言語処理で何が出来るか
－自然言語処理技術の紹介－

言語情報研究室　丸元　聡子[*1]
言語情報研究室　木田　敦子[*2]
言語情報研究室　乾　　裕子[*3]

1. はじめに

昨今、コンピュータの記憶容量が膨大になり、インターネットやワープロ、パソコンが普及するにつれて電子化された文書の量が急激に増加している。これらの多くは、プログラミング言語などの人工言語ではなく、英語や日本語のように人間が日常に用いている言語、すなわち自然言語により記述されている。これらの活用を容易にするためには、電子化された文書を事前に処理し、要約や構造化をしておく必要がある。

自然言語は曖昧さや、その時々によって様々な省略や言い替えがあること、さらに社会的な常識が必要なため、コンピュータに理解させるには高度な技術を要する。コンピュータを用いて自然言語を処理することを自然言語処理（Natural language processing）というが、電子化文書を処理する際には自然言語処理技術が重要な役割を果たすことになる。

IBS 言語情報研究室では、自然言語処理に関する基礎研究・基盤整備などを行っている。自然言語処理という語は、一般には馴染みが薄いが、その一部は既に一般に利用されている。例えば、機械翻訳・情報検索などは自然言語処理の技術を実装したものである。こういった自然言語処理分野の応用技術が世の中に出回るようになってきている。

本稿では、情報の活用という観点から、いくつかの自然言語処理技術を紹介する。

2. 自然言語処理とは

言語処理以外の分野でも、日本語を用いている以上、言語処理技術を利用することで情報の活用が容易になり、業務の効率化や新たな発想を支援することが可能である。

多くの文書が電子化されるようになったことで、電子化された文書が大量に蓄積され未整備のままになるという状況も見られる。

そこで、ここでは、溢れる情報の処理・活用を柱に言語処理技術を簡単に紹介する。

情報の処理に関しても、いくつかの異なった技術がある。

・情報検索
・情報抽出
・情報集約

情報検索については既に一般にも浸透してきているが、今後は、その他の技術の利用価値も高まるものと考える。それぞれ、次のような技術である。

情報検索

大量のデータを格納した巨大情報ベースから必要な情報だけを選別して、その全部を許容時間内に取り出すこと。文献検索や特許検索などは既に長期にわたり利用されている。また、最近では、インターネットの検索エンジンでも用いられているものがあるように、全文検索も一般にもよく利用されている。従来の検索では、指定したキーワードと完全に一致したデータしか見つけることが出来なかったが、最近では「あいまい検索」の技術が発達し、意味的に近いものを類推して探し

[*1] まるもと　さとこ（研究員）　[*2] きだ　あつこ（研究員）　[*3] いぬい　ひろこ（研究員）

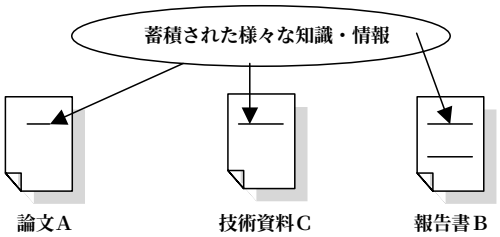
目次へ

出すことも出来るようになってきた。

情報抽出

必要な情報だけを自動的に取り出す技術

例）交通計画の手法に関する情報は？



※該当する文書を自動的に選択し、そこから必要部分だけを抽出。

計算機を用いて情報収集を支援する場合、大量の文書から検索要求に合致した文書を選び出す情報検索（Information Retrieval）と、その文書をさらに加工して必要な部分を取り出す情報抽出（Information Extraction）とがある。

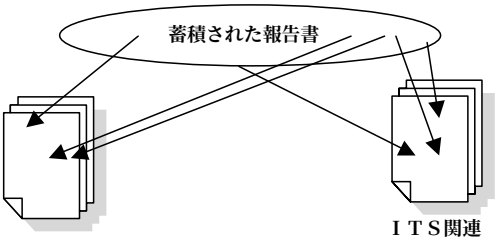
単に文字列や単語を手がかりとした情報検索では、取り出される文書は、必ずしもユーザの意図にそったものとは限らない。例えば、企業合併の情報を得たい場合に、単に「合併」を検索キーとして検索すると「合併症」の情報が取れてしまう可能性がある。このような不要な情報を排除する必要がある。また、必要とする情報がどの程度含まれるかという点で、様々なレベルの文書が混在している。そこで、取り出した文書の中から本当に必要な情報だけを抽出する処理が求められる。これが情報抽出である。

