文書分類のための Negation Naive Bayes

古宮嘉那子[†]・伊藤 裕佑^{††}・佐藤 直人^{†††}・小谷 善行[†]

本論文は、文書分類のための新手法として、Negation Naive Bayes (NNB) を提案する。NNB は、クラスの補集合を用いるという点では Complement Naive Bayes (CNB) と等しいが、Naive Bayes (NB) と同じ事後確率最大化の式から導出されるため、事前確率を数学的に正しく考慮している点で異なっている。NNB の有効性を示すため、オークションの商品分類の実験とニュースグループの文書分類の実験を行った。ニュースグループの文書分類では、一文書あたりの単語数（トークン数）を減らした実験と、クラスごとの文書数を不均一にした実験を行い、NNB の性質を考察した。NB, CNB, サポートベクターマシン (SVM) と比較したところ、特に一文書あたりの単語数が減り、クラスごとの文書数が偏る場合において、NNB が他の Bayesian アプローチより勝る手法であること、また、時には SVM を有意に上回り、比較手法中で最も良い分類正解率を示す手法であることが分かった。

キーワード：文書分類, Naive Bayes 分類器, Complement Naive Bayes 分類器

Negation Naive Bayes for Text Classification

KANAKO KOMIYA[†], YUSUKE ITO^{††}, NAOTO SATO^{†††} and YOSHIYUKI KOTANI[†]

In this study, we proposed negation naive Bayes (NNB), a new method for text classification. Similar to complement naive Bayes (CNB), NNB uses the complement class. However, unlike CNB, NNB properly considers the prior in a mathematical way because NNB is derivable from the same equation (the maximum a posteriori equation) from which naive Bayes (NB) is derived. We carried out classification experiments on products offered on an internet auction site and on the 20 Newsgroups data set. For the latter, we carried out experiments in the following two settings and discussed the properties of NNB: (1) settings in which the number of words in each document decreases and (2) settings in which the distribution of documents over classes is skewed. We compared NNB with NB, CNB, and support vector machine (SVM). Our experiments showed that NNB outperforms other Bayesian approaches when the number of words in each document decreases and when texts are distributed non-uniformly over classes. Our experiments also showed that NNB sometimes provides the best accuracy and significantly outperforms SVM.

Key Words: *text classification, Naive Bayes Classifier, Complement Naive Bayes Classifier*

[†] 東京農工大学工学研究院, Institute of Engineering, Tokyo University of Agriculture and Technology

^{††} 東京農工大学工学部, Faculty of Engineering, Tokyo University of Agriculture and Technology

^{†††} 東京農工大学工学府, Graduate School of Engineering, Tokyo University of Agriculture and Technology

1 はじめに

文書分類において Naive Bayes (NB) を利用するのは極めて一般的である。しかし、多項モデルを用いた NB 分類器では、クラス間の文書数に大きなばらつきがある場合に、大きく性能が下がるという欠点があった。そのため、Rennie, Shih, Teevan, and Karger (2003) は「クラスに属する文書」ではなく「クラスに属さない文書」、つまり「補集合」を用いることにより NB の欠点を緩和した Complement Naive Bayes (CNB) を提唱した。しかし、CNB は NB と同じ式、つまり事後確率最大化の式から導くことができない。そこで我々は、事後確率最大化の式から導くことのできる Negation Naive Bayes (NNB) を提案し、その性質を他の Bayesian アプローチと比較した。その結果、クラスごとの単語数（トークン数）が少なく、なおかつクラス間の文書数に大きなばらつきがある場合には分類正解率が NB、CNB をカイ二乗検定で有意に上回ること、また、これらの条件が特に十分に当てはまる場合には、事前確率を無視した CNB も同検定で有意に上回ることを示す。また、NNB は、Bayes 手法以外の手法であるサポートベクターマシン (SVM) よりも、時に優れた結果を示した。

本稿の構成は以下のようになっている。まず 2 節で Bayes 手法のテキスト分類の関連研究について紹介する。3 節では提案手法である NNB の導出について述べる。4 節では本研究で用いたデータと実験方法について述べ、5 節に結果を、6 節に考察を、7 節にまとめを述べる。

2 関連研究

これまでに数多くの文書分類に関する研究がなされており、これらの中でも、Bayes の手法はよく用いられている (持橋 2006)。

井筒、横澤、篠原 (2005) は html ファイルの自動分類で NB を使用し、ルールベースの手法や判別分析に比べて、プログラム開発者にかかる負担の低さとスケーラビリティの高さを指摘している。McCallum and Nigam (1998) は NB を適用してテキスト分類を行う際に使用する事象モデルとして、多項モデルと多変量ベルヌーイモデルの違いを述べ、分類結果から多項モデルの優位性を示している。Lewis (1992) は、文書分類問題において単語と句、単語と句のクラスタ化の有無、全ての索引語を用いるか一部を用いるかの違いについて、分類精度の比較を行った。花井、山村 (2005) は、NB が前提とする単語間の独立が成り立たないとし、依存性の強い 2 単語の組が同時に生起する確率を NB に適用することによって分類精度の向上を図った。Church (2000) は、単語の重み付け方法として IDF 値のかわりに “Adaptation” という概念を用い、文書に含まれていないが内容に関連している単語を “Neighbor” として定義して、テキスト内の特徴的な単語の抽出を行った。さらに持橋、菊井 (2006) は、NB におけるクラス c を未知として拡張した DM (Dirichlet Mixtures) や Infinite DM を提案し、高村、Roth (2007) は、予測的尤度

を用いて超変数を設定することなく, 加算スムージングのパラメータを求めた.

NB を発展させた研究として, 補集合を用いて学習を行う CNB が有名である (Rennie et al. 2003). CNB は, 「クラスに属する文書」ではなく「クラスに属さない文書」, つまり「補集合」を用いることにより NB の欠点を緩和した手法である.

本研究では, 多項モデルを用いた NB と CNB に注目し, その分類における特徴を考慮して, Bayes のアプローチを用いた新しい分類手法 NNB を提案する.

3 Negation Naive Bayes の導出

NNB は CNB と同様に補集合を利用して文書分類を行うが, CNB と異なって NB と同じ事後確率最大化の式から導出が可能である. その結果, 事前確率を数学的に正しく考慮することで, クラスごとの文書数が異なっているときにもより正確な文書分類を行えるようにした. 本節では NB の導出と, CNB の概念について触れた後, 提案手法である NNB の導出について述べる.

