E-30

時間の常識的判断メカニズムとその未知語処理

An Intelligent Mechanism to Understand Time and Its Unknown Word Processing

野村 理樹† **渡部 広一**† **河岡 司**†
Riki Nomura Hirokazu Watabe Tsukasa Kawaoka

１．はじめに

現在，コンピュータは人間の道具として非常に便利な存在であるが，今後，人間のパートナーとしての役割がこれまで以上に期待されると考えられる．そこでは人間とコンピュータとの双方向会話によるコミュニケーションが重要となる．本研究ではそのための基礎要素技術として，コンピュータに時間を理解させる手法を提案している．

本研究の基本は，時間を表す言葉（以下時語と記す）の知識ベース（以下ＫＢと記す）を用いて名詞から季節や時刻などの時間を判断する（以後この処理を時語理解と呼ぶ）ことである．またＫＢにない語（以下未知語と記す）については，概念についての汎用知識を集めた概念ベース[1]を利用して補完する方法を示す．また，用言ＫＢを持たずに，体言と用言の組み合わせ（単文）から既知の時語を生成し，時間を判断する方法を示す．

２．時語知識ベース[2]

まず，「元旦」など時間を表す言葉（時語）のＫＢを説明する．時語は，大きく明示的時語と暗示的時語の二つに分類できる．

明示的時語とは，「クリスマス」が 12 月 25 日のことをいうように，明らかな時間を指す言葉であり，「絶対時語」などの分類がある．

暗示的時語とは，「スキー」から冬を連想するように，その言葉自体は時間を指す言葉ではないが，暗黙に時間を示す言葉であり，一日の中での時間を暗示する「時語日」などの分類がある．表１に時語ＫＢの分類の一部を示す．

表１：時語ＫＢの内容分類

分類	説明	例	語数
絶対時語	絶対的に時間を指す語.	元日	156
時語日	一日の中での時間を暗示する語.	朝日	41
時語年	一年の中での時間を暗示する語.	紅葉	140

３．時間判断メカニズム[2]

時語ＫＢに存在する語が入力された場合，時語ＫＢを参照する事により，時語理解ができる．しかし，全ての語を時語ＫＢに格納する事はできないため，未知語の処理が必要となる．また，単文から時間を判断するためには体言と用言の組み合わせを処理する必要がある．用言ＫＢを作成する方法も考えられたが，体言と用言の組み合わせは膨大であり例外が頻発すると考えられるため，用言ＫＢを持たずに処理を行うこととする．本論文ではこれらに対して，概念ベースを用いた処理を提案する．

時間判断メカニズムの全体像は図１のようになる．

†同志社大学大学院工学研究科
Graduate School of Engineering, Doshisha University

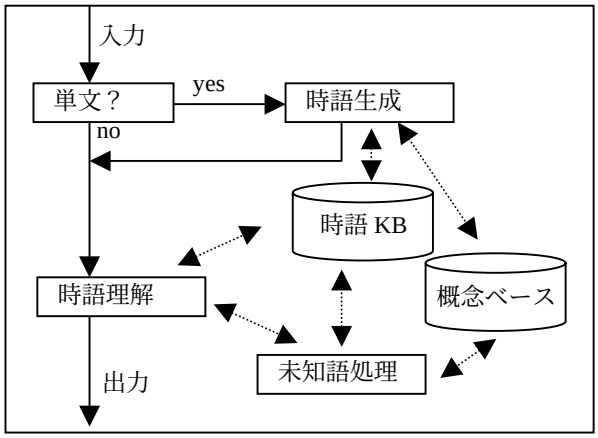

図１：時間判断メカニズム概要

４．概念の属性と重みを利用した未知語の処理

未知語が入力された場合や単文が入力された場合，概念ベースを用いて既知の時語を生成する処理を行う．

ここで概念ベースについて説明する．概念ベースとは語（概念）と意味（属性）のセットを約 10 万語分蓄積している，国語辞書等から自動構築された汎用データである．ある概念Ｘに対して，その意味特徴をあらわす属性（それらを概念Ｘに対する１次属性と呼ぶ）と，整数値で表された，その属性の重要度（重み）の組が数個～数十個ある．個々の属性はまた，概念ベースに登録されている概念であり，それに対する１次属性が登録されている．概念Ｘの１次属性の１次属性を，概念Ｘの２次属性と呼ぶ．

