

表層表現抽出と論理構造解析に基づく規程文書構造化技術の開発

布 目 光 生[†] 後 藤 和 之^{††}

近年、様々な社会的背景を元に、企業における内部統制対応が、社会的にも法的にも求められるようになった。その対応の一つとして、企業には、自社で定められている膨大な量の規程文書を、必要に応じて参照したり、また部門の改変や業務自体の変革に伴って最新の状態に更新したり、さらに一つの文書を改訂することで変更が波及する他の文書を正しく把握する必要などが求められている。こうしたニーズに対応することを目的として、我々は、企業内規程文書において閲覧、改訂などの活用支援を行う文書管理システムの試作を行った。本論文では、この文書管理システムの内部で用いられている文書構造化モジュールについてその詳細を述べる。文書構造化モジュールは、規程文書から抽出されたテキストを入力として、文書中の構造や意味に基づいたタグを付与し、事前定義されたタグ仕様に基づく XML 文書を出力することを特徴とする。モジュールの精度評価を目的として、実際の規程をターゲットとして評価実験を実施し、タグ平均で 85%の付与精度が得られたことを確認した。また、このモジュールで XML 化された規程文書を活用する GUI プロトタイプを作成した。このプロトタイプにより、当初の規程文書活用に対する閲覧・検索支援といったニーズが機能として実現されていることを GUI 上で検証した。

A document structuring system for company regulations based on keyword extraction and document structure analysis

KOSEI FUME[†] and KAZUYUKI GOTO^{††}

It is now mandatory for companies domiciled in Japan to prove that their internal controls on a consolidated basis are in compliance with laws and regulations. To assist companies, we made a prototype of a document structuring system that supports practical use such as reading of documents and the revision of the management of regulations documents in companies. In this article, we provide a detailed description of the document tagging module used in this document system. Text extracted from regulations is input to the document tagging model. Then, the module gives text a tag based on its structure and the semantic meaning of the terms in the text and finally outputs an XML document based on predefined tag specifications. To evaluate the module's precision, we performed an evaluation experiment using actual regulations and confirmed that tagging precision is 85% on the average. In addition, we made a GUI prototype to utilize regulations output as XML documents by this module. Using this prototype, we intend to develop a GUI-based document structuring system equipped with practical functions, such as browse and search support, for handling regulations.

1. はじめに

近年、世の中の経済・投資活動への関心の高まりもあり、企業の不正会計や粉飾決算などにまつわる事件

が、国内外また企業規模の大小を問わず、各種メディアに取り上げられる機会が多くなった。その結果、こうしたルール違反を未然に防ぎ、健全な企業活動や市場取引を継続してゆくことを目的として、企業には法令遵守（コンプライアンス対応）の体制が強く求められるようになってきている。企業に対し、法令を遵守させるための具体的な方策の1つとして、いわゆる日本版 SOX 法による『内部統制』構築の義務化が試みられている。

これは、日々の企業活動が正しく行われているかを企業自身が検証できるように、事前に業務プロセスを文書化し、潜在的なリスクを洗い出したり、企業自身の決まりである『規程文書』に則って、業務が正しく行われているかを確認できる仕組みを作ったり、さら

[†] 株式会社東芝 研究開発センター 知識メディアラボラトリー
〒 212-8582 川崎市幸区小向東芝町 1
Knowledge Media Laboratory, Corporate Research & Development Center, Toshiba Corporation
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 212-8582
^{††} 東芝ソリューション株式会社 IT 技術研究所
〒 183-8512 東京都府中市片町 3-22
Advanced IT Laboratory Toshiba Solutions Corporation
22, Katamachi, 3-Chome, Fuchu-shi, Tokyo, 183-8512

には外部監査の要求に応じて帳簿や領収書など必要なエビデンスを簡単に提示できる仕組みをつくり運用してゆくことを目指すものである。この『内部統制』は、一度社内にシステムを構築すればそれで完了するものではなく、そこがむしろ出発点となって、継続的に日々の業務活動が規程文書や業務マニュアルに沿ったものであることを確認したり、あるいは逆に、業務内容の変化に応じて規程文書やマニュアルを改訂したり更新してゆく必要がある。すなわち、実際の運用においては企業内の膨大な規程文書や業務マニュアルを適切に管理しておき、関連箇所をすばやく参照したり、更新や修正に伴う規程文書間の整合性を維持するといったことが求められている。

