平成 27 年度

学士学位論文

ニューラルネットワークを用いた テキストデータからの低次元特徴の抽出

Low-Dimensional Features using Neural Network for Text Data

1160303　　川上 雄仁

指導教員　　情報学群 吉田真一

2016 年 2 月 26 日

高知工科大学 情報学群

要 旨

ニューラルネットワークを用いた
テキストデータからの低次元特徴の抽出

川上 雄仁

従来，機械学習を用いたテキストマイニングの研究では，特徴量として単語ベクトルが広く用いられている．最も設計が容易な単語ベクトルは，One-Hot 表現 (1-of-K 表現) と呼ばれ，語彙数分の次元数のうち 1 単語に対する値のみが非零で，それ以外を 0 とするものである．One-Hot 表現では，ベクトルの各要素に単語が対応するため，次元が語彙数となり一般に数万次元になる．この One-Hot 表現を応用して，ある文書の特徴量を表すために，その文書で用いられる単語に対応する要素を 1（または出現回数や TF-IDF 値)，それ以外を 0 とするものがある．この特徴ベクトルを用いて機械学習を行うことを考える際に，次元が数万次元となるため「次元の呪い」の問題が発生する．このため，次元の削減を行う方法として様々なものが提案されているが，その一つに近年ニューラルネットワークの一種である Word2Vec を用いた単語ベクトルの低次元表現が提案されている．本研究の目的は，Word2Vec を用いた低次元特徴量と，One-Hot 表現を用いた高次元特徴量との識別精度の比較を行い，テキストマイニングにおける低次元特徴量の有効性を検証することである．本研究では，経済新聞記事からの株価予測，日本語メール文書からの SPAM メール判定，ある商品に対する英字レビュー文書の内容が商品についての肯定的か否定的かの判定を行う．各記事に対して，形態素解析処理を行い，日本語の名詞，動詞，形容詞を抽出する．レビュー文書のみ英語文書のため，不要となる記号等のみを削除する．訓練データに含まれる単語のみについて Word2Vec により低次元特徴を抽出する．得られた単語の特徴量を 1 記事中の単語分だけ合計することにより 1 記事の文書ベクトルとし，訓練データとテスト

データを作成する．One-Hot 表現については，単語の出現頻度を記事ごとに求め，単語ベクトルを作成する．教師ラベルとして，株価のデータセットについては，該当する記事の日の日経平均終値が 5%を越えて増加すれば上昇，5%を越えて減少すれば下降，値動きが ±5%以内であれば値動きなしとした 3 クラスとする．SPAM メールの判定については，SPAM か非 SPAM の 2 クラス，レビュー文書の判定については，肯定か否定の 2 クラスとする．結果として，Word2Vec による単語の低次元特徴量を用いたテキストデータ識別は SPAM メールのデータについては識別率が 95.2%となり，従来の単語表現より 5.2 ポイント向上することを示す.

キーワード　ニューラルネットワーク，Word2Vec，テキストマイニング，単語ベクトル

Abstract

Low-Dimensional Features using Neural Network for Text Data

Yuto KAWAKAMI

In the research area of the text mining using the machine learning, word vectors are widely used to represent text documents. Word vector expression is also called as one-hot expression (1-of-k expression), and only the value for a word is non-zero and the others are zero. Using one-hot expression, the number of dimension of the feature is over several thousands. The reason of the high dimension of word vector expression is that the dimension equals to the number of word used in all documents. The curse of dimensionality is caused by high dimension when machine learning is performed using feature vectors. In order to reduce the dimension, several methods have been proposed. Recently, low dimensional expression of word vector using Word2Vec has been proposed using neural network. The purpose of this research is to evaluate the low dimensional features using Word2Vec compared with high dimensional features using one-hot expression in text mining using machine learning. In this research, stock price prediction from newspaper, spam mail discrimination in Japanese, and reviews of shopping item in English are used to compare. First, words included in training data are input to Word2Vec and it outputs low dimensional features. Next all low dimensional word vectors in a news article are summed and the vector of the summation is treated as a feature vector of a news article. All articles are converted to article feature vectors, and they are divided into training and test data. For one-hot expression, frequency

of word appeared in an article are used to make a word vector. The training label of stock market prediction are three categories, price up, price down, and stationary price. The training label of spam mail in Japanese are two categories, spam and non-spam. The training label of reviews of shopping item in English are two categories, positive meaning and negative meaning. As a result, the accuracy using low dimensional features is higher than high dimensional features.

key words neural network, Word2Vec, text mining, word vector

目次

図目次

表目次

第 1 章

序論

近年，テキストデータを基に特徴抽出を行い，テキストデータの識別予測を行う研究が進んでいる．テキストデータを用いての株価予測においては，学習器にニューラルネットワークを用いた予測や，重回帰分析を用いた予測，単一銘柄の予測，平均株価の予測等多くの手法で研究が行われた [1][2][3]．これらの予測等に用いられてた単語の特徴量は，One-Hot 表現（1-of-K 表現）と呼ばれる単語数分の次元数を持つ高次元な特徴量の利用が一般的であった．しかし，テキストデータ中の語彙数が膨大なデータに対しての識別予測を行う際には，特徴量が高次元となり機械学習時において次元の呪いが発生するという問題があった．そこで，各研究では，次元の呪いを回避するために特徴量の次元の削減を様々な方法を用いて行った．しかし，One-Hot 表現での特徴量の次元削減を行うためには，使用する単語の語彙数を限定することで次元の削減を行う必要があるため，全単語を利用した識別予測は設計上困難であった．その中，近年注目を集めているものが，ニューラルネットワークを用いた単語の低次元表現法である．

本研究では，現在注目を集めているニューラルネットワークを用いた単語の低次元表現法である Word2Vec を用いることで，単語ベクトルの低次元化を行うことで次元の呪いを回避し，従来の単語頻度を計算した One-Hot 表現（1-of-K 表現）の高次元特徴量との識別精度を比較し，Word2Vec のテキストマイニングにおける有効性を検証することが目的である．また，識別に用いる特徴量の次元数の変化と精度の変化についても検証を行う．

本論文の構成として，第 2 章ではテキストマイニングに関する技術および Word2Vec について説明する．第 3 章では識別予測に使用する機械学習に関する手法について説明する．第 4 章では識別実験についての説明する．第 5 章では結果と考察を述べる．

