様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　24　年　5　月　31　日現在

機関番号：１２１０２
研究種目：若手研究（B）
研究期間：2009～2011
課題番号：２１７００２６８
研究課題名（和文）　日英二言語のブログ文書を対象とした意見抽出と要約生成に関する研究
研究課題名（英文）　Research on Opinion Extraction and Summarization for Blogs in Japanese and English

研究代表者
関 洋平（SEKI YOHEI）
筑波大学・図書館情報メディア系・助教
研究者番号：00348468

研究成果の概要（和文）：
本研究では，Yahoo! ブログを対象として，意見情報と意見対象タイプを付与したアノテーションについて分析を行った結果，頻出する態度評価と意見対象タイプの組み合わせについて，コンピュータ，エンターテイメント，政治などドメインに特徴的な意見を把握することができることを確認した。さらに，作成したコーパスを用いて，複数ドメインを対象としたアンサンブル学習による意見分析システムを提案した。NTCIR-8 の多言語意見分析タスクでは，日本語，英語，中国語を対象とした意見分析のための評価タスクを新たに開催した。さらに，アプレイザル理論に基づく英語の辞書から，日本語のアプレイザル辞書を構築し，日英の意見抽出における効果について検証を行った。また，コミュニティ型質問応答コンテンツ(CQA)に焦点を当てて，意見抽出や要約の基盤技術と応用事例として，良質回答の自動予測と，クエリ拡張型 Web 検索システムについて研究を遂行した。最後に，マイクロブログ（Twitter）コンテンツ中のテキストに現れる感情の推定に基づく顔文字推薦，人と実体間の関係推定，対訳表現を手がかりとした用例選択に基づく機械翻訳など，意見を介したコミュニケーション支援についての基礎研究を進めた。
研究成果の概要（英文）：
In this research, we annotated opinion and opinion targets using Yahoo! Blog corpus and found that opinions were characterized with domains such as computer, entertainment, or politics. We also proposed opinion extraction system toward multiple domains based on ensemble learning. In NTCIR-8, we conduct multilingual opinion analysis task to evaluate opinion extraction technologies in Japanese, English, and Chinese. We also extended English appraisal dictionary toward Japanese dictionary and experimented opinion extraction in Japanese and English using two dictionaries. We also focused community question answering contents (CQA) to extract and summarize user opinions, and proposed good answer estimation and query expansion type Web retrieval systems. Finally, we proposed communication supporting systems mediating personal opinions such as: (1) facemark recommendation system based on emotion estimation using microblogs, (2) relation estimation system toward personal names and surrounding entities, and (3) example-based machine translation system using alignment clues in parallel multilingual corpus.

交付決定額　（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,300,000	390,000	1,690,000
2010 年度	1,100,000	330,000	1,430,000
2011 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：総合領域
科研費の分科・細目：情報学・図書館情報学・人文社会情報学
キーワード：情報検索

１．研究開始当初の背景

個人ユーザが Web 上で日記の形式で意見や感想を発信するブログは，日本ではアクティブな利用者数が 300 万人を突破しており，国別の統計分布では，日本語，英語，中国語の上位３ヶ国語で，世界中のブログの８割を占めていた。消費者が発信するメディアは，マスメディアや企業が一方的に配信する情報と異なり，他の消費者のコメントやトラックバック（双方向リンク）に晒されるため，他の消費者に読まれるためには正直である必要がある。このため，ブログは，より個人の本音を反映した文書と言われていた。このことから，ブログを利用した評判の分析が期待されており，ブログマーケティングというビジネスモデルも立ち上がりつつあった。

２．研究の目的

日本語・英語等を対象とし，意見抽出技術を評価するデータセットを作成し，有効な技術を評価する。具体的には，ブログ等の消費者が発信するメディアに含まれる口語的表現を対象とした意見を抽出するために有効な素性（意見を選択するための属性）と，抽出した意見を読みやすい要約としてまとめるためのテキスト構造について知見を明らかにする。また，異なる言語間の意見抽出技術の差異や文書ジャンルごとの意見抽出技術の違いについて考察する。

