外国人が作成した日本語文書に対する自動校正技術

祖　国威、加納 敏行
東芝ソリューション株式会社　ＩＴ技術研究所
{so.kokui, kano.toshiyuki}@toshiba-sol.co.jp

１．はじめに

近年の企業活動の国際化により、日本語を母語としない人々が、日本語の文書を作成する機会が増加している。このような、外国人が書いた日本語文書に含まれる誤りは、日本語母語話者が犯す誤りとは異なる種類のものが多い。また発生する誤りが多様なため、結果として文書の文脈が把握しにくくなるという問題もある。

ここでは、まずオフショア開発において外国人技術者が作成した日本語文書を分析し、問題点を整理する。次に、発生頻度が最も高かった「助詞に関する誤り」に対して、これを自動的に検出する手法について述べる。最後に、開発した手法を評価するために、誤りの頻度の多い助詞「を」についての、自動検出に関する実験結果について説明する。

２．オフショア開発における日本語文書の問題点

オフショア開発とは、ソフトウェア開発を海外に所在する企業に委託することである。IT バブル崩壊以降、2000 年代に入り、日本の企業は急速に海外、特に中国への外部委託を展開してきた。オフショア開発のように、日本企業から海外の企業に開発が委託される場合、発注仕様書や納品文書が日本語のままで渡され、納品文書も日本語で書くよう求められるケースが少なからずある。このような場合、外国人技術者は自ら、或いは翻訳者を介して、日本語で発注仕様書を読み、日本語で納品文書を書く必要がある。この際、やり取りされる文書について、以下の２点の問題点が含まれていることが多い。

(1)日本側で作成した発注仕様書が､外国人の技術者にとって理解しにくい。

(2)外国側で作成した納品文書に誤りが多く、日本側での修正が必要になる。

そこで私たちは、まず課題(1)を解決するために、当社の業務文書チェック技術[1,2]に基づき、発注仕様書において不適切な表現を自動的に検出できる仕様書チェックシステムを開発した[3]。

次に、課題(2)の誤りをチェックできる手法を検討するために、まず当社が海外協力会社に発注した際、外国人技術者が作成した文書（納品文書、週報、メールなど）に対して、日本側の担当者が指摘した問題を収集・調査した。日本から中国へのオフショア開発委託が一番多いため、今回の調査対象は、2006 年 4 月～9 月に日本側が中国協力会社の通訳担当者に提示した問題点リストとした。

調査の結果、さまざまな日本語の問題が文書中に含まれていることが明らかになった。図 1 にその頻度分布を示す。

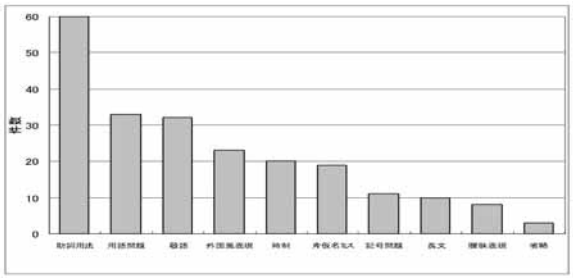

図 1. 中国人が作成した日本語文書の問題点の頻度分布

調査の結果、中国人が作成した日本語文書には、助詞の間違いが特に多いことが分かった。例えば、「が」と「を」、「を」と「に」の誤用など、日本語母語話者なら間違わないものばかりである。図１に示したデータによれば、助詞誤用の問題は全体の 33%を占めている。日本語は助詞によって文法的な関係を示す言語であるため、助詞を誤用すると、文の意味がわかりにくくなる。従って、助詞用法のチェックは、中国人が作成した日本語文書において優先的に解決すべき課題である。

間違いやすい助詞には、格助詞「を」、「が」、「に」、「で」、係助詞「は」、修飾助詞「の」など、さまざまなものがあるが、助詞誤用をさらに詳細に調査した結果、図 2 に示すように、格助詞「を」の誤用がもっとも多く、助詞誤用の 28%を占めていることが分かった。

