美術に関する常識判断システムの構築

中本　一志　　渡部　広一　　河岡　司
同志社大学工学部知識工学科

1. はじめに

本研究では，人間と日常会話のできるロボットの実現を目指している．その為には，自然言語文章を常識的判断に基づき，理解できるシステムが必要となる．人間が持っている常識をコンピュータにも持たせることが必要となり，初期には人間が与え，その後は，その分野に関する知識を自動獲得し，新たな知識として追加する学習メカニズムが必要となる．

本稿では，美術分野に関する基本的な知識を与え，美術に関する自然言語文章の意味理解・判断・知識の学習を目指す方法を提案している．

2. 連想システム

連想システムとは，概念ベース[広瀬 02]［小島 04]やシソーラスを利用し，人間と同じような想起語処理や未知語処理を実現するメカニズムであり，コンピュータに常識的な判断をさせるために必要となる．想起語処理とは，例えば，「りんご」から「赤い」「甘い」などといった言葉を連想させる処理システムである．また，未知語処理とは，「唐辛子」などの知識ベースに定義されていない言葉を，知識ベースに定義されている「香辛料」などという言葉と置き換える処理であり，この処理によって，コンピュータに与える知識を最小限に留めることが可能となる．また，これらの処理を行うには，言葉と言葉の関連性を定量的に扱う必要がある．本研究では，概念同士の関連性を評価する方法としてシソーラスと，概念ベースを用いた関連度計算方式[渡部 06]を利用している．

2.1 概念ベース

概念ベースとは，語(概念）とその意味を表す属性の集合で定義された概念を格納した知識ベースのことである．概念A は，概念の意味を表す属性a_i と，属性の重要性を表す重みw_i の対で表される．概念A の属性数をN 個とすると，概念A は以下のように表せる．

$$A=(a1, w1), (a2, w2), \quad \cdots, \quad (aN, wN)$$

ここで，属性a_i を概念A の一次属性と呼ぶ．概念ベースには，このように定義された概念が約87000語収録されている．

2.2 関連度計算方式

(1) 一致度

概念A，B の１次属性を a_i，b_j，重みを u_i，v_jとし，属性がそれぞれ L 個，M個($L \leqq M$) とすると

$$A=\{(a_1, u_1), (a_2, u_2), \cdots, (a_L, u_L)\}$$
$$B=\{(b_1, v_1), (b_2, v_2), \cdots, (b_M, v_M)\}$$

と表現する．概念A，Bの一致度 $MatchWR(A, B)$は

$$MatchWR\ (A, B)=\sum_{a_i=b_j} \min\left(u_i, v_j\right)$$

（各概念の重みの総和は１に正規化する）

と定義する．このとき，一致度は一致する属性のうち小さい方の重みとなるが，これは両方の属性に共通して存在する重み分は有効だと考えるためである．

(2) 関連度

関連度とは，２つの概念間における関係の深さを定量化した０～1の実数値で，関係が深いほど大きな値となる．概念ベースの属性と重みを再帰的に参照することにより計算される． 関連度は，対象となる全ての一次属性の組み合わせについて一致度を計算し，１次属性同士の対応を決定することにより計算する．

対応の取れた属性の組み合わせが T個の場合，概念A，B の関連度 $Rel(A,B)$は次の式で示される．

$$\mathrm{Re}\,l(A, B)=\sum_{i=1}^{T} MatchWR(a_i, b_i)\times(u_i+v_i) \times(\min(u_i, v_i)/\max(u_i, v_i))/2$$

2.3 シソーラス

シソーラスとは一般名詞・固有名詞の意味的用法を表す 2710 個の意味属性（ノード）の上位－下位関係，全体－部分関係が木構造で示されたものである．ノードに属する名詞として約 13 万語（リーフ）が登録されている．図１に例を示す．

