概念連想を用いたスケジュール表の理解方式

加島 幸, 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

人間とコミュニケーションを行うことのできるコンピュータを実現するためには，常識をふまえて自然言語文章の意味理解・判断を行うことができる常識判断メカニズムを組み込むことが必要である．その過程において，情報を知的に得ることが望まれる．世の中にはさまざまな形の情報が存在し，その一つである表には多くの情報や知識が含まれており，我々人間はそれらを的確に判断し，利用することができる．人間と同じようにコンピュータが表の内容を理解できれば，それを人間との会話に活かすことが可能となる．

表には，一般的な表と特殊な表がある．一般的な表とは，表理解システム[1]が解析の対象としている表のことであり，これを表 1に示す．特殊な表とは，表理解システム[1]が解析の対象としていない表のことであり，運賃表やスケジュール表，商業簿記で使用される表などがある．そこで本稿では，特殊な表の一つであるスケジュール表に着目し，概念ベース[2]やシソーラス[3]，関係辞書，時間判断[4]などを用いることにより，その構造やデータを理解する手法を提案する．また，スケジュール表に関する質問文を理解し，適切な回答を返すシステムの構築を目指す．

表 1 一般的な表のモデル

X \ Y	Y_1	Y_2	Y_3		Y_j
X_1	r_{11}	r_{12}	r_{13}		r_{1j}
X_2	r_{21}	r_{22}	r_{23}		r_{2j}
X_i	r_{i1}	r_{i2}	r_{i3}		r_{ij}

2. スケジュール表の理解

本稿では，スケジュール表の理解の定義を「入力した表がスケジュール表であるかを判断し，スケジュール表である場合はスケジュールについての質問に適切な回答を返すこと」とする．この定義を踏まえ，表理解システムの構成の概要を**図** 1に示す．スケジュール表の理解は，**図** 1において点線で囲まれた部分（スケジュール表理解システム）である．スケジュール表理解システムは，スケジュール表であるかの判断と，質問文に対する表の検索に大きく分かれる．

スケジュール表理解システムでは，構文解析ツール[5]や形態素解析ツール[6]，連想メカニズム，判断知識，常識判断メカニズム，会話処理メカニズムを使用している．連想メカニズムについては 3 章，判断知識については 4 章，常識判断メカニズムと会話処理メカニズムについては 5 章で説明する．

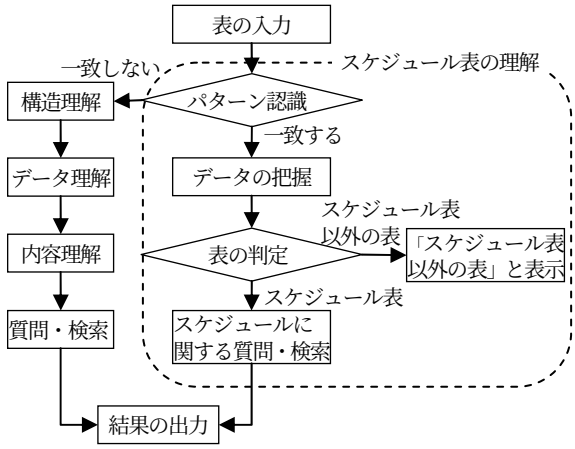

図 1 表理解システムの流れ

3. 連想メカニズム

概念ベースや関連度計算[2][7][8]といった連想メカニズムを組み込むことで，知識ベースにない語についても対応できる．

3.1 概念ベース

概念ベースとは，約 9 万語の概念が格納された知識ベースである．概念 A をその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 の対の集合として定義すると，以下のように表される．

$$A = \{(a_1, w_1), (a_2, w_2), \ldots, (a_m, w_m)\}$$

ここで，a_i を 1 次属性と呼ぶ．1 次属性 a_i は，概念ベース中の概念表記の集合に含まれているため，さらにその 1 次属性を抽出することができる．これを 2 次属性と呼ぶ．このようにして，概念ベースにおいて概念 A から n 次まで属性を抽出することができる．

