概要

ブログ記事は，近年巨大な情報源として注目されている．ブログ記事は，本文の部と0個以上のコメント部というように大きくブロックの単位で構成されている．本文部には，ブログ著者の持つ情報が記述されており，コメント部には，ブログ著者とブログ閲覧者の対話が記述されている．コメント部の対話には，本文で記述されなかった新たな事柄が追加されている．そのため，ブログ記事からの情報収集では，本文部だけでなくコメント部も参照することが望ましい．しかし，コメント部では，省略された表現が多いため，何に対する追加情報なのかが不明確である．

そこで，本研究では，対象を明確にするため，本文部へのコメントであるか，あるいは，別の先行するいずれのコメント部へのコメントであるかというブロックの単位でのコメント先の解析を行う．関連文の類似度を計算する方法の一つとして，荒牧らは，単語 n-gram に対して Okapi-BM25 を用いた．そこで，関連研究の対応付けの方法の他に，ブログ記事の慣習的特徴を利用する手法，および，ブログの意図伝達に着目する手法を決定リストで組み合わせた手法を利用してコメント先解析システムを作成した．

Ameba ブログからランダムに収集したブログ記事 32 件，ブロック数 255 件，コメント元件数は 224 件を利用して，人手により作成した正解データと解析システムによる出力を比較した．その結果，適合率 0.64，再現率 0.63，F 値 0.64 という性能評価を得た．

目 次

図 目 次

表 目 次

第1章　はじめに

ブログ記事は，近年巨大な情報源として注目されている．一般に，ブログ記事は，本文の部と 0 個以上のコメント部というように大きくブロックの単位で構成されている．本文部には，ブログ著者の持つ情報が記述されており，コメント部には，ブログ著者とブログ閲覧者の対話が記述されている．コメント部の対話には，本文で記述されなかった新たな事柄が追加されている．そのため，ブログ記事からの情報収集では，本文部だけでなくコメント部も参照することが望ましい．しかし，コメント部では，省略された表現が多いため，何に対する追加情報なのかが不明確である．そこで，本研究では，対象を明確にするため，本文部へのコメントであるか，あるいは，別の先行するいずれのコメント部へのコメントであるかというブロックの単位でのコメント先の解析（コメント先を計算機により自動で特定させること）を目標とする．

ブロック単位でのコメント先の解析は，複数文で構成されるもの同士の対応関係，すなわち，記事対応の問題と類似している．池田らは，ニュースについて言及されたブログ記事と，そのニュース記事との対応付けに，ニュース記事の特徴語ベクトルとブログ記事の特徴ベクトルのコサイン類似度を用いた [1]．一方，関連文の類似度を計算する方法の一つとして，荒牧らは，単語 n-gram に対して Okapi-BM25 を用いた [2]．ここで，特徴ベクトルと Okapi-BM25 を用いる方法を比べると，2 つの文に含まれる共通語から特徴度が計算されるという点で共通しているが，Okapi-BM25 の場合，さらに，他の文書と比べた特徴語の出現の仕方が影響するという点で異なる．本研究では，Okapi-BM25 を用いる手法を採用する．

ブログのコメントの特徴について，もう少し考えてみると，次のことが言える．

- コメント先や相手名を明示することについて，慣習的な形式がある．
- 質問-応答，伝達-感謝など意図のやりとりがある．

そこで，関連研究の対応付けの方法の他に，ブログ記事の慣習的特徴を利用する手法，および，ブログの意図伝達に着目する手法が考えられる．そこで，これらを決定リストで組み合わせた手法を，本研究で提案する．

本論文の構成は以下のとおりである．第 2 章では関連研究及び Okapi-BM25 計算式について述べる．第 3 章では提案手法，ならびに，コメント先解析システムについて述べる．第 4 章では人手により作成した正解データとコメント先解析システムによる出力結果を比較して，適合率，再現率，F 値による性能評価を行い，第 5 章にて各手法についての考察を述べる．最後に第 6 章でまとめを述べる．

第2章　関連研究

2 章では本研究と関連する二つの研究及び本研究において使用した Okapi BM25 について述べる．

2.1　ブログとニュース記事の自動対応付け

池田らはブログとニュース記事という内容の性質が異なる文書間の対応付けを行った．ブログには，ニュースを特定できる程度の情報と，主にそれに対する書き手の意見や感想が書かれている．対して，ニュース記事は，タイトルでニュースの全体像が，最初の一文で内容のサマリがそれぞれ書かれている．これらの特徴を利用してブログ中の語全てからブログベクトルを，ニュース記事のタイトルと最初の 1 文から特徴語ベクトルを生成し，両ベクトル間の類似度に基づき，ニュースについて言及しているブログというニュース記事とブログの組を求めた．

2.2　非文法的かつ断片化されたテキストの頑健な分類

荒牧らは電子カルテの一文章が一患者に対応しているという特性より，カルテからの患者の喫煙情報の抽出を，カルテの分類というアプローチで行った．まず，入力文章とトレーニングセットから喫煙に関する文を抽出し，類似度計算の結果より，最も類似した喫煙状況の分類へ分類した．類似度を計る際，尺度として編集距離，$n-gram$ ベース，統語解析の 3 種を用いて，それぞれの確信度と入力文の統計量により，喫煙状況を左右する重要な情報となる語群を手がかりとして，適した尺度を選択している．

編集距離は，S_i を入力文章の喫煙関連文，S_t をトレーニングセットとしての喫煙関連文,$|S_s|$,$|S_t|$ をそれぞれ S_s,S_t の文字数として，式 2.1 により正規化した類似度を算出する．

$$sim_{ED}(S_i, S_t) \quad = \quad \frac{\text{編集距離}\,(S_i, S_t)}{|S_i| + |S_t|} \tag{2.1}$$

