機械翻訳システムの誤り分析のための誤り箇所選択手法

赤部 晃一[†]・Graham Neubig[†]・Sakriani Sakti[†]・戸田 智基[†]・中村 哲[†]

複雑化する機械翻訳システムを比較し、問題点を把握・改善するため、誤り分析が利用される。その手法として、様々なものが提案されているが、多くは単純にシステムの翻訳結果と正解訳の差異に着目して誤りを分類するものであり、人手による分析への活用を目的とするものではなかった。本研究では、人手による誤り分析を効率化する手法として、機械学習の枠組みを導入した誤り箇所選択手法を提案する。学習によって評価の低い訳出と高い訳出を分類するモデルを作成し、評価低下の手がかりを自動的に獲得することで、人手による誤り分析の効率化を図る。実験の結果、提案法を活用することで、人手による誤り分析の効率が向上した。

キーワード：統計的機械翻訳, 誤り分析, 誤り検出, 評価尺度, コーパス

Error Selection Methods for Machine Translation Error Analysis

Koichi Akabe[†], Graham Neubig[†], Sakriani Sakti[†], Tomoki Toda[†] and Satoshi Nakamura[†]

Error analysis is used to improve accuracy of machine translation (MT) systems. Various methods of analyzing MT errors have been proposed; however, most of these methods are based on differences between translations and references that are translated independently by human translators, and few methods have been proposed for manual error analysis. This work proposes a method that uses a machine learning framework to identify errors in MT output, and improves efficiency of manual error analysis. Our method builds models that classify low and high quality translations, then identifies features of low quality translations to improve efficiency of the manual analysis. Experiments showed that by using our methods, we could improve the efficiency of MT error analysis.

Key Words: *statistical machine translation, error analysis, error detection, evaluation metric, corpus*

1 はじめに

最新の機械翻訳システムは、年々精度が向上している反面、システムの内部は複雑化しており、翻訳システムの傾向は必ずしも事前に把握できるわけではない。このため、システムによっ

[†] 奈良先端科学技術大学院大学, Nara Institute of Science and Technology

てある文章が翻訳された結果に目を通すことで、そのシステムに含まれる問題点を間接的に把握し、システム同士を比較することが広く行われている。このように、単一システムによって発生する誤りの分析や、各システムを比較することは、各システムの利点や欠点を客観的に把握し、システム改善の手段を検討することに役立つ。ところが、翻訳システムの出力結果を分析しようとした際、機械翻訳の専門家である分析者は、システムが出力した膨大な結果に目を通す必要があり、その作業は労力がかかるものである。

この問題を解決するために、機械翻訳の誤り分析を効率化する手法が提案されている (Popović and Ney 2011; Kirchhoff, Rambow, Habash, and Diab 2007; Fishel, Bojar, Zeman, and Berka 2011; El Kholy and Habash 2011)。この手法の具体的な手続きとして、機械翻訳結果を人手により翻訳された参照訳と比較し、機械翻訳結果のどの箇所がどのように誤っているかを自動的にラベル付けする。さらに、発見した誤りを既存の誤り体系 (Flanagan 1994; Vilar, Xu, d'Haro, and Ney 2006) に従って「挿入・削除・置換・活用・並べ替え」のような分類することで、機械翻訳システムの誤り傾向を自動的に捉えることができる。

しかし、このような自動分析で誤りのおおよその傾向をつかめたとしても、機械翻訳システムを改善する上で、詳細な翻訳誤り現象を把握するためには、人手による誤り分析が欠かせない。ところが、先行研究と同じように、参照文と機械翻訳結果を比較して差分に基づいて誤りを集計する手法で詳細な誤り分析を行おうとした際に、問題が発生する。具体的には、機械翻訳結果と参照訳の文字列の不一致箇所を単純な方法でラベル付けすると、人間の評価と一致しなくなる場合がある。つまり、機械翻訳結果が参照訳と同様の意味でありながら表層的な文字列が異なる換言の場合、先行研究では不一致箇所を誤り箇所として捉えてしまう。このような誤った判断は、誤り分析を効率化する上で支障となる。

本研究では、前述の問題点を克服し、機械翻訳システムの誤りと判断されたものの内、より誤りの可能性が高い箇所を優先的に捉える手法を提案する。図 1 に本研究の概略を示す。まず、対訳コーパスに対して翻訳結果を生成し、翻訳結果と参照訳を利用して誤り分析を優先的に行うべき箇所を選択する。次に、重点的に選択された箇所を中心に人手により分析を行う。誤りの可能性が高い箇所を特定するために、機械翻訳結果に含まれる n-gram を、誤りの可能性の高い順にスコア付けする手法を提案する（3 節)。また、誤りかどうかの判断を単純な一致不一致より頑健にするために、与えられた機械翻訳結果と正解訳のリストから、機械翻訳文中の各 n-gram に対して誤りらしさと関係のあるスコア関数を設計する。設計されたスコア関数を用いることで、誤り n-gram を誤りらしさに基づいて並べ替えることができ、より誤りらしい箇所を重点的に分析することが可能となる。単純にスコアに基づいて選択を行った場合、正解訳と一致するような明らかに正しいと考えられる箇所を選択してしまう恐れがある。この問題に対処するため、正解訳を利用して誤りとして提示された箇所をフィルタリングする手法を提案し、選択精度の向上を図る（4 節)。

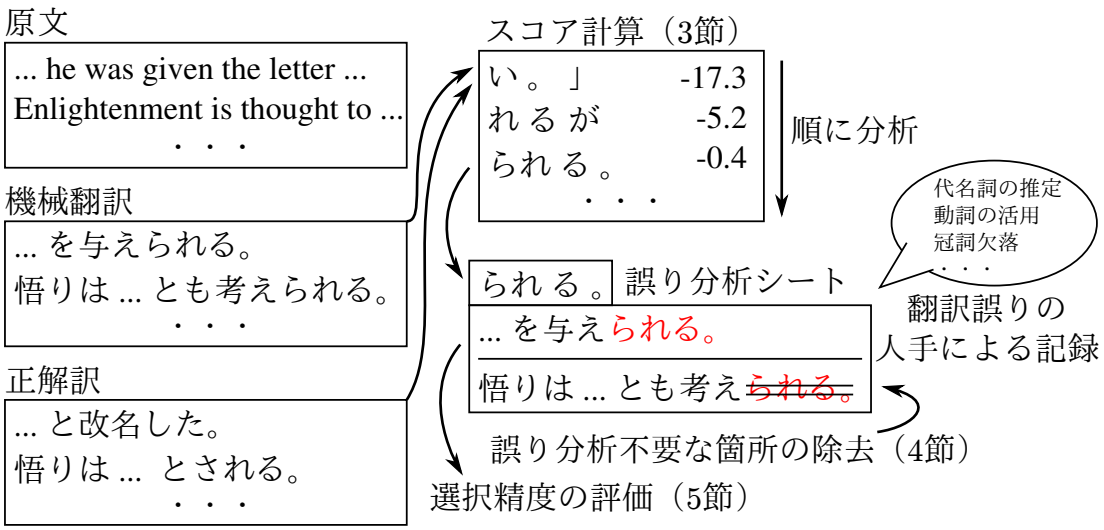


図 1: 本研究の流れ図

実験では、まず 5.1 節〜5.2 節で提案法の誤り箇所選択精度の測定を行い、単一システムの分析、及びシステム間比較における有効性の検証を行う。実験の後半では、提案法の課題を分析し（5.3 節）、提案法を機械翻訳システムの改善に使用した場合の効果について検討を行う（5.4 節）。

2 機械翻訳の自動評価と問題点

本節では、従来から広く行われている機械翻訳の自動評価について説明し、その問題点を明らかにする。

2.1 評価の手順

機械翻訳システムに「原文 $\boldsymbol{f}$」を与えることで、「機械翻訳結果 $\boldsymbol{e}$」が得られたとする。評価方法として「自動評価尺度」を用いる場合、事前に人手により翻訳された「参照訳 $\boldsymbol{r}$」を与える。自動評価尺度は、機械翻訳結果 $\boldsymbol{e}$ と参照訳 $\boldsymbol{r}$ の差異に基づき機械翻訳結果の良し悪しをスコアとして計算するものである (Papineni, Roukos, Ward, and Zhu 2002; Doddington 2002; Banerjee and Lavie 2005)。また、「品質推定」と呼ばれる技術は、参照訳を利用せずに評価を行う。具体的には、誤りのパターンを学習したモデルによって機械翻訳文の精度を推定することや (Specia, Turchi, Cancedda, Dymetman, and Cristianini 2009)、翻訳結果の精度を部分的に評価することが行われている (Bach, Huang, and Al-Onaizan 2011)。

自動評価尺度を利用する場合は参照訳を用意する必要があるが、翻訳精度の計算を翻訳シス

表 1: 機械翻訳の誤訳の例。文脈から “right” は「正しい」と訳すべきだが「右」と訳されている。また “choose” に相当する語句が機械翻訳結果では削除されている。

原文 $\boldsymbol{f}$	For this reason, it is considered crucial to choose the right zen master.
機械翻訳結果 $\boldsymbol{e}$	この ため 、 右 禅師 が 重要 で あ る と さ れ る 。
参照訳 $\boldsymbol{r}$	それ ゆえ に 正し い 禅師 を 選 ぶ こと が 肝心 と さ れ る 。
BLEU+1 スコア	0.234199

テムに依らず一貫して行える利点がある。一方、品質推定は参照訳を必要としない分、翻訳精度を正しく推定することが比較的困難である。本研究は、参照訳が与えられた状況で機械翻訳の誤り分析を行う場合を対象とする。

