# 6K-01　オントロジーに基づく電子メールからのスケジュール情報の抽出

後藤 将志　福田 直樹　新谷 虎松
名古屋工業大学知能情報システム学科
e-mail: {shoji, fukuta, tora}@ics.nitech.ac.jp

## 1　はじめに

最近，パターンマッチングを利用して電子メールからスケジュール情報を抽出し，スケジュール管理システムとの連携を図ろうとする研究が盛んに行われている [1]．パターンマッチングを利用した情報抽出では，あらかじめ記述した文章のパターンを利用して抽出を行う．この方式の特徴として，抽出した結果の適合率(抽出した結果における正解の割合)が高いという点があげられる．パターンマッチングにおいて，抽出できる文章はパターンの数に依存する．再現率(全正解数から抽出できた割合)を向上させるにはパターンを多く記述する必要がある．スケジュール情報を抽出し，ユーザに提示するシステムにおいて，情報抽出手法に要求されるのは，適合率の向上も必要であるが，再現率の向上が重要であると考えられる．しかし，パターンの記述にはたいへんな労力を必要とする．

本論文では，スケジュール情報の抽出において，オントロジーを利用した手法を提案する．この手法により，パターンマッチングでは抽出できないような情報を抽出できることを示す．本手法を用いることにより適合率を低下させることなく，再現率を向上させることができることを示す．

## 2　スケジュール情報の抽出

### 2.1　パターンマッチングの問題点

パターンマッチングを用いた手法では，日本語の助詞が持つ意味に注目する．パターンの記述には，ある特定の分野，ここではスケジュール情報の記述において助詞の使われる場面など考慮にいれる．パターンマッチングでは多様な意味の助詞を文章の構造によって吸収する．パターンマッチングでは文章一文ずつに対してマッチングを行うことが多い．パターンマッチングでは二文に分けて記されているような情報からの抽出は難しい．例えば「ミーティングを行います．日時と場所は 11 月 21 日ゼミ室です」という文章である．このような文章から抽出ができるよう，網羅的にパターンを記述するには大きな労力を必要とする．

Ontology based schedule information extraction from E-mail
Shoji GOTO
Naoki FUKUTA
Toramatsu SHINTANI
Dept. of Intelligence and Computer Science, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, 466-8555, JAPAN

### 2.2　オントロジーを利用した情報抽出

本稿ではパターンマッチングに加え，オントロジーを利用した情報抽出を提案する．

オントロジーは概念階層を基本とする．一般にオントロジーには更に，概念の意味定義と概念間の関係を明確に記述する．あるドメインに特化した情報を扱うには，このようなオントロジーは巨大すぎ，構築にたいへんな労力を必要とする [2]．情報抽出を目的としたオントロジーには，抽出の対象となる項目がどのような属性を持つかといった事物の意味定義を与える概念，及び概念間の記述が必要となる．本研究で作成したオントロジーは，クラス概念とその属性概念，概念間の関係のみを記述する．

本手法では，オントロジー中の各概念に言語表現パターンについてのルールを記述することにより抽出を行う．本研究で作成した「場所」に関するオントロジーの一部を図 1 に示す．

```
class(場所, has([地名, 建物, 部屋])).
concept(地名, endsWith('町')).
concept(地名, endsWith('村')).
concept(建物, has([建物名, 階数]).
concept(建物名, endsWith('館')).
concept(建物名, endsWith('所')).
concept(建物, endsWith('大学')).
concept(階数, [数字, '階']).
concept(階数, [数字, 'F']).
concept(数字,
        or(['0','1','2','3','4',
            '5','6','7','8','9'])).
concept(部屋, endsWith('室')).
concept(部屋, endsWith('部屋')).
```

図 1: 場所に関するオントロジーの記述例

オントロジーは述語論理式で記述する．述語classは抽出したい概念(この場合は場所)を示す．概念「場所」中の関数子hasは，概念「場所」が属性概念として「地名」「建物」「部屋」を持つことを示す．述語conceptは各属性概念の定義を記述する．ここで言う概念の定義とは言語表現パターンの記述を含む．関数子endsWithは言語表現パターンを記述するために用いられる．endsWithは，その概念を表現する単語の最後の文字が，第一引数に与えられる文字であること