我々は、こうした背景に対し、従来より自然言語処理関連技術の応用として、意味クラス解析技術¹⁾や、オフィス文書に対する構造化、また Web 文書に対する構造化技術²⁾などを開発してきた。

本論文では、ターゲットとなる文書を規程文書と設定した上で、構造化によるユーザメリットを明らかにすると共に、構造化システムの試作を通じて、システムの概要を述べる。また、実規程を対処とした試作モジュールの構造化評価結果、また今後の課題について述べる。

2. 基本方針

規程文書を活用していく上で、本システムでは規程文書を半自動で XML 化し、XML データベースに登録した XML データに対する検索や編集として、閲覧や編集支援を行うという構成を用いた。

この方針においては、下記の点が重要である。

タグ仕様の設定：規程文書に対し、様々な活用ニーズに応えることを可能とするタグの定義すなわちスキーマの策定が必要である。

構造化の精度：次にそのスキーマに基づき、システムがタグを十分な精度で付与することが必要である。

活用例の検証：タグ仕様に沿って生成された XML 文書が、当初のニーズに沿った閲覧支援や検索といった活用を支援しているかの検証が必要である。

上記を踏まえて、本システム試作にあたっては、タグ仕様の策定、文書構造化モジュールの試作と評価実験、検索 GUI プロトタイプの試作と機能検証、という 3 つのフェーズを考慮したシステム試作を行った。

2.1 規程文書を対象としたタグ仕様の検討

現在のところ、文書の構造化（タグ付け）に関しては、様々な団体・組織によって、各種の仕様が提案されている。例えば、日本語テキストを対象とした情報検索・抽出のコンテストである IREX (Information Retrieval and Extraction Exercise) では、組織名、人名、地名、固有物名、日付表現、時刻表現、金額表現、割合表現など 8 種類の固有表現が定義されている。ま

た、DCMI (Dublin Core Metadata Initiative) により定義されている Dublin Core では、ウェブ文書の書誌情報を記述する目的で、title、creator、subject、date、type など 15 種類の基本要素タイプ（さらに精密化の属性が存在する）が定義されている。また、製薬・放送・金融・流通など様々な分野で業界やタスクに応じた構造化定義がなされている。

このような構造化定義は、実際に業界標準として欠かせない役割を担っているものから、実運用でのメリットや敷居が高く、事実上形骸化している定義もある。

こうした現実を踏まえ、まず構造化のニーズ目的は何であるか、またその構造化が付与されたことによりどのようなユーザメリット、ユースケースが創出されるかを考慮し仕様策定を行った。

本章ではその概略を述べる。

まず、関係者へのヒアリングにより、文書ライフサイクル、および想定ユースケース（ニーズ）に関する要求項目の抽出および分析を行った。

2.2 規程文書のライフサイクルと要求項目

規程文書のライフサイクルに関する VoC 収集結果を図 1 に示す。この結果文書管理には、作成、登録、閲覧、改訂更新、(廃棄) の各フェーズが存在することが判明した。

また、これに対応して、獲得できた要求項目の例（抜粋）を以下に示す。

- 部署名や業務を手がかりにして文書を閲覧したい
- 作業や行動、アクションの単位を手がかりにして、文書を検索したい
- 作業や担当者の役割を手がかりにして、文書を閲覧したい
- 類似の業務を、他部署の規程から探して参考にしたい
- 文書中出现している関連文書に簡単にアクセスしたい
- 閲覧すべき文書をナビゲートして欲しい

こうした要求を実現するために必要な文書中の要素と、必要なタグ名を定義すると共に、それらを重要度と抽出コストの観点から分析を行った。

分析の結果として、下記の結論が得られた。

- 高度な意味抽出が必要とされる
 - － action … 行動、動作の内容
 - － object … 取り扱い対象とする事物
 - － subject … 行動の主体
 - － if/then … 条件記述と、その結果や指定された行動

