構文的特性に着目した可読性診断技術

Document Readability Diagnostic Technology Focusing on Syntax Characteristics

祖　　国威　　吉村　裕美子　　加納　敏行
■ZU Guowei　　■YOSHIMURA Yumiko　　■KANO Toshiyuki

業務文書は，企業の活動における情報の共有や伝達の基本となるメディアである。したがって内容が正確であるだけでなく，その内容が読み手に正しく効率的に伝わることが望ましい。そのためには，読み手にとってわかりにくい記述を減らし，業務文書の品質を確保することが求められる。

東芝ソリューション（株）は，文の構文的なわかりにくさを診断し，わかりにくい文を検出する可読性診断技術を開発している。この技術は，翻訳の前処理をはじめ，文書品質を向上させるための支援技術として活用できるもので，例えば翻訳などにおいて，作業コスト増加の要因となる，わかりにくい文を検出することができる。

As business documents are the basic media for sharing and distributing information in corporate activities, correct and efficient transmission of the contents to readers as well as correctness of the contents are required. To meet these requirements, it is necessary to eliminate ambiguities and maintain the quality of business documents to the greatest extent possible.

Toshiba Solutions Corporation has been developing a technology to evaluate the readability of sentences applying syntactic analysis of sentences and to identify unclear sentences that increase the workload of document translation. This technology can be used to support improvements in the quality of a wide range of documents, particularly for pre-editing in human and machine translation.

1　まえがき

業務文書は，業務マニュアルや業務活動に伴って発生する情報などを文書化したメディアであり，顧客やパートナー，他部門との間で，業務情報の伝達，共有，及び記録のために中心的に使われる。したがって，内容が正確であるだけでなく，その内容が読み手に正しく効率的に伝わることが望ましい。そのためには，読み手にとってわかりにくい記述を減らし，業務文書の品質を確保することが求められる。

業務文書の品質にはいくつかの観点があるが，代表的なものとして記述された文章のわかりにくさがある。わかりにくい文章は，記述された内容が読み手に誤解される危険性が大きく，内容を理解するのに時間が掛かる文章である。東芝ソリューション（株）は，文章のわかりにくさを診断し，わかりにくい箇所を検出する可読性診断技術を開発している。

ここでは，可読性の問題及びその診断技術と，研究成果の適用先として機械翻訳の前編集支援を想定した可読性診断システムの評価と実現の可能性について述べる。

2　業務文書における可読性問題

業務文書の場合，限られた時間内に文書の情報を正しく把握できることが望ましい。例えば，海外との事業では，多数の業務文書を翻訳する必要があり，しばしば翻訳者に翻訳を依頼することになるが，翻訳者がその事業や事業分野に精通しているとは限らない。そこで，翻訳者が文書の内容をまちがいなく正しく理解できるような文章が求められる。わかりにくい文章は，誤訳を生じさせ，翻訳された文書の修正や再翻訳などの作業増加を招く。また，原文の意味を理解するために翻訳者から多数の質問を受け，対応に時間を取られることもある。わかりにくい文章は，翻訳業務の作業量と時間を増加させることになる。

2.1　機械翻訳への影響

機械翻訳でも，わかりにくい文章を入力すると良い翻訳結果は期待できない。当社は，機械翻訳技術に基づく日英，日中翻訳を，機械翻訳製品やサービス（The翻訳$_{TM}$シリーズ，Eiplaza$_{TM}$/MT（Machine Translation））として提供している。これらの機械翻訳製品やサービスを用いると，いわゆる簡潔でわかりやすい日本語文章に対しては，実用レベルに達した翻訳結果が得られる。

しかし，翻訳者にもわかりにくい文章では，正しい翻訳は非常に難しくなり，翻訳の精度が低下する。機械翻訳の場合，人間より言語理解の能力が低いので，正しい翻訳のために翻訳者以上に，わかりやすい文章が求められる。そのため，利用者から“良い翻訳結果を得るためのチェックや校正ツールはないか”との声も多く寄せられていた[(1)]。

