

文の意味理解に基づく常識的時間判断システムの構築

岩瀬 元秀†

渡部 広一†

河岡 司†

†同志社大学大学院工学研究科知識工学専攻 〒610-0394 京都府京田辺市多々羅都谷 1-3
E-mail: † dtf0720@mail4.doshisha.ac.jp, ‡ {hwatabe, tkawaoka}@mail.doshisha.ac.jp

近い将来、自然な方法で要求を理解してくれるような知性を持った知能ロボットが求められる。そのようなロボットには、知的な会話機能と、様々な常識を理解する機能が必要である。時間表現の理解は会話機能の重要な要素の一つである。

本稿では、日常的な語句が暗示する時刻や季節をコンピュータが判断する手法を述べる。その際、単語のみから判断を行うのではなく文全体の意味を考慮することにより、多様な入力語に対応し結果を導く方式を提案する。

Construction of Commonsense Judgment System for Time Based on Understanding of the Meaning of Sentences

Iwase Motohide†

Watabe Hirokazu†

Kawaoka Tsukasa†

†Department of Knowledge Engineering and Computer Sciences,
Graduate School of Engineering, Doshisha University
1-3 Miyakodani Tatara Kyotanabe-shi, Kyoto, 610-0394 Japan

In the near future, we need intelligent robots which understand user's demand by the natural way. For such robots, it is important to have an interactive conversation function and have an ability to understand various commonsense. An understanding of time expression is one of the important elements of the conversation function.

This paper proposes a mechanism to understand time expression with knowledge base of usual words. Especially, this paper proposes the method that leads the result not by a word alone but by considering a meaning of an entire sentence, in order to accept very various input words.

1. はじめに

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、現状ではその利用に際して、コンピュータ特有の操作や仕組みを理解していなければ、十分に機能を引き出すことが難しい場合が多い。一方で、コンピュータの役割としては、今後、人間のパートナーとして介護ロボットや秘書ロボットなど、知能ロボットに組み込まれる知的なものはたらきがより強く期待されると思われる。

そのような利用形態においては、人間が機械の側の制約によって受けるストレスや負担を抑え、快適に意思を伝えることができるような機能が必要である。具体的には、人間と知能ロボットの双方向会話を実現する機能や、柔軟な判断機能、言い換えれば「ロボットが常識を持つこと」が必要であり、そのための知的会話メカニズムが提案されている。

知的会話メカニズムの上層には、外界の理解や人間の要求を理解した上でその要求に応じた出力を返す層が存在する。例えば広告や看板などの環境を理解し、また人間の「温かいものが食べたい」等の要求に対して飲食店の位置や情報を出力する。中間層は感覚・知覚に関する判断や感情に関する判断等、人間が日常的に行っている種々の常識的な判断を担う。例えば、先の要求文中における「温かいもの」とはどんなものを指す

か等を判断することで、上層のメカニズムに、柔軟性を持たせる。下層には概念処理を行うメカニズムを構築し、中間層の判断メカニズムに対して連想などの機能を提供する。それは人間で言えば、知らない言葉を聞いたとき過去の記憶から大体的見当をつけて類推することができる、あるいは、辞書や周囲の人間に聞く等して対応可能であることに相当する。

本稿では知的会話メカニズムにおける、時間に関する表現の理解処理を提案する。

2. 研究概要

人間がストレス無くコンピュータとの双方向会話を行うためには、コンピュータに、会話中に出現する時間に関する要素・表現を理解させる必要がある。

理解が必要な要素・表現としては表 2.1 のような種類が挙げられる。

表 2.1 時間に関する要素の理解が必要な表現

種類	会話文例
時刻・季節	「スキーに行きたい」→「冬」, 「日が沈む」→「夕方」
所要時間	「お風呂に入ってきます」→「15分」
歴史的話題	「新撰組の近藤勇が・・・」→「江戸時代」

時刻・季節は、時刻や季節に関する知識や理解が必要な場面である。人間は無意識のうちに、「スキーに行きたい」というのは、冬のことを言っている、といった判断を行っていると思われる。

所要時間は、時間の長さに関する常識である。「お風呂に入ってきます」と言われたら、2〜3分が出てくる、あるいは3時間以上出てこない、とは普通の人は思わないが、それは所要時間に関する常識を持っているからである。

歴史的な話題は広く知られている過去の出来事の共通知識が元になり、その知識が無い、あるいは話題を理解するための常識が無ければ円滑な会話を行えない。

これらの中で、明示的に時刻や日付等を意味する表現をコンピュータが扱うためのモデルなどが研究されている^[4]。また、明示的では無いが時刻や季節等を想起させるような表現からでも、人間が抱くのと同じように時刻や季節等を導出できるようなシステムが提案されている^[5]。

本稿では、時間に関する表現のうち、時刻・季節に関する判断方式を提案する。これは知的会話メカニズムの基礎要素となる。

3. 時刻・季節判断システム

人間は会話の中で、意識して、あるいは無意識のうちに、話題の中の時刻や季節についての理解を行っている。例えば、待ち合わせや、旅行の計画を立てる際、あるいは「葉が赤くなりましたね」という言葉に「もう秋も深いですね」と応じる自然な反応には、時刻や季節に関する判断が不可欠である。

