犯罪スキーマに基づく新聞記事からの事件情報抽出

金山 淳一† 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,tam}@tamlab.eis.ynu.ac.jp

1 はじめに

本研究では、一連の事件において関連する人間の相互関係としての意味構造を犯罪スキーマとして定義し、新聞記事からスキーマの各要素を抽出する手法を検討する。

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、事件、犯罪を検討し、その意味を表現する「犯罪スキーマ」を提案する。さらに、犯罪、事件ついて書かれた新聞記事（事件記事）からスキーマ・インスタンスを抽出する手法を検討する。

本研究ではまず、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）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・インスタンスの抽出を行う。インスタンスの各要素は、関連人物の役割（容疑者、被害者、警察）を示すロール、それぞれのプロフィール、犯罪の動機、犯人の供述、人物の行動からなる。各人物の行動は、人物の一連の動作を格フレームの列として表現される。抽出手法としては、主に時間セグメント分割と時間順整列に基づく手法と事件に関係する動詞の辞書を用いる。 他の要素は、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を要素に応じ組み合わせることにより抽出する。

2 犯罪スキーマ

事件は、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展開していく。そこで、我々は、事件を関係する人物の視点、個々の動作（行動）でとらえる構造として犯罪スキーマを提案する。

そこで、事件は、罪状、動機、供述、人物の4つの要素で表現できると仮定し。我々は事件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。

図1に犯罪スキーマの実現を示す。

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各要素の抽出部に分かれる。意味解析部では、一般的な意味構造を抽出する。犯罪スキーマ各要素抽出部では、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。

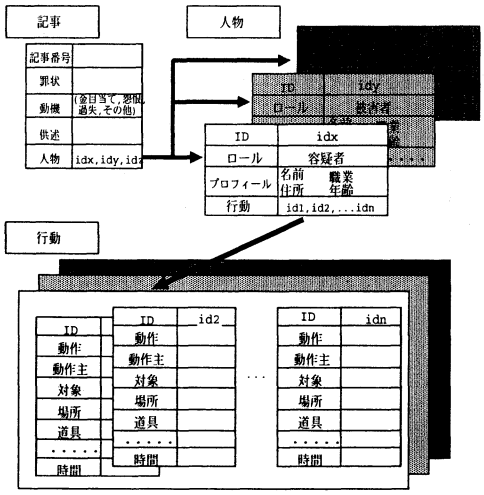

図 1: 犯罪スキーマの全体図

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[6]、構文解析には日本語構文解析ツール KNP[5] を用いている。

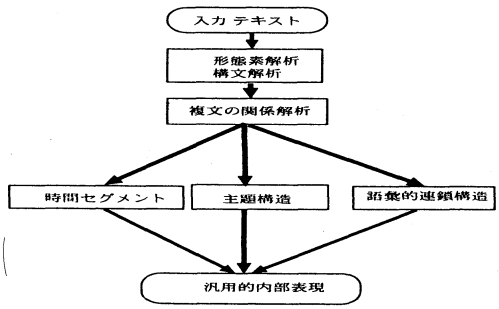

図 2: 意味構造の解析

- 主題構造

  主題構造解析では、文中の「は格」、「が格」を話題の中心 (主題) として抽出し、その連鎖関係 (主題維持、主題変化等) を決定し、文章を構造化する。

- 時間セグメント

  時間セグメント解析では、時間参照表現の出現した単文節で文章を分割し、文章をセグメント化する。

- 語彙的連鎖構造

  語彙的連鎖構造解析では、最小単位を形態素とし、同じ形態素を持つ名詞句を、語彙連鎖があるとして構造化する。

4.1 汎用的内部表現

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。

汎用的内部表現は、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 tsgmnt、語彙的連鎖関係を示す chain、単文節を示す cls 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 1のような結果が得られた。

