ブログ記事におけるコメント先の解析

津野優佑 徳久雅人 村田真樹
鳥取大学 工学部 知能情報工学科
{s072034, tokuhisa, murata} @ ike.tottori-u.ac.jp

1　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，本稿では，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目標とする．

2　関連研究

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，他の文書と比べた出現の仕方が影響するという点で異なる．本稿では，Okapi-BM25 を用いる手法を採用する．

3　提案手法

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，これらを決定リストで組み合わせた手法を，本稿で提案する．

3.1　文章中の内容語の利用

Okapi-BM25 を用いてコメント先を解析する方法を説明する．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．

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

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

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

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．

さて，本稿では，$s_{BM25}(D,q)$ 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，次式で求める．

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

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　共起語の利用

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．

3.2.1 共起度の計算

一般のブログ記事から単語を抽出し，同一記事内の単語の組を共起語とみなす．2 つの単語 x, y の共起語の度合い（共起度）s_{COON} は次式の相互情報量で求める．

$$s_{COON}(x, y) = \log \frac{P(x, y)}{P(x) \cdot P(y)}$$

$P(x, y)$ は全ての記事における 2 つの単語 x, y の同時出現確率，$P(w)$ は全ての記事における単語 w の出現確率である．

3.2.2 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．

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

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 ブログ記事の慣習の利用

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．

具体例を図 1 に示す．

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 1 ブログ記事例文

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．ただし，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．

3.4 文末表現の利用

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）」[4] 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 2 にその様子を示す．

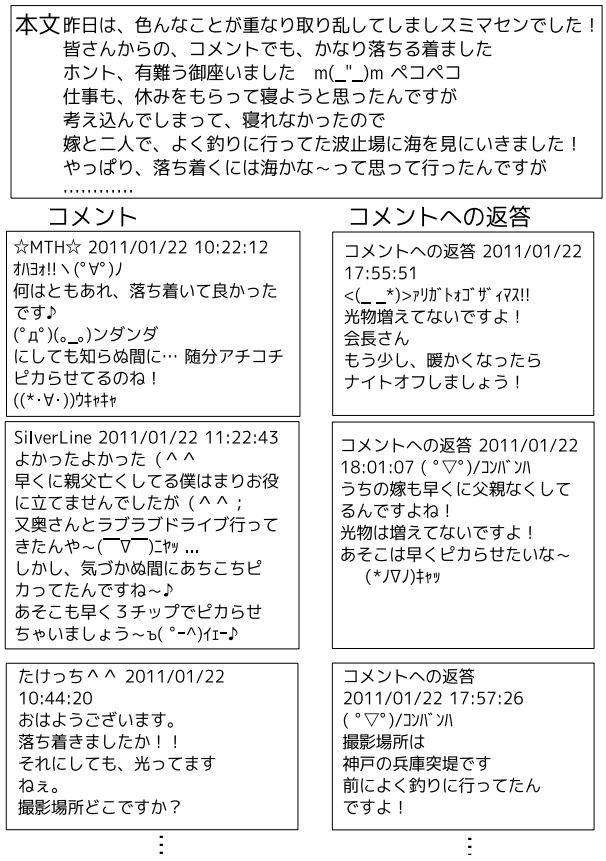

図 2 みんカラブログ外観構造

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．

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

ここで，d は定数（本稿では $d = 0.5$ とした）である．

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．

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

表 1 見出し語/共起度：共起語対の例

見出し語/共起度	:	共起語
インターネットカフェ/		
6.989335	:	騒動
6.989335	:	開発途上国
6.989335	:	クレア
6.583870	:	睡眠時間
6.073045	:	毎朝
:		:
自主回収/		
6.989335	:	増殖
6.989335	:	重荷
5.890723	:	賞味期限
:		:

