自 然 言 語 処 理 152－1
（２００２．１１．１２）

文章の構造解析による新聞記事からの事件情報抽出

金山 淳一† 北條 孝† 田村 直良††

†横浜国立大学大学院 環境情報学府 情報メディア環境学専攻
††横浜国立大学大学院 環境情報研究院
{junichi,hojo,tam}@tamlab.eis.ynu.ac.jp

本論文では、一連の出来事において関連する人間の相互関係として意味構造を定義し、特に新聞の事件記事から意味構造（犯罪スキーマ）を抽出する手法を述べる。事件スキーマの要素は、関連人物の容疑者、被害者、警察としての同定、それぞれのプロフィール、犯罪の動機、事件の進行などからなり、新聞記事から抽出される。解析、抽出処理は、スキーマの要素に応じて、パターンマッチング的な手法、構文解析、格フレーム抽出に基づく手法、主題の構造解析に基づく手法、時間セグメント分割に基づく手法などにより、犯罪スキーマとして再構成する。

Crime Information Extraction from Newspaper based on Text Structure Analysis

Junichi Kanayama† Takashi Hojo † Naoyoshi Tamura††

†Department of Information Media & Environment Sciences
Graduate School of Environment & Information Sciences
Yokohama National University

†† Graduate School of Environment & Information Sciences
Yokohama National University

{junichi,hojo,tam}@tamlab.eis.ynu.ac.jp

In this paper, we define a semantic structure as mutual relations among persons who relate a crime and we present a method to extract the semantic structure, especially from crime articles of newspaper.We call hte structure as crime scheme The scheme consists of descriptions of persons who relate the crime, identification of the persons with one of suspect, victim or police, the motivation of the suspect of the crime and theevent sequences occurred in the crime. The analysis and the extraction are based on the pattern matching, syntax analysis, case frame extraction, thematic structure extraction and so on, and are reconstructed as a scheme, according to the element of the scheme.

1　はじめに

本論文では、一連の出来事において関連する人間の相互関係として意味構造を定義し、特に新聞の事件記事から意味構造（事件スキーマ）を抽出する手法を述べる。

昨今は大量の電子化されたドキュメントが日々増え続けている。そのような中で、それらのドキュメントを人が逐一読んで処理するのは困難となってきている。そこで、必要な情報をすばやく効率よく手にいれるために、それらのドキュメントに対する自動要約や情報抽出などの自然言語処理への要求が高まってきている。このような現状において、パターン駆動による表層処理的な自然言語処理技術は、実装が容易なこととそれでいてある程度実用的な結果を得られることから、意味解析、文脈解析による「深い解析」が「王道」とは思われつつも、多くのシステムで採用されている。

意味解析、文章解析とは、割りきってしまえば、文章を構成する文字の一次元的な配列を使用目的に応じて、定義された構造へ変換することである。抽出しようとする情報は、使用目的に応じてその「意味」の形式が変わりうる。

そこで、我々は、ある程度実用規模での文書理解、情報抽出を前提とし、文章要約や二次利用可能な情報蓄積を利用目的と想定し、意味構造を検討する。実際には、「犯罪」、「事件」ついて書かれた新聞記事 (事件記事) を対象とし、「犯罪」、「事件」の意味を表現する「犯罪スキーマ」を提案する。

文章の意味的処理に関する研究としては、以下のようなものがある。

福本、安原 [4] らは、新聞社説記事に対し、接続表現、文末表現、主題から連接する 2 文間の関係の解析から、文をグループ化し、グループ間の関係を解析することにより、文章全体の構造化を行い有用性を示している。しかし、この研究では、対象が社説記事などの論説文であるため、新聞記事の構造化で有効と考えられる時間的関係、主題の省略についての解析は行われていない。

川端、原田 [5] らは、日本語要求仕様文や判例文に対し、接続表現、 EDR 電子辞書を用い連接する 2 文間の理由、条件、時間的近傍などの文間深層格を記述するシステム InSeRa を構築しその有用性を示している。しかし、連接文間に対する考察しか行っていないため、時間的関係や語彙的連鎖など文章全体に関わる意味的関係をつかむという意味では不十分であると考えられる。

