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

Approach to Development of Business Support Solutions Utilizing Japanese-Language Analysis Technologies

早川　ルミ　　松本　茂　　齋藤　佳美
■ HAYAKAWA Rumi　■ MATSUMOTO Shigeru　■ SAITO Yoshimi

東芝ソリューション（株）は，情報知識利活用技術の一部として，日常業務で作成される様々な業務文書の品質を高めるとともに，精度よく分類・整理して活用できるようにする技術の研究及び開発を進めている。

現在，“業務文書チェック技術”，“業務文書分類技術”，及び“パラフレーズ（言いかえ）検索技術”の実用化に注力しており，今回，中国でのオフショア開発(注1)向けの仕様書チェックシステムや，特許文書の自動分類システムなどを開発し，それらの評価を行っている。

Toshiba Solutions Corporation has been advancing the research and development of technologies for more practical use of business documents that are being created and stored every day. For this purpose, our technologies support improvements in the quality of documents and classification accuracy.

We are currently focusing on the utilization of business document checking technology, business document classification technology, and paraphrase searching technology. Utilizing these technologies, we are building and evaluating prototype systems such as a document checking system for offshore development with China and an automatic classification system for patent documents.

1　まえがき

東芝ソリューション（株）は，知識経営を支援するために，“情報知識利活用技術”の研究開発に取り組んでいる。

当社は，“情報知識利活用”を，「企業内外に蓄積されている多種多様な大量のデータから，有用な情報を迅速かつ的確に抽出して加工し，業務の品質や効率の向上，リスクの低減，戦略の立案や意思決定に生かすこと」と定義している。また，情報知識利活用技術は，必要な情報を収集する技術，情報を蓄積し共有する技術，情報を分析する技術，及び情報を知識に変換する技術など，情報や知識の活用を支援するために必要な技術の集合である。

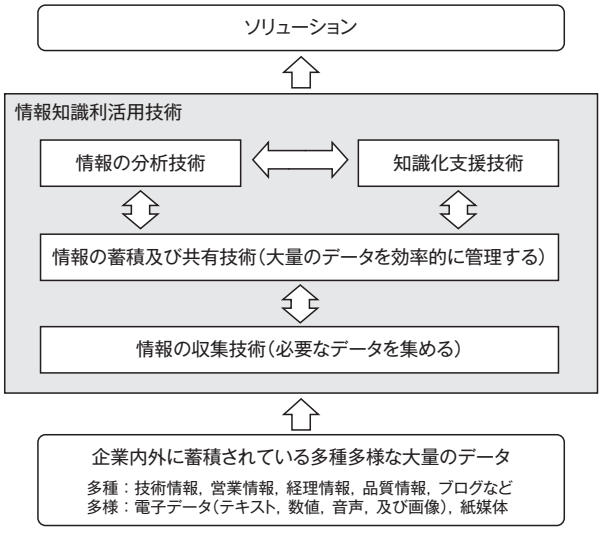


図1．情報知識利活用の全体像 ― 種々雑多な状態にあるデータは，情報知識利活用技術の適用により，企業活動の様々な問題解決に活用できる。
Overview of information knowledge utilization

図1に示すように，テキスト，数値，音声などの種々雑多な状態にあるデータは，情報知識利活用技術の適用により，企業活動の様々な問題解決に活用できると考えている。

しかし，これら情報知識利活用技術の実現には多くの課題がある。そこで，実業務に適用可能なソリューションの提供を目指し，日本語解析技術をベースとした文書処理技術の研究を進め，情報知識利活用技術の実用化に取り組んでいる。

2　情報の活用における課題

情報知識利活用における企業内外に蓄積されている多種多様な大量のデータの中で，企業活動に与える影響が大きいデータとして業務文書（テキストデータ）がある。まずは，この業務文書から有用な情報や知識を抽出し活用することが，情報知識利活用の実現に大きな効果があると考えられる。

