単語翻訳モデルを用いた翻訳後編集による湧き出し語対策

土居誉生[†,☆] 隅田英一郎[†,☆☆]

コーパスベース翻訳は高い翻訳品質を実現しうる有望な技術である．しかし，入力文と関連のない語が訳文に湧き出すなど，特徴的な翻訳誤りも観察される．本稿では，この湧き出し誤りへの対処として，後編集により自動修正するアプローチを提案する．提案手法は，単語翻訳モデルを利用して誤り語候補を検出し，修正処理を起動する．誤り語候補は，対訳コーパスから得られた用例を利用した制約の検証を経て，訳文から削除される．日英および中英翻訳を対象とした複数のシステムを使った実験で，誤り語自動削除による翻訳精度向上効果が確認された．

Post-edit of Machine Translation Using a Lexicon Model to Cope with Out-of-the-blue Words

Takao Doi[†,☆] and Eiichiro Sumita[†,☆☆]

Although corpus-based machine translation is a promising technology, one error that is typically observed is out-of-the-blue words, i.e., words that appear in the translation that are unrelated to the input information. Our approach to correct the error is an automatic post-edit of the translations. The proposed method finds error clues based on a lexicon model and triggers the correction process. The error words derived from the clues are deleted from the translation after a constraint check using translation examples. We conducted an experiment using several translation systems on Japanese-to-English and Chinese-to-English translations, whose results showed improvement on translation accuracy by automatically deleting unrelated words.

1. はじめに

近年，対訳コーパスから自動的に翻訳システムを構築するコーパスベースの翻訳技術の研究開発がさかんになってきており，その性能も向上している[1)]．しかしながら用例翻訳や統計翻訳などのコーパスベース翻訳では，翻訳知識を人間の手を介さずに自動的に構築するゆえに，ときとして人間にとって考えられないような翻訳誤りを犯す．特に，原文でまったく触れてもいない概念を表す語が訳文に現れる場合，その誤りの奇異な印象は強く，たとえ全体的な性能が高くても，システムへの不信感をユーザに与えてしまう．下の例では，入力文とは関連のない “departure” という語が訳文に現れてしまっている[2)]．

入力文：お名前とお部屋番号を教えて下さい

翻訳結果：What is the name and departure room number?

この例のように入力文と関連なく訳文中に湧き出した語を，我々は湧き出し語と名付ける．

本稿では，湧き出し語問題に対処するために，後編集により翻訳誤りを自動修正するアプローチについて論ずる．翻訳システム内部の改良ではなく，後編集のアプローチをとることで，複数種類の翻訳システムの翻訳結果への適用が可能となる．一般に後編集による修正には削除・挿入・置換の操作があるが，今回は湧き出し語の削除に焦点を当てる．提案手法では，統計翻訳で用いられる単語翻訳モデルを利用して削除すべき訳語の候補を検出し，対訳コーパスから得られた用例を使って候補を絞り込む．

以降の章では，湧き出し語の発生の仕組み，我々のアプローチと関連研究，提案手法と実験について報告する．実験では，複数の翻訳システムを対象とし，訳

† ATR 音声言語コミュニケーション研究所
ATR Spoken Language Communication Research Laboratories
☆ 現在，株式会社 CSK システムズ，神戸大学大学院自然科学研究科
Presently with CSK Systems Corporation and Graduate School of Science and Technology, Kobe University
☆☆ 現在，独立行政法人情報通信研究機構，神戸大学大学院自然科学研究科
Presently with National Institute of Information and Communications Technology and Graduate School of Science and Technology, Kobe University

語自動削除による翻訳精度向上効果を検証する．

2. 湧き出し語問題

どのような方式の機械翻訳システムでも，完全なものはなく，何らかの翻訳誤りを犯す．対訳コーパスから翻訳知識を学習するコーパスベース翻訳方式のシステムでも，個々に様々な誤りを犯すが，共通して目に付くのは湧き出し語の問題である．たとえば国際ワークショップ IWSLT-2004 [1),3)] 評価キャンペーンではコーパスベース翻訳システムは高い翻訳精度を示したが，やはり湧き出し語問題が観察される．特に好成績をあげた統計翻訳システムにおいて残された問題の中で大きなものが湧き出し語である☆．

湧き出し語の概念は，ユーザに提示される翻訳結果で観察される問題を示し，それがいかに生成されたかのシステムの内部事情を区別しない．しかしコーパスベース翻訳という枠内で考えると，湧き出し語の発生原因の特徴をとらえることができる．湧き出し語の多くは単語アラインメント[4)] に起因する．単語アラインメントとは，対訳を構成する原文と訳文の単語間の対応関係を求める処理およびその結果の情報を指す．単語アラインメントは，コーパスベース翻訳において翻訳知識の学習に必須の基本的な処理および情報と位置付けられる．

2.1 用例翻訳における問題

まず用例翻訳方式における湧き出し語について例をあげて説明する．以下，1 章で上げた例について，その発生過程を示す．

入力文：お名前とお部屋番号を教えて下さい

に対して，次の類似用例が見つかる．

原文：便名と時間を教えて下さい

訳文：What is the flight number and departure time?

入力文は，用例原文の「便名」を「お名前」,「時間」を「お部屋番号」でそれぞれ置換したものと見なされる．ここでシステムは,「便名」と「時間」の用例訳文中の対応箇所はそれぞれ “flight number” と “time” であると判断する．これらの用例訳文中の対応箇所に「お名前」と「お部屋番号」の訳語が入れられ，次の翻訳結果が得られる．

翻訳結果：What is the name and departure room number?

