

## 事象解析による要約情報の抽出

稻垣 博人

NTT ヒューマンインターフェース研究所

あらまし 文章の要約は、人間と機械間や人間同志における情報伝達活動の重要なインターフェース技術である。ヒューマンインターフェースにおいては、人間により発せられた言葉を扱うことを想定するため、部分的情報を用いて処理しなければならない。ここでは、システムが用意すべき情報の部分性、処理の部分性を考慮しながら計算機上にインプリメント可能なインターフェース技術を提案する。入力された文章を事象という単位で捉え、文章の事象構造を解析するインターフェースである。そのインターフェース処理により、要約のテーマに関する要約情報を文章中からフィルタリングして抽出することが可能となる。

Abstract Information Extraction based on Event Analysis

INAGAKI, Hirohito

NTT Human Interface Laboratories

1-2356 Take Yokosuka-Shi, Kanagawa 238-03, Japan

**Abstract** Extracting abstract is one of the important interface application for man-machine or man-man communications. In the man-machine or man-man communication, it is necessary to use natural language created by real person. Therefore, the system should take into account of partiality of information and partiality of processing in natural language. To avoid such difficult AI problems, we introduce a framework of focusing of information and information closure and event analysis. Using these frameworks, we can extract information about the theme of abstraction from a document.

## 1 はじめに

文章の要約は、ヒューマンインターフェースに欠くことのできない言語処理技術である。特に、人間と機械の対話における要約処理だけでなく、人間と人間に創造された人工物（本、新聞、TV番組など）や人間同志のインターフェースにおける情報取得支援などにおいて要約処理は重要である。インターフェース部では人間により発せられた情報を扱わなければならないため、部分的情報<sup>1</sup>だけで処理を進めなければならない。また、与えられる情報が部分的であるため、処理も完全には行えない。この情報の部分性と処理の部分性がヒューマンインターフェースにおける特徴である。つまり、インターフェース部では、情報の部分性と処理の部分性に対応できるような機構が必要とされてくる。

情報の部分性としては、例えば、ある発話をどのような世界（現実世界、心象世界、夢想世界などの意味の世界）について述べているのか、また、その世界で実体を持つ個体や個体間の関係、時空がどのようなことを指しているのかが部分的にしか明記されないこと（入力された情報の部分性）や入力された情報を処理するに必要な情報が与えられないこと（処理のために必要な情報の部分性）を意味する。また、処理の部分性としては、ある発話を、まず表現として捉えられ、次に意味を形成し、最終的には知識として知的の個体中に同化するモデルを考えると、表現、意味、知識の3段階の処理を全ての言葉に対して行えないことを意味する。

本稿では、言語をメディアとしたヒューマンインターフェースにおける情報の部分性や処理の部分性について検討し、そのインターフェース処理の実現方法について述べる。そのインターフェース処理を用いた、ある特定の分野に関する要約情報を文章中からフィルタリングして抽出するアプリケーションの実現方法について提案する。このインターフェース部では、情報の部分性に対応するために、視野や情報の囲い込みを導入し、さらに事象という単位で文章を捉えることを提案する。

## 2 情報の部分性と処理の部分性

文章の内容を理解するためには、種々の知識が必要となる。しかし、常識、専門知識などの言葉に対する知識は完全に与えられているわけではない。

<sup>1</sup>情報とは、入力される情報と処理のために必要な情報の両方をいうが、ここでは、入力された情報の部分性については検討しない。

く、与えることも不可能に近い。そのため、常識や専門知識などの知識が完全になくても処理が可能となるようにするために、ある処理を行うために最も適した部分的情報を与え、その情報だけで処理できるような枠組を考える必要がある。そこで、ある処理に必要な情報を与えた場合の情報の部分性、処理の部分性について考える。

