自然言語処理 126－16
(1998. 7. 24)

# 新聞記事からの要素属性情報の抽出

桝井 文人, 福本 淳一

沖電気工業株式会社
研究開発本部 関西総合研究所

〒540 大阪市中央区城見 1-2-27 クリスタルタワー

*{masui,fukumoto}@kansasi.oki.co.jp*

あらまし 本論文では，新聞記事における固有名詞の属性情報の抽出手法について延べる。本手法は，機械翻訳に用いる構文解析モジュールを利用してテキストの構文解析を行い，解析結果と構造パターンを照合することで属性情報を抽出する。本手法は 1998 年 4 月に開催された情報抽出技術に関する国際会議 Seventh Message Understanding Conference (MUC-7)におけるタスク定義に基づいており，同会議参加システムに実装されている。MUC - 7 テスト用データに適用した結果を用いて，本手法の性能について評価する。

キーワード 情報抽出，MUC-7，構文解析，属性情報，構造パターン

# Extracting Elements and their Attributes from Newspaper Articles

Fumito MASUI, Jun-ichi FUKUMOTO

Kansai Laboratory, Research and Development Group
Oki Electric Industry, Co., Ltd.

1-2-27 Shiromi, Chuo-ku, Osaka 540-6025, Japan

Abstract In this paper, we will discuss about an approach to extract elements and their attributes from newspaper articles. Firstly, a sentence in a text are analyzed by parsing module f MT system. Then, information is extracted using structural information patterns from the parsing results.
We implemented this extraction system based on Template Element(TE) task definition of seventh Message Understanding Conference(MUC-7). We will report the evaluation with 100 texts which were used in formal run test.

keyword information extraction, MUC-7, parsing, attribute, structural information pattern

## 1. はじめに

1998 年 4 月末に情報抽出技術に関する国際会議 Seventh Message Understanding Conference (MUC-7)が開催された[1,2]。我々は MUC-7 の 4 つのタスク— Named Entity (NE), Co-reference (CO), Template Element (TE), Template Relation (TR) — に参加し，情報抽出システムを作成した。本システムは，コンポーネントの組み合わせによって各タスクに応じた柔軟な処理を実現する [4,5]。

本発表では，上記参加システムにおける TE（固有名に関する属性情報の抽出）タスクの処理方式について述べる。第一に，抽出対象となる固有名の属性タイプや属性カテゴリ，すなわち属性情報の定義について説明する。第二に，表層的なパターンによる抽出手法およびパージングと構造的なパターンマッチングによる抽出手法の有効性について述べる。第三に，MUC-7 training data コーパスから構造的なパターンを収集し分析を行ったので，その結果について述べる。最後に，本手法を用いた実装システムを用いて MUC-7 formal run data コーパスへ適用し，抽出結果についての評価と考察を行う。

本発表では，NE（固有名の抽出）処理によって抽出された固有名の実体（Entity Name）を“要素”とよぶことにする。

## 2. 抽出対象の定義

MUC-7 における Template Element (TE) タスクとは，SGML タグが付与された新聞記事データから NE 処理によって抽出された要素の属性を抽出するものである[5]。抽出要素には実体要素 (ENTITY) と場所要素 (LOCATION) の区別があり，実体要素には人名・組織名・製品名の 3 種の属性タイプと，場所要素には地域・国家・地方・都市・水域・空港・他の場所，の 7 種の属性タイプ (ENT_TYPE) が定義されている（表 1, 2)。さらに，実体要素には各タイプごとに属性カテゴリ (ENT_CATEGORY) が定義されている。属性カテゴリは，人名には軍人・民間人，組織には官公組織・民間組織・他の組織，製品名には，陸・水・空，と合計 9 種類の項目が設定されている（表 1)。ただし，実体要素の製品名タイプについては実際には陸・海・空に関する乗物に限定されている。場所要素には要素が属する国名も抽出対象となっている。(COUNTRY)。このほか，何らかの形で実体要素について記述または参照している文字列も抽出対象となっている。抽出される情報は要素の名前や別名を含まない文字列であり，要素そのものとは区別され，属性記述 (ENT_DESCRIPTOR) として定義されている。属性記述として抽出される文字列は，複合名詞句全体