第 2 章

テキスト解析手法および特徴表現法と抽出技術

本章ではテキストマイニング技術を用いて識別予測を行うための技術の説明と，本研究で使用する技術の説明を行う．

2.1 形態素解析

形態素解析とは，自然言語処理技術であり，文章を単語に分割し，各単語の品詞を求める技術である．日本語の文章では英語の様に単語ごとの区切りが存在しないことから，文章のどの文字までが一つの単語になっているかの判断が難しい．そのため，日本語文を解析する際には，形態素解析を用いて，文章のうち意味を持つ最小の単位である形態素に分割する必要がある．今回は，形態素解析に MeCab を用いて，データセット中の記事データに対する形態素解析を行い，名詞，動詞，形容詞を抽出した．なお，英語の文章に関しては，単語ごとに区切りが存在するため，形態素解析を行う必要がないが，データ中の記号など余分な文字は削除する．図 2.1 は，形態素解析処理の例を示した図である．

2.2 One-Hot 表現と分散表現

One-Hot 表現とは単語のベクトル表現の一つであり，ある文章中に使用される語彙数分の要素があり，ある単語の表現をその単語に対応する要素のみが非零で，それ以外の次元は 0 となるものである [4]．このベクトルの次元は一般に数千〜十数万になる．この One-Hot

2.2　One-Hot 表現と分散表現

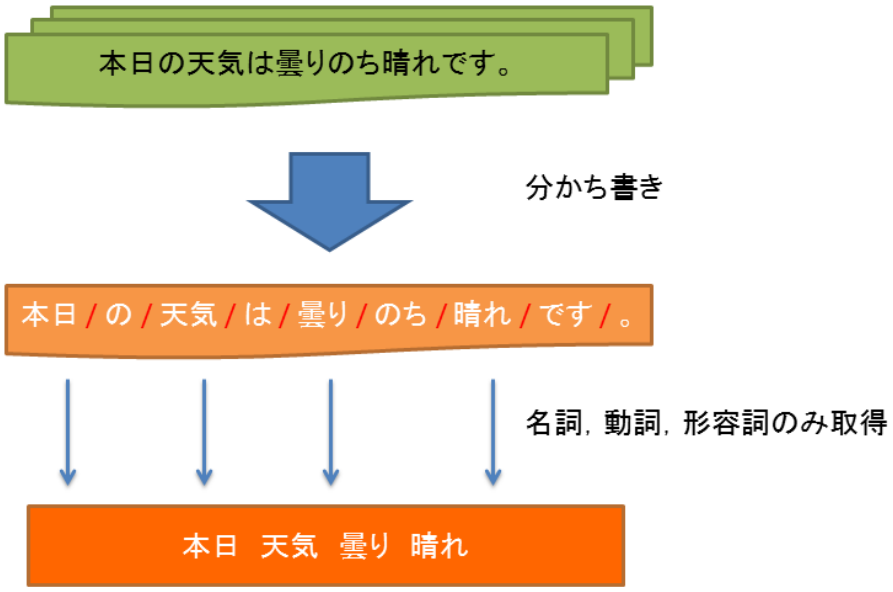

図 2.1　形態素解析の例

表現を用いて文章などを表現する際には，文章中に登場する各単語の One-Hot 表現を加算し，各単語の出現頻度を計算することで文章を表現する．このベクトルを記事数分にまとめた行列を単語文書行列という．一方，分散表現とは，単語や文章の意味などを数百次元ほどの固定長ベクトルで表現したものである．ニューラルネットワークはこの分散表現を導き出すために有効であり，テキストを基に学習させることで，各単語の分散表現ベクトルに対してどのような実数値を割り当てるかを決める．図 2.2 は，One-Hot 表現と分散表現の例を示した図である．

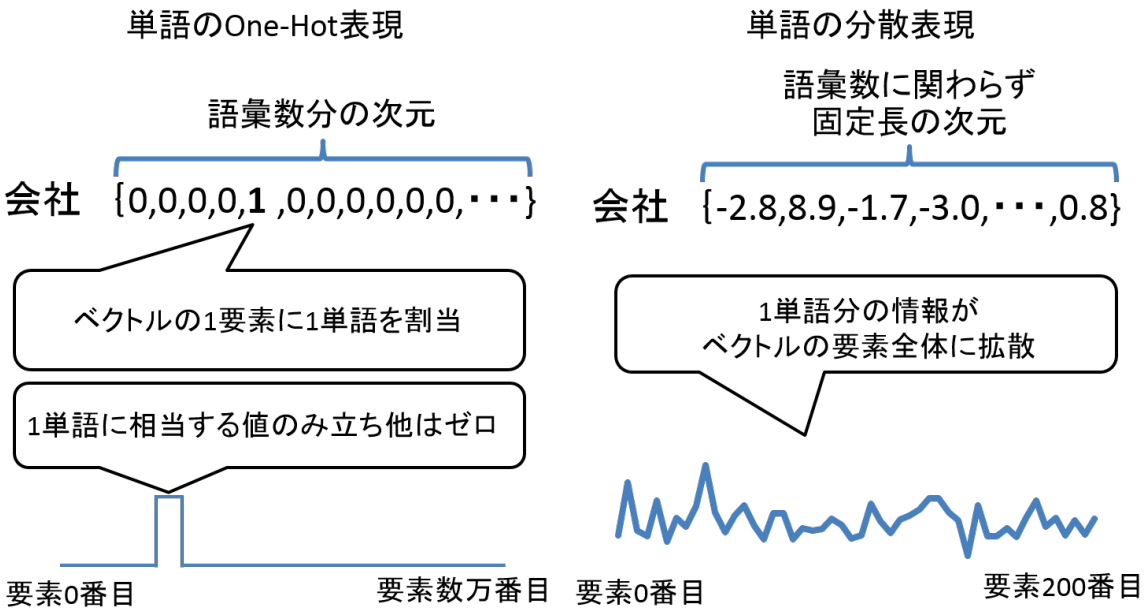

図 2.2　One-Hot 表現と分散表現の例

2.3　Word2Vec

Word2Vec[5] とは，米グーグル社の Tomas Mikolov 氏らが提案した手法を用いたオープンソースソフトウェアであり，ニューラルネットワークを用いた単語の分散表現を得られる．Word2Vec は，3 層のニューラルネットワークモデルとなっており，入力と出力が同じになるように値を学習する．テキスト中の各単語をその単語の前後の単語とともにニューラルネットワークに入力し，出力層に同じ値が現れるよう重みを学習する．これは一種のオートエンコーダのようなものである．学習後，中間層の入力値を抽出することで，各単語の単語ベクトルを取得する．ある単語の前後数単語から一つの単語を推定する方法を CBOW（Continuous Bag-of-Words）と呼び，一つの単語から前後数単語を推定する方法を Skip-gram と呼ぶ [6]．どちらの方式においても，入力層と中間層をつなぐ重み

