Vol.2014-NL-219 No.19
2014/12/17

パラフレーズを考慮した機械翻訳の誤り箇所選択

赤部 晃一[1,a)]　Graham Neubig[1,b)]　Sakriani Sakti[1,c)]　戸田 智基[1,d)]　中村 哲[1,e)]

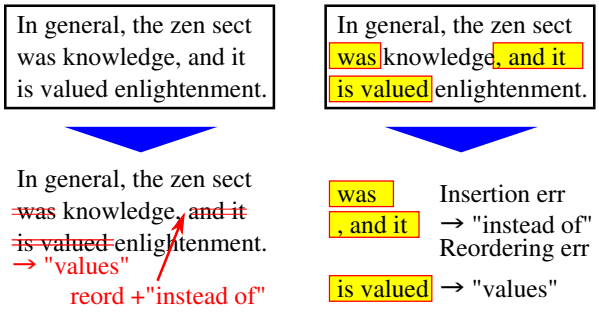
概要：
機械翻訳の誤り分析は，翻訳システムを改善する上で必要不可欠であり，誤り分析を効率化する様々な手法が提案されている。これらの手法の一例として，訳出と参照訳の差分を自動的に抽出するものや，機械学習の枠組みで誤り傾向の強い素性を抽出する手法等が提案されている。しかし，これらの先行研究では，同義語やパラフレーズを誤りとして誤選択する傾向があった。本研究ではまず，誤り箇所選択の精度を自動的に評価する手法を新たに提案する。次にパラフレーズを考慮した誤り箇所選択手法を提案する。提案法により誤り箇所選択精度が大きく向上した。

1. はじめに

機械翻訳システムの性能は年々向上しているが，一方でシステムの内部は非常に複雑化している。その結果，システムへの改良が翻訳結果に与える影響は必ずしも事前に把握できるわけではなく，実際に翻訳を行ってその結果を分析し，システムを改善することが広く行われている。翻訳結果に実際に目を通すことで，自動評価尺度だけでは分からない知見を多く得ることができる。

従来の機械翻訳の誤り分析では，まず事前に複数のテスト文に対して機械翻訳を行い，機械翻訳の専門家（以降，分析者）が専用の誤り体系 [23] などにしたがって分析を行う（図 1(a)）。しかし多くの場合，誤り傾向を捉えるには大量の文を評価する必要があり，非常に時間のかかる作業となる。さらに，目を通す文字列の多くは誤りを含んでいないか，誤りだとしてもシステム全体に影響を及ぼすとは限らないものが多く含まれている。

この中で，誤りの可能性が高い箇所に自動的にアノテーションを行うことができれば，誤り分析を行う専門家はアノテーションされた場所に集中することができ，重要な誤りをより効率的に見つけ出すことができる（図 1(b)）。このようなアノテーションを実現するため，先行研究ではルールに基づく誤り箇所の選択手法 [19] や，機械翻訳の n-best から学習された識別言語モデルの重みに基づく手法 [1] が



(a) 一般的な誤り分析　(b) 誤り箇所アノテーションあり

図 1 一般的な誤り分析と事前に誤り箇所をアノテーションした場合の誤り分析

提案されている。さらに文献 [24] は複数の尺度（頻度に基づく手法，相互情報量に基づく手法，条件付き確率に基づく手法，識別言語モデルの重みに基づく手法）で誤り候補となる n-gram を選択し，各手法の比較を行っている。

しかし，これらの先行研究にはまだ多くの課題が残る。特に本研究では 2 つの課題に着目する。まず，先行研究は単純に正解訳と機械翻訳の差異に基づいているため，表層的に異なりながら同等の意味を持つ訳出を誤って誤訳として扱うことが多い。この問題を解決するために，本研究では誤り箇所の選択時に参照訳のパラフレーズを考慮し，参照訳のパラフレーズに含まれている n-gram を誤り箇所の候補から除外する枠組みを導入する。これにより，同義語を誤りとして選択してしまうことを避け，実際に誤っている箇所を重点的に捉えることが可能となる。

次に，誤り箇所選択手法の新しい自動評価方法を提案する。今までの研究では，選択された誤り箇所を実際に人手で調査すること [1] や，誤り箇所を直接アノテーションした

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[a)] akabe.koichi.zx8@is.naist.jp
[b)] neubig@is.naist.jp
[c)] ssakti@is.naist.jp
[d)] tomoki@is.naist.jp
[e)] s-nakamura@is.naist.jp

コーパス [9] を用いて自動的に精度を評価すること [5], [10] で，誤り箇所選択法が評価されてきた。しかし，人手による分析は効率が悪く評価の揺れも大きくなり，誤り箇所を直接アノテーションしたコーパスの作成は，機械翻訳の専門家による作業でも非常に時間がかかる上，並べ換え誤りによるフレーズの移動を追うことも困難である。この問題を解決するために，本研究では代わりに後編集とアライメントに基づく手法を提案する。後編集の作成，及び機械翻訳と後編集のアライメント作業は，機械翻訳の専門家でなくても翻訳者なら比較的容易に行うことができる上，アノテーションの誤りも発見しやすく，正確な評価を行うことが可能となる。

2. *n*-gram に基づく誤り箇所選択

本節では，まず *n*-gram に基づく誤り箇所選択手法 [1], [24] について説明する。*n*-gram に基づく誤り箇所選択では，図 1(b) に示すように，誤り分析が必要な場所を訳出結果の *n*-gram 統計を用いて特定する。具体的な手続きは以下の通りである。

(1) 分析対象のデータに対して翻訳結果を生成する。
(2) *n*-gram を翻訳結果全体から取り出し，参照訳を参考にしながら *n*-gram に対してスコア付けを行う。スコア付けの手法は下記の通り様々であるが，基本的には誤り箇所で現れる可能性の高い *n*-gram には高いスコア，現れる可能性の低い *n*-gram には低いスコアを付与するように設計する。
(3) *n*-gram をスコアの降順にソートし，誤り分析を行う作業者が，各 *n*-gram が現れる翻訳結果に対して誤り分析を行う。