結果として “departure” が残り，湧き出し語となってしまう．この原因は,「時間」の対応箇所が “time” と判断され “departure” が含まれなかったこと，つまり単語アラインメントの誤りにある．ここでは用例翻訳の一方式の例を使って説明したが，単語アラインメントの誤りに起因する湧き出し語は他の用例翻訳方式でも発生する．

2.2 統計翻訳における問題

一方，統計翻訳においては，多くの対訳表現から得られる特徴が統計的に処理されるので，それぞれの対訳表現は，用例翻訳の例のように端的な動作を引き起こすのではなく，翻訳結果に見られるいくつかの傾向を強化する．ここでは単語アラインメント結果から翻訳モデルを学習する際，前節の例で類似用例として示した対訳表現が学習データに存在する場合を考える．この対訳表現において，もし単語アラインメント中で “departure” がいずれの原言語単語にも対応していないとすると，冠詞などと同様に原言語単語との明確な対応がなくても出現する語として “departure” が扱われる確率が大きくなる．そのため，“departure” が誤って訳文に挿入されてしまう可能性が出てくる．逆に「時間」の対応箇所が “departure time” であると判断された場合,「時間」と “departure” との対応確率が大きくなる．そのため “departure” とは関連のない「時間」を含んだ入力文に対する訳文に “departure” が誤って出現する可能性が出てくる．もちろん翻訳モデルは，他の多くの対訳表現から生成されるのでこれらの誤りは淘汰されることが期待される．しかし統計翻訳には，モデルではとらえ難い現象がある（モデルの問題），高度なモデルでは最適なパラメータの発見が保証されない（学習の問題），翻訳時の探索空間が大きく最適な訳文を得るのが難しい（探索の問題）といった問題が存在する[5)]．このため翻訳結果に誤りが残ってしまうことがありうる．統計翻訳における湧き出し語の具体例は，5.4 節において実験結果を考察する際に改めて取り上げる．

3. 問題へのアプローチ

3.1 後編集による修正

湧き出し語問題への対処として，個々の翻訳システム自体の改良，特にその利用する単語アラインメントの改良が考えられる．しかし本稿において我々は，機械翻訳結果を後編集により自動修正するアプローチを取り上げる．

後編集アプローチには，特定の翻訳システムに限らず複数のシステムに適用可能という利点がある．複数の翻訳システムの利用例としてセレクタ・アーキテク

☆ IWSLT-2004 で使われた，学習用，開発用，テスト用のコーパス，参照訳，各システムの翻訳結果のすべてが GSK（言語資源協会，http://www.gsk.or.jp/）を通じて一般に公開される．

チャ[6] があげられる．セレクタ・アーキテクチャでは，複数の翻訳システムの出力した複数の訳文を統計的に評価し，スコアが最良の訳を選択する．これにより，性質の異なる翻訳システムが互いに補い合い，より多くの原文に対して良い訳文を得ることが可能となる．この枠組みにおいては，個々の翻訳システムは固定ではなく他のシステムへの差し替えも考えられ，特定の翻訳システムに対象を限定しない後編集アプローチが有効となる．

3.2 関連研究

湧き出し語問題および後編集アプローチに関連する先行研究として，文献 7) の翻訳自動校正，文献 8) の翻訳文の誤り検出，および，文献 9) の統計翻訳のグリーディ・デコーディングの各研究があげられる．

文献 7) では，我々と同様に後編集により訳文を修正するが，入力文の内容をいかに正確に伝えるかという問題ではなく，いかに自然な文を生成するかという問題に重点を置く．その提案手法では，翻訳結果とそれを人手で校正した結果を使って校正規則を学習し，その規則に従って翻訳文を修正する．それに対して我々は，入力文と翻訳結果の対応関係に基づいて翻訳文を修正する．つまり入力文の内容を正確に伝える問題に重点を置く．両者のアプローチは互いに補完関係にあると考えられる．

文献 8) では，単語レベルでの翻訳文の誤り検出を行う．この研究では統計翻訳システムを対象とし，翻訳結果中の各単語について，いくつかの指標を用いて正誤判定を行う．この手法では対訳コーパスと対象翻訳システムを使った学習を行う．すなわち，対訳コーパスの原文を翻訳し N ベスト訳を出し，対訳コーパスの参照訳と比べることにより単語の正誤を判定し，この判定を基準として各指標値に対する単語の正誤分類をナイーブベイズ法で学習する．実験では N ベストの N を 1,000 として学習および評価を行っている．実験結果として，一番精度の良かった指標は，N ベスト順位で重み付けした単語の出現頻度（ランク重み付け頻度）であり，僅差で IBM モデル 1[10] の単語翻訳モデルに基づく指標が続く．

ランク重み付け頻度を使うには，翻訳システムの出す N ベスト訳が必要である．ランク重み付け頻度の出す信頼度は，翻訳システムの判断，つまり翻訳システムの作成した訳とその順位に依存する．一方，単語翻訳モデルは，特定の翻訳システムに依存せず，それを利用するためには，正誤判定実行時の N ベスト訳や N ベスト訳を使った学習は必須ではない．我々は，N ベスト訳や学習を利用せず，単語翻訳モデルを使った誤り語検出処理とその拡張を提案する．

文献 9) では，最適な翻訳文を探索する統計翻訳手法が提案されている．この手法は，翻訳結果の後編集による修正ではないが，より良い翻訳文を探索するために訳文候補に変形操作を加える．この手法では，5 種類の変形操作を定義し，適当に与えられた翻訳文の元となる単語列に対して，評価関数値を良くする操作がある限り，それを繰り返し適用する．ここで定義された変形操作はより一般的なものであるのに対し，本稿の提案の要点は，単語翻訳モデルによって検出される誤りに修正箇所を絞ることにある．また提案手法は，翻訳システムが結果として出力した文，つまりシステムが最良と判断した翻訳文を対象とする．