が，Word2Vec が出力する単語の低次元特徴量となる．そのため，中間層のニューロン数が Word2Vec による単語ベクトルの次元数となる．また，Word2Vec による単語ベクトル表現により，単語間での意味演算も可能となる（例:Paris - France + Italy = Roma）．CBOW モデルは，図 2.3 に示すように，3 層のニューラルネットワークモデルとなっており，入力層には，注目単語の周辺単語 $w(t-2)$，$w(t-1)$，$w(t+1)$，$w(t+2)$ を入力とし，$w(t)$ を出力する．この低次元特徴の次元数は，ニューラルネットワークの中間層のノード数と等しく一般的には数百程度が利用されることから，入力層における語彙数分の次元数を数百次元にまで次元削減することが可能となる．Skip-gram は，図 2.3 に示すように，CBOW の反対で $w(t)$ を入力とし，$w(t-2)$，$w(t-1)$，$w(t+1)$，$w(t+2)$ を予測する．

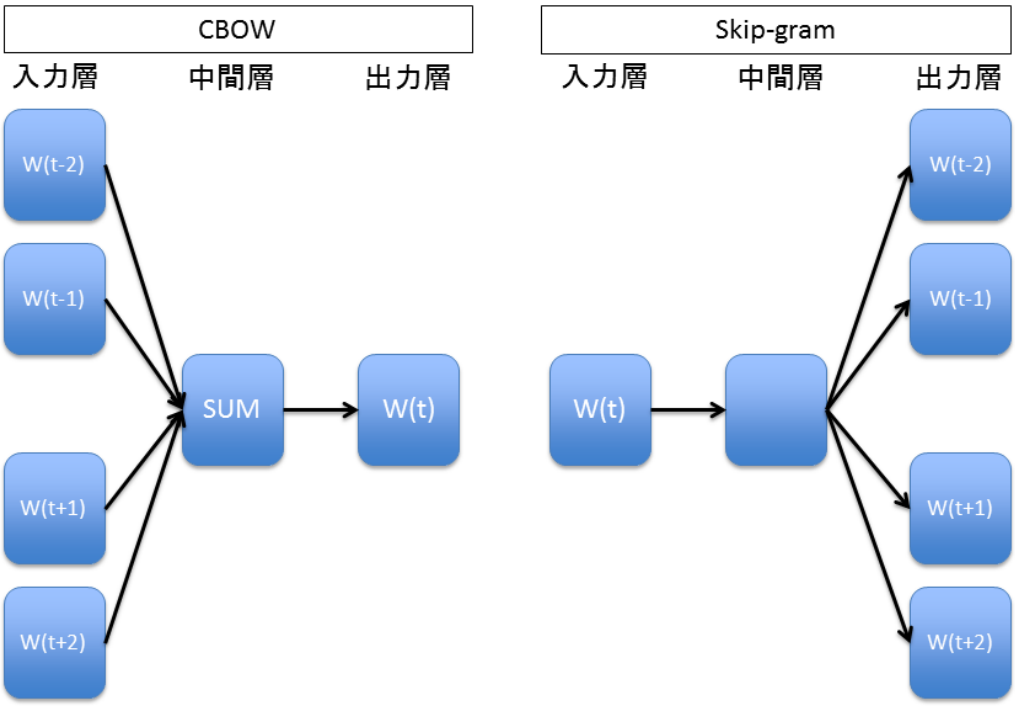

図 2.3　CBOW と Skip-gram の概略図

2.3.1　Word2Vec のニューラルネットワーク構造

Word2Vec のニューラルネットワークモデルは，上記でも記述したように入力層，中間層，出力層の 3 層のニューラルネットワークモデル（階層型ニューラルネットワーク）となっており，入力層と出力層のデータを同じものとして学習を行うオートエンコーダと同様の方法を用いて値の学習を行う．Word2Vec による単語の低次元表現法の概略図を図 2.4 に示す [7]．例では学習用のコーパスとして，数千〜数万単語ある中の “This is a pen” を例として入力する．入力層では，One-Hot 表現と同様に語彙数分の次元数となるため，数千〜数万次元の高次元となる．そのうち，“This is a pen” に対応する箇所のみが 1 となり残りの箇所は 0 という疎なベクトルとなる．この入力に対して，出力もできるだけ値が近くなるように学習をする．そして，学習が終了した後に，中間層の低次元ベクトルを取り出すことで，“This is a pen” を表す低次元のベクトル値を獲得する．

