情報学基礎 64−13
自然言語処理 145−13
(2001. 9. 11)

著者判別に有効な特徴量の推定

吉田 篤弘　延澤 志保　平石 智宣　斎藤 博昭

慶應義塾大学大学院 理工学研究科 開放環境科学専攻
Email: {atsuhiro,shiho,tomonobu,hxs}@nak.ics.keio.ac.jp

概要：
従来著者判別において単独で用いられてきた特徴量を複数用い、より汎用的かつ有効と考えられるそれら特徴量の比率を推定する。実験には近現代日本小説家14名の作品122編を著者ごとに複数組み合わせたテキストを使用し、これらのテキスト間の非類似度を算出する。それを基にして判別の正解率を求め、これを評価値とした重みの増減から安定して高い正解率を与える各特徴量の重みを算出した。その結果3gram、形態素3gram、読点前の文字の種類、行別のひらがな出現数の重みが特に高い値となり、これを基に得られた正解率は最大で60%程度だった各特徴量単独での正解率より平均で10%近く上回った。また得られた重みの有効性を確認するために行った、新規の著者を含んだ新しいテキスト群を使用しての正解率測定実験においても特徴量単独での正解率を上回る結果を出す事ができた。

キーワード: 著者判別、特徴量、文書スタイル

Effective features of authorship identification

Atsuhiro Yoshida　Shiho Nobesawa　Tomonobu Hiraishi　Hiroaki Saito

Department of Computer Science
Keio University
Email: {atsuhiro,shiho,tomonobu,hxs}@nak.ics.keio.ac.jp

Abstract :
We propose a method of authorship identification using weighted features. First we select 122 works in total by 14 Japanese modern authors and calculate the dissimilarity of these texts using ten features to determine the ideal weight value among features, where a hill-climbing method is adopted for convergence. As a result, four features, namely 3gram, morpheme 3gram, character set before comma, and *hiragana* usage show high effectiveness. The identification rate using the weighted features is higher than that of a single feature which is about 60% or less. In addition, we test a new text group which contains new authors and confirm the effectiveness of the obtained weight.

Key word: authorship identification, features, stylometry

1 はじめに

文献の真贋問題は有史以来人類が向き合って来た問題の一つである。著者判別の研究はそういった問題を解決するための一手段として19世紀半ばより行われて来た。近年ではデータ処理機器の発達に後押しされた計量文献学的手法が試みられるようになり、より大量のデータを扱う事が可能になった。しかし、それら大半の研究は一文辺りの語数や一語辺りの文字数といった単一の特徴量を用いて行われたものであり、どの特徴量を用いれば最も良い結果を得られるのか、という比較的な観点から行われた研究はまだ少ない。いかに処理機器が発達しようとも判別に寄与しない特徴量を考慮するのは無益であり、逆に著者の癖を如実に与える特徴量を発見・使用することができれば効率の面から見て非常に有益である。そこで本研究では、有益であると考えられる特徴量を発見し、それを元にした著者判別を行うことによって判別上有効な特徴量を求める。

2 特徴量

特徴量は文章上における著者の癖を示す要素であり、言語に応じて様々なものが用いられてきた。以下では日本語における特徴量と本稿で使用する特徴量について述べる。

2.1 日本語における特徴量

日本語は漢字・カタカナ・ひらがなの三形態を取る表記の複雑な言語であり、しかも欧米の言語と異なり単語の分かち書きが為されていない。そのため語彙の豊富さや特定の付属語の使用状況など、単語に関する情報を容易に用いる事ができた欧米に対して統計的な観点からの研究は遅れてきた[1][2]。しかし安本は1957年、源氏物語について質的内容によらない要素に注目しての統計的な分析を行い、この方面の研究の先鞭をつけた。これは和歌・直喩の使用度や1ページ当たりの心理描写の割合といった情報を使用したものであり、これ以後読点の使用状況や助詞・助動詞の使用状況などに着目した日本語に特化した研究が行われるようになった。最近では村上・今西らが同じ源氏物語について、作品中に出現する全26種類の助動詞に注目した研究を行っている[3]。彼らは助動詞の出現率を数量化III類で分析することにより、全54巻の類似性について調べた。さらに松浦らは文字を単位とした手法、すなわちn個の文字が隣接して生じる確率－Ngramの出現確率に注目した研究を行った[4][5]。文字の並びという無意識的に生じる情報を用いたこの手法は事前操作の必要がほとんどないため、効率その他の面で有効な特徴量として捉えられる。

