機械翻訳精度を向上させる可読性診断技術

東芝ソリューション株式会社　プラットフォームソリューション事業部参事　**熊野　明**

PROFILE　1982 年東京工業大学卒業。同年東京芝浦電気（株）（現（株）東芝）入社。2010 年から東芝ソリューション（株）プラットフォームソリューション事業部。自然言語処理システムの研究開発に従事。アジア・太平洋機械翻訳協会理事。AAMT/Japio 特許翻訳研究会委員。2007 年度から特許版・産業日本語委員会委員。

東芝ソリューション株式会社　IT 技術研究所主任研究員　**加納　敏行**

PROFILE　1992 年名古屋大学卒業。同年（株）東芝入社。2006 年から東芝ソリューション（株）IT 技術研究所。自然言語処理システムの研究開発に従事。

1 はじめに

現代社会の生産活動において、自然言語で記述する文書を作成、蓄積、検索、理解、再利用、さらには変換する必要性は、日々高まっている。変換された文書は、さらに蓄積され、もとの文書と同様に、検索や理解の対象となる。つまり、文書の目的は、情報の記録だけではなく、情報の伝達にある。今日、このあらゆる過程で、計算機による支援が行われている。

この情報伝達に要するコストを下げることは、日々の生産活動を効率化することにつながる。このためには、情報を伝達しやすい文書、すなわち理解しやすい文書が必要である。

Japio が提唱する産業日本語の目指すものは、この理解しやすい文書の作成である。それを具体化したものが、特許版・産業日本語であり、現在ライティングマニュアルの整備が進行中である。さらに、人間にとっての理解しやすさ、わかりやすさだけではなく、計算機支援における処理しやすさが、生産活動の効率を向上させる。具体的には、機械翻訳の精度を向上させることが示されている［1］。

我々は、理解しやすい日本語の作成を支援する技術として、可読性診断技術の研究開発を行ってきた。本稿では、この技術が、産業日本語の編集を支援するだけでなく、機械翻訳の精度向上を支援できることを示す。

2 文書の可読性

文書の理解しやすさにはいくつかの種類があるが、計算機処理を考慮すると、以下のレベルに整理することができる。

（1）用語・表現のレベル

文書に使われる個々の用語やその表現形式に関する理解しやすさである。単語、もしくは、句の単位が判断の対象となる。自然言語処理の要素技術としては、辞書検索や形態素解析が利用できる。

（2）構文のレベル

文の構造的な理解しやすさである。一般には 1 文全体が判断の対象となる。文の一部である節だけでも判断の対象となりうる。自然言語処理の要素技術としては、構文解析が利用できる。

（3）意味のレベル

個々の単語や文のレベルを超えた、文書全体としての理解しやすさである。文書全体、あるいは段落のまとまりで判断されるものである。自然言語処理の要素技術としては、意味解析が利用できる。

我々は、業務文書に対する可読性という観点で、その

判断基準や対応方法などを検討してきた。この可読性は、産業日本語の理解しやすさと共通する性質が多い。そこで、産業日本語が目指す機械翻訳の可能性を考慮しつつ、産業日本語の要素技術として利用できるよう考えた。

3 可読性診断技術

人にとって理解しにくい文は、ほとんどの場合、機械翻訳などの計算機処理に対しても処理しにくい文である。その中でも、用語や表現が要因となるものは、汎用のソフトウェアで指摘できる。また、意味的なレベルの分りにくさは、現象が個別的であり、規則化が容易ではない。それに対して、構文的な要因によるものは、人と計算機の両方に影響を与えており、技術的にも検出の可能性が高いと判断した。

3.1 可読性診断の処理

診断対象の文書に対して、可読性診断は以下の５段階の処理を行う［2］。

（1）文の切り出し

診断対象文書から１文を切り出す。

（2）構文解析

切り出した文に対して、構文解析処理を行う。

（3）可読性診断

構文解析結果に対して、可読性診断機能群の機能によって、文の理解しにくさを検出する。詳細は次節で説明する。

（4）診断結果のまとめ

可読性診断機能群で診断された結果に対して、指摘された要因の関係により、本質的な要因を優先して指摘するよう、出力結果をまとめる。

（5）診断メッセージの生成

まとめられた診断結果に基づいて、利用者に出力する診断メッセージを生成する。

3.2 可読性診断機能

現行のシステムが指摘する可読性に対する診断機能を表１に示す。

表１　可読性診断の機能

診断機能	内容
曖昧な係り受け	係り受け関係の解釈が複数ある箇所を指摘する。
複合語	辞書に登録されていない複合語で、直訳では訳しにくい接辞（「可」、「未」、「無」など）を含む複合語を指摘する。
	例：ペット可賃貸マンション、文書管理システム未導入部門
述語の省略	述語動詞が省略されている箇所を指摘する。
主語の省略	主語が省略されている述語動詞を指摘する。
目的語の省略	目的語が省略されている述語動詞を指摘する。
主語と述語が離れている	述語動詞から離れている主語を指摘する。
目的語と述語が離れている	述語動詞から離れている目的語を指摘する。
長い修飾部	修飾部が長い箇所を指摘する。
述語の数	述語動詞が多く含まれている文を指摘する。
助詞「は」	格の曖昧性がある助詞「は」を指摘する。

