機械翻訳における前編集規則の自動獲得

岡田 真也　　南條 浩輝　　吉見 毅彦

龍谷大学 理工学部 情報メディア学科

1　はじめに

自然な入力文を機械翻訳システムで翻訳すると望ましい翻訳が得られないことがある。むしろ直訳調の文を入力としてシステムに与える方が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。本研究では、対訳コーパスにおける第 1 言語の文を機械翻訳システムで翻訳して得られる直訳調の第 2 言語の文に現われる語句と、対訳コーパスにおける第 2 言語の文（自然な文）の語句とを確率的に対応付けることで前編集規則を獲得する。

2　前編集規則の獲得アルゴリズム

本研究では、日英機械翻訳システムの入力文の前編集規則を自動的に獲得する。前編集規則の獲得は以下の手順で行う。

1. 直訳調の文（図 1 の J_i^k）を生成する。
2. 自然な日本語文と直訳調の日本語文のペア（図 1 の J^k と J_m^k のペア）を前編集規則獲得のための学習データとする。
3. 自然な日本語文の文節と直訳調の日本語文の文節を確率的に対応付けて前編集規則を獲得する（図 2）。

これらの詳細を 2.1 節、2.2 節、2.3 節で述べる。

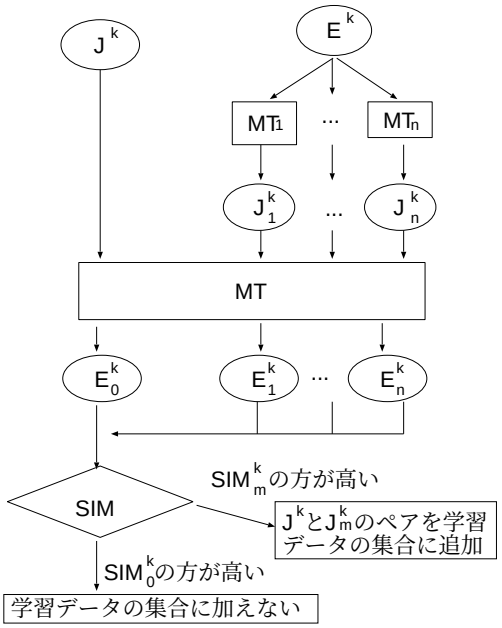

図 1: 前編集規則獲得のための学習データの作成

2.1　直訳調の文の生成

英日対訳コーパスの k 番目（$k = 1 \cdots N$）の日本語文を J^k とし、その対訳英文を E^k とする。E^k を（複数の）機械翻訳システム $MT_i (i = 1 \cdots n)$ で英日翻訳し、日本語文 J_i^k を得る。この J_i^k は直訳調の文であることが多い。

2.2　前編集規則を獲得するための学習データの作成

J_i^k の日英翻訳結果 E_i^k は J^k の日英翻訳結果 E_0^k より必ずしも良いわけではない。すなわち、E_i^k と参照訳 E^k との類似度を SIM_i^k、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 に書き換えるべきではない。これらのことより、$SIM_m^k > SIM_0^k$ を満

たす m に対してのみ J^k と J^k_m のペアを前編集規則獲得のための学習データに加える。

2.3 前編集規則の獲得

自然な日本語文の文節と直訳調の日本語文の文節を確率的に対応付けて前編集規則を獲得する。図 2 にその概要を示す。まず、自然な日本語文と直訳調の日本語文を文節単位に区切る。次に、GIZA++[1] を用いて自然な日本語文の文節と直訳調の日本語文の文節が対応する確率を求める。最後に、確率の高い対応を抽出し前編集規則とする。

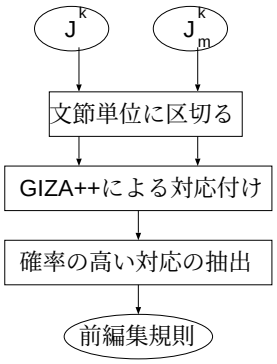

図 2: 前編集規則の獲得

自然な日本語文のある文節 A に対して、直訳調の日本語文の複数の文節が対応することがある。このときは対応確率が最も高い文節 B を選び、前編集規則（文節 A から B への書き換え規則）とする。例えば、「伝えた」という文節と「報告しました」という文節の対応確率が 0.9 で、「伝えた」と「伝染した」という文節の対応確率が 0.1 であるとき、「伝えた」を「報告しました」に書き換える前編集規則とする。