本研究では、事件記事に対し、既存の文章内に内在する意味関係の解析 (時間セグメント、主題構造解析、語彙連鎖構造解析、文末構造解析）を行い文章を構造化し汎用的内部表現を得る。そして、得られた汎用的内部表現から犯罪スキーマの抽出を行う。犯罪スキーマでは、主に表層文とのパターンマッチング、表層的フレームの解析を行うことにより、記事から、犯罪に関わる動機、犯罪のタイプ、人物、その役割、そのプロフィール、その行動などを抽出する。ある種の情報の抽出には、深い解析を行うよりも、むしろ、表層的な手がかり表現やパターンマッチングを用いることにより、容易に行うことができるものもある。一方、深い解析が必要な情報抽出もある。我々の解析システムでは、抽出する要素の性質に応じて、両者を組み合わせている。

2　犯罪スキーマ

事件記事は、犯人、警察、被害者など関連する人物、事項の相互関係が時間的進行で書かれる場合が多い。そこで、我々は、事件記事をこれらの視点でとらえる構造として犯罪スキーマを提案する。

2.1　犯罪スキーマの定義

すべての事件記事は、罪状、動機、供述、人物の 4 つの要素で表現できると仮定する。そこで、我々は事件記事をこれらの要素をもつ犯罪スキーマとして定義する。

以下では、犯罪スキーマ中の各要素について述べる。

- 罪状スロット: 記事中で犯人が問われている罪状を値として持つ。
- 動機スロット: 犯人が犯行に至る理由を値として持つ。
- 供述スロット: 犯人の取り調べ中に述べている言動を値として持つ。
- 人物スロット: 記事中での役割を示すロール、経歴であるプロフィール、その人物のとった行動を示す行動という要素を持つサブスキーマで表現される。

人物スロットの持つ各要素について述べる。

- ロールスロット: 犯人、被害者、警察のいずれかを値として持つ。
- プロフィールスロット: 人物の名前、年齢、職業、住所という要素を持つサブスキーマで表現される。
- 行動スロット: 各人物のとった行動を示す格フレームの時間順の列を値として持つ。

プロフィールスロットの持つ各要素について述べる。

- 名前スロット: その人物の名前 (警察の場合は、警察の名称) を値として持つ。
- 年齢スロット: その人物の年齢を値として持つ。
- 職業スロット: その人物の職業を値として持つ。
- 住所スロット: その人物の住所を値として持つ。

3 事件構造解析システムのアーキテクチャ

本システムは、文章の意味解析部と犯罪スキーマの抽出部に分かれる。意味解析部では、どのような新聞記事からも抽出が可能な一般的な意味構造を抽出する。犯罪スキーマ抽出部では、対象を事件記事と限定することにより、より文章の内容に即した構造化を行う。図 1に本システムのアーキテクチャを示す。

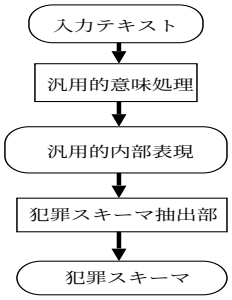

図 1: 事件構造解析のアーキテクチャ

4 文章の汎用的意味処理

汎用的意味処理は、入力テキストに対しまず構文解析を行い、その結果に対し、複文の関係解析を行う。各意味構造抽出部で意味構造を抽出し、それらの結果を統合した汎用的内部表現を出力する (図 2)。形態素解析には日本語形態素解析ツール JUMAN[7]、構文解析には日本語構文解析ツール KNP[6] を用い、表 1の関係により、必要なら複文を分割している。

以下では、各構造解析部について説明する。

4.1 複文の関係解析

本システムでは、以後の解析の複雑化を避けるために複文に対して関係 (主節、従属節) の解析を行う。文は大きく分けると単文と複文の二つに分けることができる。

- 単文: 単一の述語を中心に組み立てられる文。
- 複文: 述語を中心としたまとまり (節) が 2 つ以上集まって構成された文。

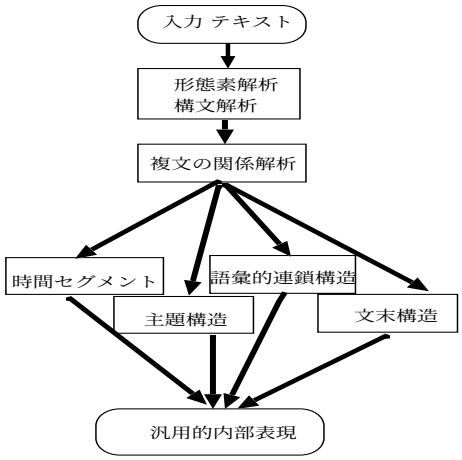

図 2: 意味構造の解析

複文の関係解析は、以下の二つの事を行う。

- 接続形式による複文の分割点の判断をする。
- 分割された従属節に対して表 1の従属節の分類をする。

主節	
従属節	時を表す連用節
	原因・理由を表す連用節
	付帯状況・様態を表す連用節
	逆接を表す連用節
	目的を表す連用節
	前提となる事実・動作を表す連用節
	事態を対比的に述べる連用節
	並列節
	従属節候補が複数存在する場合