表 1 実体要素の属性

| タイプ | カテゴリ | 例 |
|---|---|---|
| 人名 | 軍人 | Napoleon |
| | 民間人 | Zidane, Nakata |
| 組織名 | 官公組織 | FBI |
| | 民間組織 | FIFA, Air France |
| | その他 | Eskimo |
| 製品名 | 空 | Endever |
| | 地 | TGV, Eurostar |
| | 水 | Titanic |

表 2 場所要素の属性

| タイプ | 例 |
|---|---|
| 地域 | Asia, South America |
| 国家 | Japan, Argentina |
| 地方 | Provence, Bordeaux |
| 都市 | Toulouse, Marseille |
| 水域 | Atlantic Ocean |
| 空港 | Charles de Gaulle |
| その他 | Earth |

である場合や複合名詞句の部分文字列である場合など，その範囲に幅がある。

以下に属性記述の例を示す。四角で囲まれた範囲が属性記述であり，下線で示した範囲が要素である。

- 名詞句全体が属性記述となる例
  - the twin-engine, two-seat Tomcat
  - a chartered 757 aircraft
  - a plane modeled on the Boeing 737
  - the new plane
- 名詞句の部分文字列が属性記述となる例
  - the law firm Smith Blarney
  - Oki Electric Corp's Kansai subsidiary
  - President Clinton
  - the space shuttle Challenger
- 限定的でない名詞句が属性記述となる例
  - General Dynamics Corp. is a major contracting firm.

- Another potential choice for a top position in the State Department is Democratic lawyer Tom Donilon

属性記述によっては，情報の明確さにも差が生じる。しかし，抽出するものは，要素を明確に指し示しているもののみを対象としている。よって，集合を記述する場合や抽象的な参照，潜在的な真実のみについての参照や真実でないことの描写は含まない。たとえば，"it"のような代名詞，"Mr.", "Sir", "Dr."のような基本的な接辞，"company", "airplane"のような要素を特徴付けるような名詞を中心語とする単純名詞句（無冠詞や，"the, "his", "this"のような限定詞のみの修飾や非限定的企業名などを含む固有名詞）は，"抽象的" 属性記述であるため，抽出対象としない。

➢ 抽象的属性記述の例（抽出対象としない）

- Mr. Bean
- plane
- the plane
- his plane
- that plane
- the Boeing 737 plane
- KML Airline's Boeing 737 plane

## 3. パターンに基づく抽出手法

本章では，文字列の表層的な特徴パターン（表層パターン）を利用した属性情報の抽出手法および構文の構造的な特徴パターン（構造パターン）を利用した要素属性の抽出手法について述べる。図 1 のように比較的単純な属性記述を伴う要素については表層パターンが有効であるが，図 2 のようにやや複雑なものについては構造的なパターンの利用が効果的である。以下，本手法について詳述する。

### 3.1. 表層パターンに基づく抽出

図 1 の例に示すような "Cmdr. Jean-luc Picard" や "River Shannon" という文字列では，要素の特徴を示す接辞や機能語などが隣接している。この場合，前者は，"Cmdr." という文字列は軍隊の指揮官または司令官であるという職務または階級を示す接辞であり，"River" という文字列は川であることを示す機能語である。よって，機能語や接辞を特徴文字列として利用することで，前者の場合は，"Jean-luc Picard" という文字列が属性カテゴリ "軍人" を持つ属性タイプ "人名" の実体要素であるということ