しかし，業務文書を情報として活用するための技術的支援はいまだ十分ではなく，例えば，以下に示すような問題が実業務の現場に存在している。

（注1）　ソフトウェアの開発を海外の企業に委託すること。

(1) 情報の品質を整えるための作業負荷が大きい　例えば，複数人で開発用ドキュメントを作成する場合，用語のまちがいや担当者ごとに表現のばらつきがあり，修正作業に負荷が掛かる。
(2) 情報の分析作業に手間が掛かる　例えば，顧客からの問合せ情報を蓄積しFAQ (Frequently Asked Questions) を作成する業務の場合，問合せや回答の内容を分類してまとめる作業には人手に頼る部分もあり，非常に手間が掛かる。
(3) 情報を収集するための検索精度が必ずしも良くない　例えば，大量の業務文書から目的の文書を探すために検索キーワードを増やしても，自分がほんとうに必要とする業務文書を探し出すことは困難な場合が多い。

このような現場の問題を解消し，業務文書を情報として活用するために，次の技術の研究開発が必要となった。

(1) 情報としての業務文書の品質を保持向上させる，業務文書チェック技術
(2) 情報の特徴を分析して，業務文書をわかりやすく分類・整理する業務文書分類技術
(3) 必要な情報を得るために，業務文書を狙いどおりに検索するパラフレーズ検索技術

3　日本語解析技術の業務支援ソリューションでの活用

2章で述べた問題解決のために，形態素解析技術や，構文解析技術，意味解析技術などを中心とする日本語解析技術を，実業務で役だてるための研究と開発に取り組んでいる。ここでは，“業務文書チェック技術”，“業務文書分類技術”，及び“パラフレーズ検索技術”について，その技術概要と応用事例について述べる。

3.1　情報の品質を保持向上させる業務文書チェック技術

業務文書を情報として活用するためには，業務文書の品質の保持と向上を効率的に実現することが必要である。このため，業務文章中に潜む“不適切な表現”を機械的に洗い出し，文書作成者に注意や修正を促して適切な表現に修正することで，文書の品質向上に寄与する業務文書チェック技術の研究に取り組んでいる。業務文書における不適切な表現とは，具体的に以下の表現を言う。

(1) 日本語としてまちがっている表現
(2) 読み手により異なる解釈を与える可能性のある表現
(3) 矛盾を含んだ表現

当社が開発した業務文書チェックシステムの仕組みを**図2**に示す。このシステムは，形態素解析や構文解析などの日本語解析技術と文字列パターンマッチング技術を用いている。一般的なワープロのソフトウェアとは異なり，業務文書チェックシステムでは，業務文書の性質に応じてチェック項目と用語辞書を組み合わせ，業務文書中の不適切な表現を判定し，元の文書に対してコメントを提示する。チェック項目の例を**表1**に示す。

現在，試作品の評価や商品として実際の業務で行っている

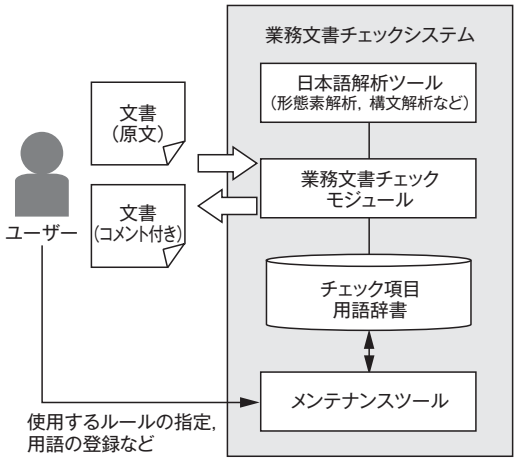


図2．業務文書チェックの仕組み — 業務文書の性質に応じてチェック項目と用語辞書を組み合わせ，業務文書中の不適切な表現を判定し，元の文書に対してコメントを提示する。
Document checking system