2.2 本稿で使用する特徴量

一般に日本語における研究は形態素解析・意味解析など、文章に対する事前操作を施した上で得られる情報を用いて行われる傾向にある。しかし意味解析のような高度な操作を誤り無く行うツールはまだ存在せず、また人手による情報の添付は極めて労力を要し効率的とは言い難い。すなわち、事前操作をあまり必要とせず、容易に情報を取得でき、かつ判別に有効な特徴量を選ぶ事が望ましいと言える。本稿ではその考えに基づき、以下の特徴量10個を選択し使用する事とした。

(a) 3gramの出現数(以下、3gram)
ある3文字が連続して出現する度合。連続する数は複数選べるが、本稿では松浦らの研究に基づき、精度が高くなると考えられる3gramを選択した。なお、取得に当ってはNgram測定用ツールCMU-Cambridge Statistical Language Modeling Toolkit v2[6]を利用した。

(b) 形態素3gramの出現数(以下、形態素3gram)
ある形態素が3つ連続して出現する度合。形態素解析ツール茶筅[7]、及び上記のNgram測定用ツールを用いて取得した。

(c) 読点前の文字(以下、読点前文字)
読点前に出現する文字の種類の度合。

(d) 各行ごとのひらがなの出現数の度合(以下、行別ひらがな出現数)
「あ」から「お」の出現数・「か」から「こ」の出現数といったように、文中におけるひらがなの出現の度合を行別に分けたもの。

(e) 各品詞の出現数 (以下、品詞出現数)
文中における形容詞、動詞などの品詞の出現の度合。茶筅を利用し取得した。

(f) 文頭における各品詞の出現数 (以下、文頭品詞出現数)
文の先頭に範囲を限定した、各品詞の出現の度合。茶筅を利用し取得した。

(g) 文末における各品詞の出現数 (以下、文末品詞出現数)
句点直前に範囲を限定した、各品詞の出現の度合。茶筅を利用し取得した。

(h) 1 文辺りの文字数 (以下、1 文の文字数)
一文中に出現する文字数の度合。

(i) 1 文辺りの漢字数 (以下、1 文の漢字数)
一文中に出現する漢字数の度合。

(j) 1 文辺りの読点数 (以下、1 文の読点数)
一文中に出現する読点数の度合。

3 評価式

特徴量を用いた判別を行う際は、これを数値化して計算する必要がある。本稿では以下の手順により数値化する。

3.1 特徴量の数値化

ある 1 つの特徴量を x で表すと、ある基準のテキスト P とその比較対象のテキスト Q における特徴量分布は各々確率分布関数 $P(x)$ 及び $Q(x)$ で表される。この時テキスト $P \cdot Q$ 双方に出現する共通項目の集合 C は

$$C = \{x | P(x)Q(x) \neq 0\} \tag{1}$$

で表せる。この時 $card(C)$ を集合 C の要素数とすると、テキスト $P \cdot Q$ 間の非類似度は松浦らの提案した以下の関数 $D(P,Q)$ で示される。

$$D = \frac{1}{card(C)} \sum_{x \in C} \left| \log \frac{P(x)}{Q(x)} \right| \tag{2}$$

この関数は $P(x)$ と $Q(x)$ に関して対称であり、値の大小はテキスト間の類似の度合に反比例する。ここで、特徴量を (a) から (j) まで変化させる事により、基準テキストと比較テキストとの間に合計 10 個の非類似度を求める事が出来る。これらを以下の式により結合する。

$$T = \sum_{i=1}^{p} \omega_i D_i \quad (\text{但し、} \sum_{i=1}^{p} \omega_i = 1) \tag{3}$$

D_i は特徴量 i における非類似度を、ω_i は Di の重みを表す。また p は現在の特徴量の総数を示し、初期値は 10 とする。以後、ある $i(1 \leq i \leq p)$ について重みの変更を行う。

$$\omega_i = \omega_i \pm a \quad (a \text{ は重みの移動量}) \tag{4}$$

$$\omega_{!i} = \omega_{!i} \mp a/(p-1) \tag{5}$$

$\omega_{!i}$ は ω_i 以外の重みを表す。この変更と式 (3) より得られた計 20 個の T の値を比較し、後述する山登り法に基づいて重みを更新する。なお、変更の結果重みが 0 になったものについては不要な特徴量であったと判断してこれを除外し、1 つ減らした状態で再試行する。

