E-016

統計的機械翻訳における翻訳誤り原因自動同定手法の提案

Automatic Diagnosis Method of Translation Errors in Statistical Machine Translation Systems

林克彦† 山本誠一†
Katsuhiko Hayashi† Seiichi Yamamoto†

1. はじめに

世界のグローバル化によってコンピュータにより異言語間の翻訳を行う機械翻訳技術が重要なものとなってきている．機械翻訳技術には様々な手法が存在するが，コーパス開発の増加に伴い，統計的機械翻訳（Statistical Machine Translation）が活発に研究されている．しかし，現実世界への広範な適用を考える際には，その精度はまだ十分に満足できるものではなく改善の余地は大きい．

そこで本研究では SMT の主要構成要素である言語モデル，翻訳モデル，デコーダの 3 つのうちから翻訳誤りを引き起こした原因を自動同定する手法を提案する．従来，翻訳機のシステム性能は BLEU[1]などに代表されるように翻訳結果と参照訳(リファレンス)との類似度に基づいて翻訳自動評価した結果から比較検証されてきた．しかし，本稿の提案手法は従来の評価手法とは異なり，統計的機械翻訳機の問題を直接的に自動同定し，定量化できるという利点がある．

2. 翻訳誤り原因自動同定手法

統計的機械翻訳ではある原言語の入力文 J を目的言語のテキスト E に翻訳する問題を最大尤度の解を発見する問題としてとらえる．

$$\hat{E} = \arg\max_{E} P(E \mid J)$$

さらにベイズの定理から上式は最大事後確率を求める問題に置き換えられる．

$$\hat{E} = \arg\max_{E} P(E)P(J \mid E)$$

ここで $P(E)$ を言語モデル，$P(J \mid E)$ を翻訳モデルと呼ぶ．また，これらのモデル尤度に基づいて解を探索する過程をデコーダと呼ぶ．言語モデルには一般に *N*-gram モデルが用いられ，目的言語のコーパスから学習される．一方，翻訳モデルの学習には原言語と目的言語のテキストが対となった対訳コーパスが利用され，単語単位の翻訳モデルでは IBM モデル 1～5[2]が一般的である．デコーダでは基本的に訳語を最適な順序に並び替える問題を解くことになる．デコーダの探索手法は様々に存在するが代表的なものとしては初期近似解から仮説を探索するグリーディアルゴリズム[3]や入力文に対して仮説幅を設けて出力文を文頭から生成していくようなビームサーチアルゴリズム[3]などがある．

ここでは統計的機械翻訳の翻訳誤りを引き起こす原因に言語モデル，翻訳モデル，デコーダの 3 つを考える．本稿では誤り原因を自動同定する方法として図 1 のような決定木による手法を提案する．提案手法の基本的な概念は音声認識システムへの適用例として提案されたもの[4]と同一であるが，音声認識とは異なり機械翻訳では解が一意に決定されないため，バイリンガルによって作成された複数の参照文を正解とする．

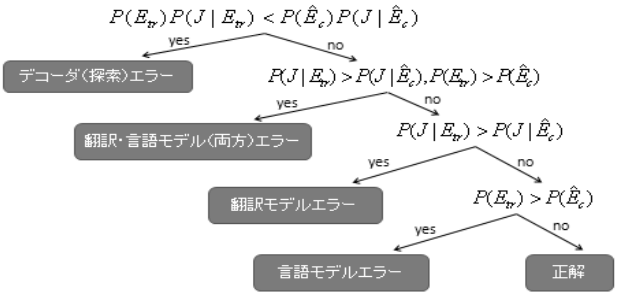

図 1 翻訳誤り原因同定のための決定木

ここで J とは原言語入力文のことであり，ある統計翻訳機によるその翻訳結果を E_{tr} とする．また，J に対する参照訳を S 種類用意し，$E_c^s (1 \le s \le S)$ と定義する．さらに，S 種類の参照訳の中から翻訳モデルと言語モデルの積が最大となる参照訳を $\hat{E}_c$ とする．これより翻訳誤り原因を自動同定する手続きは図 1 の決定木に従う．図 1 の決定木は正解文のモデル尤度の方が翻訳結果よりも良くなるという仮定に基づいて分岐を行っている．

3. 実験

実験に用いた統計的機械翻訳システムは ATR で開発されたグリーディデコーダ方式の統計翻訳機である．また，検証用にそのモジュールを利用して作成したビームサーチデコーダも使用

† 同志社大学大学院工学研究科
Graduate School of Information Engineering, Doshisha University

する．言語モデルは 3-gram，翻訳モデルは IBM モデル 4 で実装した．実験に用いるテストセットは ATR で開発された日英旅行対話コーパス BTEC(Basic Travel Expression Corpus)からランダムに選んだ日本語文 90 文である．また，日英バイリンガルによって作成された英語参照訳を日本語文に対し各 10 種類用意した（合計 900 文）．以下では，これらのテストセットを用いて日英方向の翻訳設定とし，提案手法からの実験結果を提示する．