ここで，翻訳結果全体を判断材料として *n*-gram のスコアを計算しているため，低頻度で偶然参照訳と異なった *n*-gram を除外し，何回も参照訳と異なり誤りの可能性が高い *n*-gram を重点的に分析することができる。またこの手法の利点として，選択に翻訳システムの内部情報（翻訳ルールの ID 等）ではなく訳出の *n*-gram を用いることで，特定の機械翻訳システムに依存しない誤り箇所選択が可能となる。

n-gram の選択順序はスコアの計算方法によって大きく異なり，初期に選択される箇所に誤りが多く含まれていれば，誤り分析の作業効率が上がる。一方不適切な選択を行うと，評価者は誤り分析が不要な箇所を多く見ることとなる。文献 [24] では，以下の 4 つの基準でスコアを計算し，各手法による誤り箇所の選択精度を調査している。

頻度に基づく選択　文ごとに機械翻訳に含まれ，正解訳に含まれない *n*-gram を調べ，その合計回数が最も多い *n*-gram を順に選択。

自己相互情報量に基づく選択　1-best 訳出と *n*-gram の間の自己相互情報量 [6] を計算し，その値を *n*-gram の出現回数で正規化した値が最も大きいものを順に選択。

平滑化された条件付き確率に基づく選択　*n*-gram が訳出に含まれる確率を，ラプラス平滑化（add-one）した条件付き確率として計算し，その値が最も大きい *n*-gram を順に選択。

識別言語モデルの重みに基づく選択　*n*-gram を素性とし，構造化パーセプトロン [7] により 1-best 訳出を正解訳に修正するように学習された識別言語モデルを作成し，各素性の重みが最も負に大きい *n*-gram を順に選択。

ここでの「正解訳」には参照訳を用いるか，*n*-best の中で BLEU+1[12] などの機械翻訳の自動評価尺度が最大となるオラクル文を用いる。また，手法により選択された *n*-gram が参照訳にも含まれる場合，その *n*-gram は誤選択である可能性が高いため誤りの候補から除外する。

本研究ではこれらの手法に対し，次節以降で説明する提案法を適用する。

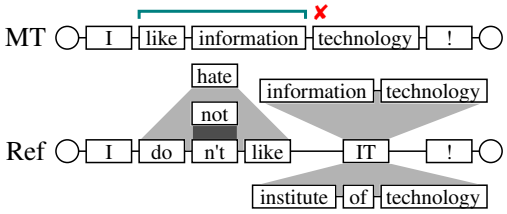


図 2　参照訳ラティスによる候補のフィルタリング

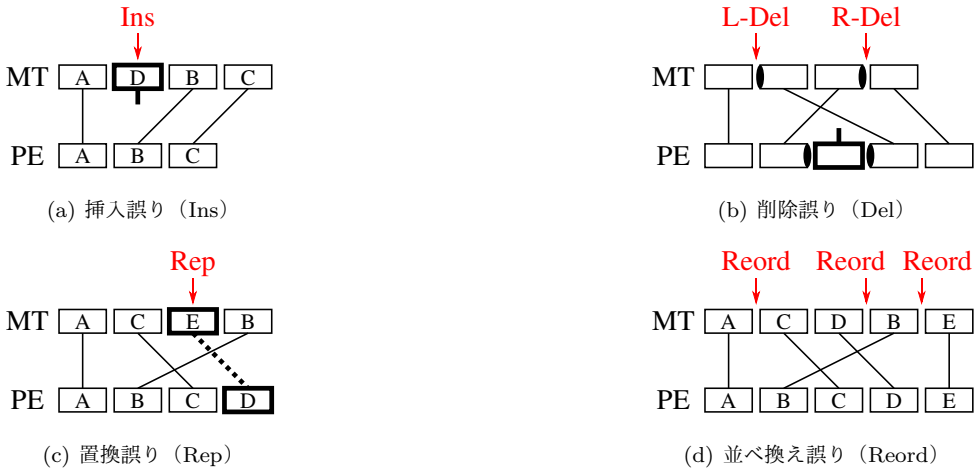
3. 言い換えを考慮した誤り箇所選択

先行研究では，正解訳と 1-best 訳出の表層的な差異に基づいて *n*-gram を選択しているが，この場合，表層的に異なりながら同等の意味を持つ単語列（パラフレーズ）も誤りとして選択されてしまう。この問題を解決するために，本研究では [24] で提案されている誤り箇所選択手法に対し，以下の 2 つの手法によるパラフレーズの導入を提案し，選択精度の向上を図る。

3.1 パラフレーズを用いた誤り箇所フィルタリング

誤りの可能性が高いとして選択された *n*-gram であっても，参照訳の中に同等の意味を持つ表現が含まれている場合，選択された *n*-gram はその文脈では誤りでない可能性が高い。本研究では，参照訳の中に含まれている *n*-gram を分析の対象から除外する先行研究の手法を拡張し，参照訳だけでなく参照訳のパラフレーズもフィルタリングの対象となる手法を提案する。具体的には，文献 [4] の手法により作成されたパラフレーズのデータベースを用いて，参照訳のパラフレーズラティス [17] を生成し，ラティス上に存在する *n*-gram を誤りの候補から除外する。

図 2 にフィルタリングの例を示す。参照訳として “I don’t like IT!” が与えられている中，機械翻訳結果が “I like information technology!” となり，“like information”



(a) 挿入誤り（Ins）　(b) 削除誤り（Del）

(c) 置換誤り（Rep）　(d) 並べ換え誤り（Reord）

図 3　アライメントによるラベル付与の例。評価者が赤矢印の位置を見ることは，分析により太線のアライメントが検出されることに対応する。

が誤りの候補として挙げられたとする。提案法では，まず参照訳に含まれる全ての部分単語列をパラフレーズデータベースの中で検索し，ある閾値以上の確率で置換可能なパラフレーズを抽出する。次に，抽出されたパラフレーズを利用して参照訳のラティスを構築する。最後にラティス上を探索し，誤りの候補として挙げられた n-gram “like information” が見つかった場合は，この n-gram を誤りの候補から除外する。