3.1 Naive Bayes 分類器

一般に確率モデルによる文書分類では, 分類対象となる文書を d , ある一つのクラスを c としたとき, 事後確率 $P(c|d)$ を最大化するクラス $\hat{c}$ を求める (Zhang 2004). NB 分類器を用いた文書分類では, 事後確率に Bayes の定理を適用する. 文書の取り出される確率 $P(d)$ はすべてのクラスについて一定であることを考慮すると, 事後確率が最大のクラスを推定することは, クラスの出現確率 $P(c)$ と各クラスでの文書の出現確率 $P(d|c)$ の積を最大化するクラスを推定することと等しくなる.

$$\begin{aligned}\hat{c} &= \operatorname{argmax}_c P(c|d) \\ &= \operatorname{argmax}_c \frac{P(c)P(d|c)}{P(d)} \\ &= \operatorname{argmax}_c P(c)P(d|c)\end{aligned}\tag{1}$$

式 (1) において, $P(c)$ は全文書中でクラス c に属する文書の割合を用いて容易に推定ができるが, $P(d|c)$ を直接推定するのは難しい. そこで, まず文書 d を単語列 $w_1, w_2, \dots, w_n$ で近似する.

$$P(d|c) \approx P(w_1, w_2, \dots, w_n|c)\tag{2}$$

次に, 各クラスで単語が独立に生起すると仮定すると, 式 (2) は

$$P(w_1, w_2, \dots, w_n|c) \approx \prod_{i=1}^n P(w_i|c)\tag{3}$$

と近似される.

したがって, d の属するクラス $\hat{c}$ は最終的に以下の式で求められる.

$$\hat{c} = \underset{c}{\operatorname{argmax}} P(c) \prod_{i=1}^n P(w_i|c) \quad (4)$$

3.2 Complement Naive Bayes 分類器

多項モデルを用いた NB 分類器では, クラス間の文書数に大きなばらつきがある場合に, 文書数の小さいクラスで $P(w_i|c)$ が大きくなる傾向がある. $P(w_i|c)$ は「そのクラス中に出てきたそのトークン w_i の数/そのクラス中に出てきたそのトークンの総数」であるため, 訓練事例の単語トークン数に大きな差ができた結果, 大きいクラスの $P(w_i|c)$ は比較的小さく, 小さいクラスの $P(w_i|c)$ はかなり大きくなることが予想できる. その結果, 小さいクラスに出現した単語を含む文書が出現した場合, その文書は, その単語をもつ小さなクラスに割り当てられることになる.

また, 文書数の少ないクラスでは, 新規文書に出現した単語がそのクラスに含まれていない割合が多くなり, データがスパースになりやすい.

そこで, 学習する文書数のばらつきを抑え, スパースネス問題を緩和するよう NB を改良したのが Rennie et al. (2003) の CNB である. 具体的には, 「クラス c に属す訓練事例」ではなく「クラス c に属さない訓練事例」すなわち「 $\bar{c}$ に属する訓練事例 (補集合)」を用いて学習を行う¹.

図 1 は, NB と CNB での学習に用いる文書数の違いを表している. 文書数 10, 10, 20, 40 の 4 つのクラスがある場合, NB ではこの文書数を自身のクラスの学習に使う. そのため, 文書数が最も少ないクラスと最も多いクラスでは学習に使用する文書数に 4 倍の差がある.

図 1 NB と CNB での学習に用いる文書数の違い

¹ Rennie らは文献の中で CNB のほかに 5 種類のヒューリスティックを導入しているが, 本研究では純粋に式の変更による違いを見るため, ヒューリスティックは使用しなかった.

一方, CNB では自身のクラスに属する文書以外の文書から学習を行うため, 学習に用いる文書数は最小のもので 40, 最大のもので 70 となり, NB に比べてばらつきが小さくなる.

CNB は, 文書内にある単語の出現確率の積から尤度を計算し, 分類するクラスを決めるという点では NB と同じである. つまり, 式 (4) を用いて文書 d の属するクラス $\hat{c}$ を推定する. しかし, CNB では $P(w_i|c)$ を最尤推定で求めるのではなく, c 以外のクラス $\bar{c}$ の尤度の積から推定する. つまり, d の属するクラス $\hat{c}$ は最終的に以下の式で求められる.

$$\hat{c} = \operatorname{argmax}_c P(c) \prod_{i=1}^n \frac{1}{P(w_i|\bar{c})} \quad (5)$$

3.3 Negation Naive Bayes 分類器

前節で説明した CNB は, NB の持つ「クラス間の文書数のばらつきによって分類結果が偏る」という特徴を緩和する手法である. しかし, CNB はヒューリスティックによる解決法であって, 事後確率最大化の式から導出することはできない.

そこで本研究では, 事後確率最大化の式から導出でき, かつ, CNB の「訓練にクラスの補集合を利用する」という長所をもつ分類器を作成する. 以下で NB と同様の, 事後確率最大化の式 (式 (1)) からの式の変形について述べる.

まず, 事後確率 $P(c|d)$ を最大化するクラス $\hat{c}$ を求める式を補集合を利用するように変形する.

$$\begin{aligned} \hat{c} &= \operatorname{argmax}_c P(c|d) \\ &= \operatorname{argmax}_c (1 - P(\bar{c}|d)) \\ &= \operatorname{argmin}_c P(\bar{c}|d) \end{aligned} \quad (6)$$

次に, Bayes の定理を用いて式 (6) を変形する.

$$\begin{aligned} \hat{c} &= \operatorname{argmin}_c \frac{P(\bar{c})P(d|\bar{c})}{P(d)} \\ &= \operatorname{argmin}_c P(\bar{c})P(d|\bar{c}) \end{aligned} \quad (7)$$

そして, 式 (7) を近似する. $P(d|\bar{c})$ は式 (2), (3) と同様に

$$P(d|\bar{c}) \approx \prod_{i=1}^n P(w_i|\bar{c}) \quad (8)$$

と近似される. したがって, 文書 d の属するクラス $\hat{c}$ を以下の式で推定する.

$$\hat{c} = \operatorname{argmin}_c P(\bar{c}) \prod_{i=1}^n P(w_i|\bar{c}) \quad (9)$$

なお, $P(\bar{c}) = 1 - P(c)$ であり, CNB と同じく最大化で表現すると以下の式になる.

$$\hat{c} = \operatorname{argmax}_c \frac{1}{1 - P(c)} \prod_{i=1}^n \frac{1}{P(w_i|\bar{c})} \quad (10)$$

式 (5) と比較すると, $\frac{1}{1-P(c)}$ の最大化の部分, つまり事前確率 $P(c)$ の部分が異なっていることが分かる. 式 (10) は事後確率最大化の式から求められたため, 事前確率を数学的に正しく考慮した式となっている. なお, Rennie らの研究では, 式 (5) において事前確率の扱いについてあまり注意を払っていないが, 我々はクラスごとに単語数の偏りが大きいデータセットについて分類を行う場合には, $P(c)$ を利用するか $\frac{1}{1-P(c)}$ を利用するかの影響は必ずしも無視して良いとは言えないと考える. また, Rennie らの研究では, $P(c)$ は $P(w_i|\bar{c})$ に比べて分類結果への影響が小さいと判断し, 事前確率は計算してもしなくても結果は同じと考え, 実際の分類には $P(c)$ を無視して $P(w_i|\bar{c})$ のみを計算しているため, $P(c)$ なしの CNB についても参考として実験を行う.