3.2 山登り法 (hill climbing)

山登り法とは最適化問題の解法の一つである [8]。ある状態 C における評価値 R をまず算出し、次いで別の状態 C' における評価値 R' を算出する。もし $R' < R$ であれば C' を次の状態とし、$R' \leq R$ ならば元の C をそのまま次の状態とする。この操作を評価値の改善される状態変化がなくなるまで繰り返し行う。

しかしこの方法は局所解に陥りがちで最適解を得にくいという欠点がある。また単に重みの移動量を大きくしただけでは精度が粗くなり、付近の頂上解に達しない可能性が出て来る。そこで可能な限りそれらの欠点を回避させるため、本稿では以下の方法を取った。

- 特徴量の重みの初期変更値を 0.01～0.2 まで、0.01 刻みで増加させる。
- 更新が停止したら現在の重み変更値の 9/10 を新規の変更値として設定し、以後 10 回までこの操作を繰り返す。

変更が行われなくなった状態の値を最終的な非類似度の値として確定する。

4 著者判別実験

以上の関数及び解法を用い、大量の電子化されたテキストを対象とした著者判別実験を行う。

4.1 使用するテキスト

実験においては本稿の最後に添付した表6に示す、近現代日本小説家13人の作品125編を用いた。これらは青空文庫[9]上で公開されており、かつ十分な量のテキストが確保される事を基準として選択した。これらの文章には作品名や章名、ルビなど文章に直接関係がないと判断できる情報は含めない。また半角英文字は一単語で一文字、全角英文字は一字で一文字と換算するものとした。

このような処理を行った上で、基準テキストと比較テキスト間で字数が異なる事での非類似度に対する影響を回避するためテキストの分量を3万字に制限する。なお、3万字より少ない作品に関しては複数の同一著者の作品を繋ぎ合わせ、余りの部分を削除する事で対処する。逆に、3万字より多い作品に関しては先頭の3万字分のみを取り出し、残りに関しては新たに繋げる事なく削除する。これはある1つの作品は文体が意図的に統一されている可能性があり、それを複数のテキストに分ける事で非類似度が低下するのを防ぐためである。この作成の様子を図にしたのが図1である。

用いるテキストには旧字体(ゐ・ゑ・いふ・なつた等)が用いられている作品と新字体(現在通常使われている字体)の作品とが混在している。しかし、字体の差によって影響があると考えられる3gramのみを使用した事前実験を旧字体・新字体同数ずつの計10テキストに対して行った所、完全正解率100%を得る事ができた。この結果から字体の差は影響しないと判断し、この点に関して考慮はしないものとする。これにより生成されるテキストは芥川4・伊藤2・梶井2・菊池3・独歩2・藤村3・太宰3・中島1・漱石4・新美2・一葉2・水野2・賢治2・鴎外4の合計36である。このまとまりを1セットとし、テキストの繋げ方を変えて合計5セット

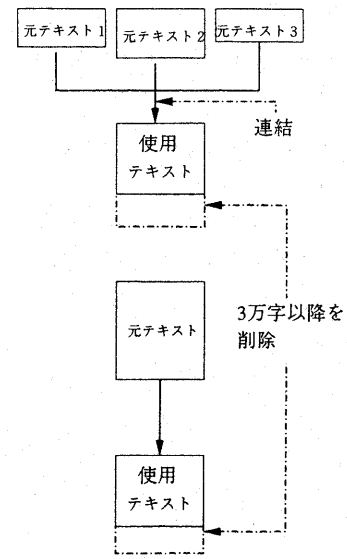

図1: 3万字のテキスト生成の例

作成、最も判別率の高い特徴量の組合せを求める。なお、1つしかない中島のテキストは著者数を増やすためのダミーであるとする。図2にセット生成の例を示す。

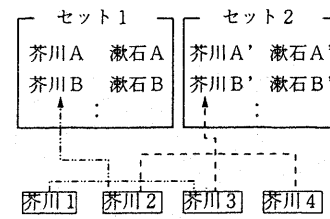

図2: セットの生成の例

また、セット1〜5とはまた繋ぎ方を変えたセット6・セット7・セット8、及びセット1〜8の著者の中からランダムに選択した8名と新規の著者12名、計20名の作品からセットαを作成し得られた組合せの有効性を確かめる。なおセットαは著者1人につき2テキストずつ、40テキストから構成されるものとする。