日本語の情報抽出の手法には、１）形態素解析（予め用意した単語辞書と文法規則を用いて、文から単語を切り出す処理）結果を利用するもの、２）表層文字列並びの特定のパターンを認識するパターンマッチング手法で処理を行うもの、がある。

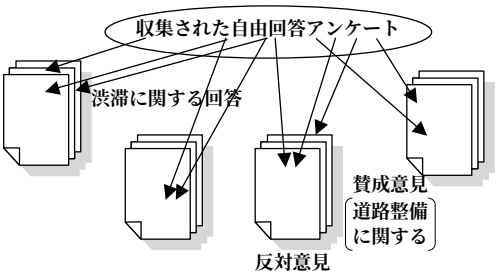
情報集約

類似する情報を自動的にまとめる技術

例）1999 年の報告書をテーマ別に分ける。



例）自由回答アンケートの分類



道路整備に関する回答

※観点・目的別の分類が可能

テキストの自動分類：

テキストの自動分類（text categorization）とは、テキストを予め決められたカテゴリに分類する、あるいはテキストにカテゴリを付与することをいう。例えば、ある新聞記事をその内容に従って「政治」「経済」「社会」などに分類することが考えられる。自動分類の基本的な手続きは次のようになる。

- **各カテゴリを予め内部表現に変換する。**
- **入力テキストを内部表現に変換する。**
- **テキストと各カテゴリ間の類似度を計算する。**
- **テキストに最も類似したカテゴリを付与する。**

テキストやカテゴリの表現形式や類似度の計算方式は用いるモデルによって異なる。代表的なモデルとしては、ベクトル空間モデル、確率モデル、規則に基づくモデル、のような

目次へ

モデルがある。

このような自動分類の手法は、新聞記事のように半定型的な表現で書かれたものだけでなく、割合に自由な表現を取ることの多い、自由回答アンケートの回答などにも適用が可能である[5)]。

IBS 言語研では、３年前から道路局・都市局が公募した道づくりのボイスレポートを対象にアンケートの自動分類に関する研究に取り組んでいる。大規模プロジェクトには必要不可欠となっている情報公開や説明責任などの際、意見集約・意見分類は重要であり、これを客観的基準をもとに自動的に行える技術は今後ますます重要になると考える。

3. 技術紹介：社内文書からのキーワード抽出

IBS 言語研では、文書情報活用支援への言語処理技術の適用を試みている。文書の有用性判定のキーとなる情報を提示するために、社内文書からキーワードを抽出した例を紹介する。

3.1 目　的

電子化された文書は、ますます増加の一途を辿っており、多量の文献やイントラネットの文書のうち、どれを読むべきかを判断するのは、困難になりつつある。

全文検索する方法もあるが、検索の場合、自分でキーワードを入力する必要があり、適切なキーワードを思いつかなかった場合や、ちょっとした表記の異なりがあった場合には、必要な文書を検索することが出来ない。

しかし、文書に既にキーワードが付与されていれば、読むべき文書の発見は容易であるし、検索者が自分では思いつかなかったキーワードに気づくこともある。従来、文書群にキーワードを付与する作業は人手で行われていたが、人手によるキーワード付与には、時間とコストがかかる、キーワード選択の基準が作業者ごとに異なるため基準が一定しない、専門分野ごとに作業者を教育する必要がある、などの問題がある。そこで、IBS 所内文書を対象に、文書からキーワードを自動的に抽出するシステムを構築する。

3.2 方　法

日本語の文章は分かち書きがなされていないため、通常、単語の切り出しには形態素解析を用いる。だが、報告書やテクニカルレポートなどの社内文書には一般に専門用語が多い。そのため、形態素解析を用いると、１）辞書登録がないと未定義語になってしまう、２）分割されて一般用語になってしまう、など処理がうまく行かない場合が多い。よって、文字種のマッチングによってキーワードの抽出を行う。

キーワードの抽出においては、助詞や代名詞、接続詞など、どのような文書においても全般的によく出現する高頻度語（不要語）を候補から除外する。これらは、平仮名・記号などが多い。

キーワード候補文字列の抽出

・漢字、カタカナ、英字、またはそれらの組み合わせが連続した文字列を抽出する。
例：都市計画･バリアフリー･情報バンク
・漢字が一文字の場合、動詞や形容詞の語幹であることが多いため、基本的に除外する。但し、後に「が」「を」など手がかりになる文字が付随していれば採用する。

候補として抽出された文字列を全てキーワードとする訳ではない。当該文字列が文書の内容を適切に表現するかどうかの度合いを測定し、一定の基準を満たす文字列だけをキーワードとする重み付けを行う。どの文書にもよく出現するような語は、異なる文書を識別する手がかりにならないため、キーワードとしない。この重み付けには、tf･idf 法を用いる。