### 2.1 情報の部分性

処理に必要な情報を部分的に持つ場合、以下のような情報の区切り方を考える。

- ある領域に関する知識のみを持つ。
- ある領域に関する全ての知識は持たない。

ある領域に関する知識のみを持つとは、ある入力情報が与えられた場合、ある領域に関する知識を持って処理を行うことを意味し、他の領域でたとえ解があったとしてもそれは無視する。（ここでは、ある領域に関する知識を持つことを知識に對してある視野を持つと呼ぶ。）例えば、図1の例を考えてみる。”掛金”という文字を見た時、2種類の解釈がされる。主に使われるのは、預金の掛け金（かけきん）であろう。一方、窓の掛け金（かけがね）という解釈も実はできる。人間の場合、文脈等から適切な視野を設定し、適切な解釈を行うが、ここでは、視野は固定する。そのため、図1のように、銀行に視野を絞り込んだ場合、預金の掛け金（かけきん）の方にしか解釈せず、窓の掛け金（かけがね）の事については、全く解釈しないことになる。

しかし、視野を導入しても、図1を見ればわかるように、その知識は膨大である。その意味である領域に関する全ての知識は持たないという制約が必要となるのである。知識の内容や形式は処理系に依存する。

現実世界の知的個体および個体の集合による活動（ここでは、事象と呼ぶ）を抽出しようとする場合、まず、事象における基本的な構成要素と事象における役割の明確化が必要となる。その処理を行うためには、事象の構成要素と役割を明確化するための知識が必要となる。事象を表すのに必要な構成要素としては、知的個体および個体の集合、その個体がある活動を実現する時使用する実体、活動、時間、空間が考えられる。つまり、以下の制約が加えられる。

- 現実世界における事象の構成要素は、個体、実体、活動、時間、空間である。

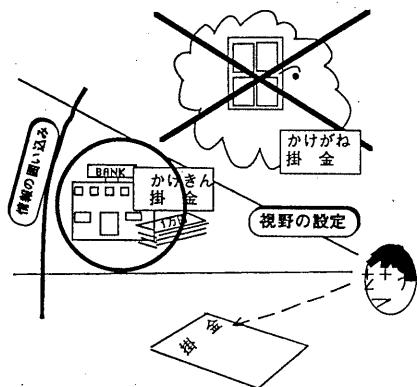


図 1: 視野と情報の囲い込みの導入

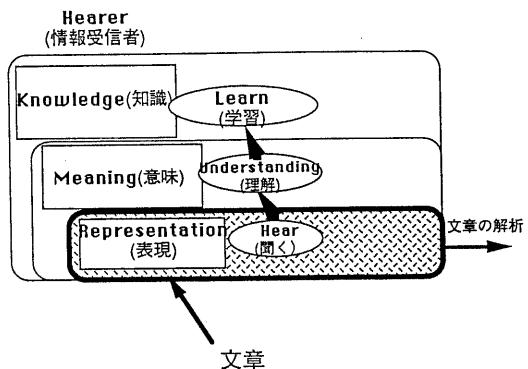


図 2: 言語における処理のレベル

## 2.2 処理の部分性

情報が部分的である場合、それに対する処理も部分的にならざるを得ない。ただ、人間と異なり、処理もある一定のレベルを想定すればよい。例えば、言葉が表現、意味、知識の 3 階層で処理されていると仮定すると、それぞれの階層に適した情報が与えられ場合、適した階層までの処理まで行うことができる事になる。そのため、処理の部分性に対する制約としては、

- 言語の特性を考慮して、表現、意味、知識の各処理レベルで適切な処理を行う。

ことが考えられる。

このような制約事項がヒューマンインターフェースにおける特徴であるが、計算機上で実現する場合、言語の処理レベルは、表現レベルでの処理が現実的である。(図 2) つまり、文章で表現している意味は捉えず、表層表現から文章の解析をすることである。それに必要な知識は、文章の解析の知識ということになる。ここで、現実世界を描写した文章が事象の集合から成り立っているとすると、文章を事象の構造に変換解析するということは、文章を事象および事象間の関係で表すことである<sup>2</sup>。この文章の事象解析に必要な知識としては、その語が記述世界においてどのような構成要素となっているかを示す素性 (feature) と feature が事象においてどのような役割 (基本格; substantial case) を持つかが必要となる。つまり、表層言語は feature にリンクし、リンクした feature と事象に関する基本格により表層的に表