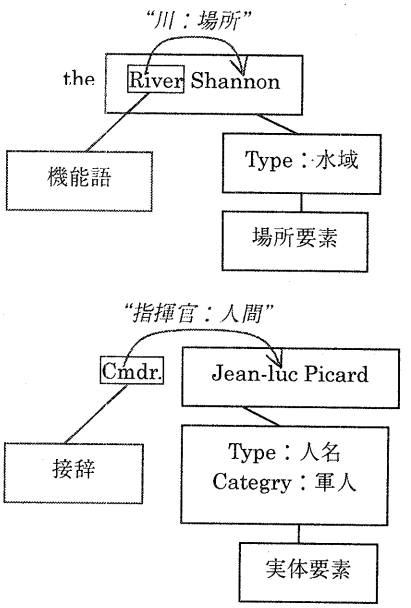


図 1 表層パターンに基づく抽出

がわかり，"Cmdr." という文字列が属性記述となる。後者の場合，"River Shannon" という文字列が属性タイプ "水域" の場所要素であることがわかる。このように，明確な特徴文字列を伴う文字列については，あらかじめ機能語や接辞のリストを作成しておき照合することで表層パターンのパターンマッチングによって，比較的簡単に属性情報を抽出することが可能である。

### 3.2. 構造パターンに基づく抽出

前章の例であげた "General Dynamics Corp. is a major contracting firm" という文は，要素と属性記述が隣接していない。もちろん，表層パターンを利用すると "Corp." という接辞が特徴文字列となり，属性カテゴリが "民間組織"，属性タイプが "組織" という属性情報が得られる。しかし属性記述となる "a major contracting firm" の抽出は困難である。また，固有名に接辞が伴わず，単に "Genera Dynamics is..." と表された場合，属性情報の認識さえも困難となる。ところが，実際の新聞記事中には "New York Times" や "Bill Clinton" のように明確な特徴文字列を伴わずに表記される例も多くみられる。よって上記のような表記にも対応できる手段が必要である。

以下，我々は構造パターンの照合を利用する属性情報の抽出手法を提案する。

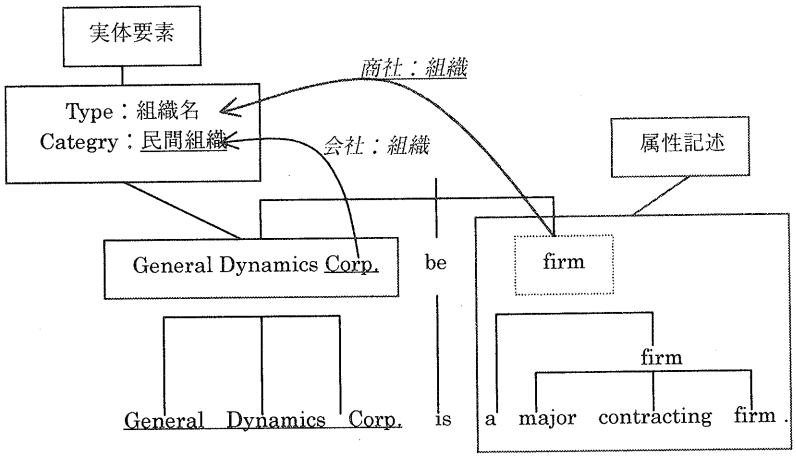


図 2　構造パターンを利用した抽出

前章の例文 “General Dynamics Corp. is a major contracting firm” に対してパージング（形態素解析および構文解析）すると，図 2 に示すような構文木で表すことができる。固有名 “General Dynamics Corp.” は未知語の集まりであり，“a major contracting firm” は名詞 “firm” を中心語とする名詞句である。二つの名詞句と be 動詞は “SVC” の第三文型の文構造を構成する。ここで，構文木の中心語ノードに注目すると，例文は “未知語（固有名候補） + be-verb + 名詞句（firm）” として扱うことができる。上のような SVC 構文が S = C の関係を表すものであることを考慮すれば，未知語と firm が同じ属性を持つべきであることがわかる。すなわち，中心語 “firm” を特徴文字列とみなすことができ，表層パターンを利用した手段と同様に，未知語 “General Dynamics Corp.” は属性タイプ “組織名” で属性カテゴリ “民間組織” という属性をもつ実体要素であることがわかり，名詞句 “a major contracting firm” は属性記述となる。このように，上記例文のようなやや複雑な場合でも，“固有名（要素）＋be-verb＋NP（属性記述：組織）” のような，構造レベルを含めた要素と属性記述の組み合わせパターンのバリエーションリストを作成して構造的な照合を行うことで，表層パターンと同様な効果を得ることが可能である。このような，構造レベルを考慮した言語パターンを構造パターンとよぶことにする。