目次へ

＊＊　tf･idf 法　＊＊

文書集合中で少数の文書に偏って高頻度で出現する語をキーワードとして抽出する。

スコア＝tf×idf

tf（term frequency）：
あるキーワードの対象文書中での出現回数

idf（inverse document frequency）：

$$\log\left(\frac{N}{n}\right)$$

N；全文書数

n；そのキーワードを含む文書数

3.3 実　験

(1) 実験対象

1995 年度版～1998 年度版までの４ヶ年分の IBS 研究調査実績一覧[7)8)9)10)]に掲載されているプロジェクト概要のうち、都市・地域研究室および旧都市計画・地域開発研究室のもの、46 文書を対象とする。

(2) 実験結果

下記ａ）の文書から、自動的にｂ）のキーワードが得られた。

ａ）抽出元文書（全文）

『長岡市道路交通円滑化方策策定調査』(平成９年)
過年度調査につづき、長岡市中心部周辺地域での TDM 施策として、①時差出勤、②相乗り、③新規バス路線（循環バス）の利用促進の計画・推進方策を検討した。時差出勤では、市役所での試行を念頭におきつつ、様々な主体から見た時差出勤の実施意向を受け、地域内調整の枠組みを取り込んだ実施方法を提案し、相乗りでは自動車通勤者の意向から相乗りを誘導するための制度や既存の HOV レーンの改善を提案、また循環バスについては利用者意向から捉まえた利用促進策を提案した。さらに、検討結果については、市の管轄する範囲以外を含む総合策としての提案となるため、委員会を通じて関連機関や民間主体との調整を図った。

⬇

ｂ）キーワード（上位３語）

時差出勤、相乗、循環バス

(3) 考　察

「長岡市」が取られていない、「相乗り」の「り」が脱落するなどの問題があるが、概ねキーワードと言える語が取れている。また、「意向」や「検討」は複数回、出現しているが、今回の方法ではキーワードとはならない。

今回の実験では、各文書で抽出されたキーワードが少なかった。これは、文書自体が短いことと共に、そのことにより、１）同じ語の繰り返しを避ける、２）文章を簡潔にするために臨時に一語になった複合語（例：環境負荷抑制型土地利用誘導）が多い、というアブストラクトの特性に起因していると考える。

4. 今後の応用例

このようなキーワード抽出結果を利用して、下記のように応用することも可能である。

◆専門用語辞書作成

取り出した単語のうち、形態素解析して未定義語になるものを集めて専門用語辞書を作る。新語にも対応が可能である。

また、こうして作成した専門用語辞書を用いて形態素解析を行うことで、より高精度の情報抽出・検索が可能になる。

◆概念ネットワークの自動生成

抽出したキーワードを用いて概念同士のネットワークを動的に自動生成し、関連のあるものを構造化して示す。これにより、問題解決の発想を支援する。

◆文書の自動分類

各文書のキーワードの類似度を計算することで、文書をクラスタリングすることが出来る。この方法は、自由回答アンケートの自動分類や意見集約にも利用できる。

5. おわりに

情報公開、合意形成の潮流から、自由回答の意見集約・分類は今後、需要を増すと考えられるし、ナレッジ・マネージメントの観点からも、職場内に蓄積された知識情報、社会

目次へ

的に蓄積された知識からの情報抽出・活用は要請が高まると考える。

情報が溢れている今日、さらなる発展のためには、情報の活用が不可欠である。付加価値のある情報を取り出す、または、大量データの処理をする場合には、言語処理技術を利用できると考える。

参考文献

1）情報処理学会編『コンパクト版情報処理ハンドブック』(1997)
2）長尾真・宇津呂武仁他『岩波講座マルチメディア情報学４　文字と音の情報処理』岩波書店（2000）
3）馬場肇『日本語全文検索システムの構築と活用』ソフトバンク（1998）
4）徳永健伸『言語と計算５　情報検索と言語処理』東京大学出版会（1999）
5）乾裕子・村田真樹・内元清貴・井佐原均「文末表現に着目した自由回答アンケートの自動分類」情報処理学会　自然言語処理研究報告　No.128（1998）
6）内元清貴・小作浩美・井佐原均「キーワードによるネットワークニュース記事群の構造化」言語処理学会第４回年次大会発表論文集（1998）
7）計量計画研究所『IBS 研究調査実績一覧・1995 年度版』(1996)
8）計量計画研究所『IBS 研究調査実績一覧・1996 年度版』(1997)
9）計量計画研究所『IBS 研究調査実績一覧・1997 年度版』(1998)
10）計量計画研究所『IBS 研究調査実績一覧・1998 年度版』(2000)