文脈を考慮した業務文書の数値不整合チェック技術

Contextual Checking System for Numerical Inconsistencies in Business Documents

谷口 裕子 ■ TANIGUCHI Yuko
祖 国威 ■ ZU Guowei
加納 敏行 ■ KANO Toshiyuki

数値情報を含んだ業務文書には，営業週報や財務報告書，社会動向に関する各種調査報告書などがある。もしこのような業務文書に金額などの数値のまちがいがあり，チェック時の見落としによってそれらが公になった場合，企業の信頼失墜につながるといったリスクが考えられる。このようなリスクを回避するためには，業務文書を漏れなくチェックし校正することが重要であるが，日々作成される多数の業務文書について人手で細かくチェックするのは企業にとって大変な労力やコストが掛かる作業である。

東芝ソリューション（株）は，このような数値情報の不整合チェックを支援する“数値不整合チェック技術”を現在開発している。この技術により，文書作成者が数値情報の整合性を確認する作業を軽減し，結果としてリスクの少ない文書の作成支援が可能になる。

Business documents such as weekly sales reports, financial reports, and social surveys contain many types of numerical information. In the event that such a document contains a numerical inconsistency that is not discovered and corrected before publication, the company's reputation will suffer negative consequences. The proofreading of documents is therefore indispensable to avoid such risks. However, the proofreading of large volumes of business documents generated by a company using human efforts alone involves a great deal of labor and high costs.

Toshiba Solutions Corporation is developing a contextual checking system for numerical inconsistencies in business documents, which will help persons responsible for such documents to eliminate numerical inconsistencies in them. This technology will result in considerable savings in labor and facilitate the publication of documents with a low risk of mistakes.

1 まえがき

企業経営において，コンプライアンスやビジネスリスク低減を実現するためには，日々の業務の中で法令やルールを確実に遵守し，規程の違反や虚偽の報告などが起こらないようにする必要がある。このためには，業務ルーチンで作成された業務文書（電子データ）に問題がないかチェックし，不適切な表現があれば担当者に通知して修正を促し，記述内容にリスク要因があれば早期に抽出してリスクを低減していくことが求められる。しかし，日々作成される多数の業務文書から，文書の書き方に関する社内規程に違反する文章や，不適切な表記などを人手で細かくチェックするのは，企業にとって大変な労力やコストが掛かる[1), (2)]。

東芝ソリューション（株）は，様々な業務文書を関連法規や，社内規程，業務知識，及びノウハウなどを基準としてチェックする業務文書チェックシステムを開発し，各種分野でプロトタイプによる評価を行っている[3)]。

例えば，オフショア開発仕様書では，日本側で作成した仕様書の内容に外国の技術者が理解しにくいあいまいな部分がある，というリスク要因がある[5), (6), (7)]。また，医療分野での業務文書の一種である“読影レポート”(注1)では，診療科医師の検査目的に応える内容が書かれていない，誤解を生じやすい表現が使われている，といったリスク要因がある[4)]。更に，法令対応のために作成されるRCM（Risk Control Matrix）では，具体性の乏しい記載や整合性のない記載がリスク要因として挙げられる。

現在，試作と評価を進めている業務文書チェックシステムは，オフショア開発仕様書や医療レポート，RCMといった種類の業務文書をチェックし，不適切な表現の指摘やリスク要因の抽出を行うことで，コンプライアンスやビジネスリスクの低減に役だてることができる。

このような業務文書の中でチェックすべき項目は，不適切な表現や文法上のまちがい，数値情報の不整合など多岐にわたる。ここでは，これらの中で数値情報の不整合に着目し，現在研究開発中の“数値不整合チェック技術”について述べる。

2 数値不整合をチェックする際の課題

数値情報を含む代表的な業務文書としては，営業週報や財務報告書，決算報告書，業績レポート，社会動向に関する各種調査報告書（例えば，環境レポート)などがある。もしこのよ

（注1） 放射線科医師がMRI（磁気共鳴断層画像法）やCT（コンピュータ断層撮影）などの検査画像から，異常の有無や考えられる疾患及び状態について報告するレポート。

