Distant Supervisionによる感性トピックの抽出

Modeling Sentiment Topics with Distant Supervision

要　旨

近年、インターネット上で個人が自由に情報発信を行うソーシャルメディアが急速に拡大している。ソーシャルメディアではさまざまな事柄に対する個人の意見が述べられており、世間一般の関心を調べるうえで貴重な情報源となる。本稿では、このソーシャルメディアから感性トピックを自動的に抽出する手法を提案する。

テキスト集合を分析する手法として、語の確率分布をトピックとして抽出するトピックモデリングと呼ばれる手法が知られている。提案手法では、トピックモデリング手法にdistant supervisionの考えを導入することにより、感性トピックの抽出が低コストで実現できることを示す。

Abstract

The recent popularity of social media has encouraged many individuals to publically express their opinions. These opinions include valuable information about a vast array of topics. This paper proposes a method of automatically extracting sentiment topics from these social media texts.

Topic modeling is a well-known method of extracting topics expressed as the probabilistic distributions of words from a text collection. We introduce an idea based on distant supervision to this topic modeling, and enable an extraction of topics associated with sentiments. By using the proposed method, we show that sentiment topics can be easily extracted with little effort.

執筆者

三浦　康秀（Yasuhide Miura）
服部　圭悟（Keigo Hattori）
大熊　智子（Tomoko Ohkuma）
増市　博（Hiroshi Masuichi）

研究技術開発本部 コミュニケーション技術研究所
（Communication Technology Laboratory, Research & Technology Group）

1. はじめに

近年、テキストデータの継続的な増加により、膨大なテキスト集合を自動的に分析する技術の需要が高まっている。このような大規模データを分析する手法の1つとして、Latent Dirichlet Allocation（以降、LDA）[1]等のトピックモデリング手法が知られている。トピックモデリング手法では通常、トピックは語の確率分布として抽出される。図1にソーシャルメディアのテキスト集合から抽出したトピックの例を示す。トピックはテキスト集合の全体像を把握するのに有用であり、手法はすでに科学分野の分析[2]、インターネットブログの分析[3]、マイクロブログの分析[4]等に用いられている。

トピックモデリング手法の問題点として、抽出されるトピックが人の直感とは一致しないという点がある。典型的なトピックモデリング手法では教師データ*1中の尤度が最大となるトピックを抽出するが、この基準は人にとって必ずしも最適な基準ではないことが指摘されている[5]。この問題の解法の1つとして、テキストに付与されたラベル*2に結びついたトピックを抽出する教師ありトピックモデリング手法が提案されている[6]-[8]。実際に、テキストと結びついた製品やサービスの評価値をラベルとして用いて、感性情報と結びついたトピック（以降、感性トピック）が抽出できることがすでに示されている[6], [9]。しかし、このような評価値はレビュー記事以外には付与されていないことが多く、提案されている手法をそのままほかの分野のテキストに対して適用するのは困難である。

トピック	語
1	食べる, 美味しい, 飲む, 屋, 料理, ラーメン, 店, コーヒー, 肉, …
2	!, ありがとう, よろしく, お願い, くださる, イイ, これから, 楽しむ, できる, …
3	～さ, 暑い, 夏, この, その, 中, 今日, 風,外, 汗, …
4	くる, 目, 痛い, 入る, 風呂, 寝る, 頭, お腹, すぎる, ない, …
…	

図1　ソーシャルメディアのテキストから抽出したトピックの例。図中では省略しているが、各語にはトピックからの生成確率が設定されている。
An example of topics extracted from social media texts. Although not shown in the figure, each word has a generation probability from its topic.