３．研究の方法

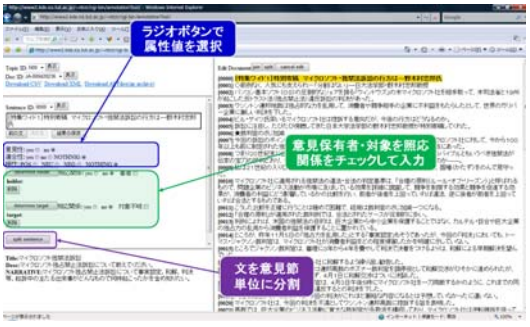
日英のブログ等を含む消費者メディアコンテンツを対象として意見を抽出するためのデータを作成し，言語，文書ジャンル，話題に応じて共通する技術と異なる技術を明らかにする。これらの意見情報の特徴の違いを整理するために，国立国語研究所の作成する現代日本語書き言葉均衡コーパス（BCCWJ）などに含まれる多様な文書ジャンルを対象として大規模な分析を進め，傾向の違いを整理し，情報アクセスに着目した応用を進める上で，必要となる意見情報 を明らかにする。今年度は，新聞記事（780 記事），Yahoo!ブログ（471 記事），国会会議録（14 記事），Yahoo!知恵袋（267 記事），Spinn3r.com（ICWSM）ブログ（80 記事），書籍（10 記事）を対象として意見情報を付与したコーパスを作成した（表 1）。本研究ではこれらを比較・分析した。

表 1　作成した意見分析コーパス

文書ジャンル	ソース	文書数	総文数	内容	作成年度
Yahoo!知恵袋	BCCWJ	251	1,924	コアデータのうち，主要7カテゴリ	平成21年度
書籍		10	407	評論，随筆など	平成21年度
Yahoo!ブログ		**471**	**6,944**	**コアデータすべて**	**平成22年度**
国会会議録		**14**	**5,812**	**本会議録**	**平成22年度**
新聞	NTCIR	780	21,391	NTCIR-6, 7 MOAT	平成21年度
Yahoo!知恵袋	API	16	118	アプリケーション用途	平成21年度
ブログ	ICWSM	80	2294	アプリケーション用途	平成21年度

意見情報としては，国際評価会 NTCIR のタスクやアプレイザル理論を参照しつつ，1. 意見性，2. 極性，3. 意見保有者，4. 意見対象，5. 態度評価 14 タイプ，6. 形勢・やり取り 9 タイプ，7. 推測・発話行為 8 タイプ，8. 体験情報（体験性，体験主，体験タイプ，意見誘発性）に加えて，拡張固有表現の定義に基づく 9. 意見対象タイプを定義し，また，文書ジャンル特有の属性として，カテゴリ適合性（Yahoo!ブログを対象，Yahoo!側で設定したカテゴリに適合した内容の文とそうでない文を判別）と会話タイプ（国会会議録を対象，質問，回答，議事進行，呼びかけ，その他を文単位で定義）の属性を追加した。6 名の判定者を雇用し，詳細なアノテーションマニュアルと図１のオンラインアノテーションツールを作成し，判定者間の付与一致率（κ 係数）を計算したところ，ほぼすべての属性について 0.4 以上，半数以上について 0.6 以上の一致率を達成した．このように，高い一致率を確認し，作業の拡張性を確認した。

図 1 アノテーションツール



これらの意見分析データセットを用いて，アプレイザル理論に基づく英語用の辞書を日本語辞書に拡張し，複数言語・複数ドメインを対象とした意見抽出システムを開発し，どのようなアプローチが有効か明らかにする。