$n-gram$ ベースでは文を，単語 $n-gram$ ($n=1..4$) の単位に分解し，分解された語列間の類似度を Okapi-BM25[3] 尺度を用いて計算を行う．最終的な出力は上位 k 個の類似度の重みつき投票により決定する．

統語解析では依存構造上で文を n 語の組み合わせとして分解した後 $n-gram$ ベースと同様の処理を行う．

2.3 Okapi-BM25

Okapi-BM25 は，文書検索に使用されるものであり，クエリ Q に対する文書 D の関連度を順位付ける機能である．次の式で関連度 $score$ を計算する．

$$score(D,Q) = \sum_{q \in Q} s_{BM25}(D,q) \tag{2.2}$$

$$s_{BM25}(D,q) = IDF(q) \cdot \frac{f(q,D) \cdot (k+1)}{f(q,D) + k \cdot (1 - b + b \cdot \frac{|D|}{avgdl})} \tag{2.3}$$

$$IDF(q) = \log \frac{N - n(q) + 0.5}{n(q) + 0.5} \tag{2.4}$$

ここで，$f(q,D)$ は，文書 D における単語 q の出現頻度，$|D|$ は文書 D の文書長，$avgdl$ は収集されたテキストの平均文書長である．k と b は自由なパラメータであり一般的には $k=2.0$，$b=0.75$ とされる．

第3章　提案手法

提案手法では，前述したBM25を利用した対応付けの手法を参考にする他，ブログ記事の慣習，共起語，文末表現を利用した合せて4つの対応付け方法で構成する．

3.1　ブログ記事の慣習の利用：M_{COM}

次の2つのブログ記事の慣習を利用する．

- **引用+相手名：** コメントのタイトルやコメント文中に “> author さん” といった記述が多い．この “author” はコメント先の相手名である．
- **“Re” 付きタイトル：** コメントのタイトルに，“Re” 付きで，コメント先のタイトルが記述されることがある．

3.1.1　判定方法

これらの慣習が見られると，それを用いてコメント先を決めることができる．引用＋相手名の場合コメント元より前にあるその相手の記述しているコメントがコメント先となる．“Re” 付きタイトルの場合そのタイトルのブロックがコメント先となる．

ただし，図3.2のように複数の候補がある場合，記述者は意図的に複数対応先を指定していると考えられるので対応付け先を複数とする．

3.1.2　慣習利用対応付け例

具体例を図3.1に示す．

ニックネーム「MLJ」さんの書いた「ココは夜がお勧めですよ」というタイトルをもつコメントに対して,「Re:ココは夜がお勧めですよ」と「Re」をタイトルに組合せることでそのタイトルのコメントに対して返信している．また,「＞MLJ さん」の記述は「引

1 ■ココは夜がお勧めですよ
ウチの会社はココからの近所なので、
夜は良く利用しています。…
MLJ 2010-10-06 22:03:16
2 ■ Re:ココは夜がお勧めですよ
＞ MLJ さん
コメント誠に有難うございます。
そうなんですね！それは良さそうですね…
ゲンゾウ 2010-10-06 22:06:55

図 3.1: ブログ記事例文

用+ニックネーム」の慣習であり相手を指定して返信を行っている．これによりコメント先が決定する．

4 ■コメントありがとうございます♪
＞ Momoi さん
そうなんです、温かいエクレアに、
キャベツのシャーベットと乾燥焼きした葉で、
見た目も楽しいし、
温度差もステキなデザートでした (^-^)
＞ misty さん
そうですよね、直前や週末は特に取りにくいようですよね。
先々の外出ご予定にあわせて、
またの機会がありますように (^-^)
＞夢見うさぎさん
そうですね、酸味や甘みの使い方が
とっても素晴らしいと思います◎
写真を撮るのは習慣になりつつありますが、
たまに間違えたりします (^-^)
ひめ 2010-09-01 00:40:27 >>このコメントに返信

図 3.2: 複数対応例文

3.2 文章中の内容語の利用：M_{BM25}

3.2.1 判定方法

本研究では，各名詞ごとのスコアを利用するため，式 2.2 で示した Okapi-BM25 の要素である式 2.3 を用いてコメント先の解析を行う．コメント元のブロックを B_s，コメント先の候補となるブロックの集合を C，ブロック B に含まれる名詞の集合を返す関数を $nouns(B)$ とする．このとき，コメント先のブロック $\tilde{B_d}$ は，式 3.1 で求める．

$$\tilde{B_d} = \underset{\substack{B_d \in C \\ q \in nouns(B_s)}}{\arg\max}\ s_{BM25}(B_d, q) \tag{3.1}$$

ただし，$s_{BM25}(B, q)$ が同点の場合，C にて先に現われたもの，すなわち，ブログ記事において，本文部やコメント部のはじめの方を優先する．また，Okapi-BM25 における全文書集合は，$C \cup B_s$ とする．

3.2.2 Okapi BM25 対応付け例

BM25 による対応付けの例を示す．まず，ブログ記事の本文，コメント A，コメント B を図 3.3 に示す．

次に，本文，コメント A，コメント B を形態素解析し，Okapi BM25 によるスコア付けを行った結果の一部を表 3.1，表 3.2，表 3.3 に示す．

特徴的な名詞である「パスタ」「メチャメチャ」などの名詞が上位に表れているのに対して，普遍的な名詞としてブログ中でよく使用される「私」,「の」といった名詞が低い特徴度として表れていることが確認できる．

これよりコメント B の対応先が本文部であるか，コメント A であるかを解析する場合を考える．まず，コメント B に存在する名詞と一致する名詞が本文またはコメント A に存在するかを調べる.「私」,「の」が本文及びコメント A 共に存在し,「てて」が本文に,「雰囲気」がコメント A に存在していることが確認できる．これらを対応先候補名詞とする．次に，コメント B 中の名詞の Okapi BM25 スコアを確認すると「雰囲気」が最も高いスコアである．よってコメント B の対応先はコメント A となる．