うな文書に金額などの数値のまちがいがあり，チェック時の見落としによってそれが公になった場合，企業の信頼失墜につながるといったリスクが考えられる。このようなリスクを回避するためには，業務文書中の不適切な表現や文法上のまちがい，数値情報の不整合など様々な項目についてチェックし，校正することが重要である。数値情報が記載された文書を作成する場合，例えば**図1**に示すサンプル文書のような不具合が発生する。

(1) 同一文書内での数値の不整合

(a) 本文に複数回記載されている数値の不整合　例えば，2000年のA国の二酸化炭素（CO_2）排出量が，本文では57億トンとなっているが，文書冒頭の要約では単位をまちがえた（57トン）。

(b) 表と本文に記載されている数値の不整合　例えば，表の数値（57億トン）を本文に転記した際，単位をまちがえた（57トン）。

(2) 複数文書間での数値の不整合　例えば，文書Aでは2000年のCO_2排出量がB国は30億トンだが，これを参照して記述した文書Bでは，39億トンと打ちまちがえた。

このような，文書中に含まれる数値の正しさをチェックする場合，数値そのものを見るだけではなく，文脈も考慮して，その数値が“いつ”，“どこの”，“何に”関するものなのかといった，数値の前後に記述されている様々なキーワードも併せて見る必要がある。例えば，“2000年のCO_2排出量は，A国が57億トンである”という文において，“57億トン”という数値が正しいかどうかを判断する場合，“2000年”や“A国”，“CO_2排出量”のような，数値を特定するためのキーワードも必要である。場合によっては，複数の文書を比較して数値を確認しなければならないケースもある。

3　数値不整合チェック技術

当社は，このような数値情報の不整合チェックを支援する“数値不整合チェック技術”を現在開発している。数値不整合チェック技術の最大の特長は，自由記述された文や非定型の表が含まれる自由度の高い文書を解析し，数値情報を抽出できることである。

数値不整合チェックは，数値情報の抽出と不整合チェックの二つのステップから構成される（**図2**）。ここでは，同一文書内の表と本文にある数値の不整合について主に述べるが，本文に複数回記載された数値の不整合や，複数文書間での数値の不整合の場合にも応用できる。図2に示すように，まず“数値の抽出”処理で，文書中の本文と表から，数値とそれが何に関する数値かを表すキーワードを，日本語解析技術と表解析技術を用いて，それぞれ見つけ出す。具体的には，本文を解析して前述の数値とキーワードを抽出し，文中の語順などの相関関係を手がかりに，数値とそれに対応するキーワードの関連付けを行う。この際，同一文中で必要なキーワード，例えば数

