概要

Web から情報の収集が行われている．そのソースは，ブログ，掲示板，SNS などである．先行研究として，石野らはブログエントリから，パターンと機械学習を用いて，お土産情報として，お土産名，および，その観光名所名のペアの抽出を行った [1]．一方，掲示板や SNS を対象とすると，1 つ 1 つの書き込み (以降レスと呼ぶ) は短く，情報が小出しになっている．そのため，幾つかのレスをつなげて読むことで情報を得るという方法をとる必要がある．

そこで，本研究ではインターネット掲示板からお土産情報の抽出を行う．特に，複数のレスから情報を集約することを目的とする．そのために，まず，明示的なレスの宛先を利用してスレッドの構造を有向グラフで表す．次に，各レスから情報を抽出する．最後に，有向辺をたどり，情報の集約を行う．実験では，集約の有無による性能の違いを確認をする．掲示板「2 ちゃんねる」のレス 100 件を対象として「商品名」,「場所名」,および,「評価情報」の 3 つ組情報を抽出して，情報の集約を行った．人手により作成した正解データとの比較において，集約を行わない場合は，適合率 0.22，再現率 0.50，および，F 値 0.27 であったところ，集約を行った場合は，適合率 0.24，再現率 0.55，および，F 値 0.30 となった．こうして，集約による抽出性能の向上を確認した．

目 次

図 目 次

表 目 次

第1章　はじめに

Web から情報の収集が行われている．観光情報においては網羅性が高く，最新の観光情報を得られることや，ローカルな口コミ情報を得られることなどから注目されている．

そのソースはブログ，掲示板，SNS などである．先行研究ではブログエントリを対象として「旅行ブログエントリからの観光情報の自動抽出」が石野らにより行われている．

一方で，最近ではマイクロブログと呼ばれる情報源も注目されている．掲示板やマイクロブログは1つ1つの書き込み (以降**レス**と呼ぶ) は短く，情報が小出しになっている．そのため，幾つかのレスをつなげて読むことで，情報を得るという方法をとる必要がある．

そこで，本研究では掲示板を対象にして，お土産情報の抽出というタスクの実現を目的とする．特に，小出しにされた情報の集約に焦点をあてる．

レスをつなげて読むために，まず，各レスにおける宛先を決定しなければならない．レスには明示的に宛先を示すことがある．これを利用することで，レスの宛先を決定することができる．次にレスからの情報抽出を行い，レスごとのお土産情報を得る．最後に，宛先を利用して各レスの情報を集約することでより正確な情報を抽出することができると考えられる．

実験では掲示板「2 ちゃんねる」を対象として，宛先の明示されたレスの割合を確認し，レスからの情報の抽出とその集約の性能の評価を行う．

最後に，実験結果を考察することで誤り解析と今後の課題を述べる．

第2章　関連研究

本章では，まず，観光情報の抽出に関する先行研究を紹介する．観光情報のうち，お土産に関する情報を本研究では扱う．そこで，お土産の名称などの固有表現抽出や評判情報抽出といった要素技術について次に紹介する．最後に，これらの関連研究に対して本研究の課題を述べる．

2.1　観光情報の抽出

ブログを情報のソースとして，観光情報を抽出する先行研究には，石野らの研究がある [1]．石野らは “旅行ブログエントリからの観光情報の自動抽出” において日記形式で綴られた旅行ブログエントリに焦点をあて，ブログデータベースから旅行ブログエントリを検出した．そこから観光情報として土産物情報および観光名所情報を抽出する手法を提案した．さらに，旅行ブログエントリからリンクを抽出することで，観光情報のリンク集を構築した．

一般ブログから旅行ブログエントリ (旅行について記述されたブログの 1 記事を旅行ブログエントリと呼ぶ) を検出し，観光情報を抽出する情報源としている．旅行ブログエントリの検出方法として，機械学習の CRF を用いている．

土産物情報，観光名所情報の抽出には，表層パターンと機械学習を用いている．土産物リスト (地域名と土産物が対となったリスト) と観光名所リスト (地域名と観光名所が対となったリスト) を Google から提供されている “Web 日本語 N グラム” データベースに表層パターンを当てはめ自動抽出を行う．このデータベースは Web 上に存在する 20 億文から抽出された N グラム (N=17) で構成されている．使用している表層パターンを以下に示す．

- 「地域名」名物「土産物」(例:東京名物東京バナナ)
- 「地域名」にある観光名所「観光名所名」(例:和歌山にある三畳敷)

抽出の結果，土産物リストには 482 対，観光名所リストには 35,827 対登録された．

機械学習の素性としては “旅行”，“ツアー” といった手掛かり語 416 個である．それらが各エントリに含まれるかどうかを機械学習器に与えている．

実験において土産物情報の抽出は，以下の式で評価されている．

$$\frac{\text{正しく抽出された地域名と (土産物 } or \text{ 観光名所) の対}}{\text{抽出された地域名と (土産物 } or \text{ 観光名所) の対}} \times 100[\%]$$

旅行ブログにおいて，地域名と土産物の対の抽出は 74.0%，地域名と観光名所の対の抽出は 71.0%という結果を示した．

2.2 個別的な情報の抽出

2.2.1 固有名の抽出

日本語固有表現の抽出における先行研究として福島らの研究がある [2]．福島らは大規模なウェブコーパスから固有名リストという形式で知識を収集し，そのリストを素性として系列ラベリングのモデルに取り込むことで，固有表現抽出の精度を向上させる手法を提案した．