本稿では、ソーシャルメディアのテキストから感性トピックを抽出するトピックモデリング手法を提案する。提案手法では、従来の教師ありトピックモデリング手法にdistant supervision*3の考えを導入し、“感性手掛かり”および“柔軟なラベル付与スキーマ”を用いた感性トピックの抽出を実現する。ソーシャルメディアのテキストは既存のメディアと内容、文章のスタイル共に大きく異なっており、他メディアを対象に開発された分析技術を適用しても高い性能が望めない[10]。提案手法は少数の感性手掛かりのみで、ソーシャルメディアに特化した感性トピックの抽出が実現できる。

本稿の以降の構成は以下にようになっている。2章では、提案手法の詳細を述べる。3章では、評価実験で用いた各種データについて説明する。4章では、評価実験の詳細を記す。5章では、まとめおよび今後の展望を述べる。

2. 手法

2.1 Partially Labeled Dirichlet Allocation

提案手法はPartially Labeled Dirichlet Allocation（以降、PLDA）[8]を教師ありトピックモデリング手法として用いる。PLDAはLDA[1]を拡張した手法である。LDAでは、文書は語の多項分布であるトピックの混合として表される。PLDAはLDAにラベルを導入し、文書のトピックの生成に対してラベルに応じた制約を掛ける。図2のグラフで表されるPLDAにおける文書の生成プロセスは具体的には図3の手順で行われる。プロセスでは、Dir（・）はディリクレ分布を意味し、Mult（・）は多項分布を意味している。

PLDAの学習は、学習データに対して$P(\mathbf{w}, \mathbf{z}, \mathbf{l} | \Lambda, \alpha, \eta, \gamma)$を最大化する$\Phi$、$\psi$、$\theta$を求

*1 機械学習手法を適用するデータ。ここではトピックモデリング手法を適用するテキスト集合を意味する。

*2 機械学習手法で学習対象となる情報。ここではテキストに人手等の何らかの基準で設定されたトピックを意味する。

*3 間接的な教師信号を用いた機械学習手法。本稿では、基本的には2章で述べる“手掛かり”を教師信号として利用する教師あり学習手法を意味する。

める問題になる。ここで、wは語、zはトピック、lは単語ごとのラベル、Λは文書のラベル、αおよびηはディリクレ分布のパラメーター、γはラベルベクトルのパラメーター、Φはトピック-語の分布、ψはラベルの分布、θはラベル-トピックの分布である。これらパラメーターの効率的な推定手法については、Ramageらの文献8)で述べられている。

2.2 提案手法

提案手法は、感性トピックをテキスト集合から抽出する3ステップの手法である。

- ステップ1：感性手掛かりの定義
  感性手掛かりを定義する。感性手掛かりとは、感性と結びつきの強いメタデータもしくは語彙的な特徴を意味する。例としては、ポジティブの感性と結びつきの強い“笑顔のエモティコン*4”、またネガティブの感性と結びつきの強い“災害に関連するソーシャルタグ*5”が挙げられる。表1に感性手掛かりの例を示す。なお、エモティコンやソーシャルタグを感性手掛かりとして用いる設定は、既存のdistant supervisionによる感性情報分析を行った研究[11)-13)]を参考にしている。
- ステップ2：柔軟なラベル付与スキーマ
  ステップ1で設定された感性手掛かりに基づいてテキスト集合に対してラベルを付与する。ラベルの付与手段は、テキストが感性手掛かりを含んでいるか否かによって変化する。感性手掛かりを含むテキストには、感性手掛かり固有のラベルと感性ラベルが付与される。表2は、表1の感性手掛かりを用いた場合にテキストへどのようにラベルが設定されるかをまとめている。
- ステップ3：教師ありトピックモデリング
  ステップ2でラベルが付与されたテキストに対してPLDAを適用する。感性トピックは、ステップ1で定義された感性手掛かりの感性と結びついて抽出される。

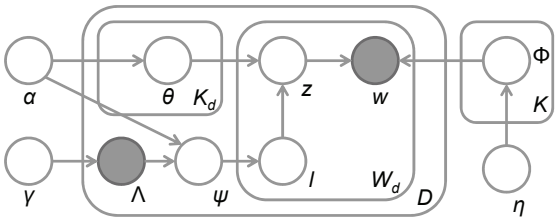

図2　PLDAのグラフィカルモデル。影付けされたノードは観測される要素を意味する。
The graphical model of PLDA. Shaded elements represent observed elements.

