社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－FI－80（3）
2005－NL－169（3）
2005／9／29

単語翻訳モデル駆動型の翻訳後編集

土居 誉生　　隅田 英一郎
ATR 音声言語コミュニケーション研究所
{takao.doi,eiichiro.sumita}@atr.jp

コーパスベース翻訳は有望な技術であるが，入力文と関連のない語が訳文に湧き出す問題など，その方式に由来した特徴的な翻訳誤りも観察される．本稿では翻訳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とる．提案手法は，単語翻訳モデルを利用して誤り候補を検出し，修正処理を起動する．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．

Post-edit of Machine Translation Driven by Lexicon Model

Takao Doi and Eiichiro Sumita
ATR Spoken Language Communication Research Laboratories
{takao.doi,eiichiro.sumita}@atr.jp

Although corpus-based machine translation is a promising technology, one error that is 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. We conducted an experiment using several translation systems 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盛んになってきており，その性能も進化している (隅田ら, 2005)．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，原文で全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，例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 (土居ら, 2004)．

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，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．修正には削除・挿入・置換の操作があるが，今回は特に湧き出し語の問題に対処するために，削除に焦点を当てる．提案手法は，統計翻訳で用いられる単語翻訳モデルを中心に利用する．

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．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，その多少に関わらず，共通して目に付くのは湧き出し語の問題である．湧き出し語の多くは単語アライメントの誤りによって発生する．

以下，第1節で上げた例について，その発生過程を説明する．これは用例翻訳方式における例である．

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みなされる．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，つまり単語アライメントの誤りによる．ここでは用例翻訳の一方式の例を使って説明したが，このような誤りは他の用例翻訳や統計翻訳でも発生する．

3 関連研究

我々のアプローチは，機械翻訳結果をより良い訳にするための後編集である．関連研究として，翻訳自動校正に関する先行研究，単語レベルでの翻訳文の誤り検出，および，翻訳文の変形操作によって最適な結果を探索する統計翻訳手法が上げられる．

翻訳自動校正に関する研究 (山本, 1999) では，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校正する．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校正する．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．

機械翻訳結果の誤り語検出に関する研究として (Blatz et al., 2004) がある．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 (Brown et al., 1993) 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と考えられる．我々は，単語翻訳モデルを使った誤り語検出とその修正を試みる．

他方，統計翻訳のグリーディ・デコーディング (Germann et al., 2001) 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，翻訳システムの出力文，つまり計算コストとの兼ね合いはあるにせよ，システムが最良と判断した翻訳文を対象とする．

4 提案手法

湧き出し語の削除を第一課題とし，機械翻訳結果を修正する後編集手法を提案する．誤り語削除処理では，まず単語翻訳モデルによって削除語候補を検出し，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 単語翻訳モデルによる削除語候補検出

湧き出し語は，入力文中の単語との対応確率の小さな語と考えるのは自然である．複雑なモデルを考える前に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．つまり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する．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．$C(e)$ が十分小さな値 (ある定数以下) となる訳語 e を削除語候補とする．

単語翻訳モデルは 2 言語間の単語の対応関係を調べるために用いるが，人手で編集された対訳辞書は信頼性の高い対応関係を示すと考えられる．対訳辞書を利用する場合は，入力文中の単語に関して辞書に示された訳語は削除の対象としないこととする．

4.2 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得る．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?

この翻訳結果は悪くはないが，「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め “check” が削除語候補となる．しかし，次の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．また注目する入力文と訳語 e について，D を入力文に現れない用例原文の単語のリスト，R を入力文に現れる用例原文の単語のリストとする．次の 3 条件を全て満たす単語 e は削除語とはしない．

- 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．

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

この条件の意図するところは，用例に現れた訳語は，原文の中で入力文に含まれない部分に対応する場合のみ削除の対象とするということである．

4.3 挿入・置換修正について

本稿の焦点は削除による翻訳文の修正であるが，修正システムとして完結するには，削除の他に挿入と置換の操作が必要となる．ここでは，置換は削除と挿入の組み合わせにより実現可能として，挿入について考える．