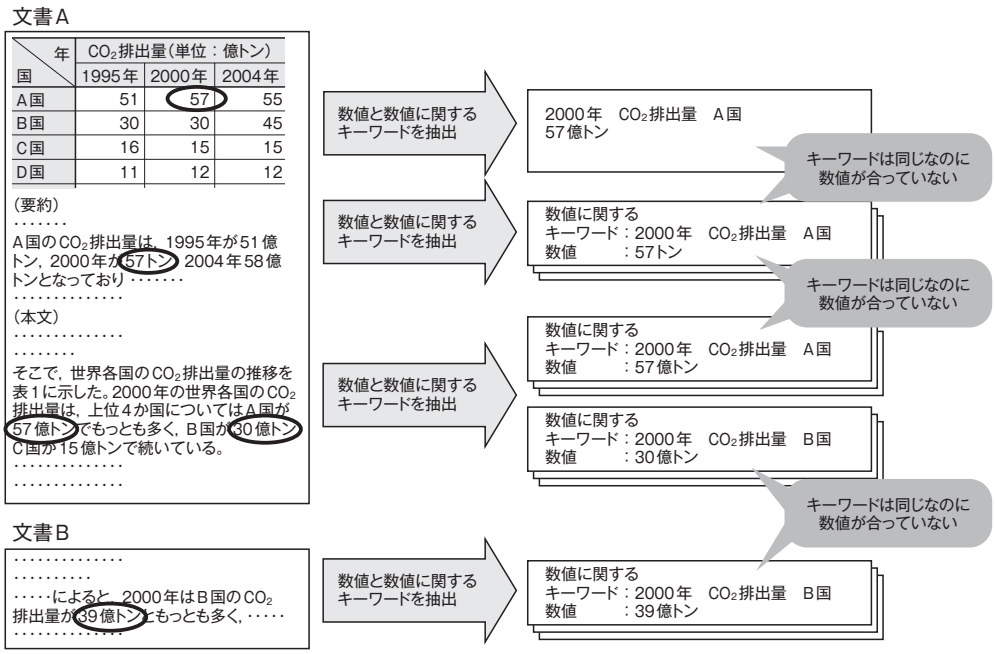

図1. 数値情報不整合の例 — 数値情報の不整合には，本文に複数回記載されている場合や，表と本文に記載されている場合，複数文書で記載されている場合などがある。

Examples of numerical inconsistencies in business document

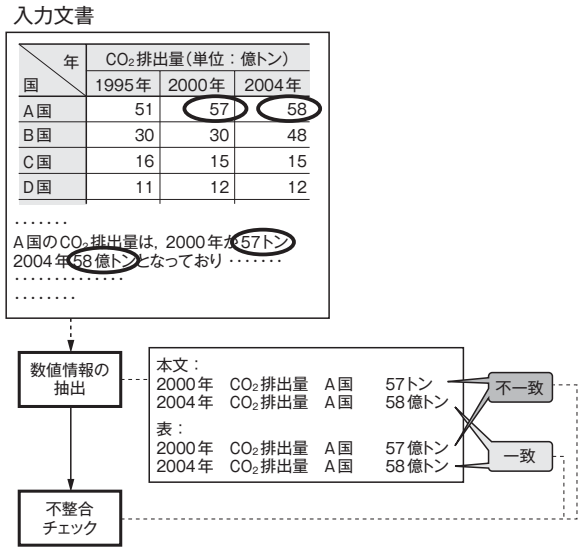

図2. 数値不整合チェック実行時の流れ — 文書中の数値情報を抽出し，不整合をチェックする。

Process flow of proofreading for numerical inconsistencies

値が“いつ”発生したかが見つからない場合，文脈の解析を行い，周辺の文から必要な情報を補完する。次に，本文から抽出した数値とキーワードの情報を使って表の構造を解析し，表に記述されている数値とキーワードを抽出する。その後“不整合チェック”処理にて，本文及び表から抽出した数値情報を照合し，数値の不整合の有無をチェックする。

ここで，この数値不整合チェックを世界各国のCO_2排出量の統計データをまとめたサンプル文書に対して行った例を**図3**に示す。この例では，文から抽出した情報と表から抽出した情報が数値不整合チェックにより照合された結果，本文と表に不整合があったものと整合性がとれていたものとが色分けされ，不整合のある箇所が視覚的に強調されている。このようなチェックを実施することで，文書作成者が数値情報の整合性を確認する作業を軽減し，結果としてリスクの少ない文書の作成支援が可能になる。

現在，この数値不整合チェック技術を用いて，金融分野における数値情報のチェックを想定した試行システムの開発を行っている。

地球温暖化の原因となる温室効果ガスには，二酸化炭素(CO_2)やメタン，フロンなど様々なものがある。その中で最大の要因を占めるのはCO_2である。そこで，世界各国のCO_2排出量の推移を表1に示した。2000年の世界各国のCO_2排出量は，上位4か国についてはA国が57トンでもっとも多く，B国が30億トン，C国が15億トンで続いている。このことから，CO_2排出量の多い国が必ずしも先進国ではないことがわかる。

2000年に比べ，2004年の世界各国のCO_2排出量は更に増加しており，A国が57億トンでもっとも多く，B国が48億トン，C国が15億トンとなっている。2000年と比較して，B国のCO_2排出量の増加が顕著である。このままB国のCO_2排出量が増加すると，A国を抜いて世界第1位になるのも遠い将来ではないと言える。