3.2 関連度計算

関連度とは，任意の概念と概念の関連の強さを概念ベースの属性と重みを用いて定量化した 0 以上 1 以下の数値のことであり，関連度が 1 になるのは二つの概念が同一のときだけである．従って，関連度は概念間の関連が強いほど 1 に近い値を取り，弱いほど 0 に近い値を取る．このように関連の強さを定量化することで，概念間の関連の強弱を判断することができる．表 2に例を示す．

表 2 関連度計算の例

基準概念	対象概念	関連度
飛行機	航空機	0.418
	自動車	0.057

4. 判断知識

スケジュール表の判定には，予定として処理される語（予定語）の知識が必要である．その知識の基準となる予定語知

識ベースは，シソーラスの一部と独自に作成した予定語 DB（データベース），関係辞書によって構成されている．予定語知識ベースと予定語 DB については 6 章で説明する．

4.1 シソーラス

シソーラスとは，一般名詞を広く同義，類義的に整理したもので，約 2700 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，ノードに属する名詞として約 13 万語（リーフ）が登録されている辞書である．

4.2 関係辞書

関係辞書は，同義と考えられるもの約 20 万 6 千組，上位と考えられるもの約 13 万 7 千組，反語と考えられるもの約 1 万 7 千組，類義と考えられるもの約 9 万 2 千組を格納した辞書である．

5. 常識判断メカニズムと会話処理メカニズム

自然言語文章に対応できるコンピュータの実現には，言語の意味や概念同士の関係を知識として習得し，常識をふまえた判断ができる常識判断メカニズム，元となる入力文に対し，適切な意図理解を行い，その意図に基づいた意味理解を行うことで応答・発話処理を行う会話処理メカニズムを組み込む必要がある．本稿では，時間判断システム[4]と質問文意味理解システム[9]，質問文意味解釈システムを用いている．

5.1 時間判断システム

時間判断システムとは，時間語知識ベースとシソーラスを用いて，入力された時間語の時間軸や期間，季節などを出力するシステムである．また，知識にない語についても時間に関する情報を連想することができる．

5.2 質問文意味理解システム

質問文意味理解システムは，質問文から質問対象語（質問文が求めている対象）とその条件（質問対象語にかかっている条件）を取得するシステムである．

5.3 質問文意味解釈システム

質問文意味解釈システムは，質問文意味理解システムによって取り出された質問対象語とその条件を用いて，単文から質問文の回答に最もふさわしい語を導き出すシステムである．

例）情報文：特急は 20 日に運転を再開した
質問文：いつ特急は運転再開しましたか？
回答　：20 日

6. スケジュール表の判定

入力された表がスケジュール表であるかをコンピュータに判断させるには，スケジュール表の構造やデータを設定しておき，それを判断知識として用いる必要がある．

6.1 パターン認識

特殊な表は構造のパターンがほぼ決まっている．そこで，スケジュール表の構造パターンをあらかじめ設定しておき，入力された表がスケジュール表の構造パターンと一致するかを判定する．一致すれば，スケジュール表理解システムへ移行する．構造パターンは，表の形（m 行 3 列など）とデータの文字種（数値，空白など）の組み合わせで決める．これは，一般的な表の中にもスケジュール表と同じ形のものが多く存在するため，表の形のみでは判定できないからである．データの文字種は，形態素解析ツールを用いることで解析する．

6.2 データの把握

パターン認識によりスケジュール表理解システムに移行してきた表が，必ずしもスケジュール表であるとは限らない．そのような表を除去するために，表のデータを理解する必要がある．一致した構造パターンによってどのデータにどの文字種があるかわかっているので，データを日付と曜日，予定の候補として把握する．同じデータ内に日付と曜日の候補が格納されている場合は，データの分離を行う．