といったタグ名に関しては、抽出コストが高い一方で、実現した場合は、既存のキーワード検索とは異なる検索手段や結果提示手段を提供することが可能となり、価値が高い。

なお、この場合の「抽出コスト」には、抽出精度

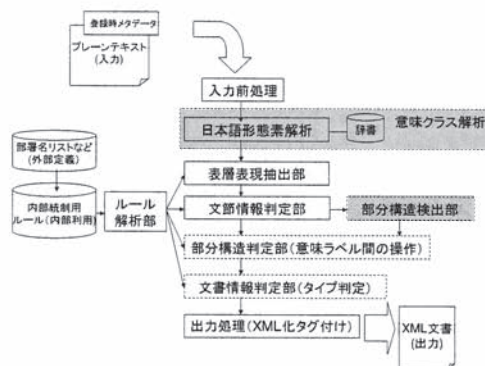
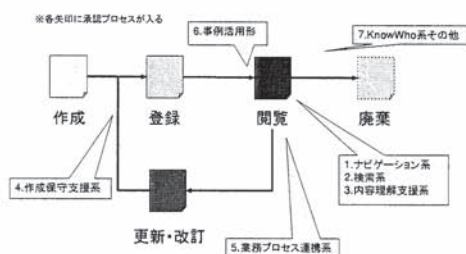
[illegible]

図 2 タグ仕様案

を向上させるために必要な分析や抽出ルール作成コストのほか、原理的に 100%の精度による付与が困難なため、そのギャップを埋めるための代替手段（例えばエディタを用いた手作業による修正）による作業コストを含む。

- 一方で、文書の属性情報である（文書タイトルや、文書 ID、改訂・登録日、作成者）などは、検索の再現性を保証するために欠かせないものであるが、定型的なヘッダ情報からの変換が高精度に可能であったり、修正として手作業による付与が必要の場合でも、文書登録時にフォームなどで容易に入力可能であるという特徴がある。

2.3 タグ仕様の設定

次に、これらの要求項目を満たすために必要となるタグ仕様の案を図2に示す。本仕様案に基づき、今回はカテゴリ A タグを評価対象としたタグを付与可能とする文書構造化モジュールを開発した。

3. 文書構造化モジュール

規程文書を入力として構造を付与する文書構造化モジュールの処理概要を図3に示す

本モジュールは次のような特徴を持つ各処理で構成されている。

3.1 入力前処理

入力文書であるプレーンテキストファイルを読み込む。また、入力文書を後段の各処理で利用可能な形式にするため、文字コードの変換や改行記号の変換などの前処理を行う。

3.2 日本語形態素解析

前段の前処理の結果得られたテキストに対して、形態素解析処理を適用し、解析結果を取得する。解析結果は、形態素ごとに表層表現、品詞、割当済みの意味ラベル、特定の文字列含有の有無などの情報と共に独自形式の内部オブジェクトとして格納される。

3.3 表層表現抽出

内部オブジェクトに対し、後段の意味ラベル付と対象となる範囲を抽出するため、事前定義されているキーワードや品詞パターンを条件としたマッチングを行う。

3.4 文節情報判定

単純に Word 文書をテキスト化した場合には、意味的に続いている「文」の途中に、レイアウト由来の改行+インデントや行そろえのスペースなどが挿入されていることがある。こうした表示上の切れ目に対し、品詞や表層表現の出現パターンから、表示上の切れ目の前後を一文として結合するか否かを判定し、行の結合もしくは分割の保持を行う。

3.5 筒条書き構造の検出

主として記号や数詞表現を含む文頭の表層表現を特定すると共に、事前定義されている制約(1→2→3、a→b→c、A、1→A、1、1)などを適用して、文書中に出現する箇条書き表現を箇条書きグループ(クラス)として抽出する。抽出されたクラスにはそ