2.2 代表的な自動評価尺度

機械翻訳の自動評価尺度は、様々なものが提案されており、尺度ごとに異なった特徴がある。BLEU (Papineni et al. 2002) は機械翻訳の文章単位の自動評価尺度として最も一般的に使われるものであり、参照訳 $\boldsymbol{r}$ と機械翻訳結果 $\boldsymbol{e}$ の間の n-gram の一致率に基づきスコアを計算する。機械翻訳結果と参照訳が完全に一致すれば 1 となり、異なりが多くなるに連れて 0 に近くなる。BLEU を文単位の評価に対応させたものに BLEU+1 (Lin and Och 2004) がある。BLEU や BLEU+1 は、$\boldsymbol{e}$ と $\boldsymbol{r}$ の表層的な文字列の違いにしか着目しないため、$\boldsymbol{e}$ が $\boldsymbol{r}$ の換言である場合にスコアが不当に低くなる場合がある [1]。

BLEU とは異なり、評価尺度自体が換言に対応したものに、METEOR (Banerjee and Lavie 2005) がある。METEOR を利用する場合、単語やフレーズの換言を格納したデータベースを事前に用意しておく。これにより、参照訳と機械翻訳結果の n-gram が一致しない場合であっても、データベース中に含まれる換言を利用することで一致する場合、スコアの低下を小さくすることが可能となる。

2.3 自動評価の課題

表 1 に、英日翻訳における原文、機械翻訳結果、参照訳の例を示す。機械翻訳システムとして、句構造に基づく機械翻訳システムを利用した。自動評価尺度の一例として、BLEU+1 スコアを示す。また、図 2 はシステムが翻訳結果を出力した際の導出過程の一部である。自動評価尺度を用いることで、翻訳システムの性能を数値で客観的に比較することが可能であるが、この例から自動評価尺度に頼り切ることの危険性も分かる。前節で述べたように、自動評価尺度は

[1] BLEU や BLEU+1 を用いる場合、複数の異なった言い回しの参照訳を与えることで、複数の言い回しに対応した評価が行える。

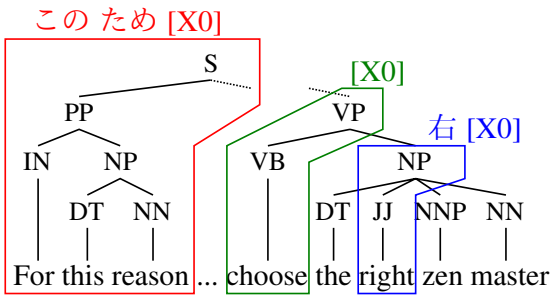


図 2: 機械翻訳システムの導出過程の例。この文脈で形容詞（JJ）“right” を「右」と訳すのは誤りである。また動詞（VB）“choose” を削除する規則をここで使用することも誤りである。

人間の評価と必ずしも一致しない評価を行う場合がある。表 1 の例では、“For this reason” が機械翻訳で「このため」と正しく翻訳されているが、参照訳では「それゆえ」と翻訳されているため、文字列の表層的な違いにしか着目しない BLEU+1 では誤訳と判断されて、スコアが不当に低くなる。METEOR を用いた場合、換言によるスコアの低下は発生しにくくなるが、逆に誤った換言が使用され、スコアが不当に高くなることも考えられる。

自動評価尺度は、機械翻訳結果の正確さを判断する上で有用であるが、その結果からシステムの特徴を把握することは困難である。しかし、このように翻訳結果に目を通すことで、自動評価尺度の数値だけからは分からない情報を把握することが可能となる。

3 スコアに基づく誤り候補 n-gram の順位付け

機械翻訳の誤り箇所を自動的に提示する際に、単純に誤り箇所を列挙するのではなく、より誤りの可能性が高い箇所から順に示すことができれば、後の人手による誤り分析の効率が上がると考えられる。本節では、機械翻訳結果に含まれる n-gram に対して、分析の優先度に対応するスコアを与える手法を提案する。分析者はこのスコアを参考にし、最初に分析する箇所を決定する。図 3 に n-gram のスコアに基づく誤り分析の例を示す。この例では、提案法によって**い。」** が最も優先的に分析すべき n-gram と判断されているため、最初に機械翻訳結果全体からこの n-gram が含まれる箇所を見つけ、誤り分析をする。分析が終了したら、次に優先度の高い**れるが** を、最後に **られる。** を分析する。

ある n-gram を提示した際に、もし分析者が機械翻訳の誤りでない箇所を分析対象としてしまうと、余計な分析作業を行うこととなり、分析効率が低下する原因となる。このため、効率的な誤り分析が行われるためには、最初に提示される n-gram ほどシステムの特徴的な誤りを捉えていることが望ましい。

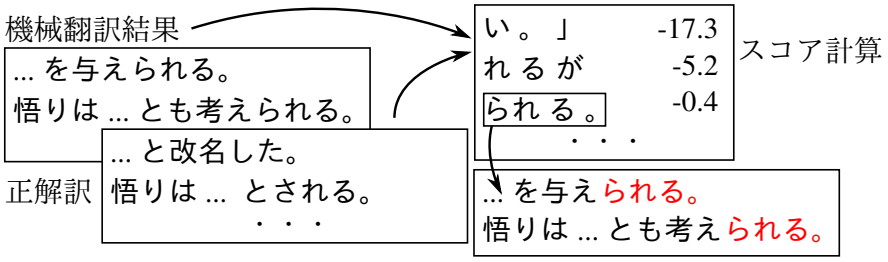


図 3: n-gram のスコアに基づく誤り分析

3.1 正解訳を用いた n-gram のスコア計算法

本節では、このようなスコア付けを行う手法を 5 つ説明する。そのうち、2 つ（ランダム選択と誤り頻度に基づく選択）はベースラインであり、3 つ（自己相互情報量に基づく選択、平滑化された条件付き確率に基づく選択、識別言語モデルの重みに基づく選択）は提案法である。

まず、すべての手法に共通する以下の関数を定義する。

$\boldsymbol{\phi}(\boldsymbol{e})$: 文 $\boldsymbol{e}$ に対する素性ベクトル。各要素は文 $\boldsymbol{e}$ に含まれる n-gram の出現頻度。

$\boldsymbol{e}_{MT}(n)$: コーパス中の n 番目の文に対応する機械翻訳結果。

$\boldsymbol{e}_{C}(n)$: コーパス中の n 番目の文に対応する正解訳（3.2 節で定義）。

これらの関数を利用し、次節以降で述べるスコア関数に従って、コーパス全体から n-gram のスコアを計算する。

3.1.1 ランダム選択

ランダム選択は、n-gram の順位付けを一切行わずに誤り分析を進めることに等しい。4 章で誤り候補のフィルタリングの説明を行うが、フィルタリングを一切行わない場合はチャンスレートとなる。また、ランダム選択と参照訳を用いた厳密一致フィルタリングを組み合わせた場合は、先行研究 (Popović and Ney 2011) で提案されている参照訳との差分に基づく分析となる。

3.1.2 誤り頻度に基づく選択

誤り頻度に基づく手法は、機械翻訳結果に多く含まれ、正解訳に含まれない回数が多い n-gram は重点的に分析すべきという考え方に基づく。n-gram のスコア計算では、ある n-gram x が機械翻訳結果 $\boldsymbol{e}_{MT}(n)$ に含まれていて、かつ正解訳 $\boldsymbol{e}_{C}(n)$ に含まれていない回数を計算し、スコアとする。

$$S(x) = -\sum_{n} \left[\phi_x(\boldsymbol{e}_{MT}(n)) - \phi_x(\boldsymbol{e}_{C}(n))\right]_+ \tag{1}$$

表 2: 機械翻訳で頻繁に起こる誤り

1-gram		2-gram		3-gram	
の	75	た 。	35	た 。 (文末)	35
に	60	る 。	27	る 。 (文末)	26
、	56	大学 (文末)	26	あ る 。	22
た	56	あ る	26	で あ る	18
が	54	年 （	26	さ れ た	16

ここで $[\ ]_+$ はヒンジ関数である。このスコアが低い n-gram から順に選択することで、誤って出現した回数が多い n-gram を優先的に分析することとなる。

しかし、頻繁に発生する誤りが必ずしも分かりやすく有用な誤りとは限らない。表 2 は、ある英日機械翻訳システムが出力した翻訳結果に含まれ、参照訳には含まれなかった n-gram を、回数が多いものから順に一覧にしたものである。この表で、右側の数字はテストコーパス内で誤って出現した回数を示している。この表を見ると、単純に頻繁に検出される誤りは目的言語に頻繁に出現するものに支配されており、この結果だけからは翻訳システムの特徴を把握しにくいことが分かる。

3.1.3 自己相互情報量に基づく選択

誤り頻度に基づいて n-gram の選択を行った場合、表 2 に示したように誤りとして検出されるものの多くは単純に目的言語の特徴を捉えたものとなってしまう。本研究ではこの問題に対処するため、出現頻度より正しく、ある n-gram が誤った文の特徴であるかどうかを判断する手法を提案する。

最初のスコア付け基準として、自己相互情報量（PMI: Pointwise Mutual Information）に基づく手法を提案する。PMI は、2 つの事象の関係性を計る尺度であり、本研究では与えられた n-gram と機械翻訳結果との関係性をスコアとして定式化する。機械翻訳結果と関係が強い n-gram は、正解訳との関係は逆に弱くなる。PMI は以下の式によって計算される (Church and Hank 1990)。

$$\begin{aligned} PMI(x, \boldsymbol{e}_{MT}) &= \log \frac{p(\boldsymbol{e}_{MT}, x)}{p(\boldsymbol{e}_{MT}) \cdot p(x)} \\ &= \log \frac{p(\boldsymbol{e}_{MT}|x)}{p(\boldsymbol{e}_{MT})} \end{aligned}$$

ここで、各原文につき機械翻訳結果と正解訳が 1 つずつ与えられるため、$p(\boldsymbol{e}_{MT}) = 1/2$ である。条件付き確率 $p(\boldsymbol{e}_{MT}|x)$ は以下の式で計算される。

$$p(\boldsymbol{e}_{MT}|x) = \frac{\sum_n \phi_x(\boldsymbol{e}_{MT}(n))}{\sum_n \{\phi_x(\boldsymbol{e}_{MT}(n)) + \phi_x(\boldsymbol{e}_C(n))\}}$$