現された文章は解析されることになる。feature は、全ての語に対して付与されているわけではなく、ある個体の活動を表現するのに必要な基本的構成要素に対してだけ記述される。ただし、feature の基本要素といっても、新たに創造される場合があるので、全ての単語に対して、feature を付与することはできないが、default でどの feature と推論するかは決定する必要がある。

このような、視野や情報の囲い込みの導入を用いた場合の利点としては、以下の点がある。

- 暗昧性の解消
- 辞書の軽量化
- 計算量の減少

## 2.3 情報、処理の部分性を考慮した事象の枠組

情報の部分性や処理の部分性を考慮した場合、事象とは、表層の言語表現で表されている現実世界で起こった変化である。つまり、表層表現で表された 1 変化とその変化を構成する要素をひとまとめにしたものを事象とする。たとえば、”太郎が本を買った。”や”次郎は昨日 9 時に寝た。”という文で表された現実世界は、ある事象がその世界に対してなんらかの変化を起こしていることが明確である。”太郎が本を買った。”という文では、太郎は、本を持っていないという状態 (シーン) から本を持っている状態 (シーン) への変化が起こったと考えられる。つまり、変化の集合で現実世界を記述するのが事象の枠組である。

<sup>2</sup> 但し、入力された情報の部分性により、全ての事象内の関係、事象間の関係が明らかになるわけではない。

この事象の解析に必要な feature 基本要素は以下のものである。

feature 基本要素	feature の持つ意味
individual	知的個体および知的個体の集合
element	活動する時に使用される実体
thing	活動により変化する個体
action	活動
time	時間
location	場所

これらの feature 基本要素は、処理すべき対象の動作主が決定されることにより、その feature の実体名が決定される。また、事象に対して付加される様相表現は、通常の feature とは別に modality とした。一方、事象の役割を記述する基本格として以下の格を考える。

substantial case(SC)	基本格の持つ意味
agent	事象における動作主
object	動作主の活動の対象
action	動作主の活動
from	基点
to	着点
time	時間
location	場所

つまり、活動を起こす動作主 (agent)、その動作主がある活動を起こす場合、活動を起こした場所 (location) と時間 (time) があり、その活動がどの状況において発生し、他の事象とどのような関係を持つかが伝達される。この 7 つの基本格以外の役割については、これら的基本格に対する制約された役割として考える。例えば、tool(道具格) は、事象における object の制約された役割を持つと考え、role(役割格) は、agent の制約された役割を持つと考える。そのため、今まで、格要素として考えられてきたものは、基本格と基本格の制約で表現されることが期待される。

### 3 情報、処理の部分性を考慮した文章解析

ここでは、事象に基づく枠組で表層言語表現を事象表現に変換解析する方法について述べる。事象解析の実現方法は種々考えられるので、ここでは、比較的簡単な実現例を説明する。

#### 3.1 形態情報と構文情報

事象解析に必要とされる形態情報と構文情報を抽出するために、まず、表層言語表現の最少単位

である形態素を抽出し、その構文的情報について解析する。

形態素の抽出手法としては、単語数最少法、文節数最少法、最小コスト法などがあるが事象解析においては、全解の適宜探索ができる処理方式がよい。

解析に必要な単語には、feature が付与される。図 3 に feature 付与例を示す。各単語に feature を付与するための辞書を feature 辞書と呼ぶ。

辞書項目名	feature
株式会社	*element*(company)
NTT	*individual*(company)
会社	*element*(company)
工場	*element*(company)
工場	*location*(company)
麻生	*location*()
所沢	*location*()
松下	*man*()
松下	*individual*(company)
増設	*action*(company)
新設	*action*(company)
建設	*action*(company)
1 億円	*money*()
資本金	*money*(company)
開始	*modality_開始*()
停止	*modality_終了*()
終了	*modality_終了*()
予定	*modality_予定*()
...	...