3.2　評価尺度へのパラフレーズの導入

2 節で述べたように，誤り n-gram のスコア付けを行う際に正解文としてオラクル文を利用する。このオラクル文を利用する際に，従来法では BLEU+1 を利用したが，BLEU+1 も n-gram の表層的な違いのみ着目する評価尺度のため，最適でないオラクルを選択する可能性がある。

本研究では，このオラクル文を選択する際に，パラフレーズを考慮した尺度である METEOR[3] を利用する。METEOR を導入した場合，正解訳として選ばれるオラクル文が参照訳の言い回しに縛られず，1-best 訳出に近い言い回しになることが期待される。このため，正解訳と 1-best 訳出の差分にパラフレーズが含まれにくくなると考える。

4.　誤り箇所選択の評価

本節では，誤り箇所選択手法を評価するための自動評価指標を提案する。機械翻訳システムの分析者が誤り分析を行う際，2 節で述べたように誤り箇所選択システムが順に出力した n-gram を中心に翻訳結果を分析する。本節で提案する評価尺度は，この作業をシミュレーションして評価を行う。具体的には，誤り箇所の正解ラベルを付与したコーパスを用意し，誤り箇所選択システムをこのコーパスに適用して，誤りと判断された箇所が実際に誤っている割合（適合率）と，コーパスに含まれている誤り全体に対して実際に検出された割合（再現率）を計算する。

表 1　機械翻訳-後編集の対応コーパスの例。この例では，「この 頃」-「今」，「風穴」-「溥傑」，「延昭」-「enshou」の置換誤り，「人物」の削除誤り，「風穴 延昭」と「enshou 溥傑」の並べ替え誤りが発生している。

原文 $\boldsymbol{f}$	the central figure around that time was enshou fuketsu .
訳出 $\boldsymbol{e_{MT}}$	今 の 中心 は enshou 溥傑 。
後編集 $\boldsymbol{e_{PE}}$	この 頃 の 中心 人物 は 風穴 延昭 。
対応付け $\boldsymbol{a}$	(1)-(1,2)　(2)-(3)　(3)-(4)　-(5)　(4)-(6)　(5)-(8)　(6)-(7)　(7)-(9)
誤り $\boldsymbol{t_{err}}$	$a_{1\text{-}1,2}$: 置換　$a_{5\text{-}8}$: 置換　$a_{6\text{-}7}$: 置換　$a_{\text{-}5}$: 削除　$p_{4.5}$: 並べ替え　$p_{5.5}$: 並べ替え　$p_{6.5}$: 並べ替え
紐付け $\boldsymbol{t_p}$	1:$a_{1\text{-}1,2}$　3.5: $a_{\text{-}5}$　5: $a_{5\text{-}8}$　6: $a_{6\text{-}7}$　4.5: $p_{4.5}$　5.5: $p_{5.5}$　6.5: $p_{6.5}$

翻訳誤りのアノテーション済みコーパス [9] は既に存在し，誤りの自動分類タスクにおいて利用されている [5], [10]。我々も本研究で同様の枠組みで評価できると考え，予備実験でコーパスの作成を試みたが，2 つの大きな欠点が明らかになった。

(1) 機械翻訳結果に対して直接誤りをアノテーションを行う作業は，機械翻訳の専門家でないと困難であり，非常に労力がかかる。

(2) アノテーションを行ったとしても，並べ替え・削除誤りがどのように発生しているのかが明確ではない。

そこで本研究では，効率的かつ正確な評価を行うために，後編集とアライメントに基づく評価手法を新たに提案する。

4.1　機械翻訳-後編集の対応コーパスの作成

本研究で用いる評価法には，表 1 に示すような，機械翻訳結果 $\boldsymbol{e_{MT}}$ と後編集結果 $\boldsymbol{e_{PE}}$ とその対応付け $\boldsymbol{a}$ からなるデータを利用する。$\boldsymbol{a}$ は，対応しているフレーズであり，

対応するものがない場合は挿入または削除すべきフレーズとしてアノテーションされている。$\boldsymbol{e_{PE}}$ は通常の人手による後編集により得られ，$\boldsymbol{a}$ は後編集を行った翻訳者に $\boldsymbol{e_{MT}}$ と $\boldsymbol{e_{PE}}$ の対応している部分をアノテーションしてもらうことで，比較的容易に入手できる。

4.2 後編集アライメントを利用した誤りラベル付与

誤り箇所選択システムは，機械翻訳結果の中で誤りと推定された n-gram を分析者に提示する。その際，提示された位置から実際の誤りを特定できなければならない。これを踏まえ，訳出 $\boldsymbol{e_{MT}}$，後編集 $\boldsymbol{e_{PE}}$，アライメント $\boldsymbol{a}$ を利用して，ある n-gram を分析者に見せた際，分析者が発見できる誤りを特定するために必要な情報を格納した $\boldsymbol{t}$（表 1 下部）を生成する。まず，本枠組みは大まかに以下の手続きからなる。

(1) 全てのアライメントに対し，誤りがある場合は，対応する誤り（挿入・削除・置換誤り）を最大で 1 個ラベル付けする。各アライメントにおけるラベルは以下の基準に従い付与する。
挿入誤り $\boldsymbol{e_{PE}}$ 側に単語が存在しないアライメント。
削除誤り $\boldsymbol{e_{MT}}$ 側に単語が存在しないアライメント。
置換誤り $\boldsymbol{e_{MT}}$ 側と $\boldsymbol{e_{PE}}$ 側でフレーズが異なるアライメント。

(2) 全てのアライメントを，関連する機械翻訳結果 $\boldsymbol{e_{MT}}$ 中の位置 $p \in \{0.5, 1, 1.5, 2, 2.5, \cdots\}$ に紐付ける（.5 は単語間を表し，文頭記号を 0 としている）。

(3) 並べ換え誤りを，機械翻訳結果 $\boldsymbol{e_{MT}}$ 中の単語間 p に直接紐付ける。

p とアライメントの紐付けは，分析者が $\boldsymbol{e_{MT}}$ 中の位置 p を見た際に，その位置から把握可能な誤りを特定することに対応し，以下の手順に従い行う。

(1) 挿入誤りと置換誤りは，アライメントが含む $\boldsymbol{e_{MT}}$ 側のすべての単語位置 $p \in \{1, 2, 3, \cdots\}$ に紐付ける。（図 3(a)，(c)）