を表す．例えば概念「建物」の場合，建物を示す概念は更に属性概念として建物名と階数を含む．また「建物名」を示す概念は表現として「館」や「所」などで終わるということを表現する．

図2に抽出処理の流れを示し，各ステップについて説明する．

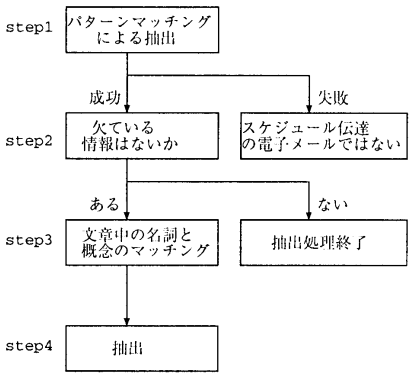


図 2: 情報抽出の流れ

**step1** パターンマッチングによる抽出処理を行う．パターンマッチングが成功した場合，対象となる文書がスケジュール情報を含む可能性が高い．

**step2** あらかじめ決めておいた抽出項目の中で，パターンマッチングで抽出できなかった情報がないかを調べる．

**step3** 概念と文章中の名詞とのマッチングを行う．マッチした名詞はその概念を包含する概念が抽出対象となっているときに抽出する．

**step4** 一つの概念につき，複数の情報が抽出された場合はパターンマッチングが成功した文章に一番近いところに記述されている情報を抽出する．

### 2.3 システムへの応用

本研究では，スケジュール抽出の応用例として，スケジューラと連携を図る情報検索エージェントを実現した．エージェントはPOP3を実装しており，定期的に電子メールを取り込む．取り込んだ電子メールを用いてスケジュール情報の抽出を行う．エージェントはサーバとして動作し，スケジューラがそのクライアントになる．TCP/IPで通信を行うことにより，クライアントはスケジュール情報の抽出結果をvCalendar形式で受け取ることができる．

## 3 評価と考察

オントロジーを利用した抽出の精度を確かめるために，本研究では「場所」に焦点を当てて実験を行った．

オントロジーの作成には，スケジュール伝達に関係があるかないかに問わず，実際に届いた電子メールの中から場所の表現を抜き出して作成した．オントロジー中の概念数

文章中からの名詞の抽出には奈良先端技術科学大学で開発された「茶筌」を利用した．

電子メール400通を利用して抽出実験を行った．そのうち，スケジュール情報を含む電子メールが67通である．抽出した結果の再現率と適合率を，パターンマッチングのみ，パターンマッチング＋オントロジーそれぞれについて表1に示す．

表 1: 場所に関する抽出実験結果

| | 適合率 | 再現率 |
|---|---|---|
| パターンマッチのみ | 72.2 % | 77.6 % |
| パターンマッチ+オントロジー | 72.2% | 83.6% |

表1により，場所に関して，オントロジーを利用した抽出をパターンマッチングに併せて利用した場合，適合率を下げることなく，再現率の向上を実現できたことがわかる．

実験において抽出できなかった項目として，固有名詞で表現される場所名あげられる．固有名詞をパターンとして記述することはできないため，概念とのマッチングに失敗したことがあげられる．これについては固有名詞で表される概念をオントロジーに追加することにより解決できると考えられる．

## 4 まとめ

本研究では抽出に利用する手法としてオントロジーを利用した手法を提案した．この手法を用いれば，パターンマッチングで抜き出せなかった情報を抽出することができる．また，実際に届いた電子メールからの抽出実験により，適合率を下げることなく，再現率を更に向上させることができたことを示した．

## 参考文献

[1] 長谷川 隆明，高木 伸一郎: 文書構造の認識と言語の特徴の利用に基づく電子メールからのスケジュールと ToDo の抽出，情報処理学会論文誌 Vol.40 No.10，pp.3694-3705，1999

[2] 廣田 啓一，佐々木 裕，加藤 恒昭: オントロジー主導による情報抽出，人工知能学会誌 Vol.14 No.6，pp.1010-1018，1999

[3] 岩爪 道明，白神 謙吾，畑谷 和右，武田 英明，西田 豊明: テキストからの情報抽出・統合化法の提案と知的情報収集・分析システム IICA の実験的評価，電子情報処理学会 DEWS '96 論文集，pp.91-96，1996