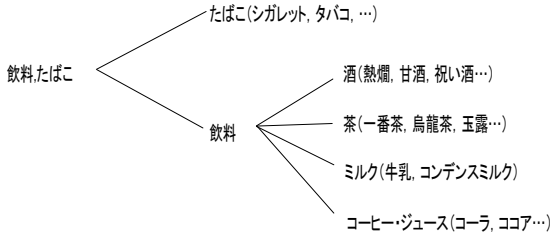

図１　シソーラス

3. 美術判断

美術に関して常識的な会話の行えるシステムを構築する．常識的な会話とは質問に対して，常識的な回答が返ってくるものともいえる．本研究では美術に関する質問に対して常識的な回答が得られるシステムを構築する．例えば「ゴッホの描いた絵はなんですか？」という質問に対して「ゴッホはひまわりを描きました」といったような文章の回答が得られるシステムである．その為には質問文を理解し，持っている美術知識と連想機能を利用して，適切な回答文を選ぶことが必要となる．これは質問文と回答文の概念同士の関係から判断する．コンピュータはこれらの自然言語の意味を人間のように理解することはできない．そのためシステムに常識的知識をもたせておくことが必要となる．

常識的知識として，美術に関する自然言語を文章で格納した「美術知識ベース」，また単語を上位－下位関係によって意味で分類し体系化した「美術シソーラス」をもちいる．これらはＷｅｂ上の美術に関する辞書から取得した自然言語表現の知識である．

また初期の知識ベースは手動で構築するものであるが，手動では労力を要するため，最初に与えた知識を基に，自動的に知

識を追加する「自動学習システム」を構築する．これにより知識を自動的に増やすことができ，より高度なシステムに発展させることができる．

美術問題に関する常識判断システムを以降「美術判断システム」と呼ぶ．

4. 美術知識ベース

美術知識ベースとは，表1に示すようなものである．

表1　美術知識ベース（一部）

知識	情報文
K1	ゴッホはオランダの画家である
K2	ゴッホの代表作にはひまわりがある
Ｋｎ	・・・

美術知識ベースはＷｅｂや美術辞書などから文章を抽出しその中で美術に関する文章で適切だと判断したものを情報文として格納し，それぞれの情報文にK1,・・・，Ｋｎという知識に分類し，格納したものである．

また美術知識ベースを人物，作品，用語，作風といった分野に分類分けする．

5. 美術概念ベース

美術に関する語は，既存の概念ベース[広瀬 02]には登録されていない固有的な分野語などが多い．そこで，新たに美術に関する概念の追加を行う．新たに追加された，美術に関して連想機能を持った概念ベースを「美術概念ベース」と呼ぶ．美術概念ベース構築の流れを図2に示す．

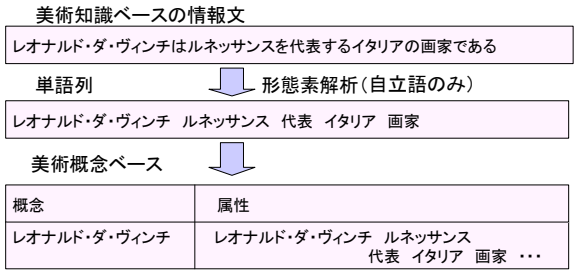

図2　美術概念ベースの構築方式

まず，Ｗｅｂから得た美術知識ベースの情報文を茶筅[奈良先端科学技術大学院大学　03]を用いて形態素解析し，自立語の単語列にする．その得られた単語一つ一つを概念とし，それぞれ自身と，文中の自分以外の単語を属性として持たせる．

その際の重み，すなわち，概念 A に新たに属性 a を追加する際の重み $w(A,a)$は，情報検索の分野で広く用いられる単語重み付け手法の一つである $tf \cdot idf$ を用いる．その際の tf は概念と追加する属性との関連度 $Rel(A,a)$とし，idf は概念ベースにおける3次属性までの出現頻度を考慮して次のような式で定義する．

$$idf = \sqrt{idf_{cb(3)}} = \sqrt{\log \frac{V_{ALL}}{V(a_i)}}$$

V_{ALL} は概念ベースに定義される全概念表記数，$V(a_i)$ は概念表記 a_i を3次属性[坂田 04]内に持つ概念数である．

よって，重み $w(A,a)$は，次のような式で定義する．

$$w(A,a) = \mathrm{Rel}(A,a) \times \sqrt{idf_{cb(3)}}$$

6. 美術シソーラス

標準のシソーラスだけでは，美術に関する知識が十分に登録されておらず，新たに美術に関する知識を別にシソーラス形式で構築する必要がある．その美術に関するシソーラスを，美術シソーラスと呼ぶ．美術シソーラスの例を図3に示す．

