機械翻訳における自動校正と日中翻訳への適用

山本 和英
E-mail: yamamoto@itl.atr.co.jp
ATR 音声翻訳通信研究所

1 はじめに

機械翻訳における後編集について述べる。自然言語処理、特に機械翻訳においては入力の前編集の重要性はこれまでたびたび議論されてきたが、出力の後編集については、人手による編集のしやすさなどが議論されているだけで、計算機による後編集の議論はあまり行なわれていない。本研究では機械翻訳における生成処理において、自然と感じる文の生成、特に校正作業によって自然な文に変更することを目指した。ただし構造の大幅な変更は本稿の対象外とする。

本稿では、校正規則の自動収集についての試みを報告する [Yam99]。翻訳結果に対して手作業により校正を行ない、この両者を比較することによって規則を作成する。規則は校正前後のDPマッチングをとることにより、局所的に変更するような形式で記述する。以上の処理を日中翻訳の中国語生成に適用した結果を報告する。

2 機械翻訳における校正処理

機械翻訳を文生成の問題として捉えた場合、原言語の内容をいかに正確に伝えるかという問題とは別に、いかに自然な文を生成するかという問題がある。この両者は共に重要な問題であるが、翻訳という処理の性質上内容の正確な伝達をより重視するため、自然な文の生成はこれまであまり活発に議論が行なわれてこなかった。現在の機械翻訳システムの出力文の自然さも十分とは言えない。

一般に、機械翻訳における文の語順などの調整は、生成部において行なっている。しかし、従来のシステムで行なわれているのは、英語生成における副詞的要素の位置の移動や疑問詞疑問文における疑問詞の移動など、構文的な理由に基づくものが多く、言いやすさ、語調をそろえるためなどの文の自然さの考慮はあまりなされていない。

以上のような動機に基づき、本研究では機械翻訳の処理結果をより自然な文に書き換えることを目指す。この処理の実現のために、人手で行なった校正結果を利用し、それを規則化することで自動的に校正を行なうことを試みる。生成される文がどのような場合に自然と感じられるかは難しい問題であるため「自然さ」の規則を直接作ることは困難であろう。また校正対象の「不自然さ」はそのシステムに依存する。これに対し、与えられた文を自然な文に校正することは比較的容易であるので、本研究では校正前後の文の比較によって校正知識を獲得するというアプローチを採る。このため本稿では、校正前のテキストは形態素および品詞の情報を持ち、校正後のテキストはこれらの情報を持たない文字列と仮定する。

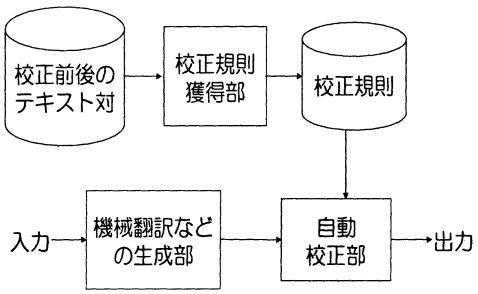

図 1: 自動校正システム概要

3 自動校正知識の獲得

3.1 概要と方針

本稿で提案するシステムの概要を図1に示す。校正システムは、規則の獲得部と適用部に分かれる。規則獲得部は翻訳出力とその人手による校正結果を組として入力し、これらから規則獲得部において矛盾のない規則の集合が出力され、保存される。規則適用部(自動校正部)は、機械翻訳などの出力結果を入力とし、予め決められた順に校正規則を適用していく。

観察可能な少数の現象から規則作成する際には、できるだけ規則を一般化してその規則の適用範囲を広くしなければならない。その一方で、現象の一般化は誤った解釈を起こす可能性も高くなるので、副作用の問題が大きくなる。これらを両立させることは観察対象が少数の場合においては特に困難となる。そこで本

```
Step 0    G = 文生成 (), P = 校正 (G),
          R = (), n = 2;
Step 1    R0 = 仮規則生成 (P, G, n);
Step 2.1  R1 = 有害規則排除 (P, G, n, R0);
Step 2.2  R2 = 矛盾規則排除 (R1);
Step 2.3  R3 = 照合検定 (P, G, R2);
Step 2.4  R4 = 重複規則排除 (R3);
Step 3    R = 規則追加 (R, R4),
          G1 = 規則適用 (G, R4);
Step 4    if (G1 == P or G1 == G)
          then Step 5;
          else G = G1, n++, goto Step 1;
Step 5    return R;
```