表1．業務文書のチェック項目例
Example of document check items

No.	チェック項目	内　容
1	“ですます”体	“です”，“ます”体で終わる文を，“だ”，“である”体に変更することを促す。
2	体言止め	体言（名詞，名詞句）で終わる文をチェックする。
3	基本の用字と用語	JIS Z8301:2005の附属書G（規定）「文章の書き方，用字，用語，記述符号及び数字」のサブセット。特定の語について，漢字又はかなのどちらで表記すべきかなどをチェックする。
4	未知語	標準的な辞書に登録されていない用語を検出し，アラームを出す。
5	用語統一	あらかじめ指定した表現が用いられていないものをチェックする。例えば“試験仕様書”を“テスト仕様書”に統一する。
6	長文	指定された文字数や文節数を超える場合，指定数以内への変更を促す。
7	英略語初出	文書で英略語が初めて記載された場合に，該当の英語原語表現の併記を推奨する。
8	参照基準のあいまい	“既存”，“必要に応じ”などに対して，具体的な対象や条件の記載を推奨する。
9	わかりにくい疑問詞	疑問調の表記の場合に，疑問詞を用いた平叙文を用いないように促す。
10	わかりにくい否定	否定語を重ねて用いる表現を使用しないように促す。例えば“ないわけではない”。
11	口語の使用	口語表現（ら抜き言葉）を使用しないように促す。
12	同形異義	中国語と日本語で同じ漢字だが意味が違うことばについて，可能なかぎり使用しないか，英語の併用を推奨する。
13	数値の整合性	本文と表の数値の一致，不一致をチェックする。例えば，本文が“利益実績94百万円”で，表が“利益実績84百万円”であれば不一致とする。

チェックには次の例がある。例えば中国でのオフショア開発では，日本語母語話者（日本人）と非日本語母語話者（中国人）の間の意思疎通が重要である。表1のNo.10や12を，日本の開発者が作成する開発仕様書に適用すれば，中国の開発者に誤解を与えにくい表現で記述された仕様書を提示することができる[1]。また，業務報告書や財務報告書などでは，“数値”情報の整合性が保たれなければならないが，表1のNo.13を用いれば，本文と表に出現する数値の一致や不一致，あるいは対応する数値の有無をチェックできる。作成者が見落としやすい数値の記述まちがいを指摘して正すことで，誤った情報を提示するリスクを回避できる[2]。

このほか，医療現場で作成される画像診断報告書に対しては，東芝メディカルシステムズ（株）及び名古屋大学医学部付属病院放射線部と協力して，専用の報告書作成システム上で動作する業務文書チェックを試作し評価している[3]。また，内部統制におけるRCM（Risk Control Matrix）(注2)を対象に，正しく漏れなく情報が記載されているか否かを業務文書チェック技術により判定する仕組みは[4]，KnowledgeMeister Succeed$_{TM}$ for Compliance(注3)の機能として搭載され，活用されている。

3.2　情報の効率的な管理に役だつ業務文書分類技術

業務文書の内容を把握し活用するためには，大量の業務文書を分類して管理することが必要になる。分類を行うためには，業務文書の活用目的に添って分類構造（カテゴリーやその階層）を決定するとともに，各文書をそこへ割り付ける作業が必要になる。しかし，この作業を人手で行う場合は，以下に示すとおり，大きく二つの課題がある。

(1)　分類構造を決めるのが難しい　　どのようなカテゴリーを作れば適切か，分類対象となる文書群の内容を網羅的に考慮してカテゴリーを決めるのは難しい。事前に必要と予想されるカテゴリーを設定しても，どのカテゴリーにも分類されない文書が多量に存在する場合は，新カテゴリーを作るなど，分類構造を見直しながら作業を進める必要がある。

(2)　大量の文書を分類するのは非常に手間が掛かる
　　各文書を分類構造へ割り付けるためには，個々の文書の内容を理解して，分類すべき適切なカテゴリーを判断する必要があり，非常に手間が掛かる。

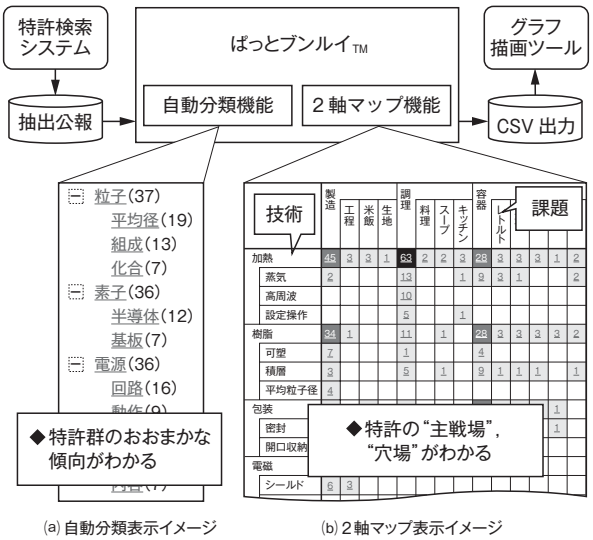
これらの課題を解決する一つのアプローチとして，文書クラスタリングをベースとした業務文書分類技術の開発に取り組んでいる。