(2) 削除誤りは，まず $\boldsymbol{e_{PE}}$ 側で左側または右側に隣接するアライメントで削除誤りでないものを見つけ，そのアライメントの $\boldsymbol{e_{MT}}$ 側における右側または左側の単語間 $p \in \{0.5, 1.5, 2.5, \cdots\}$ に紐付ける。（図 3(b)）

(3) 並べ換え誤りはアライメントに対してではなく，$\boldsymbol{e_{MT}}$ の単語間 p に直接誤りとしてラベル付けする。並べ換え誤りは，挿入・削除誤りを無視した際に，$\boldsymbol{e_{MT}}$ 側で $\boldsymbol{e_{PE}}$ と順序が異なるアライメント間 $p \in \{0.5, 1.5, 2.5, \cdots\}$ にラベル付けする。（図 3(d)）

4.3 n-gram による誤り箇所選択の評価

本節では，前節の誤りラベル付きデータを用いて，ある誤り箇所選択法により選択された n-gram を評価する方法について述べる。評価では，誤り箇所の候補として提示された n-gram を順に選択していき，選択された箇所の中で実際に誤りを含む箇所の数 T と誤りを含まない箇所の数 F，そしてコーパス中の誤りの数 A を用いて適合率 P と再現率 R を評価する。

$$P = \frac{T}{T+F}, \quad R = \frac{T}{A} \tag{1}$$

T の定義として，「n-gram により，紐付けられた場所が 1 つ以上分析対象となった誤りの数」とする。F の定義は，「誤りとなっていないが分析対象となった単語の数」とするが，一度分析した場所を再度分析する必要はなく，また単語間の誤りを分析した時に隣接する単語を選択することが誤選択扱いにならないようにするため，以下の 2 つの場合に当てはまる箇所を F の数に含まない。

(1) 現在見ている位置に紐付けられたアライメント・並べ替え誤りが，既に別の n-gram により評価されている場合。

(2) 現在見ている位置に誤りが無く，隣接している単語・単語間に誤りがあり，しかもそれらの位置が n-gram により同時に選択されている場合。

実際の選択例を図 4 に示す。(a) は 2-gram による選択で，不正解箇所の単語を 2 つ選択しているため F が 2 点増える。(b) は 3-gram による選択で，単語間の誤りを 2 つ（並べ換え誤りと削除誤り）選択しており T が 2 点，その両端の単語は不正解だが，正解ラベルに隣接するため評価しない。最右の単語は選択中の単語の中では正解ラベルが隣接していないため F が 1 点増える。(c) も同様にして T が 4 点増える。

表 2 KFTT のデータサイズ

	文数	単語数	
		英語	日本語
Train	330k	5.91M	6.09M
Dev	1166	24.3k	26.8k
Test	1160	26.7k	28.5k

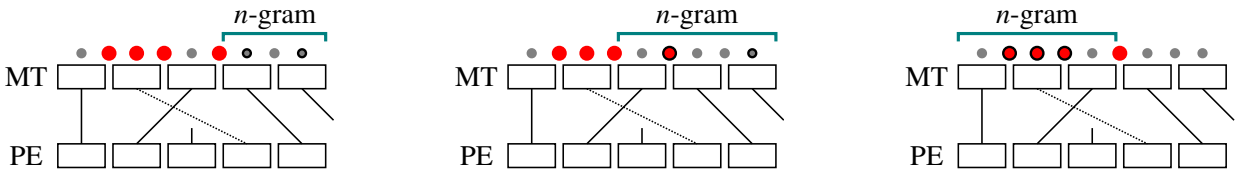
5. 実験

各手法の有効性を検証するために，機械翻訳の訳出を利用した評価実験を行った。

5.1 実験設定

すべての実験で京都フリー翻訳タスク (KFTT)[14] の英日翻訳結果を利用した。コーパスの大きさを表 2 に示す。評価対象とした翻訳システムは Travatar ツールキット [15] に基づく Forest-to-String システムである。チューニングには MERT [16] を利用し，評価尺度を BLEU[18] とした。

評価は，2 節で説明した頻度に基づく手法，自己相互情報量に基づく手法，平滑化された条件付き確率に基づく手法，識別言語モデルの重みに基づく手法の 4 つの手法に対して行った。先行研究では，モデルの学習時の正解訳とし