本稿では日常的な時間表現に注目し、限られた知識を時間の観点で使用する時刻・季節判断システムの提案を行う。その際、単語のみから判断を行うのではなく文全体の意味を考慮することにより、多様な入力語に対応する柔軟性を持たせる。

本章では、日常的な時間表現をコンピュータに理解させる、時刻・季節判断システム(本章ではこれを「本システム」と呼ぶ)を提案する。基本的な枠組みとして、時間を表現する言葉(以後これを時語と呼ぶ)を蓄積した知識ベース(以後時語 KB と呼ぶ)を用いて時間表現を理解する。しかし、すべての言葉を知識ベースに格納すると知識は膨大な量になり構築は不可能である。そこで、知識ベースに存在しない未知の言葉を既知の語に帰させる手法が必要である。これには電子化国語辞書などから構築された概念ベース^[6]と、概念間の関連の強さを計算する関連度計算^[6]を用いる。

3.1 問題の定義

まず、この研究において使用する言葉や問題を定義する。

本システムで扱う時間表現は、次の①から③、あるいはその組み合わせとする。

- ① 数字と時間単位あるいはそれに準ずるものによって表された時刻・日付
- ② 体言と用言の組み合わせ(単文)
- ③ 体言

時語は、明示的に時間を表す「明示の時語」と、暗黙的に時間を想起させる「暗示の時語」という分類に分ける。また、時間の判断結果は、地球の自転と公転に基づいて定義される、時間の基本的な周回単位である「日」の中の時刻・時間帯か、「年」の中の日付や季節、あるいはそれらの両方とする。特に基本的に重要であり、それらの中に含まれない時間帯や季節が無いような語集合を構成する語を代表語と呼ぶ。代表語に属する語は「春、梅雨、夏、秋、冬、朝、昼、夕方、夜」の9語である。システ

ムへの入力に暗示の時語の場合には、最終的に代表語を判断結果として出力する。

季節や時刻に関係あるかどうかの判断ができ、時間帯・季節を想起させるような語において、想起される時間(代表語)を出力できた場合を「時語を理解した」と定義する。本システムでは、あらゆる一般名詞、あるいは時語の組み合わせに対する時語理解を目標とする。

3.2 時語知識ベース

3.2.1 明示の時語

12月25日のことを指している「クリスマス」などのような、明らかな時間を指す言葉である時語を絶対時語と呼び、157語登録されている(表3.1)。それぞれ時語に対応する日付や時刻および代表語が収録されている。フィールド「ID」は管理用の通し番号、フィールド「名称」には時語、フィールド「ふりがな」には時語の振り仮名が登録されている。フィールド「時間軸」には、その語が指す時間の単位(月日時分秒)の種類が数値で示され、それに応じて始①、始②、終①、終②フィールドを参照することで、語句である時語から時間の数値的表現を得ることができる。フィールド「代表語」には、時語が表現する時間を包含する代表語が格納されている。

表 3.1 時語 KB における絶対時語テーブルの一部

ID	名称	ふりがな	代表語	時間軸	始①	始②	終①	終②
16	盆	ぼん	夏		7/8/13		8/16	
17	真夏	まなつ	夏		2/8		8	
130	早朝	そうちょう	朝		9	4:00		6:00

また、ある時点基準として相対的に時間を表す語を相対時語と呼び、93語登録されている(表3.2)。フィールド「時間軸」は、時語が示す時間軸の種類(年月日時分秒)を表しており、フィールド「始」「終」が示す値はその軸を単位としていることを示す。フィールド「始」「終」は基準時点から相対的にどの時点を指すか、その始まりと終わりを表す。

フィールド「基準」は基準として現在の時点しか対応しない時語ならば n 、過去の時点に対応するならば $-x$ 、未来の時点に対応するならば $+x$ 、どちらにも対応するならば x がそれぞれ格納されている。例えば「昨日」の場合は n が格納され、また「前日」の場合は x が格納される。

表 3.2 相対時語テーブルの一部

ID	名称	ふりがな	時間軸	基準	始	終
39	今	いま	分	n	0	0
48	あくる日	あくるひ	日	x	1	1
56	去年	きょねん	年	n	-1	-1
58	今月	こんげつ	月	n	0	0
71	先月	せんげつ	月	n	-1	-1

相対的な意味と絶対的な意味双方を持つ「昨晚」(昨日+晩)などの時語は、複合時語と呼び、30語が登録されている(表3.3)。元の時語を、単独で意味をなす時語の組み合わせで定義することによって、理解可能な形にする。

表 3.3 複合時語テーブルの一部

ID	名称	ふりがな	再定義1	再定義2
2	あくる朝	あくるあさ	明日の	朝
11	一昨晚	いちさくばん	一昨日の	晩
18	今夏	こんか	今年の	夏
22	今宵	こよい	今日の	晩