それぞれのトピック $k \in \{1...K\}$
　選択 $\Phi_k \sim \mathrm{Dir}(\eta)$
それぞれの文書 $d \in \{1...D\}$
　それぞれの文書ラベル $j \in \Lambda_d$
　　選択 $\theta_{dj} \sim \mathrm{Dir}(\alpha)$
　選択 $\psi_d \sim \mathrm{Dir}(\alpha)$
　それぞれの単語 $w \in \mathbf{W}_d$
　　選択 $l \sim \mathrm{Mult}(\psi_d)$
　　選択 $z \sim \mathrm{Mult}(\theta_{dl})$
　　選択 $w \sim \mathrm{Mult}(\Phi_z)$

図3　PLDAの生成プロセス
The generation process of PLDA

表1　感性手掛かりの例
Examples of sentiment clues

手掛かりID	手掛かり表現	感性
Happy Face	:–)	ポジティブ
Sad Face	:-(	ネガティブ

表2　感性手掛かりを用いたラベル付与の例
An example of how labels are set to texts

テキスト種別	Happy Face ラベル	ポジティブ ラベル	ネガティブ ラベル	Sad Face ラベル
Happy Face を含む	✓	✓		
Sad Face を含む			✓	✓
手掛かりを含まない		✓	✓	

3. データ

3.1 モティコンリスト

感性手掛かりを設定するために、エモティコンの調査を行った。日本語で広く利用されている6種類のエモティコンを選択し、各エモティコンを含むツイートを50件ずつTwitter®より収集した。収集した合計300のツイートに対し、3人のアノテーター*6がポジティブ、ネガティブ、ポジティブ・ネガティブ両方、ニュートラルのいずれかの感性を付与した。表3に、各エモティ

*4 エモティコンとは、顔文字、絵文字等の感情を表現するために用いられるテキスト表現を意味する。
*5 ソーシャルタグとは、ハッシュタグ等のソーシャルコミュニティーのユーザーが分類等に用いるタグを意味する。
*6 データに対して何らかの付加情報を付与する行為をアノテーションといい、それを行う人をアノテーターという。

表3　6種類のエモティコンと一致数が最大となった感性
Six emoticons and their largest vote polarities

エモティコン	感性
(´ ▽ `)ﾉ	ポジティブ
＼(^ o ^)／	
(^ −^)	
orz	ネガティブ
(´ ・ω・`)	
(> _ <)	

コンについて2人以上のアノテーターが一致した数が最も多かった感性を示す。

3.2　トピックモデリングデータ

トピックモデリングの対象データとしてツイートを用いた。2011年5月から2011年8月の期間にTwitter®の“public stream”ツイートをStreaming APIを用いて収集し、収集したツイートから次の3つの条件のいずれかを満たす220,000ツイートをサンプリングした。

- HAPPY
  エモティコンの“(´ ▽ `)ﾉ”（以降、EMO-HAPPY）を含む10,000ツイート。
- SAD
  エモティコンの“orz”（以降、EMO-SAD）を含む10,000ツイート。
- NO-EMO
  エモティコンを含まない*7 200,000ツイート。この条件においては、重複ツイートや内容のないツイートを減らすために、5単語以上より構成される、リツイートではないという制約も加えている。

NO-EMOのサンプリングを行う場合には、日本語形態素解析器のKuromoji[14]を用いてツイートを単語単位に分割した。表4にサンプリングしたツイートの概要を示す。

表4　トピックモデリングデータの概要
The summary of the topic modeling data

条件	ツイート数
HAPPY	10,000
SAD	10,000
NO-EMO	200,000
合計	220,000

3.3　極性判定評価データ

提案手法の評価の1つとして、極性*8判定性能を評価した（詳細は4章で述べる）。そのための評価データとしては、“ツイート”と“新聞”の2種類のデータを用意した。ツイートは、トピックモデリングデータと同様のTwitter®からのランダムサンプリングデータであり、多様な分野のテキストを含んでいる。新聞は、ニュース分野のテキストであり、ツイートとはかなり性質が異なる。