以下に未知語処理の流れを示す．ここで，未知語 X の１次属性を X_1，２次属性を X_{11} のように記す．

① 概念ベースから未知語 X の１次，２次属性を取得する．
② 後述の代替手法により，未知語 X の１次属性 $X_1 \cdots X_n$ それぞれが，未知語という分類か，代表語（春，梅雨，夏，秋，冬，朝，昼，夕方，夜）に代替される．
③ X_i が代表語に置き換わったら，X_i の，未知語 X に対する重み a_i の値だけその代表語の得票に加算する．これを１次属性全てに繰り返す．
④ 重みの合計数値が一番高かった代表語が，基準とする得票割合の下限 Th 以上を獲得している場合，未知語 X の代替語となる．得票数が同じものが複数ある場合は，関連度計算により一つに決める．

ここで，②の代替手法を，X_1 を例に挙げて説明する．

1) ２次ｆ属性（$X_{11} \cdots X_{1n}$）それぞれの語が時語ＫＢに存在するかどうかをみて，あれば代表語に，なければ未知語という分類に置き換える．
2) $X_{11} \cdots X_{1n}$ の中で置き換えられた代表語の多数決をとり，最大得票数の代表語で X_1 を代替する．同点の

場合は関連度計算により一つに決める．$X_{11}\cdots X_{1n}$ が全て未知語の場合は，X_1 も未知語とする．

なお，関連度計算とは概念間の関連性を定量的に評価する計算手法[1]である．

５． 単文からの時語生成

ここでは，「葉が落ちる」→「秋」などのように単文から時語を生成し，時間を判断する方法を述べる．なお本論文で扱う体言と用言の組み合わせは必ずしも主語・述語の関係にはなってないが，簡単のためこれを単文と呼ぶ．以下に，単文からの時語生成の流れを示す．

まず前処理として，体言が Thesaurus 上で"施設"のノード以下に含まれる場合，体言を「館」に置き換える．

次に，以下の順に時語生成を試み，どれかの段階で生成が成功すればそれを出力して終了する．時語が生成できなかった場合は「時間に無関係」と判定して終了する．

1. 体言中の漢字と用言中の漢字を組み合わせて時語を生成する．
2. 概念ベースから用言の属性を取得し，属性中の他の用言を用いて体言との組み合わせを行う．
3. 時語の中で，その属性中に体言と用言の両方を含むものを検索する．
4. 用言を無視し，体言のみから時間を判断する．体言が時語ＫＢにない場合は未知語処理を行う．
5. 体言を無視し，用言のみから時間を判断する．具体的には，未知語処理を用言に適用する．

Thesaurus とは一般名詞の意味的用法を表す 2700 の意味属性の上位下位関係，全体部分の関係が木構造で示されたものであり，約 13 万語が登録されている．

６． パラメータ調整と評価

アンケートによって，以下のデータを収集し，それぞれを時間判断システムがどれだけ正解できるか調べた．

Ａ群：時間に関係ある体言　343 個
Ｂ群：時間に関係ない体言　303 個
Ｃ群：時間に関係ある単文　256 個

４節に示した未知語処理方式中の，得票割合の下限 *Th* を，未知語 *X* の１次属性の重み合計に対する割合として 15~30%まで 5%刻みで変化させた場合の，時間判断の正解率を図 2~4 に示す．図中で「無答」とは「時間に無関係」と出力された場合である．