(a) True: 0 False: 2　(b) True: 2 False: 1　(c) True: 4 False: 0

図 4 *n*-gram による誤り箇所選択の例。赤丸は誤り，灰色はそれ以外の箇所を示す。黒縁の丸は，評価済みの箇所。

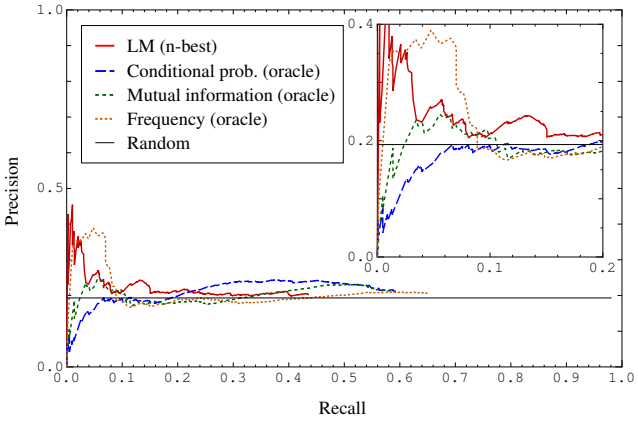


図 5 ベースライン手法の評価結果

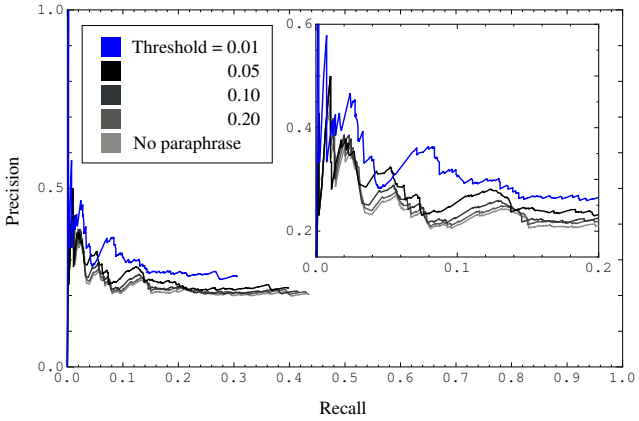


図 6 パラフレーズによる誤り候補フィルタリングを適用した結果

て，*n*-best の中で BLEU+1[12] が最大となるオラクル文を利用しており，今回の実験はこれをベースラインとした。

次に各提案法の実験設定を示す。

誤り候補のフィルタリング 3.1 節の手法では，日本語のパラフレーズとして，機械翻訳の翻訳テーブルに基づき作成された日本語言い換えデータベース 0.2.0[13] を利用し，言い換え確率の閾値を 0.01〜1.0（1.0 では言い換えしない）の間で変化させて実験を行った。

オラクル選択尺度の変更 3.2 節の手法では，オラクル文を選択する際に，評価尺度として METEOR version 1.5 [8] を利用した。METEOR の学習データは日本語言い換えデータベースと同一とし，パラメーターのチューニングは行わなかった。

n-best による識別言語モデルの学習は，反復回数を 100 回とした。先行研究と同様に L1 正則化 [22] を行い，正則化係数は 10^{-7}-10^{-2} の中から KFTT のテストデータに対して高い精度を示す値（0.00015）を利用した。すべての手法で，*n*-gram として 1-gram から 3-gram までを利用した。

後編集とアライメントのデータを KFTT の Dev セット内 200 文（4846 単語）に対して作成し，このコーパスを評価に用いた。

5.2 ベースライン手法の評価結果

先行研究の誤り箇所選択手法を，4 節で提案した評価手法に適用した場合の適合率-再現率曲線を図 5 に示す。この図でいずれの手法も再現率が 1.0 に達していないが，これは誤り箇所と同一のフレーズが参照訳にも含まれている場合，誤り箇所を検出できなくなるためである。先行研究の分析対象が日英翻訳であるのに対し，今回の分析対象が英日翻訳のため一概に比較できないが，この結果から，先行研究と同様に識別言語モデルの重みに基づく手法（LM）で初期に選ばれる *n*-gram が，実際の誤り箇所を他の手法に比べて適切に捉えていることが分かる。

先行研究では平滑化された条件付き確率に基づく手法も有効としているが，本結果を見ると初期に選ばれる *n*-gram の適合率がランダム選択よりも低いことが分かる。実際に選択された *n*-gram を調査した結果，この原因として英日翻訳特有の問題である和暦・西暦の翻訳が挙げられる。正解訳が西暦で，訳出が和暦の場合，本手法では和暦を誤りとして捉えてしまうが，実際は誤りでないため誤選択となる。この問題も一種のパラフレーズ問題であり，パラフレーズを考慮する重要性を裏付ける結果とも言える。

頻度に基づく手法は初期の適合率が高いが，再現率が 0.1 を超える前に極端に低下していることが分かる。これは選択される *n*-gram が多数の箇所に存在し，*n*-gram を 1 つ選択する度に結果が大きく影響を受けるためである。

5.3 パラフレーズによる誤り候補フィルタリングの効果

次に，識別言語モデルの重みに基づく手法に対してパラフレーズによる誤り候補フィルタリングを適用した結果を，図 6 に示す。この結果から，言い換え確率の閾値を下げて利用可能なパラフレーズを増やすほど，誤り箇所の適合率

表 3　パラフレーズによるフィルタリングの例。訳出の黒枠は，パラフレーズを考慮しない場合に誤り箇所として提示されるが，パラフレーズによるフィルタリングを適用すると候補から除外される。

1	原文	... the members of the kanoha group were ... and castles as the shogunate 's official painters ...
	訳出	··· 狩野 派 の メンバー は 幕府 の 御 用 絵師 と し て ···
	参照訳	··· 狩野 派 は 、 幕府 の 御 用 絵師 と し て 、 ···
	貢献したパラフレーズ	、 幕府 ⟶ は 幕府
2	原文	... , rinzai school started with its founder gigen rinzai at the end of the ...
	訳出	··· 臨済 kaishou 年間 の 廃仏 毀釈 の 後 、 唐 の 創業 者 の 臨済 義玄 と とも に ···
	参照訳	··· 臨済 宗 は 会昌 の 廃仏 毀釈 運動 の 後 、 唐末 の 宗祖 臨済 義玄 に よ り ···
	貢献したパラフレーズ	の 宗祖 ⟶ の 創業 者

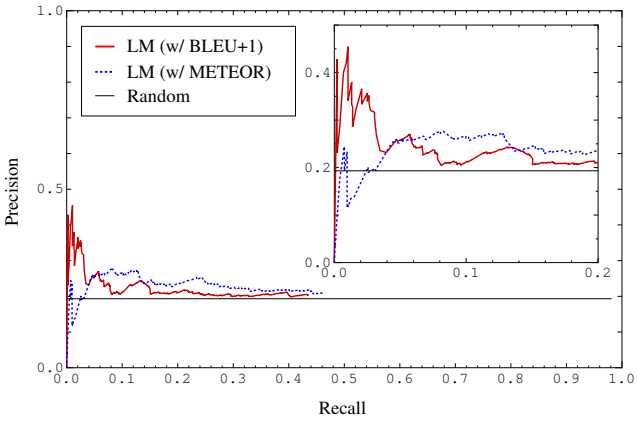


図 7　オラクル文を選択する際の評価尺度を変更した結果

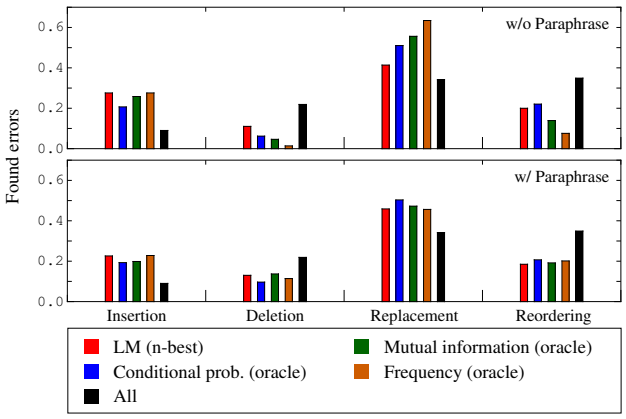


図 8　各手法により検出された誤りの内訳

が向上していることが分かる。一方，再現率の上限は閾値を下げるほど低下している。この原因は，参照訳のパラフレーズであっても誤訳を含む場合があり，正解の n-gram の一部が誤って除外されてしまうためである。