前節までに述べた訳語削除による修正処理は，単語翻訳モデルを使った基準により，入力文と対応しない翻訳文中の語を誤り候補として起動される．同様の考え方を挿入による修正に適用することができる．つまり，単語翻訳モデルを基準に翻訳文と対応しない入力文中の語を検出し，その訳語を挿入語候補として修正処理を起動する．

以下，挿入による訳文修正処理手順の一例を示す．

1. 入力文と翻訳結果を入れ替えて，前節まで説明した手法を適用し，削除語を求める．結果として，入力文中の単語が削除語となる．この語に関する情報が訳文から欠落していると仮定し，欠落語と呼ぶ．

2. 各欠落語について，その訳語を基の翻訳結果に挿入した翻訳文の候補をいくつか生成し，候補の中から一つを選択する．

ここで 2. の候補の生成では，統計翻訳モデルとして繁殖確率も持つ IBM モデル 3-5 のいずれかを使用し，次の組み合わせにより複数の候補を生成する．

- 訳語候補は，辞書の示す訳語，単語対応確率上位となる訳語
- 訳語組み合わせは，繁殖確率上位の繁殖数で，繁殖数個の訳語の順列
- 挿入場所は，元の翻訳結果の単語間で N グラムに基づく連接尤度の低い箇所

候補の選択では (Akiba et al., 2002) で使われている言語モデル確率と翻訳モデル確率との積を評価値として利用する．ここでの翻訳モデル確率の計算にも，多面的な確率を考慮した IBM モデル 3-5 のいずれかを使用する．

5 実験

4.1節で述べた単語翻訳モデルによる削除語候補の検出処理のみを実装し，その候補を全て削除する条件で予備実験を行い，その有効性を確認した．

5.1 条件

実験では，国際ワークショップ IWSLT-2004 (Akiba et al., 2004; 隅田ら, 2005) 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) (Takezawa and Kikui, 2003) 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．各トラックとも 500 文からなるテストセットを翻訳する．

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，適用前後の翻訳自動評価値を比較した．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 (Sumita et al., 2004) 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 20 万，辞書の見出し語数は約 9 万である．

翻訳自動評価指標として次の 2 つを使った．各指標のための 1 文あたりの参照訳数は 16 である．

BLEU: BLEU スコア (Papineni et al., 2002)．値が大きいほど翻訳品質は良い．

mWER: すべての参照訳との Word-error-rate のうち，最も誤り率の低いもの (Ueffing et al., 2002)．値が小さいほど翻訳品質は良い．

5.2 結果

日英 supplied トラックの評価結果を表 1に，日中 supplied トラックを表 2，日英 unrestricted トラックを表 3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．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 の BLEU 値を除く全指標で，評価値の差の大小はあるにせよ，削除後の翻訳文の方がより良い評価値を得ている．この予備実験の結果は，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．

表 1: 日英翻訳 supplied トラック

システム		削除率 (%) 語	文	WER	BLEU
[s]RWTH	ベース	-	-	0.4196	0.4515
	削除	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
[s]IBM	ベース	-	-	0.5289	0.3649
	削除	4.3	18.8	0.5171	0.3852
[s]ATR-S	ベース	-	-	0.6145	0.3645
	削除	15.4	47.6	0.6047	0.3625

表 2: 日中翻訳 supplied トラック

システム		削除率 (%) 語	文	WER	BLEU
[s]RWTH	ベース	-	-	0.4548	0.4093
	削除	2.4	12.0	0.4527	0.4139
[s]ATR-S	ベース	-	-	0.4702	0.4535
	削除	8.4	35.6	0.4599	0.4934
[s]ISL-S	ベース	-	-	0.4716	0.4152
	削除	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
[s]IRST	ベース	-	-	0.5083	0.3489
	削除	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
[s]IBM	ベース	-	-	0.5391	0.3465
	削除	4.3	16.8	0.5334	0.3567
[s]TALP	ベース	-	-	0.5564	0.2786
	削除	5.0	23.0	0.5486	0.2877
[e]HIT	ベース	-	-	0.6172	0.2089
	削除	12.2	42.2	0.6036	0.2361

