金属材料論文からの知識抽出

○中挟知延子†
星本健一‡

Knowledge Extraction from Technical Papers of Metallurgy

Chieko Nakabasami
Kenichi Hoshimoto

Abstract

We claim that it is necessary for semantic representations of technical papers to describe a word's meaning from various points of view and emphasize the semantic aspect of words.In this paper, we describe semantic representations of technical papers based on Pustejovsky's GENERATIVE LEXICON(GL). Our focus is on technical papers concerned with metallurgy because this research is carried out in cooperation with an expert in metal engineering. Our purpose for making semantic representations is twofold: (1) to detect differences and inconsistencies among papers by applying appropriate mechanisms; and (2) to integrate the content of a new paper into the content of an existing one. The semantic representations of technical papers based on the generative lexicon are modified so that they match the characteristics of the papers. In addition, we propose operations worked effectively on the semantic representation. Modifying some rules included the semantic representation for each paper makes it possible to analyze the semantic differences and similarities among the papers. The semantic representation is implemented using an object-oriented database management system on MATISSE.

1 はじめに

我々の研究の目的は，金属材料分野における研究論文の意味構造を記述するための枠組みを提案し，それを用いて論文間の意味内容の差異を抽出し，それに加えて，既に構造化した論文に新たな論文の意味構造を加えた場合に，意味内容の整合性をチェックし，矛盾を検出できるシステムを構築することである．

金属材料論文に焦点をおいた理由として，1) 他の文章に比べて専門用語が多く，語の曖昧性が少ない，2) 文章の章立てが明確になっており，文章全体の構造がパタン化しやすい，3) 著者の一人の専門分野である金属材料分野での研究支援システムを構築するため，4) 論文の記述内容についての詳しい検討ができる，ことがあげられる．そこで，我々

は金属材料論文の意味構造をドメインとして研究を進め[1][2][3]，そこで得られた知見を足がかりに，他の分野の文章に関しても研究を発展させていく予定である．

文章の意味表現については従来さまざまな方法が提案されている[4]．ただし，これらは各フレーズについての意味表現にとどまっており，複数の文にわたる意味の研究はほとんどなされていない．また，参照情報を利用して論文間の関係を抽出する研究もある[5]が，これは参照されている箇所についての局所的な情報であり，一つの論文の内容についての全体像をつかむことはできない．さらに文章の類似性の判断を，文章中に表れる語の出現頻度や共起情報などを統計的手法によって行う研究も多数見受けられる[6]が，これらはいずれも複数の文章をジャンル別に分類するものであり，論文文章についての詳細な内容記述を基に論文間の差異・矛盾などを抽出するものではない．

それに対して我々は，複数の文にまたがる意味構造を詳細に記述し，それらの記述の上に推論機構を構築することで，論文間の内容上の差異・矛盾を抽出することを試みている．意味構造の記述には，Pinker[7]の提唱する，「子供の言語獲得において語の意味からその語の統語的なふるまいが予測可能である」という仮説に基づいた「意味立ち上げ(Semantic Bootstrapping)」の流れから発展してきた Lexical Semantics の立場から提案されている Pustejovsky の Generative Lexicon（生成語彙）[8]を基に行っている．また，意味構造で表現された論文データはオブジェクト指向データベース Matisse[9]を用いて蓄積する．

また今回，対象とする金属材料分野などの実験プロセスとその結果の記述が主体である論文における内容上の差異は，「同じ実験を行って異なる結果が出る原因は何なのか」や「同じ材料（物質）について異なる振る舞いが得られる原因は何なのか」に焦点がおかれている．そのため，差異を抽出することは，それらを生じさせる原因を分析することになる．

以下，2 章では生成語彙を用いて表現された論文の意味構造を提案し，3 章では複数の論文間の差異を抽出するためのルールとそれらを操作するための機構について述べる．

2　論文の意味構造

2.1　生成語彙を用いた構造

Pustejovsky［8］の提案した生成語彙（Generative Lexicon）は，元々名詞や動詞の多義性を解消するために考えられた意味記述である．本論文で述べる金属材料研究論文の意味内容に沿うように，元々の生成語彙の記述内容を工夫している．