4. 訳語削除手法

機械翻訳の後編集処理として湧き出し語を削除する手法を提案する．本手法では，まず単語翻訳モデルによって削除語候補を検出し，次に対訳用例を使って候補を絞り込み，残った候補を削除する．

4.1 単語翻訳モデル

湧き出し語は，入力文中の単語との対応確率の小さな訳語と考えるのは自然である．この確率計算のために，提案手法では，統計翻訳モデルの原始要素である単語対応確率を利用する．

翻訳モデルは，ある言語の文 F が与えられたとき，別の言語のある文 E に翻訳される確率 $P(E|F)$ を与えるモデルである．統計翻訳でよく使われる翻訳モデルは IBM モデル[10] であり，IBM モデル 1 から 5 までのバリエーションがある．IBM モデル 1 は，原言語と目的言語の単語間の翻訳確率を示す単語翻訳モデルのみから構成される．このモデルは対訳コーパス中の対訳文における両言語の単語の共起関係から求められる．一方 IBM モデル 2 から 5 は，原言語文のある位置に現れた語が他言語の文の特定の位置に対応する確率，原言語のある単語が他言語の特定個数の単語に翻訳される確率なども考慮して組み合わせたモデルである．

今回の問題となる湧き出し語は，原文でまったく触れてもいない概念を表す語が訳文に現れることにより奇異な印象を与える問題である．我々は，湧き出し語は原文中の単語の集まりから想起されない訳語であり，単語の出現位置や対応語数に関係ないととらえ，単純な IBM モデル 1 の単語翻訳モデルを利用する．

また統計翻訳では，与えられた原文 F に対して翻訳確率 $P(E|F)$ を最大にする訳文 E を求める問題を，ベイズの法則を利用して $P(F|E)P(E)$ を最大にする

E を求める問題に置き換えることがよく行われる．ここで $P(F|E)$ はもとの翻訳方向とは逆方向の翻訳確率であり，$P(E)$ は言語モデル確率である．この変換により言語モデルを利用することが可能となる．しかし湧き出し語問題は 2 言語間の問題であり，我々は，訳文単独で評価する言語モデルは使わず，順方向の翻訳モデル $P(E|F)$ を用いる．

4.2 単語翻訳モデルによる削除語候補検出

対訳コーパスから学習した IBM モデル 1 の単語翻訳確率を利用し，翻訳文中に現れた各単語 e について，入力文と単語翻訳確率 p から見た出現数の期待値 $C(e)$ を求める．

$$C(e) = \sum_{j=0}^{m} p(e|f_j)$$

ここで，入力文は $f_1, \ldots, f_m$ の単語列であり，f_0 は NULL 単語を意味する．つまり $p(e|f_0)$ は，訳語 e が原文のいずれの語とも関連なく訳文に現れる確率を示す．$C(e)$ が十分小さな値，つまり次式のようにある定数 δ 以下となる訳語 e を削除語候補とする．

$$C(e) \leq \delta$$

この枠組みにおいて NULL 単語を使うことは，原言語の単語と明確に対応しない機能語などの訳語にある程度の大きさの確率を与え，その削除を防ぐことにつながる．

単語翻訳モデルは 2 言語間の単語の対応関係を確率として与える．一方，人手で編集された対訳辞書は信頼度の高い対応関係を与えると考えられる．今日では多くの対訳辞書が電子的に利用可能となっている．人手で編集された対訳辞書を利用する場合，入力文中の原言語単語に関して辞書に示された訳語は，原言語単語からの高い翻訳確率を持つと見なして削除の対象としないこととする．

4.3 対訳用例による削除語の制限

単語対応確率が大きくなくても，入力文全体に対応して正しいと考えられる訳語もありうる．この場合，単語対応確率のみで判断すると誤り語として削除される危険がある．これを防ぐために類似用例を利用する．

以下，例で示す．

入力文：もう一度名前を探してください

翻訳結果：Could you check my name again please?

これは良い翻訳だが,「探す」をはじめとする入力文中の語と “check” の翻訳確率は必ずしも高くない．たとえば「探す」は “look for” や “find” などの語句に対応する場合が多く “check” への翻訳確率は大きくない．このため “check” が削除語候補となる．しかし，次の対訳用例が対訳コーパスに存在することにより “check” が正しい訳語だと判断され，削除語候補から外される．

原文：もう一度探して下さい

訳文：Could you check again?

ここで改めて，用例を利用して削除語を制限する判断基準を定義する．まず対訳コーパス中の対訳文を用例とし，入力文と編集距離の近い原文を持つ用例を類似用例とする．2 文間の編集距離 $dist(s_1, s_2)$ は，次のように定義する．ここで，s_1 と s_2 は対象となる 2 つの文であり，$|s_1|$ と $|s_2|$ はそれぞれの単語数，I と D は，それぞれ 2 単語列の比較における挿入語数と削除語数を示す．置換は挿入と削除の組合せに相当し，式には現れない．

$$dist(s_1, s_2) = \frac{I + D}{|s_1| + |s_2|}$$

この距離定義に基づくコーパスからの類似用例検索は，クラスタリング[11] や単語グラフ上の A* アルゴリズム[2] の手法によって効率的に行うことが可能である．

また注目する訳語 e と各類似用例について，R を入力文に現れる用例原文の単語のリスト，D を入力文に現れない用例原文の単語のリストとする．先の例では，R は { もう, 一度, 探し, て下さい }，D は空リストとなる．ここで入力文を “名前を探してください” とすると，R は { 探し, て下さい }，D は { もう, 一度 } となる．

以上の準備のもと，次の 3 条件をすべて満たす類似用例が存在する単語 e は削除語とはしない，と判断する．

- 用例訳文が e を含む．
- 対訳辞書を利用する場合，辞書に示された D 中の語の訳語に e が現れない．
- 次の式を満たす．

$$\sum_{f \in D} p(e|f) \leq \sum_{f \in R} p(e|f)$$

これらの条件の意図するところは，用例原文のうち入力文に使われている部分に対応する訳語は削除しない，ということである．

4.4 削除の実施

削除語候補となった訳語のうち，用例による制限にかからなかった語を削除語とし，訳文から削除する．

5. 実　　験

複数の翻訳システムの翻訳結果に対して提案手法を

表 1 コーパスサイズ：supplied トラック用
Table 1 Statistics of the training corpus: supplied tracks.