固有表現抽出における精度の評価として CRL データセットを利用している．固有名詞と数値表現を区別せずに計算した F 値において，89.20%とベースラインの 89.01%を 0.28%上回っている．

2.2.2 場所名の解析

固有表現抽出において，形態素解析器 “MeCab” [4] を利用する方法がある．場所名を抽出するにあたり，“MeCab” では地名に「地域」の要素を出力するため，一般的な県名や地名であれば，“MeCab” を利用することで「場所名」の抽出とすることができる．“MeCab” に例文として「食いだおれの街、大阪。」を入力した場合の出力例を図 2.1 にのせる．

2.2.3 評判情報の抽出

Web から評判情報を抽出する先行研究に高尾らの研究がある [3]．高尾らは Web 全体から自動的かつ正確に飲食店舗の評判情報の抽出を行った．Web ページを評判情報であ

食いだおれの街、大阪。
食い 動詞, 自立,*,*, 五段・ワ行促音便, 連用形, 食う, クイ, クイ
だ 助動詞,*,*,*, 特殊・ダ, 基本形, だ, ダ, ダ
おれ 名詞, 代名詞, 一般,*,*,*, おれ, オレ, オレ
の 助詞, 連体化,*,*,*,*, の, ノ, ノ
街 名詞, 一般,*,*,*,*, 街, マチ, マチ
、 記号, 読点,*,*,*,*,、,、,、
大阪 名詞, 固有名詞,地域, 一般,*,*, 大阪, オオサカ, オーサカ
。 記号, 句点,*,*,*,*,。,。,。

図 2.1: 形態素解析例

るページと評判情報でないページ (非評判情報と呼ぶ) に分類することで飲食店舗の評判情報を抽出した.

分類は共起情報を用いた上で SVM を利用している．結果として，共起情報と SVM を単独で使用した場合よりも精度が向上している.

2.3 本研究の課題

Web からの観光情報の抽出を行う場合，ソースの種類として，ブログだけではなく，掲示板やマイクロブログも考えられる．掲示板やマイクロブログでは，レスからの情報が断片的であるので，正確な情報を抽出するためには情報の集約が必要である.

そこで，お土産名や地名の固有表現の抽出，ならびに，お土産などの評判情報の抽出については個別に抽出できることを前提として，本研究では，掲示板のレスからこれらの断片化された情報を集約することに焦点をあてる.

第3章　情報抽出の方法

本章では，まず，掲示板の構造について述べる．次に，掲示板からの情報抽出の方法を提案する．

3.1　掲示板の構造

掲示板「2ちゃんねる」は，大規模で，即時性が高く，読む価値のある情報がある程度存在する．書き込まれる話題はスレッドという単位で分かれている．スレッドは1,000件以下のレスで構成される．レスは，書き込みの本文だけでなく，通番 (以降**レス番号**と呼ぶ) が付与されている．

あるレスに対して同意や反論などのレスを書き込む際，レスの宛先をレス番号で明示することがある (図3.1)．レス番号をたどることで，小出しにされた情報を集約することができると予想される．

4 ：名無しさん@お腹いっぱい。
喜八洲のキンツバ
5 ：名無しさん@お腹いっぱい。
>>4
http://www.kiyasu.jp/s-n2.html
これも美味しいですね！
親戚の集まりでたまに食べたりしますが、
お土産リストの中には入っていませんでした。
お店もいくつかあって買いやすいのも良いですね。

図 3.1: 宛先の明示されたレスの例

3.2　情報抽出の提案

情報抽出を以下の手順で行う．

- 手順 1:スレッドの構造解析
- 手順 2:レスからの情報抽出 (断片的情報の抽出)
- 手順 3:情報の集約

次節以降に詳細を述べる.

3.2.1 スレッドの構造解析

スレッドの構造解析では，まず，スレッドをレスの単位に分割し，次に，レスとレスの関係を解析する．最後に，これらを有向グラフの形式にまとめる.

レスの単位に分割するために，正規表現を用いる．レスの開始行の特徴は「レス番号:ハンドルネーム 記述された日時 ID 番号」という形式で記述される．レスの終了は次のレスの開始をもって判定することができる．ただし，削除されたレスがあり，それは「レス番号:あぼーん:あぼーん」と記述される．以上を考慮すると，スレッドのレスの分割ができ，各レスから，レス番号，および，本文を抽出できる.

レスとレスの関係の解析とは，レスはある話題について記述する際に，同様の話題に関するレスに対して同意，反論，補足などの意図をもって書かれることがある．意図の宛先 (以降**レスの宛先**と呼ぶ) は，本文に明記されることがある．その形式は「>>レス番号」が基本となり，主に 7 通りがよく用いられる．以下に用いられる例を示す.

