第三言語翻訳システム

井佐原均
独立行政法人 通信総合研究所
isahara@crl.go.jp

１．はじめに

コンピュータ・ネットワークが世界をつなぎ、高度情報化がますます進む中で、いかに情報を得ることができるかが、各自が人間的な生活を送るための重要なポイントとなる。いわゆる主要国の発信する、あるいは主要言語で記述された情報は質・量ともに膨大なものであるが、それらが、他の多くの国々、多くの言語で共有されるかといえば、心もとない。一部の語学堪能者だけではなく、その国の多くの人々が高度情報化社会の恩恵をこうむるためには、母国語での情報の受信が必須であるが、現状ではそれは難しい。経済的な観点からは翻訳や言語処理の研究が成り立たないような国の人々が母国語で外国との情報交換をすることは将来的にも困難と思われる。

情報獲得は、母国語で行うことが望ましいが、人手による翻訳はコストが高く、必要な情報をすべて人手で翻訳することは、コストの面でも即時性の面でも不可能である。機械翻訳の精度はまだまだ不十分であり、有効とされる場面は限定されている[1]。一方、英語を効率よく学ぶ手段についての研究も行っている[2]が、全ての人々が英語で自在にコミュニケーションができるようになるとは考えにくい。しかしながら、社会が健全に発展し、全体としての生産性が向上し、物心両面での満足度が高まるためには、母国語による情報は必須であり、現実的な資源を用いて、主要言語で書かれた情報を多くの言語に翻訳する技術が必要とされよう。

２．第三言語翻訳の提案

本稿では、現実的な資源と新しい手法を組み合わせることにより、実社会で有効な高精度の翻訳を実現する方法を提案する。具体的には、日本語と英語の対訳文書を入力し、別の言語の高品位の文書を生成する技術であり、これにより、デジタルデバイドに悩む多数の国家との情報共有が可能になることが期待される。

従来の機械翻訳の課題は、解析における課題と生成における課題に分けることが出来る。

解析における課題(解析能力の不備)は文脈を用いずに文レベルで入力を処理することの限界とも言える。ひとつの文を見ただけでは十分に意味が取れず、前後の文脈を見てはじめて意味がわかるような文が存在するが、現在の自然言語処理技術は、このような文脈を扱う能力は不十分である。この解決策として、対訳を入力として用いる。つまり言語によって、明示される情報が異なることを積極的に利用するのである。別の見方をすれば、システムは文脈に頼って必要な情報を得るのではなく、人間が文脈を利用して注意深く一つの言語から他の言語に翻訳した対訳文から必要な情報を得ようとするものである。このようにして対訳文書から得られるものは主として内容面の情報である。

一方、生成側の課題は、解析結果（内容面での情報）に基づく目標言語での文生成の精度向上ということになるが、ここでは目標言語の単一言語データから言語固有の情報を獲得することを考える。

さて、これらの処理は、現実的な資源を用いて行われるものでなくてはならない。日英対訳データは、出版社や企業によって、高品位のものが今後とも継続的に人手をかけて作り出されると想定できる。ただし、この多くは単純なテキストであり、詳細な言語情報(構文的・意味的情報)が付加されるということは、少なくとも現時点では考えにくい。

一方、非主要言語に対して、このようにコストをかけた高品位の人手翻訳が大量に行われるとは考えにくい。アジア諸国の言語に代表される目標言語に関して、現地の研究機関と共同で言語資源[3]やツール[4]の開発を行ってきたが、その量はまだまだ不十分である。各言語について期待できる電子化された言語資源は、新聞や報告書等の単言語の生のテキストが中心であろう。この資源を使って、目標言語の特徴を得ることを試みる。なお、これらの資源のほかに、少量ではあるが対訳辞書(一般の英タイ辞書など)が変換規則に関する情報として使用できよう。

3．要素技術

機械翻訳システムの精度が上がらない大きな理由の一つは、言語解析の困難さによる。従来の自然言語処理技術は、ひとつの文を読んで、それを翻訳したり、要約したりするという、通常人間が行うであろう行為を模擬するものであった。しかし一方で、計算機に文脈を扱わせる技術の確立が困難であるという致命的な欠陥があった。本研究では、単言語の文書ではなく、たとえば日本語と英語の対訳文書を入力として用い、そこから情報を和や積の形で取り出し、深い意味理解を実現する。その理解結果を元に、目標言語（例えばタイ語）の単一言語コーパスから各言語固有の情報を得て、表層の文章を生成する。(図1)

