文書の種類を考慮した機械翻訳システムの構成

4K－8

島津美和子 熊野明 吉村裕美子 中村真理子
(株)東芝 研究開発センター

1.はじめに

従来、機械翻訳(MT)の用途は主にマニュアル翻訳に限定されており、実務翻訳であっても、文体・文法・構造の面で大きく異なる他の種類の文書には対応が不十分であった。MTは自動車、コンピューターといったマニュアル内における専門分野ごとの特殊な用語に各専門用語辞書を提供することで、訳語の統一という長所を活かしてきた。近年翻訳需要が高まってきているマニュアル以外の客観的文書(特許明細書、広報文書、放送文、新聞記事など)にもMTを利用するには、専門分野の指定の他に、文書の種類という視点を別次元で設ける必要がある。

今回我々は、この新たな視点を導入するにあたり、日英機械翻訳システムの従来の枠組をベースに、これまでに我々が開発したツールを活用しながら、目的言語の文書の種類(以降、ジャンルと呼ぶ)に応じた翻訳文を出力するには何が必要か、事前の比較言語的な検討を行った。本稿では、報道文と手紙文の2つのジャンルを例に、これらの間の日本語と英語の一般的な文書形態を分析し、MTの拡張に必要な開発項目を具体的な言語現象を交えて示し、ジャンルに適した訳文を出力できる方法を提案する。

2.従来方式の問題点

言語現象は一般的に成り立つものと分野やジャンルに特有なものとに分けられる。翻訳者が不得意分野の専門の文書を翻訳する際、一般の英語の文法知識に加え、専門書やスタイルマニュアルにあたって専門知識(語彙、文法)と特有の表現形態の習得に努める。MTも同様の知識追加が必要である。ここで、MTのモジュラー性を維持するために、システムを特化させずに、共通知識の上に別個の分野用・ジャンル別知識を利用する枠組が望ましい。従来はマニュアル以外の文書を翻訳する際、(1)専門用語辞書の開発、(2)一般辞書・文法規則の開発、(3)前編集作業によって対処していた。これらは各々、専門知識(語彙)の強化、専門知識(文法)の強化、一般的な表現形態への転換に相当する。

(1) 専門用語辞書の開発

市販の専門辞書から得られる情報には限度がある。翻訳者が適切な訳を判断できない時、既存の対訳文書を参考にするのと同じ発想でその分野での英語の文書から専門用語を抽出し、辞書に追加していく。この作業を自動化する試みが最近脚光を浴び、技術も向上している[4]。だが、用語辞書だけでは文レベルにとどまり、文書全体のスタイルを十分コントロールできないために(2)と(3)が必須となる。

(2) 一般辞書・文法規則の開発

マニュアルとの差異を吸収するため、現在主語なし文の訳し方や命令文の英語表現などの翻訳方式を指定するカスタマイズ変数[3]を設けているが、利用者が選択できるのは開発者が提供する項目に限られ、自由度が低い。また、分野別辞書の種類以外の項目は、選択に英語の知識を有していることが前提となっている。しかも文単位の調整しかできない。残りの部分は開発者に委ねられ、ジャンルに特有な構造を意識した訳出には利用できない。

(3) 前編集作業

人手による翻訳作業と同様、二段階から成る。第一段階は、日本語自体が悪文である場合の書き直しである。原文の日本語の意味が取れないのは、その専門性ゆえではなく、書いた本人しか分からない破格の文になっているためであることが多い。これをそのままMTにかけても良い結果は望めない。[8]などの文章の書き方の指針("Use definite, specific, concrete language." "Omit needless words." "Avoid succession of loose sentences.")に準拠して日本文を書き改める必要がある。

A Design of Machine Translation Systems Based on Text Types

Miwako SHIMAZU, Akira KUMANO, Yumiko YOSHIMURA, Mariko NAKAMURA TOSHIBA Corporation

第二段階は、文化的・社会的相違に起因し、日本語と英語ではその分野特有の慣習が存在し、双方に対応関係の見られないもの[1]に対処するための校正作業である。例えば、放送報道文の冒頭文(lead)は日本語では日付を用いるが、英語では曜日を用いるか、曜日は言わずに現在完了形で表現するのが通常である。こうした違いを念頭に置いて適切な英文作成に必要な情報を補う必要がある。しかし、現在の前編集[6]はシステムが解析に失敗しやすいと思われる文をシステムが処理しやすいように書き改めるのが主目的である。そこでは、利用者はMTの弱点を把握していなければならず、また実際の作業も負担がかかる問題があった。

このうち(1)の専門辞書構築については多くの先行研究があるため割愛する。一方、(2)と(3)での精度が翻訳文の質を直接左右することが分析結果から導けたので、以下はこれらに焦点を当てる。

