# 営業活動の蓄積情報を利用した自動要約・知識提示に関する研究

## A Study of Document Summarization for Knowledge Presentation from Stored Information of Business Activities

今野陽子[*1] Yohko Konno
中村拓哉[*1] NakamuraI Takuya
吉田優之[*2] Masayuki Yoshida
川村秀憲[*2] Hidenori Kawamura

[*1](株)調和技研 Chowagiken Corporation
[*1](株)調和技研 Chowagiken Corporation
[*2](株)北海道アルバイト情報社 HAJ Corporation
[*3]北海道大学 Hokkaido University

This study examines a method of effective utilization as knowledge in organizational stored information of business activities. As an approach, a information is presented by the document summarization using the real bulletin board system data. For a question sentence, the system retrieves information based on categories and feature words, and documents are summarized.

## 1. はじめに

Web の投稿データから，有益な情報を発見し応用する動きが活発である．一方で企業や組織の業務においても，日々多くのデータを記録している．例えばメール，電子掲示板(BBS: Bulletin Board System)，営業日報や議事録などの文書においても，重要なノウハウとなる情報が記録されている．現在，このような組織の蓄積情報に着目し，形式的な知識を可視化するための研究，特定業務における成功要因やリスク管理などのナレッジマネジメントに活用するサービスにニーズが高まっている(図 1)．これらの記録は，時系列に蓄積され，内容も多岐に渡る．キーワードによって検索しても，文書がそのまま表示される場合が多く，重要性の高い情報がどこに含まれているか把握できずに埋もれてしまう可能性が高い．本研究では，このような組織で蓄積された大量のテキストデータを対象として，社内の知識として有効活用することを目的に，その方法について検討する．本研究のアプローチとして，ユーザは質問文とカテゴリを入力し，システムはあらかじめカテゴリが付与された情報の中から，該当のカテゴリと，質問文に含まれる特徴語に基づいて，必要な情報を抽出する．抽出された情報は，自動要約により重要な部分のみを取り出し，質問への回答として提示する．情報検索の際に，特徴語と併せて，過去の掲示板データにおいて類似の意味で使用される語も検索のキーワードに含めて情報を抽出する．また自動要約では，特徴語に関係の強い文を優先的に採用する．実際に企業で蓄積された電子掲示板のデータを取り上げて評価を行った．最終的には社内で活用される自動要約・提示システムを目指している．

## 2. システムの構成

本研究で対象とするデータ は，アルバイト求人情報を扱う企業の社内スレッド型電子掲示板に投稿された 2007 年から 2017 年までの 36,000 件のテキストデータである．この掲示板は主に営業活動に関する社内の情報共有のために設けられ，効果的な事例や地域の顧客情報，企画の告知などが自由形式で投稿される．営業活動は，経験に基づく属人的成果が大きい．これまでは投稿データを整理や再利用することは無かったが，営業成果のノウハウが蓄積されていることから，営業活動の支援や標準化に役立てたいという要望がある．本研究では，質問に適した文書要約[平尾 2001]に着目し，具体的なサービスとして，質問に対して特定のキーワードをクエリとした情報検索を行い，該当情報を要約して提示するシステムを開発した．このようなシステムは，ユーザの問い合わせに対して大量のコーパスから確信度の高い回答を提示する対話エージェントへの応用も考えられる．図 2 は，自動要約・提示システムの構成であり，システムは大きく 5 つのプロセス(A)～(E)から構成される．

(A) 質問文を入力し，知りたい情報のカテゴリを選択する
(B) 質問文から特徴語を抽出する
(C) カテゴリと特徴語を用いて情報検索する
(D)検索でヒットした投稿について，自動要約を行う
(E)自動要約した結果を提示する