４．研究成果

4-1. ブログ等の口語表現を含む多様な文書ジャンルとドメインを対象とした意見分析

ここ数年,Webなどの大量の電子化テキストに現れる意見情報を抽出し，集約や可視化を行うことで，世論調査や評判分析といった応用を実現する研究が進んでいる。一方，意見情報の特徴はさまざまであり，文書ジャンル（例：新聞，ブログ，雑誌等）やドメイン（政治，映画，商品，恋愛相談等）に応じて，出現する意見情報の傾向は異なる。

本研究では,Yahoo! ブログのアノテーションから，頻出する態度評価と意見対象タイプの組み合わせについて，コンピュータ，エンターテイメント，政治などドメインに特徴的な意見を把握することができることを確認した。また，国会の本会議において，国会の本会議においては，“抽象概念”，“行為概念”，“製品名”などが，意見の対象とされていることを確認した。さらに，文書ジャンル間及び，文書ジャンル内のドメイン（話題領域）ごとに意見情報を比較し，態度評価が主にドメインに依存し，形勢・やり取りが文書ジャンルに依存することを確認した(学会発表20, 23, 25）。また，意見対象を拡張固有表現で分類し，態度評価との組合せを抽出することで，ドメインごとに有用な情報を抽出できることを確認した（学会発表11, 12, 図書1）。さらに，コミュニティQAに現れる質問に対するブログからの回答抽出（学会発表21）および意見を問う質問の分類（学会発表15），複数のドメインを対象にしたアンサンブル学習に基づく意見分析システム（学会発表6）や，ニュース・ブログの話題の相関および変遷(学会発表9）について検証した。

さらに，ここ数年で，ブログの後に発展した消費者メディアコンテンツとして，コミュニティQAとマイクロブログに着目し，応用研究を進めた。コミュニティQAに関しては，良質回答の自動予測（雑誌論文1,5，学会発表19,24）と，クエリ拡張型Web検索システム（雑誌論文2,3，学会発表8, 16）の研究を遂行した。マイクロブログに関しては，感情推定に基づく顔文字の推薦（学会発表3），話題チャンクの抽出（学会発表13），フォロー先の推薦（学会発表14）などについて研究を遂行した。

4-2. 多言語意見分析

NTCIR-8の多言語意見分析タスク（学会発表2）では，日本語，英語，中国語を対象とした意見分析のための評価タスクを新たに開催した。過去の開催では，システムの評価の正解として使う判定者間の意見の判定一致率が低いという課題があった。今回の開催では，言語間のデータフォーマットの統一をアノテーションツール，評価ツールを用いることで達成し，日本語，英語，中国語（簡体字・繁体字）の4言語で共通のXMLフォーマットによる意見分析コーパスを作成し，言語横断意見検索タスクの実現を容易にした。結果として，過去2回の開催よりも，（特に）英語，中国語の判定者間の一致率（κ係数）が向上した。

表 2 判定者間一致率（κ係数）

判定属性	英語	日本語	簡体字中	繁体字中
意見性	0.7309	0.7174	0.9722	0.4568
極性	0.7069	0.6330	0.8901	0.3521
適合性	0.6907	0.6199	0.9716	0.4003
回答性	0.6607	0.5580	0.9712	0.2901

また，アノテーションの積集合（双方とも判定属性がありと判定したものを正解とみなす戦略）に基づく評価の安定性を，κ係数の違いと，参加者のランクの分散分析に基づき調査し，κ係数が0.7以上であれば，ランキングの変動は，分散分析の有意差のない範囲のランク内にとどまることを確認した。

タスクについては，質問に対する回答意見を検索するという観点を取り入れ，言語横断意見検索タスクを新たにサブタスクとして導入した。また，データは比較的新しい2002年～2005年の記事を採用し，英語に関しては，Native の意見性を反映した言語データと言う観点から New York Times を採用した。その結果，世界各国16チームから，56の結果提出があり，参加者の半数以上が2言語に関連したタスクの結果を提出した。さらに，意見文判定サブタスクで，最良の結果を残したチームは，表4に示す通り，(1) 洗練された特徴素選択，(2) 機械学習技術，(3) 豊富な語彙リソースを利用していることを確認した。