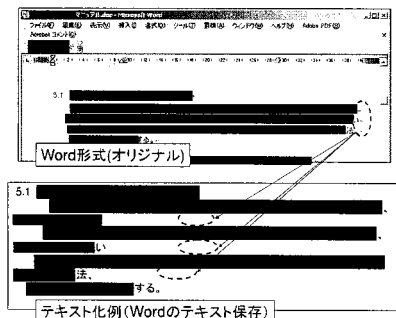


図 4 オリジナル文書のレイアウト表示に由来する改行やインデントの具体例

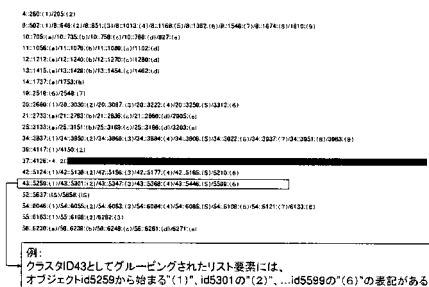


図 5 箇条書き構造検出の中間結果

それぞれ個別の内部 id が付与される。箇条書き構造検出の中間処理結果を以下に示す。各行が同一のクラスを意味する。行の表記は、先頭にクラス ID、続く数値と文字列が内部オブジェクト ID：表層表現の組になっており、さらに同一グループに属する内部オブジェクトがあれば、スラッシュ区切りで繰り返し列挙されている。

3.6 部分構造の判定

前段までの処理結果を元に、外部定義されているルールを適用して意味ラベルを付与する。ルールにより、品詞の出現パターンや表層表現の出現パターンに基づくラベルの付与、また、一旦付与された意味ラベルを制約として、排他的に他のラベルを付与したり削除する等、一部ラベルを出現条件とした定義が可能である。

3.7 構造化ルールの適用

前段までの形態素解析結果や箇条書き構造検出の結果を受けて、文書中に出現している表層表現や品詞情報

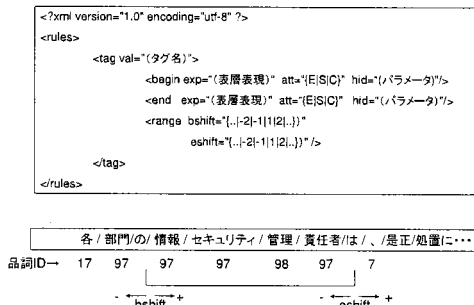


図 6 ルール書式の説明の図

報、意味ラベルの出現を条件としたルールを適用し、さらに詳細な意味ラベルを付与する。以下に現在のルール書式の例を示す。

ルールは、将来的な新規機能の追加との互換性を考慮し、XML 形式での記述としている。本ルールの基本的アプローチは、

- (1) 意味ラベルが含まれる表層表現や品詞を定義する
- (2) 意味ラベルの範囲（境界条件）を定める
- (3) 対象とする意味ラベルの近傍や隣接条件を定める

といった各条件定義を簡素に忠実に反映することである。これは、本定義で用いられている各要素は、例えばルールを人手作成する場合には、付与したい文字列に対して直感的に範囲指定を行うことが可能なため、GUI ルール作成ツール上などでのルール作成手段との対応付けが良い。また、将来的にはルール作成ツールの付加機能として、タグ付き文書を正例として入力し機械学習によってルールを自動生成することが考えられるが、この場合でも SVM や決定木などにより、タグと対象文字列（品詞・表層表現パターン）のルールを学習させる場合の fact（学習パラメータ）と対応しており、拡張が容易であると考えられる。また、形態素解析辞書の拡充や増強などにより、特定分野の辞書の拡張やプロパティ情報の付加などが行われた場合にも、上記の形式により直接的に拡張された品詞情報を利用可能である。

ここで、図を用いてルール記述を簡単に説明する。“tag” 名の要素が、ルールの最小構成単位となっており、これらを宣言的に記述してゆく。属性 “exp” は表層表現、“hid” は品詞 ID を示している。また、要素名 “begin”、“end” はそれぞれ開始点と終点の基準を定義する。“att” 属性はその詳細化定義で、指定の表現で始まる (S=StartWith)、指定の表現で終わる (E=EndWith)、指定の表現を含む (C=Contains) といった定義を行う。また、正規表現による指定 (R) も可能である。要素 “range” では、始点と終点の境界条

```

<rules>
...
<tag val="person" exp="" hid="33" />          ... (1)
<tag val="person" exp="" hid="34" />          ... (2)
<tag val="company" exp="" hid="35" />         ... (3)
...
<tag val=">title" >                          ... (4)
...
  <begin exp="T" att="E" />                    ... (5)
  <begin exp="[0-9]{1,3}" att="R" hid="1" />    ... (6)
  <end exp="%" att="R" />                      ... (7)
  <range bshift="1" eshift="-1" />            ... (8)
...
</tag>
...
</rules>

