知の共創のための自然言語処理技術
－情報マネジメント技術を俯瞰する－

今日の高度情報化社会において，知的生産性を高めて企業競争力を獲得するためには，「情報の氾濫」をいかに克服するかが課題となっている．しかし，情報の量的増加は質的欠乏を引き起こすため，人間の創造的知性に働きかけることにより，戦略的に情報を利活用する必要がある．本稿では，自然言語コンテンツとして流通する大量の情報から，「意味」や「知識」だけでなく属人的情報をも抽出し，情報を質的に体系化するための仕組みについて，技術的観点とマネジメント的側面から概観する．そして，「情報マネジメント」技術として実用レベルにある自然言語処理技術に着目し，翻訳，検索，分類，フィルタリング，マイニングといった技術の関連と役割について俯瞰する．

日本電気（株）
亀井 真一郎
s-kamei@cw.jp.nec.com

富士ゼロックス（株）
田邉 栄一
eiichi.tanabe@fujixerox.co.jp

産業技術総合研究所
和泉 憲明
niz@ni.aist.go.jp

知の共創のための言語情報とそのマネジメント

従来からの生産活動資源である「人･もの･金」に「知識」を加えて，いかに新しい価値を創造できるかが今日のきわめて重要な課題となっている．この課題に対し，インターネットとブラウザの普及は情報の量におけるボトルネックを解決しつつある．しかし一方で我々は，「情報が氾濫している」という問題に直面している．今日の社会において知的生産性を高め新たな競争力を獲得するためには，日々溢れくる情報を効果的に活用できる技術，すなわち「情報をマネジメントする」技術が必須となっているのである．

情報マネジメントに関しては，情報処理の基礎研究レベルからビジネス分野でのアプリケーション開発に至るまで，さまざまなアプローチが採られている．これらを改めて俯瞰すると，単なる対応事例の集積としてではなく，自然言語処理に裏打ちされた，人の創造性の支援と情報処理とを融合させる，新しい技術領域を予見することができる．

本稿では，まず溢れる「情報」を有効活用するパラダイムが情報マネジメントであると位置づけ，自然言語処理技術による情報氾濫への本質的な挑戦こそが，知的生産性を向上させる第一の手段であるとして，「知の共創」への到達手段を考察したい．

「情報の氾濫」が象徴する課題の本質

ここでは，まず「情報」の持つ基本的な性質を考察し，その観点から自然言語の役割を位置づけてみたい．Claude E. Shannon によれば，「情報」とは「体系が一定の不確実性を持っているとき，この不確実性を減らす働きをするもの」である[1)]．一方，現代社会には「情報が氾濫している」という実感があるが，不確実性を減らして体系を秩序立てるはずの「情報」が氾濫するとは不可解である．

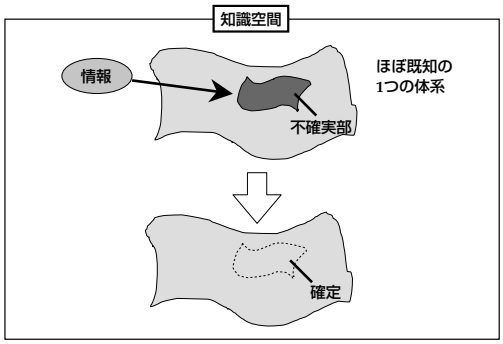

図 -1 「情報」の基本定義

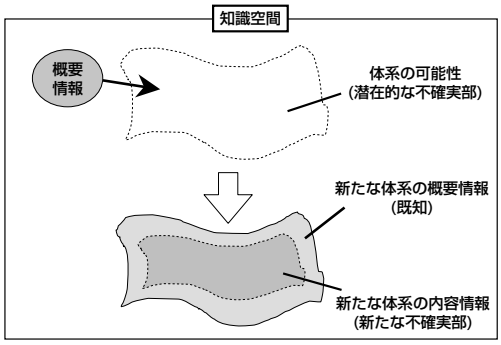

図 -2　概要情報が内容情報の欠乏を生じる

この不可解な現象を考察するために，まず，図 -1 に「情報」の基本的定義を表す模式図を示す．この図は，全体の様子がほぼ分かっている既知の体系の一部に不確実性が残っている場合を表している．ここで「体系」としては「知識空間」における「知識体系」を想定する．この場合，「情報」を受け取り，不確実部へ適合するように質的に変形させることにより，体系全体の持つ不確実性が減少する．たとえばある事柄について大体分かっている状況で，「分からない部分を他人に聞いたり書籍で調べたりして」「理解する」という行為が，この図 -1 で表される状況である．