本手法は，単純に考えれば構造パターンの規模が大きいほどカバレッジが高くなるはずである。ただし，単純にパターンのバリエーションを増やすだけではなく，利用効果の高いパターンを上手く抽象化して利用する必要がある。

表 3 属性タイプ “人名” の構造パターン

| |
|---|
| suffix + NAME |
| NAME + NOUN |
| NOUN + NAME ’s + NOUN + NAME |
| suffix + NAME + NOUN |
| NAME + “be” + NOUN |
| NOUN + “be” + NAME + NOUN |

## 4. 構造パターン

前章では，パージングと構造パターンを用いることで，表層パターンのみでは対処しきれないケースにもパターンマッチングによる属性情報抽出が有効であることを示した。本章では，利用効果の高い構造パターンを得るために新聞記事を調査したのでその

表 4 属性タイプ “製品名” の構造パターン

| |
|---|
| DET + ADJP + NAME |
| DET + NAME + NOUN |
| DET + ADJP + NAME + NOUN |
| DET + NAME + ADJP(ADJC) |
| DET + ADJP + NAME + ADJP(ADJC) |
| DET + NOUN + “of” + DET + NAME |
| NOUN ’s + NAME + NOUN |
| NAME + VERB(past participle) |
| NOUN + “called” + NAME |
| NOUN + “be” + NAME |

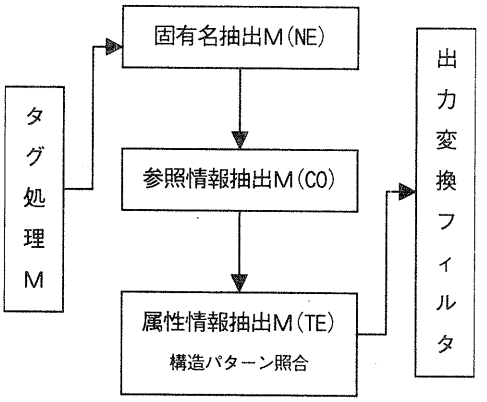


図 3 属性情報抽出システム構成

調査内容と結果について述べる。
調査対象には MUC-7 training data 30 記事（2251 文)を用いた。要素と属性記述が関連性を持つパターンを記事から人手で選び出した。人間の目で見て曖昧性が残るもの，一文を超えて対応するものは対象外とした。選び出したパターンのうち，共通するものまとめて抽象化した。その結果，属性タイプ“人名”および“製品名”に関する構造パターンについては規則性がみられたが，属性タイプ“組織名”については，明確な規則性はみられなかった。規則性がみられた属性タイプ“人名”と“製品名”については，それぞれ 55 組，43 組のパターンが得られた。これらを抽象化した構造パターンを表 3，4 に示す。
属性タイプ“人名”については，要素と属性記述が前後に隣接するパターンが多く，表層パターンマッチングでもかなり対応可能であると思われる。属性タイプ“製品名”に関しては，固有名詞要素を名詞句の一部として，冠詞を伴う名詞句として出現し，固有名詞要素の前後に特定のキーワードまたはそれを含む名詞句が隣接するパターンのバリエーションが確認できた。どちらも be 動詞または“called”のような他動詞の完了形によって要素とその説明が関連付けられているパターンもみられた。また，属性タイプ“製品名”に関しては，特定の動詞が非常に高頻度で出現していることがわかった。対象とした記事の場合，“crash”，“take off”，“fly”という三つの動詞句でほとんど網羅される。これは，記事内容が航空機の衝突に関するもので占められていたことに由来すると思われるが，これは，製品に関連する動詞はドメインによってかなり絞られるということを意味していると思われる。よって，記事内容やドメインが限定され，その動詞の出現傾向を知ることができれば，属性タイプ“製品名”の抽出手段として動詞が有効であるといえる。要素の前後に隣接する名詞句について述べると，属性タイプ“人名”では，名詞句として出現するものが 25 組，接頭辞として出現するものが 18 組であった。属性タイプ“製品名”では，名詞句として出現するものが 13 組であり，接頭辞は存在しなかった。
表 3，4 のような抽象化されたパターンは，全体的な構成が共通である。修飾語句の有無やその長さなどが変化することでバリエーションを表現できる。