	日英		中英	
	日本語	英語	中国語	英語
文数	20,000		20,000	
総語数	209,012	188,712	182,904	188,935
平均文長	10.45	9.44	9.15	9.45
語彙数	9,277	8,074	7,643	8,191

表 2 コーパスサイズ：日英 unrestricted トラック用
Table 2 Statistics of the training corpus: JE unrestricted track.

	日本語	英語
文数	224,535	
総語数	1,865,298	1,589,983
平均文長	8.31	7.08
語彙数	21,686	14,548

適用し，その有効性を確認した．

5.1 条　　件

実験では，IWSLT-2004 評価キャンペーンにおける言語資源と参加した翻訳システムの翻訳結果を利用した．当キャンペーンは，旅行会話に関するコーパス BTEC（Basic Travel Expression Corpus）[12] を用い，参加システムの翻訳結果を評価するものである．本稿の実験では，日英翻訳の supplied トラック，中英翻訳の supplied トラック，日英翻訳の unrestricted トラックを対象とした．supplied トラックでは，翻訳対象言語対に関する 2 万文の原文とその訳語からなる対訳コーパス（**表 1**）を学習セットとして与えられ，参加システムは，それ以外の言語資源を使うことは許されない．unrestricted トラックでは言語資源の制限はなく，追加の対訳コーパス，辞書，構文解析知識などの使用が許される．したがって unrestricted トラックの参加システムは supplied トラックの参加システムよりも高品質であることが期待できる．各トラックとも 500 文からなるテストセットを翻訳する．テストセットの平均文長は，日英では 8.7 語/文，中英では 7.6 語/文である．

本稿の実験では，参加各システムの翻訳結果に訳語削除手法を適用した．supplied トラックを使った実験では，キャンペーンで与えられた 2 万対訳の学習セットを使って，削除語検出に用いる IBM モデル 1 を構築した．日英翻訳の unrestricted トラックについては，当トラックに参加した ATR-H システム[13] の使ったコーパスを用いて IBM モデル 1 を構築し，また同システムの使った日英対訳辞書も削除語検出で利用した．このコーパスの対訳数は約 22 万（**表 2**），辞書の見出し語数は約 9 万である．

IBM モデル 1 の構築には GIZA++ [14] を用いた．削除語候補検出処理においては $C(e)$ が 0.01 以下となる訳語 e を候補とした．用例による削除語制限処理においても，削除語候補検出処理と同じコーパスを使った．ここでは入力文に対する距離が 0.4 以下の原文を持つ用例を類似用例とした．類似用例が多数の場合，距離の小さい方から 100 位までの用例を使った．

5.2 翻訳品質への効果

各システムの翻訳結果に訳語削除手法を適用し，適用前後の翻訳自動評価値を比較した．翻訳自動評価指標として多重参照単語誤り率[15]（multi-reference Word Error Rate，以下，mWER と記す）と BLEU スコア[16] を使用した．mWER は翻訳結果が参照訳と異なる割合を編集距離に基づいて計算する．BLEU は翻訳結果と参照訳の N グラムの一致する割合を示す．翻訳結果の品質が高くなれば mWER は小さくなり，BLEU スコアは大きくなると考えられている．mWER，BLEU ともに各テスト文につき 16 文の正解翻訳例を参照訳とした．

日英 supplied トラックの評価結果を**表 3** に，中英 supplied トラックを**表 4**，日英 unrestricted トラックを**表 5** に，それぞれ示す．表中，各システムの左肩に付いている記号は翻訳方式を示す．s は統計翻訳，e は用例翻訳，h は統計翻訳と用例翻訳のハイブリッド，r はルールベース翻訳である．表は各システムについて，訳語削除を行っていないベースとなる翻訳文と削除後の翻訳文の自動評価値を示す．削除に関しては 2 種類の結果を示している．「削除」では，対訳用例による制限を行わず，削除語候補検出処理により候補となった語をすべて削除している．「削除 (制限付)」では対訳用例により削除語候補を制限している．また削除率の欄は，翻訳語全体に対する削除された語の割合，訳文数 500 に対する 1 語でも削除された文の割合を示す．

結果として，日英 supplied トラックの ATR-S システムの BLEU 値を除く全指標で，評価値の差の大小はあるにせよ，削除語制限を行った場合と行わなかった場合のいずれでも，削除後の翻訳文の方がより良い評価値を得ている．この結果は，単語対応モデルを基準に翻訳誤りを修正するアプローチの有効性を示している．

一方，削除語制限を行った場合と行わなかった場合とでは，自動評価値の差は小さく，いずれの場合により良い値を示すかを言うことはできない．これは削除率の差が小さいためと考えられる．特に supplied トラックでは，unrestricted トラックに比べ削除率の変

表 3 翻訳品質：日英 supplied トラック
Table 3 Translation quality: JE supplied track.