機械翻訳にとってわかりにくい文とその誤訳の例を以下に示す。

［例文1］

顧客用文書をカラー，社内用文書を白黒で印刷する。

［日英翻訳結果］

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

［日中翻訳結果］

在黑白印刷顾客用文书彩色，公司内部用文书。

例文1の場合，“カラー”の後に“で印刷し”が省略されたことにより，機械翻訳は“カラー”の係り受け関係を正しく認識できず，原文とは異なる意味の文に翻訳してしまう。プリンタの機能として“カラー”と“白黒”の選択があることを知っている人にとっては，その知識で補完することによってこの文を正しく解釈できるが，機械翻訳では，“カラー”と“白黒”の関係を捉える前に，“カラー”とその直後の“社内”や“社内用文書”との並列関係も捉えてしまい，誤訳を導いている。

このようなわかりにくい文を，事前診断によって発見し，作成者にフィードバックし修正を促すことによって，人手翻訳や機械翻訳の精度と効率の向上が期待できる。

表1．文章の分かりやすさを低下させる要因

Main factors reducing document readability

要　因	文章事例	説　明
述語の欠落	顧客用文書をカラー，内部用文書を白黒で印刷する	目的語“顧客用文書を”の述語が省略されている
複合語中の接辞	文書管理システム未導入部門	複合語に要注意接辞“未”を含むため，誤解析の可能性がある
係り受けが曖昧	昨日更新されたデータが削除された	“昨日”“更新された”なのか，それとも，“昨日”“削除された”なのか，係り受けが曖昧である
主語と述語が離れている	文字情報板が，パターンごとに1～10の番号で管理するが，既設情報板では使用，未使用の状態だけ管理し，パターンバージョンが，提供系処理部からのパターン登録時に自動的に設定し，バージョン要求応答時に，このデータを送信する	主語“文字情報板が”と述語“送信する”が離れている
目的語と述語が離れている	このデータを，提供系処理部に応答を送信するときに使用する	目的語“データを”と述語“使用する”が離れている
修飾語と被修飾語が離れている	図のように，ルールに基づき解析する機能と，統計情報に基づき訳語を選択する機能に，翻訳機能は，分けることができる	修飾語“図のように”と被修飾語“分ける”が離れている
修飾部が長い	上記第1の配管と平行に上記燃焼装置内に設置される蒸気発生器側に送出される冷却材が通過する第2の配管	“配管”の修飾部分が長い
述語の数が多い	そのままの状態で使用を続けると，ホースに亀裂が発生し，ブレーキ液が漏れ，制動力が低下する	述語が四つあり，多い
主語の欠落	大規模コーパスを解析し，データベースに収録する	述語“収録する”の主語が省略されている
目的語の欠落	当部門は毎週1回実施する	述語“実施する”の目的語が省略されている
接続助詞“が”	情報をネットワーク状に表示し分析する技術が開発されているが，分析に加えて検索にも応用が期待されている	接続助詞“が”が多義のため，接続関係がわかりにくい
副助詞“は”	日本語は英語に翻訳し，これにより，観光客の利便性を向上させる	副助詞“は”の格役割が多義のため，係り受け関係がわかりにくい
埋込み表現	上述の機能を備え，特徴情報の照合の結果が同時に提示されるデータ処理装置	埋込み表現の係り先の名詞と用言との関係がわかりにくい
係り受け関係の不明確	機能を確認のうえ利用申し込みをする	必要な助詞や読点が欠落したため，係り受け関係がわかりにくい