また、明示的な時刻・日付などの表現には、時間を示す単位が用いられる。そこで、「1999年」や「2月」のように数字と組み合わせられる時間の単位 14 語を単位時語として登録する(表 3.4)。フィールド「時間軸」には、その語が表す時間のスケールが示されている。

「次の」や「前の」のように、未来方向や過去方向を示し、かつ、その方向への距離も示しているベクトル的な語を、指定時語と呼び、9 語から成る。表 3.5 にその一部を示す。フィールド「値」は、時語が表す方向が未来方向なら正、過去方向なら負の値であり、絶対値はその方向にどれだけ移動するかを表す。

表 3.4 単位時語テーブルの一部

ID	名称	ふりがな	時間軸
1	年	ねん	年
6	日	にち	日
7	時間	じかん	時

表 3.5 指定時語テーブルの一部

ID	名称	ふりがな	値
1	次の	つぎの	1
2	前の	まえの	-1
7	この	この	0

3.2.2 暗示的時語

人間であれば「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、暗黙に時間を連想する暗示的時語が存在する。一日の中での時間を暗示する語である「時語日」を 43 語と、一年の中での時間を暗示する語である「時語年」を 144 語、二種類に分けて時語 KB に格納している。表 3.6 に時語日のテーブルの一部を、表 3.7 に時語年のテーブルの一部を示す。

いずれも、フィールド「ID」には管理用の通し番号、フィールド「名称」に時語、フィールド「ふりがな」に時語の振り仮名、フィールド「代表語」に時語に対応する代表語が登録されている。なお、知識ベースにデータとして登録されている語は、すでに意味が分かっている語であるため、これ以後「既知語」と呼ぶ。

表 3.6 時語日テーブルの一部

ID	名称	ふりがな	代表語
1	起床	きしょう	朝
44	帰宅	きたく	夕方
41	鶏鳴	けいめい	朝
2	就寝	しゅうしん	夜

表 3.7 時語年テーブルの一部

ID	名称	ふりがな	代表語
97	アイスクリーム	あいすくりーむ	夏
41	紫陽花	あじさい	梅雨
110	枯葉	かれは	秋
45	肝試し	きもだめし	夏
9	蟬	せみ	夏

3.3 システムの処理の流れ

入力から出力を得るための、本システムの処理の概要を示す。

時語 KB には時語として体言のみ格納している。システムの入力は複数の語の組み合わせも対象とするため、時語知識ベースを単純には参照できない。そこで入力を解析し、時語に切り分ける必要がある。入力の解析は、前処理、形態素解析、後処理の三段階に分けられる。入力に単文が含まれていた場合、後に述べる「時語生成処理」を行う。また、知識ベースに格納されていない語には「未知語処理」を行う。

3.3.1 入力の解析—前処理

形態素解析ツールを利用する前に、数字の処理と記号の処理を行う。

まず、入力に含まれる漢数字を半角アラビア数字に置換する。また全角アラビア数字も半角アラビア数字に置換する。これは表 3.8 に示すような形態素解析結果の揺れを防ぐためである。また漢数字でも「十五夜」のように単語を形成するものは置換してはならない。このような誤りを防ぐために置換可能か否か時語知識ベースを参照する必要がある。時語知識ベースに存在する語は、漢数字を含んでも置換の対象としない。

表 3.8 形態素解析ツールの特徴例

形態素解析前	形態素解析後
7月	7, 月
7月	7月

また、入力に含まれる全角スラッシュ、全角コロンをそれぞれ半角に置換する。これは後の処理の簡易化だけでなく、形態素解析の揺れを防ぐためである。

3.3.2 入力の解析—形態素解析

前処理をした上で、一般のツールを用いて形態素解析を行い、単語・記号単位に入力を切り分ける。

3.3.3 入力の解析—後処理

形態素解析を行っただけでは単語の曖昧性が解消されない等のため、次に示す後処理を番号順に行う。

①複合時語の展開

複合時語は「今朝＝絶対時語&相対時語」のように複数の分類にまたがるものである。これを「今日の朝」の様に展開することにより、既存の分類の組み合わせに帰着させる。

②日付の曖昧性解消

「3日」という言葉は日付を示す絶対時語であるが、「3日後」であれば相対時語である。また、「3日間」であれば単に長さを示す。以上から「3日」は後ろに続く語により働きが異なる曖昧性を持っている。そこで以下に示すルールに従って曖昧性を解消する。なお、「前」と「後」を前後時語と記す。

・数字 + 単位時語 + 「間」 → 数字 + 長さ単位時語 + 長さ助詞

・数字+単位時語+前後時語 → 数字 + 前後単位時語 + 前後時語

・数字+単位時語+ not(長さ助詞, 前後時語) → 数字+日付単位時語

③前後時語と指定時語の曖昧性解消

「前」という言葉は、「3日前」と「前の月」では働きが異なる。そこで「前」の分類は前後の言葉から決定する必要がある。ここで、前後時語と指定時語の和集合を前後指定時語と記すと、曖昧性解消のためのルールは以下ようになる。

