機械翻訳の誤り箇所選択法における誤選択箇所の調査

赤部 晃一　Graham Neubig　Sakriani Sakti　戸田 智基　中村 哲
奈良先端科学技術大学院大学　情報科学研究科
{akabe.koichi.zx8, neubig, ssakti, tomoki, s-nakamura}@is.naist.jp

1　はじめに

最先端の機械翻訳システムは年々精度が向上しているが，その反面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，翻訳結果を用いた評価実験によって改善すべき点を見出だすことが広く行われている。その際，翻訳結果を一文ずつ見ることで誤りを見つけることもできるが，その作業は非常に労力がかかり，見つけた誤りがシステム全体を大きく改善できるとも限らない。しかし，誤り箇所の検出や，誤りに対する重要度の付与を自動的に行えれば，システムの不得意な翻訳現象を容易に発見でき，分析作業の効率化を図ることが可能となる。

先行研究として，機械翻訳の誤り箇所を自動的に抽出するための複数の手法が提案されている。文献 [11, 5, 3] では，機械翻訳と事前に人手で翻訳された参照訳の差分を様々な情報で分類し，それぞれの誤りを頻度順に並べる手法が提案されている。結果一つ一つを眺める前におおよその流れを定量化できるという意味で，有用な方法である。文献 [1] では，分析単位としてある大きさの n-gram を設定し，n-gram を与えられたスコアに従ってソートすることで，分析すべき箇所に優先順位を設ける手法が提案されている。次に，誤りとして特定された箇所にフィルタリングを行い，選択精度を向上させる手法が提案されている。文献 [12] では，参照訳の換言を用いて誤り箇所選択時にフィルタリングを行い，意味が正しく表層的な文字列が異なるものを誤り箇所から除外する手法が提案されている。

このように，機械翻訳の誤り箇所を特定するために様々な手法が提案されているが，最先端の手法でも 58% の適合率という比較的低精度である一方で，各手法が誤選択した箇所については十分な分析が行われていないのが現状である。本研究は，先行研究で提案されている各誤り箇所選択手法について，現状どのような誤り箇所を選択できるのか，どのような誤り箇所を選択できないのかを調査する。具体的には，まず n-gram のスコアに基づき機械翻訳の誤り箇所選択を行った際に，誤選択と判断された箇所についてその原因をアノテーションする。次に，それらの誤選択が現状提案されている手法によってどの程度改善されるかを測定する。これらの実験を通して，どのような改善が必要とされているのかを提案することを目的とする。

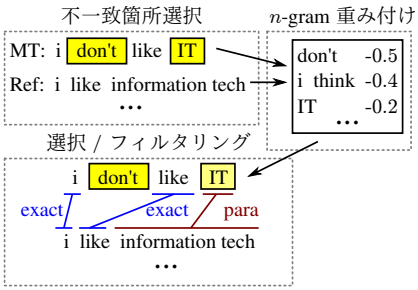

図 1　誤り箇所選択の流れ

2　機械翻訳の誤り箇所選択法

本節では，本研究が対象とする誤り箇所選択法について説明を行う。本手法は，図 1 に示すように，誤りと考えられる箇所の選択，選択された箇所のフィルタリングの 2 つの要素からなる。

2.1　参照訳との差分に基づく選択

本手法は，機械翻訳システムが出力した翻訳仮説に含まれる n-gram の内，参照訳に含まれないものを誤り候補として検出する。その際に，翻訳仮説が参照訳と異なっていても正しい場合があるため，誤り箇所として選択された箇所が実際には誤りでない可能性があることに注意されたい。参照訳は，人手により機械翻訳とは独立に翻訳された文であるため，使用される語彙に差が生じやすくなる。このため，意味が正しいにもかかわらず表層的な文字列の不一致により誤り箇所として検出されてしまう箇所が発生しやすくなる。

2.2　n-gram に基づく誤り箇所選択

前節では，翻訳仮説と参照訳の間で異なる n-gram をすべて誤り箇所の候補として抽出した。しかし複数の誤り箇所の候補がある場合に，重要と考えられる箇所，もしくは誤っている可能性がより高い箇所を優先的に分析できれば，誤り分析を効率的に行えるようになると考えられる。文献 [1] では，n-gram を素性とした際の識別言語モデルに含まれる素性の重みによって誤りの候補を並べ替える手法が提案されている。識別言語モデルは，学習データとして機械翻訳システムが出力した n-best リストを利用し，識別言語モデルは翻訳仮説を正解訳に修正するように学習される。その際，重みが負に大きくなった素性は翻訳精度を低くする原因と考えられる。このため，誤り分析の際に重みの低い n-gram を重点的に分析すれば，システムにとって重要な誤りを捉えやすくなると考えられる。