4.2 評価基準

重み変更後の式 (3) の値を昇順に並べる事により、基準テキストに類似したテキストを順に得る事ができる。この先頭のテキストの著者が基準テキストの著者と同一であれば判別は成功したと言える。また、用意したテキスト 36 個を順に基準テキストとする事でそのセットにおける判別の正解率を調べられる。この事を用い、以下の 3 つの評価基準を定める。

(1) 完全正解率
基準テキストと同一の著者の作品が、先頭から昇順に完全に並んだ場合の割合。

(2) 部分正解率
基準テキストと同一の著者の作品が、先頭に出現した場合の割合。なお、完全正解率は部分正解率に含まれる。またセット α は 1 著者につき 2 テキストずつ、すなわち基準テキスト以外のテキストはどの著者においても 1 つしかないため、部分正解率については求めない事とする。

(3) 先頭からの距離
基準テキストと同一の著者の作品の、先頭からの距離を合計したもの (図 3 参照)。

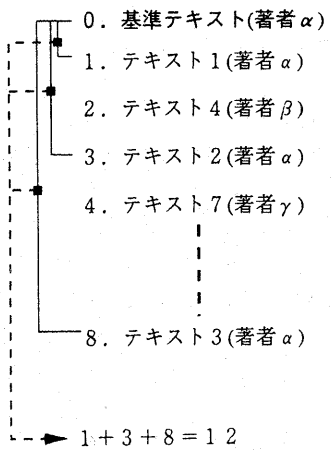

図 3: 先頭からの距離

なお、(1) − (2) − (3) の順に優先するよう設定した。

4.3 実験結果と評価

それぞれ 36 のテキストから構成されるセット 1 からセット 5 までを用いて得られた結果を以下に示す。表 1 はセット 1〜5 における、各特徴量を単独で用いての正解率の平均を示す。

特徴量を組合せ、付与した重みを変更することによって高い正解率を与えた重みの状態を各セットごとに得る事ができた。これらの重みのうち、各セットにおいて完全正解率上位 2 値を与えた重みを集計しその平均値を求めた。その数値及び各試行において重みが 0 とならずに残る確率、すなわち判別に置いて必要と判断される確率を残留率として表したものが表 2 である。

表 1: 各セットの単独正解率の平均値

特徴量	完全正解率	部分正解率
(a). 3gram	58.28	84.00
(b). 形態素 3gram	61.71	87.43
(c). 読点前文字	33.71	61.14
(d). 行別ひらがな出現数	21.71	42.86
(e). 品詞出現数	29.72	54.86
(f). 文頭品詞出現数	9.71	30.28
(g). 文末品詞出現数	12.00	22.86
(h). 1 文の文字数	9.14	20.00
(i). 1 文の漢字数	17.14	30.80
(j). 1 文の読点数	7.43	16.00

まず表 1 の結果より、これらの中では高い値を見せた 3gram や形態素 3gram でも正解率は 60% 程度に留まっている事が分かる。

次に表 2 の結果より、1 文辺りの情報は残留率がいずれも高々20% であり、重みも 1% に達しない事から大人数の中から特定の著者を見極める場合に用いるのは難しい事が判明した。一方読点前文字のように、単独での正解率は低めながら複合させた場合は残留率が 100% に達するなど、複合させて初めてその利用価値が判明する特徴量が存在する事も分かる。また試行の結果 3gram、形態素 3gram、読点前文字、行別ひらがな出現数の 4 特徴量に割り振られた重みの合計の割合は 95% 近くに達し、この 4 つが判別において重要な役割を果た

表 2: 完全正解率上位 2 値の特徴量の構成

特徴量	残留率	重み平均値
(a). 3gram	64.10	0.1801
(b). 形態素 3gram	92.31	0.4398
(c). 読点前文字	100.00	0.1468
(d). 行別ひらがな出現数	69.23	0.1602
(e). 品詞出現数	38.46	0.0361
(f). 文頭品詞出現数	35.90	0.0293
(g). 文末品詞出現数	15.38	0.0025
(h). 1 文の文字数	17.95	0.0062
(i). 1 文の漢字数	20.51	0.0022
(j). 1 文の読点数	7.92	0.0011