	cls	frame	morph	chain
平均出現数	12.46	64.47	102.37	22.28

表 1: 1601 記事に対する平均出現数

5 犯罪スキーマの各要素の抽出

本節では、犯罪スキーマ・インスタンスの各要素の抽出部について述べる。

5.1 犯罪スキーマの抽出アーキテクチャ

図 3に示す手順で犯罪スキーマを抽出する。

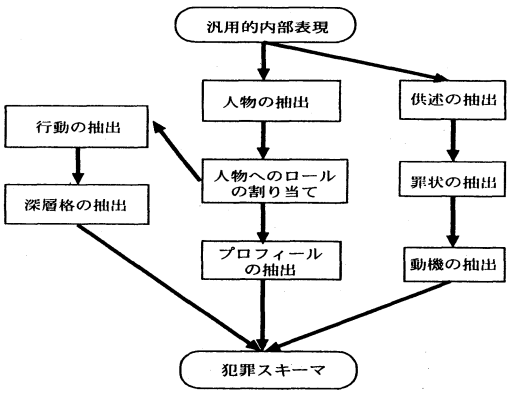

図 3: 犯罪スキーマの抽出アーキテクチャ

5.2 各過程の具体的な方法

ここでは、図 3で示したインスタンスの抽出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

  罪状は、「逮捕」、「指名手配」、「書類送検」といった述語のリストと、その語句に接続する「～容疑で」もしくは「～の疑いで」といった語句のリストを用意しておき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、「金目当て」,「怨恨」,「過失」,「その他」の4種類とし、罪状から動機を類推する手法を以下に示す。

  1. 罪状のキーワードとなるものを個々に分類しておき前段階として罪状と動機の対応リストを作成する。
  2. パターンマッチにより動機を決定する。
     - 実際に抽出された罪状に罪状のキーワードを含む場合、対応する動機を事件の動機として採用する。
     - 含まない場合、その他を動機として採用する。

  ただし、複数、動機となる候補が存在する場合、**金目当て＞怨恨＞過失**の順で採用する。

- 人物の抽出

  多くの事件記事の場合、最初に出現した人名に対し、括弧 () で年齢を括り、それを人名に付随して記述される傾向がある。

  そこで、本研究では、人名を抽出する際、まず最初に括弧で括られている数字を探索する。そして、その節の表層的に直前に隣接する節を人名として、抽出する。

  警察に関しては、警察を明らかに示唆するキーフレーズ「署」、「県警」、「警察」などを含む節を抽出する。

- 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」など容疑者を明らかに示唆する表現が付随する場合ロールを**容疑者**とする。それ以外の場合は、すべてロールを**被害者**とする。前述した人物の抽出で警察として抽出したものに対して、ロールを**警察**とする。

- プロフィールの抽出

  事件記事において、人物に対して、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。

  そこで、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。

5.3 行動の抽出

人物の行動を抽出する手法を以下に示す。

1. 抽出された人物とそのロールから各動作フレームの主格、対象格にロールを割り当てる。

2. 各動作フレームの主格、対象格のどちらかに、抽出対象のロール (容疑者, 被害者, 警察) を含む動作フレームの ID 列を抽出することにより、抽出対象のロール (容疑者, 被害者, 警察) の動作の列を抽出する。

3. 抽出された ID 列を時間順整列された時間セグメントを元に時間順に整列し、その時間順に整列された ID 列を行動として抽出する。

5.3.1 動作フレームへのロールの割り当て

行動として、各ロールに対し動作列を抽出するために、ロールを前述した3つにその他を加えた、容疑者 (S)、被害者 (V)、警察 (P)、その他 (O) の4つに分類し、各動作のフレームの主格、対象格に対して割り当てる。

各動作格フレームへのロールの割り当ての手法を以下に示す。

1. 事件記事 1601 記事から抽出された述語 1725 個に対し、その述語の態をもとに、主格・対象格のロールの分類、取りうる主格・対象格と表層格の対応を記述した辞書を前段階として作成しておく。