3.3.1　ツイート

以下の3つの条件を満たす3,000ツイートを、トピックモデリングデータと同じ2011年5月から2011年8月中のツイートよりサンプリングした。

a. ツイートが5単語以上で構成（NO-EMOと同じ条件）。
b. ツイートが形容詞、副詞、連体詞、名詞-副詞可能のいずれかを含む。この条件は、何らかの評価を含むツイートをサンプリングしやすくするように設定した。
c. 特定の品詞がツイートを構成する単語の80%以上を占めない。この条件は、名詞の列挙や特定の文字の連続が出現するツイートを除外するために設定した。

単語の品詞は、Kuromojiをツイートに対して適用し、その解析結果より取得した。

サンプリングされた3,000ツイートに対して、次の6種類のラベルのいずれかを設定した。

*7 複数のウェブサイトから収集した10,924個のエモティコンを判定に用いた。

*8 何らかの“極”に基づく性質を意味し、本稿ではポジティブ、ネガティブの極を意味する。

- ポジティブ、ネガティブ、ポジティブ・ネガティブ両方、ニュートラル、広告、解釈不能

“広告”ラベルは、広告内容のツイートをポジティブと判定しないように設定した。“解釈不能”ラベルは、文脈に強く依存し単独では解釈が困難なツイートを除外するために設定した。6種類のラベルの付与は、18人のアノテーターが10組[*9]を構成して行った。2人のアノテーターがポジティブもしくはネガティブで一致した723ツイートをアノテーション結果より抽出し、極性判定評価データとした。表5の“ツイート”は、本評価データにおける各感性のツイート数をまとめている。

3.3.2 新聞

NTCIR-7 Multilingual Opinion Analysis Task（MOAT）[15]の日本語セクションのデータを用いた。日本語セクションのデータは7,163文のニューステキストより構成されており、3人のアノテーターにより文単位で極性が付与されている。極性判定評価データとして、このデータより以下の条件を満たす434文を抽出した。

- 2人以上のアノテーターがポジティブもしくはネガティブな文として合意したもの。

表5の“新聞”は、本評価データにおける各感性の文数をまとめている。

表5　極性判定評価データの構成
The compositions of the polarity classification evaluation data

種類	感性	データ数
ツイート	ポジティブ	384
	ネガティブ	339
新聞	ポジティブ	107
	ネガティブ	327

[*9] 10ペアを構成するのに2人足りないため、2人のアノテーターは2つのペアに参加している。

4. 実験

提案手法の性能を確認するために、実験および二通りの評価を実施した。

4.1 感性手掛かり

表6に示される感性手掛かりを実験では用いた。なお、3.2節で述べたトピックモデリングデータのサンプリング条件に感性手掛かりと同じエモティコンを用いているため、EMO-HAPPYとEMO-SADを含むツイートは各10,000ツイートずつトピックモデリングデータに含まれている。

4.2 前処理

トピックモデリングデータのテキストから語を抽出するときにいくつかの前処理を実施した。

1）次のテキスト正規化処理を実施：Unicode正規化　Form NFKC[16]、3文字以上の“w”の連続を“ww”に置換、Twitter®のユーザー名（例. @user）を“USER”に置換、ハッシュタグ（例. #hashtag）を“HASHTAG”に置換、URL（例. http://example.org）を“URL”に置換。
2）テキストをKuromojiで解析し、単語とその品詞を取得。
3）次の品詞に属さない単語を削除：名詞[*10]、動詞、形容詞、副詞、連体詞、感嘆詞、フィラー、記号-アルファベット、未知語。
4）日本語で頻出する以下の単語をストップワードとして設定し削除：“する”、“なる”。
5）トピックモデリングの語として形態素解析結果の原形を取得。
6）トピックモデリングデータ中に一度しか出現しなかった語を削除。

