換言と言語変換の協調による機械翻訳モデル

山本 和英
kazuhide.yamamoto@atr.co.jp
ATR 音声言語コミュニケーション研究所

概要

換言処理を基軸とした機械翻訳／音声翻訳モデル (SANDGLASS 翻訳モデル) の提案を行なっている。これまでに、日中翻訳を対象に翻訳モデルの一部を実装し、プロトタイプを作成した。本稿では、このプロトタイプのうち原言語換言部と言語変換部における協調処理について報告する。

1 はじめに

我々人間の持つ言語運用能力は、母語に対する能力と比較して、外国語に関する能力は常に低い。そのため、母語を外国語に翻訳する際には、言語使用時に意識するかどうかに関係なく、人は母語の運用能力をできるだけ活用して外国語に翻訳を行なっている。一方、従来の機械翻訳モデルは二言語処理 (言語変換処理) が処理の中心であり、原言語で行なう解析処理はあくまでも二言語処理を行なうための前処理という位置付けでしかない。

我々は、従来の典型的な機械翻訳モデルと比較して、より人間の翻訳過程に近いと考えられる機械翻訳モデルを提案している。このモデルは、二言語知識が十分でない場合に原言語において翻訳可能な表現に換言してから翻訳するモデルであり、この意味において外国語初学者が行なう翻訳過程と同様の処理と考えられる。この「翻訳初学者モデル」とも言うべき我々の翻訳モデル (SANDGLASS モデル [Yam01b] と呼ぶ) は、工学的観点では二言語知識の低減が可能となるため多分野、多言語へ適用しやすくなるという利点を持つ。また、換言処理が入出力共に同一言語であるという処理の性格上、翻訳に限らず自然言語処理のほとんどすべての問題に換言技術の転用が可能である。

我々は現在、中日／日中の言語対に対して SANDGLASS 翻訳機構の構築を進めている。これまでに、日中翻訳においてこのプロトタイプを実装したので本稿ではこれを報告する。

2 SANDGLASS 翻訳モデル

2.1 各処理部の独立性と処理方略

SANDGLASS モデルにおいては、原言語における換言処理 (以下、単に換言処理と呼ぶ) と目的言語への言語変換処理 (以下、単に変換処理と呼ぶ) の独立性を高くして部品化することで各部の開発効率を高める。また同時に、換言部においてできるだけ汎用的な換言処理を行なわせることで機械翻訳に依存しない換言機構の構築を目指す。

変換部がどのような知識を持つかに依存しない換言処理を実現することは部品の汎用性や開発の独立性など様々な特長を持つが、その一方でどのような換言を行なえばいいのか、あるいはどのような表現が変換可能なのかという換言目標が立てづらい。これに対する一方策として、換言部において入力文のすべての可能な換言文を作成し、この全文に対して変換処理を試み、変換に成功した文のうち最良のものを変換結果とする方策がある。しかし、例えば実時間性を要求される音声翻訳においてこのような処理を実時間の範囲で実現することは考えにくく、また無目標で換言文を大量に生産することは合理的でない[1]。

これに対し、SANDGLASS モデルでは以下のような方策を立てた。すなわち、換言部と変換部は互いに独立した部品として構成するが、両者の間に制御部を設け、この制御部を介して換言部と変換部が協調することで翻訳文を生成する。つまり、翻訳を行なうための情報の流れは換言部から変換部への一方的なものとはせず、換言部と変換部の両者が与えられた入力文に関して自らの必要十分な情報を交換することで翻訳文を生成する枠組みを提案する。具体的には以下のような特徴を持つ。

1. 換言部は1回の換言要求につき必ず換言文1文を出力する。
2. 変換部は変換に失敗した場合、換言部に換言を行なうための「ヒント」を、可能な場合に出力する。
3. 換言部はヒントにできるだけ沿った換言文の生成を試みる。ただし、ヒントに従うかどうかは換言部が判断する。