本文の例
場所は西鉄高宮駅から徒歩 5 分足らず。
外観がかわいらしい感じなので内装はどうなんだろ … と思いながら入りましたがとっても落ち着いてて、お隣さんとのテーブル感覚も良い感じ。
　この日はパスタが食べたかったのでここに来たのですッ
なので 1000 円のパスタランチにしました。
パスタメニューは 8 種類位あったかな。
○○○円となるものもあり、秋らしいメニューもチラホラ。
今回はシンプルに 1000 円で OK のものを選びました。
まずは前菜盛り合わせとフォカッチャ。
秋野菜とベーコンのトマトソース。
最初はケチャップを効かしたトマトソースっぽいなぁと思っていたのですが、食べていくにつれ、トマトのみずみずしさがでてきました。
量が少なめだなと思ったのも最初だけ。
食べ終わりはそこそこ満足
　＋１００円でコーヒーをプラス。
高宮価格かな～という気もしましたが、前菜がちゃんとしてたし。
ねッ
実はこちら、平日ランチのみ、幼稚園以下のお子様連れはお断りのお店。
これには賛否両論あると思いますが、ふと一息つくお昼御飯の時間を静かに過ごしたい私には、ありがたいなって思いました
イタリア食堂　CUCCIOLO　（クッチョロ）

コメント A の例
あら～可愛いお店ですね
イタリアン好きな私は最近いった中で雰囲気がよかったな～と思ったのがサーラカリーナでした。
。
今日は、冷泉公園でやってるビアフェスにいってきましたよー初日にもかかわらずメチャメチャ人がおおくてびっくりしました。

コメント B の例
かわいらしいでしょ～でも店内はなかなか落ち着いてて、私は店内の雰囲気の方か好みでした♪サーラカリーナ、美味しいですよねッ雰囲気もよく、私も好きです。
久留米に本店があるのですよ。
でも雰囲気は断然御所ケ谷の方が良いです。

図 3.3: ブログ記事例

表 3.1: BM25 スコア（本文）

特徴度	名詞
1.52884082272606	パスタ
1.17252591054585	感じ
1.17252591054585	もの
1.17252591054585	最初
1.17252591054585	円
0.691637048535585	トマト
:	中略
0.690051367539579	なぁ
0.690051367539579	は
0.690051367539579	お
0.530442638662287	ッ
0.312174481534326	てて
0.000000000000000	私
-0.691637048535585	の

3.3 共起語の利用：M_{COON}

前節で文章中の内容語を利用する手法を述べた．コメント部に，追加情報が書かれる際，コメント先の内容語がそのまま現われるのではなく，関連する言葉が現われる．

たとえば，コメントに「レカロは何がいいかな？」があるとき，これに対するコメントに「TS-G が視点さがるからいいよ。」がある.「レカロ」は車のシートの社名であり，「TS-G」は製品名である．これらの論理的関係は，社名と製品名であるが，この関係を共起語で代用することとする．コメント先の解析としては，コメント元の文にあった語の共起語のある文はコメント先の可能性が高いと考える．

3.3.1 共起語データベース

共起語の抽出には ALAGIN の「単語共起頻度データベース」[4] を利用する．今回共起語の共起元として利用するのは，形態素解析機「茶筅」によって抽出された名詞とする．ただし「茶筅」によるタグが次の名詞は除外する．

除外対象

- 非自立

表 3.2: BM25 スコア（コメント A）

特徴度	名詞
1.58115376317169	メチャメチャ
1.58115376317169	最近
1.58115376317169	あら
1.58115376317169	てる
1.58115376317169	冷泉
1.58115376317169	イタリ
1.58115376317169	今日
1.58115376317169	中
1.58115376317169	サーラカリーナ
0.715303062153352	雰囲気
0.715303062153352	ー
0.0	私
-0.715303062153352	の

- 代名詞
- 接尾
- 助動詞語幹
- 数
- 名詞接続
- 特殊

3.3.2 判定方法

共起度を用いたコメント先の解析方法を説明する．前述と同じく，コメント元のブロックを B_s，コメント先の候補となるブロックの集合を C，関数 $nouns(B)$ はブロックに含まれる名詞の集合を返す関数である．このとき，コメント先のブロック $\tilde{B_d}$ は，次式で求める．

$$\tilde{B_d} = \underset{\substack{B_d \in C \\ w_d \in nouns(B_d) \\ w_s \in nouns(B_s)}}{\arg\max} \; s_{COON}(w_s, w_d) \tag{3.2}$$

表 3.3: BM25 スコア（コメント B）

特徴度	名詞
2.28318690529902	方
2.28318690529902	店内
1.61442175585093	本店
1.61442175585093	久留米
1.61442175585093	好み
1.61442175585093	御所ケ谷
1.61442175585093	なかなか
1.0328980159101	雰囲気
0.730353272695444	ッ
0.730353272695444	てて
0.0	私
-0.730353272695444	の

コメント元の単語（名詞）と共起度の高い共起語に注目し，コメント先候補にその共起語が存在すれば，コメント先と判定する．複数のコメント先が存在する場合は，複数をコメント先として出力する．

3.3.3 共起語利用例

共起語利用の例として次のような食べ物に関するブログのコメントがある．

まる玉らーめん６５０円。うろ覚えだけど、本店より軽めの鶏白湯スープ、極細麺はとても美味しい。あおさが美味しいねえ、チャーシューはまずまず
★★★★４．０

図 3.4: 共起語（コメント A）

うまそーっ、替え玉お願いしまーす。

図 3.5: 共起語（コメント B）

これはラーメン記事におけるコメントである．前述した BM25 ではコメント A とコメント B に共通する名詞が存在しない場合は対応付けができない．ここで，コメント B の