表6　実験で用いた感性手掛かり
The sentiment clues used in the experiment

感性手掛かり	感性
EMO-HAPPY	ポジティブ
EMO-SAD	ネガティブ

[*10] 名詞-接尾等の一部例外あり。

表7　ラベルごとのトピック数
The number of topics set to each labels

ラベル	トピック数
ポジティブ	50
ネガティブ	50
EMO-HAPPY	1
EMO-SAD	1
background	1

4.3　教師ありトピックモデリング

PLDAの実装としてStanford Topic Modeling Toolbox[17]を用いた。ラベルごとのトピック数は表7の値に設定した。表中の“background”は感性ラベルと独立して単語を生成できる特別なトピックを用意するために設定した。教師ありトピックモデリングでは、このようなトピックを設定することにより、文脈に依存しないトピックを抽出できることが知られている[4]。

PLDAのパラメーターは、前処理されたデータを教師データとして、Collapsed Variational Inference[18]で繰り返し回数をStanford Topic Modeling Toolboxのデフォルト値に設定して推定した。図4に抽出されたトピックの例を示す。

4.4　評価

4.4.1　トピックの定量評価

感性トピック抽出性能の定量評価として、極性判定性能を評価した。この評価は感性トピックの抽出性能を直接評価するものではないが、感性トピックの抽出を行う既存の研究[19), 20)]にならい実施した。

ラベル	語
EMO-HAPPY	(´▽`)ノ, USER, ない, ん, ?, の, w, ww, 笑, …, …
EMO-SAD	orz, USER, !, ー, …, °, д, 行く, …
ポジティブ #11	食べる, 美味しい, 飲む, 屋, 料理, ラーメン, 店, コーヒー, 肉, …
ポジティブ #30	!, ありがとう, よろしく, お願い, くださる, イイ, これから, 楽しむ, できる, …
ネガティブ #2	～さ, 暑い, 夏, この, その, 中, 今日, 風, 外, 汗, …
ネガティブ #48	くる, 目, 痛い, 入る, 風呂, 寝る, 頭, お腹, すぎる, ない, …
…	

図4　感性トピック抽出の例。図1と同様に各単語にはトピックからの生成確率が設定されている。
Examples of extracted sentiment topics. Like in Figure 1, each word has a generation probability from its topic.

4.3節で学習したモデルを用いて、3.3節で述べた極性判定評価データに対して文書-トピック推定を行った。推定結果に対し、以下の式（1）に基づき各ツイートのポジティブとネガティブのスコアを計算した：

$$\mathrm{score}(d,l)=\sum_{t_l} P(t_l|d) \quad (1)$$

dは文書（ツイート）、lはラベル（ポジティブもしくはネガティブ）、t_lはlのトピック、P（t_l|d）はdが選択されたという条件のうえでのt_lの事後確率である。ツイートのラベルは、式（1）を最大化するものを設定した。

提案手法のベースラインとして、Goらの手法[11]にならったサポートベクトルマシン（以降、SVM）に基づく極性判定器を用意した。3.2節のHAPPY条件のツイートをポジティブ、SAD条件のツイートをネガティブの学習データとして、Goらの手法[11]で最も高いaccuracyが得られたunigram素性のみでSVMを学習した。データの前処理には、基本的には提案手法と同じものを用いたが、EMO-HAPPYとEMO-SADの2つのエモティコンをストップワードに追加した。SVMの実装としてはLIBLINEAR[21]を用い、デフォルト設定のL2-loss linear SVMおよびコストパラメーターC=1.0を利用した。

表8に極性判定結果を示す。表中のMajority Baselineは全ての判定結果を頻出するラベルに設定したものであり、“ツイート”はポジティブで“新聞”はネガティブに設定している。提案手法は“ツイート”ではaccuracyでベースラインの70.5%に近い70.1%が得られた。“新聞”ではベースラインの71.2%に対して69.1%と低かったが、“新聞”ではMajority Baselineで最も高い75.3%が得られている。