・数字+単位時語+前後指定時語 → 数字+単位時語+前後時語

・前後指定時語+単位時語 → 指定時語+単位時語

3.4 未知語処理

未知の語に対して、連想概念ベースや関連度計算手法などを用いて、知識ベースに存在する既知語に帰着させることを未知語処理と呼ぶ。ここでは、時刻・季節判断システムで使用する未知語処理方式について説明する。

3.4.1 概念ベースと関連度計算

概念ベースとは語(概念)と意味(属性)のセットを約9万語分蓄積している、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 C に対して、その意味特徴をあらわす属性(それらを概念 C の1次属性と呼ぶ)と、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 C の1次属性の1次属性を、概念 C の2次属性と呼び、概念 C の意味はその n 次属性までの連鎖によって表される。

また、関連度計算によって、概念と概念の関連の強さを定量的に評価した値(関連度)を得ることができる。関連度は $0 \sim 1$ の実数値で表され、その値が大きいほど関連が深いことを表す。

概念ベースのデータの例を図 3.1 に、関連度の例を図 3.2 に示す。

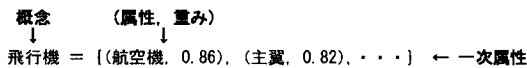


図 3.1 概念ベースのデータ例

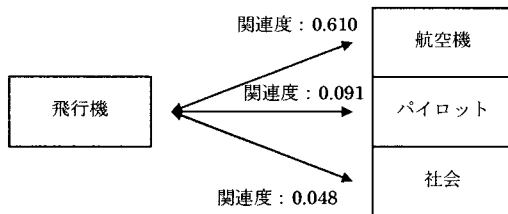


図 3.2 関連度の例

3.4.2 既知語と極めて関連の高い語への対応

特定の既知語と同義かあるいは非常に関連の深い未知語の場合、時間的観点においてもその既知語と同等な判断を行うことができると考えられる。

以下に、関連度計算を利用した、既知語と極めて関連の高い未知語に対する未知語処理の手法を示す。

1) 未知語 X と、すべての既知語との関連度計算を行う。

2) 未知語 X と最も関連度の高い既知語が分かれば、未知語 X が示す時間はその既知語に対応する代表語とする。

この処理はそのままで、関連がそれほど深くなく関連度が小さい語であっても置換されてしまうため、閾値 Th_r を設け、最も高い関連度の語であってもその関連度が Th_r 以下であった場合は置換しないこととする。

この未知語処理手法を「最高関連度語置換処理」と呼ぶ。

3.4.3 概念の属性と重みを利用した未知語の処理

この処理は、最高関連度語置換処理とは別のアプローチによる未知語処理であり、未知語の属性をもとにして、未知語と対応する代表語を特定しようとするものである。

以下に処理手順を示す。ただし、概念 X の n 個目の1次属性を X_n 、 X_n の m 個目の1次属性(概念 X の2次属性にあたる)を X_{nm} のように記す。同様に、重みを w_n 、 w_{nm} と記す。

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、1次属性 $X_1 \cdots X_n$ がそれぞれ、代表語か、未知語という分類に置換される。

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 w_i の値だけその代表語の得票に加算する。これを1次属性すべてに繰り返す。

4) 得票が一番多かった代表語の得票割合が、閾値 Th_r 以上であった場合、未知語 X の代替語とする。同得票割合の代替語が複数ある場合は関連度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。

i) X_1 の1次属性 ($X_{11} \cdots X_{1n}$) 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換する。

ii) $X_{11} \cdots X_{1n}$ が代表語に置換されたら、それぞれ X_1 に対する重みを乗数として多数決をとり、得票が一番多かった代表語が、基準とする得票割合の下限 Th_b 以上を獲得している場合、 X_1 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。 $X_{11} \cdots X_{1n}$ がすべて未知語の場合は、 X_1 も未知語とする。

以下、概念の属性と重みを利用した未知語処理のこの手法を、「二次閾値付き多数決未知語処理」と呼ぶ。

3.4.4 二段階未知語処理

最高関連度語置換処理と二次閾値付き多数決未知語処理を実装し、予備調査を行ったところ、同じ語に対して違う結果となることや、閾値の変化に対する出力の数や精度の変化に違いがあることが明らかになった(表 3.9)。

表 3.9 二種類の未知語処理方式の性質の違い

処理方式	閾値の変化に対する変化の傾向
最高関連度語置換処理	正答率: ほぼ線形に減少 精度: ほぼ線形に増大
二次閾値付き多数決未知語処理	正答率: ゆるやかに減少 精度: 低い閾値においても比較的高い精度

ここで、回答の解釈を、正しい答である正答、誤った答である誤答、出力が無かった無答の三種とし、「正解率」と「精度」を以下で定義する。

正解率 = 正答数 / (正答数 + 誤答数 + 無答数)

精度 = 正答数 / (正答数 + 誤答数)