4 実験

NNB の性能を測りその特色を調べるため, ふたつのコーパスを用いた実験を行った. 一つ目はオークションの商品分類の実験であり, 二つ目はニュースグループの文書分類の実験である. これらの二つの実験において, 形態素解析により得た表層形の bag-of-words を用いて NB, CNB, NNB の文書分類の性能を比較した. また, NNB と CNB の差は事前確率 $P(c)$ と $\frac{1}{1-P(c)}$ の部分であるため, CNB と NNB から第一項を省略した形の式 ($P(c)$ なしの CNB) でも実験を行った.

なお, 形態素解析ソフトには MeCab²を利用した.

また, スムージング手法としては, 予備実験によりラプラススムージング (marquis de Laplace 1814, 1995) と Jeffreys Perks 法 (Box and Tiao 1973) を試し, Jeffreys Perks 法の方が結果がよかったため, これを採用した.

さらに, Bayes ではない手法の比較対象として, SVM についても実験を行った. この際, 分類器としてはマルチクラス対応の SVM (libsvm³) を使用した. カーネルは予備実験の結果, 線形カーネルが最も高い正解率を示したため, これを採用した. また, 学習の素性は Bayes の手法とそろえ, 表層形の bag-of-words の頻度ベクトルを使用した. すべての実験には五分割交差検定を用いた.

² <http://mecab.sourceforge.net/>

³ <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

4.1 オークションの商品分類の実験

オークションの商品分類の実験は, Yahoo! オークション⁴の商品タイトルを商品カテゴリに分類する実験である. 詳細は佐藤, 藤本, 小谷 (2010) になった.

本実験では, 「デスクトップ」「記念切手」「赤ちゃん用の玩具」の三つのジャンルカテゴリ⁵に含まれる商品を対象に実験を行った. これらのジャンルカテゴリは, 以下のようにトップカテゴリ (オークション) から絞り込むことができる.

- オークション > コンピュータ > パソコン > Windows > デスクトップ
- オークション > アンティーク, コレクション > 切手, 官製はがき > 日本 > 特殊切手, 記念切手
- オークション > おもちゃ, ゲーム > ベビー用

ここで, > の左が親カテゴリ, 右が子カテゴリを示す. 「デスクトップ」と「記念切手」は2012年6月26日に, 「赤ちゃん用の玩具」は2012年6月29日に取得したデータである. ひとつの商品はひとつの葉カテゴリにのみ属しているものとし, 出品者によって登録されたカテゴリを正しいカテゴリとして実験を行った. なお, 例えばジャンルカテゴリが「デスクトップ」の場合, 「ASUS の 1,000 円~1,099 円」や「IBM パソコン単体 31,000 円~34,999 円」といったものが葉カテゴリである. また, 佐藤 他 (2010) にならい, 出品者個人による商品情報表記の癖などの偏りをなくするため, 各カテゴリにつき 1 人の出品者の商品は 1 つしか使用しないものとし, 商品タイトルの全単語を利用した実験と名詞のみを使用した二つの実験を行った.

表 1 にそれぞれのジャンルカテゴリ中の葉カテゴリ数, 同じ出品者による商品をひとつにする前と後の商品数を示す. ここで, 商品数 (処理前), 商品数 (処理後) はそれぞれ, 同じ出品者による商品をひとつにする前と後の商品数である. なお, 「赤ちゃん用の玩具」の商品数 (処理後) には 8205, 8204 と二つ数字があるが, 8205 は全ての品詞の分類, 8204 は名詞のみの分類による商品数を示している. 「赤ちゃん用の玩具」のデータには, 形態素解析の結果, 全ての形

表 1 それぞれのジャンルカテゴリ中の葉カテゴリ数, 同じ出品者による商品をひとつにする前と後の商品数

ジャンルカテゴリ	葉カテゴリ数	商品数 (処理前)	商品数 (処理後)
デスクトップ	59	31419	1975
記念切手	51	24399	1808
赤ちゃん用の玩具	45	44789	8205, 8204

⁴ <http://auctions.yahoo.co.jp/>

⁵ オークション > コンピュータ > パソコン > Windows > デスクトップの例では, 「デスクトップ」だけでなく, 「コンピュータ」や「パソコン」, 「Windows」もジャンルカテゴリであるが, 本実験ではデータサイズの観点から「デスクトップ」を対象として実験を行った.

態素が名詞以外の品詞に割りつけられた商品が1件あったため、このような結果になっている。

また、図2、図3、図4にそれぞれ「デスクトップ」、「記念切手」、「赤ちゃん用の玩具」のカテゴリごとの文書数、単語トークン数、名詞のトークン数を折れ線グラフにしたものを示す。横軸のカテゴリ index は、カテゴリを文書数で並べ替えたときに降順につけたものである。これらの図から、カテゴリの分類実験において、カテゴリごとに文書数、トークン数が非常に偏っていることが分かる。特に、「記念切手」が最も偏っていることが読みとれる。

また、表2にオークションのカテゴリ分類実験の訓練事例中の商品数の標準偏差、訓練事例中の文書数の標準偏差を訓練事例中の総文書数の平均で割った値（以降、標準偏差／平均と表記）、トークン数、1商品当たりの平均単語数を示す。標準偏差／平均は、カテゴリ（またはクラス）ごとの商品数（または文書数）のばらつきを見るために示した。ばらつきを測るのには