システム		削除率 (%) 語	削除率 (%) 文	mWER	BLEU
^{s}RWTH	ベース	-	-	0.4196	0.4515
	削除	2.4	14.2	0.4155	0.4566
	削除 (制限付)	2.4	14.2	0.4155	0.4566
^{s}ISI	ベース	-	-	0.4844	0.4008
	削除	1.3	6.4	0.4791	0.4061
	削除 (制限付)	1.3	6.2	0.4791	0.4064
^{s}IBM	ベース	-	-	0.5289	0.3649
	削除	4.3	18.8	0.5171	0.3852
	削除 (制限付)	4.2	18.6	0.5168	0.3858
^{s}ATR-S	ベース	-	-	0.6145	0.3645
	削除	15.4	47.6	0.6047	0.3625
	削除 (制限付)	15.4	47.6	0.6047	0.3625

表 4 翻訳品質：中英 supplied トラック
Table 4 Translation quality: CE supplied track.

システム		削除率 (%) 語	削除率 (%) 文	mWER	BLEU
^{s}RWTH	ベース	-	-	0.4548	0.4093
	削除	2.4	12.0	0.4527	0.4139
	削除 (制限付)	2.4	11.8	0.4527	0.4138
^{s}ATR-S	ベース	-	-	0.4702	0.4535
	削除	8.4	35.6	0.4599	0.4934
	削除 (制限付)	8.3	35.6	0.4595	0.4936
^{s}ISL-S	ベース	-	-	0.4716	0.4152
	削除	3.7	20.4	0.4630	0.4288
	削除 (制限付)	3.7	20.4	0.4630	0.4288
^{s}ISI	ベース	-	-	0.4872	0.3754
	削除	3.0	12.2	0.4862	0.3819
	削除 (制限付)	3.0	12.0	0.4860	0.3818
^{s}IRST	ベース	-	-	0.5083	0.3489
	削除	12.3	46.0	0.4992	0.3984
	削除 (制限付)	12.3	46.0	0.4992	0.3984
^{h}IAI	ベース	-	-	0.5330	0.3382
	削除	7.6	39.0	0.5078	0.3602
	削除 (制限付)	7.6	39.0	0.5078	0.3602
^{s}IBM	ベース	-	-	0.5391	0.3465
	削除	4.3	16.8	0.5334	0.3567
	削除 (制限付)	4.2	16.6	0.5342	0.3561
^{s}TALP	ベース	-	-	0.5564	0.2786
	削除	5.0	23.0	0.5486	0.2877
	削除 (制限付)	4.8	22.4	0.5493	0.2875
^{e}HIT	ベース	-	-	0.6172	0.2089
	削除	12.2	42.2	0.6036	0.2361
	削除 (制限付)	12.0	41.4	0.6043	0.2361

化は小さい．これは，supplied トラックでは小さなコーパスを使うため，削除語制限条件に合う類似用例が少ないことによる．つまり，この実験では削除語制限の有効性は示されていない．

5.3 翻訳妥当性との関連

訳語削除が有効に機能している場合，何らかの訳語削除がなされたベースとなる訳文は妥当な訳ではないことが期待される．ここでは訳語削除の有効性を，翻訳妥当性に関する主観評価結果を使って検証する．

IWSLT-2004 評価キャンペーンにおいて，各翻訳システムの翻訳結果の妥当性について主観評価がなされている．それぞれの訳文について，原言語を理解できる目的言語のネイティブ 3 名により 5 段階の評価が与えられている．評価は 5 から 1 までの数値で与えら

表 5 翻訳品質：日英 unrestricted トラック
Table 5 Translation quality: JE unrestricted track.

システム		削除率 (%) 語	削除率 (%) 文	mWER	BLEU
^{h}ATR-H	ベース	-	-	0.2631	0.6306
	削除	2.3	10.6	0.2608	0.6490
	削除 (制限付)	2.0	8.4	0.2596	0.6481
^{s}RWTH	ベース	-	-	0.3064	0.6180
	削除	2.4	12.8	0.2993	0.6308
	削除 (制限付)	2.1	11.4	0.3009	0.6313
^{e}UTokyo	ベース	-	-	0.4852	0.3963
	削除	9.4	33.4	0.4628	0.4484
	削除 (制限付)	8.4	29.2	0.4629	0.4464
^{r}CLIPS	ベース	-	-	0.7304	0.1320
	削除	12.4	53.6	0.6991	0.1529
	削除 (制限付)	12.1	52.6	0.6997	0.1524

表 6 翻訳妥当性：日英 supplied トラック
Table 6 Translation adequacy: JE supplied track.

システム		文数	翻訳妥当性ランク別文数 5	4	3	2	1	翻訳妥当性ランクの平均
^{s}RWTH	ベース	500	181	90	68	76	85	3.412
	削除	71	8	8	12	19	24	2.394
	削除 (制限付)	71	8	8	12	19	24	2.394
^{s}ISI	ベース	499	157	74	53	88	127	3.092
	削除	32	6	0	2	11	13	2.219
	削除 (制限付)	31	5	0	2	11	13	2.129
^{s}IBM	ベース	499	130	77	71	103	118	2.996
	削除	94	5	10	18	26	35	2.191
	削除 (制限付)	93	4	10	18	26	35	2.191
^{s}ATR-S	ベース	473	66	29	34	79	265	2.053
	削除	238	9	4	13	40	172	1.479
	削除 (制限付)	238	9	4	13	40	172	1.479

れ，大きい数字ほど妥当な訳となる．各訳文について 3 名の付けた評価の中間値が，その訳文の翻訳妥当性ランクとされ，テストセット 500 文の平均値が各システムの翻訳妥当性スコアとされる．