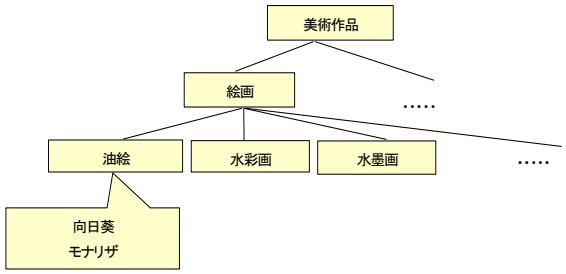

図3　美術シソーラス（一部）

美術シソーラスは美術に関するＷｅｂページを参考に手作業で構築したものである．四角で囲った語はノード，矢印のコメントボックスの語はリーフを示している．ノード「絵画」には「油絵」,「水彩画」,「水墨画」などのリーフを下位に持たせた．次にノード「油絵」に注目するとリーフが存在する．リーフには「向日葵」「モナリザ」といった知識を登録した．

美術シソーラスをこのように構築することによって，例えば「向日葵」は「油絵」であることが分かるだけでなく，「絵画」であり，また「美術作品」であるということも分かる．同じリーフに存在する「モナリザ」に関しても同じことが言える．このように美術シソーラスは美術的な包含関係を示す木構造になっている．

7. システムの構成

美術判断システムは，美術に関する問題かどうかを判断する判断部と，美術に関する問題に対して回答を行う回答部に分けられる．判断部は，自然言語の入力（質問文）に対して，それが美術に関する問題かどうかを判断する処理部分であり，回答部は美術に関する問題と判断された問題に対して答えとしてふさわしい美術知識ベース内の情報文を返す処理を行う．次の図4に美術判断システムの構成を示す．

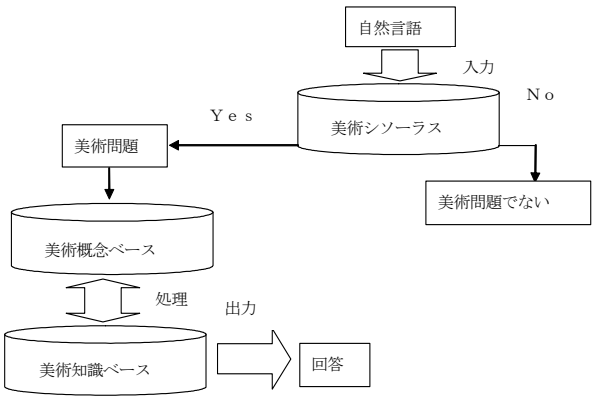

図4　美術判断システムの構成

処理の流れとして，まずシソーラスを用いて自然言語の入力文（質問文）が美術に関する問題かどうか判断を行う．この部分が判断部となる．

そして美術問題と判定されなかった場合は美術判断システムの対象外とし，美術問題と判定された場合のみ次の処理に進む．美術概念ベースを用いて自然言語と知識ベースの情報文との関連度計算[渡部　06]を行う．そして最高関連度の知識ベースの中の情報文が回答として出力される．この部分が回答部となる．

7.1 美術判断システム（判断部）

この処理は，まず入力された自然言語の文章を自立語に形態素解析する．そしてその文章の中の単語が 1 つでもシソーラスに存在すれば美術問題だと判断している．

判定部での妥当性を実験により評価した．評価データは美術に関する問題 89 文，音楽に関する問題 89 文である．美術に関する問題はシステムで美術問題と判定されれば正解として正解数/全問題数×100 を正解率とする．音楽に関する問題は美術の問題でないと判定されれば正解とする．評価結果を表 2 に示す．

表 2　美術判断システム（判断部）

	正解数	不正解数	正解率
美術に関する問題	76	13	85%
音楽に関する問題	88	1	99%
合計	164	14	92%

美術判断システム（判断部）の精度は 92%となり有効だと言える．美術に関する問題で美術に関する問題ではないと判断された文では，登録されてない美術用語を美術シソーラスに新たに追加することで解決する．美術に関する用語がない場合は関連度を用いるなど新たな手法を検討しなければならない．