2.3 選択箇所のフィルタリング

n-gram に基づく誤り箇所選択を行う場合，正解訳を考慮しなければ明らかに誤っていないと考えられる箇所を誤選択する恐れがある。文献 [1] では 2.1 節の手法と組み合わせ，選択された n-gram が正解訳に含まれている場合に，選択箇所から除外する処理を行っている。さらに文献 [12] では，正解訳のパラフレーズラティス [9] を生成し，誤り候補となった n-gram がラティス上に存在する場合に選択箇所から除外する枠組みを導入している。先行研究では正解訳として参照訳を利用しているが，本研究では機械翻訳システムが出力した候補の中で自動評価尺度が最大となるオラクル訳も利用する。オラクル訳を利用する際の効果として，参照訳を用いた場合に比べ，翻訳仮説と正解訳の表層的な異なりが小さくなることが挙げられる。参照訳は機械翻訳とは独立に人手により翻訳されるため，使用する語彙が機械翻訳とは異なる場合が多いと考えられるが，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，それら 2 つの差分を見た際，表層的に異なりながら意味が正しいものを翻訳誤りとして誤選択してしまう可能性が低くなると考えられる。

2.4 誤り箇所選択の精度評価

誤り箇所の選択精度を測ることも，本タスクの重要な要素である。文献 [1] では特定された箇所を人手により調査し，誤り箇所の適合率を計測している。しかし，各手法による選択箇所をすべて人手で調査することは労力がかかる上，選択されたものの内ごく一部しか調査できず，評価の揺れも大きくなる。文献 [12] では，この問題を解決するために，誤り箇所が事前にアノテーションされたコーパスを正解ラベルとして利用し，コーパスに対して誤り箇所選択を行うことで適合率と再現率を調査している。これにより，誤り箇所選択の精度を高速に一貫して行うことが可能となる。しかし，正解ラベルの誤りにより，この評価に誤りが生じる可能性もある。

3 翻訳誤りの誤選択箇所の調査

前節で述べた機械翻訳の誤り箇所選択法は，それぞれ失敗する可能性がある。本節では前述の誤り箇所選択法が失敗する原因の調査を行う手法について説明する。

3.1 誤選択された正しい翻訳箇所に対する分析

2.2 節の手法では，重要な誤りと判断された箇所を優先的に分析するが，重要と判断された箇所が実際には誤りでない場合がある。その原因の一つとして，ある n-gram が特定の文脈では重大な誤りを引き起こすが，別の文脈では問題とはならない場合がある。さらに，2.4 節の誤り箇所選択の自動評価では，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，精度評価が正しく行えない可能性がある。そこで本研究では，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行う。

Exact match: 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。

Paraphrase: 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。

Syntactic paraphrase: 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。

Correct-translation error: 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。

Optional: 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。

Postedit error: 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。

Others: 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。

3.2 選択されなかった誤り箇所に対する分析

2.3 節の手法により，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してしまう場合がある。例えば正解訳が誤っていたり，不適切な換言が利用されたりすることが主な原因として考えられる。そこで，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されなかった部分について，以下の基準に従って分類を行う。

Matched wrong segment: 正解訳の異なる位置に対応する n-gram に一致した。

Wrong paraphrase: 換言テーブルの不適切なルールが使用された。

Contextual wrong paraphrase: この文脈では使うべきでない換言ルールが使用された。

Contextual postedit: 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。

Correct-translation error: 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。

Postedit error: 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。

Japanese name: 日本人の名前（姓名の順序が正解訳・機械翻訳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。

4 評価実験

各手法による誤り箇所選択における誤選択箇所の調査のため，以下の評価実験を行った。

4.1 実験設定

分析対象として，京都フリー翻訳タスク (KFTT, [7]) のデータ (表 1) から学習された日英翻訳システムを利

表 1　KFTT のデータサイズ

	文数	単語数	
		英語	日本語
学習	330k	5.91M	6.09M
dev セット	1166	24.3k	26.8k
test セット	1160	26.7k	28.5k

表 2　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]
Exact match	202
Paraphrase	134
Optional	70
Syntactic paraphrase	36
Postedit error	16
Correct-translation error	11
Others	15
合計	484

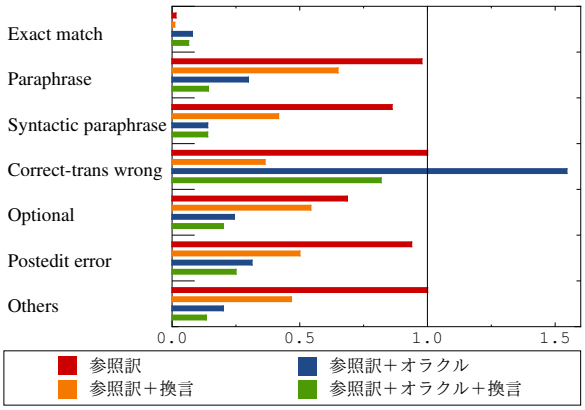

図 2　各フィルタリング適用後の誤選択箇所（フィルタリング前に対する比率）

用した。システムとして Travatar[8] に基づいて構築された forest-to-string モデルを用いた。単語アラインメントは，構文情報を用いる Nile[*1]で行い，日英それぞれの構文解析に Egret[*2]を利用した。システムのチューニングでは，機械翻訳システムの評価で一般的に利用される BLEU[10] を評価尺度として利用した。