[1] 一般的には局所的で独立性の高い換言を複数箇所で行なうことが可能な場合が多く、その結果組み合わせ的に非常に多数の換言文を生成可能である。

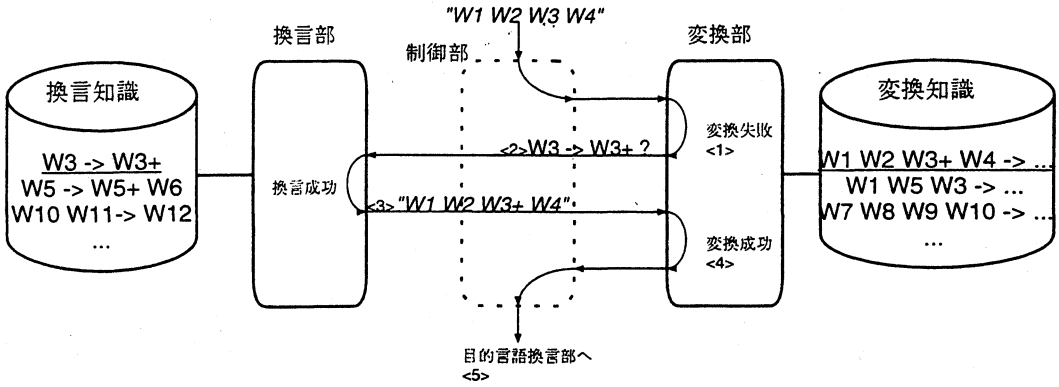

図 1: 換言部と変換部の協調による翻訳モデル

2.2 換言部と変換部の協調機構

図 1に我々の提案を図示した。翻訳機構の主要部は (原言語) 換言部と変換部から構成され、両者を制御する制御部がその間に入る[2]。この機構の特徴は (1) 換言部と変換部が対等で処理の上流、下流という関係がない (2) 換言するための換言知識と変換するための変換知識が完全に独立であり、双方が独自の基準で処理を行なう、の 2 点である。

形態素解析結果はまず制御部を介して目的言語への変換が試みられる。仮に、$W_1W_2W_3W_4$ という形態素列の変換を試みるが失敗したとする ⟨1⟩。この際、変換可能な $W_1W_2W_{3+}W_4$ という類似した表現があり、もし W_3 が W_{3+} に換言可能であれば変換可能であるという情報を変換部が換言部に伝える ⟨2⟩。この情報を受け取った換言部は換言知識を調べ、W_3 が W_{3+} に換言可能であるためこの語を換言して変換部に返す ⟨3⟩。変換部は換言後の表現の変換を試み、今度は成功するため ⟨4⟩、変換されて目的言語となった表現は以後の処理に渡される ⟨5⟩。

仮に、最初の変換処理 ⟨1⟩ で類似した表現を見つけ出すことができなかった場合は、換言部が自らの換言知識を使用することによって何らかの換言文を生成する。変換部から与えられた換言の「ヒント」の通りに換言できないと換言部が判断した場合も、換言知識を使って換言を試みる。

2.3 換言部

換言部では以下の種類の換言を以下の順に行なう。いずれかで換言事例を得ることができた時点で処理は終了し、換言結果を出力する。

1. 変換部からの「ヒント」による換言
2. 入力分割
3. 敬語の削除 [Oht01]
4. 文末表現などの単純化 [Yam01a]
5. 名詞連続の複合名詞化
6. 文の要素削除

(1) 変換部からの「ヒント」による換言は、変換処理の結果出力された換言候補の通りに換言することが可能かどうかの検証を行なう。この処理は、あらかじめ語彙レベルで換言可能な事例を用意しておき、変換部から与えられる換言候補との照合を行なう。この処理によって行なわれる換言は、ある 1 単語を同一品詞内で異なる 1 単語に換言するもののみである。換言される語は主に機能語であるが、内容語である場合もある。

原文 おもしろそうですねえ

換言文 おもしろそうですね

原文 御安心ください

換言文 ご安心ください

(2) 入力分割は、入力が 2 文以上から構成される場合にこれを分割することが処理の目的である。話し言葉は書き言葉とは異なり、句点で入力を分割することが不可能であるので、終助詞などの存在で入力を分割する規則を用意した。この規則に照合した場合、分割を行なう。なお、以下では分割点を記号 ";" で表現する。

原文 それじゃあさよなら

換言文 それじゃあ; さよなら

[2]本論に関係のない処理部 (例えば目的言語換言部) は簡単のため省略した。

原文 いくらですかそれ

換言文 いくらですか; それ

(3) 待遇表現は日本語会話表現に頻出するため、このすべての表現を翻訳対象とすることは二言語知識の増大を招き、好ましくない。このため、可能な範囲で簡単化もしくは削除を行なう ([Oht01])。

