の誤差を相殺している．つまり，適切な BLEU スコアを算出するためには，評価コーパス全体の翻訳結果が必要である．

BLEU は，参照訳を 1 原文あたり複数使用することができる．しかし本稿では，既存の対訳コーパスを用いるため，1 原文あたり 1 参照訳を使用する．また，翻訳結果が日本語の場合，単語に分かち書きする必要が生じる．本稿では，翻訳結果，参照訳ともに同一の形態素解析器[22] で分かち書きを行う．形態素解析結果には分かち書き誤りも含まれるが，参照訳も同一の形態素解析器で分かち書きすることにより，同じ文は同じ誤りを共有するため，BLEU スコアに対する影響は最小限にとどめられる．

☆ 本稿では，規則数を数える場合，原言語パターンと目的言語パターンの対を単位とする．

4. フィードバッククリーニング方式

本章では，提案方式であるフィードバッククリーニングについて説明する．本方式は，基本的には評価コーパスの自動評価値が向上するように，変換規則の追加/削除を繰り返すことで実行される（図 1）．すなわち，規則の組合せ最適化問題と位置づけられる．本稿では，タスクの特性を考慮した山登り法を用いて最適化を行う．以下その理由と手順を述べる．山登り法による最適化は，しばしば局所解に陥るが，有効なクリーニング方式が提案されていない現在，少しでも翻訳品質が向上することには意味があると考える．

4.1 組合せ最適化のコスト

本方式は，規則の追加/削除とその評価を繰り返して最適な組合せを探索するが，最も時間がかかる処理は評価である．たとえば，評価コーパスサイズ C が 1 万文であるとすると，1 つの規則の組合せ（解）に対する BLEU スコアを算出するためには，1 万回の機械翻訳を実行しなければならない．さらに，ある解の最適近傍を求めるためには，すべての近傍解の BLEU スコアが必要である．仮にルールセットの規則数 R を 10 万とし，1 規則の追加/削除状態を近傍解とすると，$CR = 1$ 万文 $* 10$ 万規則 $= 10$ 億回 の機械翻訳を実行する必要がある．この翻訳回数の削減が本手法の実現性を決める．

本タスクの特徴は，規則の追加に比べ削減が容易であるという点である．なぜなら，機械翻訳を行った時点で，使用された規則が判明する．逆にいうと，評価コーパスを一度翻訳すれば，ある規則 r が使用される文集合 $S[r]$ が決まる．r を削除したときに変化する翻訳文は $S[r]$ だけなので，他の文は再翻訳する必要がない．つまり，r が削除されたときの BLEU スコアは，$S[r]$ 回の機械翻訳で計算することができる．一方，規則を追加する場合は，追加により変化する翻訳文が不明であるため，評価コーパス全体を再翻訳する必要が生じる．

4.2 クリーニング手順

以上の議論をふまえ，本方式では，全規則を含んだ状態（これをベースルールセットと呼ぶ）を初期状態とし，規則削減に限定して探索する，山登り法を用いる．アルゴリズムを**図 3** に示す．本アルゴリズムの要点は，以下のとおりである．

(1) 最初に，評価コーパスをすべて翻訳し，使用された規則と，削除前 BLEU スコアを得る．
(2) 各規則ごとに，その規則削除後の BLEU スコアを算出し，削除前スコアとの差分を得る．こ

```
static: C_eval, 評価コーパス
        R_base, 訓練コーパスから作成されたルールセット全体．ベースルールセット
        R, 現在のルールセット．R_base のサブセット
        S[r], 規則 r が使われた文セット
        Doc_iter, 現在のルールセットで評価コーパスを翻訳した結果
---------------------------------------------------------------
procedure CLEAN-RULESET ()
    R ← R_base
    repeat
        R_iter ← R
        R_remove ← ∅
        score_iter ← SET-TRANSLATION()
        for each r in R_iter do
            if S[r] ≠ ∅ then
                R ← R_iter − {r}
                S[r] をすべて翻訳し，翻訳文 T[r] を得る
                Doc[r] ← Doc_iter から T[r] を入れ替えたもの
                規則寄与度 contrib[r] ← score_iter − BLEU-SCORE(Doc[r])
                if contrib[r] < 0 then R_remove に r を追加
        end
        R ← R_iter − R_remove
    until R_remove = ∅
---------------------------------------------------------------
function SET-TRANSLATION () returns 現ルールセット R で翻訳した評価コーパスの BLEU スコア
    Doc_iter ← ∅
    for each r in R_base do S[r] ← ∅ end
    for each s in C_eval do
        s を翻訳し，翻訳文 t を得る
        s を翻訳するのに使用されたルールセット R[s] を取り出す
        for each r in R[s] do s を S[r] に追加する end
        Doc_iter に t を追加する
    end
    return BLEU-SCORE(Doc_iter)
```