7.2 美術判断システム（回答部）

この回答部では，美術概念ベースと美術知識ベースを用いることによって，質問に対して適切な解答を得る．まず，美術概念ベースを作成したときと同様に，質問文を自立語の文字列に切り分ける．美術知識ベース中の情報文と入力された質問文を，美術概念ベースを用いて関連度計算を行う．その結果，美術知識ベースの情報文の中から最高関連度の文章を抜き出す．

応答の精度評価を実験により行った．評価方法は得られた情報文が質問文の答えであれば正解とする．

評価データは美術に関する一問一答問題 73 問である．評価結果を図 5 に示す．45 問が正解（62%），28 問が不正解（38%）という結果になった．

美術判断システム（回答部）の精度は 62%となっており今後改善すべきだと考えた．有効な属性を増やす，固有名詞同士の関連度をより高くするなど考えられる．

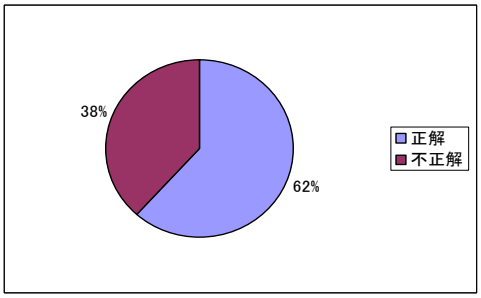

図 5　美術判断システム（回答部）の評価結果

8. 自動学習システム

美術知識ベースを手作業だけで作成するには多くの労力を要する．そこでＷｅｂなどから無作為にとってきた知識の中で美術知識だけ，美術判断システムにより抜き出し，それを美術知識ベースに追加するというシステムが自動学習システムである．次の図 6 に自動学習システムの大まかな流れを示す．

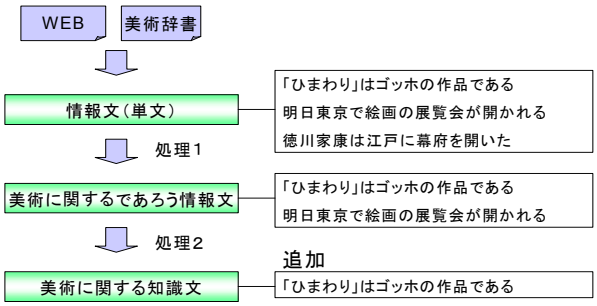

図 6　自動学習システムの流れ

まずＷｅｂや美術辞書から情報文を単文で抜き出す．その抜き出してきた情報文には美術に関係ない文章も含まれる．そこに処理 1（7.1）を行い美術に関するであろう文章だけにする．美術に関するであろう文章のなかには，美術知識ベースに格納するには不適切だと考えられる文章が含まれる．そこで次節に述べる処理 2 を行い不適切な文章を省き，適切な文章だけ新たに美術知識ベースに追加する．

8.1 自動学習システムの構成

処理 1 は 7.1 節と同様の処理となる．処理結果も同だと考えられる．処理 2 の構成を図 7 に示す．

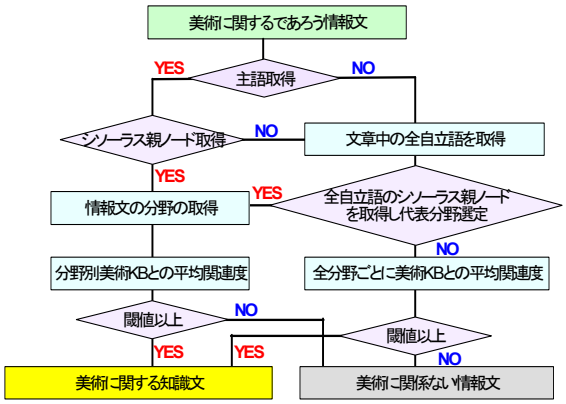

図 7　自動学習システムの構成

まず美術に関する情報文であると思われる情報文が入力されたとする．そして「主語取得」の処理に進む．主語にはその文章で重要となる単語が含まれると考えられる．例えば「ひまわりはゴッホの作品である．」といった文章では主語が「ひまわり」になっており，それがこの文章で説明しようとする美術用語となっている．

(1) まず情報文から主語を抜き出す.
(2) 次にその主語から美術シソーラスを用いて情報文が知識ベースのどの分野に判別されるか分野判定を行う.
(3) 主語の上位ノードに美術知識ベースの 4 分野のいれずれかが存在すればそれを情報文の情報分野とする.