原文 ではいかがいたしましょうか

換言文 ではどうしましょう

原文 あいにくですがございません

換言文 あいにくありません

(4) 文末表現等も日本語には多様な表現が存在するため、これを簡単化する。また、口語的表現 (「…してる」「…したんですけど」など) の削除もできる限り試みる ([Yam01a])。

原文 風邪じゃないかと思うんですけど

換言文 風邪でしょう

(5) 名詞連続の複合名詞化は、普通名詞またはサ変名詞が直接または助詞「の」を介して連続していた場合にこれを一語の名詞 (以下の例では {…} と表す) としてみなす処理である。後述する変換部で行なう処理の都合上、複合名詞と単一名詞を同一視して変換することが不可能なため、テンプレート方式の欠点を補うためにこのような処理を行なう。ただし、話し言葉の場合にこの処理を無条件で行なうことは問題があるため、処理の優先度を低くした。

原文 火曜日の午後五時なんですが

換言文 { 火曜日の午後五時 } なんですが

(6) 文要素の削除は、最後の手段として文を構成する要素のうち、比較的重要でないと考えられる副詞 (「ちょっと」「たぶん」「本当に」など) や助詞「は」「も」など (他の格助詞と同時に使われる場合) の削除を行なう。

原文 明日までにはご用意いたしますよ

換言文 明日までに用意します

原文 たぶん十分くらいだと思います

換言文 十分くらいだと思います

2.4 変換知識の作成

事前準備として、簡易の変換知識の作成を行なった。変換知識として、本研究では二言語対訳コーパスと複数の対訳が付与されている日中対訳辞書から自動作成した。

まず、我々の収集した日本語旅行会話表現約 23 万文に対し全文に中国語訳を付与した。ここで、中国語対訳は形態素情報を含む一切の言語情報は付与されていない。次に、JUMAN[3] と KNP[4] を使って日本語解析を行なった後で、日中対訳辞書を用いて単語単位の単純な日中の対応付けを行なった。すなわち、日本語解析結果の各単語に対して、その対応する中国語単語が中国語訳文の中に含まれているかどうかを調べ、照合した場合は日中の当該単語に同一の ID 番号を付与する。ある単語に対して、対訳辞書中の複数の単語に対して照合した場合は、それらのうち最長の中国語単語に対して対応付けを行なった。

ここで、ある日本語に対してある単一の中国語訳が複数箇所に見られる場合は、そのすべてに対して対応付けした。例えば、「行きませんか」に対して「去不去？」(「去」は「行く」の意) という対訳が付与されていた場合、「〈去 #538〉不〈去 #538〉？」などとして日本語の「行く」に対する同一の ID(例では #538) を付与した。

以下では、この処理によって対応づけされた原言語 (日本語) と目的言語 (中国語) 一対の表現対をテンプレート、テンプレート中で対訳辞書によって対応付けできた部分をそのテンプレートの変数、それ以外の項目を固定表現と呼ぶ。

2.5 変換部

変換部は以上の処理で作成されたテンプレートに基づく翻訳を行なう。変換処理は、テンプレート検索、テンプレート照合の二つの処理から構成される。

まず、入力文 (の形態素解析結果) に対し、入力文と全く同一の品詞列を持つテンプレートを検索する。これが存在しない場合は変換に失敗し、新たな換言文を要求する。

同一品詞列のテンプレートが検索できた場合は、次に各形態素ごとに入力とテンプレートとの照合を行なう。照合にすべて成功した場合は変換成功となり、目的言語へ変換した結果を制御部に返す。テンプレート照合に失敗した場合は、テンプレート中の一部の固定表現が入力とは異なることを意味する。つまり、両者の差異を変換部が把握しており、またこの差異が換言処理の際の目標表現になり得るので、この相違する表現列を換言のための「ヒント」として制御部に返す。ここで、テンプレート照合に失敗したテンプレートが

[3] http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[4] http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/knp.html