図 2: 校正規則獲得アルゴリズムの概略

研究では、悪影響を出さないことをより重視して規則の作成を行なった。すなわち、規則を検定する際に、悪影響が出る恐れのある規則をできるだけ排除する方針で取捨選択を行なった。これはいわば「臆病な」規則作成方針である。

さらに、規則は局所的な視野において作成した。これは、規則の表現形態として要素列の置換による表現、すなわち、ある要素列 A があればこれを要素列 A' に置換せよ、という表現が処理上便利であり、本稿の獲得対象としている文の自然さは概ね前後の局所的な情報が関係すると予想したためである。

3.2 校正規則の獲得

規則獲得部では、校正されたテキストを校正前のものと比較することで規則を作成する。比較は文字を単位とする DP マッチングによって行なう。図2に、規則獲得過程の概要を示す。規則生成部は翻訳された文の集合 G と G の各文の人手による校正結果の集合 P を入力とする。

Step 1 では、G と P の対応する各文ごとの DP マッチングによって仮規則集合を作成する。DP マッチングは文字を単位にして行なう。DP マッチング後に、連続する要素列の書き換えによって両者の差異を吸収するように規則を作成する。この際、規則作成時には G の各文が持つ単語および品詞に対して規則作成を試みる。このようにして作成された規則を、ここでは仮規則と呼ぶ。

例えば、G に属するある文 s が ⟨abcd⟩[1], ここで a,b,··· の品詞は A,B,··· とする。この文が校正の結果 s' = ⟨abxcd⟩ となり、これが P に格納されているとする。s と s' の DP マッチングの結果、x の挿入があることがわかる。従って、規則生成部ではこの違いを吸収するような仮規則を作成する。このとき、要素長パラメータ n の値によって、どの程度周囲に依存した規則を生成するかを決定する。例えば、初期設定では n が 2 であり、規則左辺の要素長が 2 である以下の 12 の規則が生成される。この時、右辺の要素長は差異の種類 (挿入、欠落、置換) によって左辺の要素長 ±1 のいずれかになる。これによって一ヶ所の修正位置に対して複数の仮規則が作成される。

```
<ab> --> <abx>,  <Ab> --> <Abx>,
<aB> --> <aBx>,  <AB> --> <ABx>,
<bc> --> <bxc>,  <Bc> --> <Bxc>,
<bC> --> <bxC>,  <BC> --> <BxC>,
<cd> --> <xcd>,  <Cd> --> <xCd>,
<cD> --> <xcD>,  <CD> --> <xCD>
```

同様に、規則の要素長 n が 3、G の一文が ⟨abcde⟩、そして対応する P の一文が ⟨abde⟩ である場合、以下の 18 の規則が仮生成される。

```
<abc> --> <ab>,  <bcd> --> <bd>,
<cde> --> <de>,  <Abc> --> <Ab>, ...
```

これらの処理は P のすべての文に対して実行され、その結果仮生成された規則集合は $R0$ に格納される。この時、重複した場合でも一つとみなし頻度は記録しない。上の例に示すように、一ヶ所の修正に対して複数の仮規則が作成されるが、矛盾したものや最終的に重複しているものは削除される。

Step 2 では、Step 1 で作成された仮規則の絞りこみを行なう。これは、「有害」な規則の排除、矛盾の排除、最終的な検査、重複規則の排除という、4 種類の処理に分かれ、これらの処理によって徐々に仮規則の絞り込みを行なっていく。

Step 2.1 では、まずすべての文を対象に DP マッチングによって不変化部を要素列形式で抽出する。例えば、正解とその校正結果がそれぞれ ⟨abcd⟩ と ⟨abxcd⟩ であった場合に、⟨ab⟩ と ⟨cd⟩ という二つの文字連鎖が抽出される。これをすべての文に対して行ない、変化させてはいけない要素列の集合を集める。次に、これら不変化部に対して操作を行なっている仮規則を「有害」な仮規則とみなし、これを排除する。この処理は、Step 2.3 で DP マッチングの回数を低減させるために採用した。

