# 合金研究論文テキストからの知識抽出

星本　健一　科学技術庁金属材料技術研究所・計算材料研究部
松尾　利行　奈良先端科学技術大学院大学(現在(株) 日本総研)
康村　昌司　奈良先端科学技術大学院大学(現在(株) 日本総研)

# Exploration of Knowledge on Metallurgy from Research Articles

Ken'ichi Hoshimoto[†], Toshiyuki Matsuo[¶] and Shohji Yasumura[¶]
[†]National Research Institute for Metals
[¶]Nara Institute of Science and Technology(Present adress: Japan Research Institute)

An information base system to support construction and exploration of an information space in metallurgy, especially in the domain of superalloys, was implemented. Information about materials development is, in most cases, obtained empirically. Use of natural language is, therefore, important to transmit such information. The developed system interprets technical papers on metallurgy and constructs information space semi-automatically. It consists of three sub-systems. The first retrieves information from incomplete keys by using thesaurus. Successful results were obtained by picking up all the idiomatic expressions appeared in the articles on the subject. The second sub-system, called METIS(METallugy papers Intelligent Surveyors) adopts a naive natural language process. The heart of METIS is a packet of domain specific knowledge called KP(Knowledge Pieces) in which procedures for extracting and structuring technological information are embedded. METIS extracts technological information from technical papers on metallurgy written in a mark-up language. Products of METIS is a variety of summaries and surveys such as structured technical summary, as well as visualization of similarities, differences of relevant papers, and cause-effect relations. The third sub-system is called KE(Knowledge Editor) adopts both syntactic and semantic analysis technologies. In the system, domain knowledge is represented using an instance-oriented and script-based method. KE represents knowledge on metallurgy by a set of objects together with a set of relations between them. KE provides two interactive facilities for exploring into information space; one is the natural language query based on semantic understanding and the other is the navigation of the relations between objects.

## 1 はじめに

近年、知識処理へのコンピュータの応用が盛んになり、知識処理の分野ではエキスパートシステム等が実用の域に達しているが、材料開発部門にこれを応用することは難しい。その理由は、知識が多くのあいまいさを含み、コンピュータ処理の技術が未だ研究段階にあること、及び知識が断片的で、体系化の作業自体が研究であることによる。最も困難な点は、現実の材料には組織、構造等を考えると同じものは決して存在せず、従って、データ、知識等の記述において対象を限定することができないことである。通常のエキスパートシステムは閉じた世界の中で有限の回答の中から最適のものを選び出すという形をとっている。しかし、新材料探索は閉じた世界で行われるものではなく、また対象を記述するモデルも確定していない。むしろ、各種のデータに適合するモデルの探索そのものが材料の研究である。そのような意味で，現段階の技術でコンピュータに創造的材料開発を行わせることは不可能である．したがって，本研究では研究者の発想を支援する目的で情報

を提供するシステムを開発することを目的としている.

本研究では合金設計の研究者支援することを目的にバーチャルコンサルタントシステムを開発した。同システムは、金属材料論文テキストを入力とし（半）自動的に情報空間を構築し、情報空間の探訪を支援する情報ベースシステムであり、以下に述べる曖昧検索, METIS, KE の三つのサブシステムから構成される.

これらのサブシステムを統合的に働かせることにより,研究者の思考の中断を招くことなく，求める情報を的確に提供するシステムの実現を目指すものである.

## 2　曖昧検索システム

第１のものはシソーラスをもとにキーワードを展開することにより、あいまいな情報から的確に文献を検索するシステムである.このシステムでは用語をいかに的確に抽出するかが鍵である. 現実に用いられる用語・表現形式は教科書に記述されるような模範的なものばかりではなく，同じことを言い表すのに人によってさまざまな語・書式を用いる. たとえば "nickel base superalloy" に対して, nickel を Ni としたもの, nickel と base の間にハイフンを入れたもの，base を based としたもの，さらにこれらの組合せと，じつに８通りの記述があり，金属関係でよく知られた文献データベース METADEX にもその全てが現れる. ニッケル基超合金に関する約 300 編の論文のフルテキストから，用いられている用語・表現をすべて拾い出し，シソーラスの整備を行った。検索システムとしては，入力されたキーワードに対して，シソーラスから同義語及び上位／下位語を選び出し，これらの出現頻度から該当論文を見いだすものである.

## 3　METIS システム

METIS システム(METallugy papres Intelligent Surveyors)は較的浅い自然言語処理技術を用いて論文の要約，類似論文の検索等を行うものである. METIS の中心的な役割を果たすのは、KP(Knowledge Pieces)と呼ぶ技術情報の抽出法と構造化法を一体化したドメイン知識のパッケージであり、文の選択・特徴の抽出・マージによる構造化・交差による構造化の４つの機能を提供する。METIS は、自然言語（英語）で記述された技術論文を入力とし、KP に記述されたキーワードやキーフレーズに着目した比較的浅い自然言語処理により、論文内容の項目別要約，因果関係の抽出等をおこなう. また各論文の解析結果を比較し,論文の類似度を視覚的に表示と比較を行うなど豊富なサーベイ情報を提供するものである。このシステムを用いて，超耐熱合金に関する国際シンポジウムのプロシーディングス(Superalloy '92: Proceedings of the 7th International Symposium on Superalloys, September 20-24, 1992, Sevens springs,)に掲載された８７編の論文について類似度を計算し，３次元座標にプロットしたところ，各セッション別の論文グループ毎に特有の配置を示し，解析の有効性が裏付けられた.

