複数の機械翻訳器による学習データの自動生成とそれに基づく統計的前編集

山本 祐司　　南條 浩輝　　吉見 毅彦*

龍谷大学 理工学部 情報メディア学科

e-mail: yyamamoto@nlp.i.ryukoku.ac.jp

1　はじめに

我々は翻訳前編集すなわち，自然な原文から翻訳しやすい文への自動変換を研究している [1]．本稿では複数の機械翻訳器（MT）を用いて，統計翻訳の枠組みに基づく前編集システムを学習する方法について述べる．

2　統計翻訳の枠組みに基づく前編集

統計翻訳の枠組みでは，前編集は，原文Sに対して$P(T|S)$が最大となる機械翻訳しやすい文$\hat{T}$を見つける問題として定式化される．

$$\hat{T} = \underset{T}{\operatorname{argmax}} \ P(T|S) \qquad (1)$$

ベイズの定理を用いて変形を行っていくと式（1）は式（2）に帰着できる．

$$\hat{T} = \underset{T}{\operatorname{argmax}} \ P(S|T)P(T) \qquad (2)$$

本稿では日英対訳コーパスと両方向のMTを用いた$P(S|T)$と$P(T)$の推定用データの自動獲得とそれを用いた統計的前編集について述べる．ここでは，$P(S|T)$を与えるモデルを前編集変換モデル，$P(T)$を与えるモデルを言語モデルとよぶこととする．

2.1　前編集用学習データの自動作成手順

前編集翻訳モデルと言語モデルの学習に必要となる自然な日本語文と翻訳しやすい日本語文のペアの生成方法について記述する．ここでは，日英翻訳を対象として，日英対訳文ペアから翻訳しやすい日本語文を生成する方法について述べる．図1はこの様子を示したものである．前編集用学習データの獲得は，1）翻訳しやすい文（図1のJ_i^k）を生成する．2）自然な日本語文と翻訳しやすい日本語文のペア（図1のJ^kとJ_i^k）から適切なものを選択して前編集用の学習データとするという手順で行う．2.1.1節および2.1.2節で，これらの手順について詳しく述べる．

*Automatic Generation of Training Data for Statistical Pre-editing with Plural Machine Translation Systems by Y.YAMAMOTO, H.NANJO, T.YOSHIMI (Ryukoku University)

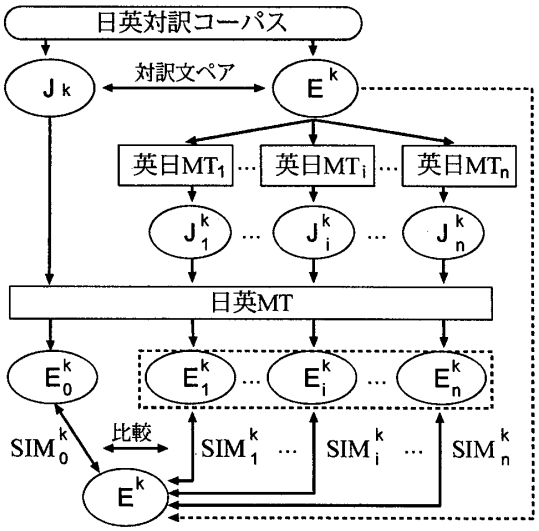

図1: 前編集用学習データの作成手順

2.1.1　翻訳しやすい文の生成

文単位で対応づいた日英対訳コーパスから，自然な日本語と翻訳しやすい日本語が文単位で対応したペアを自動生成する．具体的には以下の手順で行う．

1. 対訳コーパスのk番目（$k = 1 \cdots N$）の日本語文J^kと，その対訳英文E^kのペアを用意する．
2. E^kを複数の翻訳システムMT_i（$i = 1 \cdots n$）で英日翻訳し，日本語文J_i^kを得る．
3. J^kとJ_i^kのペアを得る．

2.1.2　翻訳しやすい文と自然な文とのペア選別

前節で得られたJ^kとJ_i^kのペアの全てが学習データとして適切とは限らない．前編集の目的は，日英翻訳の精度向上であるため，J_i^kの日英翻訳結果E_i^kがもとの自然な日本語J^kの日英翻訳結果E_0^kよりも英語として適切でない場合は，そのような変換を学習すべきではない．すなわち，E_i^kと参照訳E^kとの類似度をSIM_i^kとしたとき，$SIM_i^k \leq SIM_0^k$となるようなiに対してはJ^kをJ_i^kに前編集するべきではなく，このような対応関係を学習データから除く必要がある．具体的には，以下の手順でデータを選択する．

