Vol.2009-NL-191 No.1
Vol.2009-SLP-76 No.1
2009/5/21

機械翻訳のための統計的手法に基づく前編集

南條浩輝[†1]　吉見毅彦[†1]　岡田真也[†1]

機械翻訳の前編集について述べる．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（前編集）しようとする研究はこれまでに多く行われているが，それらは主にルールベースのものである．また，変換規則の獲得のために，自然な文と翻訳しやすい文を用意する必要があり，その作成のコストが大きいという問題もある．このような背景に基づき，本研究では統計的手法に基づく前編集を提案する．その際，統計モデルの学習データの自動作成も行う．具体的には，ある文とその対訳文を機械翻訳して得られる文のペアから，自然な文を翻訳しやすい文に変換するための学習データを作成し，統計的機械翻訳の枠組みに基づいて自動変換を行う方法について述べる．

Statistical Pre-edition for Machine Translation

Hiroaki NANJO,[†1] Takehiko YOSHIMI[†1]
and Shinya OKADA [†1]

Pre-edition for machine translation is addressed. Most of machine translation systems do not work well for natural style texts. Therefore, studies of text transformation into literal style, namely pre-edition, have been investigated. Conventional pre-editing studies adopt rule-based editing. Moreover, it is expensive to prepare parallel corpus of natural and literal style texts which is required for pre-editing systems. Based on the background, in this paper, we propose a statistical pre-editing method. Specifically, we show an automatic generation of training data for statistical models from multi-lingual parallel corpus, and then, we describe pre-editing based on statistical machine translation framework.

†1 龍谷大学理工学部
Faculty of Science and Technology, Ryukoku University

1. はじめに

現在の機械翻訳システムでは，自然性が高いテキストに対しては望ましい翻訳結果が得られないのに対して，自然性の低いテキスト，例えば直訳調の文に対してはうまく翻訳できることがある．すなわち，自然な原文を直訳調の文に書き換える（前編集する）ことで翻訳品質の向上が期待できる．このような前編集の効果は，お互いに近いヨーロッパの言語間の翻訳においてよりも，日本語と英語などの大きく異なる言語間の翻訳において大きいと考えられる．また，使用人口が少ない言語などでは研究が十分に行われていないために機械翻訳自体の性能が低く，このような言語の翻訳においても前編集の効果は大きいと考えられる．機械翻訳の品質を向上させるためには，機械翻訳システム自体の性能向上を行うことが本質的であるが，そのためには翻訳対象の言語ペアについての豊富な知識および大量のデータが必要となる．前編集を考えた場合は，翻訳元の言語に関する知識やデータおよび後段の機械翻訳システムの使用方法（入力インターフェイス）を習得していれば十分であり，翻訳品質の向上が比較的得やすいという利点がある．

前編集の研究はこれまでに多く行われているが，それらは主にルールベースのものである．さらに，変換規則の獲得のための自然な文と翻訳しやすい文のペアを用意するコストが大きいという問題もあった．例えば，山口らは前編集前後の文から前編集規則を獲得する方法を提案している[1)]．そこでは人間が実際に前編集を行う必要がある．さらに，その際作業者には，どのように前編集を行えば機械翻訳がより正しく翻訳できるかについての経験や知識が必要とされる．阿辺川らは翻訳経験の浅い翻訳者が翻訳した文である下訳と，ベテランの翻訳者が下訳を修正した後の文である修正訳を対応づけて変換ルールの獲得を目指している[2)]．この学習データの作成コストは大きい上に，これらを使って獲得した変換モデルが，実際の機械翻訳の前編集にとって適切かは保証されない．

このような背景に基づき，本研究では統計的手法に基づく前編集を提案する．その際，統計モデルの学習データの自動作成も行う．具体的には，ある文とその対訳文を機械翻訳して得られる文のペアから，自然な文を翻訳しやすい文に変換するための学習データを作成し，統計的機械翻訳の枠組みに基づいて自動変換を行う方法について述べる．統計的な文体の変換として下岡らの手法[3)]があるが，これは翻訳の前編集を目的としたものではない．また，学習データの自動獲得は行われておらず，本研究とはこれらの点において異なるものである．

Vol.2009-NL-191 No.1
Vol.2009-SLP-76 No.1
2009/5/21

2. 統計的枠組みに基づく前編集

統計的な枠組みでは，翻訳は，ある言語（原言語）の文字列 S が与えられたときに，事後確率 $P(T|S)$ が最大となる他の言語（目的言語）の単語列 T を見つける問題として式（1）で定式化される[4]．