これに対し図 -2 と図 -3 は，情報が与えられることで体系の不確実性が増大するように見える場合を表した模式図である．これらも「知識空間」における図である．図 -2 は，新規の体系の枠組みだけを与える情報が入ってきた場合を表している．たとえば「松本市内で毒ガスによるとみられる多数の死亡者がでた」という情報が与えられたとする．このような事件は可能性としてはあり得るが，実際の出現頻度はきわめて稀であるからこの情報（ニュース）の「情報量」は大きい．しかしながらこの情報はまだ 1 つの知識体系をなしていない．これは事

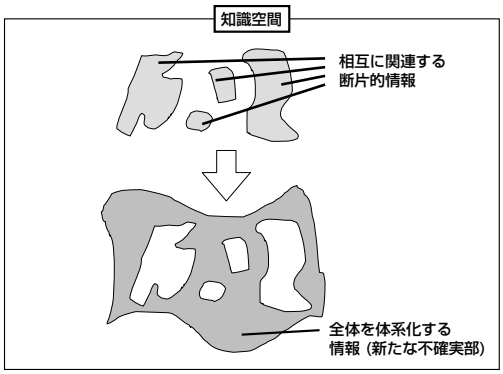

図 -3　断片データが全体体系化情報の欠乏を生じる

故なのか故意なのか．事故ならば，なぜ，どのように起こったのか．故意ならば，誰が，なぜ，どのように行ったのか．このような「内容」に関する情報が欠如している．つまり「ある事柄が生じた」という情報によって，同時に新たに「その事件に関する知識体系」が生じ，その部分の不確実さが新たに生じたわけである．

図 -3 はやはり知識空間において，断片的な多数の事実がまだ知識体系をなしていない状況を表している．全体を 1 つの知識体系に整理するためには，新たな種類の情報が必要となる．今日では，Web ページや電子メールといった情報メディアが，図 -3 に示すような知識空間を構成している．このような情報メディアでは，主に自然言語テキストで情報が記述されている．このことは，情報の不確実部を確定させるためには，テキスト情報から適切な内容を抽出し，何らかの質的な変換を行う必要があることを意味する．

このように「情報」は「不確実性を減少させる機能」を持っていると同時に，別種の情報の欠乏を露見させるというメカニズムを備えている．情報が供給されればされるほど新たな情報が供給されねばならない必要が生じ，整理がつかない状況を呈する．この状態が「情報の氾濫」と呼ばれる現象である．これは情報の質に起因するものであるため，量的補強では解決できない．言い換えると，「情報の氾濫」の解決には情報の質（内容）に踏み込む必要がある．

ここで，情報の内容を記述する情報メディアの性質を考察する．画像情報の電子化には大量のビット数を要するため「画像は情報量が多い」といわれる．実際，言語では書きつくせないような秋の紅葉の美しさでさえ，画像を用いると，具体的な形や，色，動きが表現できる．しかし，たとえば「風情」「季節感」「日本人」といった概念は，自然言語でしか表現できず，画像は補助情報で

しかない．このことからも分かるように，「意味」あるいは「知識」と呼ばれる情報内容の大部分は，自然言語で表現される．このことは「情報の質的な変換に関する課題」を解決する情報メディアとして自然言語が有望であることを示唆している．しかし，たとえある情報が言語メディアで与えられていたとしても，質的内容変換により知識として抽出することは，容易な技術ではない．

ビジネス応用としての情報マネジメント技術の登場

「情報の氾濫」が「情報の質的変換の課題」に起因することを確認したが，次に，この課題に対する企業システムにおける取り組みを概観する．

1990年代，インターネットとシームレスにつながる情報ネットワークが企業の隅々にまで行き渡るようになり，CRM（Customer Relationship Management），SCM（Supply Chain Management），ERP（Enterprise Resource Management）といった新しい概念を取り入れた情報システムが登場した．ここでの情報システムは，あらかじめ決められたフォルダの分類構造に従って，情報資源である文書を格納している．作成者やファイル名など，既定のファイル属性を識別することによってバージョン管理やアクセス管理を支援するものであり，ワークフロー管理や文書検索機能を実現していた．

