文書自動校正技術におけるチェック結果およびフィードバック情報の活用

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、企業において業務文書の品質が厳しく要求されるようになっている。筆者らは、業務文書に起因するリスクの低減を目的として、文書自動校正技術の開発研究に注力し、業務文書における不適切な表現を検出するシステム（以下「業務文書チェックシステム」）を開発した[1]。現在、内部統制文書、オフショア開発文書[2]、医療分野レポート[3]、財務会計等の数値情報を含んだ文書[4]などのさまざまな分野で適用・評価を進めている。

現状の業務文書チェックシステムは、業務知識や利用者のニーズに合わせたチェックルールの調整・カスタマイズの負荷が高いという課題がある。本論文では、チェック結果及びチェック結果に対する評価を蓄積し、分析・活用することにより、チェックルールの調整を支援する処理モデルを提案する。また本モデルに基き開発したチェックルール管理支援システムの適用評価結果についても報告する。

2. 業務文書チェックシステムの問題点

2.1. 業務文書チェックシステムの特徴

業務文書チェックシステム（図１）は、クライアント・サーバ形式で、クライアントからチェック対象文書の情報をサーバに送信し、サーバでチェックを行う。

不適切な表現を検出するためのチェックルールは、チェックモジュールと辞書により実現されている。チェックモジュールと辞書をサーバに置くことにより、同じチームのメンバーによるチェックルールの共有を可能としている。

また、チェック結果の納得性を向上させるため、利用者が自ら定義・管理できる以下の２種類のユーザ辞書がある。

・既知語辞書（利用者にとってチェックが不要な表現を登録）

・チェック対象語辞書（利用者にとってチェックが必要な表現を登録）

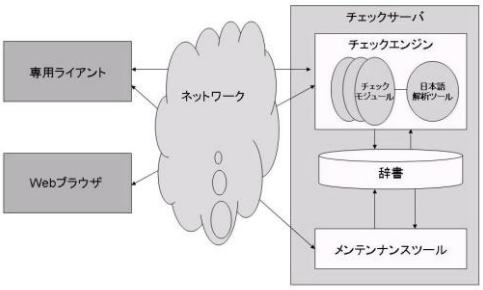

図 1.業務文書チェックシステムの構成

チェックルールの管理では、管理者が利用者の要望やニーズをヒアリングし、適用するチェックルールの選択、チェックモジュールのパラメータの調整、ユーザ辞書の編集の作業を行う。これにより、チェックルールの整合性と均一性を保ちつつ、利用者特有のチェックルールを提供できる。

2.2. 業務文書チェックシステムの課題

図２は、現状のシステムにおけるチェックルール管理の仕組みを示している。これまでの適用評価を通じ、以下の課題が明らかになった。

課題 1. ユーザ辞書作成の負荷が高い

ユーザ辞書を最初に作成するためには、システム開発者が大量の事例を分析した結果に基づき登録語を決める必要がある。事例を分析する作業だけでなく、利用者が事例を準備する作業にも時間がかかる。また、利用者が事例を提供できない場合もある。

課題 2. ユーザ辞書の編集・管理が難しい

現状のユーザ辞書管理は、管理者によって統一的に行われるが、登録する用語は、利用者の意見に基づき決定する必要がある。そのためには利用者が意識的に毎回のチェック結果からチェック対象とすべき語や対象外の語を記録し、管理者に伝えていく必要があった。