6.3 表の判定

把握したデータを用いて，入力された表がスケジュール表であるかの判定を行う．判定の基準は，

A) 日付候補に 1 以上 31 以下の整数が順に並んでいる
B) 曜日候補に曜日が並んでいる
C) 空白のデータを除いた予定候補に予定語が 60%以上ある

の 3 つである．A)と C)，もしくは B)と C)のデータがあれば，その表はスケジュール表であると判定する．

6.3.1 日付の判定

日付候補が「22」のような表記の場合は，そのまま数値の範囲と順序を調べ，「22 日」「2006/2/26」のように数値以外の文字が含まれた表記の場合は，分離することで数値のみのデータを取得して調べる．また，「22 位」のように『日』以外の単位が数値についている場合は，分離して調べなくても日付ではないことがわかる．

6.3.2 曜日の判定

曜日候補が曜日であるかの判定は，曜日を表す語を集めた曜日関係知識ベース（21 語で構成）を用いて行う．具体的には，曜日候補にあるデータが曜日関係知識ベースの名称と一致すれば，そのデータは曜日であるとする．

6.3.3 予定の判定

予定候補が予定であるかの判断には，予定語知識ベースが必要となる．予定語知識ベースは，シソーラスで予定語と考えられるノード・リーフを主として，シソーラスにはない予定語を集めた予定語 DB，予定語 DB 内の予定語と同義・類義・反対関係にある語（予定語 DB 内の予定語と関連度計算を行い，実験的に求めた閾値0.35以上の関連度がある語のみ）から構成されている．

以下の 2 つの手法によって，予定の判定を行う．

①表記一致

予定候補にあるデータは文となっていることもあるため，形態素解析ツールを用いることでデータ内の自立語を抽出す

る．それぞれ予定語知識ベースにある語と比較し，一つでも一致していればそのデータは予定語であるとする．

②未知語処理

予定語知識ベースに限らず，知識ベースには代表的な語のみ格納している．文章中に用いられる語で知識ベースにない語については未知語として扱い，連想メカニズムによって代表語の中で関連が強いものに置換する処理（未知語処理）を行う．この処理により限定された知識から様々な語に対応することを実現している．未知語処理のイメージを図 2に示す．

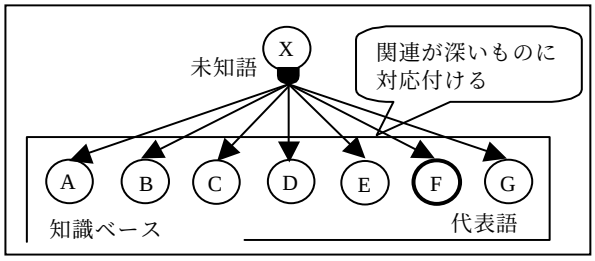

図 2　未知語処理のイメージ

ここでは予定候補のデータから抽出した自立語を未知語とし，予定語知識ベースと関連度計算を行う．関連度が実験的に求めた閾値0.35以上あれば，その未知語は予定語とする．

6.4 評価

提案した表の判定方法の評価を行う．人手によって作成されたスケジュール表15件，スケジュール表以外の表15件に対し，正しく判定が行えたかを人目で評価する．この結果を表 **3**に示す．

表 3　**表の判定の評価**

表の種類	正解率
スケジュール表	87%
スケジュール表以外の表	93%

87%の精度でスケジュール表を「スケジュール表」として正しく判断できた．誤った表は，形態素解析を行ったときに正しく分離できていなかった．英字やカタカナのデータは一つの単語と考え，未知語理解支援システム[10]を用いてシソーラスのノードを獲得し，予定語知識ベースと比較することなどが考えられる．

93%の精度でスケジュール表以外の表を「スケジュール表以外の表」として正しく判断できた．誤った表は実際に行ったことについて書いた表であり，内容が予定と似通っているため判断できなかった．時制が過去である場合はスケジュール表以外の表だと判定することで解決できると考えられる．