表 1: 従属節の分類

4.2 時間セグメント

本システムでは、時間セグメントを文章中の時間的な境界から次の時間的な境界の間に出現している文の集合であると定義している。そのため、同じセグメントに属する文は何らかの時間的な関連性を持って起きた事象について記述されている。

ここで、時間情報は何らかの時点を明示的に示している。そのため、時間情報がある文の直前で時間の連続性が途切れている。よって、時間セグメントに時間情報は多くとも一つしか存在しない。

4.3 主題構造解析

本システムでは、一文にはその文において中心の話題となる語句 (主題) が存在すると仮定している。

4.3.1 主題の抽出

主題構造解析をするために、トピックと主題の抽出を行う。トピックと主題の定義を以下に示す。

- トピック：本研究では、新聞記事の見出しに出現する**名詞句**をすべてトピックと定義する。
- **主題**と**題述** [2]：各文は、主題構造を持つと仮定し、各文は主題と題述とから構成されているとする。具体的には、**は格**、もしくは初出現の**が格**を主題と定義し、文の主題以外の残りの**名詞句** を、題述と定義する。

主題となりうる語句が複数存在する場合は、一文中でより先頭に近い語句を主題として抽出する。

4.3.2 主題の連鎖関係の種類

記事中の各文間が、下記の条件の６種類の連鎖関係のうちで少なくとも１つを満たすものとし、何らかの結束性を持っているとする。

A **主題維持**：直前の文の主題と同一か、基準以上の類似性を持つ主題を持つ場合。

B **主題変化**：直前の文の題述のいずれかと同一か、基準以上の類似性がある主題を持つ場合。

C **主題回復**：最も近い主題変化の直前の主題と同一か、基準以上の類似性がある主題を持つ場合。

D **トピックの導入**：文章のトピックと同一か、基準以上の類似性がある主題を持つ場合。

E **主題派生**：上記のいずれにも該当しない場合。この場合、直前の文やトピックとは関連性の低い文となる。

F **主題の導入**: 最初に主題が出現した場合

ただし、基準以上の類似性とは、一方の語句が他方の部分文字列になっている場合とする。

4.3.3 主題の連鎖関係の決定

主題の連鎖関係を決定するルールを示す。

- ルール１：原則として、結束関係の強さはＡ＞Ｂ＞Ｃ＞Ｄ＞Ｅとし、可能な限り結束性の高い連鎖を採用する。ただし、主題を抽出する際、主題が省略されている文に関しては、省略 (ellipsis) により結束構造 (cohesion)[2] があるものとして、主題の維持と見なす。
- ルール２：最初に主題が出現するまでの文は「トピックの導入」とし、主題を定めない。最初に主題が出現した文を「主題の導入」とし以降主題の連鎖関係を決定する。

4.4 語彙的連鎖構造

語彙的連鎖とは、語彙的結束を持つ語の連続のことをいう。語彙的連鎖は、テキスト中に存在する意味的なまとまりを示すと考えることができる [2]。

本システムでは、連鎖を作る語の最小単位を形態素としている。同じ形態素を持つ名詞句は、その形態素で語彙的連鎖があるとする。

4.5 文末構造

文章は、文末表現により**叙述文**と**意見文**に区別することができる。叙述文は現象を表す文であり、文末表現としては「〜している」,「〜という」,「形容詞の終止形」といった表現が存在する文であるとする。また、主張文は書き手の主張を表す文であり、文末表現として「〜だ」,「〜である」といった表現が存在する文であるとする。

4.6 表層格フレーム表現

前述された汎用的意味処理と汎用的内部表現を prolog により実装する。表層的内部表現は、テキストの表層表現である surf、表層格フレームである frame、各 frame の形態素情報と節情報を持つ morph、時間セグメントを示す`tmp_sgmnt`、語彙的連鎖関係を示す chain、文の各情報 (文番号、 frame の情報、主題の情報、文末の情報) を示す sent という述語により記述される。