最終的に、自己相互情報量の期待値に比例する以下の値をスコアとし、スコアが低いものから順に n-gram を選択する。

$$\begin{aligned} S(x) &= \phi_x(\boldsymbol{e}_{MT}(n)) \cdot (-PMI(x, \boldsymbol{e}_{MT})) \\ &\propto p(x, \boldsymbol{e}_{MT}) \cdot (-PMI(x, \boldsymbol{e}_{MT})) \end{aligned} \tag{2}$$

3.1.4 平滑化された条件付き確率に基づく選択

「誤り頻度に基づく選択」では、目的言語に頻繁に出現する n-gram が分析対象の上位を占めてしまう問題があった。そこで、2 つ目のスコア付け基準は、誤り頻度を全体の出現回数で正規化し、条件付き確率として定式化することを考える。平滑化された条件付き確率に基づく選択では、ある n-gram がシステム出力に含まれながら参照文に含まれない確率をスコアとし、このスコアが高いものを優先的に分析する。まず、以下の関数を定義する。

$$\begin{aligned} F_{MT}(x) &= \sum_n [\phi_x(\boldsymbol{e}_{MT}(n)) - \phi_x(\boldsymbol{e}_C(n))]_+ \\ F_C(x) &= \sum_n [\phi_x(\boldsymbol{e}_C(n)) - \phi_x(\boldsymbol{e}_{MT}(n))]_+ \end{aligned}$$

ここで、$F_{MT}(x)$ は誤り頻度に基づく選択で利用した式 (1) に等しい。また、$F_C(x)$ は n-gram が正解訳により多く出現した回数を表す。ある n-gram を選択した際、その n-gram が正解訳に多く含まれる条件付き確率は以下の通りである。

$$p(\boldsymbol{e}_{MT}|x) = \frac{F_{MT}(x)}{F_{MT}(x) + F_C(x)} \tag{3}$$

しかし、確率を最尤推定で計算すると、正解訳として出現せず、機械翻訳結果に 1 回しか出現しないような稀な n-gram の確率が 1 となり、頻繁に選択されてしまう。上述の問題点を解決するために、確率の平滑化を行う。文献 (Mackay and Petoy 1995) では平滑化の手法としてディリクレ分布を事前分布として確率を推定しており、本手法もこれに習う。平滑化を用いた際の n-gram x についての評価関数は式 (4) の通りであり、$S(x)$ が低いものを代表的な n-gram とする。

$$S(x) = -\frac{F_{MT}(x) + \alpha P_{MT}}{F_{MT}(x) + F_C(x) + \alpha} \tag{4}$$

ただし、

$$P_{MT} = \frac{\sum_x F_{MT}(x)}{\sum_x F_{MT}(x) + \sum_x F_C(x)}$$

このとき平滑化係数 α を決定する必要がある。n-gram を利用して参照文もしくはシステム出力文を選択する際、選択される文の種類がディリクレ過程に従うと仮定すると、コーパス全体に対する尤度は式 (5) で表される。

$$P = \prod_{x} \frac{\{\prod_{k=0}^{F_{MT}(x)-1}(k+\alpha P_{MT})\}\{\prod_{k=0}^{F_C(x)-1}(k+\alpha P_C)\}}{\prod_{k=0}^{F_{MT}(x)+F_C(x)}(k+\alpha)} \tag{5}$$

式 (5) の P が最大化されるような α をパラメーターとする。P は全区間で微分可能であり、唯一の極があるとき、その点で最大値となる。よって α は P の微分からニュートン法により計算できる。

3.1.5 識別言語モデルの重みに基づく選択

最後に、識別言語モデルの重みに基づくスコア付け基準を提案する。識別言語モデルは、自然な出力言語文の特徴を捉えるように学習される通常の言語モデルとは異なり、ある特定のシステムについて、起こりやすい出力誤りを修正するように学習される。さらに学習時に正則化を行えば、モデルのサイズが小さくなり、少ない修正で出力を改善するような効率的な修正パターンが学習される。誤り分析の観点から見ると、モデルによって学習された効率的な修正パターンに目を通せば、システムの特徴的な誤りを発見できると考えられる。

○ 構造化パーセプトロンによる識別言語モデル

識別言語モデルの学習は構造学習の一種である。先行研究では、構造学習の最も単純な手法である構造化パーセプトロン (Collins 2002) を、識別言語モデルの学習において有用な手法であると示している (Roark, Saraclar, and Collins 2007)。構造化パーセプトロンでは、候補集合の中で誤りの修正先として学習される目標 E^* を定める。本研究では目標として、機械翻訳結果の n-best の中で評価尺度が最も高かった文（オラクル訳、3.2 節参照）を選択する。学習では、モデルによって最も大きなスコアが与えられる現在の仮説 $\hat{E}$ と E^* の素性列を比較する。1 回の更新において、$\hat{E}$ と E^* の差分を用いて重み $\boldsymbol{w}$ を更新する。重みが更新されると、重みと素性列から計算されるスコアが変化し、仮説 $\hat{E}$ が更新される。$\hat{E}$ と E^* が等しいときは差分が $\mathbf{0}$ のため更新を行わない。重みの更新はコーパス全体に対して一文ごとに逐次的に行い、反復回数や重みの収束といった終了条件が満たされるまで反復する。学習のアルゴリズムを Algorithm 1 に示す。 ここで、$\hat{\boldsymbol{E}}(n)$ は n 番目の文に対応する機械翻訳結果の n-best リスト、T は反復回数である。また、$EV(E)$ は機械翻訳結果 E の翻訳精度を評価するための自動評価尺度である。

○ L1 正則化による素性選択

機械翻訳システムの誤り傾向をより明確にするため、重みの学習時に L1 正則化を行う。L1 正則化は、重みベクトルに対して L1 ノルム $\|\boldsymbol{w}\|_1 = \sum_i |w_i|$ に比例するペナルティを与える。L1 正則化を用いる時に、重み $\boldsymbol{w}$ の中で多くの素性に対応するものが 0 となるため、識別能力に

Algorithm 1 構造化パーセプトロンによる識別言語モデルの学習

for $t = 1$ **to** T **do**
  for $n = 1$ **to** N **do**
    $E^* \leftarrow \arg\max_{E \in \hat{\boldsymbol{E}}(n)} EV(E)$
    $\hat{E} \leftarrow \arg\max_{E \in \hat{\boldsymbol{E}}(n)} \boldsymbol{w} \cdot \boldsymbol{\phi}(E)$
    $\boldsymbol{w} \leftarrow \boldsymbol{w} + \boldsymbol{\phi}(E^*) - \boldsymbol{\phi}(\hat{E})$
  end for
end for

表 3: オラクル訳の例

原文	日本の臨済宗は、日本の禅の宗派のひとつである。	BLEU+1
参照訳	Japan's rinzai is one of zen schools in Japan.	
機械翻訳結果	The rinzai sect in Japan, and is one of the zen sects in Japan.	0.223856
オラクル訳	Japanese rinzai sect is one of the zen sects in Japan.	0.295023

大きな影響を与えない素性をモデルから削除することが可能となる。

L1 正則化された識別モデルを学習する簡単かつ効率的な方法として、前向き後ろ向き分割（forward-backward splitting; FOBOS）アルゴリズムがある (Duchi and Singer 2009)。一般的なパーセプトロンでは正則化を重みの更新時に行うが、FOBOS では重みの更新と正則化の処理を分割し、重みの利用時に前回からの正則化分をまとめて計算し、効率化を図る。

○ 識別言語モデルの素性

識別言語モデルの素性として様々な情報を利用できるが、本研究では n-gram に基づく選択を行うため、以下の 3 種類の素性を利用する。

翻訳仮説を生成したシステムのスコア: システム出力を修正するように学習するため、学習の初期においてシステムスコアによる順位付けが必要である。

翻訳仮説に含まれる n-gram の頻度: n-gram に対して重み付けをすることで、システムが出力する誤った n-gram を捉える。

翻訳仮説の単語数: 翻訳システムが利用する評価尺度が単語数によって大きく影響される場合、単語数を調整するのに用いられる。

n-gram の選択時には、識別言語モデルによって学習された重みが低いものを優先的に選択する。

表 4: 厳密一致フィルタリングの例

	n-gram	、 右
1	機械翻訳	この ため 、 右 禅師 が 重要 で あ る と さ れ る 。
	正解訳	それ ゆえ に 正し い 禅師 を 選ぶ こと が 肝心 と さ れ る 。
2	機械翻訳	その 時 、 右 の 標識 に 注意 す る 。
	正解訳	その 際 、 右 の 標識 に 注意 。

3.2 スコア計算に用いる正解訳 $e_C(n)$ の選択

機械翻訳の評価では、正解訳として事前に人手で翻訳された参照訳を利用することが多い。しかし参照訳は機械翻訳とは独立に翻訳されるため、使用する語彙が機械翻訳結果とは異なる場合が多いと考えられる。本研究では参照訳の代わりに、機械翻訳システムが出力した翻訳候補の中で、自動評価尺度により最も高いスコアが与えられた文（オラクル訳）を正解訳として利用し、参照訳を用いた場合との比較を行う。

表 3 にオラクル訳の例を示す。この例では、日本語の「宗派」が機械翻訳結果で “sect” と正しく翻訳されているが、参照訳では “school” となっているため、差分を取ると誤り n-gram として選択されやすくなってしまう。オラクル訳は機械翻訳システムの探索空間の中で、参照訳の表現に最も近い文であるため、訳出に近い表現を維持しながら正しい翻訳に近づく。この場合、誤りでない “sect” がオラクル訳でも使用されているため、差分をとった際に誤り n-gram として扱われにくくなる。このように、オラクル訳は翻訳仮説と同じシステムから出力されるため、オラクル訳は参照訳に比べて翻訳仮説との表層的な異なりが少なくなり、換言を誤り n-gram として誤選択する可能性が低くなると考えられる。一方、オラクル訳は機械翻訳システムから出力されている以上、誤訳を含む場合もあることに注意されたい。