性質の違う方式を組み合わせることで、より良い結果を得られるのではないかと考え、その考えに基づく図 3.3 に示す二段階未知語処理方式を提案する。

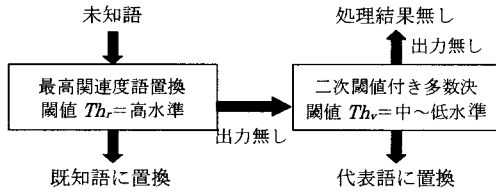


図 3.3 二段階未知語処理方式

一段階目では、閾値を高い水準に置いた最高関連度語置換処理によって既知語への置換を試みる。それで出力が得られない場合には、閾値を比較的低い水準にした二次閾値付き多数決未知語処理によって出力を試みる。

信頼性の高い処理を優先的に実行することで、それぞれの処理を単独で行うよりも高い精度・正解率を得られる。

3.5 時語生成

時を表す表現として「太陽が登る＝朝」や「葉が赤い＝秋」など、体言と用言の組み合わせがある。ここでは体言と用言の持つ意味に着目し、用言 KB によらずに処理する手法を述べる。何故ならば、新たに用言 KB を構築し、体言との組み合わせ処理を行ったのでは例外が頻発し、処理が煩雑になるからである。具体的には、入力中の漢字や概念ベースを活用し、直接時間を生成するのではなく、既知の時語へいったん帰着させることを目指す。例えば、「葉が赤い」から「秋」を生成するのではなく、既知語である「紅葉」を生成することができれば、KB を参照することにより「秋」であると理解できる。以下に時語生成の手順を述べるが、先に述べる処理から優先して行い、時語生成が成功した段階で処理を終える。

3.5.1 漢字組み合わせ

漢字の持つ意味に着目し、体言中の漢字と用言中の漢字を組み合わせることで時語を生成する方法である。「日が落ちる」を例として図 3.4 に流れを示す。生成される語は2通りであるが、時語 KB の絶対時語や暗示の時語に含まれている語を取り出す。

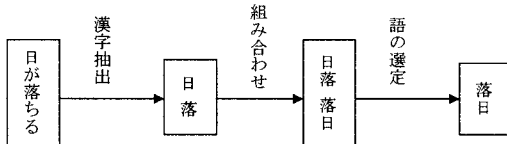


図 3.4 漢字組み合わせ処理の流れ

3.5.2 用言の代替による漢字組み合わせ

漢字の意味に着目した手法であるが、単純な漢字組み合わせが成功しない場合に用いる。概念ベースから用言の属性を取得し、属性中の用言、及び、サ変名詞を用いて体言との漢字組み合わせを行う。例えば、「日が沈む」という単文では、「日」と「沈」では漢字組み合わせ処理はうまくいかないが、用言「沈む」の属性中から「没する、入る、落ちる、…」などの語が取得できる。それによって漢字組み合わせを試みる。複数の候補があるが、重みの大きな属性を優先する。

3.5.3 属性検索

絶対時語と暗示の時語の中から、単文中の体言と用言が属性に含まれている語を検索する。例えば、単文「太陽が沈む」に対し、「太陽」と「沈む」を両方属性に含む時語「夕日」が見つかる。

なお、複数の候補があるが、単文中の語との関連度の平均が最も高いものを選択する。

3.5.4 体言のみによる判断

入力のうち、用言を無視し体言のみから時を推論する。体言が時語 KB に含まれる語ならば、直接時語 KB を参照することで時間が分かる。体言が未知語であった場合には、3.4 節で述べた未知語処理を行う。例えば、「柿を食べる」の「柿」のみから「秋」と判断する。

3.5.5 用言のみによる判断

用言に対して、3.4 節で述べた未知語処理を行う。例えば、「手がかじかむ」の「かじかむ」から「冬」と判断する。

4. 時刻・季節判断システムの分析

4.1 時刻・季節判断システムの評価

4.1.1 評価用データ

3 節で示したようなメカニズムを実装した本システムの評価のため、以下のデータを作成した。

- A群:新聞記事中の明示的時間表現(285 個)
- B群:時刻・季節を想起可能な単文(256 個)
- C群:時刻・季節を想起することが可能な語(289 個)
- D群:時刻・季節を想起しない、無関係な語(250 個)

A群は、ある全国紙の 1995 年の記事から、季節などの偏りを避けるため、ランダムに選択した 100 個の記事から、明示的時間表現を手手で抜き出したものである。B群は、アンケートによって収集した、時間を想起しうる「体言と用言の組み合わせ」であり、時語生成の評価に当たる。C群は、アンケートによる記述および、俳諧で用いられる時間を表す言葉である季語のうち日常的に使われない語などを削除したものを集めた。D群は、アンケートによって時間に関係ない語を収集したものである。

4.1.2 評価

それぞれのデータからの出力を「正答」「誤答」「無答」の三種類に分ける。A群では、時間を正しく把握できた場合を正答、残りを誤答とする。B群とC群では、そのデータから人間が想起するのと同じ代表語が得られた場合が正答、正答以外の代表語が得られた場合が誤答、出力が得られなかった場合が無答とする。D群では出力が無い場合が正答、何らかの代表語が出力された場合は誤答である。正解率、精度は 3.4.4 節での定義と同じとし、結果を表 4.1 に示す。