名詞を検索ワードとして共起語データベースにて検索すると「替え玉」の共起語に「極細麺」が存在している（図 3.6）．一般的に「細麺」は伸びやすいため大盛りにせず,「替え玉」をすると知られている．つまり「極細麺」と「替え玉」は同じラーメンという話題において共起される名詞である．これより，類似する内容のコメントであるとしてコメント B のコメント先はコメント A となる．

3.4　文末表現の利用：M_{SFX3}

日本語で話し手の意図は，文末表現に表れやすい．たとえば，質問の意図は「〜ですか？」のように助詞や助動詞で構成された文末表現に表れる．そこで，コメントでの意図のやりとりが最も自然であると解釈されるブロックの対をコメント元とコメント先の対であると仮定して，コメント先を解析する．

3.4.1　コメント-返答の対のモデル化

ここで，コメント元とコメント先の対を大量を得る必要がある．そこで,「みんカラ（みんなのカーライフ）」[5] を参照する.「みんカラ」はブログ記事のコメントに対してブログ著者からの返答が 1 対 1 対応で記述される．よって,「みんカラ」の返答付コメントを収集することでコメント-返答の対を大量に収集することが可能である．図 3.7 にその様子を示す.

次に，モデル化の方法を説明する．まず，文末表現の認定条件を以下に示す.

- 文字列の末尾側に存在する非漢字かつ非カタカナ文字列を採用
- 括弧内はすべて含む
- 全角空白は無視
- 文末の句点や全角ピリオドは無視

次に，文末表現は多様であるため，文末表現を構成する任意の 3 文字に注目し，コメントと返事のブロック対から 3 文字対を全通り抽出する．そして，その対の頻度を求める.

ここで，２つの 3 文字列 s_1 , s_2 がコメント-返事のブロック対に出現した回数を返す関数を $f_{cr}(s_1, s_2)$ とする．s_1 と s_2 がコメント-返事のブロック対に出現しやすく，逆に

s_2 と s_1 がコメント-返事のブロック対に出現しにくいことを表す関数 $s_{SFX3}(s_1, s_2)$ を次式で定義する．

$$s_{SFX3}(s_1, s_2) = \log\{f_{cr}(s_1, s_2) + d\} - \log\{f_{cr}(s_2, s_1) + d\} \tag{3.3}$$

ここで，d は定数（本研究では $d = 0.5$ とした）である．
M_{SFX3} のモデル化のために，2,980 件のブログ記事を利用した．得られた 3 文字列の組は，55,237 組であった．s_{SFX3} のスコアの高いものを幾つか表 2 に例示する．

表 3.4: コメント返答文末表現対の例

コメント文末	返答文末	スコア
めでと	ありが	4.330733
おめで	ありが	4.330733
めでと	りがと	4.317488
…	…	…
んばん	おはよ	3.583516
めでと	うござ	3.583519
でとう	うござ	3.540959
…	…	…
ですね	ますｍ	2.833213
すね!	ｗｗｗ	2.833213

3.4.2 判定方法

$s_{SFX3}(s_1, s_2)$ を用いたコメント先の解析方法を説明する．前述と同じく，コメント元のブロックを B_s，コメント先の候補となるブロックの集合を C とする．関数 $suffix_3(B)$ はブロック B から文末表現を構成する 3 文字の集合を返す関数である．このとき，コメント先のブロック $\tilde{B}_d$ は，次式で求める．

$$\tilde{B}_d = \underset{\substack{B_d \in C \\ s_1 \in suffix_3(B_d) \\ s_2 \in suffix_3(B_s)}}{\arg\max} \; s_{SFX3}(s_1, s_2) \tag{3.4}$$

おはようございます。 落ち着きましたか!! それにしても、光ってますねぇ。 撮影場所どこですか?

表 3.5: 文末表現利用：コメント先候補

(゜▽゜)/コンバンハ 撮影場所は 神戸の兵庫突堤ですよ!

表 3.6: 文末表現利用：コメント元

3.4.3 文末表現対応付け例

文末表現を利用した対応付けの例を示す.

コメント先候補の文末「ですか」とコメント元の文末「ですよ」の対が文末表現対データベースに存在するかを調べる.

表 3.7: 文末表現データベース

:	:
される	ですよ
ですね	ですよ
ですか	ですよ
にいて	ですよ
あるし	ですよ
すよね	ですよ
れます	ですよ
…	ですよ
:	:

文末表現データベースに「ですか　ですよ」の対があるのでコメント先候補をコメント元のコメント先と決定する.

3.5 コメント先解析システム

コメント先を解析するために実装したシステムの流れ図を示す.

入力を画像や絵文字を除いた文章のみのブログ記事として，本文部と各コメント部としてブロック単位に分割を行う．分割後は，後述する決定リストに従い，各手法を利用してコメント先の解析を行い，結果を出力する.