表8　極性判定結果
The polarity classification results

種類	手法	Accuracy
ツイート	Majority Baseline	53.1%
	SVM	70.5%
	提案手法	70.1%
新聞	Majority Baseline	75.3%
	SVM	71.2%
	提案手法	69.1%

4.4.2　トピックの定性評価

4.4.1節の定量評価では、感性トピックを極性判定という別問題で評価した。より直接的な定性評価として、2人の評価者が提案手法により抽出された50のポジティブトピックと50のネガティブトピックを評価した。

評価者はそれぞれのトピックについて最も確率的に関連の強い40の語と20のトピックを提示された。関連の強い語は、トピック-語分布$P(w|t_i)$の上位語を単純に選択した。関連の強いツイートは、まず文書-トピック分布をトピックモデリングデータに対して計算し、それぞれのトピックt_iについて、$P(t_i|d)$の上位ツイートを選択した。

評価者はそれぞれのトピックに対して、ポジティブ、ネガティブ、解釈不能のいずれかのラベルを設定した。“解釈不能”は例外的なラベルであり、関連する語もしくはツイートが以下の条件のいずれかを満たすものに付与した：（a）大半が日本語でない、（b）大半が感嘆語もしくはオノマトペ、（c）大半がニュートラル。

作成したデータのうち、2人の評価者のラベルがポジティブもしくはネガティブで一致した59トピックのaccuracyを計算した。表9に評価結果を示す。評価結果は、全体のaccuracyで72.9%が得られた。

表9　50ポジティブ、50ネガティブトピックの表結果。#Pと#Nは評価者がポジティブ、ネガティブと判定した数であり、#PNは2つの合計である。
The evaluation result of the 50 positive topics and the 50 negative topics. #P and #N are the numbers of topics that the two evaluators labeled, and #PN are the summations of #P and #N

ラベル	#P	#N	#PN	Accuracy
ポジティブ	24	3	27	88.9%
ネガティブ	13	19	32	59.4%
全体	37	22	59	72.9%

5.　まとめ

本稿では、感性トピックを抽出する手法を提案した。定量評価ではツイートデータで70.1%のaccuracyが得られ、新聞データで69.1%のaccuracyが得られた。これらはSVMによるベースラインの70.5%と71.2%に近い性能である。より直接的な定性評価では、全体で72.9%のaccuracyが得られた。結果は提案手法により感性トピックが抽出できていることを示唆している。

提案手法の特徴として、僅かな感性手掛かりを定義するのみで感性トピックの抽出を実現できる点がある。このため、今回対象としたツイートおよび新聞以外のテキストに対しても、低コストで提案手法を適用できる。提案手法の今後の課題としては次の2点を検討している。

1）評価側面トピックの抽出
本稿では感性トピックを抽出する手法について述べた。提案手法は手掛かりを定義できれば、感性以外のトピックへの拡張ができる。例えば、Twitter®等ではハッシュタグを用いたテキストの分類がユーザーにより行われている[4)]。今後の拡張として、ハッシュタグ等のソーシャルタグを用いて特定の評価側面のトピックを抽出することを検討している。

2）ノンパラメトリックベイズ手法の導入
提案手法の実験では、ポジティブのトピック数とネガティブのトピック数を同数に設定した。しかし、感性がどのように分布するかは分野依存であり、同数設定が最良であるとはかぎらない。今後の拡張として、最適なトピック数を自動的に決定できるノンパラメトリックベイズの手法[4), 22)]の導入を検討している。