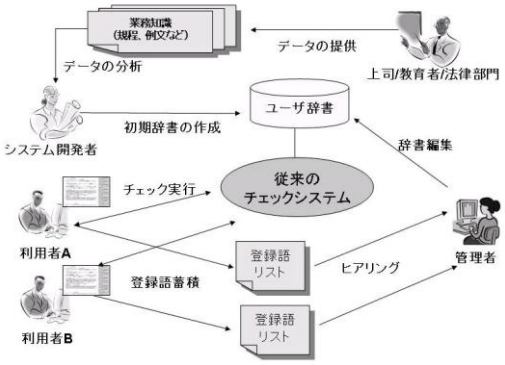

図 2.現状のチェックルール管理のイメージ

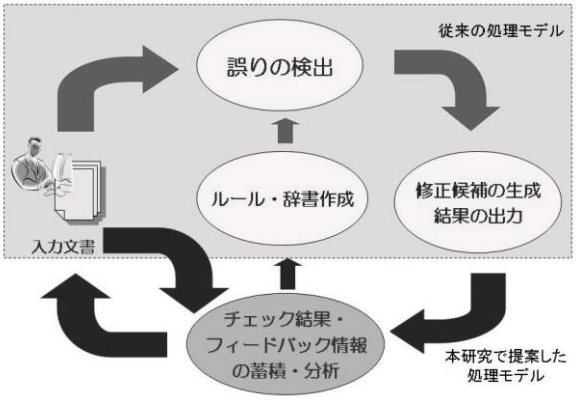

図 3.本研究で提案するチェックモデル

課題 3. 利用者の評価がチェックルールの調整に活用されていない

本システムにおいて、同じ開発チームのメンバーは同じチェックルールを共有し、利用者の特性（例えば、使用する専門用語や、文書作成上の規約など）に応じて詳細な設定をする仕組みが用意されている。このような利用者の特性やチェック結果に対する評価は、チェックルールの調整のための重要な情報である。

しかし、現状のシステムでは、利用者の評価結果を収集・分析し、チェックルールの調整に反映させる仕組みが用意されていない。

３．チェック結果とフィードバック情報の活用

1980 年代より、文書作成業務の自動化の進展に伴い、新聞・図書の出版分野をはじめとして、文書校正支援技術に関する多くの研究が行われている[5]。これらの研究の殆どは、誤りの自動検出及び修正候補の自動生成に注力しており、誤りを検出するため、予め大量のチェックルールを用意するルールベースの方法が広く用いられている(図 3 の上部)。大量のルールを事前に人手で作成することや状況の変化に伴うチェックルールの編集・管理は負荷の高い作業であった。従って現状では、チェックルールの構築・管理の省力化は、文書校正技術において難しい課題の一つとなっている。

本研究では、文書チェックにおけるチェック結果及びチェック結果に対する評価情報を活用することで、チェックルールの調整を支援する処理モデル（図３）を提案する。本モデルは、従来のチェックモデル（図 3 の上部)にチェック結果・フィードバック情報の蓄積・分析処理（図 3 の下部)を付加することによって実現する。付加された処理は、図４で示した４つの機能を備えることが特徴である。以下に概要を示す。

(1) チェック結果の蓄積機能

文書をチェックした時のチェック結果を蓄積する。蓄積されるチェック結果は、チェック内容（指摘された問題箇所(表現)、指摘メッセージ、修正候補、例文など)、ログ情報（ユーザ名、文書名、チェック日時など)、及び統計情報（チェックによる指摘頻度など）である。

(2) フィードバック情報の収集機能

利用者のフィードバック情報を収集する。利用者のフィードバック情報とは、利用者によるチェック結果に対する評価であり、意識的なフィードバック情報（納得できるかどうか、チェック漏れがあるかどうかの直接的な評価）と、無意識のフィードバック情報（対象文書に対する修正結果）が含まれる。

(3) 集計・分析機能

収集・蓄積された情報に基づいて、集計と分析を行う。例えば、既知語辞書に登録すべき語を決定するため、チェック頻度が高い指摘語を抽出する。また、利用者の納得性を把握するため、フィードバック情報を定量的に算出する。さらに、利用者の特性を明確にするため、指摘項目や専門用語の分布を分析する。