表 3 は実際にフィルタリングされた候補の例である。1 つ目の例では，訳出の中で n-gram により選択された箇所は参照訳と同等の意味であるが，参照訳には読点が含まれてしまっているため表層上は異なった文字列である。しかしパラフレーズデータベースを適用することで読点が削除され，結果的に参照訳に一致して誤りの候補から除外された。2 つ目の例は，実際に誤り箇所を捉えている n-gram が，逆にパラフレーズにより除外されてしまった例である。この例のように，本来捉えられるべき誤り箇所がパラフレーズにより除外されることは，再現率の上限が低下する原因となる。

5.4　オラクル文を METEOR に基づき選択する効果

次に，識別言語モデルの重みに基づく手法について，オラクル文を選択する際の評価尺度を変更した場合の結果を図 7 に示す。BLEU+1 を利用した場合と METEOR を利用した場合を比較すると，最初に選択される n-gram は BLEU+1 の方が高い適合率を示すものの，再現率が 0.1 を超える前に両者の違いは見られなくなる。この理由として，今回の実験で利用した翻訳システムが Dev セットに対して BLEU で最適化されていることが挙げられる。

BLEU はパラフレーズを考慮しない尺度のため，最適化されたシステムでは，n-best 中の上位の候補で似通ったフレーズが使用されやすい傾向となる。このため，正解訳と 1-best 訳出の間でパラフレーズを考慮した比較が機能しなかったと考えられる。

5.5　各手法により選択された誤りの傾向

最後に，各手法により検出された誤りの内訳を図 8 に示す。データは，言い換え確率の閾値が 0.01 で，再現率が 0.1 の時点のものを利用した。上段のグラフは，誤り箇所候補のフィルタリングを参照訳のみ用いて行った場合（ベースライン），下段のグラフは参照訳のパラフレーズを考慮した場合である。黒棒は，コーパス中に含まれるすべての誤りの割合を示し，自動的に選択された誤り箇所がこれに近ければ近いほど，自動選択による分布の偏りが少なく良い結果と言える。この結果から，すべての手法が挿入誤りと置換誤りに偏って検出する傾向にあることが分かるが，特に頻度に基づく手法において，提案法ではこの問題が改善されていることが分かる。これは，参照訳のパラフレーズによるフィルタリングにより，置換誤りが誤検出されにくくなったためと考えられる。

これを定量的に評価するために，表 4 に実際に検出された誤りの割合と，コーパス中に含まれる誤りの割合の差分

表 4 各手法で検出された誤り傾向と実際の誤り傾向の違い。スコアの小さい方を手法ごとに太字で示した。

手法	参照訳のみ	パラフレーズ考慮
識別言語モデル	0.516	**0.506**
条件付き確率	0.571	**0.530**
相互情報量	0.766	**0.479**
頻度	0.957	**0.506**

の絶対値の和を各手法について計算した結果を示す。このスコアは，実際に検出された誤りの比率がコーパス中に含まれる誤りの比率に近いほど 0 に近くなる。この結果からすべての手法において，参照訳のパラフレーズによるフィルタリングにより，実際のコーパス中に含まれる誤り傾向に近い比率で誤りを検出できたことが分かる。

6. 先行研究

本研究が行う誤り箇所選択に近い技術として，精度推定（QE: Quality Estimation）が挙げられる [2]。QE は，文単位，あるいは個々の箇所について誤訳を特定しようとしている意味では本研究と類似しているが，参照訳が与えられていない機械翻訳を対象としているため，本研究と目的も手法も大きく異なる。

参照訳が与えられた状況では，BLEU+1，METEOR，TER[21]，RIBES[11] などの機械翻訳の自動評価尺度が利用可能である。これらの自動評価尺度は分析のためのモデルを作成する必要がなく汎用性が高いが，一般に文単位の評価に限定される。

本研究で扱っている手法 [1], [20], [24] は参照訳が与えられた状況で，文単位ではなく翻訳結果の各々の箇所について評価を行うものであり，この手法を利用することで，より効率的な誤り分析が可能となる。

7. まとめ

本研究では，機械翻訳の誤り箇所を提示する際に，参照訳のパラフレーズを用いて選択精度の向上を図る手法を提案した。その結果，識別言語モデルの重みに基づく手法と条件付き確率に基づく手法において，提案法の有効性が確認された。また，本研究では誤り箇所をアノテーションしたコーパスを作成し，評価の一貫性を保ち，評価の高速化を行った。コーパスは再利用可能なため，今後の新たな誤り分析手法の研究・開発に貢献すると思われる。

今後の課題として，構文情報や品詞情報を利用した誤り分析への応用，誤り箇所選択を利用した実際の誤り修正での利用価値を示すことがある。また，今回の実験では翻訳システムは全て BLEU で最適化したものを利用したが，今後 METEOR などのパラフレーズを考慮した評価尺度で最適化した場合の，提案法の有効性を検証する。

謝辞

本研究の一部は，JSPS 科研費 25730136 と（独）情報通信研究機構の委託研究「知識・言語グリッドに基づくアジア医療交流支援システムの研究開発」の助成を受け実施したものである。