7. スケジュール表に関する質問・検索

表を解析したコンピュータが，表に対する質問に答えることが出来れば，そのコンピュータは正しく表を解析し，理解出来ているということが可能である．

7.1 スケジュール表の正規化

スケジュール表にあるデータの種類はほぼ決まっているので，表の検索を容易にするためにスケジュール表を正規化する．正規化後の表のモデルを表 **4**に示す．

表 4　**正規化後の表のモデル**

日付	曜日	時間	予定	第○曜日
1～31	日～土	0：00～23：59	文 or 空白	1～5

「日付」「曜日」「予定」の他に「時間」「第○曜日」という項目を追加している．

7.2 質問対象語とその条件の抽出

質問文意味理解システムを用いることで，質問文から質問対象語を抽出することはできるが，その条件の抽出は正しく行えないことがある．そこで，条件の抽出は本システムで行うことにする．その準備として，質問文意味理解システムから質問対象語と，質問文から質問対象語を導いた語を除いた文を受け取る．

7.2.1 質問対象語の処理

質問対象語と正規化した表との対応付けを行うために，質問対象語知識ベース（7 種で構成）を作成し，用いる．例えば，対象語が「日にち」なら日付を参照する．質問対象語が質問対象語知識ベースにない場合は，予定の一部分について質問していると考える．予定を情報文とし，質問文意味解釈システムを用いることで，適切な回答の抽出を行う．

7.2.2 条件の抽出

本システムにおける条件の抽出は，条件の一文を構文解析ツールと形態素解析ツールを用いて，以下の処理を行う．

A)　接頭詞の場合，次の単語と複合

B)　名詞の場合，

I.　①単語に『月』が含まれており，次の単語が範囲知識ベースに存在する，②単語が数値であり，次の単語が『：』，その次の単語が数値，③次の単語が名詞のとき，各単語を複合

II.　①次の単語が『の』であり，その次の単語が名詞または形容詞，②次の単語が『が』であり，その次の単語が形容詞のとき，各単語を複合

C)　形容詞の場合，次の単語と複合

D)　動詞の場合，単語を基本形に変換

この処理によって，質問文意味理解システムでは抽出できなかった条件を取り出すことができる．

7.2.3 条件の処理

取り出した条件を時間判断システム，形態素解析ツール，関連度計算などを用いて，以下の 4 種類に分類する．

A)　日付の条件：「5」のような具体的な数値に変換

B)　曜日の条件：「水」のような漢字一文字に変換

C)　時間の条件：「15 時 20 分」のように変換

D)　予定の条件：表にある単語や複合語，空白に変換（表記一致，未知語処理による）

これによって，参照するデータが決定する．

7.3 データの抽出

質問対象語とその条件を用いて，正規化した表から求めるデータを取り出す．これを質問・検索システムと呼ぶ．この抽出のイメージを図 **3**に，流れの説明を以下に示す．

まず，「ゼミは何時にありますか？」という質問文から質問対象語「時間」とその条件「ゼミ」「ある」を抽出する．質問対象語知識ベースから，この質問文は時間のデータを求めていることがわかる．次に，条件の「ゼミ」と「ある」がどの条件であるかを調べると，「ゼミ」は予定の条件に，「ある」はどの条件にも属さないものとなる．最後に，条件から予定のデータを参照し，「ゼミ」のある時間を抽出する．

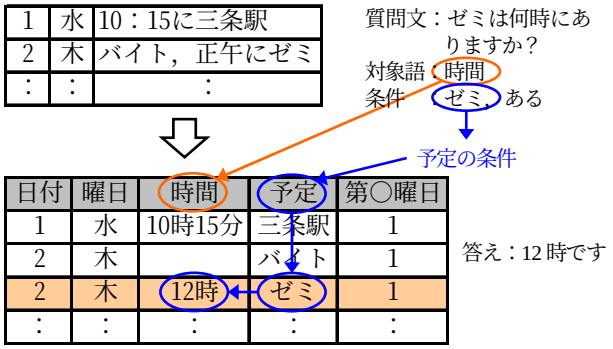

図 3　データ抽出のイメージ

7.4 評価

提案した質問・検索システムの評価を行う．人手によって作成されたスケジュール表 10 件に対し，協力者 10 人が手作業で作成した質問文，計 387 問を入力し，正しく答えを抽出できるかどうかを正解，不正解で評価する．