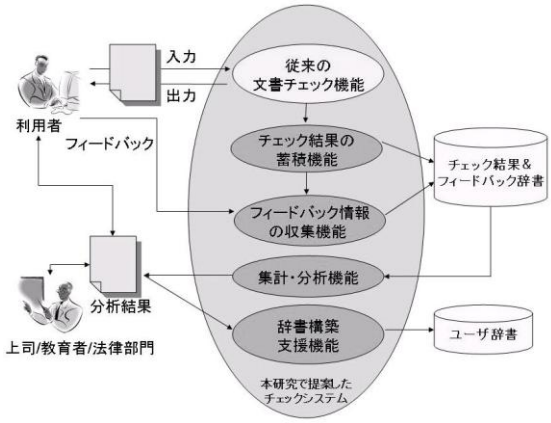

図 4.提案モデルの機能イメージ

(4) 辞書構築支援機能

集計・分析の結果を用いて、ユーザ辞書の自動登録や、最適なチェックパラメーターの選択や、利用者の特性に合うチェックルールの選択などのチェックルールの調整を支援する。

以上（1）～（4）の機能を備える業務文書チェックシステムの実現により、チェックルール管理のイメージが図 5 のようになる。従来のチェックルール管理の流れ(図 2)と比べると、以下の変化があり、効率は全体的に向上する。

(1)システム開発者による初期辞書の作成が不要となる

運用しながら、チェック結果やフィードバック情報によって辞書が更新されるので、システム開発者が初期辞書を作成する必要がなくなる。初期辞書が必要となる場合でも、他の利用者により蓄積された情報を参照することができる。(課題 1)

(2)管理者によるユーザ辞書の管理が不要となる

ユーザ辞書を更新するためのデータが自動的に蓄積されるので、利用者による意識的な収集作業が不要となる。また、辞書エントリが自動的に登録されるので、管理者による辞書の管理が不要となる。(課題 2)

(3) チェックルール調整のためのヒアリングが不要となる

利用者によるチェック結果に対する評価を随時収集できるため、利用者の要望を素早くシステムに反映できる。(課題 3)

(4) 利用者の特性を把握できる分析情報が生成される

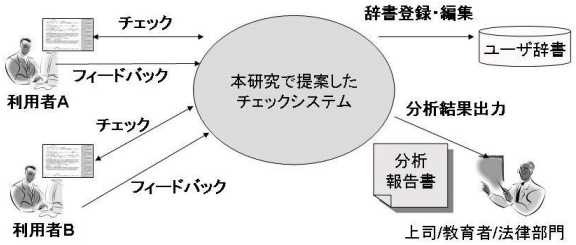

図 5.新しいチェックルール管理のイメージ

指摘頻度集計などの統計・分析情報を出力するので、利用者の特徴を把握でき、より詳細なチェックルールの設定が可能となる。(課題 3)

4． チェック結果蓄積機能の適用評価

4.1. チェック結果蓄積機能の実装

第 3 章で述べた 4 つの機能に対して、今回はチェック結果を蓄積し分析を行う機能を備えるチェックシステムのプロトタイプを実装した（図 6)。

チェック結果として蓄積する情報は、利用者名、対象文書名、チェック日時、指摘表現、例文、チェック項目、指摘表現の出現頻度など 12 種類の情報である。

また、蓄積された情報を分析するため、チェック結果の検索・閲覧ツールを作成した。このツールを通じて、指定された条件に基づき、チェック結果情報を表示し、既知語や登録語の抽出を支援する。

4.2. 適用評価

中国オフショア開発において作成された仕様書(20 件)を対象として、本システムの適用評価を行った。

表 1 は蓄積されたチェック結果の例である。蓄積されたデータの活用により、以下の効果があることが確認されている。

(1)既知語辞書に登録するための候補が作成できる

オフショア開発における仕様書をチェックする場合、過剰チェックを抑制するために、わかりにくいカタカナ語と未知語の指摘において既知語辞書を使用している。この既知語辞書には開発チーム毎にチェック不要な用語が登録