3 実験と結果

本章では、前編集規則を獲得する実験を行い、獲得できた前編集規則の評価を行う。

本実験ではロイター英日対訳コーパス [2] の英文と日本語文を用いた。形態素解析には Chasen[3]、文節区切りには CaboCha[4] のバージョン 0.53 を使用した。

3.1 学習データの作成

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 m を 1 とした。類似度 SIM^k_0 と SIM^k_1 の算出には自動評価尺度 NIST[5] を用いた。

31580 文中 SIM^k_1 の方が高くなった文数は 29596 文（93.7%）であった。これは、対訳コーパスがあれば、前編集規則を獲得するための学習データが得られることを示している。

3.2 前編集規則の獲得

3.1 節で得られた学習データの日本語文を文節に区切った。文節区切りの結果得られた文節の延べ数と異なり数を表 1 に示す。

表 1: 学習データでの文節の延べ数と異なり数

	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

次に、獲得した前編集規則が正しい規則であるかどうかを評価した。評価は 3 段階で行った。「○」は正しい書き換えであると判断できるもので、「×」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 2 に示す。

表 2: 前編集規則の評価例

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	×

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前編集規則を評価した。結果を表 3 に示す。出現頻度が高い文節ほど前編集規則の正解率が高いことが分かる。今回の実験の学習データは、29596 文であっ

たが、さらに学習データ量を増やすことで正しい前編集規則が獲得できると考えられる。

表 3: 前編集規則の評価結果

閾値	規則数	○	△	×
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%)

4 前編集規則の適用

実際に獲得した前編集規則を学習データ 1000 文に適用し、評価を行う。自然な日本語文での出現頻度が閾値以上であり、かつ対応確率も閾値以上である文節に対して前編集規則を適用した。適用対象の文節の前編集規則が存在しない場合、その文節はそのまま出力した。また、文節の順序の入れ換えは行わなかった。前編集規則を適用した日本語文を日英翻訳した英文と参照訳 E^k との類似度を算出した。1000 文での平均類似度を表 4 に示す。前編集規則を適用せずに日英翻訳した場合の平均類似度は 3.75 であった。

前編集規則を適用しても平均類似度の向上は見られなかった。内訳を分析したところ、36.5~39.0%の文については、適用しない場合に比べて類似度が向上していることがわかった。なお、翻訳支援システム、すなわち、ユーザが入力した文とそれを前編集した文それぞれの翻訳結果をユーザに提示し、適切な英文を選択させるような支援システムでは、平均類似度が向上しなくても半分程度の文で改善がみられることは有益と考えられる。学習データ量を増やせば、平均類似度の向上と、類似度が向上する文数の増加が期待できる。

表 4: 前編集規則を学習データに適用した結果

出現頻度の閾値	対応確率の閾値	平均類似度	類似度が向上した文数
30	0.1	3.67	389(38.9%)
30	0.2	3.67	389(38.9%)
30	0.3	3.68	389(38.9%)
30	0.4	3.69	390(39.0%)
30	0.5	3.70	381(38.1%)
40	0.1	3.71	378(37.8%)
40	0.2	3.71	378(37.8%)
40	0.3	3.71	378(37.8%)
40	0.4	3.73	378(37.8%)
40	0.5	3.72	365(36.5%)

5 結論

本稿では、対訳コーパスにおける第 1 言語の文を機械翻訳システムで翻訳して得られる直訳調の第 2 言語の文に現われる語句と、対訳コーパスにおける第 2 言語の文（自然な文）の語句とを確率的に対応付けることで前編集規則を獲得することを提案した。実験では、出現頻度が高い文節についての前編集規則が正しいかどうか検証した。その結果、出現頻度が高いほど正しい書き換え規則を獲得できることが確認できた。実際に獲得した前編集規則を学習データ 1000 文に適用したところ平均類似度の向上は見られなかったが、36.5~39.0%の文については類似度が向上した。

今後の課題として、学習データを増やすことと、前編集での文節の入れ換えを検討することが挙げられる。

参考文献

[1] Och.F.J and Ney.H. A systematic comparison of various statistical alignment models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51, 2003.

[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.

[3] 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. Technical report, 奈良先端科学技術大学院大学 情報科学研究科自然言語処理学講座, 2003. http://chasen-legacy.sourceforge.jp/.

[4] 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. http://chasen.org/ taku/software/cabocha/.

[5] G. Doddington. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd Human Language Technologies Conference (HLT)*, pp. 128–132, 2002.