金属材料研究論文において，論文中で述べられている内容を表す語句として以下の 5 つがあげられる．

(1)　材料：material
(2)　装置：apparatus
(3)　実験条件：condition
(4)　前処理：pre-process
(5)　実験結果：experimental result

これら 5 つの概念は，論文における 5 つの異なる観点からの内容を表しているといえる．そこでこれら 5 つの意味構造を生成語彙の考えを取り入れて設定したものが以下の

記述である．記述中における ARGSTR 等の記述についての説明を余白に記す．

test

ARGSTR ARG1: x=material

ARG2: y=result //ARG: test を行うにあたっての必須の要素

D-ARG1: z1=condition (e.g., temperature, stress)

D-ARG2: z2=apparatus (e.g., extensometer, microscope)

//D-ARG: test を行うにあたり暗黙のうちに想定されている要素

EVENTSTR E1: e1=test_act (e.g., test, perform, conduct)

// test の事象を表す動詞

QUALIA AGENTIVE=test_act(e1, x, y)

// test を成立させるための事象と要素

LEXICAL-INHERITANCE (e.g., evaluation, measurement, observation)

// test の下位概念

material

ARGSTR ARG1: x=mat_name // 材料の種類，名前

EVENTSTR E1: e1=mat_act // 材料に施される事象

QUALIASTR CONSTITUTIVE=mat_structure // 材料の成分

FORMAL=mat_property // 材料の外見，性質

AGENTIVE=mat_process(x) // 材料に施される前処理

LEXICAL INHERITANCE (e.g., specimen, alloy, sample) // 材料の下位概念

mat_process

ARGSTR ARG1: x=material // 前処理に必要な要素

ARG2: y=condition (e.g., time of exposure)

EVENTSTR E1: e1=process (e.g., forge, expose, anneal) // 前処理を表す動詞

QUALIASTR AGENTIVE=process(e1, x) // 前処理を成立させるための事象と要素

result

ARGSTR ARG1: x=material // 結果を評価するために必要な要素

ARG2: y=factor (e.g., tensile strength, strain rate, elongation)

EVENTSTR E1: e1=res_act (e.g., explain, reflect, obtain, indicate, exhibit)

// 結果を表す動詞

QUALIASTR FORMAL=quality(x) $\vee$ quantity(x)

// 結果を生じる要因の質的要素と量的要素

AGENTIVE=res_act(e1, x, y) // 結果を生じるための事象と要素

2.2 Matisse におけるオブジェクトスキーマ

2.1 で提案した意味記述について，5 つの観点をそれぞれオブジェクトクラスとする．ARGSTR 等の表現内容をオブジェクトの属性あるいはオブジェクト間の関係で表し，図 1 のようなオブジェクトスキーマを構成した[7]．図中の→は継承を表し，●—●はクラス間の関係を表している．Matisse においては，クラス間の関係は常に双方向であり，例えば test クラスから condition クラスに cond 関係があれば，condition クラスから test クラスに