2.2　文章をわかりにくくする要因

わかりにくい文章事例を実際の業務文書から網羅的に洗い出し，要因の分析を行った。

これらの要因から人手翻訳と機械翻訳の両方の観点で，翻訳精度に影響がある14項目を選んだ（**表1**）。そのうち，重要度の高い項目を対象として開発を行っている。

2.3　従来の研究

これまでに当社は，企業内で作成される各種業務文書に対して，事前に作成したパターン辞書に基づいて，不適切な表現があるかを診断する技術の開発と検証を行ってきた[(2)－(4)]。この技術は，事前にバリエーションを把握できる禁止用語や曖昧用語の検出には有効である。しかし，辞書の作成にコストが掛かり，また，接続関係の曖昧さなどによる構文的にわかりにくい文の検出はできないという課題があった。

これらの研究は，主に語彙（ごい）レベルでの文字や形態素の情報に基づいたものである。一方，表1で示した要因のほとんどは，構文に関するもので，これまでの研究で提案された手法を適用できない。

そこで当社は，機械翻訳の前編集支援を題材として，構文解析(注1)技術を導入し，構文的な要因によるわかりにくい文を検出することを目的とした研究を行っている。

3　文章のわかりにくさを診断する可読性診断技術

この研究では，構文解析結果に基づいて，文の構成要素とその関係を表す情報を用いて，文のわかりにくさを診断し，わかりにくい文を検出する技術を開発している。

ここでは，文のわかりにくさを診断し，わかりにくい文を検出することを可読性診断と呼ぶ。可読性診断を自動的に行う技術を可読性診断技術と呼び，それに基づいて診断するシステムが可読性診断システムである（**図1**）。

3.1　可読性診断の流れ

診断対象となる文章に対して，システムは以下に示す五つの処理を実施することにより，わかりにくい文を検出する。

（1）　文の切出し　　診断対象文章から文を切り出す。

（2）　構文解析　　切り出した文に対して，構文解析を行う[(5), (6)]。

（3）　可読性診断　　構文解析結果に対して，可読性診断機能群にある全ての診断機能によって，文のわかりにくさを順に診断する。詳細は3.2節で述べる。

（注1）　単語や字句で構成される文を，定義された文法に従って解釈し，文の構造を明確にすること。

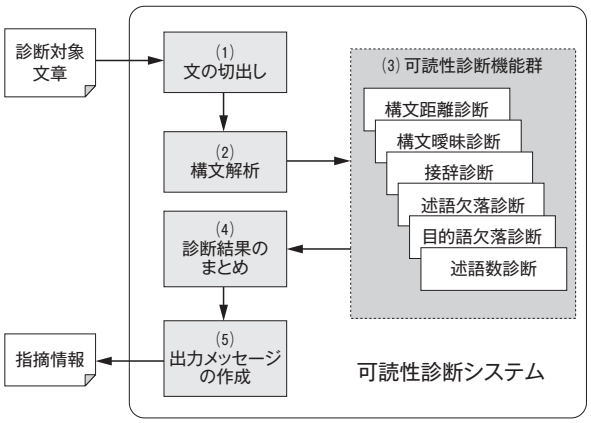

図1. 可読性診断システムの構成 — 構文解析結果を用いて可読性診断機能群の各診断機能により，文のわかりやすさを診断する。

Configuration of document readability diagnostic system

(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，指摘された要因間の関係によって，本質的な要因を優先して指摘するように，出力結果をまとめる。詳細は3.3節で述べる。

(5) 出力メッセージの作成　　まとめられた診断結果に基づいて，利用者に出力する指摘メッセージを生成する。生成した指摘メッセージには，具体的な問題点，指摘理由，及び修正の方針が含まれる。

3.2 可読性診断機能

最初のステップとして取り組んだ可読性診断機能を**表2**に示す。述語欠落診断を例に，可読性診断の仕組みについて以下に述べる。

述語の欠落は，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，及び前接の名詞句の格助詞が省略される現象である。同じ述語の繰返しを避けるために多用される表現形式である。述語の欠落によって，欠落した述語に前接する名詞句の係り先がわからなくなり，関係ない文節と接続する解釈を導きやすい。これによって，2.1節で述べた例文1のような誤訳が発生する。