図 3 フィードバッククリーニングアルゴリズム
Fig. 3 Feedback cleaning algorithm.

れを規則寄与度と呼ぶ．

(3) 規則寄与度が負（削除により BLEU スコアが向上）であれば，その規則を削除する．

(4) 以上 (1)〜(3) を，BLEU スコアが向上しなくなるまで繰り返す．

なお，本アルゴリズムでは収束を早めるため，削除される規則どうしは独立であると仮定し，1 回の繰返しで規則寄与度が負であるすべての規則を削除している．

本アルゴリズムを使った場合，仮に 1 文の翻訳に平均 5 規則が使用されたとすると，1 繰返しに必要な翻訳回数は，$C + 5C = 10,000 + 5 * 10,000 = 60,000$ 回となる．

5. 交差クリーニング方式

一般的には，訓練コーパスに比べ，評価コーパスはサイズが小さい．そのため，評価コーパスだけではすべての規則をテストできず，クリーニング漏れが発生する．この問題を回避するため，交差検定と同様な考え方を導入し，訓練コーパス自体を用いてクリーニングを行う．これを本稿では，交差クリーニングと呼ぶ（**図 4**）．

交差クリーニングの手順は以下のとおりである．

(1) まず，訓練コーパス全体から，ベースルールセットを作成する．

(2) 次に，訓練コーパスを均等に N 分割する．

(3) 分割コーパスのうち，1 つを評価用，残り (N−1) つを訓練用とし，ルールセットを作成する．したがって，N 個のルールセット/評価コーパスが作成される．各ルールセットは，ベースルールセットのサブセットとなる．

(4) 各ルールサブセットを，図 3 に示した方法でクリーニングする．なお，その際，削除された規則であっても，規則寄与度は記録しておく．このステップの目的は，規則寄与度を得るためのものである．

(5) ベースルールセット中の規則 1 つ 1 つについて，ルールサブセットから得られた規則寄与度

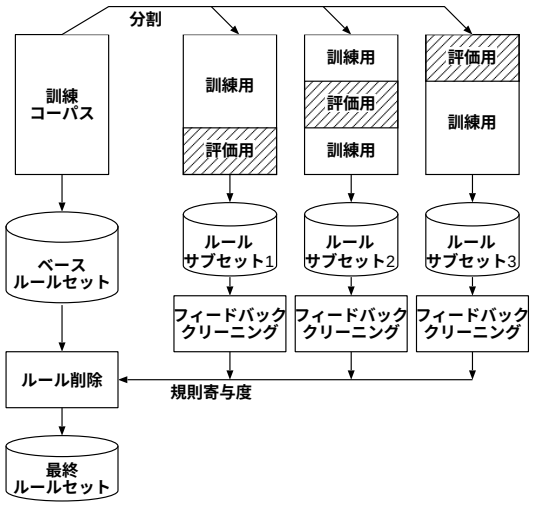


図 4 交差クリーニングの構成（3 分割交差クリーニングの場合）
Fig. 4 Structure of cross-cleaning (in the case of three-fold cross-cleaning).

を総和する．もし，総和が負であれば，その規則をベースルールセットから削除する．

交差クリーニングと交差検定の主な相違点はステップ (5) である．交差クリーニングの場合，ステップ (4) で規則が削除されるため，直接サブセットを併合することができない．そのためステップ (5) では，フィードバッククリーニング結果から，規則寄与度だけを取得し，総和を算出する．各サブセットの規則寄与度は，評価サブコーパスに対するものであり，その総和は訓練コーパス全体への寄与度の近似値と見なすことができる．交差クリーニングはこの近似値を用いてベースルールセットから規則を削除する☆．