文書クラスタリングは，あらかじめ分類構造を決めずに，内容が類似する文書どうしをグルーピングしていくことにより，ボトムアップ的に分類構造を作り上げる手法である。この技術をベースとすることにより，ユーザーが分類対象文書群の内容を知らなくても，システムが分類構造を生成し，分類結果を提示することができる。

当社は，文書クラスタリング技術を活用した特許自動分類システム“ぱっとブンルイ$_{TM}$”[5]を開発し，自社内の特許調査活動の効率化や，特許マップ作成時の分類軸検討などで活用している。ぱっとブンルイ$_{TM}$の機能構成概要と，二つの主要機能による表示イメージを**図3**に示す。

ぱっとブンルイ$_{TM}$は，外部の特許検索システムで粗く絞り込んだ数千～数万程度の抽出公報を対象に，特許公報中の“課題”と“技術”の要素それぞれについて，特許文書に含まれる単語間の関係に基づき分類構造を自動生成したうえで，各特許文書を自動分類する（図3(a)）。分類構造の生成では，特許情報で適切な分類結果を得るため，重要語・不要語辞書などの工夫を加えているほか，課題や技術など複数の観点で，それぞれ有用なカテゴリーが生成されやすくなるように文書クラスタリングを改良している。また，どのような課題にどのような技術が利用されているかを俯瞰（ふかん）する，2軸マップを自動生成する機能を備えている（図3(b)）。

ユーザーは，自動分類の結果から自分の目的に合ったカテゴリーを選別していく過程で，事前には思いつかなかった課題や技術のカテゴリーを発見できる。また，重要語や不要語，関連語などのパラメータを追加して手軽に分析を繰り返し，精



(a) 自動分類表示イメージ　　(b) 2軸マップ表示イメージ

CSV：Comma Separated Values

図3．特許自動分類システム“ぱっとブンルイ$_{TM}$”— ぱっとブンルイ$_{TM}$は特許文書から自動生成した分類構造を基に特許文書を自動分類し，更に，技術と課題の2軸マップを自動生成できる。これにより，効率的で質の高い特許調査を支援する。

PatBunrui$_{TM}$ automatic classification system for patent documents

(注2)　業務プロセス上のリスクと各リスクに対するコントロール（統制活動）を記述したもの。
(注3)　東芝ソリューション（株）の知識継承ソフトウェアのJ-SOX法（日本版金融商品取引法）対応オプション。

度を高めることができるので，効率的で質の高い特許調査ができる。

ぱっとブンルイ$_{TM}$の文書クラスタリング技術及び2軸マップ生成機能は特許文書以外にも適用可能であり，機能や性能面での見直しを加えたうえで，KnowledgeMeister Succeed$_{TM}$の文書分類と分析機能に搭載している。

一方，文書クラスタリングは文書群全体の傾向を手軽に把握できる反面，全自動で分類構造を決定するため，すべてのカテゴリーがユーザーにとって有益なものになるとは限らない。そこで現在，システムが提示する自動分類結果をユーザーが自由に編集でき，編集操作を通じてユーザーの分類意図をシステムに伝え，意図に添った形に自動的に再分類し，システムとユーザーの協調作業により分類構造を完成させていく対話型文書分類技術の開発に取り組んでいる。対話型文書分類技術については，この号のR&D最前線“ユーザーの意図を反映した対話型文書分類技術”(p.58–59)を参照願いたい。

3.3 狙いどおりの情報を検索するパラフレーズ検索技術

ビジネスか日常生活かを問わず，多くの場面で検索機能が利用されており，様々な検索技術が開発され，利用者のニーズに応えている。しかしなお，検索者の意図に対して，本来出てほしかった文書が検索結果に含まれない，期待と異なる文書が検索結果に含まれるといった問題が生じる場合がある。特に業務においては，詳細な情報の収集が必要となり，情報を狙いどおりに検索したいというニーズが大きい。