質問・検索システムは質問対象語知識ベースにある語については適切な回答を，ない語については情報文の抽出を目的としているため，システム自体の正解率と，質問文意味解釈システムを用いた質問・検索システムの正解率の 2 パターンを評価し，結果を表 5に示す．

表 5　**スケジュールに関する質問・検索の正解率**

評価システム	正解率
質問・検索システム	60.5%
＋質問文意味解釈システム	50.3%

質問・検索システム自体は約 60%の正解率を得ることができ，質問文意味解釈システムを用いた場合は約 50%の正解率となっている．これは，質問・検索システムが抽出した情報文は正しいが，質問文意味解釈システムにおいて，「友達」なら人物，「駅」なら場所といった情報文の意味的分割が適切に行われていなかったためである．

不正解となったのは，A)条件が複数あることによって正解以外のデータも抽出したもの，B)質問対象語が「時間」であり，質問文に「いつ」という語が含まれている場合において，人間ならば日付を求めていることがわかる文でもコンピュータには判断できず，時間を抽出したもの，その他として「最初のバイトから最後のバイトまでの日数」といった複雑な質問文などである．A)，B)の失敗への対処法としては，以下のようなものが考えられる．

A)　条件に優先度を付加して余分なデータを抽出しない

B)　「いつ」が「日付」であるか「時間」であるかの判断は人間でも難しいため，この場合は日付と時間の両方を抽出

8. おわりに

本稿では，コンピュータに特殊な表を理解させる手法を提案した．連想メカニズムや常識判断メカニズムなどのメカニズムを組み込むことで，少量の知識から多様な表現に対応することができた．これまで一般的な表にしか対応できなかった表理解システムを，多様な表に対応できるシステムへと拡張することを目的としており，本稿ではスケジュール表に着目し，その理解を行った．他の特殊な表についても構造パターンを設定し，データの把握などを行うことで理解することが可能であると思われる．

コンピュータが様々な表を理解できるようになれば，人間と同じ判断を行うことも可能となり，人間との会話に活かすことができると考えられる．

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．

[参考文献]

[1] 権 東旭，古川 成道，渡部 広一，河岡 司，“常識判断に基づく知能ロボットのための表理解システム”，信学技報，NLC2004-11，pp.1-6，2004

[2] 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成法－属性信頼度の考え方に基づく属性重みの決定”，自然言語処理，Vol.9，No.5，pp.93-110，2002

[3] NTT コミュニケーション科学研究所監修，「日本語語彙体系」，岩波書店，東京，1997

[4] 土屋 誠司，奥村 紀之，渡部 広一，河岡 司，“連想メカニズムを用いた時間判断手法の提案”，自然言語処理，Vol.12，No.4，pp.111-129，2005

[5] 奈良先端科学技術大学院大学　構文解析システム「CaboCha」

[6] 奈良先端科学技術大学院大学　形態素解析システム「茶筅」

[7] 笠原 要，松澤 和光，石川 勉，“国語辞書を利用した日常語の類似性判別”，情報処理学会論文誌，Vol.38，No.7，pp.1272-1283，1997

[8] 渡部 広一，河岡 司，“常識的判断のための概念間の関連度評価モデル”，自然言語処理，Vol.8，No.2，pp.39-54 ，2001

[9] 古川 成道，渡部 広一，河岡 司，“概念ベースを用いた知的検索における曖昧な質問文の意味理解”，人工知能学会全国大会，2D1-10，2004

[10] 伊藤 俊介，渡部 広一，河岡 司，“情報検索における未知語理解支援方式～未知語のシソーラスノードへの分類～”，情報処理学会自然言語処理研究会資料，2004-NL-159，pp.61-66，2004