図 3: 単語の feature 付与例

但し、man および money は feature 基本要素の thing にあたる。

feature の記述形式は、\*<feature 基本要素名>\*<領域> で、領域に何も記述されていない場合、その単語はいかなる領域においてもその feature を持つことを意味する。領域としては、企業 (company)、VIP 等の動作主が記述される。

action という feature が付与された単語は、さらに事象の基本格を決定するための選択制約が記述される。(この選択制約を記述した辞書を SCD 辞書と呼ぶ。) 図 4 に "建設" の記述例を示す。まず、action 名とその action の事象名が記述されている。"建設" は、事象名が "建築"

ACTION: 建設　事象:”建設”	
	concrete_action
	agent: *individual*(company)
	object: *element*(company)
	time: *time*()
	location: *location*()
	abstract_action
	in: *money*()
	out: *element*(company)

図 4: SCD 辞書記述例

である。辞書記述で示される concrete\_action と abstract\_action は、事象における直接的な基本格を記述しているのが concrete\_action の箇所であり、外界に対しての抽象的格を abstract\_action で記述する<sup>3</sup>。 concrete\_action、abstract\_action 両方とも、選択制約は feature によって記述される。

形態素の抽出後、形態素情報を用いてそれぞれの形態素の構文を決定する。ここでは、係り受け解析を用いて構文を決定する。

係り受けは、係り受けの非交叉と格の非重複の 2 つの原則がある。この原則に基づいて文法的属性で係りと受けを決定する。係り受け候補の尤度は、意味連結パターン [2] の手法を用いて決定する。意味連結パターンは、文章内で実際に使用された係り受けのパターンを記憶し、そのパターンの出現頻度から係り受け候補の尤度を決定する手法である。

名詞句の並列では、並列構文のパターンに着目した処理により並列候補を選出する。並列構文では、係り受けの非交叉と同様な、並列構文の非交叉がある。つまり、並列構文は必ずネスト構造になる。動詞句の並列パターンについては、並列を構成している動詞句の属性(自動詞、他動詞の区別)により並列が区別される。

並列構文のパターンにより並列候補が決定されたのち、候補間の尤度は、その並列構文を構成する並列要素の意味的類似度<sup>4</sup>や文章特有の並列構文

<sup>3</sup> concrete\_action の扱いは、fillmore[1] が提唱した格の概念に近いように思われるが、構文における特定の品詞に対して付与されるのではなく、事象のアクトに対して適用されるものであるという立場をとる。一方、abstract\_action は、事象を事物の変化であると捉え、抽象化した場合の入力と出力について記述する。抽象化した入力は要素表現では IN で表され、出力は OUT で表現される。

<sup>4</sup> ここでは、意味カテゴリ番号という、単語の類似性をト

バターンに応じて決定される。

### 3.2 構文の変換

事象解析では、形態素情報との構文情報から、文章を事象に変換解析する。事象解析に適した構造を生成しないため、係り受け解析の結果を変換する必要性がある。最も問題となる点は、事象の action の考え方と文法属性の違いである。例えば、サ変名詞として扱われている単語が action であったり、通常の文の中で、動詞表現されているが、実は action ではなく、単なる modality であったりする場合である。

構文構造から事象構造への変換では、係り受け関係から、事象の定義に照らし合わせて構造を変換する。基本的に考慮しなければならない構造変換としては、動詞 → 事象における modality への構造変換である。非動詞表現 → 事象の action への構造変換もあるが、構文構造の変更の必要はなく、解析時に、非動詞表現を action と認識して処理を進めるだけである。基本的には、どちらの構造変換も事象の action に対して構造を適正化することである。構造変換の例を図 5 に示す。