情報文の主語取得，もしくは主語の上位ノードの取得の 2 つの処理のどちらかが失敗した場合次の処理を行う.

(1) 情報文の全自立語を抜き出し，それぞれの語の上位ノードを取得する.
(2) 最も多い分野を情報文分野とする.
最も多い分野が複数ある場合は情報文分野も複数である.
分野に関する語が一語も得られない場合は全分野とする.

情報文の分野が判定された後，美術に関する情報文かを判断する．情報文分野の分野別美術知識ベースの全情報文と関連度計算を行い，その平均関連度が閾値 0.006 以上であれば美術に関する知識文とする．閾値 0.006 は実験により良好な結果が得られた値である.

8.2 自動学習システムの評価

評価の取り方としては,

(1) 美術以外の文章 40 文
(2) Ｗｅｂから取得してきた美術知識ベースに含まれる文章とは類似していない美術に関する文章 40 文
(3) 知識ベースにある美術の文章 40 文

を用いて自動学習システムにかけ，美術以外の文章が美術に関する知識文ではないと判断されれば正解，また美術に関する文章が美術に関する知識文であると判断されれば正解とする.

図 8 に評価結果を示す.

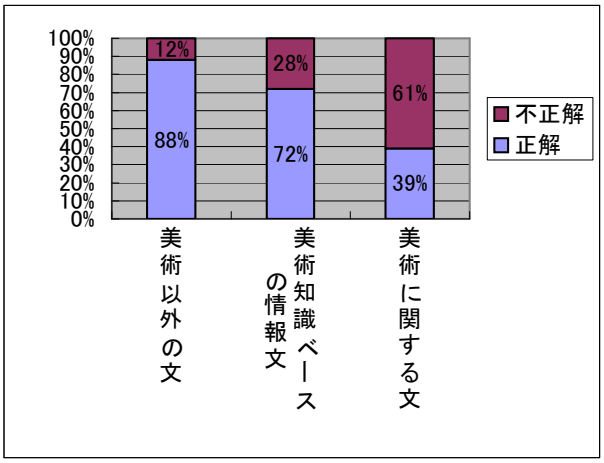

図 8　自動学習システム評価

自動学習システムの精度は 39%と低く今後改良していくべきである．より良い分類分けの方法，閾値について検討が必要である.

9. おわりに

本研究では美術に関しての会話システムと知識の自動学習システムを構築した．会話システムでは，入力した質問文に対して，持っていない知識の中から常識的と考えられる応答を抽出し返すことができた．自動学習では，現存の美術知識を基にして，Ｗｅｂなどから新たに美術に関する知識を自動的に取得するシステムを構築した．精度をあげるには，閾値の付け方，分野分けの方法について更に検討を深める必要がある.

謝辞

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行ったものである.

参考文献

[広瀬 02] 広瀬幹規，渡部広一，河岡司，“概念間ルールと属性としての出現頻度を考慮した概念ベースの自動精錬手法”, 信学技報，TL2001-49，pp. 109-116，2002.

[小島 04] 小島一秀，渡部広一，河岡司，“連想システムのための概念ベース構築法ー語間の論理的関係を用いた属性拡張”，自然言語処理，Vol. 11，No. 3，pp. 21-38，2004.

[渡部 06] 渡部広一，奥村紀之，河岡司，“概念の意味属性と共起情報を用いた関連度計算方式”, 自然言語処理, Vol. 13, No. 1, pp. 53-74，2006.

[坂田 04] 坂田光広，渡部広一，河岡司，“関連度と属性の情報価値を考慮した概念ベースの自動精錬手法”，同志社大学 理工学研究報告，Vol. 45，No. 1，pp. 14-22，2004.

[奈良先端科学技術大学院大学 03], 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座，http://chasen.naist.jp/hiki/ChaSen/, 2003