## 5. システム概要

これまで述べてきたような構造パターンを利用した属性情報認識処理(TE)システムの構成を図 3 に示す。本システムは MUC-7 への参加システムである。本システムは，(1)タグ処理モジュール，(2)固有名抽出モジュール，(3)参照情報抽出モジュール，(4)属性情報抽出モジュール，(5)出力変換フィルタ，の 5 つのモジュールから構成されている。各モジュールは，MUC-7 の他のタスクにも利用される。3 章で述べた表層パターン処理は，固有名認識（NE）処理のために実装されたパターン処理と共通であり，固有名抽出時に属性情報も同時に認識するようなっている[6]。構造パターンを利用した手法は属性情報抽出モジュールにおいて実装した。
以下，属性情報抽出モジュールについてさらに詳しく述べる。

### 5.1. 属性情報抽出モジュール

属性情報抽出モジュールは，(1)要素認識部，(2)属性記述認識部，(3)関連情報認識部，(4)情報選択部，(5)構文木変換部，の 5 つの処理部より構成される。5 つの処理部は簡単な制御 I/F 部によって構文の種類によって使い分けられる。以下，各処理部について詳細に述べる。

#### 5.1.1. 要素認識部

本モジュールは，要素または要素の可能性があるノードをマーキングする。マーキングした名詞句が前処理によってすでに要素として認識されている場合は，そのまま要素として認定する。属性タイプや属性カテゴリが決定されている場合もその情報を採用する。NE 処理では，要素としての固有名の抽出と，人名・組織・製品・場所の区別が可能である。

#### 5.1.2. 属性記述認識モジュール

本モジュールは，属性記述または属性記述の可能性があるノードをマーキングする。マーキングしたノードがすでに前処理によって属性記述として認識されている場合は，そのまま属性記述と認定する。属性タイプや属性カテゴリが決定されている場合もそ

の情報を採用する。マーキングしたノードにパージング時の辞書引きによる意味カテゴリ情報の人・組織・人工物・場所のいずれかが付与されている場合，属性記述として認識する。属性記述候補としてマーキングされている名詞句または前置詞句ノードが存在していれば，属性記述となる名詞句中心語の特徴文字列パターンの照合を行う。パターンが一致した場合は，パターン毎の属性タイプと属性カテゴリが付与され，属性記述として認定される。

#### 5.1.3. 関連情報認識部

本モジュールは，典型的な属性記述および，要素と属性記述の構造パターンを認識する。要素または要素候補としてマーキングされている名詞句ノードと属性記述または属性記述候補としてマーキングされている名詞句や前置詞句のノードが存在している場合，構造パターンと照合され，お互いの品詞や意味カテゴリ，属性情報を参照する。照合が一致した場合は，属性タイプや属性カテゴリが決定し，候補ノードは要素および属性記述として認定される。構文木内に要素と属性記述ノードが存在する場合は，構造パターンと照合する。構造パターンと一致すると，参照関係情報を利用して通番を付与する。この通番をたどることによって，同一の実体を指す要素および属性情報は統合可能となる。
ただし，今回は時間的制約のため，構造パターンは表 5 に示すように 4 種類のみを実装，属性記述の中心語の特徴文字列は実装しなかった。よって，属性記述の認識は，固有名詞認識に伴う特徴文字列パターンおよび，構造パターンに依存する。