表 4.1 各語群における評価

群	正答数	誤答数	無答数	正解率	精度
A群	235	50	—	82.5%	—
B群	123	57	76	48.0%	68.3%
C群	218	23	48	75.4%	90.5%
D群	244	6	—	97.6%	—

4.1.3 考察

A～D 群すべてにおいて高い正解率、精度を得ることができた。入力を単語、単文に限定した場合、時刻・季節判断システムは的確な返答が可能となった。またB群では、「夏が終わる」など、用言が時間の推移を表す時間表現などで誤答が目立った。このような表現についての対応策は後述する。

4.2 複数の時語を含む文における評価

4.1 節で時刻・季節判断は単語、単文の入力に対して適切な返答が可能であることが示された。しかし、一般的な文章に出てくるような長い文や複雑な文の入力に対しては、的確な返答ができない場合が多いことが予備調査から明らかになった。そこで、複雑な文に対しての精度の調査を行い、時刻・季節判断システムの問題点を明らかにする。

4.2.1 評価用データ

時刻・季節判断システムにおいて複雑な文への対応を調査するため、評価用データを作成した。その際、時間における複雑な文の定義を「複数の時語を含む文」と定めた。「5月12日から24日まで」、「去年の暮れ」といった複数の時語を含む文を、新聞記事等を参考に100個を人手で抽出し、それを評価データとする。

4.2.2 評価

それぞれのデータからの出力を「正答」「誤答」に分ける。今回の調査ではすべての入力に対して何らかの出力が得られたので、「無答」は存在しない。よって精度のみを記載することとする。その結果を図4.1に示す。

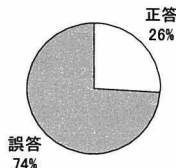


図 4.1 複数の時語を含む文における精度

図 4.1 で明らかになったように複数の時語を含む文における精度は 26% と大変低い値であった。正解例としては、「5月19日の朝」に対して「5月19日4時～9時」や、「今年の春」に対して「2007年3月～5月」などが挙げられる。

5. 文の意味理解に基づく時刻・季節判断システム

4.2 節から複数の時語を含む文における精度が低いことが明らかになった。つまり、現状の時刻・季節判断システムでは複雑な文に対して的確な返答が行われていない。その解決法として、失敗例をいくつかのカテゴリに分類し、そのカテゴリごとに対応策を検討していくこととする。また、単文における時間の推移表現についても検討を行う。

5.1 時語における優先順位

最初のカテゴリにおける失敗例として、「5時に星を見る」に対して「18時～3時」、「冬に雨が降る」に対して「6月～7月」となってしまったことが挙げられる。このカテゴリにおける問題を調査したところ、時刻・季節判断システムにおいて入力文が形態素解析された後、同じ時間軸上に複数の時語が存在する場合、後の名詞が優先されるという点の原因であると判明した。例を挙げて説明すると、「5時に星を見る」という文の場合、「5時」という時語から時間「5時」、「星を見る」という文から時間「夜(18時～3時)」が得られる。この2つの時語はどちらも「時」の時間軸上に存在しているため、後の名詞が優先されるという法則に従って、「夜(18時～3時)」が出力される。このように重要な情報が欠落してしまう点が問題点であると考えた。

5.1.1 重要度

このような問題点への対応として、すべての時語に異なる「重要度」を与える手法を考案した。時語の中でもより重要であると考えられる語には高い重要度を、あまり重要ではない語には低い重要度を付与した。重要度の付与方法を表したのが表 5.1 である。

表 5.1 重要度の付与方法

重要度	対応する時語
4	数字+時間単位
3	絶対時語
2	上記以外の明示の時語
1	暗示の時語

数字+時間単位とは「17時」「3月」などの語である。絶対時語、明示の時語、暗示の時語については 3.2 節で説明した通りである。

重要度の具体的な利用方法について説明する。入力文中において同じ時間軸上に複数の時語が存在した場合、その重要度を調べ比較し、重要度の値が高い方を結果として出力する。例として「5時に星を見る」という文の場合、「5時」「星を見る」という語からそれぞれ「5時」「夜(18時～3時)」という時間が得られる。どちらも「時」の時間軸上に存在する語であるので、重要度の考え方を利用する。「5時」は数字+時間単位であるので重要度は4、「星を見る」は暗示の時語であるので重要度は1である。よって重要度の値の大きい「5時」を結果として出力する。

重要度の利用により重要な情報の欠落を防ぐことが可能となった。

5.2 期間表現への対応

次のカテゴリの失敗例として、「今日の4時から9時まで」「11月10日から20日の昼」などが挙げられる。このカテゴリに共通する点は期間に関する表現であるということである。このような期間表現への対応手法について検討を行う。

期間に関する表現について調査を行ったところ、助詞「から」を含む表現を中心として多様な時間表現が存在していることが明らかになった。期間に関する表現として、

- ・今週から来週にかけて(2語間の期間)
- ・12月26日から1週間(基準からの期間)
- ・今日から3日後(基準からの時点)