1. 「>>」+「レス番号 (半角)」の組み合わせ (>>2)
2. 「＞」+「レス番号 (全角)」の組み合わせ (＞ 2)
3. 「>>」+「レス番号 (半角)」+「さん」の組み合わせ (>> 2 さん)
4. 「>」+「レス番号 (半角)」の組み合わせ (>2)
5. 「>>」+「レス番号 (半角)」+「-」+「レス番号 (半角)」の組み合わせ (>>2-3)
6. 「レス番号 (半角)」+「さん」の組み合わせ (2 さん)
7. 「>>」」+「レス番号 (半角)」+「>>」」+「レス番号 (半角)」複数宛先を持ち，連続で記述 (>>2>>3)

そこで，この形式を参照してレス間のつながりを解析する．

有向グラフは，各レスを頂点とし，レスの宛先の関係を有向辺とする．頂点には，レスに存在する情報が対応付けられる．辺については，始点を宛先となるレス，終点を同意や反論を行ったレスとする．辺の向きは断片的な情報が伝播する方向を表す．

3.2.2 レスからの情報抽出

各レスから単独で得られるお土産情報を抽出する．お土産情報は以下の 3 種類とし，3 つ組の形式で，グラフのレスごとに記録する．

- 商品名 (souvenir)
- 場所名 (location)
- 評価情報 (evaluation)

ここで，得られない情報は，ϕ と記載する．

図 3.1 から得られる 3 つ組を表 3.1 に示す．レス 4 では,「商品名」として「キンツバ」, 場所名として「喜八洲」が得られる.「評価情報」は記述されないため 3 つ組としては (キンツバ，喜八洲，ϕ) が作られる．レス 5 では「商品名」と「場所名」は得られず,「評価情報」として「美味しい」と「お店もいくつかあって買いやすいのも良い」という 2 つの情報が得られるため 3 つ組として (ϕ，ϕ，美味しい)，(ϕ，ϕ，お店もいくつかあって買いやすいのも良い) の 2 つが作成される．

表 3.1: レスから抽出した 3 つ組情報の例 (断片的情報)

レス番号	3 つ組情報
4	(キンツバ, 喜八洲, ϕ)
5	(ϕ, ϕ, 美味しい)
	(ϕ, ϕ, お店もいくつかあって買いやすいのも良い)

3.2.3 情報の集約

スレッドの構造を表すグラフの辺をたどることで，断片的な情報を集約する．

辺の始点にある 3 つ組情報 $\boldsymbol{t_1}$ を，辺の終点にある 3 つ組情報 $\boldsymbol{t_2}$ に重ねて，集約された 3 つ組情報 $\boldsymbol{t_3}$ を作成する (式 3.1)．なお，$t_{i,j}$ は $\boldsymbol{t_i}$ における j 番目の要素を表す．

$$t_{3,j} \underset{(j=1,2,3)}{=} \begin{cases} t_{1,j} & (t_{2,j} = \phi \text{のとき}) \\ t_{2,j} & (t_{2,j} \neq \phi \text{のとき}) \end{cases} \tag{3.1}$$

集約された情報の例を表 3.2 に示す．レス 5 において，レス 5 からは得られなかった「商品名」および「場所名」の情報は，宛先であるレス 4 を参照することで，新しく (キンツバ, 喜八洲, 美味しい)，(キンツバ, 喜八洲, お店もいくつかあって買いやすいのも良い) というより正確な情報となる．レス 5 の例では 3 つ組情報が 2 つ集約して作成された．このように，情報の集約は基本的にはレスの単位で定まる．

表 3.2: 集約された 3 つ組情報の例

レス番号	3 つ組情報
4	(キンツバ, 喜八洲, ϕ)
5	(キンツバ, 喜八洲, 美味しい)
	(キンツバ, 喜八洲, お店もいくつかあって買いやすいのも良い)

第4章　実装

本章では，掲示板のスレッドを入力するとお土産情報の抽出と集約を行うシステムの実装について説明する．

4.1　システムの概要

システムは大きく分けて 3 つの処理に分かれている．

- レスの分割および宛先の決定
- レスからの情報抽出
- 抽出した情報の集約

4.2　実行条件

スレッドは，全てのレスを Web ブラウザで表示した上で，全選択とコピー＆ペーストにより取得する，テキスト形式とする．始まりと終わりの行から，レス以外の表現を削除する．