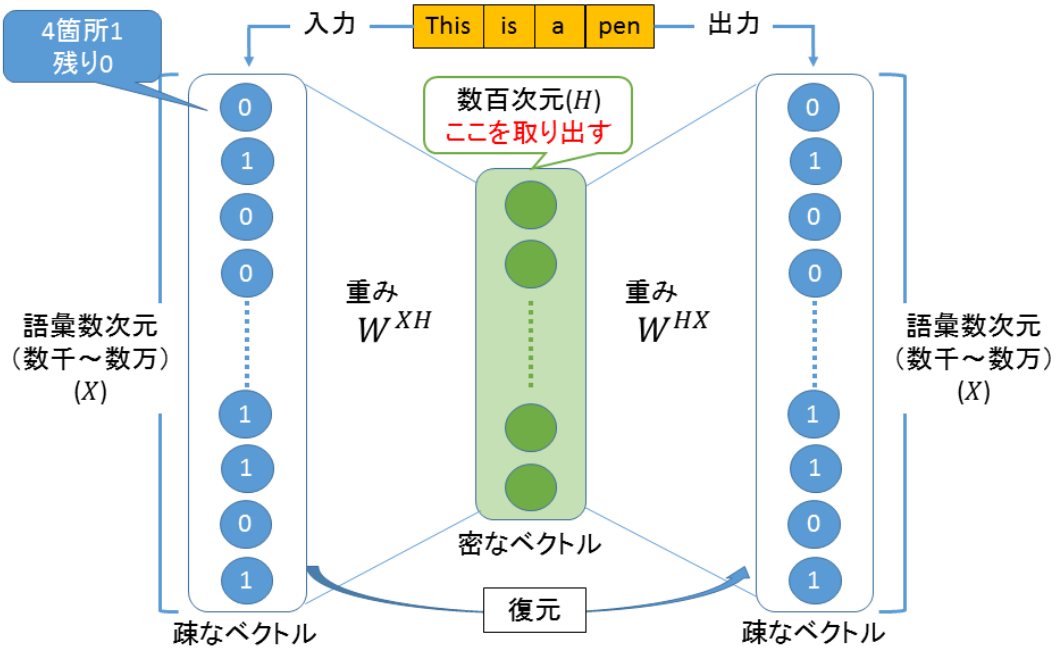

図 2.4　Word2Vec による単語の低次元表現

2.3 Word2Vec

Word2Vec では，入力層と出力層の値が同じになるように最急降下法を用いて学習を行い，誤差逆伝播法を用いてネットワークの重みの修正を行う．これにより学習が行われたネットワークに対して，中間層を取り出すことで，単語のベクトル値として利用する．

第 3 章

機械学習手法およびモデル評価法

本章では識別予測に用いる機械学習アルゴリズムについての説明を行う．また，機械学習の性能評価に用いる交差確認法についても説明する．

3.1　サポートベクターマシン

サポートベクターマシン (Support Vector Machine) とは，Vladimir N. Vapnik らが 1992 年に提案した教師あり機械学習法の一つである．サポートベクターマシンは，高次元特徴空間において線形関数による識別関数を求めるアルゴリズムの一つであり，その学習アルゴリズムの内部では最適化が行われる [8]．サポートベクターマシンの特徴として，訓練サンプルの各データ点との距離が最大となるように分離超平面を求める．これをマージン最大化と呼び，高い汎化性能が得られる．マージン最大化の概略図を図 3.1 に示す．このクラス間のマージンを最大にするために最適化処理を行う．また，サポートベクターマシンはカーネル関数を用いることで非線形分離問題にも優れた性能を発揮することが可能となる．今回は，カーネル関数として RBF カーネルのみを用いた．これは，RBF カーネルを用いることで無限次元の特徴空間を表現できるためである．

サポートベクターマシンを用いる際，特徴に RBF カーネルは，識別関数の自由度が高いため，適切にハイパーパラメータを設定する必要があるが，最適となるパラメータはデータセットによって異なる．そこで，グリッドサーチによるベストパラメータの探索を行う．

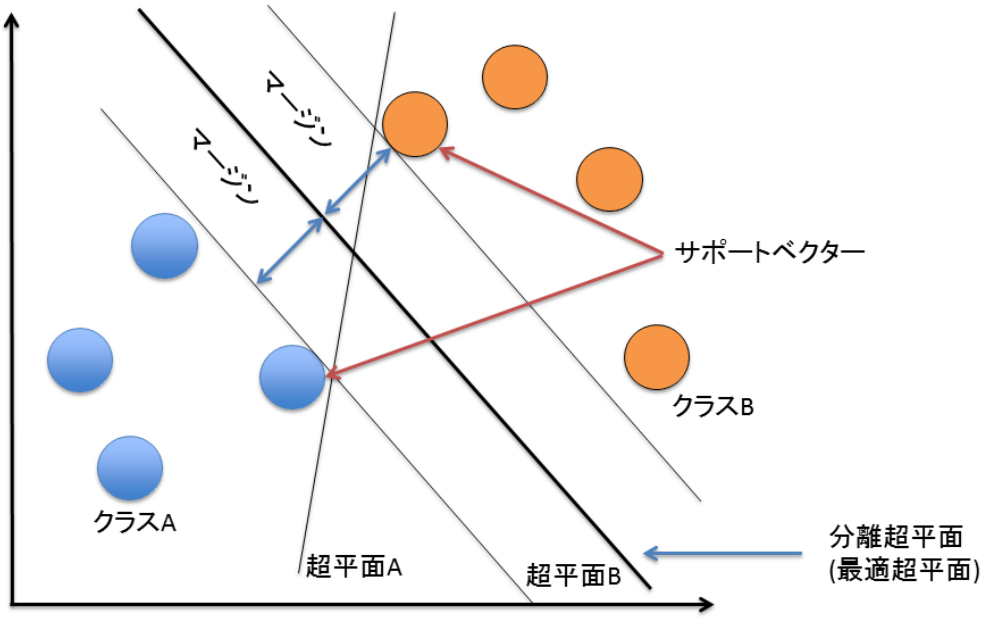

図 3.1　サポートベクターマシンの分離超平面

3.1.1　グリッドサーチ

グリッドサーチ (Grid Search) とは，パラメータの範囲を指定することによって，最適なパラメータの組を探索する方法である．サポートベクターマシンにおいては，訓練時における誤分類の許容範囲を決定する C と，RBF カーネルの識別関数の自由度を決定する γ の範囲を指定することで，学習データでの最適パラメータの組を探索する．C の値が小さいほど，誤分類を許容するように，大きいほど誤分類を許容しないようにする．γ とは，決定境界の複雑さを決める際に用いるパラメータである．γ の値が小さいほど，単純な決定境界となり，大きいほど複雑な決定境界となる．

3.2　交差確認法

交差確認法とは，学習器の汎化能力を評価するための手法である．学習は，一般に学習データに対する識別関数の出力値と教師データとの誤差が最小になるように，識別関数のパラメータを調整することである．一方，汎化能力とは，未知のデータに対する識別能力のことであり [9]，学習で得られた識別関数が学習データに含まれていない未知のデータに対しても有効であるかは不明である．そこで，学習データから取り除いておいたテストデータを用いて性能評価を行い，未知のデータに対する識別率で性能評価を行う必要がある．