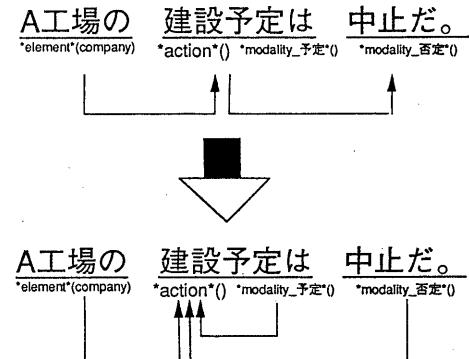


図 5: 構文構造の変換例

別の問題として、文節の問題がある。事象解析では、文節と事象の要素が 1 対 1 に対応しないことがある。そのため、通常は、文節単位で処理するが、同一文節内に複数のユニット(事象での最小単位)がある場合、文節内の係り受けも考える。

リー状に配置し、そのトリー構造を数値であらわした番号を使用する。

### 3.3 事象の解析

SCD 辞書の中には、事象に必要な基本格に対する選択制約が記述されており、その制約を満足する候補を抽出するのが事象の解析である。基本格としては、action(動作)、agent (動作主)、object (対象)、など7個が設定されている。事象解析では、これらの SC とその事象に対する入出力が明確化されることが期待される。具体的には、語、句、等のユニットに feature 辞書により feature が付与されており、その feature が選択制約を満足するかどうかを判断する。どのユニットに対して選択制約を試みるかであるが、これは構文情報から決定されることになる。つまり、構文情報の中の係り受け情報と文節情報から判定する。係り受け関係<sup>5</sup>がある場合、そのユニット間での制約を検査することが可能であるとする。また、係り受け関係だけでなく、文節内が複数のユニットで分割される場合、同一文節内のユニット間の検査は可能とする。

1 個の候補の選択制約が満足できない場合、他の係り受け候補を検査し、次に形態素解析候補を検査する。全ての候補を検査しても適当な候補がなければ、もっとも制約を満足する候補を抽出する。

また、係り受け関係があっても、事象における役割が決定されないものもある。これは、事象内の解析が事象における基本格を中心に解析するためである。これらの候補については feature も付与されず、どのような要素表現を持つかは明記されない。

modality も、この SCD 辞書を用いた解析時に処理される。modality は任意であるため、action に対して係り受け関係にある modality は全てそのアクトに関係する modality として処理される。新聞記事のような、著者が直接の動作主ではない文章では、図 6 のような modality を考慮する。

### 3.4 スクリプト KB 等を用いた事象間解析

事象間の解析では、

- 同一の事象関係
- 連続的な事象の関係

などの関係を見つける。事象間の関係は、スクリプト KB、事象の基本格のスロットにより解析さ

<sup>5</sup> 実際、係り受け関係では、係りと受けの依存関係があるが、ここでは、その依存関係は全く考慮しない。

テンス	
過去	た る
現在	る
未来	る
アスペクト	
開始	始める、開始
継続	ておる、続ける、ているところ
終了	終わる、始める、終了
...	...
伝達される情報のモーダル	
否定	ない、ず、は不可能
可能	れる、られる、ことができる
意思	う、つもりだ、うとしている
計画	計画、方針、考え、図る
...	...
発話のモーダル	
義務 (+/-fact)	なければならない
必然 (+fact)	必要がある、ことになる
良好 (+/-fact)	ほうがよい
断定 (+fact)	のだ、のである、はずだ
...	...