図2 「デスクトップ」のカテゴリごとの文書数、単語トークン数、名詞のトークン数

図3 「記念切手」のカテゴリごとの文書数、単語トークン数、名詞のトークン数

図 4 「赤ちゃん用の玩具」のカテゴリごとの文書数, 単語トークン数, 名詞のトークン数

表 2 オークションのカテゴリ分類実験の訓練事例中の商品数の標準偏差, 標準偏差/平均, トークン数, 1 商品当たりの平均単語数

ジャンルカテゴリ	デスクトップ	記念切手	赤ちゃん用の玩具
訓練事例中の商品数の標準偏差	46.40	100.78	241.81
訓練事例中の商品数の標準偏差/平均	1.44	2.84	1.33
訓練事例中の全単語のトークン数	28,221	23,760	90,104
1 商品当たりの平均の全単語のトークン数	14.29	13.14	10.98
訓練事例中の名詞トークン数	24,737	18,310	63,705
1 商品当たりの平均の名詞トークン数	12.53	10.13	7.77

標準偏差を用いるのが一般的であるが, 本実験のように全商品数 (または文書数) が異なる実験設定同士を比べる際には, 全商品数 (文書数) の絶対値が大きくなるにつれて標準偏差も大きくなるという問題があったためである⁶. 標準偏差を平均で割ることによって, 全商品数 (文書数) のスケールに左右されないデータのばらつきを測った. 表 2 のこの「訓練事例中の商品数の標準偏差/平均」を見てみると, 「記念切手」の値が「デスクトップ」や「赤ちゃん用の玩具」より高いことから, 図 2～図 4 から読みとったように, 「記念切手」が最も偏っていることが読みとれる.

⁶ 例えば, 1, 2, 3, 4, 5 の五つの値の標準偏差は約 1.58 になるが, 0.1, 0.2, 0.3, 0.4, 0.5 の五つの値の標準偏差は約 0.158 になる.

4.2 ニュースグループの文書分類の実験

ニュースグループの文書分類の実験のコーパスには 20 Newsgroups⁷を利用した. 20 Newsgroups は, 全 20 クラス, 18774 件のニュース記事からコーパスが構成されており, 文書数はどのクラスもおおよそ 1000 件である (cf. 表 3).

4.2.1 一文書あたりの単語数を減らした実験

CNB と NNB の式, 式 (5) と式 (10) を比較してみると, 事前確率 $P(c)$ と $\frac{1}{1-P(c)}$ の部分が異なっていることが分かる. 残りの $\prod_{i=1}^n \frac{1}{P(w_i|c)}$ については等しい. しかし, 単語数が少ない文書を分類する際には, 単語数が多い文書を分類する際よりも, 相対的に事前確率の影響が大きくなることが予想される. そのため, 一文書あたりの単語数を減らして実験を行い, その分類正解率を比較する. この際, 一文書あたりの単語数を n とし, パラメータを x とすると, 一文書あたりの単語数を $n/2^x$ (ただし割り切れない場合には 1 を加算する) として実験した. パラ

表 3 クラスごとの文書数を不均一にした実験のクラスごとの文書数

パラメータ	(オリジナル)	1	2	3	4
1	798	798	798	798	798
2	970	970	485	324	243
3	963	482	241	161	121
4	979	327	164	109	82
5	958	240	120	80	60
6	982	197	99	66	50
7	964	161	81	54	41
8	987	141	71	47	36
9	993	125	63	42	32
10	991	111	56	37	28
11	997	100	50	34	25
12	989	90	45	30	23
13	984	82	41	28	21
14	987	76	38	26	19
15	985	71	36	24	18
16	997	67	34	23	17
17	909	57	29	19	15
18	940	56	28	19	14
19	774	43	22	15	11
20	627	33	17	11	9
合計	18,774	4,227	2,518	1,947	1,663

⁷ <http://people.csail.mit.edu/jrennie/20Newsgroups/>

メータ x は, 0~4 を試した. x が 0 のときには, オリジナルの 20 Newsgroups と等しい.

ニュースグループの分類実験において一文書あたりの単語数を減らした実験の訓練事例中の単語数, 1 文書当たりの平均の単語トークン数を表 4 に示す. なお, このとき訓練事例中の総文書数はすべて 18,774 であり, 訓練事例中の文書数の標準偏差はすべて 95.95 である. また, 訓練事例中の文書数の標準偏差/平均は 0.10 になる.

4.2.2 クラスごとの文書数を不均一にした実験

NNB は事前確率を数学的に正しく考慮しているため, 文書分類ではクラスごとの文書数が不均一である際に効果を発揮すると考えられる. そのため, 20 Newsgroups において, クラスごとに文書を間引きすることによって, クラスごとの文書数を不均一にして実験を行い, 比較する. この際, クラスのインデックスを $i = 1 \dots 20$, パラメータを y とすると, $i = 1$ だけは常にオリジナルの文書数を保ったままとし, $i = 2$ から文書数を $1/((i-1)y)$ に減らして実験を行った. 例えば, $y = 1$ のときには, $i = 2, 3, 4$ の時にはそれぞれ $1/1, 1/2, 1/3$ となり, $y = 2$ のときには, $i = 2, 3, 4$ の時にはそれぞれ $1/2, 1/4, 1/6$ となる. パラメータ y は, 1~4 を試した.

表 3 に, クラスごとの文書数を不均一にした実験のクラスごとの文書数を示す. また, 表 4 にニュースグループの分類実験においてクラスごとの文書数を不均一にした実験の訓練事例中の総文書数, 文書数の標準偏差, 標準偏差/平均, 単語トークン数, 1 文書当たりの平均の単語トークン数を示す. また, ニュースグループの分類実験においてクラスごとの文書数を不均一にした実験の訓練事例中の総文書数, 文書数の標準偏差, 標準偏差/平均, 単語トークン数, 1 文書当たりの平均の単語トークン数を表 5 に示す.