誤り箇所選択に用いる識別言語モデルの学習には，構造化パーセプトロン [2] を利用し，反復回数を 100 回，n-gram 素性の最大長を 3 とした。学習には，KFTT の dev セットの機械翻訳の 500-best を利用し，オラクル訳の選択では，評価尺度に文単位の尺度である BLEU+1[6] を利用した。正則化係数が 1.0×10^{-7} から 1.0×10^{-2} の範囲で test セットの BLEU+1 が最大となるようにモデルを学習した。分析対象とする n-gram は，識別言語モデルの重みにより優先的に分析すべきと判断されたもので，コーパス全体に対し 5% の誤

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

表 3　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
Japanese name	24	43
Postedit error	11	40
Wrong paraphrase	0	18
Contextual wrong paraphrase	0	12
Matched wrong segment	0	8
Contextual postedit	1	7
Correct-translation error	1	4
合計	38	136

りを捉える上位の n-gram とした。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。オラクル訳には，500-best の中で BLEU+1 が最大となるものを利用した。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）[4] の XL サイズ（43.2M ルール）を利用した。

4.2　誤選択された n-gram の統計

誤り箇所アノテーションコーパスに対し，識別言語モデルの重みに基づく誤り箇所選択を行った。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 2 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実際には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，各フィルタリング手法を適用した場合に検出された誤選択箇所の，フィルタリングを使用しない場合に対する比率を図 2 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る。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，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によって 8 割以上の誤選択を回避できることが分かる。同様に，正解訳の統語的な換言であっても，オラクル訳を正解訳とすることによって 8 割以上回避できることが分かる。

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 “Correct-trans error（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，Exact match 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てしまうためである。

4.3　フィルタリングで除外された誤り箇所の統計

誤り箇所アノテーションコーパスに対し，各フィルタリング手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 3 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

表 4　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 。
Paraphrase	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 → can be said	機械翻訳	it is also thought that this from such backdrop , ...
	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 。
Syntactic paraphrase	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] → [X] in 1392	機械翻訳	in 1392 , started by muso soseki .
	後編集	in 1392 , started by muso soseki .

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る。

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不必要なフィルタリングが 2 割以上発生したことが分かる。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る。

表 4 に誤り箇所の誤選択例を示す。“Syntactic paraphrase” に分類された例を見ると，機械翻訳の “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と同じ文の構造をしており，“began” を “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することができる。

5　まとめ・今後の課題

本稿では，機械翻訳の誤り箇所選択法が誤選択した箇所について分析を行った。その結果，誤選択箇所の多くは参照訳にも含まれている n-gram であり，単純に参照訳を使ってフィルタリングを行うだけで多くの誤選択を回避できることが分かった。さらに換言テーブルを用いたフィルタリングでは，単純に正解訳のフレーズ置換に基づく換言を考慮するだけでも，正解訳の統語的な換言となっている誤選択箇所を選択候補から除外できることが分かった。一方で，換言テーブルを用いたフィルタリングでは，誤った換言ルールが使用されたり，換言ルールが誤った文脈で使用されたことにより，本来選択されるべき機械翻訳の誤り箇所を除外してしまう場合が明らかになった。また，誤り箇所アノテーションコーパスの精度に起因する問題が明らかとなった。特に後編集の誤りにより，正しく翻訳されている箇所が誤り箇所としてラベル付けされており，誤り箇所選択の再現率が本来より低く見えてしまう場合があることが分かった。

今後の展望として，機械翻訳の誤り箇所をより適切に捉えるために，より高精度な換言テーブルの使用や，文脈を考慮した換言を用いたフィルタリング手法が必要とされる。また，誤り箇所選択法の精度評価のために，より正確な後編集・アライメントを行うための枠組みが必要と考える。

謝辞

本研究の一部は，（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。

参考文献

[1] K. Akabe, G. Neubig, S. Sakti, T. Toda, and S. Nakamura. Discriminative language models as a tool for machine translation error analysis. In *Proc. COLING*, pp. 1124–1132, 2014.

[2] M. Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proc. EMNLP*, pp. 1–8, 2002.

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

[4] J. Ganitkevitch, B. Van Durme, and C. Callison-Burch. PPDB: The paraphrase database. In *Proc. NAACL*, pp. 758–764, 6 2013.

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

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

[7] G. Neubig. The Kyoto free translation task. `http://www.phontron.com/kftt`, 2011.

[8] G. Neubig. Travatar: A forest-to-string machine translation engine based on tree transducers. In *Proc. ACL Demo Track*, pp. 91–96, 2013.

[9] T. Onishi, M. Utiyama, and E. Sumita. Paraphrase lattice for statistical machine translation. In *Proc. ACL*, pp. 1–5, 2010.

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

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

[12] 赤部, G. Neubig, S. Sakti, 戸田, 中村. 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), 東京, 12 2014.