辞書データは，テストデータのスレッドから人手により「商品名」,「場所名」と判断できる名詞をリストとして作成する．

今回実装に利用したプログラミング言語は Ruby1.8 である．

4.3　システム

4.3.1　システム構成図

システムの構成図を図 4.1 に示す．次節以降で各処理部の詳細を述べる．

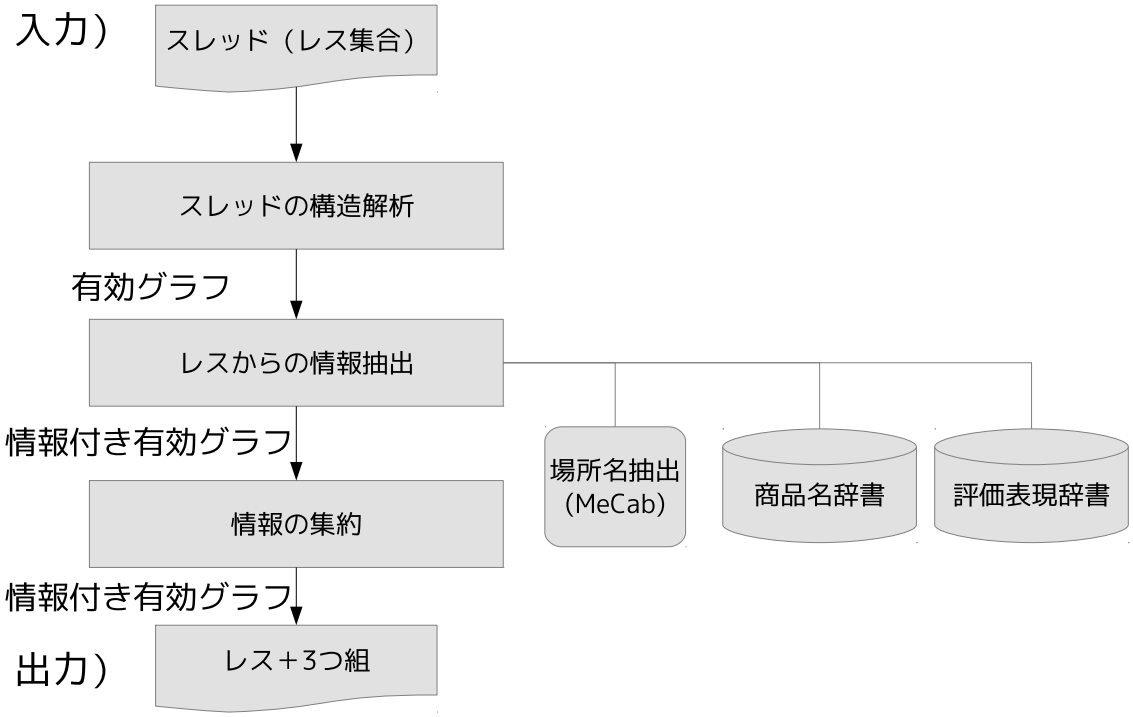

図 4.1: システム構成図

4.3.2 レスの分割および宛先の決定

レスの分割においては，取得したテキスト形式のスレッドに対して，正規表現を用いてプログラムにより，自動的にレス番号ごとに分割を行う．

実際のスレッドのレスと使用した正規表現を図 4.2，および図 4.3 に示す．

各レスに分割をした後，レスに対して 3.2.1 節にて説明した「記号」+「レス番号」を利用して各レスの宛先を決定し付与を行う．

4.3.3 レスからの抽出

レスからの情報の抽出を行うために,「商品名」,「評判情報」については，人手により，あらかじめレスから情報の抽出を行い，正解データとしてリスト化をした．

リスト化する際,「商品名」と判断した定義をは次の通りである．

- 地名，人名ではない固有名詞
- 店名+一般名詞

1 ：名無しさん@お腹いっぱい。：2005/11/09(水) 14:23:47 ID:193bCBjj0
食いだおれの街、大阪。
しかしコレと言ったお土産がないような。
たこ焼きや豚まんも、家族への土産には良いかもしれませんが
会社やちょっとした知り合いなどに渡せる、こじゃれたお土産が欲しい。
お勧めのお土産を教えてください。
2 ：名無しさん@お腹いっぱい。：2005/11/09(水) 14:29:06 ID:193bCBjj0
言い出しっぺの私のお勧めは、
サントリー山崎蒸留所で売っているウィスキーのケーキです。
http://www.suntory.co.jp/whisky/factory/yamazaki/guide/factoryshop.html#factoryshop6
南部の人には不便ですし、出張帰りに買えるわけではありませんが、
おいしいですし、大阪人の自分が遠方に行くときにはコレを持って行っています。

図 4.2: 実際のスレッド

```
^(\d+)\s：(.*)：(\d{4}\/\d\d\/\d\d\([月火水木金土日]\))
                    (\d\d:\d\d:\d\d)(?:\.\d\d)? ID:(.+)$
```

図 4.3: 使用した正規表現

- 一般的な商品名

実際に作成した商品名辞書を図 4.4 に示す．総商品名数は 86 件である．

評価情報についても同様に人手によりリスト化を行う．評価情報として判断した定義は次の通りである．

- 商品に対する感想
- 商品に対する情報

図 4.5 に実際に作成された評価表現辞書を示す．評価表現数の総数は 78 件であった．

作成したリストを用いて，一致した情報を抽出するプログラムを作成した．これにより，各レスにおける「商品名」と「評価情報」の抽出ができた．

「場所名」の抽出は 2.2.2 節にて述べた “MeCab” を用いて,「地域」となる要素を場所名として抽出するプログラムを作成した．

各レスの「商品名」,「場所名」,「評価情報」，および「ϕ」，から全ての組み合わせで 3 つ組を作成する．図 4.6 に出力 3 つ組例に使用したレスの一部を，図 4.7 に実際に作成された 3 つ組の例を示す．

たこ焼き	肉まん	米
豚まん	餃子	ムジカの紅茶
ウィスキーのケーキ	じゃがりこ	赤福
キンツバ	きんつば	神宗の昆布
こしぬけうどん	小エビの天ぷら	泉州の水茄子漬け
丸福珈琲のゼリー	あんプリン	泉州タマネギ
堂島プリン	粟おこし	大阪みやげ
今井のおうどん	みたらし団子	大阪バナナ
ツマガリのクッキー	チーズケーキ	タコ焼き
生菓子	御座候	鉄板付きのたこ焼
丸福珈琲	通天閣コーヒー	リクロおじさん
珈琲	高級佃煮	まつのはこんぶ
ケーキ	明石焼き	京都の抹茶プリン
堂島プリン	トンカツ定食	バターサンド
たこまさのたこやき	焼き餅	北海道六花亭マルセイバターサンド
てんてんの餃子	タコ焼き羊かん	タマネギ
モノレールうどん	たこ焼き饅頭	茜丸のどら焼き
五色どらやき	酒饅頭	林檎
点天の餃子	喜矢州	檸檬
水ナス漬け	パチパチパンチ	若ごぼう
田辺大根	大阪の臭い水道水	ハチエモン
お好み焼き	ミネラルウォーター	観音屋のチーズケーキ
いかやき	たこ焼	デンマークの生のチーズ
茜丸	洋菓子	カステラ生地
焼き菓子	ロールケーキ	自由軒の卵カレー
ロールケーキ	乾き物	インデアンカレー
たこべぇ	ヒロタのシュークリーム	大阪寿司
パイン飴	明石焼き	サザエボン
御堂筋弁当	水	たこ焼きプリッツ
	お米のルーロ	