図 6: 事象における modality

れる。

スクリプト KB による解析とは、事象の必須的な関係について着目し、その間の関係をスクリプトを用いて解析する処理である。例えば、時間について着目すると、時間は不可逆であり、同じ事象が起こりえず、また、ある事象が生起したら、必ずその事象に関係することが起こるという性質を考慮する。

時系列的な現象を扱う最もプリミティブなスクリプトは、開始 → 終了、開始 → 停止 → 終了などのようなアスペクトで表されるスクリプトである。さらに、抽象的な上位の時系列的スクリプトとしては、Schank のスクリプトに近い事象の時系列的スクリプトも考えられる。

スクリプトでは、先に説明したようなシステム定義のスクリプトとユーザ定義のスクリプトを考える。ユーザ定義のスクリプトは、事象解析をユーザが希望する処理に変更することができる。たとえば、ある企業が新製品の開発を企画または、開発した段階で、それに対して銀行が融資を開始しているのかを調べることを想定する。このとき、事象「開発」と事象「融資」が時系列的

に連続性をもっているかどうかを判定するだけである。つまり、着目する企業が”A社”であるとすると、事象 A(action=“開発”、agent=“A社”、time = t1) かつ事象 B(action=“融資”、agent = one;<sup>3</sup> oneλ(銀行(one))、time = t2) かつ  $t1 \leq t2$  のような条件を満足するような事象を探索する。

同一の事象関係は、2つの事象を構成する基本格のスロットを比較することにより判断される。同一の基本格に対しては、全く同じ語が対応しなければならない。ただし、企業名の略称等、表層言語上同じ文字ではないが、同じものを指している場合、この限りではない。また、2つの事象においてどちらか一方しか持たない基本格については、比較の対象とはしない。

スクリプト KB 以外に事象間の関係を求める方法として、事象の基本格のスロットの変化に着目した処理が行われる。

#### 4 事象解析を用いた要約情報の抽出

要約処理には、文章の理解、重要概念の判定、重要概念の表層表現変換など、非常に複雑な言語処理技術の組み合せが用いられる。しかし、ここで述べた事象解析は、文章の解析に対する枠組みであるため、重要概念の判定、文生成などの技術については、処理する機構は持ち合わせていない。その意味では、要約処理には不十分であるが、要約するための情報を抽出するための一処理方法として使用できる。つまり、事象解析を要約のテーマに関する情報の取捨選択に利用することである。この情報の取捨選択により、擬似的な重要概念の判定を実現する。例えば、企業の活動について事象を解析した場合、文章中の企業活動に関係のない記述については無視され、企業活動に関係のある記述だけが抽出されることになる。

要約の出力形式としては、一般的な文章で出力する要約[3]、ビジュアルな要約[4]や COGITO のような表形式の要約[5; 6]などがある。ここでは、解析した事象関係を統合して文章として表示し、さらに、要約を文章の様相表現によって分類した表示を行う。

要約情報の抽出処理の概要を図 7 に示す。

##### 4.1 要約情報抽出の初期設定

初期設定としては、対象の設定と要約テーマの決定が必要となる。対象の設定とは、対象とする文章の特徴を記述したスタイルファイルの設定で

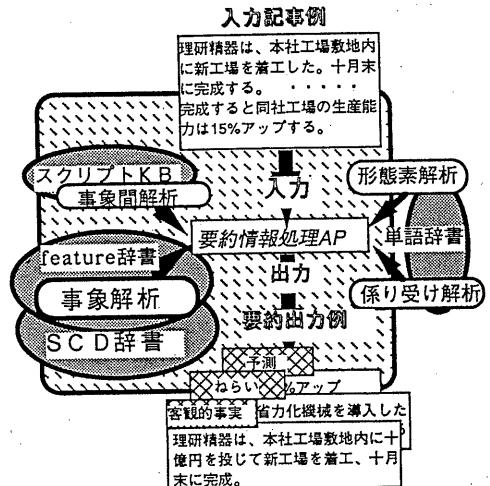


図 7: 要約情報抽出処理の概要

ある。対象が変更された時点で、スタイルファイルを変更する。このスタイルファイルには、対象文章特有の文章構造形式、対象文章特有の書法などを記述する。

要約のテーマの設定は、抽出したい情報に関する動作主のみを決定する。例えば動作主として、人間、動物、組織、機関等。

次に辞書の設定であるが、feature 辞書、SCD 辞書、スクリプト KB の中で必要な項目のみ使用する。辞書の構成については、先に説明したように、使用される領域が明示され、テーマに応じて適する項目を抽出することができる。例えば、要約のテーマが”企業”であれば、企業の活動状況を解析するのに必要な辞書情報を辞書中から抽出し、使用する。