作業工程や管理プロセスに従って独立した部門ごとに管理される情報システムでは，情報の更新が分散的に行われることとなり，これに起因する課題として，分類観点や用語，フォーマットなど，企業単位での管理ポリシーの統一がきわめて困難である，ということが認識されるようになった．同時に，冒頭に挙げた情報システムは，企業内の各部門に散在していたデータを関連付けて活用した一方で，目的に応じたシステムのカスタマイズを加速し，システム間での情報の重複，あるいは売上・利益管理，人事管理といった基幹系システムとの相互運用を困難なものとしてしまった．

そこで生まれてきたのがEAI（Enterprise Application Integration）やEDI（Electronic Data Exchange）という考え方である．これらは統合された通信プロトコルのもとに，XMLなどの標準フォーマットで文書情報のオブジェクト化を図り，独立したアプリケーション間の連携，あるいは離散した情報の相互利用を可能なものとしている．さらに，最先端のマーケティング部門などでは情報のオブジェクト化のニーズはさらに深化し，顧客ニーズに即応するかたちで，あるいは新しい観点に基づくカテゴリで，顧客や製品などの情報を再構成することが求められている．たとえば，最近流行の地域限定お菓子であれば，従来，子供向け・大人向け・女性向けといった客層カテゴリや，甘味・苦味・辛味といった味カテゴリでグルーピング可能としていた．これらのものを，地域名産物との相似性や大型化のしやすさ，あるいは販売店と顧客動線との適合性といった，新しい観点でグルーピングしなおしてみる機能である．この機能を備えた情報マネジメントシステムの実現は，情報にアクセスする人に新しい“気づき”の創造を提供するものと位置づけられる．

以上をまとめると，ビジネス界におけるこれらめまぐるしい情報システムの変遷は，当初は一貫して技術環境の変化への対応としての業務プロセスの改革や業務の効率化にあった．しかし1990年代後半から2000年代に入ると，属人的な知識を含めた情報自体の活性化へと移行してきた．背景には，手元に集められた情報をコンピュータが扱えるようにコード化し，体系的に処理をするだけでは十分な顧客価値を生み出すことができなくなってきたのである．言い換えると，人の創造性をサポートするという属人的な役割が求められるようになってきたのである．これはシステムを中心とした情報処理を，人を中心とした情報マネジメントに変換するというコペルニクス的発想の転換を促すこととなった．

情報マネジメントの企業システムとしての展開

では，人中心の情報マネジメントという目標は，実際，企業システムでは具体的にどうやって実現されてきたかを整理する．まず，個人を中心とした作業形態として，Etienne Wengerらが唱える「実践の場（Community of Practice）」を重視するという経営科学的分析がある[2]．さらに，John Seely Brownはプラクティス（Practice）をプロセス（Process）の補完的な対語として捉え，プロセスがあらかじめ決められた手順に従う固定的な業務であるのに対して，プラクティスとはこの決められた手順間で状況依存に実現される知的な生産活動であると位置づけた[3]．このような「実践の場」で必要となる情報マネジメントには，たとえば，ある職位に従って意思決定を行うというよりは，状況に応じてアドホックな意思決定フローを作成し，衆知によってコンセンサスを構築するようなシステムである．

これに対して，今アメリカでは，サービス業を中心に積極的に取り組まれているEKP（Enterprise Knowledge

Portal）というものがある．これは先に述べたようなプロセスやプラクティスの知識を領域ごとに俯瞰できるポータル（情報の HUB）を作成し，
・企業内にどんな情報があるのかを閲覧したり（Know What, Know How），
・非定型な情報や定性的な情報を誰が知っているのかを案内したりして（Know Who），
必要な知識に人がたどり着くための機能を提供している．ここでのポイントは，人が必要とする知識そのものを扱う代わりに，人が知識を得られるように支援していることである．これにより企業は，コード化された情報だけでなく，属人的な知識までを有効に活用可能としたのである．

このような EKP の普及は，新しい情報マネジメントの方向性として，個人を中心に据えた実践のコミュニティ活性化の支援という方向性を示した．同時に，企業における情報マネジメントシステムの展開は，言語情報の内容に踏み込むことなくシステムの有用性を高めることができる，という成功事例を与えたのである．