交差クリーニングは特別な評価コーパスを用いないにもかかわらず，クリーニングに訓練コーパスのすべての文を使用するため，大規模な評価コーパスを用いた場合とほぼ同じ効果が得られる．

6. 評　　価

本章では，フィードバッククリーニングの有効性を，英日翻訳の翻訳品質で評価する．6.1 節では実験条件について述べ，6.2 節では評価コーパスによるクリーニングを実施し，フィードバッククリーニングの基本的特徴を確認する．次に 6.3 節では，評価コーパスによるフィードバッククリーニング，交差クリーニングと，従来法の比較を行う．最後に 6.4 節で，削除された規則の実例を示す．

☆ 図 3 において，規則寄与度を評価コーパス全体を対象に算出している理由の 1 つは，交差クリーニングのステップ (5) でその値の総和をとる必要があるからである．

表 1 コーパスサイズ
Table 1 Corpus size.

コーパス名	項目	英語	日本語
訓練コーパス	文数	149,882	
	総形態素数	868,087	984,197
	異なり語彙数	11,288	17,575
評価コーパス	文数	10,145	
	総形態素数	59,533	67,554
	異なり語彙数	4,013	4,986
テストコーパス	文数	10,150	
	総形態素数	59,232	67,193
	異なり語彙数	4,030	5,040

6.1 実 験 条 件

6.1.1 対訳コーパス

本稿で用いたコーパスは，ATR の旅行会話基本表現集[10),21)] で，旅行会話に頻出する表現を集めたものである☆☆．このコーパスを，**表 1** に示すように，訓練，評価，テストコーパス（オープンテストによる翻訳品質測定用コーパス）に分割して実験した．なお，訓練コーパス全体から作成された変換規則（ベースルールセット）は，105,588 規則である．

6.1.2 評 価 方 法

評価方法には，以下の 2 つの方法を用いた．

(1) **テストコーパス BLEU スコア**
テストコーパスを用い，BLEU を用いて評価した．フィードバッククリーニングと同様に，参照訳を各文 1 つ使用した．

(2) **主観評価**
主観評価では，上記テストコーパス中の 510 文を使い，筆者以外の日本語ネイティブ話者 1 名が一対比較法で評価した．具体的には，ベースルールセットによる翻訳結果に対し，クリーニング後のルールセットによる翻訳結果を 1 文ずつ比較し，どちらの翻訳が良い翻訳であるのか，あるいは同品質であるのか，判定する形で行った．主観評価における翻訳品質（主観品質）は，以下の式で表す．したがって，これはベースルールセットに対する相対評価である．

$$\text{主観品質} = \frac{\text{改善文数} - \text{改悪文数}}{\text{テスト文数}} \qquad (1)$$

6.2 評価コーパスを用いたフィードバッククリーニング

まず，フィードバッククリーニング方式の基本的特徴を確認するため，評価コーパスによるクリーニングを実施した．結果を**図 5** に示す．本グラフは，繰返

☆☆ 現在約 20 万対訳を収集し，さらに拡張中である．

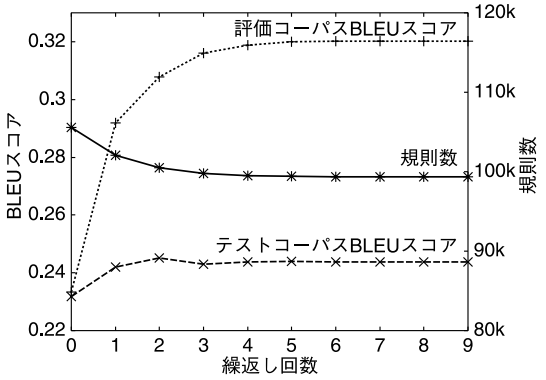


図 5 評価コーパスを用いたフィードバッククリーニング時の繰返し回数と BLEU スコア/ルール数の推移
Fig. 5 Relationship between number of iterations and BLEU scores/number of rules.