表 3 NTCIR-8 多言語意見分析タスクにおける成績優良チームの戦略

チームID	言語	特徴素選択	機械学習	言語リソース
UNINE	英語	Z スコア	ロジスティック回帰分析	SentiWordNet
PKUTM	簡体字中	反復分類器	SVM (better than NB, ME, DT)	自前, NTU, Jun Li辞書
CityUHK	繁体字中	教師あり辞書学習	アンサンブル学習	NTUSD, LCPW, LCNW, CPWP, SKPI

さらに，アプレイザル理論に基づく英語の辞書から，日本語のアプレイザル辞書を構築し，日英の意見抽出における効果について検証を行った。アプレイザル理論に基づく辞書は，言語学の詳細な分析に基づいて構成されるため，語彙項目のカバー率は限定されたものになる。本研究では,WordNet, JWordNet に基づき同義語を展開することで辞書のカバー率を向上させると同時に,SentiWordNet のスコアに基づきフィルタリングを行うことで，辞書の精度を向上させることを試みた。NTCIR-8 多言語意見分析タスクのテストコレクションを用いた実験から，この辞書は意見分析を行う上で，高い再現率を実現できる十分なカバレッジを得ていることを確認した。また，人手で判定をやり直し，辞書の精度の

向上を試みた（学会発表17，18）。また，多言語意見分析への応用のため，用例に基づく機械翻訳システムを開発した（学会発表5）。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 5 件）

① Atsushi Otsuka, Yohei Seki, Noriko Kando, and Tetsuji Satoh: QAque: Faceted Query Expansion Techniques for Exploratory Search using Community QA Resources, In Proceedings of Workshop on Community Question Answering on the Web (CQA2012), in conjunction with the 21st International World Wide Web Conference (WWW), 2012, 査読有, http://www2012.wwwconference.org/proceedings/companion/p799.pdf

② 石川大介，酒井哲也，関洋平，栗山和子，神門典子：コミュニティ QA における良質回答の自動予測，情報知識学会論文誌，査読有， Vol.21, No.3, pp.362-382, 2011, https://www.jstage.jst.go.jp/article/jsik/21/3/21_21-041/_pdf

③ 大塚淳史，関洋平，神門典子，佐藤哲司：情報要求の言語化を支援するクエリ拡張型 Web 検索システムに関する一検討，情報処理学会論文誌データベース，査読有，Vol. 4, No. 3, pp.1-11, 2011, http://www.bookpark.ne.jp/cm/ipsj/search.asp?flag=6&keyword=IPSJ-TOD0403002&mode=

④ Masaharu Yoshioka, Noriko Kando, and Yohei Seki: Evaluation of Interactive Information Access System using Concept Map. In Proceedings of the 4th International Workshop on Evaluating Information Access (EVIA), 2011, pp.20-23, 査読有, http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/EVIA/04-EVIA2011-YoshiokaM.pdf

⑤ Tetsuya Sakai, Daisuke Ishikawa, Noriko Kando, Yohei Seki, Kazuko Kuriyama, and Chin-Yew Lin: Using Graded-Relevance Metrics for Evaluating Community QA Answer Selection, In Proceedings of The Fourth ACM Conference on Web Search and Data Mining (WSDM 2011), Kowloon, Hong Kong, February 2011, pp.187-196, 査読有, http://dl.acm.org/citation.cfm?id=1935864

〔学会発表〕（計 25 件）

① 関洋平：情報アクセスと IFAT，インタラクティブ情報アクセスと可視化マイニング研究会キックオフ・イベント＆第一回研究会，東京大学駒場キャンパス，2012 年 4 月 14 日.