表 4 ニュースグループの分類実験において一文書あたりの単語数を減らした実験の訓練事例中の単語トークン数, 1 文書当たりの平均の単語トークン数

パラメータ	0 (オリジナル)	1	2	3	4
訓練事例中の単語トークン数	2,416,445	1,231,278	629,977	329,077	78,661
1 文書当たりの平均の単語トークン数	128.71	65.58	33.56	17.53	9.52

表 5 ニュースグループの分類実験においてクラスごとの文書数を不均一にした実験の訓練事例中の総文書数, 文書数の標準偏差, 標準偏差/平均, 単語トークン数, 1 文書当たりの平均の単語トークン数

パラメータ	(オリジナル)	1	2	3	4
訓練事例中の総文書数	18,774	4,227	2,518	1,947	1,663
訓練事例中の文書数の標準偏差	95.95	255.79	191.31	179.87	176.67
訓練事例中の文書数の標準偏差/平均	0.10	1.21	1.52	1.85	2.12
訓練事例中の単語トークン数	2,416,445	507,891	387,843	256,630	223,923
1 文書当たりの平均の単語トークン数	128.71	120.15	154.03	131.81	134.65

5 結果

表6と表7に全ての品詞を使用したオークションのカテゴリ分類実験の分類正解率のマイクロ平均とマクロ平均をそれぞれ示し、表8と表9に名詞だけを使用したオークションのカテゴリ分類実験の分類正解率のマイクロ平均とマクロ平均をそれぞれ示す。また、表10と表11にニュースグループの分類実験において、一文書あたりの単語数を減らした実験の分類正解率のマイクロ平均とマクロ平均をそれぞれ示し、表12と表13に同分類実験において、クラスごとの文書数を不均一にした実験の分類正解率のマイクロ平均とマクロ平均をそれぞれ示す。なお、正解率は、(分類に成功したもの) / (実験データ数) として求めた。同じ文書集合の実験で、NB, CNB, NNB のうちで最も良かった正解率を太字で示した。さらに、次に良かった正解率との差がカイ二乗検定で有意だったものに関しては下線を引いた。また、 $P(c)$ なしの CNB と SVM に関しては、上記の三手法のうち最も良かった手法と同じか、それよりも良いものは太字で示し、その優劣にかかわらず、差がカイ二乗検定で有意だったものに関しては下線を引いた。さらに、参考として最頻出カテゴリ (クラス) を答えた場合の正解率も併記した。

表6と表8から、オークションのカテゴリ分類実験において、マイクロ平均を比較した際、NNB が常に NB と CNB を有意に上回っていることが分かる。また同じ二つの表から、NNB が $P(c)$ なしの CNB よりも大抵 (名詞だけの実験の「デスクトップ」が例外である) 上回っていることが分かる。このうち、「記念切手」の実験では、全単語使用した場合、名詞だけを使用し

表6 全ての品詞を使用したオークションのカテゴリ分類実験の分類正解率 (マイクロ平均)

ジャンルカテゴリ	デスクトップ	記念切手	赤ちゃん用の玩具
NB	2.42%	6.58%	4.16%
CNB	18.64%	33.68%	34.82%
NNB	26.48%	39.99%	46.36%
$P(c)$ なしの CNB	26.40%	<u>34.79%</u>	45.86%
SVM	50.70%	60.56%	62.30%
最頻出カテゴリ	8.37%	39.21%	12.76%

表7 全ての品詞を使用したオークションのカテゴリ分類実験の分類正解率 (マクロ平均)

ジャンルカテゴリ	デスクトップ	記念切手	赤ちゃん用の玩具
NB	3.72%	4.86%	3.42%
CNB	6.94%	4.50%	13.48%
NNB	11.51%	11.41%	25.70%
$P(c)$ なしの CNB	11.75%	13.19%	26.38%
SVM	30.38%	26.28%	45.43%
最頻出カテゴリ	1.69%	1.96%	2.22%

表 8 名詞だけを使用したオークションのカテゴリ分類実験の分類正解率 (マイクロ平均)

ジャンルカテゴリ	デスクトップ	記念切手	赤ちゃん用の玩具
NB	2.26%	10.01%	6.52%
CNB	20.28%	37.67%	38.94%
NNB	29.84%	45.96%	52.15%
$P(c)$ なしの CNB	30.21%	40.87%	52.04%
SVM	53.49%	62.28%	62.12%
最頻出カテゴリ	8.37%	39.21%	12.76%

表 9 名詞だけを使用したオークションのカテゴリ分類実験の分類正解率 (マクロ平均)

ジャンルカテゴリ	デスクトップ	記念切手	赤ちゃん用の玩具
NB	1.45%	4.36%	5.37%
CNB	7.85%	6.20%	16.26%
NNB	14.44%	18.12%	34.04%
$P(c)$ なしの CNB	15.21%	20.06%	34.88%
SVM	34.03%	33.47%	47.04%
最頻出カテゴリ	1.69%	1.96%	2.22%

表 10 ニュースグループの分類実験において一文書あたりの単語数を減らした実験の分類正解率 (マイクロ平均)

パラメータ	0 (オリジナル)	1	2	3	4
NB	73.28%	55.78%	42.70%	34.22%	29.94%
CNB	78.20%	67.92%	55.91%	43.93%	34.45%
NNB	78.33%	68.04%	56.14%	44.35%	34.65%
$P(c)$ なしの CNB	78.32%	68.07%	56.14%	44.34%	34.66%
SVM	<u>69.29%</u>	<u>59.88%</u>	<u>54.24%</u>	46.56%	38.84%
最頻出クラス	5.31%	5.31%	5.31%	5.31%	5.31%

表 11 ニュースグループの分類実験において一文書あたりの単語数を減らした実験の分類正解率 (マクロ平均)

パラメータ	0 (オリジナル)	1	2	3	4
NB	72.73%	55.08%	41.89%	33.48%	29.40%
CNB	77.21%	66.93%	54.97%	43.04%	33.73%
NNB	77.41%	67.11%	55.31%	43.60%	34.08%
$P(c)$ なしの CNB	77.40%	67.15%	55.32%	43.60%	34.10%
SVM	68.81%	59.41%	53.83%	46.17%	38.44%
最頻出クラス	5.00%	5.00%	5.00%	5.00%	5.00%

表 12 ニュースグループの分類実験においてクラスごとの文書数を不均一にした実験の分類正解率 (マイクロ平均)

パラメータ	0 (オリジナル)	1	2	3	4
NB	73.28%	35.96%	18.23%	6.01%	1.80%
CNB	78.20%	67.85%	68.47%	70.83%	73.36%
NNB	78.33%	70.52%	70.89%	73.65%	73.72%
$P(c)$ なしの CNB	78.32%	70.57%	71.25%	73.75%	73.72%
SVM	<u>69.29%</u>	<u>62.41%</u>	<u>62.31%</u>	<u>64.00%</u>	<u>67.05%</u>
最頻出クラス	5.31%	22.95%	31.69%	40.99%	47.99%

表 13 ニュースグループの分類実験においてクラスごとの文書数を不均一にした実験の分類正解率 (マクロ平均)