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

ベイズ則を用いると式（1）は式（2）に変形できる．

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

ここで，分母 $P(S)$ は T に無関係であるので省略でき，式 (3) が得られる．

$$\hat{T} = \underset{T}{\operatorname{argmax}}\ P(S|T)P(T) \tag{3}$$

この $P(S|T)$ は異なる言語の文字列の対応スコア，すなわち翻訳スコアであり，これを与えるモデルを翻訳モデルとよぶ．$P(T)$ は目的言語の文字列としての現れやすさ（言語スコア）を表しており，これを与えるモデルを言語モデルとよぶ．

本研究では，この枠組に基づいた前編集（統計的前編集）を行う．統計的前編集は，前編集前の自然なテキストを S とし，前編集後のテキスト（直訳調のテキスト）を T として，$P(S|T)$ と $P(T)$ の積を最大とする T を求める問題と説明できる．この様子を**図 1** に示す．

式 (3) からもわかるように，統計的前編集のためには $P(S|T)$ と $P(T)$ を適切に学習することが重要となる．本論文では，これらの学習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．なお，本論文では，$P(S|T)$ を与えるモデルを前編集変換モデル，$P(T)$ を与えるモデルを直訳調言語モデルとよぶこととする．

3. 前編集用学習データの自動獲得手順

本章では，前編集変換モデルと直訳調言語モデルの学習データを自動獲得する方法について述べる．ここでは，日英機械翻訳を対象として，自然な日本語から直訳調の日本語に変換するための学習データを自動的に獲得する．**図 2** はこの様子を示したものである．学習データの獲得は，1) 直訳調の文（図 2 の J_i^k）を生成し，2) 自然な日本語文と直訳調の日本語文のペア（図 2 の J^k と J_m^k のペア）から適切なものを選択する，という手順で行う．3.1 節および 3.2 節では，これらの手順について詳しく述べる．

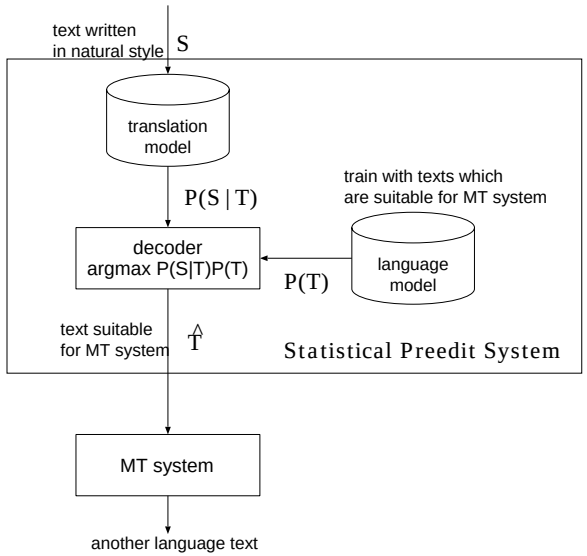

図 1 統計的前編集の枠組
Fig. 1 Framework of statistical preedit

3.1 直訳調の文の生成

日英機械翻訳の前編集モデルの学習のためには，自然な日本語の文とそれに対応する直訳調の日本語の文のペアが必要である．本研究では，文単位で対応づいた日英対訳コーパスから，自然な日本語と直訳調の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) 英日対訳コーパスの 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）機械翻訳システム $\mathrm{MT}_i (i = 1 \cdots n)$ で英日翻訳し，日本語文 J_i^k を得る．

(3) もとの日本語文 J^k と自動生成された日本語文 J_i^k のペアを得る．

なお，こうして得られる J_i^k は直訳調の文であることが多い．

Vol.2009-NL-191 No.1
Vol.2009-SLP-76 No.1
2009/5/21

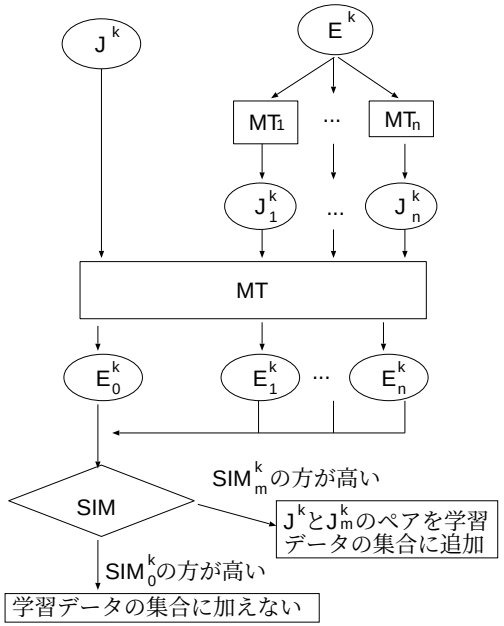

図 2 前編集用の学習データの作成手順
Fig. 2 Block diagram of automatic generation of training data for preedit