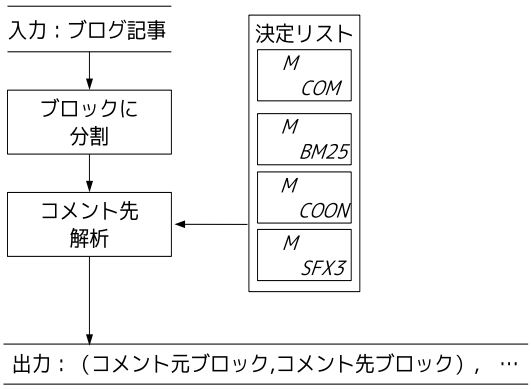

図 3 対応付けシステム

4 実装

コメント先を解析するシステムを図 3 に示す．ブログ記事を，本文部と各コメント部のブロックに分割する．決定リストに基づき，3 章で示した 4 つの解析手法を順に実行する．決定リストの順と各解析手法の表記を次のとおりとする．

1. M_{COM}: 3.3 節で示した慣習に基づく手法
2. M_{BM25}: 3.1 節で示した内容語に基づく手法
3. M_{COON}: 3.2 節で示した共起名詞に基づく手法
4. M_{SFX3}: 3.4 節で示した文末表現に基づく手法

この順位は，経験的に定めた．

ここで，M_{COON} のモデル化（共起度の算出）のために，ブログ記事 4,340 件（約 40 万文）を参照した．これは，ブログサイト「ココログ」[6] の 2008 年 8 月 1 日の一部である．共起語の対は 4,417,961 件，収録語は 12,969 語である．表 1 に幾つかを例示する．

次に，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．

5 実験

5.1 実験方法

テストデータは，Ameba ブログ [5] 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

5.2 実験結果

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．

表 2　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
・・・	おおお	3.583519
でとう	うござ	3.540959
…	…	…
ですね	ます m	2.833213
すね!	w w w	2.833213

5.3　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．

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

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

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

5.3.1　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 3 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 3　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}	0.5	0.22	0.31	（50; 100）
M_{SFX3}	0.29	0.22	0.25	（49; 167）

5.3.2　手法を総合した場合

M_{COON}，M_{SFX3} の性能が低かったので，手法を総合した実験では，次の 4 つの場合を比較する．

- L_1：決定リストにて，使用順序を「$M_{COM} \rightarrow M_{BM25}$」とする場合
- L_2：決定リストにて，使用順序を「$M_{COM} \rightarrow M_{BM25} \rightarrow M_{SFX3}$」とする場合
- L_3：決定リストにて，使用順序を「$M_{COM} \rightarrow M_{BM25} \rightarrow M_{COON}$」とする場合
- L_4：決定リストにて，使用順序を「$M_{COM} \rightarrow M_{BM25} \rightarrow M_{COON} \rightarrow M_{SFX3}$」とする場合

結果を表 4 に示す．4 つの手法を全て総合した場合 L_2，L_3，L_4 が同評価で最も性能が高かった．

表 4　総合手法の性能

手法	適合率	再現率	F 値	（一致数; 出力数）
L_1	0.67	0.60	0.63	(134; 201)
L_2	0.67	0.63	0.65	(141; 212)
L_3	0.67	0.63	0.65	(141; 212)
L_4	0.66	0.65	0.65	(144; 217)

6　考察

実験の結果，M_{COON} の性能は，あまり良くなかった．これは，共起語を収集する規模が不足していたためである（ブログ 1 日ぶんにも満たない量であった）．共起語について大規模化を行い再実験することが今後の課題である．

最も F 値 が高かったのは L_2，L_3，L_4 であった．L_2，L_3，は L_1 と比べて，適合率を保ちつつ，再現率を高めることができている．M_{SFX3} は，適合率と再現率が著しく低かったが，効果があったといえる．その理由は，M_{BM25} が内容語を処理対象とすることに対して，M_{SFX3} が付属語を処理対象としたことにあると考える．

なお，M_{COM} と M_{BM25} を単独で用いるよりも，それらを組み合わせた L_1 の方が性能が良かった．これは，M_{COM} が網羅性は低いものの正確性は非常に高いとても特殊なルールであったためである．

7　おわりに

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システムを構築した．今回有意差の計算を行っていないので有意性については断定出来無いが，4 つの手法を，M_{COM}，M_{BM25}，M_{COON}，M_{SFX3} という優先度順に使用した場合が最も性能が良く，F 値で 0.6545 を得た．

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

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

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