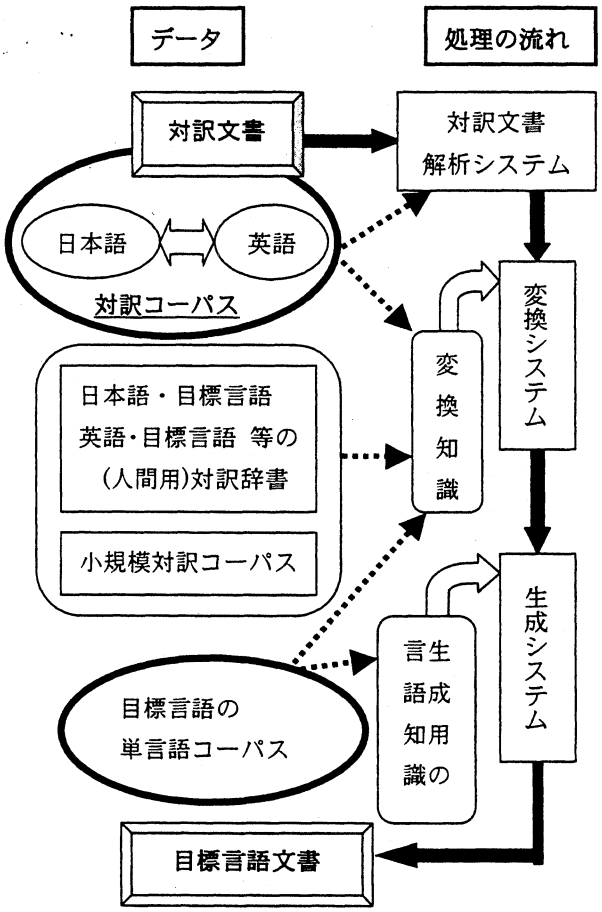

図1　第三言語翻訳システム

1）2言語解析技術

解析の困難さというのは、曖昧性の解消ができないということであるが、対訳を用いることにより、曖昧性の解消が可能になる場合がある。たとえば、日本語だけを見ていては、あるものが複数であるかどうかはわからないが、英語を見れば、その語が単数形か複数形かで判断できる。一方、英語ではその語の意味的役割がわからないが、日本語では助詞がついているので、たとえば「場所」をあらわす情報であるということが分かることがある。これは日本語と英語のように言語の体系が大きく異なる言語同士を利用することで、特に有効となる。

構文解析においては日英2言語の対訳の入力を前提に、語と語(あるいは日本語の文節のようにもう少し大きい単位)の間の依存関係を解析する。このモデルは二つの語(あるいは文節)が依存関係にあるかないかを機械学習する[5]。依存関係は学習されたモデルによって計算される確率の積が一文全体で最も高くなるように決定するが、この決定に際し、日英双方からの確率値を利用することが可能となる。

なお、本稿ではシステムが利用する2言語(例えば日本語と英語)のデータを対訳文書と表現しているが、その対応の強さによって、いくつかの段階がある。文レベルの対応が見られるparallel corpus から、同内容を扱うcomparative corpus、さらには同じ分野の文書を集めたものなどがある。Parallel corpus は元の言語で書かれた内容を確実に別の言語に移すことを目的としており、対応関係は強い[6]。このような文書は政府刊行物や取扱説明書などで着実な増加が期待できる。一方、同内容あるいは類似内容の文書の対は、新聞の英文紙と和文紙との関係などに見られ、情報源および研究用データとして大量の文書が利用可能である。これらの中間のものとして、小説などのように、意訳された文書が存在する。

文レベルの対応が少ない場合には、複数言語それぞれの文書群の中から、対応する内容をもつ文書を検索し、対応付ける技術が必要となる。

2) 単言語コーパスからの言語情報の獲得

生成に関する研究は、従来あまり系統だって行われてこなかったが、作成した文書を人間が直接読む場合、その精度は人間の「読もうとする意欲」に直結する。ここでは、2言語処理によって得られた解析結果の精度向上を踏まえて、生成の精度向上を行う。単言語コーパスから単語の用法に関する情報を得る技術と、構文構造に関する情報を得る技術を開発する。

2言語の情報を用いて理解された結果を第三言語の文にする場合には、語順・訳語の決定といった目標言語自体での課題が存在し、当然その言語についての知識が必要となる。生成される文章の質の向上のためには、このような言語固有の情報を得る必要がある。しかし、これをその言語の研究者の持つ言語直観によって規則化していくのは、膨大な作業であり、主要な言語以外でこのような作業を行うということは現実的ではない。そこで、本システムでは、個別の言語についての情報は、個別の言語の単言語データを元に自動獲得することを試みる。