などが挙げられる。このような表現の違いは「から」「まで」「以降」といった助詞や時語の組み合わせによって起きている。そこで助詞に注目した場合分けを行い、期間表現に対応することとした。

5.2.1 2語間の期間

形態素解析を行った際、「時語から時語まで」「時語～時語」などの文が含まれている場合は、2語間の期間表現と考えて対応する。

例として入力文「元旦の23時から翌日の4時まで」が与えられた際の、「時語から時語まで」に対応する手法の説明を行う。まず「から」の直前の時語「元旦の23時」を開始時点として時間フレームに格納する。時語 KB を利用して、時間フレームには月「1」、日「1」、時「23」が格納される。次に「まで」の直前の時語「翌日の4時」を終了時点として時間フレームに格納する。同様に時語 KB を利用して、時間フレームに月「1」、日「2」、時「4」が格納される。その処理の概要を図 5.1 に示す。

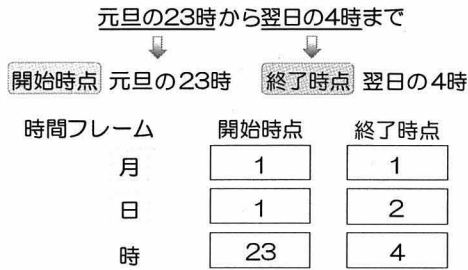


図 5.1 2語間の期間



図 5.3 基準からの時点

5.2.2 基準からの期間

形態素解析を行った際、「時語から期間を表す時語」といった文が含まれている場合は、基準からの期間表現と考えて対応する。期間を表す時語とは「日間」「週間」などの時語である。

例として入力文「今年の12月26日から1週間」が与えられた際の、基準からの期間に対応する手法の説明を行う。まず「から」の直前の時語「今年の12月26日」を開始時点として時間フレームに格納する。時語 KB を利用して、時間フレームには年「2007」、月「12」、日「26」が格納される。次に「から」に続く期間を表す時語「1週間」を開始時点に加えた時点を終了時点として時間フレームに格納する。すると時間フレームに年「2008」、月「1」、日「1」が格納される。その処理の概要を図 5.2 に示す。



図 5.2 基準からの期間

5.2.3 基準からの時点

形態素解析を行った際、「時語から時点を表す時語」などの文が含まれている場合は、基準からの時点の表現とと考えて対応する。時点を表す時語とは前後時語を含む語である。

例として入力文「今年の11月29日から3日後」が与えられた際の、基準からの時点に対応する手法の説明を行う。まず「から」の直前の時語「今年の11月29日」を基準時点として時間フレームに格納する。時語 KB を利用して、時間フレームには年「2007」、月「11」、日「29」が格納される。次に「から」に続く時点を表す時語「3日後」に注目する。「後」なので基準時点に3日を加えた時点を補正時点として時間フレームに格納する。すると時間フレームに年「2007」、月「12」、日「2」が格納される。その処理の概要を図 5.3 に示す。

5.3 範囲表現への対応

次のカテゴリの失敗例として、「今週前半」「今月中旬」が挙げられる。このカテゴリに共通している点は範囲に関する表現を含んでいることである。このような範囲表現への対応手法について検討を行う。

範囲表現に対応するために時間の範囲を示す表現をまとめて知識ベース化することとし、時語 KB に範囲テーブルを作成した。範囲テーブルの一部を表 5.2 に示す。

表 5.2 範囲テーブルの一部

ID	名称	ふりがな	始	終
1	前半	ぜんはん	1	50
2	全期	ぜんき	1	100
5	中間	ちゅうかん	45	55
11	前期	ぜんき	1	50

フィールド「ID」には管理用の通し番号、フィールド「名称」に時語、フィールド「ふりがな」に時語の振り仮名、フィールド「始」に範囲の開始ポイント、フィールド「終」に範囲の終了ポイントを格納している。「始」「終」は時間軸を100分率で区切った際の区間である。

入力された文中に範囲テーブル内の語が存在する場合に利用する。「今年の成人の日を含む週の前半」という文が入力された場合を例として説明を行う。時語 KB を利用して成人の日を含む週は2007年1月8日から1月14日であるとわかる。「前半」は範囲テーブルに存在する語であるため、範囲テーブルの参照を行う。すると「前半」の範囲「始 1、終 50」が得られる。週7日を100分率で区切り「前半」の範囲 1～50 を調べると、「月曜～木曜」が「前半」に含まれるという結果が得られる。よって1月8日～1月14日の中の月曜～木曜である「1月8日から1月11日」を結果として出力する。週の範囲指定の概要を図 5.4 に示す。

100分率で区切った値

月	火	水	木	金	土	日
0	14	28	42	57	71	85
0	14	28	42	57	71	85

図 5.4 時間軸上での範囲指定

また月や年の場合も同様に100分率による範囲指定方法を利用する。

5.4 時間の推移を表す動詞への対応

4.1 節での評価データ B 群(時刻・季節を想起することが可能な単文)を入力した際の失敗例として、「春が始まる」「夏が終わる」が挙げられる。この例における出力はそれぞれ「春(3月～5月)」「夏(6月～8月)」となっており、「始まる」「終わる」という動