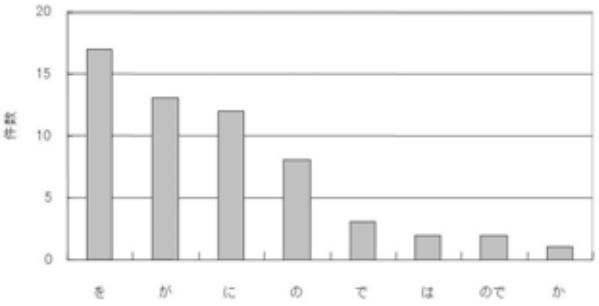

図 2. 中国人が間違いやすい日本語助詞

格助詞「を」の誤用は、主に以下の３種類の類型がある。

(1) 他の助詞を付けるべきところに、「を」を付けてしまう

特に、「立つ」－「立てる」のような、自動詞と他動詞が対応している動詞（いわゆる有対動詞）の場合、格助詞「が」と「を」を誤用しやすい。例えば、「人を集まる。」（人が集まる。）

(2) 「を」を付けるべきところに、他の助詞を付けてしまう

特に有対動詞の「が」と「を」を誤用しやすい。例えば、「考え方が変える。」（考え方を変える。）

(3)「を」格の欠落

助詞「を」と対応する目的語が欠落する。例えば、「あわせて頂きたい。」（認識をあわせて頂きたい。）

このような問題を人手で修正しようとすると膨大な時間がかかってしまう。そこで、外国側で作成した日本語文書の誤りを自動的に発見する機能を持った自動文書校正技術が期待されている。

３．日本語助詞誤用の自動検出手法

本研究では、前述の自動文書校正技術として、機械翻訳用の係り受け解析ツールで出力した格情報を活用することによって、助詞用法の適切性を機械的に判断する手法を提案する。使用した係り受け解析ツールは、「The 翻訳」® [4]の構文解析モジュールである。この構文解析モジュールは、文の係り受け関係に加えて、動詞に接続できる格情報（格助詞のリスト）も出力できる。この格情報を用いて、実際に使われた助詞との照合によって、助詞用法の適切性を判断する。

本研究で提案した手法は、以下の４つのステップから構成される。

Step1：日本語解析処理　入力文に対して、係り受け解析を行う。

Step2：助詞抽出処理　Step1 の解析結果から格助詞に関する係り受け部分（名詞-助詞-動詞の組）を抽出する。

Step3：格情報出力処理　Step2 で抽出した動詞に対して、接続できる格情報を出力する。

Step4：正誤判断処理　用意された判断ルールを参照し、実際に使われている格助詞と、動詞に接続できる格情報を照合し、格助詞の用法が適切であるか否かを判断する。

なお「判断ルール」とは、表 1 のように、判断条件、判断結果およびユーザ向けのメーセージが

関連付けられたルールのことを指す。

表 1. 判断ルールの例

番号	判断条件	判定結果	出力メーセージ
R1	①助詞が「を」である ②格情報に「を」がない。	助詞誤用	○○助詞の用法に誤り
R2	①「を」が使われていない ②格情報に「を」がある	「を」格欠落	目的語が省略されている

４．評価実験と考察

本研究で提案した助詞誤用の自動検出手法を評価するために、誤りの頻度が高い助詞「を」に対して、評価実験を行った。

4.1. 実験データ

実験データとしては、特に誤用しやすい有対動詞の格助詞「が」と「を」について、表 2 に示したようなオフショア開発で良く使われる有対動詞 20 組を採用した。これらの組に対して、以下の４通りの例文を作成し、全部で 80 件の例文データを用いて評価実験を行った。

- 「が」格＋自動詞
- 「を」格＋自動詞
- 「を」格＋他動詞
- 「が」格＋他動詞（「を」格の欠落）

表 2.評価実験データ例

番号	動詞	例文	日本語母語話者判断	機械判断	評価
1	付く	冷房が付く。	○	○	●
		冷房を付く。	×	×	●
	付ける	冷房を付ける。	○	○	●
		冷房が付ける。	×	△	▲

4.2 実験内容

評価実験は次の手順で行った。

Step1：正解の付与

日本語母語話者が、例文の正しさを判断し、判断結果を評価正解として付与する（表 2 の「日本語母語話者判断」欄)。日本語母語話者の判断結果は以下の表記で示す。

「○」(許容)、「△」(「を」格の欠落)、「×」(誤文)

Step2：機械判断

日本語解析結果と格情報を用いて、表１に記載した判断ルール（Ｒ１と Ｒ２）に基づいて、助詞用法の正誤を判断する。判断結果は表 2 の「機械判断」欄に記入する。機械判断の結果は人間判断と同じ表記で示す。