この学習データとテストデータに分ける方法として代表的な手法にホールドアウト法がある．ホールドアウト法は，データを二つに分割し，一方を学習データとして利用し，もう一方をテストデータとして利用する．この手法は，データが十分にある場合には有効となる．データ数が限られている場合は，訓練データが多ければ，テストデータが少なくなるため，性能評価は悪くなる．一方，訓練データが少なければ，十分に学習ができなくなるため，学習の精度が悪くなる．そのため，データ数が十分に無い場合では，十分な性能評価を行うことができないという欠点がある．この欠点を補う方法が交差確認法である．図 3.2 のように，データをそれぞれ分割することで性能予測を行う．以下に交差確認法のアルゴリズムの概要を示す．なお，本研究では，ハイパーパラメータ（サポートベクターマシンの C および RBF カーネルの γ パラメータ）を決定するために，下記のように交差確認法を用いて性能を計測する．最終的な汎化性能評価には，株価が時系列であることを考慮し，交差確認法による評価は行わない．

1. データを a 個に分割する．
2. 1 つ目のデータを評価用データ，残りを訓練データとして，分類の評価を行う．
3. 2 つ目のデータを評価用データ，残りを訓練データとして，分類の評価を行う．この際，
   1. で用いた評価データは訓練用データとして用いられる．
4. 同様に，a 個目までのデータを評価用データとして評価するまで繰り返す．
5. 4. までの評価で得られた正答率などの評価における指標値を平均し，そのデータセッ

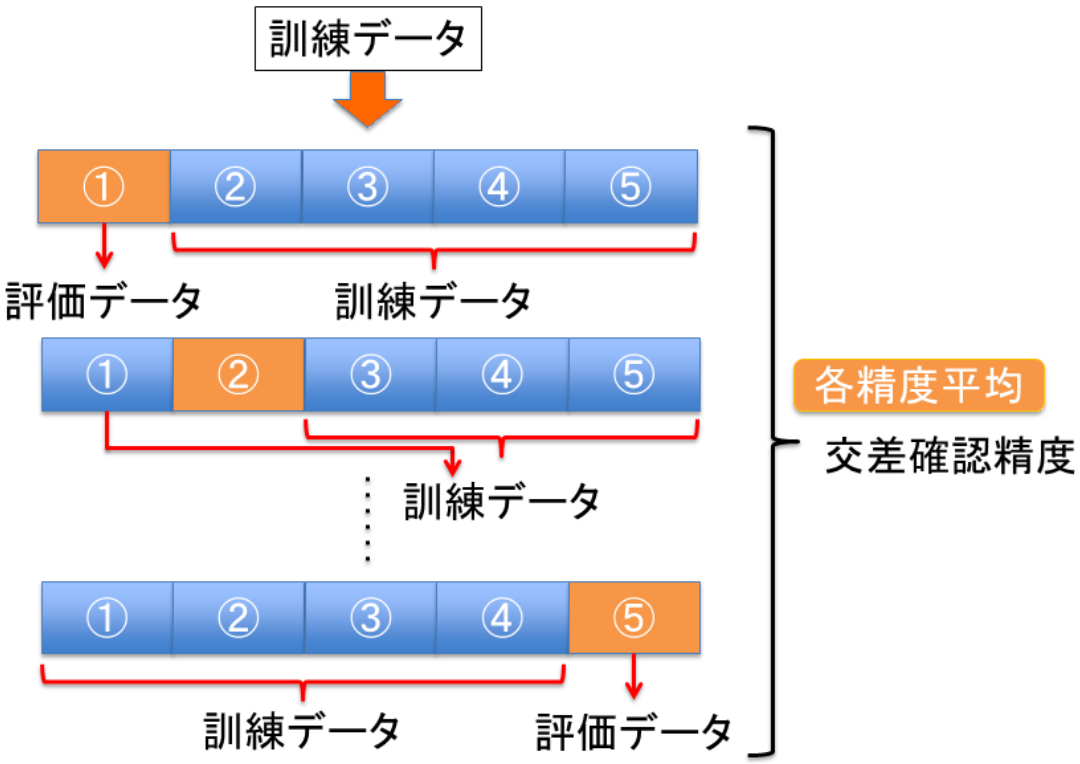

図 3.2　交差確認法の概略図

トに対する最終的な指標値とする.

第 4 章

テキストデータによる識別実験

本章では学習用データの作成手順とそれに合わせてのプログラムの変更および実験手法についての説明を行う.

4.1 実験環境

本実験を行うにあたって用意した環境は表 4.1 のとおりである.

表 4.1 実験に用いたハードウェア・ソフトウェア

OS	Ubuntu 14.04.3 LTS
CPU	Xeon E5540 2.53GHz
メモリ	32GB
利用ソフトウェア	Python 2.7.6
	sklearn 0.17 (Python ライブラリ)
	MeCab（形態素解析ソフト）

4.2 実験手順

本研究では，Word2Vec による低次元特徴量と One-Hot 表現による高次元特徴量を用いてテキストデータ識別の比較を行うことで，Word2Vec による低次元特徴量の有効性を検証する．与えられたテキストデータに対し，形態素解析処理を行い日本語の名詞，動詞，形容詞のみを抽出する．今回使用するデータ中には，英字データも含まれるため英字データに関

しては，不要となる記号を削除する処理を行う．形態素解析処理を行ったテキストデータに対し，１記事中の単語の頻度を計算した単語頻度ベクトルを生成することで，One-Hot 表現での高次元特徴量を獲得する．これに対し，本研究ではテキストデータに対し，記事中の各単語を Word2Vec により単語の低次元特徴を計算し，１記事で使用する単語の低次元特徴量を合計することにより，１記事に対する低次元特徴量を獲得する．学習には，サポートベクターマシンを使用し，サポートベクターマシン内のパラメータはグリッドサーチによりベストパラメータを決定する．評価法には，データセットをあらかじめ訓練データとテストデータに分割し，訓練データのみを用いて交差確認法を用いて学習し，テストデータにより識別を行うようにした．実験デザインを図 4.1 に示す．