図 4.4: 商品名辞書

4.3.4 情報の集約

各レスにおいて 3 つ組を作成した後，4.3.1 節で決定したレスの宛先を利用して情報の集約を行う．

幾つかのレスの例を図 4.8 に示す．レス 16 について，3 つ組は表 4.1 のようになる．表から，情報の集約を行わない場合は「場所名」として「東京」が出力されるのみで,「商品名」にあたるものが何であるかが分からない．しかし，情報の集約を行うことで,「茜丸の五色どらやき」が「東京」にあるということが情報として得られる．

このことからレス 16 では断片的な情報が，間に存在するレス 15 の影響を受けずに集約されていることが確認できる．

家族への土産には良い
こじゃれたお土産
お勧めのお土産
南部の人には不便
出張帰りに買えるわけで
はありません
お勧め
お店もいくつかあって
買いやすいのも良い
重くて
クール便で送る
のがいい
お土産向き
ネーミングだけではなく
、味も評価高い
良さそう
女性が多いところへの
お土産だとかなり喜ばれそう
上品
まずい
あんまりいない
ギネス
長い
ウケ狙い
おススメ
チョコ味
日持ちする
美味しい
冷凍食品として売ってる
普通
便利

ウマー
好き
臭ス
ちょっとだけ辛い
小さい
困ったら
不評
手軽
安い
味もソコソコ
ツッコミ期待
日持ちの面で微妙
うけた
良し悪しはわからない
お歳暮
美味い
バターで焼いてて
中にこんにゃく
冷めてもウマー
会津◯なんて食えねぇ
マイナー
大好評
死ぬほど不味かった
美味しかった
有名
悪名高い

匂いはまし
コピペ
見た目も綺麗
味おちてない？
おいしかったのは昔の話
味が劣化
悲しい
ウマー
めいぶ〜つ♪
一番の御馳走
安くて美味しい
大層気に入られた
ウマー
味が似てます
名物
砂糖で甘すぎで美味しくはない
香料臭くてたまらん
おすすめは絶対しない
食べられない
冷蔵状態でないと持ち歩くのは難しい
微妙
結構好きな味
ロクなものがねえや
個包装
配りやすい
インパクトがあって
食える

図 4.5: 評価表現辞書

表 4.1: レス 16 における集約の様子

集約の有無	3 つ組情報
無	(ϕ, 東京, ϕ)
有	(茜丸の五色どらやき, 東京, ϕ)

24 ：名無しさん@お腹いっぱい。：2005/12/06(火) 23:29:18 ID:3e4m5sq50
茜丸別に普通。
ツマガリは大丸梅田などにも入っていて便利です。
焼き菓子 (゜д゜) ウマー
ロールケーキとかも一個から買えるし、どれもおいしい。
25 ：つまらないものですが名無しです：2005/12/14(水) 12:25:35 ID:ckOb3KTT0
阪神デパートの近くとかで、全国の御土産が売ってあるところ
あれって儲かっているの?
買っている人ほとんど見たこと無いんだけど
26 ：つまらないものですが名無しです：2005/12/16(金) 11:07:23 ID:EuUVBeex0
たこべぇ好きｄ
27 ：つまらないものですが名無しです：2005/12/18(日) 21:32:49 ID:9lWWdmQKO
>>25
以前買いに行った時、沖縄の所で、パイン飴を大量に包ませてたリーマンが
いたが、あれはなんだったんだろ。

図 4.6: 出力 3 つ組例に使用したレス (一部)

24 (ケーキ, 梅田, 普通);(ケーキ, 梅田, 便利);(ケーキ, 梅田, ウマー);
(ケーキ, 梅田,ϕ);(ケーキ,ϕ, 普通);(ケーキ,ϕ, 便利);
(ケーキ,ϕ, ウマー);(ケーキ,ϕ,ϕ);(茜丸, 梅田, 普通);(茜丸, 梅田, 便利);
(茜丸, 梅田, ウマー);(茜丸, 梅田,ϕ);(茜丸,ϕ, 普通);(茜丸,ϕ, 便利);
(茜丸,ϕ, ウマー);(茜丸,ϕ,ϕ);(焼き菓子, 梅田, 普通);(焼き菓子, 梅田, 便利);
(焼き菓子, 梅田, ウマー);(焼き菓子, 梅田,ϕ);(焼き菓子,ϕ, 普通);(焼き菓子,ϕ, 便利);
(焼き菓子,ϕ, ウマー);(焼き菓子,ϕ,ϕ);(ロールケーキ, 梅田, 普通);(ロールケーキ, 梅田, 便利);
(ロールケーキ, 梅田, ウマー);(ロールケーキ, 梅田,ϕ);(ロールケーキ,ϕ, 普通);
(ロールケーキ,ϕ, 便利);(ロールケーキ,ϕ, ウマー);(ロールケーキ,ϕ,ϕ);(ロール, 梅田, 普通);
(ロール, 梅田, 便利);(ロール, 梅田, ウマー);(ロール, 梅田,ϕ);(ロール,ϕ, 普通);
(ロール,ϕ, 便利);(ロール,ϕ, ウマー);(ロール,ϕ,ϕ);(ϕ, 梅田, 普通);(ϕ, 梅田, 便利);
(ϕ, 梅田, ウマー);(ϕ, 梅田,ϕ);(ϕ,ϕ, 普通);(ϕ,ϕ, 便利);(ϕ,ϕ, ウマー)
25 (ϕ,ϕ,ϕ)
26 (たこべぇ,ϕ, 好き);(たこべぇ,ϕ,ϕ);(ϕ,ϕ, 好き)
27 (パイン飴, 沖縄,ϕ);(パイン飴,ϕ,ϕ);(ϕ, 沖縄,ϕ)