三木監督:8.08723 バリカタ:8.01454 松岡監督:7.77279 草ぶき:7.75509 博多系:7.73088 かわらさん:7.72752 キェシロフスキ:7.72684 リリー氏:7.66985 豚骨臭:7.60588 ヴェラ･ドレイク:7.55627 キルトシユー:7.556 固麺:7.50962 コールド・マウンテン:7.50252 首席補佐官:7.49354 ココシリ:7.4422 政治家志望:7.43335 マトリックス　リローデッド:7.42995 硬麺:7.42642 ヘイフラワー:7.40927 ファレリー兄弟:7.36554 イブラヒムおじさん:7.35797 北九州弁:7.34478 元祖長浜屋:7.32236 クリッピングサービス:7.3202 ジョン・トラボルタ主演:7.31198 老主人:7.26154 マイ・ビッグ・ファット・ウェディング:7.25268 ホテル・ハイビスカス:7.23886 辛子高菜:7.221 豚骨系:7.21597 ツォツィ:7.1513 長浜ラーメン:7.14916 アバウト・シュミット:7.14333 ＣＯＤＥ４６:7.14189 ケヴィン・スミス:7.14014 極細麺:7.13233 仮釈放中:7.13018 １５００マイル:7.12614 本紙川上:7.12487 えびボクサー:7.10886 魁龍:7.09559 マー油:7.09474 ラストキング・オブ・スコットランド:7.07173 デッドコースター:7.06844 四元奈生美:7.04883 北京ヴァイオリン:7.04846 上映案内:7.04248 ブラックブック:7.04158 アナライズ・ミー:7.0151 グエムル漢江:6.99562 ラクエル:6.98546 綴り字:6.98369 レーザー銃:6.98232 元ダレ:6.97887 熊本系:6.9787 久留米ラーメン:6.97025 替玉:6.96805 ブロークン・フラワーズ:6.96413 トンコツラーメン:6.96069 ラーメン自体:6.95429 スープ割:6.93878 ラーメンそのもの:6.91524 若林組:6.9147 マイ・ボディガード:6.90659 悲愁:6.9058 中太麺:6.90541 替え玉受験:6.89329 マンダレイ:6.88963 スティーブン・タイラー:6.88672 マシンガン・トーク:6.88615 唐そば:6.88384 トンコツスープ:6.87935 味玉子:6.86968 百万両:6.8645 ８Ｍｉｌｅ:6.85887 矢谷君:6.85659 ラーメンスレ:6.85567 低加水:6.84779 ストレート麺:6.84687 エンターテインメント業界:6.84593 バッド・エデュケーション:6.84097 トーク・トゥ・ハー:6.8347 九州系:6.8325 御天:6.82657 マシニスト:6.82575 大砲ラーメン:6.82179 福のれん:6.81875 大盛りラーメン:6.8009 康竜:6.79692 ジャーヘッド:6.79467 ロード・オブ・ザ・リング二つ:6.78919 ちゃぶ屋:6.78044 ジャヨン:6.77947 ベティ・サイズモア:6.77231 ダニー・ボイル監督:6.75731