表 3: 日英翻訳 unrestricted トラック

システム		削除率 (%) 語	文	WER	BLEU
[h]ATR-H	ベース	-	-	0.2631	0.6306
	削除	2.3	10.6	0.2608	0.6490
[s]RWTH	ベース	-	-	0.3064	0.6180
	削除	2.4	12.8	0.2993	0.6308
[e]UTokyo	ベース	-	-	0.4852	0.3963
	削除	9.4	33.4	0.4628	0.4484
[r]CLIPS	ベース	-	-	0.7304	0.1320
	削除	12.4	53.6	0.6991	0.1529

5.3 削除例

次に我々の実験システムで見られた湧き出し語を含んだ翻訳結果に訳語削除処理を適用した例を表4に示す．ただし表中の訳には，湧き出し語以外の誤りのある例や，湧き出し語を含まない例も混ぜてある．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取れる．

6 おわりに

コーパスベース翻訳の誤りに対処するために，後編集により翻訳結果を修正する手法を提案した．提案手法では，特に湧き出し語の問題に対処するために，単語翻訳モデルを中心に利用する．複数の翻訳システムを対象とした訳語削除の実験では，単純に単語翻訳モデルを使うだけでも，翻訳精度の向上効果があることが確認された．

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．

表 4: 湧き出し語を含んだ訳文に対する訳語削除の実行例

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.]
お名前とお部屋番号を教えて下さい
→ What is the name and [departure] room number?
ワインの赤はありますか → Do you have wine [free] red?
あそこの出口を出てすぐです → Take that exit over there and [turn] right please.
あちらの出口を出て下さい → Could you [tell] me [where] the exit over there?
目の前です → It's in front of the [station].
駅は二階にあります → Does the station is on the second floor?
駅は二階にあります → Station to your [room] is on the second floor.
食べたいですか → Do you want to eat [ice] [cream]?
野菜の料理を食べたいのですが → I'd like to have some [local] food.
野菜料理を食べたいのですが → I'd like to eat some [Chinese] food.
肉料理を食べたいのですが → I'd like a meat dish.
彼女はベジタリアンです → She is a vegetarian meal.
日本が韓国に負けています
→ There are some [similarities] [between] Japan and Korea.
エアコンが煩いです → There's [no] air conditioning.
エアコンがうるさいです → The air conditioner is very noisy.

参考文献

Y. Akiba, T. Watanabe, and E. Sumita. 2002. Using language and translation models to select the best among outputs from multiple mt systems. *Proc. of COLING 2002*, pages 8–14.

Y. Akiba, M. Federico, N. Kando, H. Nakaiwa, M. Paul, and J. Tsujii. 2004. Overview of the iwslt04 evaluation campaign. *Proc. of IWSLT 2004*, pages 1–12.

J. Blatz, E. Fitzgerald, G. Foster, S. Grandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2004. Confidence estimation for machine translation. *Proc. of COLING 2004*, pages 315–321.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–312.

土居 誉生, 隅田 英一郎, 山本 博史. 2004. 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, 45(6):1681–1695.

U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada. 2001. Fast decoding and optimal decoding for machine translation. *Proc. of ACL 2001*, pages 228–235.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *Proc. of ACL 2002*, pages 311–318.

E. Sumita, Y. Akiba, T. Doi, A. Finch, K. Imamura, H. Okuma, M. Paul, M. Shimohata, and T. Watawabe. 2004. EBMT, SMT, hybrid and more: ATR spoken language translation system. *Proc. of IWSLT 2004*, pages 13–20.

隅田 英一郎, 佐々木 裕, 山本 誠一. 2005. 機械翻訳システム評価法の最前線. 情報処理, 46(5):552–557.

T. Takezawa and G. Kikui. 2003. Collecting machine-translation-aided bilingual dialogues for corpus-based speech translation. *Proc. of EUROSPEECH*, pages 2757–2760.

N. Ueffing, F.J. Och, and H. Ney. 2002. Generation of word graphs in statistical machine translation. *Proc. of Conf. on Empirical Methods for Natural Language Processing*, pages 156–163.

山本 和英. 1999. 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, pages 21–24.