#### 5.1.4. 情報選択部

本モジュールは，2～4 の処理で認識付与された要素と属性情報の組み合わせをチェックし，不当な組み合わせを解除する。

#### 5.1.5. 構文木変換部

本モジュールは，テキスト（SGML）出力形式に対応するための処理を行う。各処理で認識した要素属性情報にタグ情報を付与し，構文木構造をフラット化する。

### 5.2 システムの動作

ここでは，本章で説明した属性情報抽出システムの動作について述べる。
タグ処理モジュールは，SGML タグ付きの MUC コーパステキストを入力とし，タグを認識してテキスト解析への影響を避けるためのガード処理を行う。固有名抽出モジュールは，表層パターン処理部と構文解析処理部から成る。表層パターン処理部は，固

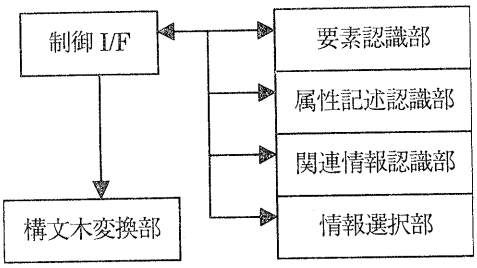


図 4 属性情報抽出モジュール構成

有名を認識して解析への影響を避けるために固有名をガードする。その際に接辞や機能語などキーワードが隣接していた場合に属性情報を抽出する。構文解析処理部は，入力テキストを構文木に変換して構文構造を認識する。参照情報抽出モジュールは，MUC でいう CO 処理にあたり，一文を超えて同一の実体を示名詞の参照関係を認識し，構文木の該当ノードにマーキングを行う。属性情報抽出モジュールは，

表 5 システムに実装した構造パターン

| |
|---|
| NAME + NOUN |
| NAME + “be” + NOUN |
| NOUN + NAME |
| NOUN + “be” + NAME |

構造パターンと構文木のパターンマッチングを行い，表層パターン処理で成功しなかった属性情報の認識を行う。出力変換フィルタは，表層パターンおよび属性情報抽出モジュールによって認識された属性情報のマーキングを取り出して，テキスト出力へ変換する。このとき，参照関係情報も利用して，同一の実体に関する情報を統合する。

## 6. 評価と考察

以上述べてきたシステムを MUC-7 コーパスに適用し，評価を行った。コーパスの内訳は構造パターンの調査に用いた MUC-7 training 用記事および，同 formal run 用テキストであり，両記事とも 100 記事約 8,000 文である。ドメインは training data が “衝突” に関する記事，formal run が “ロケット” に関する記事である。評価には，再現率，適合率を用い，再現率は式 1，適合率は式 2 を用いた。
全体的な結果は，training data に対する再現率が 33%，適合率が 88%，formal run data に対する再現率が 31%，適合率が 71%であった（表 6，7）。両者とも，再現率が低かったが，適合率は高い値を示

式 1　再現率の計算式

$$再現率 = \frac{システムのアウトプット中の正解数}{テキスト中の正解の総数}$$

式 2　適合率の計算式

$$適合率 = \frac{システムのアウトプット中の正解数}{システムのアウトプットの総数}$$

run data に対する再現率が 31%，適合率が 71%であった。両者とも，再現率が低かったが，適合率は高い値を示した。両者の数字を比べると，未知のテキストである formal run data についても，再現率を維持したまま高い適合率を得た。このことは，少なくとも同じ種類の新聞記事であれば，本手法がある程度ドメインに依存せず有効であることを示したといえる。しかし，前者と後者の適合率の差は 17%あり，これは前者のコーパスで適用できる知識が後者では適用できない場合があったを示している。したがって，本手法においても，特定ドメインに有効な構造パターンの分離やドメインによって違った属性を示す構造パターンの設定など，ドメインに依存する知識を意識しておく必要がある。