パラメータ	0 (オリジナル)	1	2	3	4
NB	72.73%	42.49%	27.47%	14.53%	5.69%
CNB	77.21%	46.30%	40.04%	33.98%	33.51%
NNB	77.41%	52.97%	45.70%	43.89%	37.86%
$P(c)$ なしの CNB	77.40%	53.23%	46.53%	44.20%	38.15%
SVM	68.81%	42.76%	<u>32.29%</u>	<u>27.29%</u>	<u>23.79%</u>
最頻出クラス	5.00%	5.00%	5.00%	5.00%	5.00%

た場合に拘わらず、カイ二乗検定によりその差が有意であった。しかし、これらの実験において最も良い結果なのは SVM であり、NNB を有意に上回っている。

また、表 7 と表 9 から、オークションのカテゴリ分類実験において、マクロ平均においても、有意ではないながら、NNB が常に NB と CNB を上回っていることが分かる。また同じ二つの表から、 $P(c)$ なしの CNB が有意ではないものの、NNB を上回っていることが分かる。さらに、SVM は「デスクトップ」と「記念切手」においては有意に、「赤ちゃん用の玩具」では有意ではないものの、NNB を上回った。

また、表 10 と表 12 から、ニュースグループの分類実験においても、マイクロ平均で比較した場合、常に NNB が NB と CNB を上回ることが分かる。ただし、その差が有意なのは、クラスごとの文書数を不均一にした実験 (表 12) のパラメータが 1 のときと 3 のときだけである。また、これらの実験において、 $P(c)$ なしの CNB はしばしば NNB を上回っているが、有意に上回っていることは一度もなかった。さらに表 10 の一文書あたりの単語数を減らした実験では、パラメータ 0, 1, 2 のとき、NNB の分類正解率が SVM を有意に上回っている。しかし、パラメータ 3, 4 のときは SVM が最高であり、NNB と比較してその差は有意であった。一方、表 12 から、クラスごとの文書数を不均一にした実験では、全ての実験設定のときに NNB が SVM を上回っている。この差はカイ二乗検定により有意であった。

これに対し, 表 11 と表 13 から, ニュースグループの分類実験において, マクロ平均においても, NNB が常に NB と CNB を上回っていることが分かる. また同じ二つの表から, $P(c)$ なしの CNB が NNB を上回っていることが分かるが, これらの差はいずれも有意ではない. さらに, 表 11 の一文書あたりの単語数を減らした実験では, 常に有意でないながらも NNB の分類正解率が SVM を上回っている. しかし, パラメータ 3, 4 のときは, マイクロ平均と同様に SVM が最高であり, NNB と比較してその差は有意であった. 一方, 表 13 から, クラスごとの文書数を不均一にした実験では, 全ての実験設定のときに NNB が SVM を上回っている. この差はパラメータ 0 と 1 のとき, カイ二乗検定により有意であった.

6 考察

オークションのカテゴリ分類実験の結果とニュースグループの分類実験の結果を総合して NNB の特色について考察する. まず, 表 6~表 13 から, 全ての実験を通して, NNB は NB と CNB を上回っていること, また $P(c)$ なしの CNB に有意に勝っていることはあっても有意に負けていることはないことが読みとれる⁸. このことから, NNB は他の Bayes の定理を利用した文書分類手法に比較しても引けを取らない文書分類手法であると言える. 特に NB と比較したときには, マイクロ平均では常に有意に, マクロ平均ではオークションのカテゴリ実験において「デスクトップ」と「記念切手」の全品詞を使った分類実験以外の実験で有意に勝っていた. なお, マクロ平均ではサンプル数の減少から, NNB と比較した際, SVM と NB だけにしか有意差が認められなかったため, 今後は主にマイクロ平均について考察する.

オークションのカテゴリ分類実験とニュースグループの分類実験の実験設定の違いのうち, NNB の式に大きく関わりそうな点は二点あるだろう. 一つ目は第一項の事前確率に関わる, クラスごとの文書数 (またはカテゴリごとの商品数) のばらつき, 二つ目はそれ以外の部分に関わる, 一文書ごとの単語トークン数である. 4.2 節でも述べたように, CNB と NNB の式, 式 (5) と式 (10) を比較してみると, 事前確率 $P(c)$ と $\frac{1}{1-P(c)}$ の部分が異なっており, 残りの $\prod_{i=1}^n \frac{1}{P(w_i|c)}$ については等しい. NNB は事前確率を数学的に正しく考慮しているため, 文書分類ではクラスごとの文書数が不均一である際に効果を発揮すると考えられる. また, 単語数が少ない文書を分類する際には, 単語数が多い文書を分類する際よりも, 相対的に事前確率の影響が大きくなることが予想される.

クラスごとの文書数のばらつきを見るために, 図 5 に同実験においてクラスごとの文書数を

⁸ ただし, 有意ではないものの, 表 7 と表 9 では常に $P(c)$ なしの CNB が NNB を上回っている. $P(c)$ は事前確率であるため, カテゴリ間のデータ数に偏りがあり, なおかつ単語トークン数が少ない場合には, NNB は大きなカテゴリに分類されやすいことが予想できる. そのため, マイクロ平均は $P(c)$ なしの CNB を有意に上回ったが, マクロ平均は $P(c)$ なしの CNB が高くなる傾向がある可能性がある.

不均一にした実験の、クラスごとの文書数の標準偏差／平均を横軸とした分類正解率の散布図を示す。

また、単語トークン数の影響を見るために、図6にニュースグループの分類実験において一文書あたりの単語数を減らした実験の、一文書あたりの単語トークン数を横軸とした分類正解率の散布図を示す。その上でNNBの特徴をNB、CNB ($P(c)$ のないものを含む)、SVM の順で比較しつつ考察する。

図5から、NBはクラスごとの文書数のばらつきが多い際にその分類正解率が著しく低下することが分かる。また、商品のカテゴリ分類実験において、NBの分類正解率がとても低いのも同じ原因によるものであることがうかがえる。商品のカテゴリ分類実験において、標準偏

図5 ニュースグループの分類実験においてクラスごとの文書数を不均一にした実験の、クラスごとの文書数の標準偏差／平均を横軸とした分類正解率の散布図

図6 ニュースグループの分類実験において一文書あたりの単語数を減らした実験の、一文書あたりの単語トークン数を横軸とした分類正解率の散布図

差／平均は「デスクトップ」, 「記念切手」, 「赤ちゃん用の玩具」がそれぞれ 1.44, 2.84, 1.33 と高く, カテゴリごとの商品数のばらつきが大きいからである. この原因として, クラスごとの $P(w_i|c)$ のデータスパースネスの差が考えられる. ここで NB の式を再掲する.