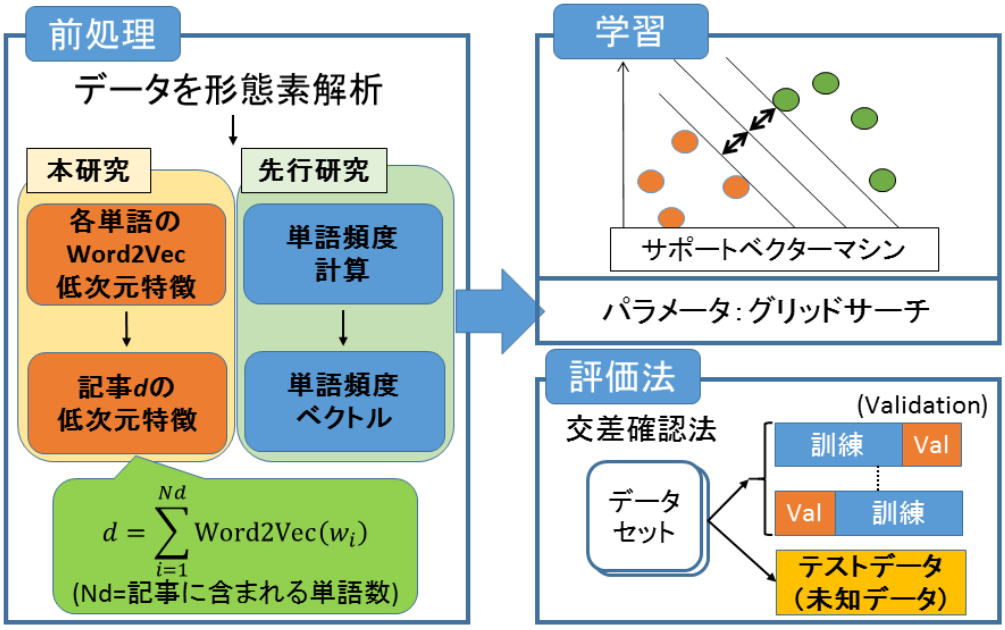

図 4.1　実験手順図

4.2.1 Word2Vec による低次元特徴量抽出

Word2Vec による出力の Word.csv と WordValue.csv を出力するように，オリジナルの Word2Vec.c プログラムを書き加え，Python ライブラリの機能であるディクショナリ配列に格納することで，連想配列を作成する．あらかじめ，形態素解析時に出力した記事中に含まれる単語を記事ごとにファイルとして生成する．１記事中の使用単語を参照し，単語の低次元特徴量を加算することで１記事分の特徴量（合計値ベクトル）とする．ある文書 r を単語列 $w_1, w_2, \cdots w_{N_d}$ としたとき，r に対する文書ベクトル d は，次式で与えられる [10].

$$d = \sum_{i=1}^{N_d} \text{Word2Vec}(w_i) \quad (N_d = \text{記事に含まれる単語数})$$

以下，低次元特徴量の単語ベクトル生成までの手順を以下に示し，その概略図を図 4.2 に示す.

1. データセット全体に形態素解析処理を行う
2. 1 データ毎に使用単語を抽出し，text 化する
3. 訓練データ分の記事を traintext.txt として 1 つにまとめる
4. Word2Vec による単語ベクトル化を行う
5. 出力された word.csv，wordvalue.csv を用いて連想配列を作成する
6. 2 で作成した text から，各データの合計値ベクトルを連想配列から計算する　※ 1
7. alldata.csv として，全データを 1 つ csv ファイルにまとめる

※ 1:テストデータ中のみに出現する新単語に関しては，Word2Vec によるベクトル化が行われていないため，加算されない.

4.2.2 One-Hot 表現を用いた高次元特徴量抽出

本研究で比較対象とする One-Hot 表現を用いた単語ベクトル生成までの手順と概略図を図 4.3 に示す.

1. データセット全体に形態素解析処理を行う

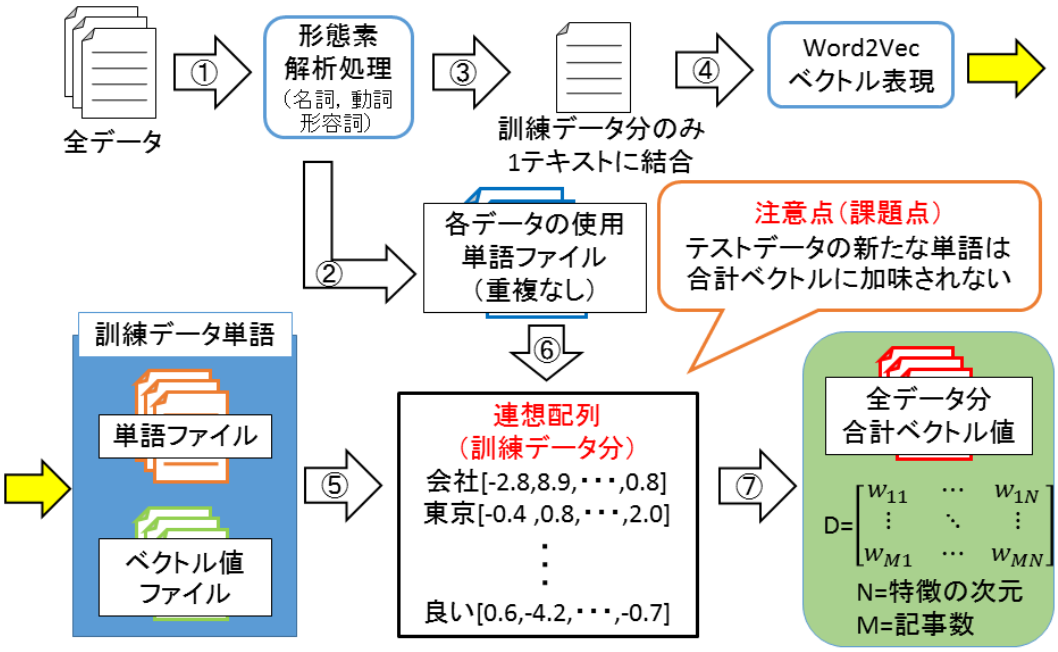

図 4.2　Word2Vec を用いた低次元特徴量獲得手順

2. 全データ中で使用される単語の頻度を計算する

3. 単語文書行列として csv ファイルにまとめる

4.2.3　サポートベクターマシンのハイパーパラメータ

今回使用したサポートベクターマシンのハイパーパラメータは以下の通りである．

使用カーネル：RBF

グリッドサーチ：C:1, 10, 100, 1000, gamma: 100, 1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-10