その他，得られた数値を個別にみると，属性記述，場所要素についての精度が悪く，再現率低下の主要因ともなっている。実体要素名については，タイプ“製品名”の要素抽出の精度が良くなかった。これらの問題点の原因はほとんど共通しており，構造パターンの問題，パージングの問題，モジュール間のフィルタリングの問題，と大きく三つに分けられる。以下これら三つの問題について考察する。

## 6.1.　構造パターンの問題

今回の実装では時間的制限から調査によって得られた構造パターンすべてを実装できなかった。そのため，構造パターンのカバレッジが狭すぎたことが全体的に再現率に影響した。表 3，4 の構造パターンすべてを反映させれば，もう少し再現率の向上は可能であると思われる。

属性記述については，ほとんど結果が得られなかった。構造パターン処理では，構造パターンにマッチした構文木の複合名詞句の中心語と，処理内の特徴文字列との文字列または属性情報の照合を行って属性記述を認識する。ところが，照合されたノードを頂点とする下位の構文木構造を属性記述のスコープとするためのマーキング処理を行ったあと，マーキングされた部分が，構文木変換部の処理中に消去されてしまう場合があった。これが原因となって本来認識された構造パターン処理の結果がうまく反映されなかった。さらに，固有名抽出処理で適用される表層パターン処理に用いる接辞や機能語などの特徴文字列の知識を受け渡す I/F がうまく働かなかったために，表層パターン処理結果も反映されていない。同じ記事に対して表層パターン知識を用いて，人手によって処理のシュミレートを行った結果，再現率，適合率とも 40%台の値が得られた。現在，構文パターンを用いたシュミレートを進めているところである。本手法の実際の属性記述抽出の精度は，上記の表層パターン処理のシュミレート結果に，構文パターン処理のシュミレート結果を加えたえたものと考えて良い。

上では実装上の原因について述べたが，つぎに，実質的な原因について考えてみると，構文構造は似ているが構造パターンに上手くマッチしないケースがある。本手法では，構造パターンの中心語のマッチング条件は，語レベルの完全マッチが科せられている。そのため，たとえ構造パターンと構造が同じであっても，パターンに記述されている“airplane”とは同様の意味を持つが表記が異なる“airliner”や“jet”にはマッチせず，属性情報の抽出が失敗してしまう。対策として，類義語や同義語を記述した辞書や類義語計算手段を用いて構造パターン照合時に中心語の範囲を類義語や同義語へ拡大する手段が考えられる。

## 6.2.　パージングの問題

本システムに用いたパーザは機械翻訳エンジン用に

表 6 属性情報抽出結果（training data）

単位:%

| | 再現率 | 正解率 |
|---|---|---|
| 実体要素名 | 55 | 100 |
| 属性タイプ | 42 | 100 |
| 属性記述 | 0 | 0 |
| 属性カテゴリ | 42 | 100 |
| 場所要素名 | 8 | 25 |
| 属性タイプ | 17 | 50 |
| 国 | 33 | 100 |
| TOTAL | 33 | 88 |

表 7 属性情報抽出結果（formal run data）

単位：%

| | 再現率 | 正解率 |
|---|---|---|
| 実体要素名 | 37 | 79 |
| 属性タイプ | 46 | 98 |
| 属性記述 | 0 | 0 |
| 属性カテゴリ | 34 | 72 |
| 場所要素名 | 41 | 80 |
| 属性タイプ | 5 | 9 |
| 国 | 26 | 49 |
| TOTAL | 31 | 71 |

開発されたものを一部改造したものである。本来翻訳を前提としているために，構文の変形や目的言語ノードの生成という処理が設定されている。また，構文解析初期段階から翻訳に特化した処理があり，これらに起因する木構造の変形やノードの消去などが構造パターン照合ミスに繋がり，再現率低下の原因になっていると思われる。対策として，パーザのMT ベースのルールを完全にリプレースしてしまうか，他の汎用パーザを利用することが考えられる。