詞の意味が考慮されていない。このように現在の時刻・季節判断システムは名詞の持つ意味のみで判断が行われている場合が多い。文の意味を考慮したシステム実現のため、時間の推移を表す動詞が入力された場合、出力に反映させる。

時間の推移を表す動詞への対応策として、5.3節で述べた範囲テーブルと特定の動詞を対応させる。例として、「始まる」に対応させる範囲として「初め(始 1～終 5)」、「終わる」に対応させる範囲として「末(始 95～終 100)」などが挙げられる。

「冬が終わる」という文を例として具体的な利用法について説明する。まず時語 KB を利用して「冬」という時語から「12月～2月」が得られる。次に「終わる」は範囲テーブル内の語「末(始 95～終 100)」に対応しているので区間 95～100 が得られる。12月～2月を100分率に区切って区間 95～100 を対応させると、2月のみ区間に含まれることがわかるので、結果として「2月」を出力することが可能になる。処理の一部を図 5.5 に示す。

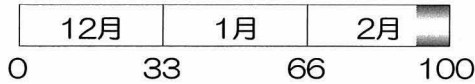


図 5.5 時間軸上での範囲指定

6. 評価と考察

5 節で述べた処理を用いて評価を行った。評価データは複数の時語を含む文(100 個)と時刻・季節を想起することが可能な単文(256 個)である。これは 4.1 節, 4.2 節で用いたデータと同じものである。評価結果を図 6.1, 図 6.2 に示す。

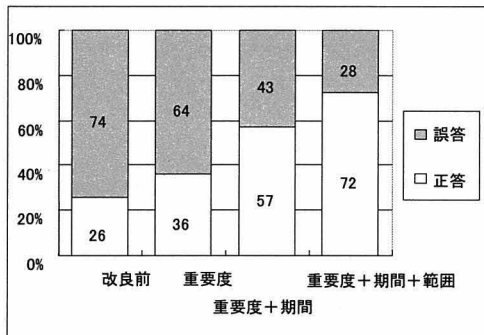


図 6.1 複数の時語を含む文における評価

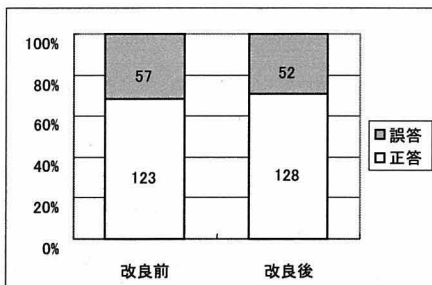


図 6.2 単文における評価

複数の時語を含む文においては、重要度の利用、期間表現への対応、範囲表現への対応を行うことですべての場合におい

て精度が向上していることが明らかになり、改良前 26%であった精度を 72%まで上げることに成功した。失敗例としては、「昨年9月頃」が「2006年9月11時～15時」と判断されたことが挙げられる。この例の場合、「頃」と「昼」の関連度が高いため失敗してしまった。このような関連度計算に起因する失敗に対応するためには、特別な処理が必要と思われる。また、「1995. 10. 1～1996. 3. 31」という入力文に対しては出力が得られなかった。このような単位が省略された表現は前後の文脈に依存している場合が多い。省略された表現の補完が必要である。

単文における評価では改良前後の無答の数と同じであったので、図 6.2 には正答と誤答のみを記した。その結果、時間の推移を表す動詞への対応を行ったことで精度は 68.3%から 71.1%となり、わずかであるが向上した。

以上より、時刻・季節判断システムに改良を加えたことで、より複雑な文に対応が可能となったといえる。よって今回、文の意味理解を行うために考案したアルゴリズムが時刻・季節判断システムの知的化に大いに貢献できたといえる。

7. おわりに

本稿では、人間のパートナーとして有用なコンピュータの実現に必要な知的会話メカニズムの基本要素として、時間に関する文の意味理解手法の提案を行った。

時刻・季節判断システムにおいて、時語における優先順位の決定、期間・範囲表現への対応を行うことで、より複雑な文への対応が可能となり、会話システム内の有用性を高めることに貢献した。また、動詞の意味に基づく処理を行うことで主語と述語の関係を出力に反映させることに成功した。

今後、時刻・季節判断システムにおいては、動詞の時制の変化(過去形、現在形、未来形)や文中の語句から、長い文章中の時間の推移を記憶するようなシステムが必要であると考えている。また今後、文中に時間的な矛盾が生じていないかを調べる時間表現の整合性についても考える必要がある。最終的には単文ではなく文章単位での判断を可能とし、知的会話メカニズムとの融合を目指す必要がある。

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「人間と生物の賢さの解明と、その応用」における研究の一環として行った。

参考文献

- [1]溝淵昭二, 住友徹, 泓田正雄, 青江順一: “日本語時間表現の一解釈法”, 情報処理学会論文誌 Vol.40, No.9, pp.3408-3419, 1999
- [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”, 電子情報通信学会技術研究報告, 人工知能と知識処理, Vol.100, No.530, pp.1-6, 2001
- [3]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