図 4.7: 出力 3 つ組例

14 ：名無しさん@お腹いっぱい。
茜丸の五色どらやきは？
15 ：名無しさん@お腹いっぱい。
大阪じゃなく神戸だが …
http://konigs-krone.co.jp/
がおすすめ。チョコ味が良いよ！結構日持ちする。
16 ：名無しさん@お腹いっぱい。
>>14
最近は東京の am/pm でも売られている。

図 4.8: 入力したレスの一部

第5章　実験

本章では，掲示板からのお土産情報の抽出とその集約について，提案手法の性能を評価するために実験を行う．

5.1　実験の目的

個々の情報の抽出が行われることを前提とし，断片的な情報の集約が，提案手法により実現できる割合を，実験により確認する．したがって，集約の有無により各レスから得られる情報が，理想的な情報と一致する割合で評価する．これに関して，レス間の関係の割合を参考情報として収集しておく．

5.2　実験の条件

実験に用いるスレッドは「2ちゃんねる」の「みやげ物・特産物板」といわれる板 (スレッド集合) から 2012 年 7 月に取得した.「大阪のお土産」をタイトルとするスレッドである．総レス数は，595 件であった．本実験では，テストデータとしてレス 1〜100 番までのレスを対象とする．

正解データ，すなわち，集約で得られるべき 3 つ組は，レスごとに手作業で定める．1 つのレスから 0 件以上の正解データを定めることができる.「大阪のお土産」においては，100 レス中 228 件であった.

情報抽出の処理の際，商品名，評判情報についての表現は，あらかじめ辞書化しておいたものを利用する．場所名については，MeCab[4] を利用して形態素解析結果が「地域」となっているものを場所名として利用する．したがって，商品名の固有表現抽出および評価表現抽出は単独でみるとクローズドな実験となっている．

5.3 宛先の明示されたレスの割合

実際に対象となるレスがどれだけあるかを確認する．宛先が明示的に示されている割合を表 5.1 に示す．

表 5.1: 宛先の明示されたレスの割合

スレッド	明示レス数	レス総数	割合 (%)
大阪のお土産	21	100	21

5.4 実験の進め方と評価方法

テストデータを用いてプログラムによりレスの宛先をデータ構造化し，用意した辞書を利用して，レスからの情報抽出を行い，情報を集約する．情報の集約を行わなかった結果と情報の集約を行った結果を出力し，正解データとそれぞれ比較することで，集約の効果を確認する．

評価方法は，各レスごとに正解データの 3 つ組と出力の 3 つ組を比較し，出力数，一致数を出す．さらに，出力数および一致数から以下の式により適合率，再現率，および F 値を計算する．

$$\text{適合率} = \frac{\text{理想的な 3 つ組と出力との一致数}}{\text{出力数}}$$

$$\text{再現率} = \frac{\text{理想的な 3 つ組と出力との一致数}}{\text{理想的な 3 つ組数}}$$

$$F \text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$$

各レスごとの適合率，再現率，および F 値を求めて，それらからマクロ平均を計算し，性能の評価として利用する．

マクロ平均とは，各レスの適合率，再現率，および F 値を計算してから，それらを平均することである．すなわち $\frac{1}{N}\sum_{i=1}^{N} x_i$ で求める．ここで，N はレス数，x_i は i 番目のレスにおける，適合率，再現率，または F 値である．

なお，全てのレスの出力数，一致数の合計を計算して，それらから適合率，再現率，および F 値を計算するのがマイクロ平均である．

5.4.1　3 つ組情報の抽出性能

結果を表 5.2 に示す．各レスでの適合率，再現率および，F 値を求め，そのマクロ平均を示している．集約の効果が確認できた．

表 5.2: 3 つ組情報の抽出性能

集約の有無	出力数	一致数	適合率	再現率	F 値
無	1,022	99	0.22	0.50	0.27
有	1,519	110	0.24	0.55	0.30

第6章　考察

本章では，実験結果に見られた誤り箇所について考察する．そして，その一部について，理想状態での実験を追加し，今後の展開のための分析とする．

6.1　過剰な3つ組が生成される問題

6.1.1　集約による3つ組の増大する問題

適合率のマイクロ平均を求めると 0.07(=110/1,519) であり，表 6.3 と大きな差がある．これは，特定のレスに 3 つ組が大量に作られたためである．