以下では、述語の省略診断を例にして、診断機能の仕組みを述べる。

述語の省略は、並列表現において連用中止形で表される述語、あるいはその活用語尾、および前接する名詞句の格助詞が省略される現象である。同じ述語の繰り返しを避けるために用いられる。述語の省略によって、それに前接する名詞句の係り先がなくなり、誤った解釈を導きやすい。

述語の省略の形式は、省略部分の形態により、次の３種類に分類できる。

（1）名詞で中止

連用中止の述語とその直前の助詞が省略されたもの

［省略例１］

顧客用文書をカラー、社内用文書を白黒で印刷する。

（2）助詞で中止

連用中止の述語だけが省略されたもの

［省略例2］

顧客用文書をカラーで、社内用文書を白黒で印刷する。

（3）動詞語幹で中止

連用中止の述語（サ変動詞）の活用語尾が省略されたもの

［省略例3］

顧客用文書をカラーで印刷、社内用文書を白黒で印刷する。

述語の省略診断は、格要素である名詞句と述語の接続関係、読点の情報、副助詞の情報を合わせて、述語の省略の有無、省略の形式、補足すべき情報を診断する。

省略例1の文を例にして、処理の詳細を述べる。

（1）構文解析による認識

構文解析処理では、同じ格助詞の繰り返しをキーとして、述語の省略が認識できるケースがある。このレベルで検出できるのは、文中の名詞句がシンプルなケースに限定される。まず、構文解析結果からこのパターンで述語が省略できるか検査する。省略例1は、この述語の省略は認識できないので、次の処理に進む。

（2）副助詞による判別

日本語の並列表現は、一部の副助詞（“は”、“では”、“でも”など）の繰り返しによって表されることがある。同様な副助詞が複数使われ、読点によって分割できる場合、並列表現による述語の省略と認識する。

省略例1では、この副助詞がないので、次の処理に進む。

（3）目的語による判別

単独の述語に、格助詞が同じ複数の語が係る文は、文法的に正しくない。本来このような解釈は正しくないが、このような現象が認識できた場合は、連用中止の述語が省略されている可能性が高い。省略例1では、述語“印刷する”の前に、“顧客用文書”と“社内用文書”が同じ格助詞“を”を伴っているので、この診断対象となる。

（4）省略の形式の判別

読点の直前にある文節の品詞、および付属語により、省略の形式を判別する。省略例1では、読点の直前の文節（“カラー”）の品詞が普通名詞であり、付属語がないことから、省略の形式として“名詞で中止”と判断する。

（5）修正候補の推定

述語の省略は、重複を避けるために、文末述語と同じ述語が省略されることで生じる場合が多い。省略例1の場合、文末述語“印刷する”の連用形“印刷し”と、文末述語に係る助詞“で”が省略された推定する。

これらの処理によって、省略例1に対して次のような診断メッセージを出力する。

“カラー”の直後に“で印刷し”が省略された可能性がある。

4 可読性診断を利用した機械翻訳精度の向上

機械翻訳の誤訳の原因は、原文に内在する曖昧性を正しく解消できないことであることが多い。我々日本人が丁寧に書いた文でも、計算機にとっては曖昧性があり、正しく解釈できないことがある。曖昧性には、語レベルのものもあるが、ここでは構文レベルのものを考える。代表的なものには、係り受け判定や、並列句認定がある。一般に、1文中に動詞が複数あれば、構文レベルの曖昧性が存在すると考えてほぼ間違いない。

4.1 高コストな後編集

曖昧性を解消できない場合、機械翻訳ができるのは次の2つの手段である。

（1）すべての解釈を提示する

係り受けや並列句の解釈を、可能な限りすべて提示する。機械翻訳の場合、係り受けの解釈を分りやすく提示するのは容易ではない、というより、提示した解釈を利用者が理解することは一般に困難である。したがって、可能な限りの曖昧性をすべて保存し、それらに対して訳文を生成し、すべて提示する必要がある。曖昧な係り受けが1～2箇所で、出力可能な訳文が数例なら、その中から正しいと思うものを選択することは可能だろう。し

かし、曖昧な係り受けが４箇所を超えると、可能な訳文は数十通りとなり、そのすべてを提示し、その中から正しいものを選択するのはコストが高く、機械翻訳の後編集作業として現実的ではない。