このような問題の解決方法の一つとして当社が取り組んでいるのが，パラフレーズ検索技術[6), (7)]である。パラフレーズとは，同じ意味を表す様々な表現，すなわち言いかえのことである[(8)]。

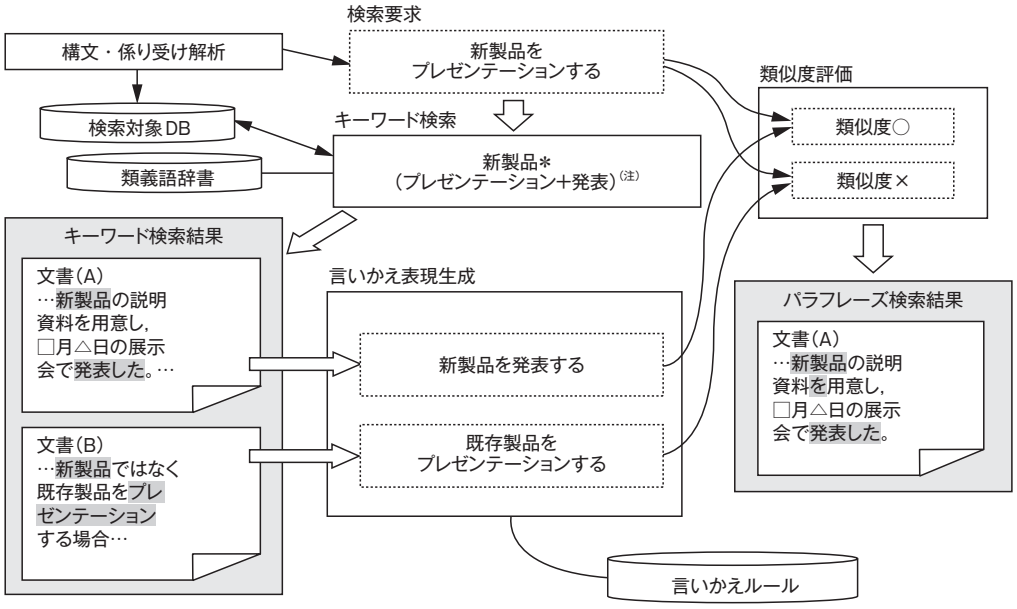
例えば，以下に示すような表現はいずれもパラフレーズに当たる。

文書を印刷する ⇔ 文書が印刷される
雑誌を購入する ⇔ 雑誌を買う
蛍光灯のスイッチを切る ⇔ 蛍光灯を切る

パラフレーズを適切に処理できれば，どの表現がパラフレーズであり，どの表現がパラフレーズでないかを判断することができ，検索の再現率（本来出てほしかった文書が検索結果に含まれる割合）や適合率（検索結果の文書の中で正解の割合）の向上につながる。

当社のパラフレーズ検索システムの概略を**図4**に示す。このシステムの特長は，まず，いったん入力された検索要求からキーワード検索を行い，次に，キーワード検索結果を別の表現に言いかえて，言いかえ結果が検索要求と類似しているかどうかを調べることで，検索意図に合致した文書を選択できる仕組みとなっていることである。

このシステムにおいて，言いかえ表現の生成には，これまでに培った日本語解析技術（形態素解析，構文・係り受け解析）による，より高品質な結果を利用している。特許文書に対して行った評価実験では，検索要求に対し，検索結果がキー



DB：データベース
（注）＊はAND条件，＋はOR条件

図4．パラフレーズ検索システム — 入力された検索要求からキーワード検索を行い，次いでキーワード検索結果を別の表現に言いかえる。言いかえ結果が検索要求と類似しているかどうかを調べる。