##### 4.2 文章の事象単位分割

大雑把に入力文章の情報を取捨選択し、さらに、表層言語表現を事象単位に分割する。つまり、要約テーマにより関係する情報のみを抽出し、表層言語表現を1事象単位に分割する。ここでは、要約のテーマに関係する action が記述されているかどうかで要約に関係するかどうかを判定する。

### 4.3 文章の解析

要約のテーマに関して抽出した文を文章構造解析、事象解析する。文章の構造は、形態素解析、係り受け解析により取得する。

事象解析においては、SCD 辞書と feature 辞書を用いて事象内の解析を行い、スクリプト KB 等によって事象間の関係を解析する。事象解析では、要約のテーマに関係することのみが解析される。

### 4.4 要約情報の表示

要約情報は、事象群中に示された様相表現(図 6)に応じて分類し、表示する。つまり、個々の事象に付与された modality 情報に応じて、事象を分類し、表示する。modality の中で、テンス、アスペクトについては分類せず、伝達される情報のモーダルと話者のモーダルとで分類する。次に、モーダルに応じて分類した結果を生成系に渡す。生成系では、関係のある事象を複文にして生成する機能を有する。独立した事象の場合、入力記事と同様な機能語が付加されて、文末だけ終止形または、体言止めで表示される。事象間の関係が示されている場合、生成系では、それを複文として表示する。この場合、原文が平叙文で表現されていれば、その原文の表現をそのまま使用し、文末は終止形とする。複文の接続には、生成用テンプレートから事象間の関係に応じた接続表現を抽出し、それを用いて文を生成する。もし、原文が平叙文であらわされていない場合、生成用テンプレートの機能語表現を使用して、文を生成する。

生成用テンプレート		
要素表現	機能語表現	接続表現
agent	”は”	—
object	”を”	—
to	”から”	”状態へ”
from	”へ”	”状態から”
time	”に”	—
location	”で”	—
through	”を通って”	”て”
role	”という役割で”	—
...	...	

### 5 まとめ

本稿では、文章の要約をインターフェース技術として捉え、そのインターフェースで問題となる情報の部分性や処理の部分性について検討した。情報

や処理の部分性に対処するため、視野や情報の囲い込みを導入し、さらに伝達される情報を事象単位で捉えることにより文章の事象解析を行う手法を提案した。また、この事象解析を用いた要約情報の抽出法を提案した。これは、要約のテーマを事象解析の領域に一致させ、要約に関係する情報だけを抽出、表示する処理技術である。

今後は、この枠組を実際に計算機上にインプリメントして、評価を行う。

### 参考文献

- [1] Charles John Fillmore. Toward a modern theory of case. *Prentice-Hall*, pp.361-375, 1969.
- [2] 稲垣博人, 壁谷喜義, 小橋史彦. 意味連結パターンを用いた係り受け解析. 情報処理学会自然言語研究会, NL67-5, 1988.
- [3] 内海功朗, 重永実. 英語文章の大意生成. 情報処理学会自然言語研究会, NL54-8, 1986.
- [4] 田淵篤, 原良憲, 笠原裕. 文章のビジュアルな要約についての一考察. 情報処理学会第 40 回全国大会, 5F-9, 1990.
- [5] 北研二, 小松英二, 安原宏. 要約支援システム COGITO. 情報処理学会自然言語研究会, NL58-7, 1986.
- [6] 小松英二, 加藤安彦, 安原宏, 椎野努. 要約支援システム COGITO - 文章の構造解析 -. 情報処理学会自然言語研究会, NL64-11, 1987.