3.ジャンルごとの分析

まず、日本語文書と専門家による英訳を比較し、日英間で相当する表現はあるが文法属性などの対応が不完全である現象、片方の言語にしか見られない現象を洗い出した。

(1) 報道文(図1)

報道文はleadと経過、背景説明から成る。構成部分ごとに照らし合わせると、次の対応関係が見られた。

A. 対応関係が不完全である現象

・leadの必須要素5WIHのうちwhenに相当する「10日」は訳さず、主動詞を現在完了形にすることで出来事が既に終わっていることを示している。・馴染みのない固有名詞は正式名称で出す必要が特にない場合、分かり易さを優先し、実体を示す簡単な名称にする。(「人民代議員大会」)・leadを簡潔にするため、日本語原稿にあった細かい事実は第2文に回す。(「事実上、人民代議員大会の解散を狙った」の部分は第2文で述べられている)・日本語では引用(〈2〉)になっていても、英語では間接話法にする。・日本語では繰り返し述べられていても、英語では一つに纏めている。(「3分の1以上の賛成が必要ですが」と「3分の1以上の賛成を得るのはむずかしいため」)・日本語では人名は姓だけでも良いが、英語では初出の場合、必ず名前も添える。(「エリツィン大統領」)

B. 片方の言語にしか見られない現象

・情報源(attribution)が明らかでない見解は省くことがある。(〈4〉の「議会との全面対決の姿勢を示したものと言えます。」)

(2)手紙文(図2)

手紙文は日本語では前文、本文、結語から成るが、英語では前文に相当するものはなく、本文から始まる。

A. 対応関係が不完全である現象

・動詞で示される動作のagentとrecipientは日本語では敬語、英語では格で表す。(〈4〉〈5〉〈6〉)・依頼の表現が英語では感謝の表現になっている。(〈6〉)

B. 片方の言語にしか見られない現象

・英語では主文の起こし言葉「つきましては」と前文の挨拶文〈2〉は不要である。

4.評価と検討

次に、機械翻訳の翻訳結果と模範訳を比較し、日本語表現をキーに目標訳を生成するには、個々の現象をどの段階で扱うべきかを検討した。訳文の質を忠実度と可読性に基づいて評価したところ、報道文は共に容認可能だが、手紙文は解析に失敗し、訳文だけでは意味が通じていない。これは現システムが「参ります」や「申し上げます」などの動詞の敬語表現には未対応であったためである。そこで、手紙文には$+\alpha$の文法を整備することが不可欠である。

4.1. 文法規則での対応

この部分を利用者に開放することは問題がある。3.で検出された事項は通常翻訳者がノウハウとして蓄積しているか、スタイルックの形で纏められている。現在、文書中の言語現象から翻訳規則を導く研究[2]も行われているが、スタイルブックの利用も考えるべきであろう。これをもとに開発者が知識の追加を行い、ここで新たに作成した文法規則は個々の現象をジャンルごとに指定し、適用できるようにする。これを図式化したのが図3である。

さらに、翻訳方式の指定は、文書全体の構成を考慮してなされると利用者の便宜に供するであろう。つまり、翻訳処理の前に、あらかじめ各文の表す内容が見解か事実か、段落が本文か前書きかなどを識別し、翻訳処理でこの情報を使うのである。なお、この識別技術の開発の問題はSGMLなどの構造化文書の普及で解決できそうである。これにより、2章の(2)で指摘した問題が軽減すると思われる。

4.2. 前編集での対応

3.で見た現象を整理すると、今までカスタマイズ変数として動詞のテンスなどは指定はあったが、実際訳を作る場合を考えると、動詞や名詞の型は文書全体ではなく、個々の文に応じて指定ができると便利なことが分かった。また、各システムとも、翻訳不要の指定はあるが、これは原文そのものを訳文側に出力するものである。そうではなく、削除するという選択肢もあってよい。このように、2章の(3)の問題は、選択肢の数を増やしたり、適用の範囲を指定できるようにするとある程度緩和されると思われる。この他、報道文・手紙文共通で改良を行うべき重要な言語事象は、前置詞である。日本語の助詞と英語の前置詞、助詞で示される日本語の格と英語の格は類似していることはあっても、確実な対応関係はない。これには、日本語文を意味解釈した上で、文書内から語の用法を抽出して得られる文書内情報が、適切な英文生成に利用できると考えている[5,7]。

5.まとめ

これまでに開発してきたツールをフルに利用し、本稿で示した指針に従って新規のジャンルのための開発を行うことで、モジュール性を保ちつつ、その分野・ジャンルに適した訳出を行うことができるMTの構成を明らかにした。今後、この構成をシステムに実際に取り入れ、評価を行っていく。