4 誤り候補 n-gram のフィルタリング

n-gram に基づく誤り箇所選択では、n-gram のスコアはコーパス全体から計算される。このため、コーパス全体を見た際に分析すべきと判断された n-gram であっても、ある特定の文では誤りとは考えにくい場合がある。本節では、選択された箇所に対してフィルタリングを適用することにより誤選択を回避し、機械翻訳の誤り箇所選択率の向上を行う。

4.1 厳密一致フィルタリング

このフィルタリングは、機械翻訳結果中のある n-gram が誤り箇所として選択された際に、その n-gram が正解訳の一部に厳密一致するかどうかを確認し、一致する場合は選択を行わない

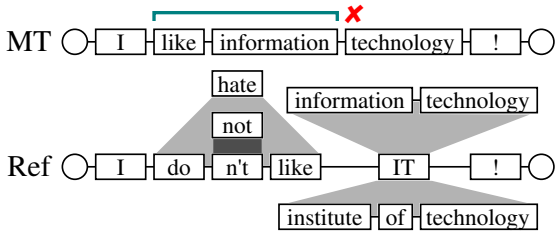


図 4: パラフレーズラティスによる誤り箇所候補のフィルタリング

ようにする。フィルタリングの具体例を表 4 に示す。n-gram「、 右」が誤り箇所の候補とされた際、1 つ目の例では機械翻訳結果の一致箇所が選択されるが、2 つ目の例では正解訳に同一の n-gram があるため、誤り箇所の候補から除外される。これは、正解訳に含まれている文字列は翻訳誤りではないだろうという直感に基づく。

4.2 換言によるフィルタリング

機械翻訳結果と正解訳の文字列が、表層的に異なりながら意味が等しい場合、厳密一致フィルタリングを用いただけでは選択された箇所が正解訳に含まれず、誤選択を回避することができない。この問題を解決するため、本研究では正解訳の換言を用いたフィルタリングを行う。

換言によるフィルタリングの例を図 4 に示す。正解訳として “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” が誤りの候補として挙げられたとする。換言によるフィルタリングでは、まず正解訳に含まれる全ての部分単語列を用意した換言データベースの中から検索し、ある閾値以上の確率で置換可能な換言を抽出する。次に、抽出された換言を利用して参照訳のパラフレーズラティス (Onishi, Utiyama, and Sumita 2010) 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。

4.3 フィルタリングに用いる正解訳の選択

3.2 節で、スコア計算に用いる正解訳として参照訳またはオラクル訳を利用するが、フィルタリングの際にも正解訳として参照訳のみ用いた場合と、参照訳に加えてオラクル訳を用いた場合で比較を行う。オラクル訳の選択では、機械翻訳の自動評価尺度を用いるが、本研究では以下の 2 つの評価尺度で選択を行った場合の比較を行う。

BLEU+1: 機械翻訳の自動評価に一般的に用いられる尺度である BLEU を、文単位の評価に対応させたもの。換言を考慮しない。

METEOR: BLEU+1 は、参照訳と機械翻訳結果の表層的な文字列の違いにしか着目しないた

め、換言に対して不当な罰則を行ってしまう。METEOR は事前に与えられた換言テーブルを用いるため、換言に対する罰則が BLEU+1 に比べて小さくなる。METEOR を用いた場合、BLEU+1 を用いた場合に比べ、オラクル訳と参照訳の違いは表層的に多くなると考えられるが、逆に機械翻訳結果との表層的な違いが少なくなり、換言の誤選択が発生しにくくなると考えられる。

5 実験

本節では、各実験を通して、提案法を利用することで機械翻訳の誤り分析をより効率的に行えることを示す。まず、各スコア基準に従って単一の機械翻訳システム（5.1.2 節）及び複数の機械翻訳システム（5.1.4 節）の誤り箇所選択を行い、人手評価を行う。これにより、提案法の選択精度とシステム間比較における有効性を検証する。次に、誤りとして選択された箇所のフィルタリングを複数の手法によって行い、フィルタリングの効果を自動評価によって測定する（5.2 節）。さらに、提案法が翻訳誤りでない箇所を誤選択する場合についても分析を行い、提案法が抱える課題を明らかにし、その改善策について検討する（5.3 節）。また、提案法によって発見された翻訳誤りを修正した際の効果について検討する（5.4 節）。

5.1 選択された誤り箇所の調査

本節では、各手法によって順位付けされた誤り n-gram を人手で分析する。人手評価の方法は赤部, Neubig, Sakti, 戸田, 中村 (2014a) に従い 2 段階で行う。まず、各誤り箇所選択手法によって選択された箇所に対し、分析者はその箇所が機械翻訳の誤り箇所を捉えているかどうかをアノテーションする。これにより、優先的に選択された上位 k 個の n-gram について、誤り箇所の適合率を測定することが可能となる。次に、誤り箇所を捉えている場合は、以下に示す誤りの種類をアノテーションする。

文脈依存置換誤り： 別の文脈では正しい翻訳だが、この文脈では不適切な翻訳。

文脈非依存置換誤り： いかなる文脈であっても、不適切な翻訳。

挿入誤り： 不必要な語句の挿入。

削除誤り： 必要な語句の不適切な削除。

並べ換え誤り： 選択された箇所が語順の誤りを捉えている。

活用誤り： 活用形が誤っている。

これにより、選択された誤り箇所の誤り傾向を把握する。これらの結果を元に、翻訳システムの比較を行う。

表 5: KFTT のデータサイズ

	文数	単語数	
		英語	日本語
学習セット	330k	5.91M	6.09M
開発セット	1166	24.3k	26.8k
テストセット	1160	26.7k	28.5k

表 6: 実験に用いた誤り箇所選択手法

		学習データ		
		翻訳結果+参照訳	翻訳結果+オラクル訳	翻訳結果 n-best+オラクル訳
スコア計算法	誤り頻度	○	○	
	自己相互情報量	○	○	
	条件付き確率	○	○	
	識別言語モデル			○
	ランダム選択			

5.1.1 実験設定

すべての実験で京都フリー翻訳タスク（KFTT）(Neubig 2011) の日英翻訳を利用した。コーパスの大きさを表 5 に示す。単一の機械翻訳システムを用いた実験では、Travatar ツールキット (Neubig 2013) に基づく forest-to-string（F2S）システムを利用した。システム間比較では、F2S システムに加え、Moses ツールキット (Koehn, Hoang, Birch, Callison-Burch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Bojar, Constantin, and Herbst 2007) に基づくフレーズベース翻訳（PBMT）システム及び階層的フレーズベース（HIERO）システムを利用した。

翻訳システムを構築する上で、F2S システムでは単語間アラインメントに Nile[2] を利用し、構文木の生成には Egret[3] を利用した。PBMT システムと HIERO システムでは、単語間アラインメントに GIZA++ (Och and Ney 2003) を利用した。チューニングには MERT (Och 2003) を利用し、評価尺度を BLEU (Papineni et al. 2002) とした。

n-gram の選択には 3 章で説明したスコア計算法を利用した。実験を行ったスコア計算法とスコアの学習に利用したデータの組み合わせを表 6 に示す。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。学習時に FOBOS (Duchi and Singer 2009) による L1 正則化を行った。正則化係数は 10^{-7}-10^{-2} の中から選び、KFTT のテストセットに対して高い精度を示す値を利用した。学習には 1-gram から 3-gram までの n-gram を長さによる区別を行わずに利

[2] http://code.google.com/p/nile/
[3] http://code.google.com/p/egret-parser/

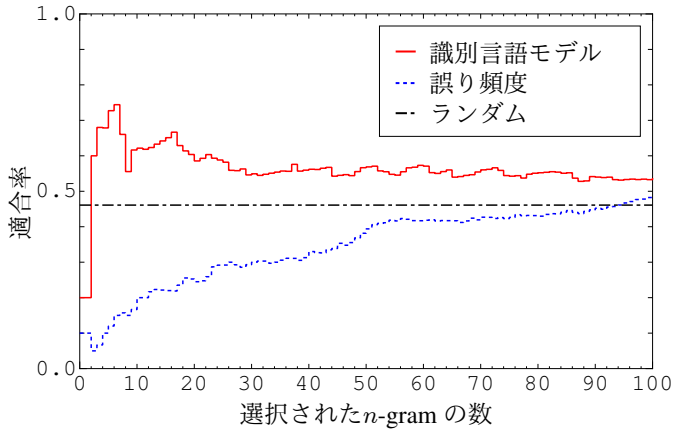


図 5: 分析対象となった n-gram の誤り箇所選択率。横軸は選択した n-gram の種類数、縦軸は誤り箇所適合率。

用した。

オラクル文の選択には BLEU+1 を利用し、選択される n-gram の誤り傾向を分析した。各手法で、参照訳を用いた厳密一致フィルタリングを行った。

5.1.2 選択する n-gram の個数と適合率の関係

まず、識別言語モデルの重みに基づく選択、機械翻訳結果と参照訳から計算された誤り頻度に基づく選択、ランダム選択の比較を行った。図 5 に、上位の n-gram を順に選んだ際の誤り箇所適合率を示す。この結果から、識別言語モデルの重みに基づき選択された n-gram が、ランダム選択、誤り頻度に基づく選択の 2 手法に比べ、機械翻訳の誤り箇所を高い精度で捉えていることが分かる。