すと言う事ができる。そして、表 2 の平均重みをセット 1 からセット 5 に適用して得られた結果が表 3 である。

表 3: 重みの平均値を用いての各セットの正解率

セット No.	1	2	3	4	5
完全正解率	68.57	71.43	77.14	74.29	62.86
部分正解率	94.29	94.28	97.14	100.00	88.57
距離合計	320	233	250	234	232

完全正解率の平均値は 70% と、特徴量を単独で使用した結果を 10% 上回る結果を得る事ができた。そしてセット 1〜セット 5 で得られた重みの有効性確認のため、セット 6〜8 及び新規の著者 12 名を含めたテキスト 40 編からなるセット α に対して得られた平均重みを適用して正解率を求めた。また対比のため、特徴量単独での新規セットにおける正解率を求めた。その結果が表 4 及び表 5 である。なお表 4 で用いた特徴量は全て 3gram である。

このように学習したデータ以外でもそれまでと同様の高い正解率を得る事ができた。新規の著者の作品を含めたセット α においても特徴量単独での正解率を大きく上回る値を出せた事から、この組合せを使用して安定した著者判別を行えると言える。

表 4: 特徴量単独での使用による新規セットの正解率

セット No.	6	7	8	α
完全正解率	65.71	66.89	68.57	52.50
部分正解率	80.00	88.57	97.14	—
距離合計	295	275	218	331

表 5: 重みの平均値を用いての新規セットの正解率

セット No.	6	7	8	α
完全正解率	65.71	68.57	71.43	60.00
部分正解率	88.57	91.43	100.00	—
距離合計	256	211	209	294

5 終わりに

今回の実験においては近現代日本小説家の文章を題材として用い、著者判別において有効な働きを示す特徴量を、複数を重みづけによって組み合わせて使用することにより求めた。これにより、判別にほとんど貢献しない特徴量の識別、及び複数の著者の中から特定の著者の作品を 60%以上の確率で判別可能な重みを算出することができた。先行研究で用いられた特徴量 (3gram) 単独では最高で 70%弱、最低で 50%弱程度であり、しかも今回得られた重みによる正解率を上回らなかった事と比較してその有効性を証明できたと言える。しかし、今回は特徴量 10 個を全て用いて実験を行ったが、明らかに重要度の低い 1 文辺りの特徴量は無視して考えた方がより良い結果を生むという可能性も否定できない。そのため、用いる特徴量の組合せについては更なる改善の余地があると考えられる。また著者判別は作品の贋作・不明著者特定問題といった文学面のみならず、法廷に供するための脅迫状の執筆者特定など様々な分野に応用する事が可能である。このことから、専門家による作品という限定した範囲を脱し、一般人の日記や論文などありふれた文章を対象としてもなお今回の特徴量が有効であるか、より少ない文字量においても有効な判別が行われるか、といった点も今後解決すべき問題として熟考しなくてはならない。そ

して、今回の実験においては品詞情報を用いた特徴量は高い正解率を得ることができなかった。その理由として「品詞すべてを対象として比較する」という設定に問題があった可能性を無視できない。これより、特徴量全般の設定に関しても今後さらに熟考して行く必要があると言える。

参考文献

[1] 村上 征勝：“真贋の科学–計量文献学入門–”, 朝倉書店, 1994

[2] Tony McEnery, Michael Oakes: “Authorship Identification and Computational Stylometry”, in Handbook of Natural Language Processing, pp.545–562, Marcel Dekker Inc, 2000

[3] 村上 征勝, 今西 祐一郎：“源氏物語の助動詞の計量分析”, 情報処理学会論文誌, Vol.40, No.3, pp.774–782, 1999

[4] 松浦 司, 金田 康正：“n-gram 分布を用いた近代日本語小説文の著者推定”, 情報処理学会自然言語処理研究会報告, NL Vol.134, No.5, pp.31–38, 1999

[5] 松浦 司, 金田 康正：“近代日本小説家 8 人による文章の n-gram 分布を用いた著者判別”, 情報処理学会自然言語処理研究会報告, NL Vol.137, No.1, pp.1–8, 2000

[6] CMU-Cambridge Statistical Language Modeling Toolkit v2, http://svr-www.eng.cam.ac.uk/ prc14/toolkit.html