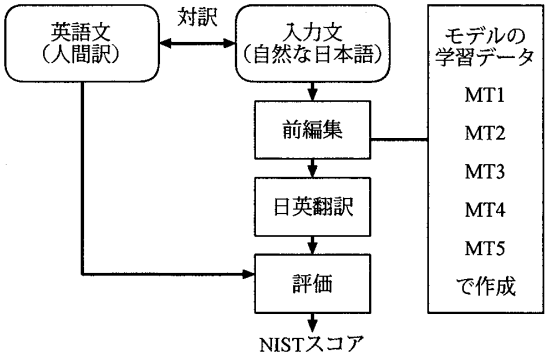

図 2: 日英翻訳における前編集の処理手順

1. J^k を日英翻訳し，英文 E_0^k を得る．
2. J_i^k を日英翻訳し，英文 E_i^k を得る．
3. 参照訳 E^k と E_i^k（$i = 0 \cdots n$）との類似度 SIM_i^k を計算する．
4. $SIM_i^k > SIM_0^k$ を満たす i に対してのみ J^k と J_i^k のペアを学習データとする．すべての i に対して，$SIM_i^k \leq SIM_0^k$ のときは，J^k と J^k のペアを学習データとする．

こうして得られた文のペア集合を前編集変換モデルの学習データとし，自動生成された日本語文の集合を言語モデルの学習データとする．

3 実験と結果

3.1 実験データ

モデル学習用のデータとして，ロイター日英対訳コーパス [2] の 31580 文対を用いた．英日 MT の数を 5 とした．類似度 SIM_i^k には NIST を用いた．2.1.2 節の方法で得られた J^k と J_i^k のペア集合から GIZA++を用いて前編集変換モデルを学習した．言語モデルは，J_i^k の集合から IRST LM Toolkit を用いて単語 5-gram モデルを学習した．評価データにはロイター日英対訳コーパスの 1000 文対を用いた．これは学習データには含まれていない．

3.2 前編集の効果

学習した前編集変換モデルと言語モデルを用いて前編集を行った．式 (2) の日本語列 $\hat{T}$ を求めるために Moses SMT Decoder を用いた．日英翻訳における前編集の効果を表 1 に示す．図 2 にはその処理を示す．

前編集なしの日英翻訳品質 NIST は 3.57 であった．学習データ作成に用いた翻訳器が 1 つの場合は，MT1 と MT4 の場合をのぞき，前編集の効果が平均としてみられなかった．ただし，40%から 50%程度の文に改善がみられていた．5 つ全ての翻訳機を学習データ作成に用いた場合は，NIST は 3.63 であり，NIST の向上がみられた文も 51.3%と効果が最も大きかった．

表 1: 前編集の効果

学習データ作成に用いた翻訳機	日英翻訳品質 (NIST)	NIST の向上/低下がみられた文の割合
前編集なし	3.57	
MT1	3.63	50.5%/47.8%
MT2	3.36	42.0%/56.5%
MT3	3.31	40.3%/57.6%
MT4	3.60	49.3%/48.6%
MT5	3.55	48.1%/51.2%
全ての翻訳機	3.63	51.3%/48.0%

表 2: 英文品質の向上が大きい学習データのみを用いた前編集の効果

閾値	学習データ数* (文対数)	日英翻訳品質 (NIST)
0	144078	3.63
1	125197	3.64
2	98384	3.67
3	71792	3.65

*学習データの作成には MT1~5 全てを使用

3.2.1 英文品質の向上が大きいデータのみでの前編集モデルの学習

前編集学習データを選別する際，E_0^k より少しでも優れていた場合に，学習データとして選択していた．次に，英文品質を明らかに大きく向上できるデータのみを学習データとして用いることを考える．具体的には，2.1.2 の手順 4 の式を以下で置き換える．

$$SIM_i^k > SIM_0^k + \alpha \tag{3}$$

本実験では $\alpha = 1, 2, 3$ を試した，結果を表 2 に示す．英文品質が大きく向上しているデータを用いる効果が確認できる．今回は閾値 2 の場合に最も高い効果が得られた．大きな閾値を用いると学習データが減るため，精度が低下したと考えられる．

4 おわりに

翻訳品質の向上を目的として，自然な原文を翻訳しやすい文に自動変換する方法について述べた．対訳コーパスと両方向の機械翻訳器から前編集システムが自動で作成できることがわかった．

参考文献

[1] 南條浩輝, 吉見毅彦, 岡田真也. 機械翻訳のための統計的手法に基づく前編集. 情報処理学会研究報告, 2009-SLP-76-1, 2009.

[2] M. Utiyama and H. Isahara. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics(ACL)*, pp. 72-79, 2003.