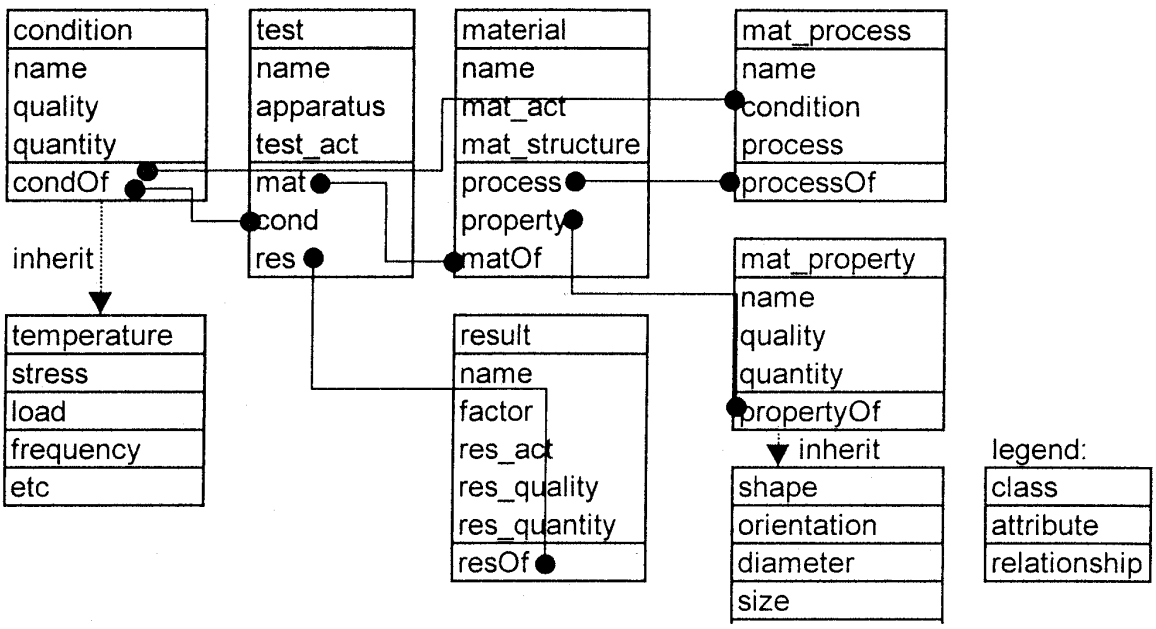

図 1 論文内容のオブジェクトスキーマ

condOf 関係がある.

図 1 にあるスキーマを用いて, 現在では約 200 文に対して, 人手で Matisse 上にオブジェクトを構築している. また, 論文文章から実験方法と実験結果についての記述を選び出して行っている.

3 Agentive role

構築したオブジェクトに対して, 2.1 節で述べた意味記述における Agentive 部分に注目して, 記述間の差異やそれを生じた原因を抽出する. 各クラスの Agetive に現れる引数の表すクラスオブジェクトをたどっていくことにより, 差異を生じた原因を推論する.

例として以下の二文をあげる. これらを入力文とする.

[1] The results of tensile tests showed that PTA welded specimens exhibited 96% nominal ultimate tensile strength of IN-738LC base materials.

[2] However, for tensile tests at room temperature, the joint PTA specimen exhibited 80% of the ultimate tensile strength of the base metal.

これら 2 つの文を図 1 に表したオブジェクト構造で表し, 共通なオブジェクトをまとめて図示すると図 2 のようになる. 図 2 で一つにまとめられず, オブジェクトの属性が異なるのは result と mat_process の箇所である. これらのクラスの Agentive role に着目して次に示す処理を行う.