表 6，**表 7**，**表 8** は，それぞれ日英 supplied トラック，中英 supplied トラック，日英 unrestricted トラックに関して，訳語削除の行われた文と翻訳妥当性の関係を示す．各表のベースの欄は各システムについて，空出力を除く翻訳結果のあった文数，翻訳妥当性ランクごとの文数，および，翻訳妥当性ランクの平均値を示す．削除の欄は，1 語でも訳語削除がなされた文についてのみ，ベースから翻訳妥当性ランクを抽出したものである．つまり，いずれの欄でも，個々の文の翻訳妥当性ランクとしては，削除を行わない元の訳文のランクを使い，欄ごとに対象となる訳文の集合が異なっている．削除に関しては対訳用例による制限の有無により 2 種類の結果を示している．

結果として，削除のあった訳文での翻訳妥当性ランクの平均値は，ベースに比べ，削除語制限を行った場合と行わなかった場合のいずれでも，明確に悪い値になっている．その差は，日英 unrestricted トラックの ATR-H システムなど，ベースの平均ランクの良いシステムで特に大きくなっている．このことから削除の有無を元の翻訳妥当性の信頼度と見なすことができる．場合によっては，削除のある訳文は妥当な訳ではないとして棄却するという提案手法の利用法も考えられる．この利用法は，3.1 節で触れたセレクタ・アーキテクチャにおいて，削除のある訳と削除のない訳が混在している場合に有望である．

一方，削除語制限を行った場合は，行わなかった場合に比べ，翻訳妥当性ランクの平均値は確実に悪くなっている．supplied トラックでは，削除語制限の有無による削除語数の差異は小さく，まったく差のない場合もある．しかし差のある場合は削除語制限を行った場合の方が悪くなっている．つまり削除語制限を使うことにより，より妥当性の低い翻訳文を絞り込むこ

表 7 翻訳妥当性：中英 supplied トラック
Table 7 Translation adequacy: CE supplied track.

システム		文数	翻訳妥当性ランク別文数					翻訳妥当性
			5	4	3	2	1	ランクの平均
[s]RWTH	ベース	500	174	86	66	83	91	3.338
	削除	60	12	5	9	14	20	2.583
	削除 (制限付)	59	11	5	9	14	20	2.542
[s]ATR-S	ベース	495	143	69	54	93	136	2.980
	削除	178	21	19	26	38	74	2.298
	削除 (制限付)	178	21	19	26	38	74	2.298
[s]ISL-S	ベース	496	149	79	50	95	123	3.073
	削除	102	14	14	11	26	37	2.431
	削除 (制限付)	102	14	14	11	26	37	2.431
[s]ISI	ベース	499	142	89	64	80	124	3.090
	削除	61	7	10	11	13	20	2.525
	削除 (制限付)	60	7	9	11	13	20	2.500
[s]IRST	ベース	500	142	87	64	87	120	3.088
	削除	230	30	34	32	52	82	2.470
	削除 (制限付)	230	30	34	32	52	82	2.470
[h]IAI	ベース	500	127	72	75	95	131	2.938
	削除	195	20	22	31	44	78	2.292
	削除 (制限付)	195	20	22	31	44	78	2.292
[s]IBM	ベース	499	116	84	69	100	130	2.912
	削除	84	10	15	15	18	26	2.583
	削除 (制限付)	83	9	15	15	18	26	2.554
[s]TALP	ベース	495	123	89	80	97	106	3.053
	削除	115	12	12	23	29	39	2.383
	削除 (制限付)	112	10	11	23	29	39	2.224
[e]HIT	ベース	500	104	110	89	104	93	3.056
	削除	211	19	41	43	57	51	2.621
	削除 (制限付)	207	17	41	42	57	50	2.604

表 8 翻訳妥当性：日英 unrestricted トラック
Table 8 Translation adequacy: JE unrestricted track.

システム		文数	翻訳妥当性ランク別文数					翻訳妥当性
			5	4	3	2	1	ランクの平均
[h]ATR-H	ベース	500	310	85	41	27	37	4.208
	削除	53	9	12	7	9	16	2.792
	削除 (制限付)	42	1	11	6	8	16	2.357
[s]RWTH	ベース	500	290	81	45	40	44	4.066
	削除	64	12	9	10	15	18	2.719
	削除 (制限付)	57	8	8	9	15	17	2.561
[e]UTokyo	ベース	500	183	76	60	78	103	3.316
	削除	167	25	26	15	35	66	2.455
	削除 (制限付)	146	15	22	13	34	62	2.274
[r]CLIPS	ベース	500	58	70	113	133	126	2.602
	削除	268	12	18	61	81	96	2.138
	削除 (制限付)	263	11	17	60	81	94	2.125

とができている．

5.4 統計翻訳への効果

当手法は IBM モデル 1 に基づく単語翻訳モデルを利用して誤り語を検出する．統計翻訳システムでは同様の翻訳モデルが利用されると考えられるが，その翻訳結果に対しても効果が確認された．つまり評価対象の大半は統計翻訳システムであったが，それらのシステムにおいて翻訳品質への効果，翻訳妥当性との関連が認められた．

最近の統計翻訳では，句レベルの翻訳モデルの有効性が認められ，それを利用する方式が主流となっている．今回対象とした統計翻訳システムのすべて[13),17)~22)]が句レベルの翻訳情報を扱い，そのうち 1 つを除くシステムが句翻訳モデルを使っている．また多くのシス

表 9 湧き出し語を含んだ訳文に対する訳語削除の実行例
Table 9 Examples of word deletion for translations that contain out-of-the-blue words.