6.参考文献

1. 池原他,「言語表現体系の違いに着目した日英機械翻訳試験項目の構成」,『人工知能学会誌』, vol. 9, no. 4, pp. 569-579, 1994
2. 宇津呂他,「二言語対訳コーパスからの動詞の格フレーム獲得」,『情報処理学会論文誌』, vol. 34, no. 5, pp. 913-924, 1993
3. 熊野他,「機械翻訳文法のカスタマイズ」, 情報処理学会NL研, NL-84-11, pp. 79-86, 1991
4. 熊野他,「言語情報と統計情報を用いた対訳文書からの機械翻訳辞書作成」, 情報処理学会NL研, NL-100-12, pp. 89-96, 1994
5. 島津他,「機械翻訳における翻訳対象文書内の情報の利用」, 人工知能学会研究会資料 SIG-SLUD-9301, pp. 35-42, 1993
6. 野村他,「機械翻訳前編集支援ツールの開発」,情報処理学会45回全国大会、vol. 3, pp. 91-92, 1992
7. James Pustejovsky et al. "Lexical Semantic Techniques for Corpus Analysis," Computational Linguistics, vol. 19, no. 2, pp. 331-358, 1993
8. William Strunk and E. B. White, The Elements of Style, MacMillan, 1979

ロシアのエリツィン大統領は10日、現在の保守、中道派が優勢な人民代議員大会は改革を妨害し、クーデターを企てていると非難し、事実上、人民代議員大会の解散を狙った国民投票を来年1月に実施するよう提案し、議会との全面対決の姿勢を示しました。〈1〉エリツィン大統領は、「大統領の進める改革路線を選ぶのか、人民代議員の進める過去への逆戻りの路線を選ぶのか来年1月に国民投票を実施するよう提案する」と述べました。〈2〉国民投票を実施するには人民代議員の3分の1以上の賛成が必要ですが、エリツィン大統領を支持する勢力は少数派で、3分の1以上の賛成を得るのはむずかしいため、エリツィン大統領は国民投票を実施するもう一つの方法として、100万人の国民の署名を集めるよう支持者に呼びかけました。〈3〉

今回のエリツィン大統領の国民投票の呼びかけは、国民に自分に対する信任を問う形になっていますが、事実上、保守・中道派が優勢な現在の人民代議員大会の解散を狙ったもので、議会との全面対決の姿勢を示したものと言えます。〈4〉

Russian President Boris Yeltsin has proposed holding a nationwide referendum in January next year, accusing Congress of hampering his quest for reform and attemping a coup. He is aiming to dissolving Congress, in which conservative and middle-of-the-road legislators predominate.

The referendum will enable the Russian people to decide whether they support President's reform plan or Congress's desire to go slower.

Holding a referendum would require either the approval of over one-third of the legislators in Congress or a petition signed by one-million Russian people. President Boris Yeltsin has asked his supporters to organize a petition because his supporters in Congress are in a minority.

図1 報道文の例

拝啓〈1〉ますますご隆盛の趣お喜び申し上げます。〈2〉

さて、この度香港の株式会社A社は、当社製品をD/A条件で出荷してほしいとの希望を寄せ、同社の財政面での堅実さに関しての照会先として貴行香港支店の名を挙げて参りました。〈3〉

つきましては、同社の財政状態及び取引の程度に関し内密に御意見をお聞かせ下されば誠に幸甚でございます。〈4〉

なお、調査をお願いいたしますことについては固く秘密を守り、ご迷惑をおかけしないことをお約束いたします。〈5〉

なにとぞよろしくお願い申し上げます。〈6〉　敬具〈7〉

Dear Sir,

A Co. in Hong Kong has sent us a request for supplies of our products on D/A terms, and given us the name of your branch office in the same locality as a reference in regard to their financial stability.

We should be greatly obliged if you would favor us with your confidential opinion as to their financial standing and the scope of their transactions.

Any information you may give us will be treated as strictly confidential. We thank you for your assistance.

yours very truly,

図2 手紙文の例

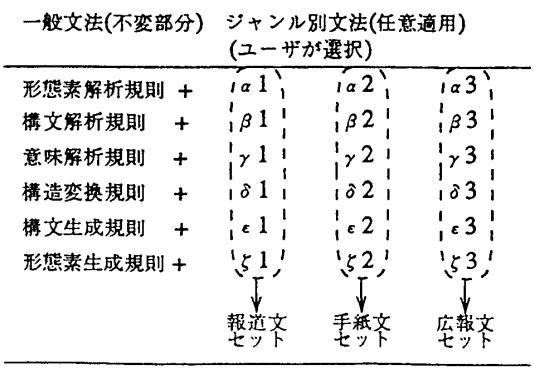

図3 文法規則の構成