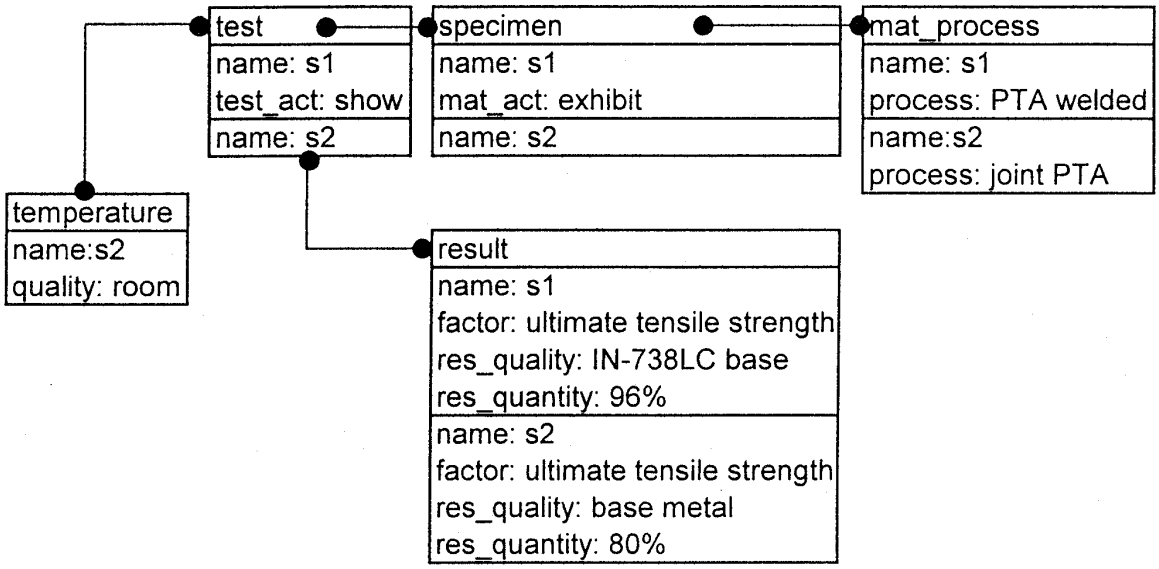

図 2 二つの文の内容を合成した後のオブジェクト構造

Step 1. result class, mat_process class について, Agentive role(AR)を抽出する.

result:AR → res_act(e1:res_act, x:material, y:factor)

mat_process:AR → process(e1:process, x:material)

Step 2. 抽出した複数の AR に出現する引数について, 対応するクラスがあれば, そのクラスの AR で置き換える.

material:AR → mat_process(x: material's name)

Step 3. 抽出した複数の AR に出現する述語について, 対応するクラスがあれば, そのクラスの AR で置き換える.

mat_process:AR → process(e1:process, x:material)

Step2, Step3 を置き換えられたクラスが一つになるか, 複数回行っても結果が同じである状況になるまで繰り返す.

Step1 から Step3 を行った結果, 得られた AR のクラスが, 入力文の間で内容の差異を生じさせる原因となっていることがらとなる.

4 今後の課題

(1) 論文間の差は何らかのテストについての条件や結果の違いとして現れ, 導入部分や結論部分に現れる著者の動機や考えなどについてはまったく考慮されていない. 論文の内容を表現するためには, これらの箇所についての構造の構築も必要である.

(2) 論文の内容を現在は人手で Matisse に入力しているが, 今後は自動化する必要がある. このためには, パーサなどのタグ付けを行うツールを用いて語句の切り出しをした後, 概念辞書や専門用語辞書を参照するシステムを構築する必要がある.

参考文献

[1] 中挾知延子，星本健一 (1999): 生成語彙による合金研究論文からの知識抽出．情処研報 Vol.99, No.2. pp.125-132.

[2] 松尾利行，西田豊明，星本健一 (1997): 金属材料論文からの技術情報空間の構築と探訪の知識支援．人工知能学会誌．Vol.12,No.1,pp.68-77.

[3] 星本健一，松尾利行，康村昌司 (1998): 合金研究論文テキストからの知識抽出．情報知識学会第6回研究報告会講演論文集．pp.47-50.

[4] 例えば，Jackendoff, R. (1990): Semantic Structures. MIT Press.

[5] 難波英嗣，奥村学 (1989): 論文間の参照情報を考慮した学術論文要約システムの開発．情処研報 Vol.98,No.82, pp.79-85.

[6] Lang, K. (1995): Newsweeder: Learning to filter netnews. In Prieditis and Russell (Eds.), Proceedings of the 12th International Conference on Machine Learning (pp. 331-339). San Francisco: Morgan Kaufmann Publishers.

[7] Pinker, S. (1989): Learnability and Cognition: The Acquisition of Argument Structure. MIT Press.

[8] Pustejovsky, J. (1995): The Generative Lexicon. MIT Press.

[9] Matisse OOS ODL Programming Guide (1996): A.D.B.S.A., Paris.

†東洋大学国際地域学部
Regional Development Studies, Toyo University
‡科学技術庁金属材料科学研究所
National Research Institute for Metals