② 江村優花，関洋平：テキストに現れる感情，コミュニケーション，動作タイプの推定に基づく顔文字の推薦，情報処理学会第 85 回デジタルドキュメント研究会・第 106 回情報基礎とアクセス技術合同研究会合同研究発表会，白百合女子大学(東京都調布市)2012 年 3 月 26-27 日.

③ 堂前友貴，関洋平，神門典子：Web から収集した学習データを用いた人物と実体間の関係の推定，言語処理学会第 18 回年次大会，広島市立大学(広島市)， 2012 年 3 月 13-16 日，pp.1027-1030.

④ 大山鉄郎，関洋平：対訳表現を手がかりとした用例の選択手法の提案，言語処理学会第 18 回年次大会，広島市立大学(広島市)，2012 年 3 月 13-16 日，pp.479-482.

⑤ 高村慎太郎，吉岡真治，関洋平：複数ドメインの意見分析コーパスを用いたアンサンブル学習による意見分析システムの提案，言語処理学会第 18 回年次大会，広島市立大学(広島市)， 2012 年 3 月 13-16 日，pp.235-238.

⑥ 枝 隼也，島田 諭，関洋平，神門 典子，佐藤 哲司：複数人での Web 協調探索のための探索履歴可視化手法の提案，DEIM 2012，神戸市，2012 年 3 月 3-5 日.

⑦ 大塚 淳史，関洋平，神門 典子，佐藤哲司：コンテキスト切替による多様な情報要求に対する Web 検索手法の提案，DEIM 2012，神戸市，2012 年 3 月 3-5 日.

⑧ 小池 大地，横本 大輔，牧田 健作，鈴木 浩子，宇津呂 武仁，河田 容英，吉岡真治，神門典子，福原 知宏，中川裕志，清田 陽司，関洋平：ニュース・ブログにおける話題の相関と変遷の分析 --- 震災に関する話題を例題として ---，DEIM2012，神戸市，2012 年 3 月 3-5 日.

⑨ 吉岡真治，神門典子，関洋平：複数国の新聞サイトを比較分析する NSContrast の実験的分析，情報処理学会第 152 回データベースシステム研究会・第 103 回情報基礎とアクセス技術研究会合同研究発表会，立命館大学（朱雀キャンパス），2011 年 8 月 2-3 日.

⑩ 関洋平：BCCWJ を利用した意見分析コーパスの構築について，「現代日本語書き言葉均衡コーパス」完成記念講演会予稿集，東京，2011 年 8 月 2-3 日，pp.125-130.

⑪ 関洋平，神門典子，佐野大樹，柏野和

佳子，稲垣陽一，栗山和子：研究活動・成果の総括:意見情報班 多様な文書ジャンルを対象とした意見分析コーパスの作成に関する研究,科学研究費補助金 特定領域研究「日本語コーパス」平成 22 年度公開 WS 予稿集，東京,2011 年 3 月 14-16 日.

⑫ 新谷歩生，関洋平，佐藤哲司：投稿間隔に基づくマイクロブログからの話題チャンク抽出に関する一検討，第 3 回データ工学と情報マネジメントに関するフォーラム（DEIM 2011），静岡県伊豆市，2011 年 2 月 27 日-3 月 1 日，A1-2.

⑬ 康大樹，島田諭，関洋平，佐藤哲司：属性伝搬モデルを用いたマイクロブログのフォロー先推薦法，第 3 回データ工学と情報マネジメントに関するフォーラム(DEIM 2011)，静岡県伊豆市，2011 年 2 月 27 日-3 月 1 日，A1-3.

⑭ 渡邊直人，島田諭，関洋平，神門典子，佐藤哲司:QA コミュニティにおける質問者の期待に基づく質問分類に関する一検討，第 3 回データ工学と情報マネジメントに関するフォーラム（DEIM 2011），静岡県伊豆市，2011 年 2 月 27 日-3 月 1 日，B5-1.