```

図 7 ルール記述の例

件を精密化する。属性 “bshift” で始点を、“eshift” で終点の精密化を行うが default 値は 0 である。負の値で文書の先頭方向で、正の値で文書の後方を指定することを意味する。なお、“tag” 以外の要素は省略可能であり、図「ルール記述の例」では、(1) – (3) ではそれぞれ品詞 ID を直接対応付けている例である。また、(4) – (8) では、他の要素を用いた記述例であり、正規表現を用いて始点終点を定義している。

3.8 出力処理 (XML タグ付け)

以上の処理結果に基づき、内部オブジェクトに対し意味ラベルが割り当てられていることになる。出力処理では、最後にこれを XML タグとしてテキストと組み合わせて XML 文書を出力する。文書前方から、意味ラベルが割り当てられているオブジェクトを走査し、意味ラベルが適用されている範囲の検出を行う。XML として開始タグと終了タグを挿入する位置、また内部定義されているタグ名の優先順序に基づきテキストにタグの付与 (挿入) を行い、XML 文書生成する。付与結果は既存の XML パーサ (今回の実装例では MSXML) を用いて整形形式のチェック経た後、ファイルに出力される。

4. 構造化に基づく応用の具体例

以上の構造化モジュールにより、規程文書に対する半自動のタグ付与が可能になるが、この構造化結果が、当初のニーズに沿っているかを検証することを目的として、ユースケースに基づく活用支援例を表示可能な GUI プロトタイプを試作した。試作した GUI ツールによる具体例を抜粋して示す。

4.1 概観表示機能

図 8 では、規程文書を構造化した結果に対し、書誌情報の他、文書中から抽出できた参照元や参照先の表示、また規程本文に対しては見出し表現や章構造といった領域に対し、スタイル定義により装飾を施し、ブラウザを介してユーザへ提示した例である。

ヘッダとして書誌情報が明示できるほか、文書中の記述から抽出された参照元リンク・参照先リンクを提

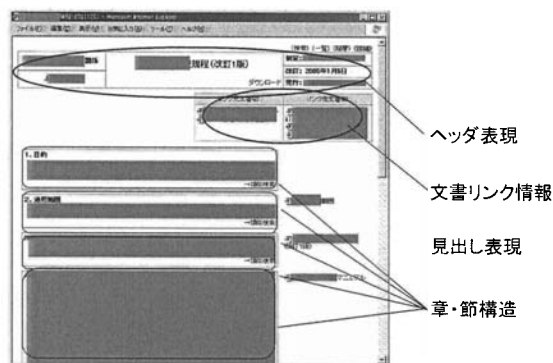


図 8 構造化結果の HTML 表示の例



図 9 規程文書間の参照関係を提示した例

示することにより、文書の詳細とその関連情報を概観することができる。

4.2 文書間リンク

図 9 では、各規程文書に対して参照されているリンク元、および規程文書から参照しているリンク先の文書間関係を枚举した表示例である。

閲覧者や文書管理者は、各文書から参照関係にある文書を連鎖的に辿ることによって目的の文書に到達することができる一方で、把握していなかった参照関係の確認を行うことができる。

4.3 役割名に基づく検索支援機能

図 10 は、検索支援の例である。従来のキーワードによる検索手段に加えて、キーワードの意味 (ここでは役割名として判定されている) を提示することにより、検索結果に対して、さらに役割名の観点から関連している文書をハイライトしたり、絞り込み検索を行うことが可能である。