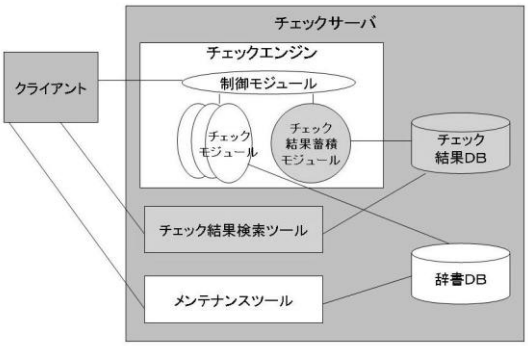

図 6. プロトタイプシステムの構成

されている。しかし、登録用語の収集をするのは、利用者にとって負担となっていた。

チェック結果蓄積機能を用いると、実際に指摘されたカタカナ語（例えば、表 1 の 3 行目の「サービスアプリケーション」）の一覧が得られ、出現頻度に基づき、登録すべき用語を簡単に選出できるので、既知語辞書作成の効率が向上した。

⑵蓄積された指摘表現の出現頻度から、辞書登録語が共通用語か特定用語かを判定できる

ユーザ辞書に登録する用語には、全利用者に共通の一般用語および、特定の利用者に特有の特定用語がある。蓄積されたチェック結果には、利用者情報と出現頻度情報が含まれるので、利用者毎の頻度比較によって、共通用語と特定用語を判別できる。

⑶例文を蓄積するので、チェック結果の妥当性を評価できる

チェック結果には例文情報（指摘された表現を含む原文）が含まれるので、指摘された表現の前後の文脈を参照することにより、チェック結果の妥当性を評価できる。

５．今後の検討課題とまとめ

本論文では、チェック結果やフィードバック情報を取り扱う処理モデルを提案し、プロトタイプシステムによる適用評価結果について報告した。今後は、実利用者による適用評価を進め、システムの有効性を検証する予定である。

チェック結果のみを分析対象とする場合は、過剰チェックを発見し、それを低減する既知語辞書の構築を支援することができる。しかし、チェック漏れは発見できない。また、利用者の意見が含まれないので、利用者の実情に応じたチェックルールの調整まではできない。この課題に対して、フェーズ２ではフィードバック情報の自動収集機能を実現する予定である。利用者の評価・意見を直接に把握でき、妥当性が高いチェックルールの調整が期待される。

今後は、図４で示したフィードバック情報の収集機能、集計・分析機能、辞書構築支援機能の実現により、辞書構築の一層の省力化やチェック精度の向上を図っていく。

表 1.蓄積されたチェック結果の例

利用者名	文書名	日時	項目	指摘表現	例文	頻度
User1	文書A	YY/MM/DD	未知語	等于	ユーザID等于'0008'的。	1
User1	文書A	YY/MM/DD	長文	40文節の文章	ファイル名、ファイル情報、フォルダパス等をファイルの先頭(これ以降、ファイルヘッダと明記する)に書き込むことにより、デアーカイブを可能とする。	1
User2	文書B	YY/MM/DD	難解カタカナ	サービスアプリケーション	AAA株式会社のサービスアプリケーション「XXXServer」(これ以降、XServと明記する)から呼び出すAPIを開発する。	10

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、2005.12

[2] Guowei, ZU, et al. “The Supporting Technology of Business Document Proofreading based on Intercultural Differences”, CEC’07 and EEE’07, pp.91-98, 2007.7 Tokyo

[3] 牧野恭子「医療分野向けテキストマイニング技術」、東芝レビューVol.60 No.9、2005.9

[4] 谷口裕子、祖国威、加納敏行「文脈を考慮した数値不整合チェック技術」、東芝レビュー Vol.63 No.2、2008.2

[5] 池原悟、小原永、高木伸一郎「文書校正支援システムにおける自然言語処理（〈特集〉自然言語処理技術の応用）」情報処理学会論文誌、Vol.34, No.10, 1993, pp. 1249-1258