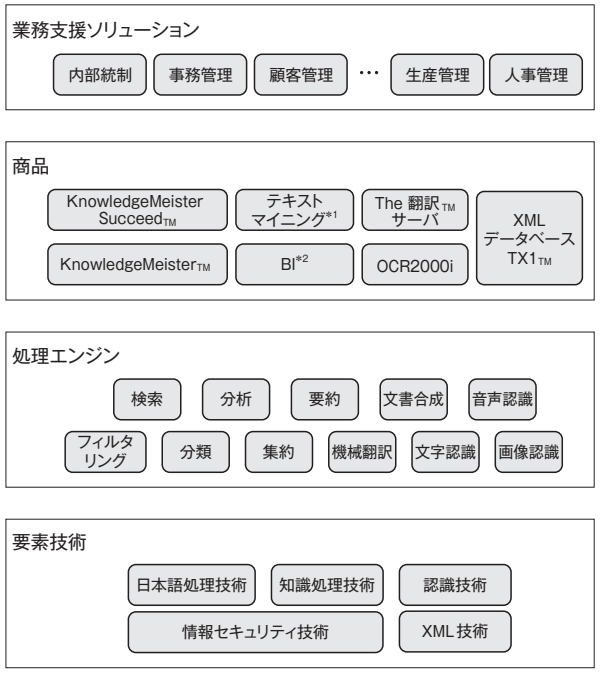
Document retrieval system based on paraphrasing

ワード検索の約半分に絞り込まれ，かつ目的とする文書の約90 %の検索に成功する，という結果が得られた。このシステムは，現在，特定業務向けシステムのモジュール機能として利用中である。

4　情報知識利活用ソリューション体系

3章で述べた技術はいずれも，東芝 研究開発センターにおける要素技術研究の成果を活用している。

これまで述べた情報知識利活用ソリューションの開発は，要素技術がベースにあり，それらを活用した処理エンジンを商品に適用し，業務支援ソリューションとして提供する，という体系の下で取り組んでいる。その体系の概要を**図5**に示す。ここで述べた3種の技術は，処理エンジン部分に相当する。



*1：文書データを意味や内容で分析する手法を用いたシステム
*2：ビジネスにおける意志決定の高度化と迅速化を支援する情報システムの総称
BI　：Business Intelligence
OCR：Optical Character Reader
XML：Extensible Markup Language

図5. 情報知識利活用ソリューション体系とコア技術 — 知識経営を支援するため，情報知識利活用のベースとなる要素技術を用いて処理エンジンを開発し，更に，それらを取り入れた商品や業務支援ソリューションを提供する。
Core technology and solution system for information knowledge utilization

5　あとがき

ここでは，日本語解析技術をベースとした文書処理技術の研究成果として，業務文書チェック技術，業務文書分類技術，及びパラフレーズ検索技術について述べた。

今後も当社は，研究開発の成果や社内外での活用で得られたノウハウを活用し，情報知識利活用という概念の下で各種技術の実用化を進め，業務支援ソリューションを提供していく。

文　献

(1) 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.
(2) 谷口裕子, ほか. 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.
(3) 早川ルミ, ほか. “読影レポートを対象とした文書チェック技術”. 言語処理学会第14回年次大会発表論文集. 東京, 2008-03, 言語処理学会. 2008, p.404-407.
(4) 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.
(5) 平　博司, ほか. 特許調査に役だつ特許情報分類技術. 東芝レビュー. **62**, 2, 2007, p.68-71.
(6) 斎藤佳美, ほか. “言い換え処理技術の文書検索システムへの適用”. 言語処理学会第14回年次大会発表論文集. 東京, 2008-03, 言語処理学会. 2008, p.794-796.
(7) 齋藤佳美. 言い換え処理を適用した文書検索. 東芝レビュー. **63**, 2, 2008, p.74-75.
(8) 乾健太郎, ほか. 言い換え技術に関する研究動向. 自然言語処理. **11**, 5, 2004, p.151-198.

早川　ルミ　HAYAKAWA Rumi
東芝ソリューション（株）IT技術研究所 研究開発部主任。情報知識利活用の研究・開発に従事。日本統計学会，人工知能学会，言語処理学会会員。
Toshiba Solutions Corp.

松本　　茂　MATSUMOTO Shigeru
東芝ソリューション（株）IT技術研究所 研究開発部参事。情報知識利活用の研究・開発に従事。
Toshiba Solutions Corp.

齋藤　佳美　SAITO Yoshimi
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。自然言語処理分野の研究・開発に従事。情報処理学会，言語処理学会会員。
Toshiba Solutions Corp.