(1) ボストン美術館に行くつもりです
→ I'm going to visit boston museum [after] [leaving] the [U.S.]
(2) お名前とお部屋番号を教えて下さい
→ What is the name and [departure] room number?
(3) ワインの赤はありますか → Do you have wine [free] red?
(4) あそこの出口を出てすぐです → Take that exit over there and [turn] right please.
(5) あちらの出口を出て下さい → Could you [tell] me [where] the exit over there?
(6) 目の前です → It's in front of the [station].
(7) 駅は二階にあります → Does the station is on the second floor?
(8) 駅は二階にあります → Station to your [room] is on the second floor.
(9) 食べたいですか → Do you want to eat [ice] [cream]?
(10) 野菜の料理を食べたいのですが → I'd like to have some [local] food.
(11) 野菜料理を食べたいのですが → I'd like to eat some [Chinese] food.
(12) 彼女はベジタリアンです → She is a vegetarian meal.
(13) 日本が韓国に負けています
→ There are some [similarities] [between] Japan and Korea.
(14) エアコンが煩いです → There's [no] air conditioning.
(15) エアコンがうるさいです → The air conditioner is very noisy.
(16) 明日ポーランドに行きます → I'll go to Portland tomorrow.
(17) 寝違えてしまいました → I [missed] my [station].
(18) 食中毒ではありませんか → Is this the [Washington] [library]?
(19) アトピーの症状が悪化してしまいました → I've got a [flat] [tire].
(20) あごが外れてしまいました → The [battery] is [dead].

テムでは，IBM モデル 1 に基づく単語翻訳モデルを，学習の初期データとして，あるいは訳文の適切性を示す指標の 1 つとして利用している．これらのシステムに対して提案手法が効果を持つことは，高度なモデルを利用する統計翻訳においても，単語翻訳モデルのさらなる有効利用法があることを示している．より具体的には，ある単語を訳文中で使用するかどうかの判断において，単語翻訳モデルに基づく確率値に関する条件を，必要条件として使うことの有効性を示唆する．

以下，実験結果から湧き出し語の削除例を示す．これは日英 supplied トラックの IBM システムの翻訳結果で見られた例である．このシステムは句翻訳モデルを使い，また IBM モデル 1 による単語翻訳モデルを学習の初期データとして利用している．

入力文：航空券を家に忘れてしまいました
翻訳結果：i have a [return] ticket i left my [glasses] in the house☆

この翻訳結果にはいくつかの問題があるが，不要と思われる語句に下線を引いている．また提案手法によって削除された語を [ ] で囲んでいる．特に入力文から見て唐突な湧き出し語と見られる “return” と “glasses” が提案手法によって削除されている．

☆ 英語の正書法に従わず大文字や句読点を使っていないのは，訳文の評価が lower-case only, no punctuation marks という条件下で行われたためである．

学習セットにおいて，上記の入力文中の語と湧き出し語が共起するいくつかの対訳が観察された．そのうち 2 つの対訳を下に示す．

原文 1：今朝十時にボストンを出た電車にメガネを忘れました
訳文 1：i left my glasses on the train that left boston at ten this morning
原文 2：帰りの航空券を見せてください
訳文 2：your return ticket please

それぞれの対訳において網がけした部分の間でなんらかの対応がとられ，それが誤りの発生に影響を与えたと推測することができる．単語翻訳モデルでは，入力文中の各語と “return”，“glasses” との対応確率は十分小さくなり，これらの語を誤りであると判断することができた．

5.5 削 除 例

いくつかの日英翻訳実験システム[2),6),23)~25)] の使用者から湧き出し語を含んだ翻訳結果として報告された例について，訳語削除処理を適用した結果を**表 9** に示す．湧き出し語と他の誤りとで区別し難い部分もあるが，下線を引いた部分が入力文と関連のない余分な語句と考えることができる．これらの翻訳文に対する訳語削除処理には 5.1 節の日英 unrestricted トラックに対する設定を使った．[ ] で囲んだ部分が実際に削除

された語である．これらの例からは，削除による誤り訂正が有効に働いていることが見てとれる．

一方，表 9 の例からは今後の課題も浮かび上がる．たとえば (1) や (6) では，削除語に付随する不要な “the” が消されずに残ってしまう．この問題への対処として，浅い構文解析[26] などの手法により句を括り出し，削除語に付随する語もともに削除するなどの削除範囲を調整する方法が考えられる．

また，たとえば (19) では，訳文を構成する主要な語が削除され，結果として意味のない文になっている．このような訳を適切な訳になるように修正するには，湧き出し語の削除だけでなく，適切な語への置換や訳文に不足する語の挿入などの操作が必要になる．一方で，訳語削除の結果として意味のない文となるような翻訳結果は 5.3 節で示唆した削除語のある訳文は棄却するというアプローチが有効な例といえる．

5.6 ま と め

翻訳自動評価値および翻訳妥当性ランクに関するデータからは提案訳語削除手法の有効性が示された．一方，訳語削除の実例からは，訳語削除手法が湧き出し語に対する誤り訂正として有効に働いていることがうかがえる．

6. おわりに

コーパスベース翻訳の典型的な誤りである湧き出し語問題に対処するために，後編集により翻訳結果を修正する手法を提案した．本手法では，単語翻訳モデルを利用して入力語との対応度の小さい訳語を削除語候補として検出し，対訳用例を利用して候補を制限した後，訳文から削除する．このように本手法は，単語翻訳モデルを中心に利用し，単語翻訳モデルから得られた削除語候補を基に修正処理を起動する．日英翻訳および中英翻訳において複数の翻訳システムを対象とした実験では，誤り語自動削除による翻訳精度向上効果が確認された．訳語削除例は，湧き出し語に対する誤り訂正が有効に働いていることを示唆している．