解析過程で得られた語と語の依存構造からの自然な並びの表層文生成を試みる[7]。自然な並びであるかどうかは、語順モデルによって決定する。このモデルは同じ語を修飾する複数の修飾語があるとき、修飾語間での自然な順序を学習するもので、機械学習モデルを用いて実現される。人手をかけない学習によって、従来のぎこちない計算機生成の文書を、コーパスに示される実際の文章での流暢さに基づいたレベルにまで向上させることが可能となろう。

3) 対訳データから第3言語への翻訳知識獲得

計算機を用いてある言語の情報を別の言語に変換するためには計算機処理に適した形態の規則が必要である。これらを人手で作ることは、やはり入出力言語を理解する専門家による膨大な作業を必要とするため主要言語対以外で行うことは現実的ではない。また、大量の対訳コーパスからこれらの言語情報を自動獲得する手法も提案されているが、主要言語対以外では大量の対訳コーパスを前提とすることはできない。

本研究では翻訳元である2言語の対訳コーパスと翻訳先言語の単言語コーパス、および、翻訳元言語と翻訳先言語との間の小規模対訳辞書等を組み合わせることによって、コーパスの言語間対応づけや、言語情報の獲得を目指す。

4）変換に基づかない生成

図１では、対訳入力で解析した結果は変換システムを介して、目標言語化され、それを高品位の文書にすることが想定されている。ここで、Comparative corpus を入力とした場合を考えると、英語において表現されるべき内容と日本語において表現されるべき内容から情報を得、それを元に第三言語において表現するべき内容を提示することになる。ここでは、解析結果として得られた情報（キーワード）から直接文章を生成するという手法［8］が有効となるのではないかと考えている。

4．これから

ここで述べた第三言語翻訳システムは、従来の機械翻訳を超えた実用的な精度で目標言語の文書を生成するため、人手により作成された高精度の日英対訳文書から内容面での情報を得、対訳辞書等から変換規則を得、目標言語の文書から言語的特徴を得て、目標言語の的確な文章を生成する技術である。

別の視点から見れば、この提案は、既存の、あるいは今後とも順調に増加するであろう、日本語と英語の対訳文書を利用し、この文書の持つ情報に他の国が母国語でアクセスできるような手段を開発するものである。これにより、発展途上国等への母国語による情報提供が可能になる。また、この手法が確立すれば、新しい言語への対応は、その言語に関する言語情報の獲得が主たる開発要素となり、それぞれの国でも対応できると思われる。

ここまで、日英対訳文書から第三言語への翻訳という視点から述べてきたが、この手法を用いて、英語の文書を日本語化することも可能である。英語の文書を人手翻訳のコストの安い国で、その国の言語に人手翻訳し、その言語と英語を入力として、日本語を生成すれば、比較的安価に実用的な英日翻訳が実現できよう。

この研究開発は、通信総合研究所けいはんな情報通信融合研究センターにおいて、広く民間企業等との連携で行う予定である。計算機科学のみでなく言語学の知見をも積極的に取り入れる予定である。また通信総合研究所が開設予定のバンコク研究センター(仮称)において、生成側のアジア圏言語の研究開発を行う予定である。

参考文献

［1］富士秀 他:機械翻訳システムの有効性の評価－どのような人にとってMTは役立つか－，言語処理学会第８回年次大会，2002.

［2］齋賀豊美 他:日本人英語学習者コーパス作成とその利用可能性，言語処理学会第８回年次大会，2002.

［3］井佐原均 他：ORCHID: Building Linguistic resources in Thai, Literary & Linguistic Computing (J. of the Association for Literary and Linguistic Computing) Vol. 15, No. 4, 2000.

［4］馬青，井佐原均：長さ可変文脈を用いたマルチニューロタガー，自然言語処理，６巻１号，pp. 29-42，1999.

［5］内元清貴，関根聡，井佐原均：最大エントロピー法に基づくモデルを用いた日本語係り受け解析，情報処理学会論文誌 40巻 9号，1999.

［6］井佐原均，春野雅彦：Japanese-English aligned Bilingual corpora, Parallel Text Processing - Alignment and Use of Translation Corpora, J. Veronis (ed.), KLUWER, 2000.

［7］内元清貴 他：コーパスからの語順の学習，自然言語処理 ７巻 ４号，2000.

［8］内元清貴，関根聡，井佐原均：キーワードからのテキスト生成，言語処理学会第８回年次大会，2002.