交差確認回数：5

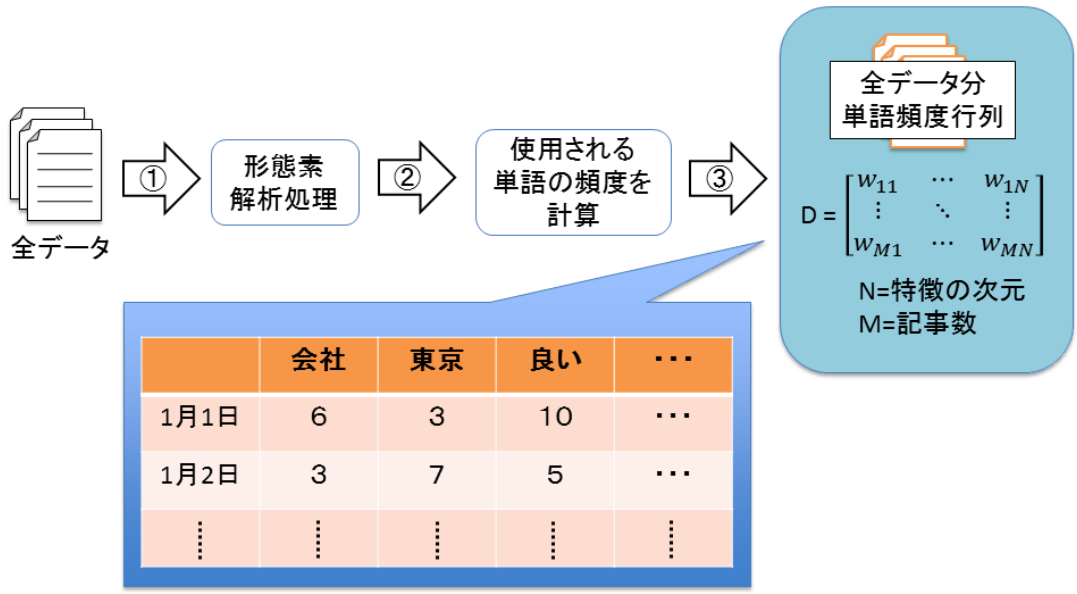

図 4.3　One-Hot 表現を用いた高次元特徴量獲得手順

4.2.4　データセット

本研究では，3 つのデータセットを用いてテキストデータから低次元特徴量を抽出し，各データに対する識別予測を行う．1 つ目は，日本経済新聞の 2014 年朝刊記事 234 日分を対象とし，株価の値動きの識別予測を行う．訓練には，1 月から 11 月までの新聞記事データを用い，テストには 12 月の新聞記事データを用いた．2 つ目は，独自に用意した日本語メールデータを対象とし，SPAM メールであるか非 SPAM メールであるかの判定を行う．3 つ目は，Amazon が公開する英字レビューデータを対象とし，ある商品に対する英字レビュー文書の内容が肯定的であるか否定的であるかの判定を行う．データセットの内訳を表 4.2 に示す．

表 4.2　データセット内容

データセット	総データ数	訓練データ数	テストデータ数
2014 年新聞記事	234	213	21
日本語メールデータ	5556	5000	556
Amazon 英字レビューデータ	1000	800	200

4.2.5　教師ラベルの生成

株価の値動きは，日経平均株価の終値が前日比 +0.5% 以上の増加を上昇，前日比 −0.5% 以上の減少を下降，その間を変動なしとした 3 クラスを教師ラベルとして用いる．SPAM メールでは，SPAM であるか非 SPAM であるかの 2 クラスを教師ラベルとして用いる．レビュー文書では，ある商品の肯定文書であるか否定文書であるかの 2 クラスを教師ラベルとして用いる．各教師ラベルの内訳を 4.3 に示す.

表 4.3　教師ラベル内訳

データ	ラベル値	ラベル意味	訓練ラベル数	テストラベル数
株価	-1	下降	62	7
	0	停滞	84	8
	1	上昇	67	6
メール	0	SPAM	2500	278
	1	非 SPAM	2500	278
レビュー	0	否定	391	109
	1	肯定	409	91

4.3　Word2Vec のプログラム変更

本研究では，C 言語で配布されている Word2Vec のプログラムに一部ソースコードを書き加えた．追記内容としては，読み込んだ単語を CSV ファイルに出力する処理，単語に対してのベクトル値を CSV に出力する処理である．今回は，単語と単語ベクトル値の出力に合わせ，Vector.bin を出力させることにより，Word2Vec のサンプルコードにある distance によって，単語間の学習が出来ているか確認できるよう設定した．以下，追記したソースコードを図 4.4 に示す.

```
//Original_Word2Vec.c の void TrainModel() 関数内に以下の内容を追記.
  //Original 546 行目追記
  FILE *foword;
  char *fword = ‘‘word.csv’’;
  FILE *fovalue;
  char *fvalue = ‘‘wordvalue.csv’’;

  //Original 563 行目追記
  fprintf(foword, ‘‘%s\n’’, vocab[a].word);

  //Original 564 行目追記
  fprintf(fovalue, ‘‘%lf ‘‘, syn0[a * layer1_size + b]);

  //Original 612 行目追記
  fclose(foword);
  fclose(fovalue);
```

図 4.4　Word2Vec ソースコード

4.4　Word2Vec のパラメータ

本研究では，One-Hot 表現を用いた特徴量と比較するため，Word2Vec による単語削減を極力行わないようにした．そのため，Word2Vec.c 中の min_count 変数の初期値を 1 に

設定することで，出現回数が 1 回の単語も出力するようにしている．Word2Vec の実行時のパラメータは下記に示す．シェルスクリプトにて，任意の次元数まで処理を繰り返すように設定する．

```
for i in Dimension
do
./word2vec -train traindata.txt -output vectors.bin -cbow 1
                      -size \$i -window 8 -negative 25 -hs 0
                      -sample 1e-3 -threads 20 -binary 1 -iter 15
done
```

図 4.5　Word2Vec 実行コマンド

各種パラメータの説明は下記に示す.