し回数を変えたときのテストコーパス BLEU スコア，評価コーパス BLEU スコア，規則数の推移を表したものである．

結果的に，9 回の繰返しで完全に収束し，延べ 6,220 規則が削除された．評価コーパス BLEU スコアは，繰返し回数が増加するに従い向上しており，山登り法による組合せ最適化が有効に働いていることを示している．テストコーパス BLEU スコアは，繰返し 2 回のとき，ピーク値 0.245 を示し，その後若干低下した．これは，過学習が起こったためであるが，最終的にはピーク値とほぼ同じ 0.244 で収束した．

評価コーパス BLEU スコアに比べ，テストコーパス BLEU スコアの向上が少ないが，これは，テストコーパスで使用される規則を，評価コーパスでは網羅的にチェックできなかったためである．評価コーパスサイズを拡大すれば，テストコーパス BLEU スコアも向上すると考えられる．

なお，翻訳回数は，1 繰返しあたり平均 36,988 回であった☆．これは，1 文の翻訳時間を 1 秒としても，10 時間強で 1 繰返しが終了することを意味する．規則のクリーニングは実時間処理を必要としないので，これは十分許容できる時間である☆☆．

6.3 クリーニング方式の比較

次に，従来方式との比較のため，クリーニング方式を変えて翻訳品質を測定した．クリーニング方式は，以下の 5 方式である．

(1) **ベースライン**
ベースルールセットによる翻訳結果．

(2) **頻度による足切り**
頻度 2 未満の規則をベースルールセットから削除し，翻訳を行った場合．この閾値は，テストコーパス BLEU スコアが最高値を示すよう，実験的に決定した．

(3) **χ^2 検定**
文献 8) の実験と同様に，χ^2 検定を実施した場合．本実験では，95%の信頼点以上の規則を採用した（$\chi^2 \geq 3.841$）．

(4) **フィードバッククリーニング**
表 1 の評価コーパスを用いたフィードバッククリーニングを実施した場合（6.2 節と同じ）．

(5) **交差クリーニング**
本実験では，5 分割交差クリーニングを行った．

結果を**表 2** に示す．提案方式（フィードバッククリーニング，交差クリーニング）はいずれも，テストコーパス BLEU スコア，主観品質とも向上し，従来方式を大幅に上回った．

主観品質について着目すると，フィードバッククリーニング，交差クリーニングともにベースラインから改悪した文がある．本方式は，評価コーパスのすべての文を正しく翻訳するのではなく，規則の削除によって改悪する場合がある．しかし，評価コーパス全体を俯瞰して，良い翻訳の割合を高めるよう動作しているため，主観品質が向上した．

規則数について着目すると，フィードバッククリーニングは交差クリーニングに比べ，残った規則数が多く，これは局所解であったことが分かる．しかし，交差クリーニングでも，頻度による足切りに比べ，3 倍程度の規則数が残っており，交差クリーニングでも局所解である可能性が高い．もし，大域最適解を発見することができれば，さらに翻訳品質が改善できると推測される．

6.4 削除規則の例

評価コーパスを用いたフィードバッククリーニングで削除された規則と，その際変化した翻訳結果の実例を**図 6** に示す．これらには，以下の傾向がある．

- 規則 1 は，変換規則自動獲得のエラーによって作られた，誤った規則である（‘*admission*’ が翻訳されない）．このような規則が機械翻訳に適用されると，必要な要素が欠落した翻訳文が生成される．しかし，翻訳結果と参照訳との類似度が低くなるため，これらは削除される．
- 規則 2 は，英語動詞句 “*include tax*” を日本語の述語句「税込みだ」に翻訳する規則である．この

☆ 1 文の翻訳に使用された規則数は，各繰返しごとに異なり，平均 2.56〜2.66 規則，最大 24 規則であった．
☆☆ この実験では，Pentium(R) 4，2 GHz のコンピュータを用い，延べ約 80 時間で収束した．

表 2 クリーニング方式別翻訳品質
Table 2 Translation quality vs. cleaning methods.

	ベースライン	従来方式		提案方式	
		頻度による足切り	χ^2 検定	フィードバッククリーニング	交差クリーニング
規則数	105,588	26,053	1,499	99,368	82,462
テストコーパス BLEU スコア	0.232	0.234	0.157	**0.244**	**0.277**
主観品質		+2.35%	−5.88%	**+5.69%**	**+11.18%**
改善文数		87	119	79	107
同品質文数		348	242	381	353
(同一翻訳)		(257)	(114)	(266)	(234)
改悪文数		75	149	50	50