表1　世界各国のCO_2排出量の推移

年 / 国	CO_2排出量(単位：億トン)		
	1995年	2000年	2004年
A国	51	57	58
B国	30	30	48
C国	16	15	15
D国	11	12	12
E国	8	10	11
F国	9	8	8
G国	5	5	6
H国	5	5	5
その他	76	84	93
合　計	211	226	256

先進国のうちCO_2排出量がもっとも少ないのは，H国の5億トンである。

コメント［本文と表が不一致1］：
本文　：2000年　CO_2排出量　A国　57トン
表1　：2000年　CO_2排出量　A国　57億トン

コメント［本文と表が一致2］：
本文　：2000年　CO_2排出量　B国　30億トン
表1　：2000年　CO_2排出量　B国　30億トン

コメント　［本文と表が一致3］：
本文　：2000年　CO_2排出量　C国　15億トン
表1　：2000年　CO_2排出量　C国　15億トン

コメント［本文と表が不一致4］：
本文　：2004年　CO_2排出量　A国　57億トン
表1　：2004年　CO_2排出量　A国　58億トン

コメント　［本文と表が一致5］：
本文　：2004年　CO_2排出量　B国　48億トン
表1　：2004年　CO_2排出量　B国　48億トン

コメント　［本文と表が一致6］：
本文　：2004年　CO_2排出量　C国　15億トン
表1　：2004年　CO_2排出量　C国　15億トン

コメント　［本文と表が一致7］：
本文　：2004年　CO_2排出量　H国　5億トン
表1　：2004年　CO_2排出量　H国　5億トン

図3. 数値不整合チェック結果の例 — 本文から抽出した数値と数値に関連するキーワードを用いて表との照合を行い，その結果をコメントとしてユーザーに提示する。

Image of diagnostic comments provided by contextual checking system

4　あとがき

今後は，文から抽出した数値情報と，同一文書中の表から抽出した数値情報との照合に加え，文から抽出した数値情報どうしの照合や，外部の文書やデータベースの数値情報との照合も可能にすることで，更なるチェック精度の向上を実現したいと考えている。更に，チェック結果の表示方法を工夫したり，ユーザーからのフィードバックを積極的に活用したりすることで，より導入しやすく，利用しやすいシステムを目指す。また，これらの活動を進めることで，数値情報を含む様々な分野の業務文書に対応したチェックシステムとして，製品化を目指して展開していく。

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. **60**, 12, 2005, p.36－39.
(2) 岩田誠司. ITによるビジネス文書処理を取り巻く動向と課題. 東芝ソリューション テクニカルニュース. **8**, 冬季号, 2006, p.2－3.
(3) 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. **8**, 冬季号, 2006, p.12－13.
(4) 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. **60**, 9, 2005, p.46－47.
(5) 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.
(6) 祖　国威，ほか. “外国人が作成した日本語文書に対する自動校正技術”. 言語処理学会第13回年次大会論文集. 大津, 2007-03, 言語処理学会. 2007, S2-4.
(7) Zu Guowei, et al. “The Supporting Technology of Business Document Proofreading based on Intercultural Differences”. CEC’07 and EEE’07. Tokyo, 2007-07, IEEE. 2007. p.91－98.

谷口　裕子　TANIGUCHI Yuko
東芝ソリューション（株）IT技術研究所 ビジネスインテリジェンスラボラトリー。文書チェック技術の研究・開発に従事。
Toshiba Solutions Corp.

祖　　国威　ZU Guowei
東芝ソリューション（株）IT技術研究所 ビジネスインテリジェンスラボラトリー。文書チェック技術の研究・開発に従事。言語処理学会，ACM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 ビジネスインテリジェンスラボラトリー主任。ビジネスインテリジェンス技術の研究・開発に従事。言語処理学会，日本OR学会会員。
Toshiba Solutions Corp.