謝辞 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) 隅田英一郎，佐々木裕，山本誠一：機械翻訳システム評価法の最前線，情報処理，Vol.46, No.5, pp.552–557 (2005).
2) 土居誉生，隅田英一郎，山本博史：編集距離を使った用例翻訳の高速検索方式と翻訳性能評価，情報処理学会論文誌，Vol.45, No.6, pp.1681–1695 (2004).
3) Akiba, Y., Federico, M., Kando, N., Nakaiwa, H., Paul, M. and Tsujii, J.: Overview of the IWSLT04 Evaluation Campaign, *Proc. IWSLT 2004*, pp.1–12 (2004).
4) Manning, D.C. and Schütze, H.: *Foundations of statistical natural language processing*, The MIT Press, London, England (1999).
5) 渡辺太郎，今村賢治，隅田英一郎，奥乃 博：階層的句アラインメントを用いた統計的機械翻訳，電子情報通信学会論文誌，Vol.J87-D-II, No.4, pp.978–986 (2004).
6) Akiba, Y., Watanabe, T. and Sumita, E.: Using Language and Translation Models to Select the Best among Outputs from Multiple MT Systems, *Proc. COLING 2002*, pp.8–14 (2002).
7) 山本和英：機械翻訳における自動校正と日中翻訳への適用，言語処理学会第 5 回年次大会，pp.21–24 (1999).
8) Blatz, J., Fitzgerald, E., Foster, G., Grandrabur, S., Goutte, C., Kulesza, A., Sanchis, A. and Ueffing, N.: Confidence Estimation for Machine Translation, *Proc. COLING 2004*, pp.315–321 (2004).
9) Germann, U., Jahr, M., Knight, K., Marcu, D. and Yamada, K.: Fast Decoding and Optimal Decoding for Machine Translation, *Proc. ACL 2001*, pp.228–235 (2001).
10) Brown, P.F., Della Pietra, S.A., Della Pietra, V.J. and Mercer, R.L.: The mathematics of Machine Translation: Parameter estimation, *Computational Linguistics*, Vol.19, No.2, pp.263–312 (1993).
11) Cranias, L., Papageorgiou, H. and Piperidis, S.: Example retrieval from a translation memory, *Natural Language Engineering*, Vol.3, No.4, pp.255–277 (1997).
12) Takezawa, T. and Kikui, G.: Collecting Machine-Translation-Aided Bilingual Dialogues for Corpus-Based Speech Translation, *Proc. EUROSPEECH*, pp.2757–2760 (2003).
13) Sumita, E., Akiba, Y., Doi, T., Finch, A., Imamura, K., Okuma, H., Paul, M., Shimohata, M. and Watawabe, T.: EBMT, SMT, Hybrid and More: ATR Spoken Language Translation System, *Proc. IWSLT 2004*, pp.13–20 (2004).
14) Och, F.J. and Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, Vol.29, No.1, pp.19–51 (2003).
15) Ueffing, N., Och, F. and Ney, H.: Generation of Word Graphs in Statistical Machine Transla-

tion, *Proc. Conf. on Empirical Methods for Natural Language Processing*, pp.156–163 (2002).
16) Papineni, K., Roukos, S., Ward, T. and Zhu, W.: Bleu: A Method for Automatic Evaluation of Machine Translation, *Proc. ACL 2002*, pp.311–318 (2002).
17) Lee, Y. and Roukos, S.: IBM Spoken Language Translation System Evaluation, *Proc. IWSLT 2004*, pp.39–46 (2004).
18) Bertoldi, N., Cattoni, R., Cettolo, M. and Federico, M.: The ITC-irst Statistical Machine Translation System for IWSLT-2004, *Proc. IWSLT 2004*, pp.51–58 (2004).
19) Ettelaie, E., Knight, K., Marcu, D., Munteanu, D.S., Och, F.J., Thayer, I. and Tipu, Q.: The ISI/USC MT System, *Proc. IWSLT 2004*, pp.59–60 (2004).
20) Vogel, S., Hewavitharana, S., Kolss, M. and Waibel, A.: The ISL Statistical Translation System for Spoken Language Translation, *Proc. IWSLT 2004*, pp.65–72 (2004).
21) Bender, O., Zens, R., Matusov, E. and Ney, H.: Alignment Templates: The RWTH SMT System, *Proc. IWSLT 2004*, pp.79–84 (2004).
22) Gispert, A.D. and Marino, J.B.: TALP: Xgram-based Spoken Language Translation System, *Proc. IWSLT 2004*, pp.85–90 (2004).
23) Watanabe, T. and Sumita, E.: Example-based decoding for statistical machine translation, *Proc. MT Summit IX*, pp.410–417 (2003).
24) 今村賢治，大熊英男，渡辺太郎，隅田英一郎：統計翻訳指標を導入した構文トランスファに基づく用例翻訳，情報処理学会研究報告，Vol.2004-NL-162, pp.71–77 (2004).
25) Imamura, K., Okuma, H. and Sumita, E.: Practical Approach to Syntax-based Statistical Machine Translation, *Proc. MT Summit X*, pp.267–274 (2005).
26) Sha, F. and Pereira, F.: Shallow Parsing with Conditional Random Fields, *Proc. HLT-NAACL 2003*, pp.213–220 (2003).

(平成 17 年 10 月 17 日受付)
(平成 18 年 4 月 4 日採録)

土居　誉生（正会員）

1984 年京都大学理学部卒業．同年（株）CSK システムズ入社．2002 年より 2005 年まで ATR 音声言語コミュニケーション研究所に出向．自然言語処理（特に機械翻訳），自動推論システムおよびプログラミング言語の実装方式，ソフトウェアのモデル化手法の研究に従事．技術士（情報工学部門）．人工知能学会会員．

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．現在，ATR 音声言語コミュニケーション研究所室長．情報通信研究機構知識創成コミュニケーション研究センター音声言語グループ研究マネージャ，神戸大学大学院自然科学研究科連携教授兼務．機械翻訳，情報検索，e ラーニングの研究に従事．電子情報通信学会，言語処理学会，日本音響学会，ACL，IEEE 各会員．