⑮ 大塚淳史，関洋平，神門典子，佐藤哲司：情報要求の言語化を支援するクエリ拡張型 Web 検索システム，第 3 回データ工学と情報マネジメントに関するフォーラム（DEIM 2011），静岡県伊豆市，2011 年 2 月 27 日-3 月 1 日，F6-3.

⑯ 関洋平：言語を横断したアプレイザル辞書の構築と多言語意見分析タスクにおける評価．情報アクセスシンポジウム 2010，東京大学工学部，2010 年 9 月 27 日.

⑰ 関洋平，神門典子，佐野大樹，柏野和佳子，稲垣陽一，栗山和子：平成 22 年度研究進捗状況報告：意見情報班 多様な文書ジャンルを対象とした意見分析コーパスの作成に関する研究，科学研究費補助金 特定領域研究 日本語コーパス領域平成 22 年度全体会議， 国立国語研究所，2010 年 8 月 30-31 日.

⑱ Yohei Seki, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, and Noriko Kando: Overview of Multilingual Opinion Analysis Task at NTCIR-8 - A Step Toward Cross-Lingual Opinion Analysis, Proc. of the Eighth NTCIR Workshop, NII, Japan, June 2010, pp.209-220 学術総合センター(東京)2010 年 6 月 15-18 日.

⑲ 石川大介，栗山和子，酒井哲也，関洋平，神門典子: Q&A サイトにおけるベストアンサー推定の分析とその機械学習への応用，情報知識学会 第 18 回（2010 年度）年次大会，東京大学本郷キャンパス，2010 年 5 月 15-16 日，pp.73-85.

⑳ 関洋平,神門典子,稲垣陽一,栗山和子：新聞，コミュニティ Q&A，ブログ，書籍を対象とした詳細な意見分析コーパスの作成と比較分析，科学研究費補助金 特定領域研究「日本語コーパス」平成 21 年度公開 WS 予稿集，国立国語研究所，2010 年 3 月 14-16 日，pp.257-264.

㉑ 関洋平：文書ジャンルを横断した回答意見の検索のための分析，情報処理学会創立 50 周年記念（第 72 回）全国大会，東京大学 本郷キャンパス，2010 年 3 月 8-12 日.

㉒ 伊藤誉将，関洋平，青野雅樹：LDA による国会会議録を対象にしたセグメンテーションの一手法,情報処理学会創立 50 周年記念（第 72 回）全国大会，東京大学 本郷キャンパス,2010 年 3 月 8-12 日.

㉓ 関洋平,神門典子,稲垣陽一,栗山和子：新聞記事とコミュニティ QA を対象とした詳細な意見分析コーパスの作成と分析，情報処理学会情報学基礎・自然言語処理合同研究会 FI-97/NL-195，筑波大学東京キャンパス,2010 年 1 月 28-29 日.

㉔ 石川大介,栗山和子,関洋平,神門典子：Q&A サイトにおけるベストアンサー推定可能性の検証，情報処理学会情報学基礎・自然言語処理合同研究会 FI-97/NL-195，筑波大学東京キャンパス，2010 年 1 月 28-29 日.

㉕ 関洋平,神門典子,稲垣陽一,栗山和子：平成 21 年度研究進捗状況報告：意見情報班 多様な文書ジャンルを対象とした意見分析コーパスの作成に関する研究,科学研究費補助金 特定領域研究 日本語コーパス領域 平成 21 年度 全体会議,東京，2009 年 9 月 5-6 日，pp,45-52.

〔図書〕（計 1 件）

① 関 洋平：9 章 コーパスと意見分析，松本裕治・奥村学「コーパスと自然言語処理」，朝倉書店（出版確定）.

６．研究組織

(1)研究代表者

関 洋平（SEKI YOHEI）

筑波大学・図書館情報メディア系・助教

研究者番号：00348468