参考文献

[1] Akabe, K., Neubig, G., Sakti, S., Toda, T. and Nakamura, S.: Discriminative Language Models as a Tool for Machine Translation Error Analysis, *Proc. COLING*, pp. 1124–1132 (2014).
[2] Bach, N., Huang, F. and Al-Onaizan, Y.: Goodness: A Method for Measuring Machine Translation Confidence, *Proc. ACL*, pp. 211–219 (2011).
[3] Banerjee, S. and Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (2005).
[4] Bannard, C. and Callison-Burch, C.: Paraphrasing with bilingual parallel corpora, *Proc. ACL*, pp. 597–604 (2005).
[5] Berka, J., Bojar, O., Fishel, M., Popovic, M. and Zeman, D.: Automatic MT Error Analysis: Hjerson Helping Addicter, *Proc. LREC* (2012).
[6] Church, K. W. and Hank, P.: Word association norms, mutual information, and lexicography, *Computational Linguistics*, pp. 22–29 (1990).
[7] Collins, M.: Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms, *Proc. EMNLP*, pp. 1–8 (2002).
[8] Denkowski, M. and Lavie, A.: Meteor Universal: Language Specific Translation Evaluation for Any Target Language, *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation* (2014).
[9] Fishel, M., Bojar, O. and Popović, M.: Terra: a Collection of Translation Error-Annotated Corpora., *Proc. LREC*, pp. 7–14 (2012).
[10] Fishel, M., Bojar, O., Zeman, D. and Berka, J.: Automatic translation error analysis, *Text, Speech and Dialogue*, Springer, pp. 72–79 (2011).
[11] Isozaki, H., Hirao, T., Duh, K., Sudoh, K. and Tsukada, H.: Automatic Evaluation of Translation Quality for Distant Language Pairs, *Proc. EMNLP*, pp. 944–952 (2010).
[12] Lin, C.-Y. and Och, F. J.: Orange: a method for evaluating automatic evaluation metrics for machine translation, *Proc. COLING*, pp. 501–507 (2004).
[13] Mizukami, M., Neubig, G., Sakti, S., Toda, T. and Nakamura, S.: Building a Free, General-Domain Paraphrase Database for Japanese, *Proc. COCOSDA* (2014).
[14] Neubig, G.: The Kyoto Free Translation Task, `http://www.phontron.com/kftt` (2011).
[15] Neubig, G.: Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers, *Proc. ACL Demo Track*, pp. 91–96 (2013).
[16] Och, F. J.: Minimum Error Rate Training in Statistical Machine Translation, *Proc. ACL*, pp. 160–167 (2003).
[17] Onishi, T., Utiyama, M. and Sumita, E.: Paraphrase Lattice for Statistical Machine Translation, *Proc. ACL*, pp. 1–5 (2010).
[18] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.:

BLEU: a method for automatic evaluation of machine translation, *Proc. ACL*, pp. 311–318 (2002).

[19] Popović, M.: Hjerson: An open source tool for automatic error classification of machine translation output, *The Prague Bulletin of Mathematical Linguistics*, Vol. 96, No. 1, pp. 59–67 (2011).

[20] Popović, M. and Ney, H.: Towards automatic error analysis of machine translation output, *Computational Linguistics*, Vol. 37, No. 4, pp. 657–688 (2011).

[21] Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J.: A study of translation edit rate with targeted human annotation, *Proc. AMTA*, pp. 223–231 (2006).

[22] Tibshirani, R.: Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, pp. 267–288 (1996).

[23] Vilar, D., Xu, J., d'Haro, L. F. and Ney, H.: Error analysis of statistical machine translation output, *Proc. LREC*, pp. 697–702 (2006).

[24] 赤部晃一, Neubig, G., Sakti, S., 戸田智基, 中村　哲：機械翻訳システムの詳細な誤り分析のための誤り順位付け手法, 情報処理学会第 216 回自然言語処理研究会 (SIG-NL), 東京 (2014).

正誤表

原稿執筆時に行った実験の結果に誤りがあったため，以下の通り訂正を行う。

誤	正

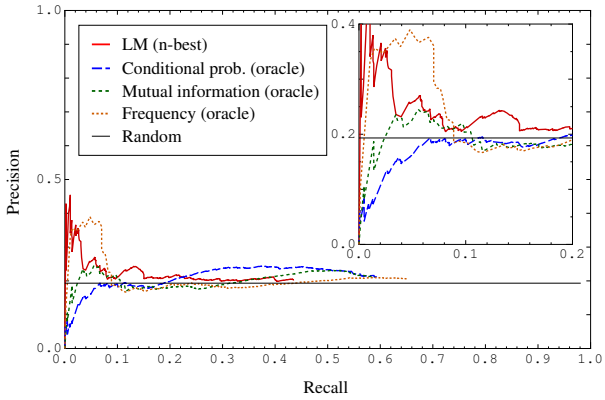
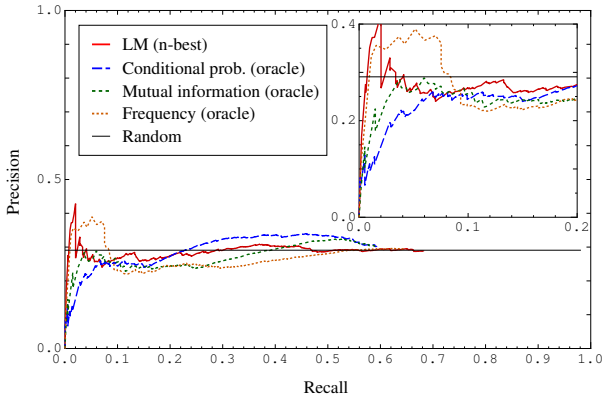

	

図 5　ベースライン手法の評価結果	**図 5**　ベースライン手法の評価結果
5.2　ベースライン手法の評価結果	
先行研究と同様に識別言語モデルの重みに基づく手法 (LM) で初期に選ばれる n-gram が，実際の誤り箇所を他の手法に比べて適切に捉えていることが分かる。	ごく初期に選択される n-gram に着目すると，識別言語モデルの重みに基づく手法 (LM) と誤り頻度に基づく手法において，他の手法に比べて誤り箇所の適合率が高いが，いずれの手法も再現率が 0.1 に達する前にランダム選択と大差がなくなることが分かる。先行研究 [1] や [24] では，識別言語モデルの重みに基づく手法が優位と述べられているが，本結果では先行研究とは異なる結果が得られた。この原因として，評価方法が先行研究と変わったことが挙げられる。本研究では，誤りアノテーションコーパスを用いることで再現率と適合率の間で評価を行っているが，先行研究では初期に選ばれる n-gram の一部に対してのみ人手による評価を行い，選択された n-gram の個数と適合率の間で評価を行っていたためと考えられる。