そこで，過剰な 3 つ組生成の問題を分析しよう．大量に 3 つ組の作成されたレスの例を図 6.1 に示す．お土産情報には，分かりやすくするために下線を引いた．

48 ：つまらないものですが名無しです
地下鉄 大阪$_L$港$_L$駅$_L$ と朝潮橋の間くらいにある 八幡$_L$ 商店街の中にある たこ焼$_S$き$_S$ 屋 美味い$_E$ よ。いつもその辺に住んでいる知合いのところに行くときはそこのを御土産としてもっていく。一応 明石$_L$ 焼って名前で売っているけど、出汁を付けて食うんじゃなくて生地に出汁が練り込まれているヤツ、俺はソースよりこっちの方が 好き$_E$。
49 ：つまらないものですが名無しです
>>48 貰って食べたことがあって、電車を乗りかえしてわざわざ買いに行った。バターで焼いてて$_E$、中にこんにゃく$_E$ が入ってるよね。冷めてもウマー$_{EE}$。あれ食ったら 会津$_L$ ○なんて食えねえ。

図 6.1: 3 つ組大量レス

3 つ組の作成を「商品名$_S$」,「場所名$_L$」,「評価情報$_E$」全ての組み合わせで行うため，3 つ組が過剰に作成されていた．具体的には，レス 48 で出力された数は,「商品名」2 つ,「場所名」5 つ,「評価情報」2 つであった．3 つ組を作る際，それぞれに ϕ を加え，かつ，(ϕ, ϕ, ϕ) を除くので，3 つ組は $3 \times 6 \times 3 - 1 = 53$ 通り作成される．

さらに，レス 49 においては，集約を行う前は「場所名」として「会津」,「評価情報」として「バターで焼いてて」,「中にこんにゃく」,「冷めてもウマー」，および「ウマー」の 4 つから 3 つ組が作成され，組み合わせは 9 通りとなるが，レス 48 との集約により，組み合わせの空欄であった「商品名」および「場所名」にレス 48 からの情報が全ての組み合わせで書き込まれることで組み合わせが増大し，93 通りに増えてしまった (表 6.1).

例えば，t_1 に (たこ焼き，大阪，ϕ)，(たこ焼き，八幡，ϕ) という 2 つの 3 つ組が含まれており，t_2 に (ϕ，ϕ，ウマー)，(ϕ，ϕ，冷めてもウマー) という 2 つの 3 つ組が含まれているので，t_3 は (たこ焼き，大阪，ウマー)，(たこ焼き，八幡，ウマー), (たこ焼き，大阪，冷めてもウマー)，(たこ焼き，八幡，冷めてもウマー) というようにこれらの部分からだけでも 4 つの 3 つ組に増える．

表 6.1: 3 つ組の増加する様子

レス番号	集約の有無	3 つ組数
48	無	53
49	無	9
49	有	93

6.1.2 言語解析力の問題

「商品名」,「場所名」,「評価情報」を抽出する際，全ての組み合わせで 3 つ組を作成するため，1 つのレスに複数の情報が存在する場合 3 つ組が大量に作成されてしまう．

別のテストセットより，典型的な例を図 6.2 に示す.

1 ：名無しさん@お腹いっぱい。
四国の土産について語りましょう。
2 ：駅前
八重山土産だぎゃ～
3 ：名無しさん@お腹いっぱい。
徳島：金長饅頭、鳴門金時、スダチ
香川：うどん、和三盆糖
愛媛：坊ちゃん団子、タルト
高知：カツオ?
高知って何かある?

図 6.2: 複数記述例

表 6.2: 大量な 3 つ組の具体例

レス番号	3 つ組情報
1	(ϕ, 四国, ϕ)
2	(ϕ, 八重山, ϕ)
3	(うどん, 徳島, ϕ);(うどん, 鳴門, ϕ);(うどん, 香川, ϕ);(うどん, 愛媛, ϕ); (うどん, 高知, ϕ);(うどん, ϕ, ϕ);(カツオ, 徳島, ϕ);(カツオ, 鳴門, ϕ); (カツオ, 香川, ϕ);(カツオ, 愛媛, ϕ);(カツオ, 高知, ϕ);(カツオ, ϕ, ϕ); (スダチ, 徳島, ϕ);(スダチ, 鳴門, ϕ);(スダチ, 香川, ϕ);(スダチ, 愛媛, ϕ); (スダチ, 高知, ϕ);(スダチ, ϕ, ϕ);(タルト, 徳島, ϕ);(タルト, 鳴門, ϕ); (タルト, 香川, ϕ);(タルト, 愛媛, ϕ);(タルト, 高知, ϕ);(タルト, ϕ, ϕ); (和三盆, 徳島, ϕ);(和三盆, 鳴門, ϕ);(和三盆, 香川, ϕ);(和三盆, 愛媛, ϕ); (和三盆, 高知, ϕ);(和三盆, ϕ, ϕ);(和三盆糖, 徳島, ϕ);(和三盆糖, 鳴門, ϕ); (和三盆糖, 香川, ϕ);(和三盆糖, 愛媛, ϕ);(和三盆糖, 高知, ϕ);(和三盆糖,, ϕ); (坊ちゃん団子, 徳島, ϕ);(坊ちゃん団子, 鳴門, ϕ);(坊ちゃん団子, 香川, ϕ); (坊ちゃん団子, 愛媛, ϕ);(坊ちゃん団子, 高知, ϕ);(坊ちゃん団子, ϕ, ϕ); (金長饅頭, 徳島, ϕ);(金長饅頭, 鳴門, ϕ);(金長饅頭, 香川, ϕ);(金長饅頭, 愛媛, ϕ); (金長饅頭, 高知, ϕ);(金長饅頭, ϕ, ϕ);(鳴門金時, 徳島, ϕ);(鳴門金時, 鳴門, ϕ); (鳴門金時, 香川, ϕ);(鳴門金時, 愛媛, ϕ);(鳴門金時, 高知, ϕ);(鳴門金時,, ϕ); (ϕ, 徳島, ϕ);(ϕ, 鳴門, ϕ);(ϕ, 香川, ϕ);(ϕ, 愛媛, ϕ);(ϕ, 高知, ϕ)