図 3.6: 「替え玉」共起語

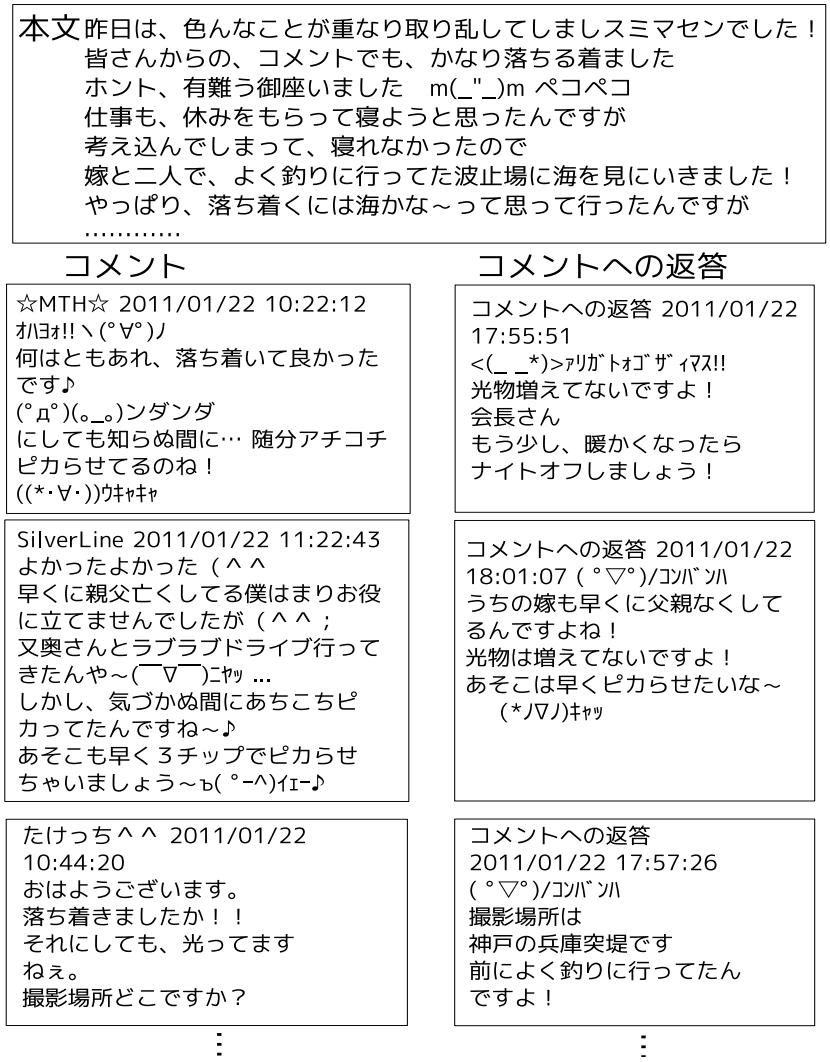

図 3.7: みんカラブログ外観構造

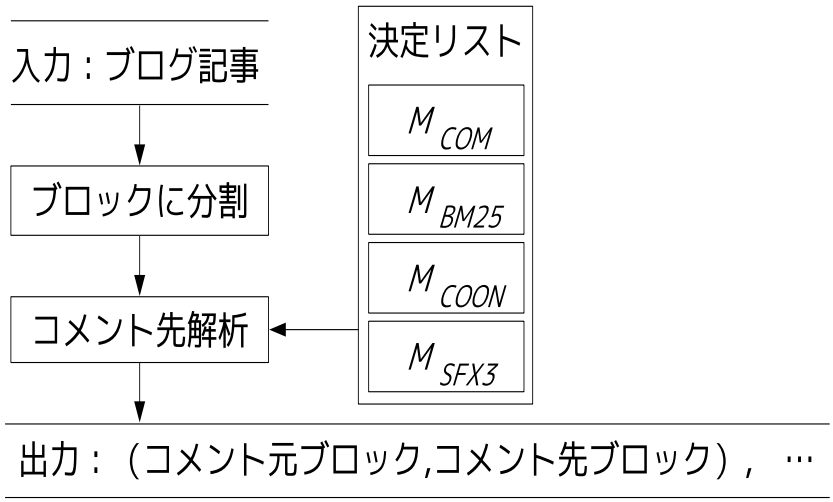

図 3.8: コメント先自動解析システム

第4章　実験

本章では人手により作成した正解データとコメント先解析システムにより出力された結果の比較による性能評価を行う．

4.1　実験

4.1.1　実験条件

テストデータは，Ameba ブログ [6] からブログ記事を抽出して作成する．コメント先の正解データは，人手で作成する．ブログ記事の抽出において，次の点に注意する．

- 内容のあるコメントが書かれている（たとえば,「ペタ」と呼ばれるブログ閲覧の形跡のみに関するコメントは内容が無い）
- Ameba ブログのジャンル別ランキングを参照し，異なる複数のジャンルからテストデータを構成する

テストデータは，ブログ記事 32 件から作成した．ブログ本文部とコメント部によるブロックの数は，255 件であった．第一コメントのコメント先は必ずしもブログ本文部とは限らない．例えば，著者と知り合いの人がブログ記事に書かれていない情報についてコメントしている場合などが挙げられる．従って，テストしたコメント元の件数は 224 件である．一方，正解のコメント先の数（理想的なコメント先の数）は，223 件である．

4.1.2　評価方法

正解のコメント先が複数ありうるので，適合率 P，再現率 R，F 値を用いて評価する．各式は以下のとおりとする．

$$P = \frac{\text{正解のコメント先と出力のコメント先の一致した数}}{\text{出力のコメント先の数}} \tag{4.1}$$

$$R = \frac{\text{正解のコメント先と出力のコメント先の一致した数}}{\text{正解のコメント先の数}} \tag{4.2}$$

$$F\text{ 値} = \frac{2 \cdot P \cdot R}{P + R} \tag{4.3}$$

4.2 実験結果

4.2.1 単独手法の場合

3 つの手法を単独で用いた場合について評価をまとめると表のとおりとなる．

適合率の順位については予想通りの結果となった．しかし，M_{SFX3} の F 値は期待したほどの性能ではなかった．

表 4.1: 単独手法の性能

手法	適合率	再現率	F 値	(一致数,出力数)
M_{COM}	0.88	0.27	0.41	（ 60 , 68 ）
M_{BM25}	0.58	0.43	0.49	（ 95 , 164 ）
M_{COON}	0.50	0.22	0.31	（ 50 , 100 ）
M_{SFX3}	0.29	0.22	0.25	（ 49 , 167 ）

4.2.2 手法を総合した場合

手法を総合した場合では経験的に定めた決定リストにて使用順序を「$M_{COM} \rightarrow M_{BM25} \rightarrow M_{COON} \rightarrow M_{SFX3}$」とした．これは，$M_{COM}$ は直接対応先を指定しているので直感的にもっとも正しい対応付けを行うと考え，次にコメント中に含まれる名詞を利用した M_{BM25}，共起される名詞を利用した M_{COON} がつぎに高いと思われ，M_{SFX3} は「質問ー回答」といったきれいな文末表現対以外にも多数の口語的な表現等曖昧な表現対も多かったため優先度を低くした．

表 4.2: 総合手法の性能

手法	適合率	再現率	F 値	(一致数,出力数)
総合手法	0.64	0.63	0.64	(142 , 219)

4.3 追加実験

M_{SFX3} の効果が他に比べて低いので効果の確認をした．M_{SFX3} を総合手法から抜いた場合と総合手法の性能を比較した．

表 4.3: 総合手法の性能

手法	適合率	再現率	F 値	(一致数,出力数)
M_{SFX3}	0.29	0.22	0.25	(49 , 167)
M_{COM}+M_{BM25}+M_{COON}	0.65	0.63	0.639	(141 , 218)
総合手法	0.64	0.63	0.643	(142 , 219)

第5章　考察

5.1　考察

5.1.1　慣習の利用に関する考察

ブログ記事の慣習の利用：M_{COM} は再現率は全てのブログ記事に慣習が利用されているわけではないため低いものの，実験前の予想通り最も高い適合率となった．再現率の低さは慣習を利用していないブログの存在や次のようなニックネームの特殊な例が存在する．

マヨネーズ　どばーっと　いこうよ　冷やし中華は　やっぱり　マヨね（^皿^）
とらさんの場合は　こしあんでもいいよ　（ ´ ▽ ‘ ）ノ

図 5.1: ニックネーム特殊例

この特殊例のコメントにはニックネームの慣習利用として「とらさん」が存在するが，このコメント先となる人物の本来使用しているニックネームは「とらきち」である．この例のように，親しい間柄である場合ニックネームの短縮や「misty」のような英字を片仮名読みで「ミスティさん」や「みっちゃん」などと呼び元のニックネームとは異なる記述がなされる．この場合コメント先を決定することが出来無い．また，ニックネームの問題点としては「ミ」，や「そうかも」文字列が利用されて，名前と認識しづらく，一致を調べる際に正規表現を利用出来無い文字列が利用されていたりする場合もあ．この問題に対しては，正規表現を利用して一致を調べるのではなく全慣習を網羅することで完全一致により正解と比較することで解消されると考えられる．