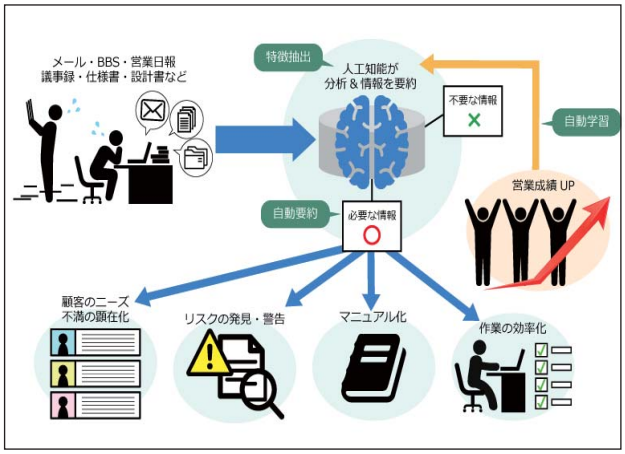


図 1　蓄積文書の活用イメージ

### 2.1 情報検索

上記(A)(B)について，効率的に情報を取得するためには，ユーザは必要なものは何かを示すことになり，また，多くの情報の中から必要なものを取り出すためには，キーとなる特徴を決めることが条件となる．特徴語は文を特徴づけるために掲示板データから抽出した語である．この特徴語の機能は，質問文からの検索で抽出する情報を的中させること，および質問に応じて要約内容を推定することである．

連絡先:[*1] 今野陽子，株式会社調和技研，北海道札幌市北区北 21 条西 12 丁目 2 北大ビジネススプリング 204 号室，ykonno@chowagiken.co.jp

(C)において，検索する掲示板のデータは，ひとつの投稿を 1 レコードとして，クエリに該当するレコードを全て抽出する．質問文に例えば「派遣」という特徴語が含まれる時，それは「求人情報」の話を指しているのか，あるいは「人材紹介」の話か，「派遣」という語だけで推測することは難しい．そのため，特徴語のみを検索の条件にするのではなく，あらかじめレコードごとにカテゴリを付与し，特徴語でレコードを抽出する前に，検索の候補となるレコードを絞り込むようにした．そして(D)において，レコードごとに特徴語を考慮してリアルタイムに自動要約した結果を表示する．また詳細に投稿の内容を確認できるように，アコーディオン形式で本文を表示できる機能を持つ．本システムを実現するために必要な要素技術として，大きく 3 つの機能(1)特徴語抽出，(2)カテゴリ分類，(3)自動要約が求められる．それぞれについて 4 章，5 章，6 章で説明する．

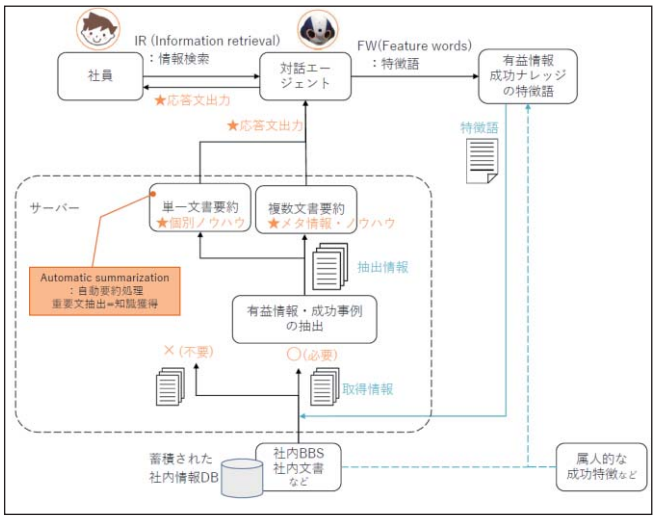


図 2　自動要約・提示システムの構成

## 3. 本研究の位置づけ

本研究における提示の部分となる自動約に着目して，要約の方法を検討する上で，その目的を明確にする[奥村 2005]．一般に要約の目的は「指示的要約」と「報知的要約」に大別される．指示的要約は，元の文書を読むべきかどうかを判断するための要約の事を示す．例としては新聞の見出しの一文が挙げられ，必要に応じて本文を読むことを前提としている．一方で，報知的要約は，元の文書の概要を伝える要約で，テレビニュースの字幕などが例として挙げられる．本研究で想定するサービスでは報知的な場合を取り扱う．次に，入力が単一文書か複数文書かによって分けられる．単一の場合とは，一つの文書から一つの要約を作成することであり，一方で，複数文書から作成する場合とは同じ事柄に対する複数の文書から，一つの要約を生成することをいう．本研究では，ある話題について複数の投稿が存在する社内文書を用いるが，現在は該当する投稿ごとに要約することを想定しているため，単一文書からの要約作成を対象としている．また，文書の要約には，その文書の概要を示す一般的な要約が必要な場合と，利用者がキーワードなどを指定してそのキーワードに関する要約を文書から生成する場合が存在する．ユーザから何らかの指定を受ける場合をクエリ依存，それ以外の場合をクエリ非依存と呼び，本研究ではクエリ依存型をベースとして要約文書を取得するシステムを想定する．そして，要約手法については要約対象の文書の中から重要と思われる文を抽出する抽出的要約と，意味を汲み取り抽象化を行った上で適切な要約を作成する生成的要約が存在する．後者に対しては意味表現の生成文の整合性等の問題があり非常に困難であるため，本研究では抽出的要約を行う．