### 6.3. フィルタリングの問題

最後に，属性記述については実質的な評価ができなかった。これは，出力形態に変換するフィルタリングに起因するところが大きい。本システムの処理は全般的に構文木の状態で進められる。構文木の該当ノードに必要な情報を付与する形で情報の認識・抽出を行う。最後に変換フィルタを通して構文木の単語情報および各属性情報を取り出して BDF として出力している。ところが，属性情報が複数の単語にわたって付与されている場合や，同じノードに複数の情報が輻輳している場合に，情報取り出しに失敗することが考えれられる。この件も時間的制約による実装の問題と絡むところが大きく，他のモジュール間にも生じる問題であるが，出力そのものを変換するフィルタリングの段階が最も影響する。現在，フィルタリングに修正を加え，実装上の問題を排除した状態での評価を進めている。

以上の問題の他に，場所要素に関する処理結果が悪い点については，辞書的知識をほとんど用いなかったことが原因とみられる。構造パターンに利用できる構造的規則性がみられなかったため，本手法による効果がほとんど見られなかった。場所要素はについては，部分的なパターン処理で対応できる場合が少ない。よって，辞書的なデータの拡充によって対応することが妥当であると思われるが，本システムでは辞書的知識はあまり用いられていない。場所の名前はドメインに依存する要素は小さいので，辞書的知識を使えば比較的簡単に現率・適合率にとも向上できるであろう。もちろん，辞書的知識は製品名や組織名など，他の要素にも有効なはずである。

## 7. おわりに

情報抽出技術に関する国際会議 MUC−7 に参加したシステムを構成する属性情報抽出および，属性間関係情報抽出技術について報告した。表層パターンのみでは対応しきれない場合にも有効な構造パターンを利用した手法を提案した。コーパス中の構造パターンを調査した結果，属性タイプ“人名”および属性タイプ“製品名”に関する構造パターンについては規則性が認められた。得られた構造パターンを抽象化し，一部を属性情報抽出モジュールへ実装した。さらに，MUC-7 training data と formal run data を対象に処理を行った。処理結果は再現率が 3 割，適合率が 7~8 割であった。再現率の低さについては，MT 用パーザを流用したために情報抽出に合った情報が得られない場合があったことと，構造パターンのカバレッジが十分でなかったことが大きな原因と思われる。ただし，適合率はある程度高い数値を得ているので，構造パターンの実装を充実させることによって再現率向上は可能であろう。

今後は，処理性能向上を目指して，パーザの改良，フィルタ機能の強化，および構造パターンの拡充によるシステムの精緻化を進める。また，日本語新聞記事を対象にした属性情報抽出の研究を行う予定である。固有名抽出については，日・英のシステム間の比較検討を行ったが[6]，属性情報抽出についても同様に，どのような点が共通技術として利用でき，どのような点で新たな技術が必要なのかを検討する予定である。

## 参考文献


[1] TIPSTER TEXT PROGRAM Phrase III, DARPA, 1996

[2] Proceedings of Seventh Message Understanding Conference(MUC-7), DARPA, 1998

[3] Chindhor, N. “MUC-7 Information Extraction Task Definition,” Ver.4.2, 1998

[4] 福本，下畑，桝井，佐々木“パターン処理に基づく情報抽出システムの概要”言語処理学会第 4 回年次大会発表論文集, pp.230−233, 1998

[5] Fukumoto,J., Masui,F., Shimohata, M. and Sasaki, M. “Oki Electric Industry : Description of the Oki System as Used for MUC-7,” Seventh Message Understanding Conference (MUC-7), DARPA, 1998

[6] 福本，下畑，桝井“固有名詞抽出における日本語と英語の比較”，信学技法，NLC98(1998-07), Vol.98, 1998

[7] 若尾“英語テキストからの情報抽出”，信学技法，NLC96-20(1996-07), Vol.96, 1996