1993 年の日経新聞から抽出した事件記事 1601 記事に対して汎用的内部表現を抽出し、文書の一記事に対する平均出現数を調査したところ、表 2のような結果が得られた。

	sent	frame	morph	chain
平均出現数	8.87	64.47	102.37	22.28

表 2: 1601 記事に対する平均出現数

5 犯罪スキーマの抽出

本節では、犯罪スキーマの抽出部について述べ、実際に抽出された犯罪スキーマを示す。

5.1 犯罪スキーマの抽出アーキテクチャ

以下の図 3に示す手順で犯罪スキーマを抽出する。

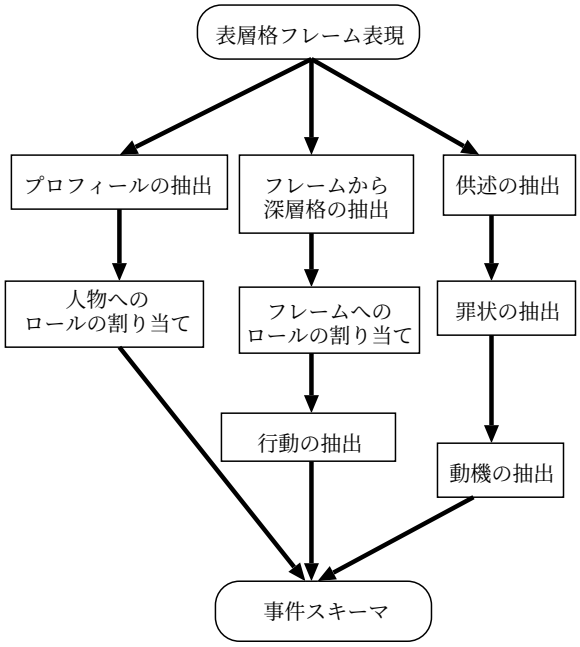

図 3: 犯罪スキーマの抽出アーキテクチャ

5.2 各過程の具体的な方法

ここでは、図 3で示した犯罪スキーマの各過程の具体的な手法について述べる。

- 供述の抽出

  事件記事では犯人の供述が鍵括弧によって括られているものが多い。そこで、そのようなものを、供述という述語をキーにその前に出現する鍵括弧を供述としてパターンマッチングにより抽出する。

- 罪状の抽出

  事件には少なくとも犯人と警察が関係していると考えられる。警察が犯人に対して行う動作としては、「逮捕」,「指名手配」,「書類送検」等が挙げられ、それらの言葉の前には「〜容疑で」,「〜の疑いで」といった語句が存在する。新聞記事によっては、罪状と逮捕の間に犯人の名前などが入る場合もあるため、表層のみでの判断ではなく、KNP によって判定された係り受けの関係を用いる。抽出は、「逮捕」「指名手配」「書類送検」といった述語を検索し、その語句に接続する「〜容疑で」もしくは「〜の疑いで」といった語句を検索する。双方にマッチした際に、前に存在する単語を罪状名として抽出する。

- 動機の抽出

  動機は、基本的には抽出された罪状から類推する。抽出された罪状は「強盗殺人」「大麻取締法違反」,「現住建造物放火」,「業務上過失致死」といったものであり、その表層表現に対して類推を行う。現段階では精密な分類は行わず、すべての罪状に対し、「金目当て」,「怨恨」,「過失」,「その他」として分類を行う。例えば、強盗殺人のように「強盗」を含む罪状は金目当てであり、「殺人」,「傷害」,「放火」などは怨恨の可能性が高いと判断する。また、供述が含まれる記事に対しては、「金」,「邪魔」といった単語の存在も動機を判断する手がかりとする。供述は、より犯人の感情を表現している可能性が高いため、罪状と供述で異なる判定結果が出た場合、供述から判定された結果を採用する

- プロフィールの抽出

  多くの事件記事の場合、人名に対して、名前、年齢、職業、住所などの経歴を最初に出現した個所にまとめて記述する傾向がある。さらに年齢に関して言えば、括弧でくくる傾向が強い。そこで、最初に年齢の出現する個所を特定し、節の情報、形態素の情報をもとにパターンマッチングにより、経歴の書かれている個所を特定しプロフィールを抽出する。警察に関しては、「〜署」、「〜県警」「〜地検」などをキーワードとして、パターンマッチングを行うことにより、その名前を抽出する。

- 人物へのロールの割り当て

  抽出された人名に対し、それらに「〜容疑者」、「〜被告」など言う表現が付随する場合ロールを**犯人**とする。「〜署」、「〜県警」などが含まれる場合ロールを**警察**とする。それ以外の場合は、すべてロールを**被害者**とする。