5.1.3 選択された n-gram の統計

次に、各手法で上位 30 個に選ばれた誤り n-gram を選択した際の誤り箇所適合率を調査した。結果を表 7 に示す。

表から、誤り箇所選択率が高い手法は、平滑化された条件付き確率に基づく手法と、識別言語モデルの重みに基づく手法であることが分かる。その他の手法はランダム選択を下回っており、誤り箇所選択率が高いとは言えない。上位 3 つの手法を比較すると、参照文を利用した条件付き確率に基づく手法では、置換誤りや挿入誤りを多く捉えているが、削除誤りをほとんど捉えていないことが分かる。一方、識別言語モデルの重みに基づく手法では、他の手法に比べて 2 倍以上の削除誤りを捉えていることが分かる。

表 7: 各手法により選択された n-gram の内訳。誤り箇所選択率がランダムよりも高い 3 つの手法を太字で示した。また、3 つの手法の中で各種類の誤りについて多く検出されたものを太字とした。

種類	ランダム	誤り頻度		自己相互情報量		条件付き確率		識別言語モデル
学習データ		ORA	REF	ORA	REF	**ORA**	**REF**	**n-best**
誤り箇所選択率	0.483	0.290	0.323	0.427	0.410	**0.607**	**0.713**	**0.598**
文脈依存置換	0.124	0.460	0.258	0.500	0.520	0.218	**0.449**	0.332
文脈非依存置換	0.166	0.023	0.052	0.086	0.049	0.016	**0.140**	0.067
挿入	0.111	0.195	0.278	0.078	0.260	**0.537**	0.364	0.164
削除	0.062	0.103	0.093	0.016	0.041	0.042	0.023	**0.245**
並べ替え	0.319	0.218	0.299	0.305	0.106	0.176	0.023	**0.192**
活用	0.093	0	0.021	0.016	0.024	0.011	0	0

識別言語モデルの重みに基づく手法以外で削除誤りの検出率が悪い原因として、削除誤りを検出する際には削除された単語列ではなく、その前後の文脈を見る必要があることが挙げられる。削除誤りが発生する前後の文脈は原文によって大きく変わるため、n-gram の発生頻度が小さく候補から外れやすくなる。しかし識別言語モデルの重みに基づく手法の場合、同じ文脈における削除誤りが n-best 中の複数の候補に発生するため、削除誤りが修正されるまで n-gram の重みを大きくしようとする。結果として、識別言語モデルの重みに基づく手法では、削除誤りも多く捉えることができる。

各手法とも、ランダム選択に対して捉えられた誤りの分布は大きく異なる。この点から、別々の手法によって捉えられた誤りの分布を比較することはできないことが分かる。また、あるシステムの分析結果に対して、どの誤りが多い、あるいは少ないという絶対的な評価はできず、システム同士の相対的な評価にしか利用できないことに注意されたい。

識別言語モデルの重みによって選択された箇所の例を表 8 に示す。この結果を見ると、誤り頻度に基づいて選択した場合（表 2）に比べ、選択された n-gram が目的言語の言語現象に支配されていないことが分かる。

5.1.4 システム間比較

分析対象とするシステムによって、含まれる誤りの分布が異なる。本節では、提案法によって検出される誤りが、本来の誤り分布を適切に捉えることを確認する。具体的には、PBMT、HIERO、F2S の 3 つの翻訳システムで日英・英日の両方向に対して翻訳を行い、単一システムの評価を行った際と同様に、識別言語モデルの重みに基づく誤り箇所選択法を利用して抽出された上位 30 個の誤り n-gram に対し、分析を行った。その結果を表 9 に示す。

この結果から、PBMT と HIERO の両システムでは、並べ換え誤りが上位の誤りとして検出されている一方、F2S システムの特に英日翻訳では下位の誤りとして検出された。一般的に、統

表 8: 識別言語モデルによって選択された上位の n-gram。枠で囲まれた部分は選択された箇所および選択箇所に対応する箇所を示す。

n-gram	重み	例文
) of	-7.50950	Src ··· 、 後醍醐 天皇 [の] 諱 ・ 尊治 （ たかはる [）] の 御 一 字 を 賜 り 、 ···
		Ref ... by emperor go-daigo , and he was awarded the letter (尊) , which came from the emperor ['s] real name takaharu (尊治 [ )] , so ...
		MT ... , imina (たかはる [ ) of] emperor godaigo , and ...
		Eval 並べ換え誤り
senior	-6.52024	Src ··· 、 [中高] など の 学校 教員 から ···
		Ref ... consists of teachers of [junior high , high] , and ...
		MT ... , and is a organization consisting of teachers such as [senior] .
		Eval 文脈非依存置換誤り
the ko clan	-6.52021	Src この 時 、 [高 氏] の 側室 の 子 ・ 竹若丸 が ···
		Ref in this fighting , takewakamaru , the son of [takauji] 's concubine , was ...
		MT on this occasion , ... , the son of a concubine of [the ko clan] .
		Eval 文脈依存置換誤り

語情報を使った翻訳システムは並べ換え誤りに強いことが知られており、本結果はこれを裏付けることとなった。次に、日英翻訳では挿入誤りが多く検出され、逆に英日翻訳では日本語で多様な活用誤りが多く検出されていることが分かる。

このように僅か 30 個の誤り n-gram に目を通すだけで、各翻訳システムが苦手とする分野に目を通すことがある程度できたことが分かる。

5.2 選択された箇所に対するフィルタリングの効果

本節では、翻訳誤りとして選択された箇所に対し、各フィルタリング法を適用した際の効果について、誤り箇所アノテーションコーパスを用いた自動評価により検証する。自動評価には、先行研究で提案されている機械翻訳結果を後編集した際の編集パターンを利用した手法 (赤部，Graham，Sakriani，戸田，中村 2014b) を利用する。評価の際は、事前に選択精度評価用の機械翻訳結果を後編集したコーパスを作成する。後編集のパターンから、機械翻訳結果の各部分に対して、挿入誤り、削除誤り、置換誤り、並べ換え誤りのラベルを付与することが可能である。これを誤り箇所の正解ラベルとし、評価用の機械翻訳結果に対して各誤り箇所選択法を適用した際に、誤り箇所の正解ラベルをどの程度予測できるかを適合率と再現率により評価する。

表 9: 3 種類のシステムで両方向の翻訳を行った際の比較。太字は各システムで発生した誤りの上位 3 種類。

誤りの種類	Ja → En			En → Ja		
	PBMT	HIERO	F2S	PBMT	HIERO	F2S
誤り箇所適合率	0.58	0.60	0.55	0.81	0.64	0.48
文脈依存置換	**0.41**	**0.33**	**0.36**	**0.10**	**0.17**	**0.52**
文脈非依存置換	0.03	0.08	0.15	**0.55**	0.03	**0.12**
挿入	**0.26**	**0.22**	**0.17**	0.06	0.13	**0.15**
削除	0.10	0.09	**0.18**	0.07	0.14	0.06
並べ替え	**0.13**	**0.28**	0.14	**0.19**	**0.32**	0.04
活用	0.07	0	0	0.04	**0.20**	**0.12**

5.2.1 実験設定

人手評価の際と同様に、機械翻訳システムとして京都フリー翻訳タスク（KFTT）(Neubig 2011) の日英データで構築された F2S システムを利用した。誤り候補のフィルタリングでは、正解訳として参照訳のみ利用した場合と、参照訳とオラクル訳の 2 つを利用した場合で比較を行った。オラクル訳の選択には、評価尺度として BLEU+1 または、METEOR version 1.5 (Denkowski and Lavie 2014) を利用し、それぞれ 500-best の中で評価尺度が最大となるものを選択し、比較を行った。パラフレーズラティスの構築のため、英語換言データベース（PPDB）(Ganitkevitch, Van Durme, and Callison-Burch 2013) の XL サイズ（43.2M ルール）を利用した。また日本語のラティス構築のため、日本語換言データベース (Mizukami, Neubig, Sakti, Toda, and Nakamura 2014) の XL サイズ（11.7M ルール）を利用した。

誤り箇所選択率の評価のため、KFTT の開発セットを日英翻訳した結果 503 文（12,333 単語）、英日翻訳した結果 200 文（4,846 単語）に対して後編集を行い、誤り箇所アノテーションコーパスを作成した。

5.2.2 参照訳による厳密一致フィルタリングの効果

予備実験として、コーパス全体をランダムに選択した場合と、参照訳によるフィルタリングを行った場合で、誤り箇所の選択精度がどのようになるか確認を行った。表 10 はすべての箇所をランダムに分析した場合の結果である。「フィルタリングなし」はチャンスレート、「フィルタリングあり」は分析の際にフィルタリングを行った結果である。[4] この表から、参照訳によるフィルタリングを行うだけでも、再現率の低下を抑えつつ適合率が大きく改善したことが分

[4] フィルタリングなしの再現率は 1.0 とならない。これはコーパスの中に機械翻訳結果を見ただけでは発見不能な削除誤りが含まれる場合があるためである。

表 10: 参照訳によるフィルタリングの効果

	適合率	再現率
フィルタリングなし	0.427	0.977
フィルタリングあり	0.492	0.959

表 11: 各フィルタリング法における適合率と再現率

正解訳	Ja→En				En→Ja			
	換言なし		換言あり		換言なし		換言あり	
	適合率	再現率	適合率	再現率	適合率	再現率	適合率	再現率
参照訳のみ	0.492	0.959	0.523	0.928	0.291	0.981	0.353	0.923
参照訳+オラクル（BLEU+1）	0.555	0.326	0.578	0.273	0.338	0.452	0.426	0.362
参照訳+オラクル（METEOR）	0.566	0.331	0.582	0.275	0.356	0.478	0.446	0.373

かる。

5.2.3 換言を考慮した正解訳の効果

各正解訳（参照訳、参照訳+オラクル訳）を用いたフィルタリング法を、換言あり・なしの場合について適用した実験を行った。表 11 は各設定における誤り箇所適合率と再現率の結果である。この表から、フィルタリングに用いる正解訳として、参照訳のみを用いた場合に比べて BLEU+1 によるオラクル訳を加えた方が適合率が高く、また評価尺度として METEOR を用いた場合は、BLEU+1 を用いた場合に比べ更に適合率が高くなったことが分かる。このことから、METEOR により選択されたオラクル訳は、機械翻訳の 1-best 出力で利用される語彙に似ており、換言表現が含まれにくくなっていることが分かる。