## 4. 特徴語の抽出

本研究における特徴語は，掲示板に投稿されるそれぞれの内容を特徴づける語として位置づけられる．内容を特徴づけることから，特徴語の役割は，情報検索の際に，ユーザの質問文に含まれる語と検索したい情報のマッチングを図ること，また抽出した情報を自動要約する際に，質問文に対して特徴語に関わる文が優先的に取得されるように調整することである．この際，特徴語はユーザの問いの意図を汲み取っている必要がある．本研究のシステムでは，入力情報として質問文に含まれる特徴語を抽出し，ユーザの求める投稿文を検索するためのキーワードとして使用する．特徴語を獲得する方法として，ベース手法として TF-IDF を用いる．また，情報を検索するにあたり，質問文に含まれる語だけではなく，掲示板において似たような使われ方をする語を類義語として，検索のキーワードとして追加する．これは社内用語や業界用語と一般的な語を結ぶ意味でも効果がある．類義語を定義するための手法として，文書の中に現れる単語間の距離を計算してベクトルで表現する Word2Vec を用い，掲示板の全文書を学習する．近傍の語を Cosine 類似度で算出して類義語を取得する[Word 2013]．

検索の性能向上を図るため，いくつかの工夫が必要となる．一つ目に，質問文の内容から，カテゴリを推測して検索対象のカテゴリに加える．この方法として，予めカテゴリ分けした投稿データ(5 章) から，各カテゴリに頻出する語を抽出する．質問文の中に，あるカテゴリと関わりの強いカテゴリ頻出語が含まれる場合，ユーザによって該当のカテゴリが選択されていなくても(図 3)，検索対象として加える．また，例えば「活動」などの広く使われる語を曖昧語として定義し，曖昧後が質問に含まれている場合には，検索のキーワードから除くようにする．このようにして抽出された特徴語は，類義語の組を or で結び，類義語の組同士を and としてクエリ検索を行う．各語には，文書間の類似度に基づく優先度を付与して，自動要約処理における文の重要度計算に利用する[森 2002]．例えば，ある特徴語の優先度=1.0 の場合，要約処理では基本重要度=1.0 と設定し，優先度を加算して重要度=2.0，すなわち基本の 2 倍の重み付けを行った．

## 5. カテゴリ分類

本研究では，情報検索の際に，特徴語による検索と併せてカテゴリを指定することで，取得する情報の厳選を行う．そこで，あらかじめ掲示板の投稿内容に応じて分類を行い，投稿ごと(レコードごと)にカテゴリを付与する．掲示板の投稿内容がどのようなカテゴリに分けられるか，その種類については，掲示板データ提供元の企業にヒアリングを行い，以下の 10 種類のカテゴリを決定した．

(Category01) 営業活動報告
(Category02) 反応事例
(Category03) 採用事例
(Category04) 競合情報
(Category05) 人材紹介
(Category06) 地域情報
(Category07) イベント
(Category08) 特集
(Category09) 業務連絡
(Category10) 求人受注

カテゴリ分類のための手法として，ランダムフォレスト(Random Forest)に勾配ブースティング(Gradient Boosting)を用いるアンサンブル学習を行う．ランダムフォレストは，複数の決定木を組み合わせて，各決定木の分類結果から多数決で決定する手法である．また勾配ブースティングは，前回の学習結果を逐次的に取り入れながら学習モデルを生成する．分類のための教師データは，人手によりカテゴリを付与した 500 件の投稿を用意する．予備実験を行い，分類の精度の比較検証によって，実際に投稿データにカテゴリを付与した．比較した 3 種類の方法は，(Type1)ランダムフォレストにより，投稿ごとに最高スコアだったカテゴリを一つ付与，(Type2)ランダムフォレストにより，投稿ごとにスコアの閾値により複数のカテゴリを付与，(Type3)ランダムフォレストの(2)をベースに勾配ブースティングのアンサンブル学習である．S 分類精度の検証は，人手により正解のカテゴリを付与した投稿 20 件について，(Type1)～(Type3)の実行結果と比較した．(Type1)9.5%，(Type2)71.4%，(Type3)81.0%の精度で分類することが判り，現在のカテゴリ分類には(3)の方法を採用している．