タイトルの「Re」を利用する方法の問題点としては次のような場合がある．

図 5.2 の場合，コメントタイトルが「無題」と「Re:無題」ばかりなためどの「無題」に対しての返信かをタイトルによって決定することは難しい．このように同じタイトルが使用されている場合はタイトルによる慣習利用は難しい．解決法としては完全ではない

コメント 1:無題
コメント 2:無題
コメント 3:Re:無題
コメント 4:無題
コメント 5:Re:無題
コメント 6:Re:無題

図 5.2: Re:特殊例

ものの，通常返信は記述された順に返していると考えられるので，各コメントに記されているコメント番号を利用して最も近いものを正解とする，等が考えられる．

5.1.2 Okapi-BM25 利用に関する考察

Okapi BM25 により，特徴度の高い名詞を利用して対応付けを行うことで各コメント文において注目すべき名詞を利用できている．BM25 の問題点としては次のような場合が存在する．本文とコメント A，コメント B を図 5.3，図 5.4，図 5.5 に示す．

次に本文，コメント A，コメント B それぞれのスコアを表 5.1，表 5.2，表 5.3，表 5.4 に示す．

コメントの対応先としてはコメント B の対応先は本文が正しいコメント先である．しかし，コメント B の対応先決定の際に使える名詞は「アヂト」のみである．今回「アヂト」という名詞は本文中だけでなくいくつかのコメントにおいても見られ，多くのブロックで使用されている名詞であるため IDF が低く，最も特徴度の低い名詞となっている．コメント A の「アヂト」の特徴度は本文（0.122）より高い値（0.2721）である．これによりコメント B は間違った対応先を持つこととなる．このように，盛り上がった話題の中で利用される中心的な名詞は特徴度が低くなるため他に利用できる名詞が存在しない場合このような誤りが発生する．

また，BM25 は形態素解析を行い「名詞」を利用しているため形態素解析に誤りがある場合も性能が低下するという問題点がある．

BM25 においては特徴度をブログ記事内で算出するのではなく，一般的な特徴度として web 上の記事全てを対象とした特徴度を算出する方法，形態素解析において「未知語」と判断された語も含むことで語の数を増加させる方法等が考えられる．

5.1.3　共起語の利用に関する考察

共起語の利用はBM25の「対応先と同じ名詞が存在しない」という問題を解決する一つの手段である利用できる名詞を増加させている．しかし，この問題点としては共起される名詞が正しいものばかりでなく，間違った名詞も存在するという点である．ブログを対象としたコメント先の解析が目的なので，それに応じた共起データベースを利用することが最も望ましいと考えられる．今回間違った対応付けの原因としては表5.5のような異常に共起度の高い普遍的な共起語の存在が挙げられる．これらの共起語は多くの名詞から共起され，かつ，高い共起度を持つ．これにより間違った対応を行ってしまう場合がある．

5.1.4　文末表現の利用に関する考察

M_{SFX3} の F 値は最も低いが，総合手法に組込むことで，F 値が “0.639” からわずかに向上した “0.64”（正確には0.643）．これは M_{SFX3} の判定の特性が文末表現依存であり，他の判定の特性である名詞依存と異なるためと考える．現段階では誤差程度であるが，文末表現対データベースの充実など，今後の改良の余地があると考えている．

5.1.5　総合手法に関する考察

各手法の性能をうまく取込み最も性能が高い結果となった．一度対応付けが行われたコメント元に対しては間違った対応付けであっても優先度の低い手法を適用することができないという問題がある．これに対して決定リストを利用するのではなく，全手法の対応先をから，総合的にもっともらしいコメント先を出力するという方法が考えられる．

今日は午前中に病院に行って来ました。
お年寄りが多く順番待ちが長そうだったので、受付票をもらって「仕事が忙しいから一度事務所に戻ってまた来ます」と言って、その間に徒歩 1 分のプリモディーネでカプチーノタイム（笑）。
今朝はまた素晴らしいメンバーがそろっていました。
オーナーシェフ、パテシエ、写真家など良くお会いするメンバーですが、実に楽しくお話が出来ました。
病院の待合室でじっとしてるより有意義な時間が過ごせました。
病院 … と言っても体調不良ではなく定期的な血液検査の結果を聞きに行ったんです … (;^_^A。
今日のランチはアヂトに行って来ました。
今週末が忙しくて行く時間がなくなりそうだったのでフライングです（笑）。
アヂトでは長年放置状態にあったウェブサイトがついにリニューアル OPEN いたしました (^_^)v
http://adito.jp/←アヂト
今回アヂトまではウォーキングで行ったので、食前にジャスミンティー風味の巨峰のさっぱり氷を頂いて身体をスッキリさせてから定食を頂きました。
大粒の巨峰がゴロゴロ入った氷は汗ばんだ身体を冷やして食欲が湧いてきます。
定食はセロリの風味が実に旨い肉汁たっぷりのミンチカツと、ヘルシーな麦飯の上に乗せられた玉葱とセロリのスライスがアヂトっぽくて旨い丼です。
付け合せのサツマイモの天ぷらや切り干し大根も優しい味でホッとします。
ここでひとつアヂト関連
でお知らせがひとつあります。
アヂトの姉妹店というより母娘店（笑）が京都の嵐山にある人気店「松籟庵」です。
この松籟庵のオーナーで、国内外で人気の高い書道家でもある小林芙蓉氏が講師を務める番組「大人の習い事：嵐山書庵」が明日１０月１６日土曜日１０：３０から BS ジャパンで放送されます。
アヂトの姉妹店って書くと他人行儀ですが … めちゃめちゃアヂトファミリーで素敵な店ですよ。
（笑）
松籟庵は僕も妻も大好きな店なので、このブログでもまた紹介したいと思います。
http://www.bs-j.co.jp/naraigoto/←「大人の習い事」
明日は早起きして妻とドライブデート（笑）の予定です (^_^)v。
前から行きたいと思っていた静岡のクレマチスの丘にある「リストランテ・プリマヴェーラ」で黒羽徹シェフと久しぶりにお会いして、自慢のお料理を頂いてから美術館を観たり温泉に寄ったりして帰って来ます。
スペインの超人気店「エル・ブリ（elBulli)」での修業経験もあるシェフのお料理は、また紹介します