次に正解訳の換言を用いた場合の結果を見ても、選択箇所の誤り箇所適合率が高くなっていることが分かる。これらから、正解訳の換言を用いたフィルタリングを行うことによって、機械翻訳の誤り箇所がより適切に捉えられるようになったことが分かる。

一方、再現率について注意しなければならない点がある。特にオラクル訳を正解訳として利用した場合に、誤り箇所選択の再現率が大きく低下している。これは、オラクル訳は機械翻訳システムが出力した文であり、1-best と同様の誤りが発生する場合があるためである。しかし、今回提案した各手法は、コーパスの中の少なくとも 20%の誤り箇所を捉えており、提案法を利用する際には大きな問題とはならないと考えられる。誤り分析を効率的に行う際には、適合率の高い手法から先に利用し、選択された箇所を全て分析してしまった場合は順次再現率の高い選択法に切り替えることが可能である。

表 12 にフィルタリングされた箇所の例を示す。1 つ目の日英翻訳の例では、“foundation of” が誤り箇所の候補として選択されている。しかし参照訳に含まれる “a foundation for” は換言デー

表 12: 換言によりフィルタリングされた n-gram の例

1	Src	··· 、 狩野 派 様式 の 基礎 を 築 い た 。
	機械翻訳	... , and laid the foundation of the kano school 's style .
	参照訳	... , and built a foundation for the style of the kanoha group .
	換言	a foundation for ⟶ a foundation of
2	Src	... the members of the kanoha group ... and castles as the shogunate 's official painters ...
	機械翻訳	··· 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 、 ···
	換言	、 幕府 ⟶ は 幕府

表 13: 異なるドメインの PPDB を利用した場合の結果

正解訳	ドメイン外		ドメイン内	
	適合率	再現率	適合率	再現率
参照訳	0.523	0.928	0.547	0.881
参照訳+オラクル（BLEU+1）	0.578	0.273	0.596	0.207
参照訳+オラクル（METEOR）	0.582	0.275	0.603	0.210

タベースによると “a foundation of” に置き換えることが可能である。その結果、“foundation of” は誤りの候補から正しく除外された。2 つ目の英日翻訳の例では、換言データベースにより句点「、」が削除されたことで、不適切な選択箇所が正しく除外された。

この際注意すべきこととして、生成されたパラフレーズラティスが言語的に正しいものとは限らないという点が挙げられる。このため、誤った翻訳が発生している箇所が候補から除外される可能性もあることに注意されたい。

5.2.4 換言テーブルのドメインの影響

次に、日英翻訳において異なる換言データベースを使用した際の選択精度の調査を行った。前節の実験で利用した英語 PPDB には、分析対象である KFTT のデータが含まれていない。このため、KFTT のデータが含まれている日本語 PPDB の構築データを利用して英語の PPDB を新たに作成した。前者を「ドメイン外」、新しく作成した後者を「ドメイン内」とし、評価結果を表 13 に示す。

この表から、分析対象のドメインのデータが含まれた換言データベースを利用することで、誤り箇所選択の適合率が向上したことが分かる。換言データベースは機械翻訳のパラレルデータがあれば容易に作成可能なため (Bannard and Callison-Burch 2005)、誤り分析で利用する際には独自に作成することが望ましいと言える。

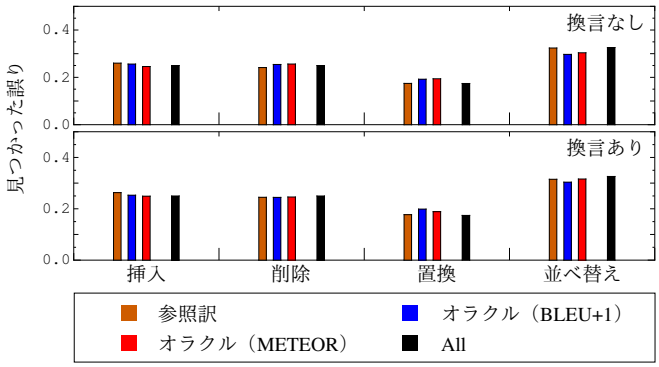


図 6: 各手法で選択された誤り箇所の分布。“All” は誤り分析コーパスに含まれる誤りの分布を示す。

表 14: コーパスに含まれる誤り分布と見つかった誤り分布の間の KL ダイバージェンス

正解訳	換言なし	換言あり
参照訳	0.00015	0.00003
オラクル（BLEU+1）	0.00102	0.00106
オラクル（METEOR）	0.00085	0.00033

5.2.5 選択された誤り箇所の分布

誤り箇所選択法によって見つかった誤りの傾向が、本来の誤り傾向と異なる場合、機械翻訳システムの傾向を正しく把握できないことにつながる。このため、誤り分析コーパスに含まれる誤りの分布と、各誤り箇所選択法によって見つかった誤りの分布の比較を行った。

各手法によって見つかった誤りの統計を図 6 に示し、表 14 に KL ダイバージェンス (Kullback and Leibler 1951) $D_{\mathrm{KL}}(P_{\mathrm{corpus}} \| P_{\mathrm{select}})$ を示す。ここで、P_{corpus} はコーパスに含まれる誤りの分布、P_{select} は各手法によって見つかった誤りの分布である。この結果から、参照訳を用いたフィルタリング法によって検出される誤りが、翻訳システムの誤り傾向を最も正確に捉えていると言えるが、他の手法でも KL ダイバージェンスの値が 0.001 程度に収まっている。この結果から、いずれの手法においても選択された誤りの種類に大きな偏りが生じず、機械翻訳システムの誤り傾向を適切に捉えていることが分かった。

5.3 誤選択箇所の分析

5.1 節及び 5.2 節の実験から、各誤り箇所選択法が誤って正しい翻訳箇所を選択する場合、またはフィルタリングによって誤り箇所が選択できなくなってしまう場合が存在することが明らかとなった。また、誤り箇所選択の自動評価の際、後編集結果に基づいて誤り箇所がアノテー

ションされたコーパスを用いるが、そもそもこのコーパスに誤りが含まれている場合は、精度評価が正しく行えないと考えられる。本節では、5.2 節と同様に、日英翻訳の誤り箇所アノテーションコーパスを用いて誤り箇所選択を行い、自動評価によって誤選択と判断された箇所について原因の調査を行った。

5.3.1 誤選択された正しい翻訳箇所に対する分析

誤り箇所選択法は、優先的に分析すべきと判断された箇所を選択するが、選択された箇所が本当は誤りでない場合がある。このような誤選択の分析を行うため、各手法により誤り箇所選択を行い、さらに選択箇所の自動評価を行った際に、誤選択と判断された部分について以下のアノテーションを人手で行う。

厳密一致: n-gram は正解訳にも存在し、換言を利用しないフィルタリングによって除外可能。

換言: n-gram は正解訳の局所的な換言であり、適切な換言ルールを利用できれば、除外可能。

統語的換言: n-gram は正解訳の換言だが、局所的な換言が困難と考えられる。文全体に影響する複雑な換言を利用できれば、除外可能。

正解訳の誤り: 正解訳が誤っているため、フィルタリングによって除外されない。

無くても良い: n-gram は正解訳に一致せず、フィルタリングできない。しかし、その n-gram が正解訳に含まれていなくても正解訳は誤りでない。

後編集誤り: 正解ラベルの誤り（後編集誤り）により誤選択と判断されたが、実際は適切な選択。

その他: 上記以外の誤り。n-gram が長すぎるためフィルタリングできない等。

○ 実験設定

京都フリー翻訳タスク（KFTT）の日英データで構築された F2S システムに対し、「識別言語モデルの重みに基づく誤り箇所選択」を行い、再現率が 5%となる上位の n-gram について分析を行った。選択の際、各手法によりフィルタリングを行った。オラクル訳を選択する際の評価尺度として BLEU+1 を利用した。

○ 実験結果

まず、選択箇所のフィルタリングを一切しない場合に検出された誤選択箇所の内訳を表 15 に示す。誤選択と判断された箇所の内、誤り箇所アノテーションコーパスの誤りであり、誤選択ではなかったものが僅か 3%であり、後編集による自動評価が十分効果的であると言える。

次に、各フィルタリング法を適用した場合に検出された誤選択箇所の個数を表 16 に示す。この結果から、識別言語モデルの重みに基づく誤り箇所選択を行った際に、参照訳を用いたフィルタリングを行うことによって 4 割以上の誤選択を回避できることが分かった。また、誤選択された箇所が正解訳の換言に含まれる場合、参照訳の換言を用いたフィルタリングによって 3

表 15: 誤選択された n-gram の内訳。正解訳を「参照訳」とした場合の統計。

誤選択の内容	誤選択箇所 [個]
厳密一致	202
換言	134
無くても良い	70
統語的換言	36
後編集誤り	16
正解訳の誤り	11
その他	15
合計	484

表 16: 各フィルタリング適用後の誤選択箇所の個数。括弧内の数字は、フィルタリングを行わなかった場合（-参照訳, -オラクル）に対する比率（%）。

種類	-参照訳	+参照訳					
	-オラクル	-オラクル				+オラクル	
	—	-換言		+換言		-換言	+換言
厳密一致	202	3	(1.5)	2	(1.0)	16 (7.9)	13 (6.4)
換言	134	131	(97.8)	87	(64.9)	40 (29.9)	19 (14.2)
統語的換言	36	31	(86.1)	15	(41.7)	5 (13.9)	5 (13.9)
正解訳の誤り	11	11	(100.0)	4	(36.4)	17 (154.5)	9 (81.8)
無くても良い	70	48	(68.6)	38	(54.3)	17 (24.3)	14 (20.0)
後編集誤り	16	15	(93.8)	8	(50.0)	5 (31.3)	4 (25.0)
その他	15	15	(100.0)	7	(46.7)	3 (20.0)	2 (13.3)
合計	484	254	(52.5)	161	(33.3)	103 (21.3)	66 (13.6)