レス 3 において人による理想的な 3 つ組では一行単位で 3 つ組を作成することで，過剰な 3 つ組の作成を抑えることが可能であるが，全ての組み合わせの場合は 3 つ組が大量に作成されてしまう．

6.2 店名の抽出が不完全である問題

適合率，再現率が低い主な原因としては，場所名の抽出は MeCab の形態素解析による地域名を利用しているため，人手の正解データで使用される「店名」や一部の「場所名」が抽出できないことが挙げられる．

例として，図 6.3 のレス 22 では場所名として「店名」である「阪神百貨店」が使用されている．こうした例は多数みられたため，場所名の抽出を改善すれば適合率，再現率は向上すると考えられる．

図 6.4 に実際に MeCab による出力を示す．MeCab による出力では「地域」とは出力されず,「組織」と出力される．

22 ：名無しさん@お腹いっぱい。
阪神百貨店のいかやき
23 ：名無しさん@お腹いっぱい。
>>22
冷凍食品として売ってるよ

図 6.3: 店名

阪神百貨店のいかやき
阪神百貨店 名詞, 固有名詞,組織,*,*,*, 阪神百貨店, ハンシンヒャッカテン, ハンシンヒャッカテン
の 助詞, 連体化,*,*,*,*, の, ノ, ノ
いか 副詞, 一般,*,*,*,*, いか, イカ, イカ
やき 動詞, 自立,*,*, 五段・カ行イ音便, 連用形, やく, ヤキ, ヤキ

図 6.4: MeCab による出力

6.3 店名の抽出が完全である場合の追加実験

「場所名」を「商品名」,「評価情報」と同様に人手により，辞書を作成した場合の実験を行った.

表 6.3: 3 つ組情報の抽出性能

集約の有無	出力数	一致数	適合率	再現率	F 値
無	1,094	131	0.28	0.65	0.34
有	1,505	168	0.32	0.78	0.40

場所名が正しく抽出できた場合，表 5.2 と比べて，再現率は 0.78 と高い値となる．しかし，過剰な 3 つ組の生成のため適合率は 0.32 と依然低い水準であった.

これより，店名の抽出が不完全である問題より，過剰な 3 つ組の生成が今後の課題として，重要な問題であると考えられる.

第7章　おわりに

本研究では，インターネット掲示板からのお土産情報の抽出というタスクにおいて，掲示板に記述されるレスの宛先を利用することで，複数のレスから情報を集約する手法をとりいれた．

情報を集約する手法として，正規表現を用いてスレッドをレスごとに分割をした．次に，レスの宛先を有向グラフとしてデータ構造にした．そして，各レスからの情報の抽出を行い，有向グラフをたどることで断片的な情報を集約した．

掲示板の 1 つである「2 ちゃんねる」を対象とした実験において，集約を行わない場合の適合率 0.22，再現率 0.50，F 値 0.27 に比べ，集約を行うことで適合率 0.24，再現率 0.55，F 値 0.30 と性能の向上を確認することができた．

残された問題として，過剰に 3 つ組が生成される問題があげられる．具体的には，集約により 3 つ組が増大する問題および，言語解析力の問題があげられた．今後の課題は，過剰に出力された 3 つ組を除去することで性能を向上させることである．

謝辞

最後に，２年間に渡り，本研究のご指導を頂きました鳥取大学工学部知能情報工学科計算機工学講座Ｃ研究室の村田真樹教授，村上仁一准教授，徳久雅人講師に深く感謝すると共に，厚く御礼申し上げます．そして，日常の議論を通じて多くの知識や示唆を頂いた同研究室の皆様方，参考にさせていただいた著書の著者の方々に，感謝の気持ちと御礼を申し上げたく，謝辞にかえさせていただきます．

参考文献

[1] 石野 亜耶，難波 英嗣，竹澤 寿幸: “旅行ブログエントリからの観光情報の自動抽出”, 知能と情報, pp.667-679, 2010.

[2] 福島 健一, 鍛治 伸裕, 喜連川 優: “日本語固有表現抽出における超大規模ウェブテキストの利用”, 第 6 回日本データベース学会年次大会, 2008.

[3] 高尾 美代子, 酒井 浩之, 増山 繁: “Web からの飲食店舗の評判情報抽出”, 言語処理学会第 17 回年次大会発表論文集, pp.268-271, 2011.

[4] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: “Applying Conditional Random Fields to Japanese Morphological Analysis”, Proc. EMNLP, pp.230-237, 2004.