Step 2.2 では、仮規則のうちで矛盾した規則を持つものを排除する。例えば、ある二つの仮規則が ⟨bc⟩

[1] ⟨···⟩ は文字列、a,b,··· は単語、x は文字である。

--〉〈bxc〉と〈bc〉 --〉〈byc〉であった場合、同一の文字列から別の操作を行なうことになり、規則適用の際にどちらを採用すべきか決めかねる。そのため、この両規則はその条件部に問題がある「悪い規則」であると考え、この両者を仮規則から排除する。

Step 2.3 では、最終的な規則の検査を行なう。ここでは、これまでの処理で排除されなかった各仮規則をすべての訓練文に対して実行する。その結果、DP距離が悪化した文が1文でもあれば、他の文でDP距離が改善されていてもその文は「有害」であると判断し、排除する。

Step 2.4 においては、残された仮規則に対して重複を調査し、これらに対してより一般的な規則を採用する。例えば仮規則中に〈bc〉 --〉〈bxc〉と〈BC〉 --〉〈BxC〉(単語bの品詞をB、単語cの品詞をCとする)の二つの規則があった場合は、後者のほうがより一般的であり、前者の規則が適用される場合には後者も必ず適用される。このことから後者の規則がある場合には前者の規則は不要であるため、このような重複規則のチェックを行ない、該当した場合には個別的な規則を削除する。

Step 3 では、以上の処理で残された仮規則を正式にパラメータ n における規則として採用し、校正対象文集合 G に対してこれら規則をすべて適用する。この際、適用する順序はどれからでもよく、任意である。また、競合する規則もない。全規則適用後の文集合を $G1$ とする。

最後に、Step 4 において終了条件がチェックされる。アルゴリズムの終了条件は、Step 3で校正後の文集合 G が最終校正結果 P と完全に一致するか、または、今回の校正規則適用によって校正対象 G が何も変化しなかった場合である。このどちらかを満たした場合にアルゴリズムは終了し、Step 5 で校正規則集合 R を返して終了する。そうでない場合には G を $G1$ と再定義し、パラメータ n を1増加して再び Step 1 に進む。

3.3 規則の適用

一般に、実際のシステムにおいては規則適用の際の処理時間を考慮することが重要である。すなわち各規則はできるだけ短時間で適用されることが望ましい。また、規則適用の際には規則相互の関係が明確であり、適用の際に適用順を明確に規定できることが望ましい。提案手法においては、各規則は入力とのパターンマッチングによって実行されるので、比較的短時間で処理できる。また、規則は互いに競合することのなく、n が同一であればすべて同時に適用できるように作成しているため、適用時には規則の選択やその適用順の選択を行なう必要がない。

4 TDMT 日中翻訳

変換主導翻訳 (Transfer-Driven MT、以下 TDMT) では、変換処理を中心に形態素解析、解析処理、生成処理など各処理が翻訳のために協調した処理を行ない、翻訳結果を作り出す翻訳手法である [Yam96]。

TDMT は様々な言語対に対して有効性を確認することを目的として、これまで日英、日独、日韓 [Yam96]、英日、韓日における話し言葉に対して研究を行なっており[2]、今回新たに第六の言語対として日中翻訳への研究を開始した。本稿で提案する自動校正処理はこの日中を言語対とする TDMT に対して実装し、現在検討を行なっている。翻訳の対象は旅行会話であり、ATR 旅行会話データベース [Tak98] を実験に使用した。

5 検証実験

提案手法の有効性を確認するため、手法を計算機上に実現し、小規模な実験を行なった。本節ではこの概要と結果を報告する。

実験では、まずコーパス中から 432 文を選択し、これらの文を TDMT 変換知識のみで出来るだけ自然に翻訳させることを試みる。次に、これらを人手によって校正を行なう。この際、校正する必要のない文も存在するが、このような翻訳出力も以降の校正処理の対象とする。

以上のようにして準備した翻訳出力とその校正結果の文を校正知識作成部の入力として、前述した校正知識獲得処理を行なう。この際に作成された仮規則数および各処理で排除される規則数をまとめたものを表1に示す。