$$\hat{c} = \operatorname{argmax}_c P(c) \prod_{i=1}^n P(w_i|c) \quad (11)$$

NB ではクラスごとの文書数のばらつきが大きい際には, クラスによって訓練事例の単語トークン数に大きな差ができる. 例えば最も大きなクラスでは, その訓練事例となるトークン数が一万となり, 小さなクラスでは 10 トークンといった具合である. その結果, 小さなクラスではデータがスパースになり, より頻繁にスムージングが行われる. そのため, NB では, 個々の分類問題と相性の良いスムージング手法を用いることが求められるが, 今回のジェフリー・パークス法はラプラス法よりは良いとはいえ, まだ実際よりも大きな値を小さなクラスに与えていたと思われる⁹. このように, NB ではデータスパースネスによって誤分類が起きて分類正解率が著しく下がることもあるが, この問題を補集合を用いることで解決したのが CNB であり, この点について NNB は CNB と全く同じ特色を持っている.

次に, NNB と CNB の差について考察する. 図 5 を見てみると, $P(c)$ を考慮しない CNB と NNB の差はあまりないが, NNB は CNB より若干良いことが分かる. これは, クラスごとの文書数に偏りが出てくると, CNB と NNB の違いである事前確率が異なってくるため, その分類正解率に差がつくからであると考えられる.

次に図 6 を見てみると, 一文書当たりの平均の単語数が減ると, どの手法を用いても全体的に分類正解率が低下することが分かる. これは, 単語数が減ることで, 統計の材料となる w_i が減っているためであると考えられる. これに対し, 一文書当たりの単語数を変化させても, CNB と NNB の差が大きくなることはなかった. これは, ニュースグループのオリジナルのコーパスでは, そもそもクラスごとの文書数に偏りがあまり見られないため, 事前確率に偏りがなく, CNB と NNB がそう変わらない結果になったためであると考えられる.

ここで, オークションのカテゴリ分類実験の実験設定も共に比較してみる. 図 7 に縦軸を標準偏差／平均, 横軸をクラスごとの単語数にした, 実験設定ごとの CNB と NNB の差の散布図を示す. 左の図が事前確率を考慮した CNB と NNB の差が有意であるかを表しており, 右の図が事前確率を考慮しない CNB と NNB の差が有意であるかを表している. なお, 両方の図において, 縦軸の値が 0.10 である五つの点が, ニュースグループの一文書当たりの単語数を減らした一連の実験であり, 横軸の値が 128.71 である五つの点が, ニュースグループのクラスごとの文書数を不均一にした一連の実験である. オークションのカテゴリ分類実験の実験設定では,

⁹ 佐藤 他 (2010) のスムージング (頻度 0 の時には $0.1/N$ を用いる. この N は訓練事例中の全単語トークン数.) を用いると NB だけ飛躍的に正解率が上昇した. ただし, 本論文の主張との矛盾はない結果であった.

図 7 横軸を標準偏差／平均、縦軸をクラスごとの単語数にした、実験設定ごとの CNB と NNB の差の散布図

ニュースグループの実験よりも、ひとつの文書（商品）ごとのトークン数が少なく（一文書当たりの単語数を最も減らした実験と同じくらいである）、なおかつクラス（カテゴリ）ごとの文書数（商品数）の偏りが、クラスごとの文書数を不均一にしたニュースグループの実験と同じくらい、またはそれ以上に偏っていることが分かる。

図 7 により、標準偏差／平均が小さい時には CNB と NNB の差はほとんどないこと、また、標準偏差／平均が大きくなるにつれて NNB が有意に CNB を上回るようになる傾向がうかがえる。また、一文書あたりの単語数が少ない時に、なおかつ標準偏差／平均が大きければ、NNB が有意に事前確率を考慮しない CNB に対しても有意に上回ることが分かる。オークションのカテゴリ分類実験の「記念切手」（左上の二点）のときには、両方の条件が共に十分当てはまったため、NNB の分類正解率が、CNB にも事前確率を考慮しない CNB にも有意に上回ったものと思われる。

次に、SVM との比較を行う。残念ながら、NNB の分類正解率が、CNB にも事前確率を考慮しない CNB にも有意に上回った「記念切手」の実験設定を含む、オークションのカテゴリ分類実験では、SVM の分類正解率がいつも他手法を有意に上回った。しかし、ニュースグループの文書分類実験では、一文書あたりの文書数を減らした場合のパラメータが 3 と 4 の実験を除き、NNB が SVM を有意に上回った。上述したように、 w_i の数が減ってしまうと、Bayesian アプローチの正解率は下がることが一文書当たりの単語数が減った際に SVM を有利にしたと考えられる。しかし、ニュースグループのオリジナルのコーパスの実験では、CNB との差は有意ではないながらも、NNB が全ての分類器の中で最も高い分類正解率となっている。このことから、NNB は時には SVM を有意に上回り、他手法と比較しても最も良い分類正解率を示しうる

表 14 ニュースグループの分類実験にかかった時間

手法	NB	CNB	NNB	$P(c)$ なしの CNB	SVM
時間	3 分 19 秒	28 分 1 秒	27 分 53 秒	27 分 59 秒	11 分 59 秒

手法であることが分かる¹⁰.

また、最後に、提案手法の速度についての目安を知るために、最も実行時間がかかるオリジナルのニュースグループの分類実験において、それぞれの手法の実行時間を測った。表 14 にその結果を示す。SVM は C 言語により実装されたツールによるものであり、それ以外の Bayesian アプローチは Perl により個人的に実装したものであるため、一概に比較は難しいが、NNB が CNB とほぼ同じ速度で実行されることが分かる。

7 まとめ

本稿では、文書分類のための新手法として、Negation Naive Bayes (NNB) を提案し、その性質を他の Bayesian アプローチである Naive Bayes (NB), Complement Naive Bayes (CNB) と比較した。NNB は、CNB と同様にクラスの補集合を用いるが、NB と同じ事後確率最大化の式から導出されるため、事前確率を数学的に正しく考慮している点で異なっている。NNB の性質を見るために、ふたつのコーパスを用いた実験を行った。一つ目はオークションの商品分類の実験であり、二つ目はニュースグループの文書分類の実験である。このうち、ニュースグループの文書分類の実験では、一文書あたりの単語数を減らした実験と、クラスごとの文書数を不均一にした実験を行い、オリジナルの文書集合の実験に加え、それぞれパラメータを変えて四通りを試した。その結果、すべての実験において NNB と CNB が NB の分類性能を上回ること、また、一文書あたりの単語数が減り、クラスごとの文書数が偏るときにマイクロ平均で NNB は CNB を上回ることが分かった。事前確率を無視した CNB と比較しても、これらの条件が共に当てはまる際には、NNB が CNB を有意に上回った。これは、CNB と NNB の違いは事前確率にあるため、クラスごとの文書数が偏るときにその影響が見られ、なおかつ一文書あたりの単語数が減る際には相対的に事前確率の影響が大きくなるためであると考えられる。また、CNB または事前確率を無視した CNB が NNB を有意に上回ることはなかった。さらに、ニュースグ