割以上、さらにオラクル訳やオラクル訳の換言を用いたフィルタリングを合わせることで 8 割以上の誤選択を回避できることが分かった。

オラクル訳を正解訳としてフィルタリングを行った場合の「正解訳の誤り」がフィルタリングをしなかった場合に比べて多く現れている。これは、オラクル訳に含まれる誤りが参照訳に対して多いためである。また、「厳密一致」に分類される誤選択箇所であっても、フィルタリングで除外されない誤りがある。これは短い n-gram で一致していても、長い n-gram では一致しない場合にフィルタリングを通過し、誤選択されてしまうためである。

表 17 に誤り箇所の誤選択例を示す。「統語的換言」に分類された例を見ると、機械翻訳結果の “1392 , started” が誤り箇所として選択されている。これは参照訳の統語的な換言であり、実験で使用した PPDB では対応できないため、参照訳のみを正解訳とした場合は誤り箇所として扱われてしまう。しかし、オラクル訳は機械翻訳結果と同じ文の構造をしており、“began” を

表 17: 各種類の誤選択例。枠で囲まれた文字列は誤選択箇所を示す。

誤選択の種類	原文	こ う し た 時代 背景 か ら 、 … と い う 見 方 も あ る 。
換言	参照訳	given such an historical backdrop , it can be said that ...
	オラクル訳	from such backdrop , believe that mainly made efforts ...
is also thought	機械翻訳	it is also thought that this from such backdrop , ...
→ can be said	後編集	it is also thought that , against such a backdrop , ...

誤選択の種類	原文	1392 年 、 夢窓 疎石 に よ り 始 ま る 。
統語的換言	参照訳	the sect began by soseki musou in 1392 .
	オラクル訳	in 1392 , began by muso soseki .
in 1392 , [X]	機械翻訳	in 1392 , started by muso soseki .
→ [X] in 1392	後編集	in 1392 , started by muso soseki .

“started” に置き換えるだけで選択箇所に一致する。このため、オラクル訳の換言を使ったフィルタリングによって分析対象から除外可能となる。

5.3.2 選択されなかった誤り箇所に対する分析

誤り箇所選択によって選択された箇所に対してフィルタリング法を適用することで、正解訳に一致する n-gram や正解訳の換言に含まれる n-gram を誤り箇所から除外することができる。しかしそれらの手法によって、逆に正しく選択されるべき機械翻訳の誤り箇所を誤り箇所の候補から除外する場合があり、再現率の低下として現れている。このような問題の分析を行うため、誤り箇所アノテーションコーパスで誤りとされている箇所で、フィルタリングにより選択できなくなる部分について、以下の基準に従って分類を行う。

誤った部分に一致: 正解訳の異なる位置に対応する n-gram に一致した。

誤った換言: 換言テーブルの不適切なルールが使用された。

文脈的に誤った換言: この文脈では使うべきでない換言ルールが使用された。

文脈的後編集: 文脈に依存する誤り箇所。後編集の表現方法を変えれば、誤り箇所ではなくなる。

正解訳の誤り: 正解訳が誤っているため、誤り箇所がフィルタリングによって除外された。

後編集誤り: 正解ラベルの誤り（後編集誤り、または不要な後編集）により誤り箇所とされているが、実際は適切な翻訳。

日本人の名前: 日本人の名前（姓名の順序が正解訳・機械翻訳結果と後編集の間で異なる）。コーパス特有の問題であり後編集誤りに分類できるが、多く含まれているため特別に分類を行う。

○ 実験設定

5.2 節で利用した日英機械翻訳の誤り箇所アノテーションコーパスについて、「参照訳を用いた厳密一致フィルタリング」及び「参照訳のパラフレーズを用いたフィルタリング」を適用し、

表 18: フィルタリングで除外された誤り箇所の内訳

	正解訳: 参照訳	
換言	なし	あり
日本人の名前	24	43
後編集誤り	11	40
誤った換言	0	18
文脈的に誤った換言	0	12
誤った部分に一致	0	8
文脈的後編集	1	7
正解訳の誤り	1	4
合計	38	136

選択されなくなってしまう誤り箇所の調査を行った。

○ 実験結果

各フィルタリング法を適用することによって選択されなくなった誤り箇所の統計を表 18 に示す。この結果から、参照訳のみによるフィルタリングを行った場合、選択されなくなる箇所の約 3 割は誤り箇所アノテーションコーパスの誤りによるもの、約 6 割は姓名の順序の違いに起因する誤りであり、実用上問題となる誤り箇所がほとんど除外されていないことが分かった。次に、参照訳の換言によるフィルタリングを適用した場合の結果を見ると、間違った換言が使用されたことによる誤選択が 20%以上あることが分かった。

また、誤り箇所アノテーションコーパスの誤りにより誤り箇所として誤判断された箇所が約 3 割検出されており、各選択法の再現率を評価する際、無視できないほどの影響が出ることが分かった。

5.4 誤り箇所選択の誤り分析における効果

本節では、実際の誤り分析を想定し、各誤り箇所選択法を用いて一定時間分析を行った際の効果を検証する。

図 7 は本節の実験の手順を示す。まず、3 章で述べた各手法によって n-gram にスコアを与え、優先的に分析すべき n-gram を順に抽出する。次に、機械翻訳の訳出の中で各 n-gram が含まれている文を列挙し、n-gram に一致する箇所を選択する。その際、4 章で述べたフィルタリング処理を行う。分析シートは、n-gram が正解訳に含まれないものを先に表示し、正解訳に含まれるものを後に表示するようにした。このようにすることで、分析者はフィルタリングの対象とならなかった結果を優先的に分析しつつ、分析者の時間が許せば、誤ってフィルタリングされた機械翻訳文も分析対象とすることができる。分析者は各 n-gram が選択した箇所について誤

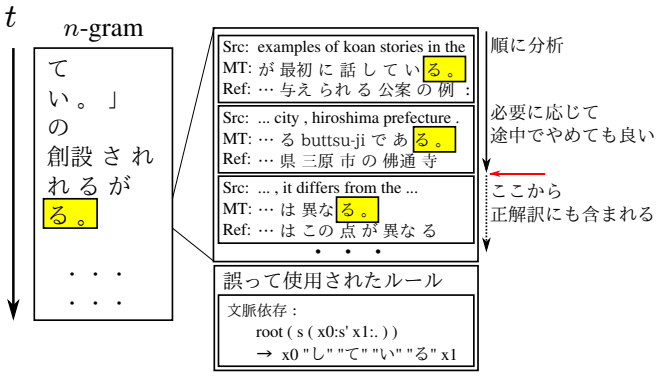


図 7: 分析時間と誤り発見数の関係の調査

り分析を行い、翻訳時に誤って使用された翻訳ルールを記録する。その際、誤り箇所が 5.1 節で述べた「文脈依存誤り」か「文脈非依存誤り」かを記録しておくことで、翻訳ルールそのものが誤っているのか、あるいはモデル化が誤っているのかが把握可能となる。1 個の n-gram により複数の文が選択された場合は、実際の誤り分析と同様に、分析者の判断ですべての文を見ずに分析を中断しても良いこととする。最後に、各 n-gram 毎に誤り分析に要した時間を記録する。

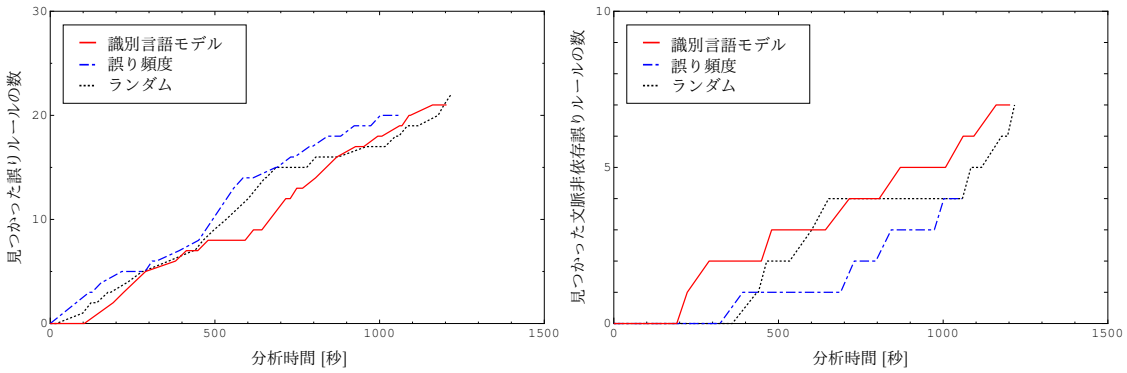
5.4.1 実験設定

機械翻訳システムとして京都フリー翻訳タスク（KFTT）で構築された f2s 英日翻訳システムを利用した。n-gram のスコアリングに「ランダム」、「誤り頻度」、「識別言語モデルの重み」に基づく 3 つの手法を利用し、自動評価で F 値が最大となった「参照訳の換言によるフィルタリング」を利用した。KFTT の開発セットに対して誤り箇所選択を行い、誤って使用された翻訳ルールを記録した。

5.4.2 実験結果

各手法を利用して誤り分析を行った際に、経過した分析時間と誤って使用された翻訳ルールが発見された個数の関係を図 8(a) に示す。また図 8(b) は発見された誤りの中でも文脈非依存誤りの原因となるルールが見つかった個数を示す。グラフの傾きが大きいほど、誤りルールを効率的に発見できることを意味する。これらの結果から、各手法とも分析時間と誤りルール発見数の間に大きな違いは見られなかった。一方で、文脈非依存誤りの原因に限って見れば、識別言語モデルの重みに基づく誤り箇所選択では、他の手法に比べて早い段階から誤りが見つかることが分かった。