表2. 可読性診断機能

Document readability diagnostic functions

機　能	対応する要因	機能説明
構文距離診断	主語と述語が離れている 目的語と述語が離れている	主語や目的語が，述語と離れている文を検出する
構文曖昧診断	係り受けが曖昧	係り先が一意に決まらず，かつ，どれが正しいのか判断しにくい文を検出する
接辞診断	複合語中の接辞	辞書に登録されていない複合語に，要注意な接辞を含む文を検出する
述語欠落診断	述語の欠落	連用中止となる述語，述語の語尾，あるいは述語直前の格助詞が省略された並列構造文を検出する
目的語欠落診断	目的語の欠落	目的語が省略された文を検出する
述語数診断	述語の数が多い	一つの文に複数の述語があるため，係り受け関係がわかりにい文を検出する

述語欠落の欠落形式は，欠落部分の形態によって，次の3種類に分けられる。

(1) 名詞で中止　　連用中止の述語とその直前の助詞まで省略された文

［前述の例文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　　連用中止の述語そのものが省略された文

［例文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　　連用中止の述語の活用語尾が省略された文

［例文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。

述語欠落診断は，目的語と述語の接続関係，読点の情報，及び副助詞の情報を合わせて，述語の欠落の有無や，欠落の種類，補足すべき情報を診断する。

例文1に基づいて，処理の詳細を以下に述べる。

(1) 構文解析による認識　　構文解析の処理では，同じ格助詞の繰返しパターンをキーとして，述語欠落の一部が認識できるケースがある。基本的に，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。したがって，まず，構文解析の処理結果から，述語欠落があるかを検査する。述語欠落が認識できた場合には，診断対象とする。例文1の場合，構文解析で述語欠落が認識されなかったため，次の処理(2)に進む。

(2) 副助詞による判別　　日本語の並列文は，副助詞(“は”，“では”，“でも”など)の繰返しによって並列を表すことがある。この特徴を利用し，同様な副助詞を複数使い，かつ読点によって分割された場合，並列文の可能性があると認識して診断対象とする。例文1の場合，副助詞が使われていないので，次の処理(3)に進む。

(3) 目的語による判別　　日本語文において，一つの述語に二つ以上の目的語が係る場合は非文法的となる。このように解釈された文は，連用中止の述語が欠落している可能性が高い。したがって，述語に係る目的語を検査し，二つ以上の目的語が係っている文を診断対象とする。例文1の場合，述語“印刷する”と接続する目的語は，“顧客用文書”と“社内用文書”の二つがあり，診断対象となる。

(4) 欠落種類の判別　　読点の直前にある文節の品詞及び

付属語によって，欠落の種類を判別する。例文1の場合，読点の直前の文節が普通名詞“カラー”であり，付属語が付いていないことから，欠落の種類は“名詞で中止”となる。

(5) 修正候補の推定　述語欠落文は，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によって生じる場合が多い。したがって，文末の述語の語幹や，活用語尾，述語に係る助詞から，欠落した述語を推定できる。例文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”が欠落したと判断することができる。

これら五つの処理によって，例文1に対して次のような診断結果が得られる。

(1) “カラー”の直後に述語が欠落している。

(2) “カラー”の直後に“で印刷し”を補足すべきである。

利用者は，診断結果に基づいて，例文1に対して“で印刷し”を補足し，例文4のように修正することができる。修正された結果に対して機械翻訳すると，原文と同じ意味の訳文が得られる。

［例文4］

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。

［日英翻訳結果］

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

［日中翻訳結果］

以彩色印刷顾客用文书，以黑白印刷公司内部用文书。

3.3 診断結果のまとめ機能

可読性診断機能は，個別の要因に着目し，それぞれ独立に文のわかりにくさを診断する。したがって，同じ箇所に対して複数の要因に基づく指摘が生成されることがある。

［例文5］

以下の説明では，リング型接続式に帰属するノードをゲートウェイと，ネットワーク上のゲートウェイ以外の通信機器をユーザー端末と呼ぶ。

例文5では，述語欠落診断によって，“ゲートウェイと”の直後の述語が欠落していると診断される。一方，構文距離診断によって，目的語“ノードを”と，述語“呼ぶ”が離れていると指摘される。