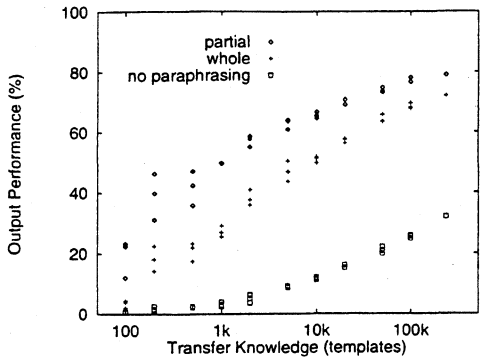

図 2: システムの出力性能

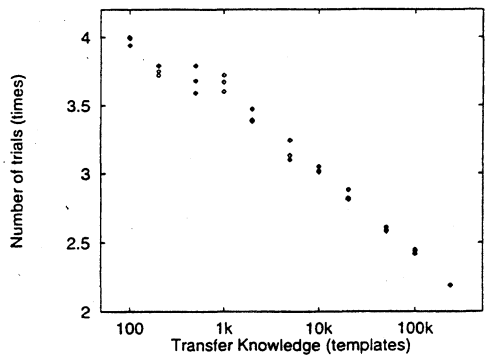

図 3: 換言試行回数

複数ある場合は、それらすべてに対して差異を抽出し、列挙して返す。

3 実験

以上のシステムに対して、変換知識量と出力性能の関係を見る実験を行なった。変換知識は、旅行会話に関する約 23 万文の日中対話文と複数語訳が付与された日本語約 51000 項目の日中辞書を使用して対応付けを行なった。この対訳文集合から無作為に 100 文～ 23 万文の範囲で変化させた実験を 3 回行なった。評価文は、変換知識とは独立に 10 形態素以下の 1000 文を無作為に選んだ。

図 2に実験結果を示す。ここで、縦軸は翻訳文の出力率であり、訳質の評価は行なっていない。グラフにおいて、部分出力 ('partial') と完全出力 ('whole') の差異は主に文分割処理による貢献を、完全出力と換言処理を一切行なわない場合 ('no paraphrasing') の性能差は換言処理そのものの貢献、すなわち本稿で提案した協調処理による貢献と考えることができる。この図より、変換知識の非常に少ない場合でも換言処理の貢献によって約 20% の文を部分的に出力可能にしており、また変換知識の増大に伴って (実験の最大変換知識量程度では) 換言処理の貢献はむしろ増大している。

ところで、本研究で行なったテンプレート変換のような、形態素単位の客観的な一般化のみで得られる変換知識では、実験の最大規模 (23 万文) で 3 割程度しか未知文を翻訳できないことがわかる。すなわち、全文を出力させるためには残り 7 割の文に対する知識を何らかの手段で補わねばならないとも解釈できる。これに対して人手で作成した現在の換言知識と協調処理はこのうち約 4 割を補うことが可能であることを実験では示した。一方、このような機構なしで、すなわちコーパスから得られる情報のみを用いた汎化処理あるいは統計処理だけでこの 7 割を補うことは容易ではないと考える。すなわち、翻訳処理の際には、例えば本稿で示した換言知識のような原言語知識も重要なのではないか。

この実験において換言試行回数も測定した。この結果を図 3に示す。この図より、変換知識の増大に伴って換言試行回数は (対数的に) 単調減少していることがわかる。さらに詳細を観察すると、翻訳に成功しているものの多くは換言回数が 1 回または 2 回程度のもので、それ以上の換言を行なっているものの多くは最終的な翻訳結果の生成に失敗している。すなわち、換言回数は少数 (出力成功) と多数 (出力失敗) に大きく分かれているようであり、さらに分析が必要である。

4 結論

機械翻訳問題の多くを原言語の換言問題とするための機械翻訳モデルの一事例を示した。本稿の機構に従えば、大量で高品質の変換知識を期待できない多言語翻訳などの状況であっても、限定的な変換知識を原言語の換言知識である程度補うことが可能と考えている。現在実装している換言処理はまだ十分なものとは言えないため十分な翻訳精度を得ることができないが、扱う換言現象の増大に伴って翻訳精度も向上していくと考える。

本研究は通信・放送機構の研究委託により実施したものである。

参考文献

[Oht01] OHTAKE, K. and YAMAMOTO, K.: Paraphrasing Honorifics, In *Proc. of NLPRS2001 Workshop on Automatic Paraphrasing: Theories and Applications*, pp. 13–20 (2001).

[Yam01a] YAMAMOTO, K.: Paraphrasing Spoken Japanese for Untangling Bilingual Transfer, In *Proc. of NLPRS2001*, pp. 203–210 (2001).

[Yam01b] YAMAMOTO, K., SHIRAI, S., SAKAMOTO, M., and ZHANG, Y.: SANDGLASS: Twin Paraphrasing Spoken Language Translation, In *19th International Conference on Computer Processing of Oriental Languages (ICCPOL2001)*, pp. 154–159 (2001).