文脈非依存誤りは、その誤りを修正しようとした際に文脈を考慮する必要がないため、文脈



(a) すべての種類 (b) 文脈非依存誤りのみ

図 8: 分析時間と記録された誤りルール数の関係

表 19: 文脈非依存誤りの原因として記録されたルールが KFTT のテストセット翻訳時に使用された回数

スコアリング手法	使用された回数
ランダム	0
誤り頻度	0
識別言語モデルの重み	1

依存誤りに比べて誤りを容易に修正できる。このため、識別言語モデルの重みに基づく手法を利用することで、修正が容易な誤りを早期に発見することができ、システムの改善を比較的効率良く行うことができると言える。

次に、文脈非依存誤りの原因として記録された翻訳ルールを機械翻訳システムから削除することによって、システムをどの程度改善できるかを見積もった。KFTT のテストセット 1,160 文を機械翻訳した際、21,080 個の翻訳ルールが使用された。この内、各手法で文脈非依存誤りの原因として記録されたルールが使用された回数を表 19 に示す。

この結果から、文脈非依存の誤りの原因となるルールを具体的に記録しても、そのルールが機械翻訳システムで使用されることは稀であることが分かる。翻訳システムの誤りを修正する際には、見つかった誤りルールを 1 つずつ修正するのではなく、見つかった誤りルールを一般化し、テストセットにおけるカバー率を向上させる必要がある。

6 おわりに

本論文では、機械翻訳システムの比較・改善のための誤り分析を効率的に行うことを目的として、機械学習の枠組みを利用した機械翻訳の誤り箇所選択法、及び選択箇所のフィルタリング法を提案した。その結果、人手評価において従来法に比べて高い精度で適切な誤り箇所を捉えることに成功した。また、優先的に選択された少量の誤り箇所を分析するだけで、各システムの誤り傾向を捉えることができ、システム間比較の効率化に貢献した。

次に、機械翻訳の誤り箇所選択法が誤選択した箇所の分析を行ったところ、オラクル訳や換言を利用したフィルタリングは適合率の向上に効果的であるが、誤った換言が使用されることによる再現率の低下が明らかとなった。

最後に、今回の提案法を実際の誤り分析に利用した場合の効果を検証した。その結果、翻訳システムを容易に修正可能な文脈非依存誤りについては、提案法により比較的早い段階から捉えることが可能であることが分かった。一方ですべての種類の誤りについて見ると、各手法とも誤りの発見数に大きな違いが見られなかった。この理由として、各手法によって選択された誤り箇所の特徴が挙げられる。誤り頻度に基づき選択された誤り箇所は、識別言語モデルの重みに基づいて選択された箇所に比べ、目的言語に頻繁に出現する n-gram を多く含む。このため、識別言語モデルの重みに基づく手法を利用した際、誤り分析者が比較的効率良く選択箇所に目を通すことができたと考えられる。今回の実験では、目を通した文の数については記録を行っていないため、今後の調査項目として検討する必要がある。

また、発見したルールを単独で見ても、システム全体から見ればそのような翻訳ルールが使用されることはごく稀であることが分かった。一方で、具体的な誤りルールを一般化することで、同様の翻訳ルールをまとめて修正することは可能と考えられる。

今後の課題として、見つかった具体的な誤りをどのように一般化するかを検討する必要がある。具体的には、見つかった翻訳ルールを品詞列などのより抽象的な情報に自動的に変換することや、誤ったルールを元に、人手によって複数の修正ルールを列挙する手法が考えられる。

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

Bach, N., Huang, F., and Al-Onaizan, Y. (2011). "Goodness: A Method for Measuring Machine Translation Confidence." In *Proc. ACL*, pp. 211–219.

Banerjee, S. and Lavie, A. (2005). "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Bannard, C. and Callison-Burch, C. (2005). "Paraphrasing with bilingual parallel corpora." In *Proc. ACL*, pp. 597–604.

Church, K. W. and Hank, P. (1990). "Word association norms, mutual information, and lexicography." *Computational Linguistics*, **16** (1), pp. 22–29.

Collins, M. (2002). "Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms." In *Proc. EMNLP*, pp. 1–8.

Denkowski, M. and Lavie, A. (2014). "Meteor Universal: Language Specific Translation Evaluation for Any Target Language." In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Doddington, G. (2002). "Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics." In *Proc. HLT*, pp. 128–132, San Diego, CA.

Duchi, J. and Singer, Y. (2009). "Efficient online and batch learning using forward backward splitting." *Journal of Machine Learning Research*, **10**, pp. 2899–2934.

El Kholy, A. and Habash, N. (2011). "Automatic Error Analysis for Morphologically Rich Languages." In *Proc. MT Summit*, pp. 225–232.

Fishel, M., Bojar, O., Zeman, D., and Berka, J. (2011). "Automatic translation error analysis." In *Text, Speech and Dialogue*, pp. 72–79. Springer.

Flanagan, M. (1994). "Error classification for MT evaluation." In *Proc. AMTA*, pp. 65–72.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). "PPDB: The Paraphrase Database." In *Proc. NAACL*, pp. 758–764.

Kirchhoff, K., Rambow, O., Habash, N., and Diab, M. (2007). "Semi-automatic error analysis for large-scale statistical machine translation systems." In *Proc. MT Summit*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). "Moses: Open Source Toolkit for Statistical Machine Translation." In *Proc. ACL*, pp. 177–180.

Kullback, S. and Leibler, R. (1951). "On Information and Sufficiency." *The Annals of Mathematical Statistics*, **22**, pp. 79–86.

Lin, C.-Y. and Och, F. J. (2004). "Orange: a method for evaluating automatic evaluation metrics for machine translation." In *Proc. COLING*, pp. 501–507.

Mackay, D. J. and Petoy, L. C. B. (1995). "A Hierarchical Dirichlet Language Model." *Natural Language Engineering*, **1**.

Mizukami, M., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (2014). "Building a Free, General-Domain Paraphrase Database for Japanese." In *Proc. COCOSDA*.

Neubig, G. (2011). "The Kyoto Free Translation Task." `http://www.phontron.com/kftt`.

Neubig, G. (2013). "Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers." In *Proc. ACL Demo Track*, pp. 91–96.

Och, F. J. (2003). "Minimum Error Rate Training in Statistical Machine Translation." In *Proc. ACL*, pp. 160–167.

Och, F. J. and Ney, H. (2003). "A Systematic Comparison of Various Statistical Alignment Models." *Computational Linguistics*, **29** (1), pp. 19–51.

Onishi, T., Utiyama, M., and Sumita, E. (2010). "Paraphrase Lattice for Statistical Machine Translation." In *Proc. ACL*, pp. 1–5.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). "BLEU: a method for automatic evaluation of machine translation." In *Proc. ACL*, pp. 311–318.

Popović, M. and Ney, H. (2011). "Towards automatic error analysis of machine translation output." *Computational Linguistics*, **37** (4), pp. 657–688.

Roark, B., Saraclar, M., and Collins, M. (2007). "Discriminative n-gram language modeling." *Computer Speech & Language*, **21** (2), pp. 373–392.

Specia, L., Turchi, M., Cancedda, N., Dymetman, M., and Cristianini, N. (2009). "Estimating the sentence-level quality of machine translation systems." In *Proc. EAMT*, pp. 28–37.

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

赤部晃一，GrahamNeubig，SakrianiSakti，戸田智基，中村哲 (2014a). 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.

赤部晃一，GrahamNeubig，SakrianiSakti，戸田智基，中村哲 (2014b). パラフレーズを考慮した機械翻訳の誤り箇所選択. 情報処理学会 第 219 回自然言語処理研究会 (SIG-NL), 神奈川.

略歴

赤部 晃一：2015 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了。同大学院博士後期課程に在学中。機械翻訳、自然言語処理に関する研究に従事。

Graham Neubig： 2005 年米国イリノイ大学アーバナ・シャンペーン校工学部コンピュータ・サイエンス専攻卒業。2010 年京都大学大学院情報学研究科修士課程修了。2012 年同大学院博士後期課程修了。同年奈良先端科学技術大学院大学助教。機械翻訳、自然言語処理に関する研究に従事。

Sakriani Sakti： 1999 年インドネシア・バンドン工科大学情報卒業。2002 年ドイツ・ウルム大学修士、2008 年博士課程修了。2003〜2011 年 ATR 音声言語コミュニケーション研究所研究員、情報通信研究機構主任研究員。現在、奈良先端科学技術大学院大学情報科学研究科助教。2015〜2016 年フランス INRIA 滞在研究員。統計的パターン認識、音声認識、音声翻訳、認知コミュニケーション、グラフィカルモデルの研究に従事。JNS、SFN、ASJ、ISCA、IEICE、IEEE 各会員。

戸田 智基：1999 年名古屋大学工学部電気電子・情報工学科卒業。2003 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了。同年日本学術振興会特別研究員-PD。2005 年奈良先端科学技術大学院大学情報科学研究科助手。2007 年同助教。2011 年同准教授。2015 年より名古屋大学情報基盤センター教授。工学博士。音声情報処理の研究に従事。IEEE、電子情報通信学会、情報処理学会、日本音響学会各会員。

中村 哲：1981 年京都工芸繊維大学工芸学部電子工学科卒業。京都大学工学博士。シャープ株式会社。奈良先端科学技術大学院大学助教、2000 年 ATR 音声言語コミュニケーション研究所室長、所長。2006 年（独）情報通信研究機構研究センター長、けいはんな研究所長などを経て、現在、奈良先端科学技術大学院大学教授。ATR フェロー。カールスルーエ大学客員教授。音声翻訳、音声対話、自然言語処理の研究に従事。情報処理学会喜安記念業績賞。総務大臣表彰、文部科学大臣表彰、Antonio Zampoli 賞受賞。IEEE SLTC 委員、ISCA 理事、IEEE フェロー。