## 6. 自動要約

本研究における自動要約は，3 章で述べたように報知的，単一文書，クエリ依存型，抽出的要約を対象とする．要約の方法として，元の文書を文分割し，要約に含める重要文を選択し，それらをつなげることで要約文を作成した[Luhn 1958]．

要約の処理は，形態素解析，構文解析，そして要約処理の順で行う．形態素解析器に Juman，構文解析器に KNP を用いる[笹野 2013]．文の重要度を計算する方法として，構文解析の出力結果に基づいて，各単語に付与された品詞情報，構文関係情報を用い，単語ごとに重みを付与し，文ごとの重要度計算を行う．また，入力された質問文の特徴語の優先度も，重みに加算する．各文のうち重要度の高い文を，要約文字数制限の範囲で，時系列に取得する．要約文字数は可変として 7 章で評価する．

リアルタイムな要約処理のための高速化の工夫として，蓄積データに対して，KNP による構文解析までの計算を事前に実行しておくことで，要約時に時間短縮を図った．また，自明な定型文，すなわち，お疲れ様です，おはようございます，よろしくお願いしますなどの文は，形態素解析の事前に，入力文から除去した．

## 7. 評価実験

図 3 は本研究で開発した自動要約・提示システムの入力画面である．ユーザが入力時に指定する項目は以下の 4 項目である．

(Input1) 質問文入力 (検索キーワードのみでも可)
(Input2) カテゴリ選択
(Input3) 類義語有／無
(Input4) 投稿期間

自動要約の一般的な評価として，内容，読み易さ，外的評価が挙げられる．本システムを評価する観点として，大きくは 3 つのポイントがあり，(1)要約内容が投稿原文に対して適当であるか，(2)ユーザの質問の意図を汲み取り，有益な情報を提示しているか，(3)システム全体として合理的であるかが考えられる．それらを考慮して，以下の項目について検証を行った．

(Test1) 投稿原文に対する要約文の妥当性(要点を得られる)
(Test2) 要約文の字数制限はどの程度が適当か
(Test3) 類義語を加味する効果はあるか
(Test4) 特徴語に応じて要約内容が調整されるか
(Test5) 重要で有益な情報を提示できるか

上記の(Test1)と(Test2)は関連があり，字数が多いほど原文に近い内容を表すことが可能であるが，あまり重要ではない内容の文も抽出される可能性が高くなる．反対に字数制限を少なくすると，より核心部分を抽出する必要があり，また，文をつなげた時に不自然になる可能性が高まる．検証方法として，社員 10 人へのヒアリングにより 100 字，200 字，300 字について確認を行った．その結果，100 字の要約が咄嗟に見やすいという意見に集約された．投稿原文から抽出される要約文の傾向としては，(1)顧客名，営業署名，個人名が含まれる文が要約文に採用されやすい，(2)100 字の制限により要約文は 2～3 文になることが多い，(3)話題について表す冒頭の文，(4)全体のまとめや結論を表す文を採用する，といった要約結果が大半となった．要点を獲得できるかについて，上記(1)(3)の顧客や話題に関する文，(4)まとめの文を採用することで，100 字内に概ね要点がまとめられた．しかし，掲示板への投稿は自由形式であり，必ずしも 1 つの投稿に 1 つの話題とは限らない．投稿文に顧客名が複数登場したり，複数の話題を扱う場合には，これらの文書の要点を 100 字の制約の中で抽出することが難しい．要約処理を改善するためには，文書分割や，生成的要約の導入が見込まれる．


図 3　自動要約・提示システムの入力画面

図 4 は，掲示板に投稿された情報の中から「美容関係の反応について知りたい」という質問文，および全カテゴリ選択，期間指定なしで検索した場合の実行結果である．上段は類義語=有の場合で，[美容 or エステ or ファッション]and[関係 or 築く or 信頼]による検索結果を示す．下段は類義語=無の場合で，[美容]and[関係]による検索結果を示す．(Test3)について，類義語=有の場合に，検索結果の拡張性を検索ヒット数で計算した場合，類義語=無の場合に対して 5 割増加する．また，図 4 の例で，掲示板データから作成した類義語の辞書において，[美容]の近傍にある語は [エステ][ファッション]と，いずれも美容に近い意味で使われる語を指している．一方で，質問に[美容関係]と入力したにも関わらず，[関係]の近傍にある語は[築く][信頼]と，いずれも類義語ではなく[関係を築く]や[信頼関係]といった共起性の高い語になっている．類義語の辞書を改良するためには，掲示板データの語を学習するだけではなく，語の取捨選択が必要である．