3.2 学習データの選択

前節の方法で得られた自然な日本語文 J^k と直訳調の日本語 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．自然な日本語を直訳調の日本語に変換する前処理の目的は，日英翻訳結果がより適切な英語となるようにすることであるため，直訳調の日本語の日英翻訳結果 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ものの場合は，そのような変換を学習すべきではない．別の書き方をすると，E_i^k と参照訳 E^k との類似度を SIM_i^k，E_0^k と参照訳 E^k との類似度を SIM_0^k としたとき，必ずしも $SIM_i^k > SIM_0^k$ が成り立たつわけではない．したがって，$SIM_m^k \leq SIM_0^k$ となるような $m(1 \leq m \leq n)$ に対しては J^k を J_m^k に変換（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．

(1) 自然な日本語文 J^k を日英翻訳し，英文 E_0^k を得る．

(2) 直訳調の日本語文 J_i^k を日英翻訳し，英文 E_i^k を得る．

(3) 参照訳 E^k と $E_i^k(i = 0 \cdots n)$ それぞれとの類似度 SIM_i^k を計算する．

(4) $SIM_m^k > SIM_0^k$ を満たす m に対してのみ J^k と J_m^k のペアを前編集用の学習データとして選択する．

こうして得られた文ペア集合を前編集変換モデルの学習データとし，直訳調の日本語文の集合を直訳調言語モデルの学習データとする．

本提案手法は言語に依存しない手法であり，対象とする言語ペアの対訳コーパスと双方向の機械翻訳システムから，自動的に前編集のための学習データを構築するものである．

4. 学習データの自動生成と前編集モデルの評価

本章では，前編集のための学習データの自動生成の結果について述べる．はじめに実験に使用したデータについて述べる．本研究ではロイター英日対訳コーパス[5]の英文と日本語文を用いた．前編集では，自然な日本語の語句と直訳調の日本語の語句の変換モデルを学習する必要があり，本研究では語句の単位として形態素単位と文節単位を使用した．形態素は形態素解析器 ChaSen[6] の出力結果に基づいており，文節は CaboCha[7] の出力結果に基づいている．

4.1 学習データの自動作成

ロイター英日対訳コーパスの英文 31580 文から直訳調の日本語文を作成した．今回は機械翻訳システムの数 n を 1 とした．類似度 SIM_0^k と SIM_1^k の算出には自動評価尺度 NIST[8] を用いた．

31580 文中 SIM_1^k の方が高くなった文数は 29596 文（93.7%）であった．この結果は，対訳コーパスからほぼ同じサイズの前編集用学習データが得られることを示している．今回は使用した機械翻訳システムの数は 1 つであったが，この数を増やすことで，より多くの前編集用の学習データが獲得できると考えられる．

4.2 前編集変換モデル

4.1 節で得られた自然な日本語文の文節（単語）と直訳調の日本語文の文節（単語）を GIZA++[9] を用いて確率的に対応付けて，前編集変換モデルを学習する．統計モデルであるため学習データ量が大きな問題となる．学習データの日本語文を単語および文節に区切った．得られた単語と文節の延べ数と異なり数を**表 1** および**表 2** に示す．ここでは，異なり数が多い文節単位での対応付けにおいて正しい対応付けが行われているかを調査した．具体

Vol.2009-NL-191 No.1
Vol.2009-SLP-76 No.1
2009/5/21

表 1 学習データにおける単語の延べ数と異なり数
Table 1 Specification of training data; total number of words and total number of unique words

	自然な日本語文	直訳調の日本語文
延べ数	約 1.02M	約 1.11M
異なり数	約 22.3k	約 26.1k

表 2 学習データにおける文節の延べ数と異なり数
Table 2 Specification of training data; total number of phrases and total number of unique phrases

	自然な日本語文	直訳調の日本語文
延べ数	約 307k	約 356k
異なり数	約 125k	約 127k

表 3 対応の評価例
Table 3 Example of alignment evaluation

自然な日本語文の文節	直訳調の日本語文の文節	評価
いう.	言いました.	○
９７年の	１９９７年の	○
可能性が	ことが	△
年間	１年あたり	△
今回の	会社は,	×
する	ために	×

的には GIZA++を用いて自然な日本語文のある文節 s と直訳調の日本語文の文節 t が対応する確率 $t(t|s)$（Lexicon モデルスコア）を求め，その最大値を与える t と s のペアが正しい対応であるかを評価した．

評価は 3 段階で行った.「○」は正しい対応と判断できるものであり,「×」は正しくないと判断できるものである.「△」は文脈によっては正しいと判断できるものを意味する．評価例を**表 3** に示す.