2. 辞書を元に、対応する述語の主格と対象格を抽出し各動作フレームに対し主格、対象格のロールを割り当てる。具体的な、割り当てアルゴリズムは、図5に示す。

図5中の格の表層表現とロールを示す語の比較では以下に示す3つの語のリストをロールを示す語としてマッチングを行っている。

- 格の表層表現中にロールを明示する語 (署、容疑者など)
- 人物の名前の形態素
- 図5中でロールを示す語として蓄積されたリスト

5.4 犯罪スキーマの各要素の抽出実験と犯罪スキーマの例

日本経済新聞から事件記事 20 記事を無作為に選び、犯罪スキーマの各スロットの抽出実験、行動の抽出でもっとも重要であると考えられる動作フレームへのロールの割り当て実験を行った。そして、その結果を表 2、表 3に、実際に抽出された行動の例を図 4、犯罪スキーマにより構造化された事件記事の例を図 6示す。

```
----------容疑者----------
フレーム:1
[pred:id11:緊急逮捕した:[aaa,タ形], de:id12:疑いで, wo:id14:（３６）を,
 無格:id19:三十一日, 未格:id20:愛知県警新城署は]
表層:愛知県警新城署は三十一日豊田市緑ケ丘五豊田市緑ケ丘大工豊田市緑ケ丘
     岡田国彦容疑者（３６）を強盗傷害の疑いで緊急逮捕した。

フレーム:2
[pred:id27:停車させた:[aaa,タ形], wo:id28:タクシーを,
de:id34:建設工事現場で:[aaa,タ形], 未格:id40:岡田容疑者は]
表層:岡田容疑者は二十九日午後十一時十分ごろ愛知県南設楽郡作手村の建設工事現場で
     豊田市堤町上町一〇五「豊田交通」社員豊田市堤町上町一〇五
     岡下猛さん（４５）のタクシーを停車させた

フレーム:3
[pred:id55:負わせ:[heiretu,タ形], wo:id56:けがを, 連用:id58:して,
 連用:id63:そして]
表層:そして岡下さんの顔を素手で殴るなどして軽いけがを負わせ

フレーム:4
[pred:id51:奪った:[syusetu,タ形], wo:id52:カバンを, 連用:id63:そして]
表層:そして売上金など約十五万円入りのカバンを奪った
```

図 4: 容疑者の行動の例

	全出現数	再現率	適合率
名前	44	0.86	0.97
年齢	46	0.80	1.00
ロール	50	0.76	0.97
住所	41	0.70	0.96
職業	41	0.66	0.96
警察	29	0.93	0.79
罪状	28	0.75	1.00
動機	20	0.85	0.85
供述	7	1.00	1.00

表 2: 各要素の抽出結果

6 まとめと今後の展望

事件という現象を対象とし、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、新聞の事件記事を対象とし、記事から犯罪スキーマ・インスタンスを抽出した。意味解析部は新聞記事 1601 記事対し動作を確認した。犯罪スキーマの各要素については、プロフィール、供述、動機、罪状、行動について抽出し評価を行った。

参考文献

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

[2] 永野 賢. 文章論総説, 朝倉書店, 1986.

[3] 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.

[4] 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.

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

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

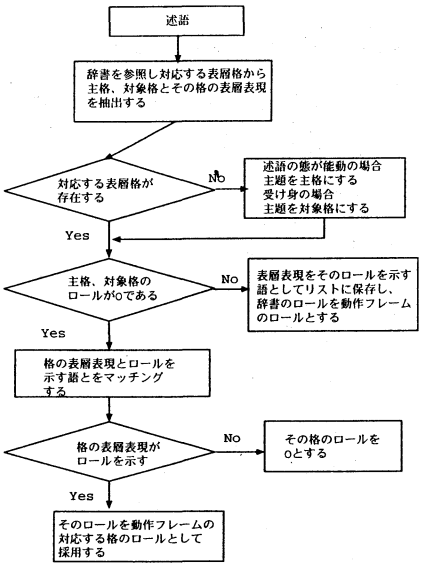

図 5: 動作フレームのロール割り当てアルゴリズム

```
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]).

dframe(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]).

dframe(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).
dframe(id3,動作:負う,動作主:岡下さんの顔,対象:軽いけが).

dframe(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 6: 犯罪スキーマの例

		警察	容疑者	被害者	その他
主格	出現数	69	104	7	21
	再現率	0.86	0.74	0.57	0.71
	適合率	0.86	0.93	1.00	0.33
	F 値	0.86	0.82	0.72	0.45
対象格	出現数	11	65	32	93
	再現率	0.36	0.85	0.50	0.96
	適合率	0.57	0.88	0.94	0.77
	F 値	0.44	0.86	0.65	0.85
合計	出現数	80	169	39	114
	再現率	0.80	0.78	0.51	0.91
	適合率	0.84	0.92	0.95	0.65
	F 値	0.82	0.89	0.66	0.76

表 3: 動作フレームへのロールの割り当ての結果