図１に、すべての解釈を示す場合の翻訳過程を示す。

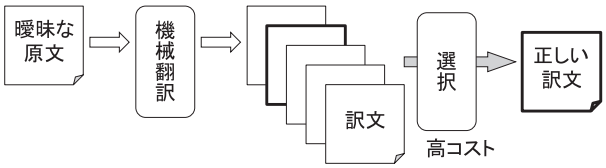

図１　すべての解釈を提示する場合の翻訳過程

（2）一つの訳文を提示する

複数の解釈の中から、システムが最も確からしいと判断した結果に対して、訳文を提示する。現在の多くの機械翻訳システムは、この方法を用いている。

提示された訳文の解釈が正しければ問題はない。しかし、1 か所でも正しくない解釈を反映した訳文であれば、正しい訳文を得ることはできない。正しくない訳文を編集して、本来の訳文を作成する必要がある。

図２に、一つの訳文を提示する場合の翻訳過程を示す。

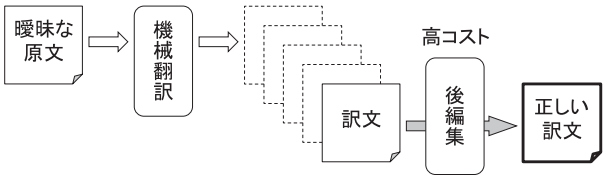

図２　一つの訳文を提示する場合の翻訳過程

省略例１に対する、日英機械翻訳、日中機械翻訳の出力例を示す。これは、システムが解釈を１つ選んだ結果である。省略部分の解釈を誤り、正しくない訳文を出力している。

［省略例１の英訳例］

Color and the document for in-house use are printed for the document for customers by black and white.

［省略例１の中国語訳例］

在黑白印刷顾客用文书彩色，公司内部用文书。

いずれの場合も、出力された訳文を後編集して正しい訳文を得るには、必要となるコストが大きい。

４.２　前処理によるコスト低減

上述した後編集のコストを削減するために、これまで述べた可読性診断を利用することを考える。

曖昧性のある原文に対して、可読性診断を行う。ここで、係り受け曖昧性などが検出できると、可読性診断のメッセージ出力機能を利用して、利用者に知らせる。この時点では、まだ訳文を出力しない。

診断メッセージによって、可読性を損なう現象、すなわち曖昧性のある箇所が認識できると、利用者はその曖昧性を解消するよう、明瞭な文に修正することができる。利用者自身が作成した原文であるなら、その文で伝達したい内容は明らかであり、その意味を保持するように原文を書き換えることには、大きなコストを要しない。

図３に、可読性診断によって前編集する場合の翻訳過程を示す。

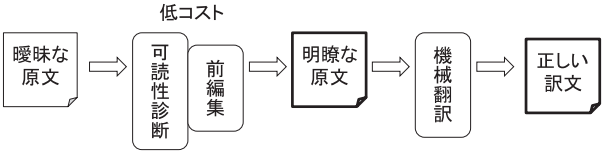

図３　可読性診断で前編集する場合の翻訳過程

省略例１に対して利用者は、３章で示した診断メッセージをもとに、曖昧な日本語文を明瞭な日本語文に修正する。修正した日本語文と、それに対する、日英機械翻訳、日中機械翻訳の出力例を示す。

［修正例］

顧客用文書をカラーで印刷し、社内用文書を白黒で印刷する。

［修正例の英訳例］

The document for customers is printed in a color and the document for in-house use is printed by black and white.

［修正例の中国語訳例］

以彩色印刷顾客用文书，以黑白印刷公司内部用文书。

ここで示したように、利用者の原文修正によって曖昧性が解消すれば、機械翻訳は正しい解釈による正しい訳文を出力することが可能になる。仮にまだ他の曖昧性があれば、再度可読性診断を実行し、新たな診断結果にしたがって、再度原文を修正すればよい。

機械翻訳を利用して正確な訳文を作成するには、対話的な操作が不可欠である。これまでは、機械翻訳が出力した訳文に対して後編集するため、コストの高い作業を必要としていた。これに対し、可読性診断技術の実現により、その後編集作業を低コストの前編集で置き換えることが可能になってきた。結果的に、機械翻訳による訳文作成の精度を向上させることができる。

5 おわりに

ここで述べた可読性診断技術の判定基準は、特許版・産業日本語ライティングルールとの共通性が高い。そこでこれまで、産業日本語を実現する手段としてその可能性を示してきた［3］。

しかし、この技術は日本語文を診断することによって、原文の解釈の可能性を示し、修正を促すものであり、原文を直接書き換えるものではない。利用者の簡単な編集操作によって、理解しやすい日本語文を作成することが目的である。したがって、ユーザインタフェースが重要である。機械翻訳や、産業日本語ライティングの精度、効率を向上させるために、インタフェースを改良する必要があると考えている。

今後は、特許版・産業日本語の特許ライティング支援環境の一技術として、可読性診断技術の有効性の評価を進める。その結果に基づき、技術の改良を進める予定である。

参考文献

［1］熊野　明他：産業日本語の構想と特許文の言い換え実験、情報処理学会第 190 回自然言語処理研究会（2009）

［2］祖　国威他：構文的特性に着目した可読性診断技術、東芝レビュー Vol.66 No.4（2011）

［3］東芝ソリューション：可読性診断技術、第３回産業日本語研究会・シンポジウム予稿集（2012）