[7] 形態素解析ツール茶筅, http://cl.aist-nara.ac.jp/lab/nlt/chasen.html

[8] 長尾 真, 佐藤 里史, 黒橋 禎夫, 角田 達彦：“自然言語処理”, 岩波書店, 1996

[9] 青空文庫, http://www.aozora.gr.jp/main.html

表 6: セット 1～セット 8 において使用したテキスト

作者	作品名
芥川竜之介	或敵打の話、或日の大石内蔵助、 地獄変、邪宗門、俊寛、 忠義、 南京の基督、杜子春、羅生門、河童 西郷隆盛、
伊藤左千夫	水害雑録、奈々子、野菊の墓、浜菊、 姪子、守の家
梶井基次郎	ある崖上の感情、ある心の風景、 交尾、城のある町にて、雪後、泥濘、 のんきな患者、冬の日、檸檬、 Kの昇天、橡の花
菊池寛	仇討三態、入れ札、大島が出来る話、 勲章を貰う話、ゼラール中尉、 船医の立場、父帰る、身投げ救助業、 吉良上野の立場、出世、俊寛、
国木田独歩	運命論者、牛肉と馬鈴薯、源おじ、 酒中日記、空知川の岸辺、武蔵野、 恋を恋する人
島崎藤村	岩石の間、旧主人、刺繍、新生、船、 芽生、藁草履
太宰治	老ハイデルベルヒ、ヴィヨンの妻、 駆込み訴え、グッド・バイ、姥捨、 富嶽百景、右大臣実朝、お伽草紙、 犯人、人間失格
中島敦	狐憑、悟浄歎異、山月記、名人伝、 南島譚・幸福、文字禍
夏目漱石	永日小品、教育と文芸、彼岸過迄、 博士問題とマードック先生と余、 満韓ところどころ、明暗、夢十夜、 吾輩は猫である、こころ、坊っちゃん
新美南吉	赤い蝋燭、牛をつないだ椿の木、嘘、 狐、ごんごろ鐘、手袋を買いに、 花のき村と盗人たち、病む子の祭 川、ごん狐
樋口一葉	うつせみ、大つごもり、たけくらべ、 にごりえ、ゆく雲、わかれ道
水野仙子	嘘をつく日、輝ける朝、四十餘日、 響、白い雌鶏の行方、道
宮沢賢治	オツベルと象、銀河鉄道の夜、 雁の童子、猫の事務所、風の又三郎、 ひかりの素足、ガドルフの百合、 双子の星
森鴎外	阿部一族、假名遣意見、かのように、 寒山拾得、栗山大膳、舞姫、妄想、 左橋甚五郎、山淑大夫、百物語、 キタ・セクスアリス、高瀬舟、 護持院原の敵討

表 7: セット α において使用したテキスト

作者	作品名
芥川竜之介	河童、地獄変
有島武郎	生まれいずる悩み、カインの末裔
伊藤左千夫	野菊の墓、水害
植松眞人	神楽坂奇譚、逢瀬までの。、 新世界交響曲、コーヒーメーカー、 神さんが降りてきた。、 続・神さんが降りてきた
小栗虫太郎	白蟻、人外魔境 天母峰
梶井基次郎	城のある町にて、橡の花、冬の日、 のんきな患者
菊池寛	仇討三態、吉良上野の立場、俊寛 勲章を貰う話
国木田独歩	運命論者、酒中日記、武蔵野
黒島伝治	渦巻ける鳥の大群、橇、電報、 豚群、「紋」、二銭銅貨
佐々木直治郎	アッシャー家の崩壊、黒猫、 モルグ街の殺人事件 (翻訳)
島崎藤村	岩石、新生
武田麟太郎	一の酉、釜々崎、現代詩、 日本三文オペラ
辻潤	え〝りと え〝りたす、錯覚自我説、 錯覚した小宇宙、自分だけの世界、 惰眠洞妄語、ふもれすく、浮浪漫語、 変なあたま
南部修太郎	S中尉の話、疑惑、女盗、 猫又先生、ハルピンの一夜
葉山嘉樹	淫賣婦、海に生くる人々、氷雨、 万福追想
樋口一葉	たけくらべ、にごりえ、行く雲
平井肇	外套、鼻 (翻訳)
二葉亭四迷	予が半生の懺悔、余が翻訳の標準、 四日間 (翻訳)、あいびき (翻訳)、 私は懐疑派だ、小説総論
南方熊楠	十二支考 (1)、神社合祀に関する意見
水野仙子	嘘をつく日、四十餘日、響、道、 白い雌鶏の行方
夢野久作	オシャベリ姫、巡査辞職