このような複数の指摘を全て利用者に出力すると，全体の指摘数が増えるだけではなく，どの指摘に対して修正すればよいか，利用者の混乱を招くことにもなる。例文5の場合，述語が欠落しているため，欠落した述語の目的語の係り先が何語も飛び越えた先の文末述語と誤解析される。したがって，述語欠落が本質的な原因であることから，利用者にとっては，この本質的な指摘だけが出力されることが望ましい。

この問題に対して当社は，要因間の関係を考慮し，重要度が高い診断機能ほど優先度を高く設定することにした。診断結果のまとめ機能が，設定された優先度を参照し，同じ箇所に対する指摘が複数ある場合には，優先度によって出力する診断結果を選別する。

4 翻訳業務への適用

4.1 翻訳業務との連携ワークフロー

可読性診断技術の適用先として，翻訳業務との連携を想定している。翻訳業務への適用後のワークフローのイメージを**図2**に示す。

このワークフローでは，翻訳対象文書を翻訳者や機械翻訳に送る前に，可読性診断システムによる診断を実施する。利用者は診断結果に基づき対象文書を編集し，その編集済みの文書が翻訳者や機械翻訳へのインプットとなる。このワークフローを通すことで，従来のプロセスに比べて翻訳者との内容に関するQ&Aの発生量を減らし，誤訳を減少させ，総合的な翻訳効率を高める効果が期待される。

4.2 評価

翻訳業務への適用可能性を検証するために，日本語の技術文書を用いてこの技術の有効性について内部評価を行った。評価者が翻訳者及び機械翻訳処理の立場を想定し，可読性診断システムの診断結果に対して，翻訳の精度や効率の向上に役だつかどうかを確認した。

その結果，可読性診断システムの指摘が，機械翻訳で誤訳が発生しやすい問題の発見に役だつことが示された。その例を以下に述べる。

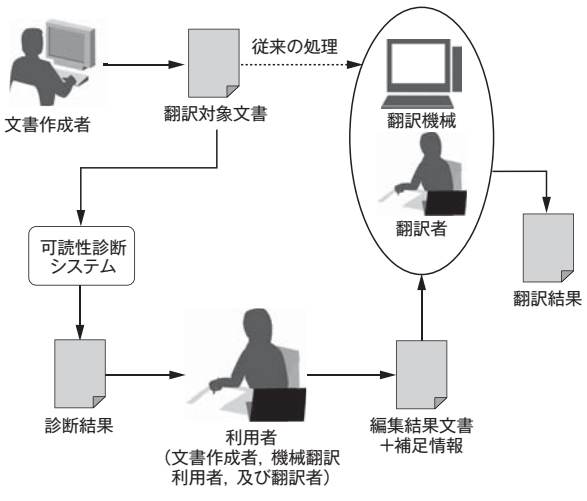

図2．翻訳への適用イメージ— 翻訳の前処理として，対象文書を翻訳者や機械翻訳に送る前に，可読性診断システムによる診断を実施する。
Image of document readability diagnostic system applied to translation

［例文6］

特定すべきディスクの障害

例えば，例文6の場合，“特定すべき”の修飾先が“ディスク”か“障害”かは，構文上では曖昧である。機械翻訳は解釈を一つしか出せないので，文中から解釈を決定付けられる情報が得られない場合は，“特定すべき”に近い語句を係り先とする。

一方，可読性診断では，係り受けが曖昧であるとして検出され，“特定すべき”と“ディスク”の他に，“特定すべき”と“障害”の接続関係もあることが指摘される。人間の判断として“特定すべき”と“障害”のほうが意味的に係り受けが合っており，機械翻訳の解釈のまちがいを可読性診断が検出できるケースである。