知にアプローチする情報マネジメント技術の展開

企業システムにおける人中心の情報マネジメントシステムの成功事例は，知識として文書情報だけでなく，属人的な情報への取り組みであり，「情報の氾濫」にみたように，情報の量的な充実は，質的内容の変換プロセスを支援することで達成される，ということでもある．すなわち，情報マネジメントの成功は，属人的なものも含めた知識への取り組みの結果である．このような知識の内容を追求する情報マネジメント技術は，主に，知識工学の分野で先導的に研究されてきた．ここでは，その技術の具体的内容を俯瞰する．

言語メディアとして電子化された情報は，人間には容易に質的内容変換することができるが，機械的な処理にてその変換を達成することは困難である．この課題の解決を目指す手法は，
a）量を質の向上へ転化させるアプローチ
b）質をなんらかの代替で補完するアプローチ
に大別することができる．以下では，それぞれでいかに属人的な知識を取り扱ってきたかを述べる．

a）は，情報検索とデータマイニングに代表される知識の発見と精錬の技術である．ここでのポイントは，「知識や情報をシステムにより操作可能なものへと機械的に変換するのではなく，データとして電子化された大量の情報の変換は人間が行うが，その人間の作業をサポートすることである．

b）は，Web コンテンツにメタデータを付加したり，リンクの参照構造に着目したりすることにより，運用にて知識の質を高める技術である．ここでのポイントは，「情報の質的内容変換を機械的に達成可能となるように，補助情報を付加して情報の運用をサポートするということである．

具体的には，a）は，機械学習技術が Web により展開した技術であると捉えることができる．この代表としては，データマイニングにおける可視化手法や領域知識の積極導入，前処理と後処理の強化，メタ学習技術，学習方法論のマネジメントなどの技術が挙げられる．

また，データマイニングにおける新展開としては，専門知識や専門家の積極介在が提唱されている．領域知識を積極利用するというパラダイムとしては，Ryszard S. Michalski の知識マイニング（Knowledge Mining）が代表的で，領域知識やマイニング結果を新たなマイニングの参照知識としてスパイラル型でマイニングプロセスを継続させる，というビジョンである．専門家の積極介在としては，Pat Langley の会話型知識（Communicable Knowledge）が代表的で，機械的には困難でも，専門家に気づきを与えるようなインタラクションが知識獲得システムの重要な機能である，という考えである．

b）は，ドキュメント処理とオントロジー技術がメタデータ技術として展開したものであり，Web における知識工学の展開であると考えられる．これらの代表は，Google の検索エンジンやセマンティック Web に代表される．ここでは，Web コンテンツそのものの内容よりも，人向けや機械向けの意味標準として，リンク構造やメタデータの補助情報を用いるものである．

機械推論（Web エージェント）のためのオントロジーは，セマンティック Web 分野にて先導的に検討されたが，機械だけでなく人とも共有可能な情報基盤として浸透しつつある．このような情報基盤としてのメタデータは，近年充実しつつあり，その代表は，Dublin Core（書誌情報）や MeSH（医療分野），MPEG-7（マルチメディアコンテンツ）などである．また，これらのメタデータ記述法として RDF が普及しつつあり，RSS（RDF Site Summery）などは，新しい Web の潮流でもある．

情報の内容分析技術としての自然言語処理

企業システムや知識工学からの情報マネジメント技術

は属人的な情報をいかに取り扱うかにフォーカスしているが，情報の質的内容を直接的に記述する唯一の手段は自然言語である．したがって情報の内容を分析し断片情報を再体系化する技術は，自然言語処理技術の発展に支えられていることになる．図 -4 は自然言語処理に関連する技術をマップ状に配置した俯瞰図である．この図では自然言語テキストが中央に配置されており，4 つの線でエリアが分割されている．左の縦線は「単言語情報か異言語情報か」を分ける線である．この異言語の壁を乗り越える技術として機械翻訳が位置づけられる．下の横線は「単一文書か大量ドキュメントか」を分ける線である．この大量情報の壁を乗り越える技術として情報検索技術が位置づけられる．右の縦線は「情報をそのまま扱うか密度を高めて扱うか」を分ける線である．テキストの中から有益な情報を取り出す「要約」や「抽出」の技術がこの情報密度圧縮の線を越える技術として位置づけられる．上側に位置する横線は「テキスト情報か異なるメディア情報か」を分ける線である．音声認識・合成技術にはその第 1 フェーズとして言語処理（辞書・形態素解析）技術が組み込まれている．したがってこれらは異メディア統合技術として位置づけられる．