自然な日本語文での出現頻度が閾値以上の文節について対応付けを評価した．結果を**表 4** に示す．出現頻度が高いほどより正しい対応付けが得られており，学習データが多いほど統計モデルが正しく学習できることがわかる．

4.3 直訳調言語モデル

直訳調言語モデルは，4.1 節で得られた自然な日本語文と直訳調の日本語文のペアのうち，直訳調の日本語文のみの 29596 文を用いて学習した．ここでは，単語 5-gram モデルおよび文節 5-gram モデルを学習した．学習は IRST LM Toolkit[10] を用いて行った．

表 4 文節単位での対応付けの評価結果
Table 4 Evaluation of phrase alignment

閾値	文節数	○	△	×
10	3446	1240(35.9%)	407(11.8%)	1799(52.2%)
20	1473	746(50.6%)	174(11.8%)	553(37.5%)
30	863	470(54.4%)	121(14.0%)	272(31.5%)
40	631	361(57.2%)	87(13.7%)	183(29.0%)
50	488	284(58.2%)	68(13.9%)	136(27.8%)
60	373	225(60.3%)	53(14.2%)	95(25.4%)
70	298	184(61.7%)	42(14.0%)	72(24.1%)
80	259	158(61.0%)	38(14.6%)	63(24.3%)
90	226	139(61.5%)	33(14.6%)	54(23.8%)
100	201	127(63.1%)	27(13.4%)	47(23.3%)

5. 統計的枠組みに基づく前編集

実際に学習した前編集変換モデルと直訳調言語モデルを利用して前編集を行った．式（3）を最大化する直訳調の日本語文字列 T を求めるためのデコーダには Moses SMT Decoder[11] を用いた．

評価データにはロイター英日対訳コーパスの 1000 文（**open-1000**）を用いた．評価データには学習データは含まれていない．比較のために学習データに含まれる 1000 文（**closed-1000**）での評価も行った．

文節単位と単語単位での前編集の結果を**表 5** および**表 6** に示す．文節単位で前編集を行って日英翻訳を行ったところ，学習データに対しては 68.5%の文で英文品質（NIST スコア）の改善がみられ，平均でも 3.75 から 4.72 と大きく改善した．評価データに対しては，37.9%の文で英文品質の改善がみられたものの，平均では向上がみられなかった．単語単位での前編集では，文節単位に比べて学習データに対する精度は低かった．評価データに対しては，英文品質が改善した文の割合は多かったものの，平均スコアはほぼ変わらなかった．これらの結果は，文節単位ではパラメータ数に対して学習データが少ないため，うまく学習が行われなかったことを示唆している．

次に，単語単位での前編集において，デコーディングパラメータのチューニングを行った．本研究では，評価データ 1000 文のうちの 200 文をパラメータチューニングのための開発データ（**dev-200**）として用いた．チューニングは前編集結果と直訳調の日本語が近くなるように行った．具体的には，直訳調の日本語文を参照訳とした前編集結果の BLUE ス

Vol.2009-NL-191 No.1
Vol.2009-SLP-76 No.1
2009/5/21

表 5 文節単位での前編集とそれに基づく日英翻訳
Table 5 Phrase-based preedition and its effect on Japanese-to-English translation

評価データ	前編集なし（NIST）	前編集あり（NIST）	改善した文の割合
closed-1000	3.75	4.72	68.5%
open-1000	3.31	3.14	37.9%

表 6 単語単位での前編集とそれに基づく日英翻訳
Table 6 Word-based preedition and its effect on Japanese-to-English translation

評価データ	前編集なし（NIST）	前編集あり（NIST）	改善した文の割合
closed-1000	3.75	4.21	61.2%
open-1000	3.31	3.09	40.7%

表 7 パラメータチューニングの効果（単語単位の前編集とそれに基づく日英翻訳）
Table 7 Effect of parameter tuning (phase-based preedit)

評価データ	前編集なし（NIST）	前編集あり（NIST）	改善した文の割合
dev-200	3.39	3.17	42.0%
open-800	3.29	3.07	40.4%