Th の値を高く設定するほど無答率が高くなり，時間に関係ある場合の全体の正解数は漸減する．しかし，出力がある場合の正解率は上がっていき，誤答率が下がる，すなわち出力の信頼性が向上することがわかる． *Th* を 25%とした場合，A 群の正答数は 144，誤答数は 14 で，出力がある中での正解率は 91%に達する．また B 群の正解率は 94%と非常に高い．C 群の正答数は 113，誤答数は 50 で，出力がある中での正解率は 69%と若干低いが，全体としては高い出力の信頼性が得られたと考えられる．

７． おわりに

時間判断メカニズムにおいて，誤答率を増やさないようにしつつ，概念の２次属性と重みを使って未知語処理を行う方法を新たに提案した．それによって，時語ＫＢに無い語の入力があっても，信頼性の高い出力が期待できる．

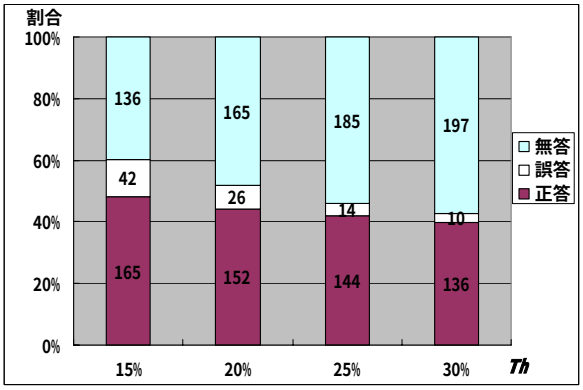

図２：Ａ群「時間に関係ある体言」の結果

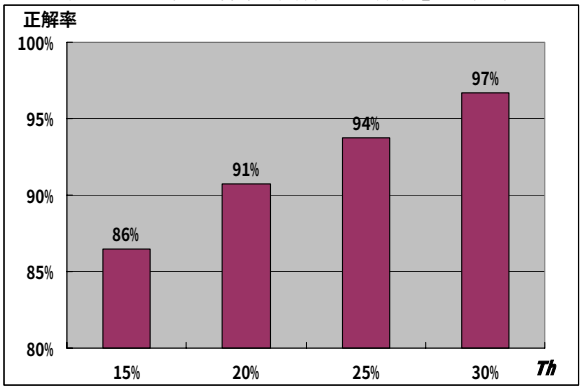

図３：Ｂ群「時間に関係ない体言」の結果

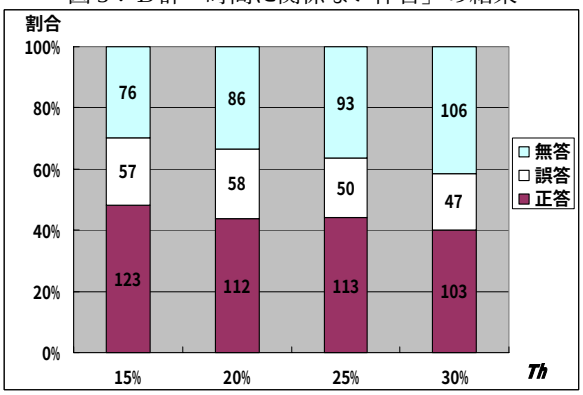

図４：Ｃ群「時間に関係ある単文」の結果

Th を全体に対する割合ではなく得票数の定数とする方法も考えられるが，属性の数や重みの合計は概念によって変化するため，重みの合計に対する割合とするほうが適切であると考えられる．しかしその割合の調整は，全体の正解率と出力の信頼性のトレードオフ関係があるため，他の手法との組み合わせを含めて今後の研究が必要である．

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「知能情報科学とその応用」における研究の一環として行った．

参考文献

[1]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 (2001.4)

[2]小畑陽一，渡部広一，河岡司：単文の名詞と動詞から時間／季節を判断するメカニズム，信学技報，AI2000-56，pp.1-6 (2001.1)