番号	訓練コーパス上の頻度	変換規則と評価コーパスの翻訳例			
		構文カテゴリ	原言語パターン	目的言語パターン	用例
1	3	NP	*the admission* X_N ⇒	X'	((*fee*) ...)
		入力文	*What is the admission fee?*		
		削除前翻訳文	料金はいくらですか.		
		削除後翻訳文	入場料はいくらですか.		
2	44	VP	*include* X_{NP} ⇒	X' 込み だ	((*tax*) (*gas*) ...)
		入力文	*Does it include tax?*		
		削除前翻訳文	税込みれていますか.		
		削除後翻訳文	税金は含まれていますか.		
3	7	S	*please* X_{VP} ⇒	X' たい の です が	((*send*) (*receive*) ...)
		入力文	*Please cash this traveller's check.*		
		削除前翻訳文	このトラベラーズチェックを現金にしたいのですが.		
		削除後翻訳文	このトラベラーズチェックを現金にしてください.		
4	710	S	*could you* X_{VP} ⇒	X' てください	((*check*) (*find*) ...)
		入力文	*Could you tell me how to fill in this form?*		
		削除前翻訳文	この書類の書き方を教えてください.		
		削除後翻訳文	この書類の書き方を教えていただけませんか.		
		入力文	*Could you give me a hand for a second?*		
		削除前翻訳文	ちょっと手伝ってください.		
		削除後翻訳文	ちょっと手伝います.		

図 6 評価コーパスを用いたクリーニングにおいて削除された規則の例
Fig. 6 Example of removed rules by feedback cleaning using evaluation corpus.

規則自体は誤りとはいえない．しかし，他の規則と組み合わせたとき，結果的に誤訳となる場合は，削除される．

- 規則 3 は正しい規則であり，削除前，削除後，どちらの翻訳文も正しい．このように，正しい規則どうしが競合している場合は，評価コーパス中の多数派表現に近づくように削除される．
- 規則 4 も正しい規則であるが，規則 3 と同様に削除された．しかしその際，一部の翻訳結果が誤訳になる現象が観察された．本稿で用いた翻訳エンジンは，規則選択の基準として入力文と用例との意味距離を基にしているため，削除後に利用される規則は一定ではない．そのため，入力文によっては誤訳となる場合がある．しかし，前節でも述べたように，フィードバッククリーニングは，評価コーパス全体として見て，良好な翻訳結果の割合が高くなるように規則の削除を行う．

7. 議　　論

7.1 他の自動評価法

BLEU 以外にも機械翻訳の自動評価法はいくつか提案されている．そのうち，評価結果をスコアとして出力する方式には，文献 1)，4)，20)，24) がある．このうち文献 1)，20)，24) は，機械翻訳結果と人が翻訳した参照訳との類似度を DP マッチング（編集距離）で算出している．文献 4) は BLEU と同様に，類似度を N-gram 一致数を基に算出している．これらの自動評価法も，フィードバッククリーニングに利用可能である．

BLEU とこれらの共通点は，参照訳との類似度を 1 文単位で測定したときの誤差を少なくするため，評価コーパス全体の翻訳結果が必要だという点である．し

たがって，これらの自動評価法を用いて翻訳規則のクリーニングを行う場合にも，4.1 節で述べたような翻訳回数削減は必須である．

7.2 他の機械翻訳方式における最適化

コーパスベースの機械翻訳の手法として，自動獲得した規則を用いる方法のほかに，単語や句の翻訳確率に基づいた統計翻訳[3),16)]がある．統計翻訳における，自動評価を用いた最適化方法としては，Och の提案[15)]がある．これは，対訳コーパスから統計モデルを学習する際に用いる素性関数の重みを，自動評価値を最大にするよう最適化するものである．

本稿の提案方式と比較すると，自動評価値を最大にするような最適化を行うことにより，翻訳品質の向上を目指しているという，コンセプトは同じである．しかし，最適化すべき変数が異なる．文献 15) では，8 つの実数値を最適化しているが，本方式は，約 10 万規則の組合せ（有効，無効の二値）を最適化している．そのため，文献 15) の方法を翻訳規則のクリーニングに適用した場合，変数の数が膨大になり，解が求まらない．また，本方式を統計翻訳に適用するには，二値を実数値に変える必要がある．したがって，文献 15) の方法を翻訳規則のクリーニングに適用すること，および本方式を統計翻訳の最適化に適用することは，直接的にはできない．