パラメータチューニングなし

評価データ	前編集なし（NIST）	前編集あり（NIST）	改善した文の割合
dev-200	3.39	3.23	42.5%
open-800	3.29	3.12	40.7%

パラメータチューニングあり（200 文）

コアが最大となるようにチューニングを行った．チューニングした上で前編集を行い，日英翻訳した結果を**表 7** に示す．開発データと評価データ 800 文（**open-800**）の両者とも，チューニングを行わない場合に比べてやや改善がみられた．ただし，前編集を行わないときと比べた場合，平均としての英文品質の改善は得られなかった．今回は，パラメータチューニングは前編集によって直訳調の日本語らしさが高くなるように行った．今後は，最終的な英語の品質が向上するようにパラメータチューニングを行う必要があると考える．

前編集を行っても平均としての英文品質の改善が得られなかったものの，40%程度の文では英文品質の改善が得られることがわかった．**図 3** に前編集によって日英翻訳の品質が改善した例を示す．機械翻訳を翻訳支援システムとして利用すること考えた場合，前編集をした場合としない場合の両方の翻訳結果を提示して，それらのうち適切なものを利用者に選択してもらうという利用が考えられる．このようなタスクでは，平均としての改善が得られなくても半分程度の文で改善が得られることは有益と考えられる．

自然な日本語文

銅は高寄りしたあと続伸し，大半の限月が，この日の高値付近で引けた．

→ Do the stretch in succession of copper after it ..height.. came near, and the greater part of contract monthes were closed in the vicinity of high price of this day. (NIST: 1.18)

前編集後（直訳調）の日本語文

銅は，より高く開きました．登り続けていて，ほとんどの契約が高値の近くで結ばれました．

→ As for copper, it opened higher, and it kept climbing, and most contracts were made near high price within a day. (NIST: 3.14)

図 3 前編集の効果がみられた例
Fig. 3 Example of translation improvement with preedition

6. 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直訳調の文に統計的翻訳の枠組みに基づいて自動変換（前編集）する方法を提案した．その際，機械翻訳システムを利用して，前編集の学習データを対訳コーパスから自動的に獲得する方法を提案し，対訳コーパスからほぼ同じサイズの前編集用学習データが得られることを示した．前編集モデルを学習して前編集を行ったところ，平均的には英語品質の向上はみられなかったものの，約 40%の文で英文品質の向上が得られることがわかった．今後は学習データを増やして実験を行う予定である．

Vol.2009-NL-191 No.1
Vol.2009-SLP-76 No.1
2009/5/21

参 考 文 献

1) 山口昌也，乾　伸雄，小谷善行，西村恕彦：前編集結果を利用した前編集自動化規則の獲得，情報処理学会論文誌，pp.17–28 (1998).
2) 阿辺川武，影浦　峡：下訳と修正訳を用いた訳文修正パターンの発見，言語処理学会第 13 回年次大会発表論文集，pp.919–922 (2007).
3) 下岡和也，南條浩輝，河原達也：講演の書き起こしに対する統計的手法を用いた文体の整形，自然言語処理， Vol.11, No.2, pp.67–83 (2004).
4) Brown.P.F, Pietra.V.J, D., Pietra.S.A., D. and Mercer.R.L.: The Mathematics of Statistical Machine Translation: Parameter Esti mation, *Computational Linguistics*, Vol.19, No.2, pp.263–311 (1993).
5) Utiyama, M. and Isahara, H.: Reliable Measures for Aligning Japanese-English News Articles and Sentences, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.72–79 (2003).
6) 松本裕治，　北内啓，山下達雄，平野善隆，　松田寛，高岡一馬，浅原正幸：形態素解析システム『茶筌』version 2.3.3 使用説明書 (2003). http://chasen-legacy.sourceforge.jp/.
7)　工藤拓，松本裕治：チャンキングの段階適用による係り受け解析，情報処理学会論文誌， Vol.43, No.6, pp.1834–1842 (2002).
8) Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics, *Proceedings of the 2nd Human Language Technologies Conference (HLT)*, pp.128–132 (2002).
9) Och, F.J. and Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, Vol.29, No.1, pp.19–51 (2003).
10) Federico, M. and Cettolo, M.: Efficient Handling of N-gram Language Models for Statistical Machine Translation, *Proceedings of the Second Workshop on Statistical Machine Translation*, pp.88–95 (2007).
11) Hoang, H. and Koehn, P.: Design of the Moses Decoder for Statistical Machine Translation, *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pp.58–65 (2008).