- フレームから深層格の抽出

  文の意味の内部表現として、表層格フレーム表現を深層格フレームに変換する。深層格フレームは表 3に示すようなスロットを持つとする。

動作主格	動作主または状態の主格
対象格	変化や移動の対象
道具格	動作の原因となること
場所格	出来事が起きる場所
時間格	出来事が起こる時間
源泉格	変化や移動の起点
目標格	変化や移動の終点
状態格	主体の状態。
動作	主体の動作

表 3: 深層格フレームのスロット

- フレームへのロールの割り当て

  フレームに関係するすべてのロールを割り当てる手法として以下の３段階を考える。

  1. 事件記事 1601 記事から抽出された述語 1725 個、1439 のサ変動詞に対し、人手で述語を犯人、警察、被害者、その他の４つに分類し、その分類を元に出現する述語表現からフレームにロールを割り当てる。
  2. 動作主格が人物の場合、そのロールをフレームのロールに割り当てる。
  3. 対象格が人物の場合、そのロールをフレームのロールに割り当てる。

- 行動の抽出

  時間格の時間表現よりフレームをソートし、行動の抽出を行う。時間表現から、順序関係がわからない場合は、出現順を時間順とする。

5.3 犯罪スキーマの抽出例と各スロットの評価

現在実装されているスロットについて評価を行った。表 4に評価結果を示す。この表からみても分かるように概ね８割程度の正解率が得られている。今回、目的とすることは犯罪スキーマを用いて、事件記事を構造化することにあるため、抽出精度は、現段階では十分であると考えている。

犯罪スキーマの結果を図 4に示す。現段階では、深層格の抽出、フレームへのロールの割り当ての２段階以降は手動で行っている。

	全出現数	システム	正解率
ロール	51	50	98.0%
名前	64	51	79.7%
年齢	64	51	79.7%
住所	44	50	88.0%
職業	35	44	79.5%
罪状	30	25	83.3%
供述	7	7	100.0%
動機	30	20	66.7%

表 4: 30 記事に対する抽出結果

6 まとめと今後の展望

事件記事を対象とし、文章全体を構造化する人物の相互関係および時間的進行の観点から犯罪スキーマを提案した。実際に意味解析部は新聞記事 1601 記事対し動作を確認した。犯罪スキーマ抽出部もプロフィール、供述、動機、罪状については自動で抽出し評価を行った。

今後は、深層格フレーム抽出、フレームの人物の同定の自動化を行っていく予定である。

```
kiji('930101-2027',[罪状:強盗障害,動機:金目当て],
     [id1,id2,id3]).
sem(id1,ロール:犯人,プロフィール:[名前:岡田国彦,年齢:３６歳,
    職業:大工,住所:豊田市緑ケ丘五],行動:[id1,id2,id3,id4]).

sem(id2,ロール:被害者,プロフィール:[名前:岡下猛,年齢:４５,
    職業:「豊田交通」社員,住所:豊田市堤町上町一〇五],
    行動:[id2,id3,id4]).
sem(id3,ロール:警察,プロフィール:[名前:愛知県警新城署],
    行動:[id1]).


cls(id1,[動作:緊急逮捕する,動作主:愛知県警新城署,
    対象:岡田国彦容疑者,道具:強盗傷害の疑い]).
cls(id2,動作:停車させる,動作主:岡田容疑者,
    対象:岡下猛さんのタクシー,場所:作手村の建設工事現場,
    時間:二十九日午後十一時十分ごろ).
cls(id3,動作:負う,動作主:岡下さんの顔,対象:軽いけが).
cls(id4,動作:奪う,動作主:岡田容疑者,
    対象:売上金など約十五万円入りのカバン).
```

図 4: 犯罪スキーマの例

参考文献

[1] J.R. キンラン. ＡＩによるデータ解析. トッパン, 1985.

[2] M. A. K.Halliday. An introduction to functional grammar second edition. くろしお出版, 2001.

[3] 永野 賢. 文章論総説, 朝倉書店, 1986.

[4] 福本 淳一, 安原 宏. 文の連接関係解析に基づく文章構造解析. 情報処理学会研究報告, 92-NL-88,1992.

[5] 川端 崇央, 原田 実. 日本語文間の意味関係解析システム InSeRA の開発研究, 情報処理学会研究報告01-NL-142,2001.

[6] 黒橋 禎夫. 日本語構文解析システム KNP version 2.0,1998.

[7] 黒橋 禎夫 長尾 真. 日本語形態素解析システム JUMAN version 3.6,1998.