¹⁰ Gabrilovich and Markovitch (2004) では、ニュースグループの分類実験において素性選択をする前でも 76.9%、素性選択後は 85.3%という正解率を報告している。また、Siolas and d'Alché-Buc (2000) では、同実験において素性選択を行うと 86.44%となっている。このように、SVM は素性選択などによってもっと性能をあげることが可能であるため、それらの手法を用いれば SVM の方が性能が良くなる可能性が高い。また、五分割交差検定の分割法を変えて実験してみると、SVM の正解率が 75.79%となったことから、SVM の性能は、選択されるテストセットの選び方によって変化することが分かった。NNB などのベイズの手法の正解率も、五分割交差検定の分割法により変化することが予想される。

ループの分類実験においては、その際の CNB との差は微小ながら、参考として比較したサポートベクターマシン (SVM) をカイ二乗検定で有意に上回り、比較手法中で最も良い分類正解率を示す結果も見られた。

これらのことから、特に一文書当たりの単語数が減り、クラスごとの文書数が偏る場合において、NNB が他の Bayesian アプローチより勝る手法であること、また、時には SVM を有意に上回り、比較手法中で最も良い分類正解率を示す手法であることが分かった。

謝 辞

本研究の一部は、文部科学省科学研究費補助金 [若手 B (No: 24700138)] の助成により行われた。ここに、謹んで御礼申し上げる。また、本論文の内容の一部は、Recent Advances in Natural Language Processing 2011 で発表したものである (Komiya, Sato, Fujimoto, and Kotani 2011)。

参考文献

- Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.
- Church, K. W. (2000). “Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to $p/2$ than p^2 .” In *Proceedings of the COLING '00*, pp. 182–191.
- Gabrilovich, E. and Markovitch, S. (2004). “Text Categorization with Many Redundant Features: Using Aggressive Feature Selection to Make SVMs Competitive with C4.5.” In *Proceedings of the The Twenty-First International Conference on Machine Learning*, pp. 321–328.
- 花井拓也, 山村毅 (2005). 単語間の依存性を考慮したナイーブベイズ法によるテキスト分類 (類似性の発見). 情報処理学会研究報告 自然言語処理研究会報告, 2005(22), pp. 101–106.
- 井筒清史, 横澤誠, 篠原健 (2005). Web 文書タイプ自動分類手法の比較評価と適用. In *IPSI SIG Notes 2005(32)*, pp. 25–32.
- Komiya, K., Sato, N., Fujimoto, K., and Kotani, Y. (2011). “Negation Naive Bayes for Categorization of Product Pages on the Web.” In *Proceedings of the RANLP 2011*, pp. 586–591.
- Lewis, D. D. (1992). “An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task.” In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 37–50.
- marquis de Laplace, P. S. (1814). *Essei philosophique sur les probabilites*. Paris: Mme. Ve. Courcier.
- marquis de Laplace, P. S. (1995). *Philosophical Essay On Probabilities*. New York: Springer-

Verlag.

- McCallum, A. and Nigam, K. (1998). “A Comparison of Event Models for Naive Bayes Text Classification.” In *Proceedings of the AAAI/ICML-98 Workshop on Learning for Text Categorization*, pp. 41–48.
- 持橋大地 (2006). 自然言語処理におけるベイズ統計. 電子情報通信学会技術研究報告 NC, ニューロコンピューティング, pp. 25–30.
- 持橋大地, 菊井玄一郎 (2006). 無限混合ディリクレ文書モデル. 自然言語処理研究会報告 2006-NL-172, pp. 47–53.
- Rennie, J. D. M., Shih, L., Teevan, J., and Karger, D. R. (2003). “Tackling the Poor Assumptions of Naive Bayes Text Classifiers.” In *Proceedings of the ICML2003*, pp. 616–623.
- 佐藤直人, 藤本浩司, 小谷善行 (2010). ウェブ上の商品情報を利用した商品のカテゴリ分類. 人工知能学会 第 87 回知識ベースシステム研究会, pp. 7–10.
- Siolas, G. and d’Alché-Buc, F. (2000). “Support Vector Machines based on a Semantic Kernel for Text Categorization.” In *Proceedings of the international Joint conference on Neural Networks*, pp. 205–209.
- 高村大也, Roth, D. (2007). Predictive Naive Bayes Classifier の提案と言語処理への適用. 言語処理学会第 13 回年次大会発表論文集, pp. 546–549.
- Zhang, H. (2004). “The Optimality of Naive Bayes.” In *Proceedings of the the 17th International FLAIRS conference (FLAIRS2004)*, pp. 562–567.

略歴

古宮嘉那子：2005 年東京農工大学工学部情報コミュニケーション工学科卒. 2009 年同大大学院博士後期課程電子情報工学専攻修了. 博士（工学）. 同年東京工業大学精密工学研究所研究員, 2010 年東京農工大学工学研究院特任助教, 現在に至る. 自然言語処理の研究に従事. 情報処理学会, 人工知能学会, 言語処理学会各会員.

伊藤 裕佑：2012 年東京農工大学工学部情報工学科卒. 在学中に自然言語処理の研究に従事.

佐藤 直人：2009 年東京農工大学工学部情報工学科卒. 2011 年同大大学院博士前期課程情報工学科修了. 在学中にゲーム研究と自然言語処理の研究に従事.

小谷 善行：東京農工大学大学院教授. 工学研究院先端情報科学部門・情報工学科. 人工知能, 知識処理, ゲームシステム, ソフトウェア工学, 教育工学の研究に従事. 情報処理学会, 電子情報通信学会, 人工知能学会等会員. コンピュータ将棋協会副会長. 最近では, 多量データからの知識獲得に興味を

持っている.

(2012 年 10 月 12 日 受付)

(2012 年 12 月 26 日 再受付)

(2013 年 3 月 6 日 採録)