Step3：評価

日本語母語話者の判断結果と機械の判断結果を比較し、表 2 の「評価」欄に記入する。評価基準は次の通りである。

(1) 日本語母語話者と機械の判断結果が一致した場合、適切な検出と判定し、「●」で表記する。

(2) 日本語母語話者が×（誤文）と判断したが、機械は△（「を格」の欠落）と判断した場合、部分検出と判定し、「▲」で表記する。

(3) 日本語母語話者が×（誤文）と判断したが、機械は○（許容）と判断した場合、検出漏れと判定し、「×1」で表記する。

(4) 日本語母語話者が○（許容）と判断したが、機械は×（誤文）と判断した場合、誤検出と判定し、「×2」で表記する。

4.3 実験結果

評価結果を用いて、式１と式２に基づいて、再現率(recall)と適合率(precision)を算出した。

再現率 = (X1+X2-Z1)/(X1+X2)×100%　　(式 1)

適合率 = (Y1+Y2-Z2)/(Y1+Y2)×100%　　(式 2)

ここで、X1 は日本語母語話者が「×」と判断した文数、X2 は日本語母語話者が「△」と判断した文数、Z1 は評価結果が「×1」となった文数である。Y1 は機械が「×」と判断した文数、Y2 は機

械が「△」と判断した文数、Z2 は評価結果の「×2」となった文数である。評価実験結果を表 3 に示す。

表 3. 評価実験結果

適切な検出	部分検出	誤検出	検出漏れ	再現率	適合率
60 件	16 件	1 件	3 件	97.2%	97.6%

4.4 考察

表 3 で示した実験結果によると、以下のことが分かる。

(1) 適切な検出と部分検出を正解とすると、再現率は 97.2%に達した。再現率が高いことは、見落としが少ないことを意味する。

(2) 適切な検出と部分検出を正解とすると、適合率は 97. 6%の精度が得られた。適合率が高く、誤判断が少ないことを示した。

(3) 部分検出は、第 2 章で示した「を」助詞の問題 2（を-が誤用）と問題 3（「を」格の欠落）を、格情報のみを用いた手法では区別できないことを意味している。

5．おわりに

本研究では、オフショア開発を行う際の課題の一つである外国人が作成した文書品質を向上させることを目的として、日本語文書における文法的な誤りの機械的な検出の可能性を検討した。評価実験結果より、格情報を活用することによって、一番間違いやすい「を」の誤用と欠落を機械的に検出できることを明らかにした。今回の検討は、中国人技術者が作成した日本語文書を対象としたが、ここで検討した手法は、外国人が作成した日本語文書に共通に適用できると考えられる。

考察(３)の問題について、対象名詞の属性（例えば、有情名詞か、無情名詞か）が把握できれば、誤りの区別ができると思われる。例えば、表 2 で示したルール 2 に、「③対象名詞が有情名詞である場合、「を」格欠落にする。対象名詞が無情名詞である場合、助詞誤用にする。」という条件を追加すると、「学生（有情名詞）が並べる」の場合、「を」格欠落であり、「椅子（無情名詞）が並べる」の場合、「が」助詞誤用と、より高精度な判断ができるようになると推定される。名詞属性を活用する自動校正支援技術は、今後の研究課題として残されている。

さらに、ここでは、基本文型の文に対して、高い精度で機械的にチェックできることを確認したが、今後は、受動・使役・否定などの様々な文型に対する自動チェックを検討していきたいと考えている。

今後は、「を」以外の助詞(「が」、「に」、「で」など)の誤りの検出に対する手法の研究も行い、オフショア開発支援や、日本語学習者に対する文書作成支援ツールとしての活用を図りたい。

謝辞

本研究にあたり、株式会社東芝　研究開発センター　熊野主任研究員をはじめとして、知識メディアラボラトリーの皆様から、様々なサポートやご教示をいただきました。この場を借りて御礼を申し上げます。

参考文献

[1] 岩田誠司、「企業経営におけるコンプライアンスのための業務文書チェック」、東芝レビュー Vol.60 No.12、2005 年 12 月

[2] 牧野恭子、「不適切表現を発見しリスクを低減する、業務文書のチェックシステム」、東芝ソリューション　テクニカルニュース、2006 年冬季号

[3] 祖国威、「中国オフショア仕様書チェックシステム」、東芝レビューVol.62 No.1、2007 年 1 月

[4] 東芝ソリューション株式会社,「The 翻訳」® 製品説明,
http://www.toshiba-sol.co.jp/ccc/products/translation.htm