3.1 グリーディデコーダによる実験結果

翻訳機の実験設定は以下の 3 通りである．

(1) 言語モデル：BTEC英語文16万文の3-gram
探索空間：探索作業が収束するまで（探索大）

(2) 言語モデル：BTEC英語文16万文と英語正解文900文の3-gram
探索空間：探索作業が収束するまで（探索大）

(3) 言語モデル：BTEC英語文16万文と英語正解文900文の3-gram
探索空間：探索作業を途中で打ち切る（探索小）

翻訳モデルは BTEC 日英対訳文 16 万文で学習し，実験中これは変化させない．これらの実験設定による結果を図 2 に示す．

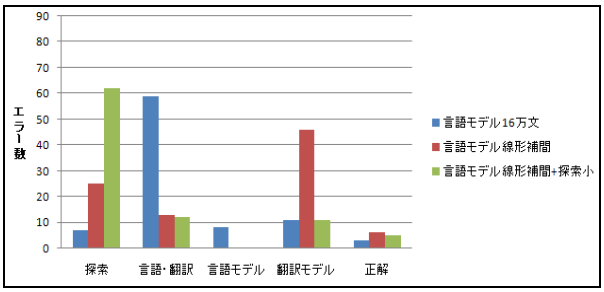

図 2 グリーディデコーダでの誤り原因

図 2 からは言語モデルの改善によって言語モデルに誤りの数が減少していることがわかる．また，探索空間を縮小することでデコーダ（探索）の問題が増加していることがわかる．

3.2 ビームサーチデコーダによる実験結果

ここではビームサーチデコーダによる実験結果を提示する．実験設定は以下の 3 通りで，翻訳モデルは 16 万文から学習したもので変化させない．

(1) 言語モデル：BTEC英語文16万文の3-gram
探索空間：ビーム幅20

(2) 言語モデル：BTEC英語文16万文と英語正解文900文の3-gram
探索空間：ビーム幅20

(3) 言語モデル：BTEC英語文16万文と英語正解文900文の3-gram
探索空間：ビーム幅 100

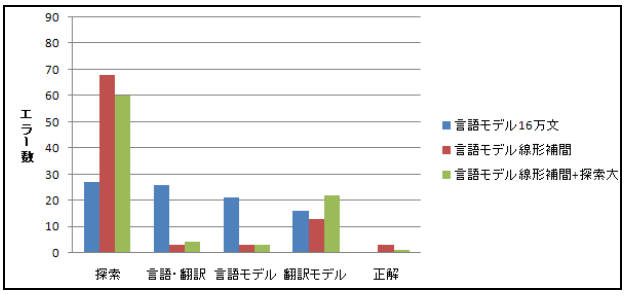

図 3 ビームサーチデコーダでの誤り原因

図 3 では図 2 と同様に言語モデルの改善によって言語モデルのエラー数が減少していることがわかる．また，探索空間を拡大することでデコーダの問題が減少していることもわかる．

4. まとめ

本稿では統計的機械翻訳機の翻訳誤り原因を自動同定する手法を提案した．また，いくつかの実験設定に基づきその有効性を検証した．実験結果からは翻訳機のある問題点を改善すれば，その問題が減少することが示された．しかし，改善後に正解となる翻訳結果が大きく増加することはなく，翻訳機の問題となる箇所が現状で最も問題のある箇所に表れるため，提案手法では翻訳機の性能の優劣を測ることはできないと考えられる．その原因としては正解文の量が充分でないことが考えられる．また，言語モデルが文長によって揺れる問題も分類性能向上にとって大きな問題となる．

今後の課題としては分類性能の向上と翻訳機の性能評価も行えるような改善が必要であると考えられる．改善方法としては正解文の数を増加させることが考えられる．正解文の数を増加させる手法としては換言による方法，翻訳モデル，言語モデルの尤度が向上するように正解文をグリーディ方式で改善する方法等が考えられ，現在それらの手法の組み込みを検討中である．

参考文献

[1] Papinei, K., Roukos, S., Ward, T. and Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine Translation, *Proc. ACL 2002*, pp.311-318 (2002).

[2] Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., Mercer, R.L.: The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, 18, 4, pp. 263–311 (1993).

[3] Germann, U., Jahr, M., Knight, K., Marcu, D., Yamada, K.: Fast Decoding and Optimal Decoding for Machine Traslation, *Proc. ACL 2001*, Toulouse, France (2001).

[4] 南條 浩輝，李 晃伸，河原 達也：大語彙連続音声認識における認識誤り原因の自動同定， 音声言語情報処, Vol.99, No.64, SLP-27-6, Jul 1999