しかし，本方式は，規則等を用いる機械翻訳であれば，どのような方式にも適用可能である．たとえば，用例翻訳[13)]は，入力文に類似した対訳文（または部分対訳文）をコーパスから検索し，目的言語側を修正することにより翻訳を行っている．類似した対訳文が競合した場合には，シソーラス等を用いて曖昧性解消を行っているが，その際，誤った候補を選択する場合がある．このような場合にも本方式を適用することにより，誤訳の原因となる対訳文を削除することが可能である．

また，規則を人手で作成する場合（たとえば文献 6)）においても，ルールセットが大規模になるに従い，追加規則が既存規則の適用を妨げる（いわゆる副作用が起こる）場合がある．このような場合にも，追加規則の規則寄与度を測ることにより，副作用を最小限にとどめることができると考えられる．

7.3 ドメイン適応への応用

機械翻訳のドメインを変更する場合，新ドメインの対訳コーパスを用意する必要がある．しかし，対訳コーパス収集コストを考えると，一般的には元コーパスと同程度の量のコーパスを準備できない場合が多い．

本稿で提案したクリーニング方式は，規則を評価コーパスに適応させていることに相当する．したがって，新ドメインの対訳コーパスを評価コーパスと見なし，フィードバッククリーニングを実行すれば，容易にルールセットをドメインに適応させることができる．つまり，新ドメインのコーパスが小規模でも適応できる可能性が高い[19)]．ただし，本方式は規則を削除することにより適応を実現しているため，ベースルールセット中に新ドメインの翻訳文を構成するのに必要な規則が含まれている必要がある．

8. おわりに

本稿では，翻訳品質の自動評価を機械翻訳規則の取捨選択に利用する，フィードバッククリーニング方式を提案した．自動評価方法には BLEU を用い，組合せ最適化方法には，タスクの特徴を利用し，全規則から削除のみを行う山登り法を用いた．さらに，評価コーパスサイズの影響を削減するため，交差クリーニング法を提案した．実験では，交差クリーニングの場合，BLEU スコアが 0.045 改善し，主観評価では 11%の文について翻訳品質が向上した．これは従来法に比べ，大幅な改善である．

謝辞 類語新辞典分類体系の研究利用を許可してくださった，(株）角川書店に感謝いたします．

また，本研究を進めるにあたって有意義な議論をさせていただいた ATR 音声言語コミュニケーション研究所の皆様，および奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座の皆様に感謝いたします．

なお，本研究は情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものです．

参考文献

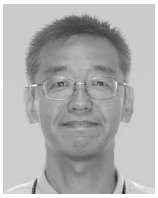
1) Akiba, Y., Imamura, K. and Sumita, E.: Using Multiple Edit Distances to Automatically Rank Machine Translation Output, *Proc. Machine Translation Summit VIII*, pp.15–20 (2001).
2) 荒牧英治，黒橋禎夫，佐藤理史，渡辺日出雄：用例ベース翻訳のための対訳文の句アライメント，自然言語処理，Vol.10, No.5, pp.75–92 (2003).
3) Brown, P.F., Pietra, S.A.D., Pietra, V.J.D. and Mercer, R.L.: The Mathematics of Machine Translation: Parameter Estimation, *Computational Linguistics*, Vol.19, No.2, pp.263–311 (1993).
4) Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-

