# メールからのイベント情報抽出によるスケジュール管理

大橋　菜津美　(指導教員：小林一郎)

## 1　研究背景と目的

インターネットの普及に伴い，電子メールによる予定の調整が昨今増えている．その際，メールの交換によって決まったイベントの日程などを手帳やスケジュール管理ツールに転記するという手間が発生する．その手間を省くため，メールを自動的に読み取り，スケジュールに関連する必要な情報を抽出し，カレンダーに書き込むシステムが望まれる．この様な背景から先行研究として，長谷川ら [6] は，イベント通知タイプのメールからイベント情報を抽出し，カレンダーに記入する手法を提案している．本研究においては，主に二人のユーザによるメールでの対話的な交渉によって最終的に決定されたイベントの日付を抽出する手法を提案する．

## 2　提案手法による処理の流れ

提案する手法の処理の全体図を図 1 に示す．

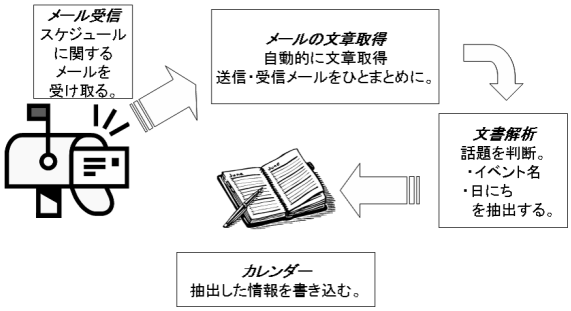


図 1: システムの全体図

システム処理のながれは大きく分けて三段階となる．

**step1. メーラーからメール本文取得**
スケジュールに関してやりとりを行ったメールを取得する．

**step2. 文章解析**
**step1.** において取得されたメールの文章を解析し，話題ごとにカレンダーに書き込むための必要情報を抽出する．

**step3. カレンダーへの書き込み**
**step2.** で抽出された情報をカレンダーに書き込む．

本研究では提案する枠組みの中核となる **step2. 文書解析**について主に取り扱う．

## 3　メール文書解析

### 3.1　メールのタイプ分類

スケジュールを決定するメールを本研究では内容の記述表現に関する特徴に基づき，以下の二つのタイプに分類して考える．

- **お知らせタイプ**
  一方的にスケジュールを知らせるタイプ．
  特徴：日付・イベント名などが明記されている
- **対話タイプ**
  二人のユーザ間でスケジュールを決定するタイプ．
  特徴：
  - 返信メール中に，自分が送った文章が，リダイレクト（>）などの記号を使って相手のメッセージの中に入る．
  - 話題を変える際には 1 行以上の改行を挟む．
  - 相手の問いかけに対する返答には相手が使っている語を使う傾向がある．

「対話タイプ」のメールから必要な情報の抽出が出来るようになれば,「お知らせタイプ」のメールについても同じ手法が適用できると考え，本研究では「対話タイプ」のメールを対象に最終的なイベントの日付を抽出する．

### 3.2　文書解析の全体の流れ

「対話タイプ」のメールでは，メールのやりとりをする際にスケジュール以外の話題も同時に記述することがある．そのため，本研究では交換されたメールを分析し，同一の話題について記述しているものごとに一つにまとめ，そこから日付・イベント名といった必要情報を抽出する手法を提案する．

### 3.3　同一メール内の話題分割

話題分割の方法として，ベクトル空間法を用いて話題分割を行う手法がある [4]. しかし「対話タイプ」のような談話構造を持つメールには適していないため，本研究では「メールの話題を変えるときには見やすくするために，改行をはさむ」などの内容記述表現に関するヒューリスティック知識を利用した分割方法を行う．

### 3.4　分割された文書の同一話題結合

話題ごとに分割した文章は，同じ話題どうしを結合し，一つの文章にする．同一話題の判定は文書ベクトルの類似度で判断する．本研究では文書を特徴付ける語彙として名詞に注目し，それらをベクトルの成分としている．名詞抽出には形態素解析システム「茶筅」 [3] を用いた．

一番高い類似度の文書を同じ話題の文書と判断し，文書を結合する．もし，どの話題とも類似していなければ，そのメールのみの話題として結合は行われない．具体例を図 2 に示す．

#### 3.4.1　話題ごとに必要情報を抽出

同一話題どうしを結合した文書に対し，正規表現を使用し，日付・イベント名を取り出す．イベント名抽出には，カレンダーに登録したいイベント名をあらかじめファイルに保存しておき，その情報に基づき抽出を行なう．指定したイベント名が文書中に存在しなければ,