図 5 は入力が異なる質問文に対する，同じ投稿原文の要約結果である．(Test4)について，)図 5 の例では，上段の質問文は「美容」についての反応，下段の質問文は「帯広」の「美容」についての反応を問い合わせした．要約結果の 1 文目は，それぞれ美容に関する顧客であるが，複数の顧客が併記されている場合，下段では帯広に特化することで帯広店の顧客を取り上げ

る. 2 文目はユーザが体験したプラスになる情報と成功体験について述べている. 投稿原文の中でも重要度の高い文といえる. これは質問が異なれば必ずしも要約内容が変わるということではなく, 要約処理の文の重要度計算に依存する.


図 4　類義語を含む検索(上)
含まない検索(下)の実行結果


図 5　質問の違いによる要約結果
美容の反応(上)と帯広の反応(下)

(Test5)については, 要約により, 投稿文書からユーザに有益であり知識となる情報を提示できることが目標である. そのための検証方法として, 実際に社員に対して要約結果の検証を行った. 掲示板は営業活動における情報共有のための投稿であることから,「営業活動の支援になる情報が掲示されている要約文」にチェックするシンプルな指標とした. 結果として, 質問文 10 問中 20%の要約が有益な情報を提示していると評価された. また, 評価されなかった要約内容ついては, (1)営業ポイントではない原文の要約, (2)現状報告, (3)部分的な内容の抜粋, といった内容に大別された.

今後の取り組みとして, 検証結果に基づいて, 詳細に調査を行う. (A)有益な情報の提示に向けた分析を行う, (B)要約の評価手法について検討し, 例えば検索キーワード, 有益に関わるキーワードの出現確率を指標とすることで, 要約結果を評価し, スコアで順位付けして表示することや, 厳選した情報を表示することが考えられる. また, (C)分析, 評価に基づいて, ユーザが意図する情報と質問と関連付けるための特徴語と優先度の算出方法について検討を行う.

## 8. まとめ

本研究では, 組織で蓄積された大量文書について, 社内の知識として有効活用することを目的に, 自動要約・提示システムの検討を行った.

(1) 営業活動の情報共有ツールである電子掲示板の蓄積情報から, ユーザのノウハウとなる有益な情報を獲得するための方法について検討した.

(2) 提案システムは, ユーザが質問文とカテゴリを入力することで, 入力情報に基づいて履歴情報を検索するともに, 投稿された文書の自動要約により重要な部分を回答として提示する.

(3) 自動要約処理分類上の位置づけは, 報知的で単一文書によるクエリ依存の抽出型要約を対象として, 処理方法の検討を行った.

(3) 情報を検索する際には, 特徴語と併せて蓄積された過去データに基づく類義語を加味し, 自動要約においては, 特徴語に関係の強い文を重要な文として優先的に提示するようにした.

(4) 実際の電子掲示板データを用いて検証を行い, 要約内容文の妥当性, 文字数, 特徴語の効果, 有益情報の獲得について評価と考察を行った.

(5) 要約内容の検証結果に基づいて, 今後, 有益情報の可視化に取り組み, 有益情報の評価方法や表示方法について検討することで, システムの改良を行う.

## 参考文献

[平尾 2001]　平尾努, 佐々木裕, 磯崎秀樹: 質問に適応した文書要約手法とその評価, 情報処理学会論文誌 42(9), pp.2259-2269, 2001.

[奥村 2005]　奥村学, 難波英嗣: テキスト自動要約, オーム社, 2005.

[Word 2013]　Word2Vec: https://code.google.com/p/word2vec/, 2013.

[森 2002]　森辰則: 検索結果表示向け文書要約における情報利得比に基づく語の重要度計算, 自然言語処理 9(4), pp.3-32, 2002.

[Luhn 1958]　Luhn, Hans Peter: The automatic creation of literature abstracts, IBM Journal of research and development 2(2), pp.159-165, 1958.

[笹野 2013]　笹野遼平, 河原大輔, 黒橋禎夫, 奥村学: 構文・述語項構造解析システム KNP の解析の流れと特徴, 言語処理学会第 19 回年次大会発表論文集, pp.110-113, 2013.