Occurrence Statistics, *Proc. HLT Conference*, San Diego, California (2002).
5) 古瀬 蔵，山本和英，山田節夫：構成素境界解析を用いた多言語話し言葉翻訳，自然言語処理，Vol.6, No.5, pp.63–91 (1999).
6) Ikehara, S., Shirai, S., Yokoo, A. and Nakaiwa, H.: Toward an MT System without Pre-Editing—Effects of New Methods in ALT-J/E, *Proc. MT Summit III*, pp.101–106 (1991).
7) 今村賢治：構文解析と融合した階層的句アライメント，自然言語処理，Vol.9, No.5, pp.23–42 (2002).
8) Imamura, K.: Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment for Pattern-based MT, *Proc. 9th Conference on Theoretical and Methodological Issues in Machine Translation* (*TMI-2002*), pp.74–84 (2002).
9) Kaji, H., Kida, Y. and Morimoto, Y.: Learning Translation Templates from Bilingual Text, *Proc. COLING-92*, pp.672–678 (1992).
10) Kikui, G., Sumita, E., Takezawa, T. and Yamamoto, S.: Creating Corpora for Speech-to-Speech Translation, *Proc. EuroSpeech 2003*, pp.381–384 (2003).
11) Menezes, A. and Richardson, S.D.: A best first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora, *Proc. 'Workshop on Example-Based Machine Translation' in MT Summit VIII*, pp.35–42 (2001).
12) Meyers, A., Kosaka, M. and Grishman, R.: Chart-Based Translation Rule Application in Machine Translation, *Proc. COLING-2000*, pp.537–543 (2000).
13) Nagao, M.: A Framework of Mechanical Translation between Japanese and English by Analogy Principle, *Artificial and Human Intelligence*, Amsterdam: North-Holland, pp.173–180 (1984).
14) Nagao, M., Tsujii, J. and Nakamura, J.: Machine Translation from Japanese into English, *Proc. IEEE*, Vol.74, No.7, pp.993–1012 (1986).
15) Och, F.J.: Minimum Error Rate Training in Statistical Machine Translation, *Proc. 41st Annual Meeting of the Association for Computational Linguistics*, Hinrichs, E. and Roth, D. (Eds.), pp.160–167 (2003).
16) Och, F.J. and Ney, H.: Discriminative Training and Maximum Entropy Models for Statistical Machine Translation, *Proc. 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.295–302 (2002).
17) 大野 晋，浜西正人：類語新辞典，角川書店 (1984).
18) Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a Method for Automatic Evaluation of Machine Translation, *Proc. 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.311–318 (2002).
19) Paul, M., Imamura, K., Sumita, E. and Yamamoto, S.: Topic-adaptation of Pattern-based MT Systems Using Feedback Cleaning, *Proc. Recent Advances in Natural Language Processing* (*RANLP-2003*), Borovets, Bulgaria, pp.364–368 (2003).
20) Su, K.-Y., Wu, M.-W. and Chang, J.-S.: A New Quantitative Quality Measure for Machine Translation Systems, *Proc. COLING-92*, pp.433–439 (1992).
21) Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H. and Yamamoto, S.: Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World, *Proc. 3rd International Conference on Language Resources and Evaluation* (*LREC 2002*), pp.147–152 (2002).
22) 山本和英，河井 淳，隅田英一郎，古瀬 蔵：単語と品詞の混合 n-gram を用いた形態素解析，情報処理学会第 54 回全国大会，1C-02 (1997).
23) 山本 薫，松本裕治：統計的係り受け結果を用いた対訳表現抽出，情報処理学会論文誌，Vol.42, No.9, pp.2239–2247 (2001).
24) 安田圭志，菅谷史昭，竹澤寿幸，山本誠一，柳田益造：対訳コーパスを用いた翻訳品質自動評価法，情報処理学会論文誌，Vol.43, No.7, pp.2108–2117 (2002).

(平成 15 年 10 月 22 日受付)
(平成 16 年 6 月 8 日採録)

今村 賢治（正会員）

1985 年千葉大学工学部電気工学科卒業．同年日本電信電話株式会社入社．2000 年より ATR 音声言語コミュニケーション研究所主任研究員，現在に至る．主として自然言語処理の研究・開発に従事．電子情報通信学会，言語処理学会，ACL 各会員．



隅田英一郎（正会員）

1982年電気通信大学大学院計算機科学専攻修士課程修了．1999年京都大学工学博士．ATR音声言語コミュニケーション研究所主任研究員．機械翻訳，情報検索，eラーニングを研究．ACM/Transactions on Speech and Language Processingの Associate Editorを担当．電子情報通信学会，言語処理学会，日本音響学会，ACL会員．

松本　裕治（正会員）

1977年京都大学工学部情報工学科卒業．1979年同大学大学院工学研究科修士課程情報工学専攻修了．同年電子技術総合研究所入所．1984年～1985年英国インペリアルカレッジ客員研究員．1985年～1987年（財）新世代コンピュータ技術開発機構に出向．京都大学助教授を経て，1993年より奈良先端科学技術大学院大学教授，現在に至る．工学博士．専門は自然言語処理．人工知能学会，日本ソフトウェア科学会，言語処理学会，認知科学会，AAAI，ACL，ACM各会員．