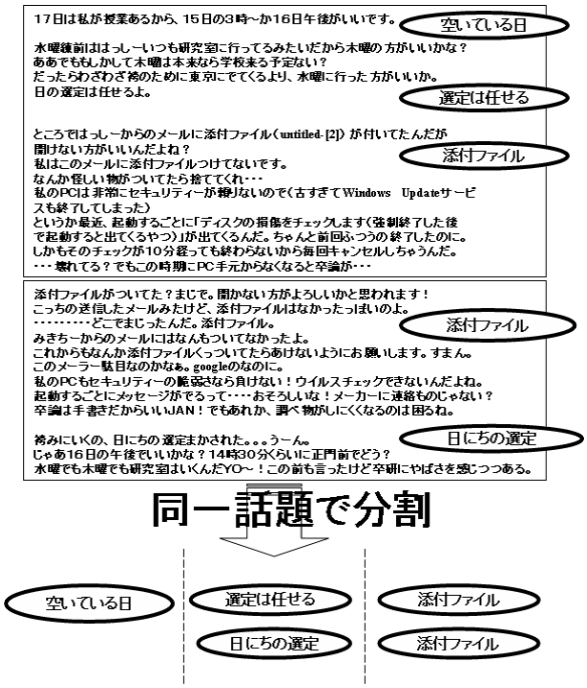


図 2: 話題ごとの分割と結合

他のイベントの情報があったとしてもカレンダーに入力すべきスケジュール情報ではないと判断され，その話題は破棄される．また，指定したイベント名があったとしても，日付情報が全く存在しない場合にはスケジュールに関連しない話題と判断され，同様に破棄される．

日付については「7 日」のような絶対表現以外に，表 1 のような相対表現がある [5, 6]．このような表現に一致した場合も，日付情報として取り扱う．カレンダーに登録する月に関しては，文書中に「月」という語があれば取り出し，無ければメール送信月を採用する．また，日付情報抽出の際には，最終的に決定されたスケジュールの日付のみを抽出する必要がある．本研究では，メール対話の一番最後に出現する日付情報を，最終的に決定されたスケジュールの日付と判断し，カレンダーに登録する日付として採用した．

表 1: 日付表現の置換表

| 相対表現 | 絶対表現 |
|---|---|
| 明日 | 送信日+1 |
| 次の日 | 送信日+1 |
| 来月の | 送信月+1 |
| 来週のＸ曜日 | 送信日の次週のＸの日付 |

# 4 結果と考察

実際に提案手法を基に構築したシステムにより，イベントに無関係な対話も含めたメールに対して処理を行った．その結果を表 2 に示す．

対話タイプメール 147 通（26 対話），お知らせタイプメール 55 通の内，指定したイベントのイベント情報を抽出できたのは，対話タイプが 23 対話，お知らせタイプが 49 通であった．誤答が起きた理由を以下に記す．

表 2: 提案手法によるイベント情報抽出正解率

| メールタイプ | 使用したメール数 | 正解率 |
|---|---|---|
| 対話タイプのメール | 147 通（26 対話） | 88.5% |
| お知らせタイプのメール | 55 通 | 89.0% |
| 計 | 202 通 | 87.2% |

- 最後に出現する日付が，イベント情報に無関係であった為，誤った日付を抽出した．(お知らせタイプ：5 通　対話タイプ：2 対話)
- 一つのイベントに関して，日付に関する話題が分割されてしまっ為，正解のイベント情報と共に誤ったイベント情報を抽出した．(対話タイプ: 1 対話)
- イベント名がメールに含まれていない為，抽出不可能だった．(お知らせタイプ：1 通)

文章の前後関係や係り受け関係等を把握し，分割方法を更なるヒューリスティックで正確に行なえるようになれば，さらに精度をあげることが出来ると考える．

# 5 まとめ

本研究では，電子メールにより，スケジュールを決定する際に最終決定された日付をカレンダーに入力する手間を省くため，電子メールの内容記述表現に着目し，話題ごとにメール文書を分割・結合することにより，あらかじめ設定しておいたイベントに対して，日付やイベント名といった必要情報を抽出する手法を提案した．本研究では，話題分割をヒューリスティック知識のみで行っている．そのため，一つの話題しかないメールに対しても改行があれば別の話題として扱ってしまうという欠点があるが，改行によって分割された各パラグラフは，相手のメールの分割された全てのパラグラフに対して，類似度の高いものと結合を行うため，結果的に一つの話題として扱うことが出来る可能性が高く，また，実験結果の正解率からも提案手法の有効性が確認で出来る．この点においては，文書内の手がかり語や語彙の結束関係を捉えるなど，より正確な話題分割方法を使用すれば精度が上がることが期待される．また,「対話タイプ」のメールのみ取り扱ったが，本研究で必要情報の抽出方法は「お知らせタイプ」のメールにも応用できる．これらのことから話題結合方法に関しては，文書ベクトルを使用した類似度に基づいており，元のメールに対して引用返信を行っている場合にとても有効といえる．

# 参考文献


[1] http://mail.google.com/mail/

[2] http://code.google.com/apis/gdata/calendar.html

[3] 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座 (松本研究室), http://chasen.naist.jp/hiki/ChaSen/

[4] 内海慶, 藤井敦, 田中和世. 分析区間長を可変としたテキスト分割手法. 言語処理学会第１２回年次大会,D1-8,2006.

[5] 土田誠司, 奥村紀之, 渡部広一, 河岡司. 連想メカニズムを用いた時間判断手法. 自然言語処理 Vol.12 No.5, pp.111-129, 2005.

[6] 長谷川隆明, 高木伸一郎, 電子メールコミュニケーションにおけるスケジュール情報抽出. No.1997-NL-123-10,1997.