## 4　ＫＥシステム

KE(Knowledge Editor)システムは構文解析／意味解析の技術を用いて論文内容を理解，構造化し意味的な内容を含む問い合わせに対応するものである. 従来のデータベースシステムでは、自然言語で記述されたテキストから自動的に情報空間を構築することは困難であり、また単純なキーワードによる全文検索であるため検索内容が膨大になることが多

くその内容を理解するのに多大な労力を必要としていた。KEシステムでは論文内容を理解した情報空間を半自動的に構築できるだけでなく、問い合わせの内容を理解した検索結果を提示することを目指している．KEではオブジェクトとその関連という形式で合金設計の知識を表現し、システムにとって未知のオブジェクトが登場した時はそのオブジェクトの入力を支援する機能を提供する。材料設計の知識形式として提案した、インスタンス指向による知識表現とスクリプトをベースにした知識表現方法は、合金設計だけでなく他のドメインへの適用できる見込みが得られた。このシステムにおける言語処理について以下にやや詳しく述べる．

1）合金研究における情報

・研究者の行為に関する情報

合金研究者が行う行為には、試料製造、熱処理、実験があり、これらについての情報（条件温度、時間、装置）を抽出し構造化する必要がある。

＊ 熱処理には視点として温度、時間が存在し、それぞれ℃、hour や増加、減少などの語彙で状態が表現される。

＊ クリープ試験には視点として温度、時間が、結果として引っ張り強度などが存在し、それぞれ℃、Pa や増加、減少などの語彙で状態が表現される。

・合金特性と微細構造に関する情報

合金研究者は合金を熱処理や実験を行って、合金の特性（力学的特性、化学的特性）や微細組織を観察しており、これらの特性や組織の状態を表している情報を抽出し構造化する必要がある。

＊クリープ特性にはクリープ強度や破壊時間などの視点が存在し、それぞれ Pa,hour や増加、減少などの語彙で状態が表現される。

＊ 相には量などの視点が存在し、％や増加、減少などの語彙で状態が表現される。

・研究の流れを用いた情報の抽出と構造化

合金研究には研究対象の合金が存在し、さらに試料製造→熱処理→実験という一連の研究の流れ（研究フレーム）が存在する。研究内容を情報ベース化する際にはこのような研究の流れに沿って個々の行為を時系列に抽出し、更に個々の行為において観察された合金の特性や微細構造に関する情報を抽出し、構造化する必要があると考えられる。逆に言えば、合金研究者は以上のような情報、体系化された知識をもっており、それらを利用して個々の論文から情報を抽出、構造化することで合金研究内容を理解していると考えられる。

2）合金研究における概念構造情報ベース

合金研究論文の文章から上記のような情報を抽出、構造化することを考える。合金研究で注目される概念を示す語彙、更にそれらの状態や事象を表わす語彙を用いて文を以下のようなフレームで表現することを考える。

・注目される概念を示す語彙をオブジェクト

・その概念に対する視点を示す語彙を属性

・その視点に対する値を示す語彙を値

文→（（オブジェクト　（属性　値）（属性　値）（属性　値））

（オブジェクト　（属性　値）（属性　値）（属性　値））
（オブジェクト　（属性　値）（属性　値）（属性　値））
（例）"The alloy containg 7.5% cobalt shows an anomalously high primary creepdue to inhomogeneous {111}<222> slip." というセンテンスを以下のように構造化する.

・alloy
(cobalt (amount 7.5%))
(primarycreep　high)
(slip (face ({111})) (direction <222>) (homogeneity inhomogenious))

このような情報の抽出と構造化を行なうには、「コバルトには量という視点がある」,「数字%は量(amount)を表わす」,「すべりには面、方向という視点がある」,「{数字 数字 数字}は面を表わす」,「<数字 数字 数字>は方向を表わす」,「high は primary creep の状態（値）を表わす」,「inhomogenious は homogeneity の状態（値）を表わす」といったような概念構造が必要となる。

これらの概念構造は合金研究者が持っている概念構造であり、情報を構造化する際には、上記のようなの視点（例えばamount,face)を表わす用語が明記されてなくてもこれらの用語（視点）を抽出して構造化するべきである。また、定義することが難しい視点もある。例えば「creep strength」のような特性を表す用語（概念）と「high」などの状態を示す用語（値）を関連付ける視点を抽出し、構造化するのは困難である。よって概念と値だけの用語で構造化できる（必ずしも視点を必要としない）ような柔軟な概念構造をもつ情報ベースを構築する必要がある。

3）論文構造

知的情報ベース化するためには、合金研究論文を単なる文書情報としてではなく、論文構造を用いて情報ベース化する必要がある。即ち、論文を章の集合、章をパラグラフの集合、パラグラフを文章の集合と捉え、またヘッダ部には著者、所属機関（住所）等が含まれる. 合金研究論文を知的な情報ベースにするためには、このような論文構造に基づいて各文章を抽出して論文として構造化し、更に各文章から合金研究情報を抽出して概念情報ベース化する必要がある。特に，図表，数式等は論文中で重要な役割を担っているが，本文中にはそれらを指示，引用する文章が頻繁に現れる．たとえば，"Fig. 1 shows ...."というものである．したがって論文の構造も知識として持たせておく必要がある.

4）おわりに

以上のごとき検討を行いながら,超耐熱合金に関する知識検索システムのプロトタイプを試作した．現在は，質問としてオブジェクト，属性，条件の三つを個別に与えて知識検索を行っているが,将来は簡単な自然言語対話によりコンサルテーションを行うシステムの開発を目指している.