1990 年代には，異言語，大量，密度圧縮，異メディア統合といった各領域の技術が出揃い，次段階の技術融合に進展する準備が整った．1990 年代半ばになると，これらの軸の重なる領域に新しい技術が発達してきた．1 つは，分類，フィルタリング，パーソナライゼーションといった技術である．単なる検索ではなく，より細かく整理する技術である．さらにテキストマイニングの技術が発達してきた．さらに，外国語情報にもシームレスにアクセスするための多言語情報検索技術が 1990 年代の終わりから研究され始めた．

大量のデータを検索・分類・フィルタリングし，新たな情報をマイニングし，抽出して，それを目に見える体系として可視化する．これら自然言語処理の諸技術は情報マネジメント技術の主要要素である．さらに機械翻訳あるいは多言語情報検索技術をトータルに含めれば，多言語情報マネジメントの構成技術となる．

情報マネジメントの基盤としての自然言語処理

本稿では，まず「情報の氾濫」という現象を手がかりに，情報の基本性質を考察した．情報の量的な増加は新たな情報の欠乏を生じるために，その解消は単なる情報の量的増加では達成されない．すなわち，情報を 1 つの

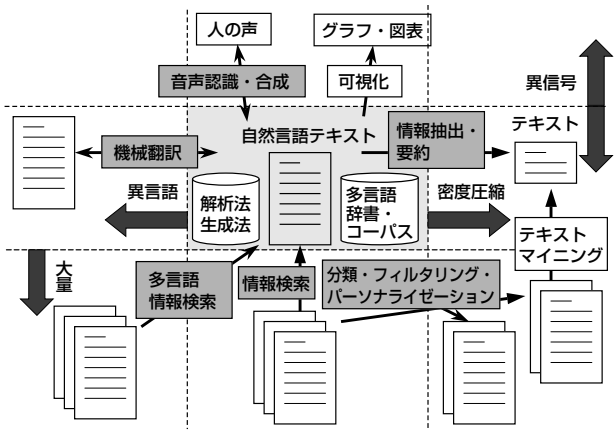

図 -4　自然言語関連技術マップ

体系にまとめあげるための情報の質的な変換が必要である．その鍵となるのが，属人的な情報をも取り込む環境とそれを支える自然言語処理技術である．ここで，情報の質的変換は容易ではないため，情報技術一辺倒では知的生産性は向上しないという指摘は的を射ている．これに対して，近年の情報マネジメントのシステムは，利用者が組織内の散在情報を再体系化することを支援する方向に展開している．同時に，知識工学の新たな展開も，専門家に気づきを与える方向に向かっている．このことから，情報の質的変換の困難を克服しようとする技術発展が求められ，結果として，情報マネジメントという新しい IT パラダイムの中核に自然言語処理技術が位置づけられる．つまり，情報の「意味」や「知識」の重要な部分は自然言語で記述されているから，利用者に情報の質的変換をうながすには，自然言語の処理技術が中心的役割を担うのである．「情報の氾濫」の克服は，自然言語処理の諸技術を活用し，その利用者の持つ創造的知性と融合させることによってのみ可能性を持つのである．

以上のような背景のもと，翻訳，検索，分類，フィルタリング，マイニングといった自然言語処理の諸技術は，いまや現代の情報社会に不可欠な基盤技術として認知され，広い分野で実活用されている．これらの実例は本特集の各記事に述べられている．

参考文献

1）Shannon, C. E.: The Mathematical Theory of Communication, The University of Illinois Press（1949），邦訳：長谷川，井上：コミュニケーションの数学的理論，明治図書（1969）．

2）Lave, J. and Wenger, E.: Situated Learning: Legitimate Peripheral Participation, New York, Cambridge University Press（1991），邦題：状況に埋め込まれた学習．

3）Brown, J. S. and Duguid, P.: Balancing Act: How to Capture Knowledge without Killing It, Harvard Business Review（May-June 2000），邦題：知のダイナミズムと BPR の融合．

（平成 15 年 9 月 11 日受付）