6. 商標について

7. 参考文献

1) D.M. Blei, A.Y. Ng, and M.I. Jordon, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp. 993-1022 (2003).
2) T.L. Griffiths and M. Steyvers, “Finding Scientific Topics”, Proceedings of the National Academy of Sciences, Vol.101 (Suppl 1), pp. 5228-5235 (2004).
3) Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, “Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs”, Proceedings of the 16th International Conference on World Wide Web, pp. 171-180 (2007).
4) D. Ramage, S. Dumais, and D. Liebling, “Characterizing Microblogs with Topic Models”, Proceedings of the Fourth International AAAI Conference on We Blogs and Social Media, pp. 130-137 (2010).
5) J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D.M. Blei, “Reading Tea Leaves: How Humans Interpret Topic Models”, Neural Information Processing Systems, Vol.22, pp. 288-296 (2009).
6) D.M. Blei and J.D. McAuliffe, “Supervised Topic Models”, Neural Information Processing Systems, Vol. 20, pp. 121-128 (2007).
7) D. Ramage, D. Hall, R. Nallapati, and C.D. Manning, “Labeled LDA: a Supervised Topic Model for Credit Attribution in Multi-labeled Corpora”, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 248-256 (2009).
8) D. Ramage, C.D. Manning, and S. Dumais, “Partially Labeled Topic Models for Interpretable Text Mining”, Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 457-465 (2011).
9) I. Titov and R. McDonald, “A Joint Model of Text and Aspect Ratings for Sentiment Summarization”, Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies, pp. 308-316 (2008).
10) 奥村 学, “マイクロブログマイニングの現在”, 電子情報通信学会技術研究報告, NLC, 言語理解とコミュニケーション Vol.111, No.427, pp. 19-24 (2012).
11) A. Go, R. Bhayani, and L. Huang, “Twitter Sentiment Classification With Distant Supervision”, Technical Report, Stanford University (2009).
12) J. Read, “Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification”, Proceedings of the ACL Student Research Workshop, pp. 43 - 48 (2005).
13) D. Davidov, O. Tsur, and A. Rappoport, “Enhanced Sentiment Learning Using Twitter Hashtags and Smileys”, Proceedings of the 23rd International Conference on Computational Linguistics, pp. 241-249 (2010).
14) http://www.atilika.org/ [Kuromoji (Atilika)]
15) Y. Seki, D.K. Evans, L-W. Ku, L. Sun, H-H. Chen, and N. Kando, “Overview of Multilingual Opinion Analysis Task at NTCIR-7”, Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access

Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access, pp. 185－203 (2008).
16) http://unicode.org/reports/tr15 [UAX # 15 Unicode Normalization Forms (Unicode)]
17) http://www-nlp.stanford.edu/software/tmt/tmt-0.4/ [Stanford Topic Modeling Toolbox (The Stanford Natural Language Processing Group)]
18) A. Asuncion, M. Welling, P. Smyth, and Y.W. Teh, "On Smoothing and Inference for Topic Models", Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 27－34 (2009).
19) C. Lin, Y. He, . Everson, and S. Rüger, "Weakly Supervised Joint Sentiment-Topic Detection from Text", IEEE Transaction on Knowledge and Data Engineering, Vol.24(Issue 6), pp. 1134-1145 (2012).
20) Y. Jo and A. Oh, "Aspect and Sentiment Unification Model for Online Review Analysis", Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 815-824 (2011).
21) http://www.csie.ntu.edu.tw/~cjlin/liblinear/ [LIBLINEAR (Machine Learning Group at National Taiwan University)]
22) D.M. Blei and M.I. Jordan, "Variational Inference for Dirichlet Process Mixtures", Bayesian Analysis, Vol.1, pp. 121-144 (2005).

筆者紹介

三浦　康秀
研究技術開発本部 コミュニケーション技術研究所に所属
専門分野：自然言語処理

服部　圭悟
研究技術開発本部 コミュニケーション技術研究所に所属
専門分野：自然言語処理

大熊　智子
研究技術開発本部 コミュニケーション技術研究所に所属
専門分野：自然言語処理、国語学

増市　博
研究技術開発本部 コミュニケーション技術研究所に所属
専門分野：自然言語処理