改善を要する課題として，“過剰指摘がある”ことと，“見せ方に工夫すべき点がある”ことが挙げられた。

［例文7］

制御手段は，リストを表示し，選択を促す。

過剰指摘の例としては，例文7のように，動詞句の並列表現に主語“制御手段は”が一つある文に対して，主語がどの範囲に係るかが曖昧とする指摘がある。構文上は，“～表示し”だけに係るのか“～表示し，～促す”全体に係るかは曖昧であるが，文のパターンとしてほぼ確かな解釈が得られ，機械翻訳も誤らないケースである。これらが常に可読性診断システムで指摘されることが評価者には過剰と映る。特に人手翻訳では不要となる指摘であり，指摘の質を上げるためには，解釈の確信度を高める取組みが必要となる。

また，見せ方については，指摘メッセージに記載された情報が不十分であったり，人間の直感と合わないという問題もあり，出力方式全体が利用者にとって使いにくいという課題がある。

5　あとがき

可読性診断システムが指摘する問題点は，必ずしも原文を書き換えることで解消できない（しにくい）ものもある。その場合は，原文書中に補足情報を付加することになる。人手翻訳へのインプットとしては，補足情報の付加方法が問題になることはないが，機械翻訳へのインプットとしての用途では，補足情報をうまく翻訳エンジンに伝える機構の構築が必要である。この課題と，4.2節で述べた見せ方の問題を併せて，現在当社は，次の三つの研究開発に取り組んでいる。

(1)　診断システムが指摘する問題点に対する回答を，利用者に無理なく促すインタフェース

(2)　(1)で得られた利用者の回答という付加情報を，文書中にタグなどによって埋め込む機構

(3)　付加情報が埋め込まれた原文書を翻訳エンジンが読み取って，翻訳に生かす機構

また，文章がわかりにくい問題は，機械翻訳に限らず，読み手の文章理解にも影響を与える。しかし，読み手にとってわかりにくい文章と，機械翻訳にとってわかりにくい文章は，必ずしも同じではない。したがって，読み手に対する文章の品質向上に役だつように，この研究で開発した技術が，どこまでの問題をカバーできるか，どのような問題が機械翻訳と異なるか，今後，事例分析に基づいて調査していく。

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. 東京，2009-03. 情報処理学会. p.15－20.

(2)　祖　国威. 中国でのオフショア仕様書チェックシステム. 東芝レビュー. 62, 1, 2007, p.70－71.

(3)　谷口裕子 他. 文脈を考慮した業務文書の数値不整合チェック技術. 東芝レビュー. 63, 2, 2008, p.70－73.

(4)　早川ルミ 他. 日本語解析技術を活用した業務支援ソリューション開発への取組み. 東芝レビュー. 64, 2, 2009, p.30－34.

(5)　平川秀樹. 最適解探索に基づく日本語意味係り受け解析. 情報処理学会論文誌. 43, 3, 2002, p.696－707.

(6)　鈴木博和.“文書全体の情報の利用による機械翻訳の高精度化”. 情報処理学会 FIT2006 第5回情報科学技術フォーラム. 福岡，2006-09，情報処理学会. 2006, p.207－208.

祖　国威　ZU Guowei, D.Eng.

東芝ソリューション（株）IT技術研究所 研究開発部主任，工博。文書品質向上技術の研究・開発に従事。情報処理学会，言語処理学会会員。

Toshiba Solutions Corp.

吉村 裕美子　YOSHIMURA Yumiko

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 クラウドサービス商品技術部参事。機械翻訳など情報活用ソリューションの企画提案や関連サービスに従事。情報処理学会会員。

Toshiba Solutions Corp.

加納　敏行　KANO Toshiyuki

東芝ソリューション（株）IT技術研究所 研究開発部研究主務。情報知識利活用技術の研究・開発に従事。日本OR学会，言語処理学会会員。

Toshiba Solutions Corp.