今回行なった実験では、左辺が6要素の規則が要素数として最大であり、7要素の処理において一つも校正規則が採用されずアルゴリズムが終了した。本稿のように、一つの校正箇所に対して複数の規則を作成することを認めた場合、一般的に n が増加すると組み合わせが非常に多くなる。このため、n が小さいうちに多くの校正を行なわないと規則数は n の増加によって爆発的に増加してしまう。しかし本稿でのアルゴリズムでは、比較的消極的な規則作成の方針を採ったにもかかわらず、多くの校正箇所が2要素や3要素の規則

[2] これら従来の言語対に対する翻訳性能については [Fur97] を参照されたい。

表 1: 規則獲得時における規則数

規則の左辺要素列長 n	2	3	4	5	6	7	合計
作成された仮規則数 (Step 1)	3291	2603	1881	1131	546	179	9631
有害規則の排除数 (Step 2.1)	621	858	593	390	95	159	2716
矛盾規則の排除数 (Step 2.2)	1505	720	388	357	434	19	3423
照合検査での排除数 (Step 2.3)	318	193	186	81	1	1	780
重複規則による排除数 (Step 2.4)	283	429	467	261	14	0	1454
最終獲得規則数	564	403	247	42	2	0	1258
n の仮規則数に対する採用率	17.1%	15.5%	13.1%	3.7%	0.4%	0%	13.1%
全規則数に対する割合	44.8%	32.0%	19.6%	3.3%	0.2%	0%	100%

表 2: 自動校正による自然さの改善

	既知	(%)	未知	(%)
改善	386	89.4%	185	41.7%
悪化	0	0 %	65	14.6%
混合	0	0 %	19	4.3%
無変化	46	10.6%	175	39.4%
合計	432	100%	444	100%

によって校正されている様子が表1からわかる。以上の観点から、校正規則の獲得処理は規則数の爆発を抑えることができ、有効に機能していると考えられる。

機械翻訳の出力結果に対し自動校正処理を適用し、適用前後でどれだけの文が改善されたかを表2にまとめた。表の未知欄は、TDMT 日中翻訳の翻訳知識に対しても入力文は未知であるオープンテストである。また改善欄は校正位置が以前よりも改善されたことを意味し、それ以外の部分の訳質を問わない。混合欄は改善と悪化の両者があった文および訳質変化に無関係の校正があった文を指す。

校正知識の獲得対象とした文に対する改善状況では、約 10% 程度の文が同一出力の文となった。これは校正不要文と校正状況が前文に依存しているなどに伴う無改善文に分かれるが、前者の方が多い。また規則作成方針通り悪化した文がないことを確認した。

未知の文に対する校正状況では、15% 程度が悪化したものの、40% 以上の文で改善が観察された。また規則が全く適用されなかった文が 40% 程度あるが、この中には校正を要する文がかなり含まれている。規則獲得に使用した文がまだ少ないため、悪化した文や同一出力の文の割合が多いが、今後は悪化文の割合をより減らすよう、改良していきたい。

6 まとめ

機械翻訳における自動校正について議論し、処理手法を提案した。これを話し言葉多言語翻訳 TDMT の日中翻訳部に導入し評価実験を行なった結果、小規模ではあるがある程度有効に機能することを確認した。今後は大規模化に伴う問題、悪影響の低減などに取り組んでいく。

参考文献

[Fur97] 古瀬蔵, 美馬秀樹, 山本和英, PAUL, M., 飯田仁：多言語話し言葉翻訳に関する変換主導翻訳システムの評価, 年次大会講演論文集, 第 3 回, pp. 39–42, 言語処理学会 (1997).

[Tak98] TAKEZAWA, T., MORIMOTO, T., and SAGISAKA, Y.: Speech and Language Database for Speech Translation Research in ATR, In *Proc. of 1st International Workshop on East-Asian Language Resources and Evaluation - Oriental COCOSDA Workshop*, pp. 148–155 (1998).

[Yam96] 山本和英, 古瀬蔵, 飯田仁：用例に基づく日韓の対話翻訳処理機構, 全国大会講演論文集, 第 53 回, 4L–10, pp. 2/71–72, 情報処理学会 (1996).

[Yam99] YAMAMOTO, K.: Proofreading Generated Outputs: Automated Rule Acquisition and Application to Japanese-Chinese Machine Translation, In *Proc. of 1999 International Conference on Computer Processing of Oriental Languages (ICCPOL'99)* (1999).