図 5.3: BM25 特殊例（本文）

病院 !! ドキッといたしました
良好で、安心いたしました
本日のアヂトも、大変魅力的です
アヂトファミリーは、究極なこだわり人でなりたっているのですね

図 5.4: BM25 特殊例（コメント A）

フィレンツェ生まれの画家の代表作。
その名を冠した料理店、興味津々です
ああ、アヂト
アヂトの巨峰味 ^^
２２日、
何とか行けるかな～（日程調整します

図 5.5: BM25 特殊例（コメント B）

表 5.1: BM25 スコア特殊例（本文 1）

特徴度	名詞
1.0558282966179	メンバー
1.0558282966179	今日
1.0558282966179	時間
1.0558282966179	結果
0.910679374062154	病院
0.668710501348339	血液
0.587516364469457	オーナー
0.587516364469457	これ
0.587516364469457	明日
0.587516364469457	今週
0.587516364469457	年
0.587516364469457	待合
0.587516364469457	聞き
0.587516364469457	ため
0.587516364469457	尿酸
0.587516364469457	先生
0.587516364469457	プリモディーネ
0.587516364469457	痛風
0.587516364469457	何度
0.587516364469457	8
0.587516364469457	明後日
0.587516364469457	いっぱい
0.587516364469457	フライング
0.587516364469457	パテ
0.587516364469457	お話
0.587516364469457	カプチーノタイム

表 5.2: BM25 スコア特殊例（本文 2）

特徴度	名詞
0.587516364469457	今朝
0.587516364469457	血管
0.587516364469457	長年
0.587516364469457	ラ・ビコッカ
0.587516364469457	一
0.587516364469457	順番
0.587516364469457	間
0.587516364469457	てるよ
0.587516364469457	チェック
0.587516364469457	ウェブサイト
0.587516364469457	受付
0.587516364469457	1
0.587516364469457	こちら
0.587516364469457	プリマヴェーラ
0.587516364469457	オプション
0.587516364469457	事務所
0.587516364469457	体調
0.587516364469457	ん
0.587516364469457	以前
0.587516364469457	徒歩
0.587516364469457	食べ
0.587516364469457	コレステロール
0.587516364469457	リニューアル
0.587516364469457	多く
0.587516364469457	ランチ
0.372104407405276	写真
0.372104407405276	今回
0.122145549210176	アヂト

表 5.3: BM25 スコア特殊例（コメント A）

特徴度	名詞
2.35253747624165	安心
2.35253747624165	ドキッ
2.35253747624165	魅力
2.35253747624165	本日
1.48998328631425	病院
0.272157871685694	アヂト

表 5.4: BM25 スコア特殊例（コメント B）

特徴度	名詞
2.16154538731997	名
2.16154538731997	2 2
2.16154538731997	代表作
2.16154538731997	画家
2.16154538731997	フィレンツェ
2.16154538731997	巨峰
2.16154538731997	何
2.16154538731997	料理
2.16154538731997	興味
2.16154538731997	ああ
0.250062580556572	アヂト

表 5.5: 普遍的共起語

共起語	共起度
今日	15149
気	13973
中	30840

第6章　おわりに

本研究ではコメントの対象を明確にするため，本文部へのコメントであるか，あるいは，別の先行するいずれのコメント部へのコメントであるかというブロックの単位でのコメント先の解析を行う方法を提案し，システムを作成した．本手法はブログ記事の慣習，ブログ記事の内容語，共起語，文末表現に着目する手法を，決定リストで組合せた手法である．コメント先の自動解析システムは，ブログ記事を入力として，本文部および各コメント部というブロックに分割を行い，各コメント部のコメント先を自動的に解析する．評価実験では，Ameba ブログからランダムに収集したブログ記事 32 件，ブロック数 255 件，コメント元件数は 224 件を利用して，人手により作成した正解データと解析システムによる出力を比較した．その結果，4 つの手法を組合せることで，適合率 0.64，再現率 0.63，F 値 0.64 と各手法を単独で用いるよりも高い性能評価を得ることができた．

謝辞

本研究を進めるに当たり，種々の御助言を頂きました村田真樹教授，および，村上仁一准教授に心から御礼申し上げます．

また，徳久雅人講師には，終始に渡り研究の進め方や本論文の書き方など，細部に渡る御指導を頂きました．ここに深く感謝いたします．

その他様々な場面で御助力をいただいた計算機工学講座 C 村田研究室の皆様に感謝の意を表します．

参考文献

[1] 池田大介, 藤木稔明, 奥村学: “blog とニュース記事の自動対応付け”, 言語処理学会第 11 回年次大会発表論文集, pp.1030-1033, 2005.

[2] 荒牧英治, 今井健, 美代賢吾, 大江和彦: “非文法的かつ断片化されたテキストの頑健な分類”, 電子情報通信学会データ工学ワークショップ (DEWS2007), 2007.

[3] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M.: “ Okapi at TREC 3 ”, Proc. of the 3rd Text REtrieval Conference, 1994.

[4] ALAGIN“単語共起頻度データベース (Version 1.1)”,Advanced LAnGuage INfomation Forum, 2011.

[5] “みんなのカーライフ”,
http://minkara.carview.co.jp/

[6] “Amebablog ランキング”,
http://ranking.ameba.jp/

[7] “ココログ”,
http://www.cocolog-nifty.com/