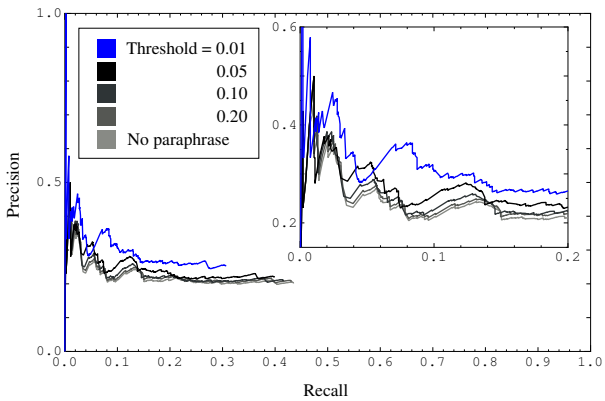
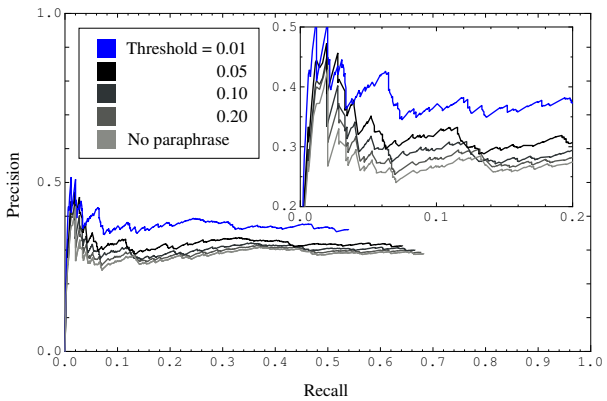

	

図 6　パラフレーズによる誤り候補フィルタリングを適用した結果	**図 6**　パラフレーズによる誤り候補フィルタリングを適用した結果

誤

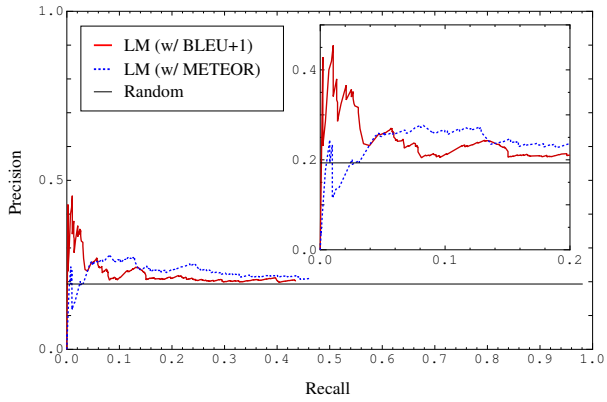


図 7　オラクル文を選択する際の評価尺度を変更した結果

正

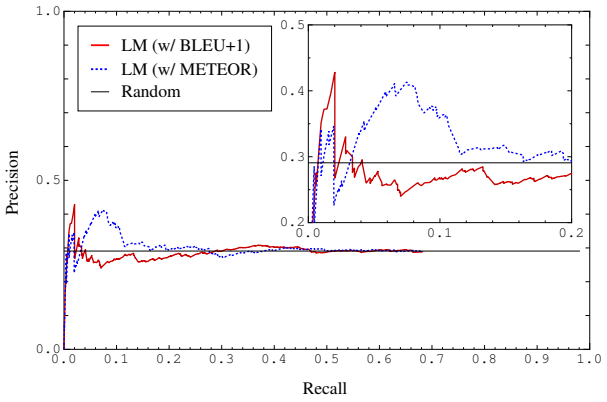


図 7　オラクル文を選択する際の評価尺度を変更した結果

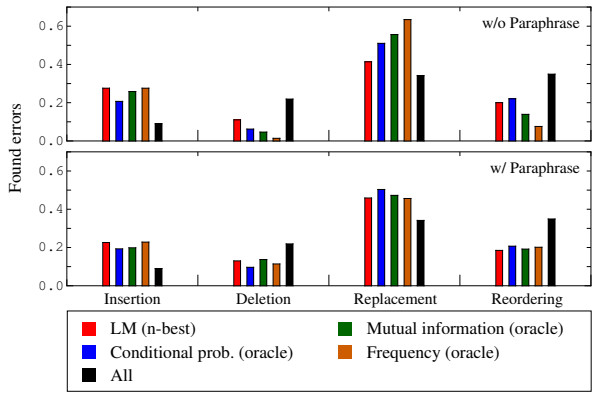


図 8　各手法により検出された誤りの内訳

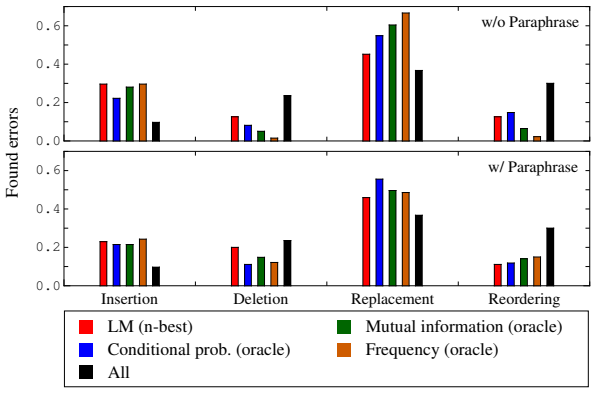


図 8　各手法により検出された誤りの内訳

表 4　各手法で検出された誤り傾向と実際の誤り傾向の違い

手法	参照訳のみ	パラフレーズ考慮
識別言語モデル	0.516	**0.506**
条件付き確率	0.571	**0.530**
相互情報量	0.766	**0.479**
頻度	0.957	**0.506**

表 4　各手法で検出された誤り傾向と実際の誤り傾向の違い

手法	参照訳のみ	パラフレーズ考慮
識別言語モデル	0.567	**0.449**
条件付き確率	0.612	0.612
相互情報量	0.841	**0.494**
頻度	0.997	**0.529**