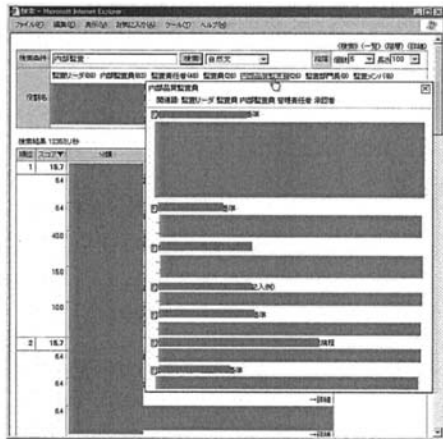


図 10 役割名に基づく検索支援の例

5. 評価実験

文書構造化モジュールの評価実験として、企業内規程文書をターゲットとした精度評価を実施した。まず準備として、対象文書のオリジナルは Word 形式のため、Word でファイルオープン後、テキストファイルとして保存を行った。保存時の設定は以下の通りである。

- 保存形式：テキスト（書式なし）
- エンコード方法：Windows(既定値)
- 改行の挿入：なし（既定値）
- 文字の置換を認める：なし（既定値）

対象ファイルは既定の 8 カテゴリーから、カテゴリを代表する規程文書のほか、ランダムに加えた 20 文書を選定した。入力文書のファイルサイズ（テキスト変換後）は（2KB～51KB）である。

この文書に対して、カテゴリ A タグの付与を計測した。図 11 に、タグ別の付与精度グラフを示す。

実験の結果、まず、基本情報タグの一部や文書 ID 表記に関しては、文書中に該当表記が出現するパターンが比較的類似しており、高精度な抽出が可能となった。また、日付表現とその近傍の表記を手がかりとした、制定日や改訂日なども、抽出対象が同様な性質を持っていることから、同等の性能が得られた。一方で、文書中の文書名や、冒頭のタイトル表記に関しては、文書毎の差異を吸収できないケースが出現している。

5.1 改良方針と課題

以上のように基本情報を対象としたタグでは比較的精度が得られている一方で、カテゴリ B タグのより高度な意味的表現の抽出を実現する上では、以下の項目が課題として挙げられる。

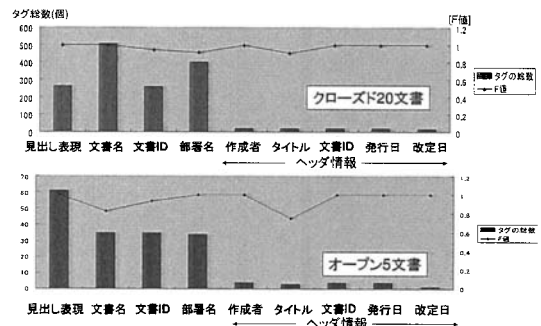


図 11 クローズド 20 文書およびオープン 5 文書に対する試作システムの評価結果

- 複数文書の統計処理によるダイナミックなノイズリダクション（一部、箇条書き表現の抽出などでは実装済み）
 - 掛かり受け解析など高度な日本語処理技術の適用
 - 初期、運用ルール作成編集支援、知識構築支援機能
- 今後の課題としては、本モジュールによるルール記述のロバスト化や記述力の検証、またルール作成コストの観点からみた、有用性の検証などがある。

6. まとめ

規程文書の構造化技術開発として、以下を実施し、規程文書構造化に対する知見を得ることができた。

- (1) 規程文書のライフサイクルや利用形態に基づくニーズヒアリングを実施し、実規程文書を入力サンプルとした構造化の仕様策定を行った
- (2) 企業内規程文書を対象としたタグ付けモジュール、ルールの試作を行った
- (3) タグ付けモジュールの出力結果を容易に確認可能とするデモアプリの試作を行った
- (4) 既定文書サンプル 20 文書に対し、クローズド評価で適合率 0.93、再現率 0.86 を実現した

今後、試作システムの仮運用フェーズを通じて、新たなニーズを獲得しシステムの精錬化を継続的に実施してゆく必要がある。

参考文献

- 1) 市村由美, 齋藤佳美, 酒井哲也, 國分智晴, 小山誠, 質問応答と、日本語固有表現抽出および固有表現体系の関係についての考察, 情報処理学会研究報告, Vol.2004, No.47, pp.17-24, 2004.
- 2) 布目光生, 石谷康人, 表層表現抽出と文書構造化解析に基づく XML 文書変換システム, 情報処理学会研究報告, 2004-DD-46, pp.1-8, 2004.