- 学習モデル:CBOW
- 次元数:Dimension で定義する次元数
- ウィンドウサイズ:最大 8 単語
- ネガティブサンプリング:25
- ソフトマックス関数:使用しない
- ダウンサンプリング:10^{-3}
- スレッド:20
- バイナリ出力:する
- 学習の反復回数:15

第 5 章

結果・考察

5.1 結果

実験結果を表 5.1 に各データセットに対する識別率のチャンスレベルを表 5.2 に示す．また，各識別結果の混同行列を表 5.3〜表 5.8 に，Word2Vec での 1 次元から 50 次元までの識別率結果を図 5.1 に示す.

表 5.1 各手法の識別率 (±S.D. 値)

手法	データ	識別率 [%]	次元数	学習時間 [s]
Word2Vec	株価	40.0(±2.3)	9	0.33
	SPAM	95.2(±0.9)	24	32.0
	レビュー	60.5(±4.5)	37	2.2
One-Hot	株価	38.0(±0.0)	54935	231
	SPAM	90.0(±0.0)	17726	19647
	レビュー	70.0(±0.0)	2118	93

表 5.2 各データセットのチャンスレベル

株価予測	SPAM メール予測	商品レビュー予測
38.0%	50.0%	54.5%

各データセットのチャンスレベルとは，各データにおいて識別結果がすべて 0 と識別した際の識別率である．新聞記事からの株価予測においては，テストデータ数が 21 件に対して

表 5.3 Word2Vec 株価予測

予測\真	下降	停滞	上昇
下降	1	6	0
停滞	0	8	0
上昇	0	6	0

表 5.4 One-Hot 株価予測

予測\真	下降	停滞	上昇
下降	0	7	0
停滞	0	8	0
上昇	0	6	0

表 5.5 Word2Vec SPAM メール

予測\真	spam	非 spam
spam	267	11
非 spam	15	263

表 5.6 One-Hot SPAM メール

予測\真	spam	非 spam
spam	249	29
非 spam	24	254

表 5.7 Word2Vec 英字レビュー

予測\真	否定	肯定
否定	54	55
肯定	24	67

表 5.8 One-Hot 英字レビュー

予測\真	否定	肯定
否定	70	39
肯定	21	70

(Word2Vec の結果は 5 回の試行の平均値を四捨五入した値)

停滞ラベル数が 8 件のため，すべて 0 と識別した識別率は，21 件中 8 件の正答となり識別率は 38%となる．その他データセットにおいても同様の計算方法となる．

Word2Vec を用いた低次元特徴量による識別率は，ニューラルネットワークによる学習の段階でランダム値が割り当てられるため，単語ベクトルに誤差が生じることから 5 試行同じ識別を行った後，平均値を算出した．結果としては，単語の低次元表現により次元削減が行われ，学習時間の大幅な短縮が見られた．また，識別率に関しては SPAM メール識別において次元数 24 のとき識別率 95.2%という結果となったことから，一部データに対しては識別率の向上が見られた．Amazon 英字レビューの識別に関しては，混同行列の結果を見ると，否定文書識別では識別率 50%，肯定文書識別で識別率 74%により全体の識別率が 60%前後という結果になっていることが分かる．これは，否定文書では内容を細かく書こうとすることから，否定文書での使用単語にばらつきが生じてしまい，識別が困難となったのではないかと考える．新聞記事からの株価予測の識別率に関しては，識別率のチャンスレベ

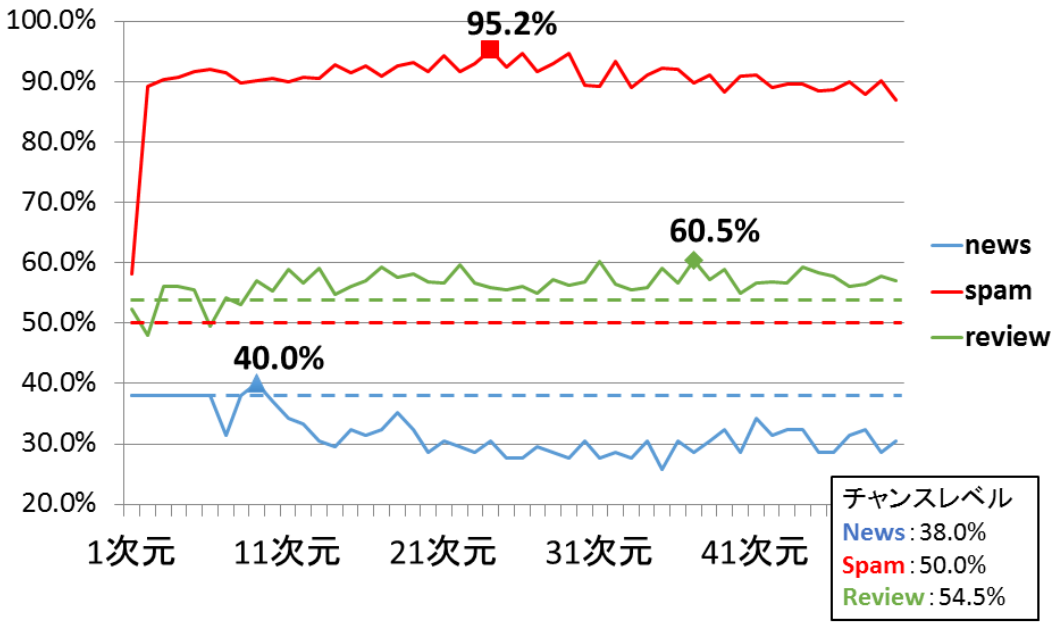

図 5.1　1 次元から 50 次元までの識別率

ル 38.0%とほぼ同等という結果となった．これは，使用語彙数と指定した次元数の関係から，Word2Vec による単語の低次元表現時の学習が上手く行えていないことが予測される．そのため，新聞記事からの株価予測に関しては，前処理においての工夫を行い再実験を行った．

5.1.1　追加実験:相関係数を用いた単語選出

各単語の出現回数と教師ラベルとの相関係数を求めることで，算出した相関係数の絶対値の降順に単語ベクトルを並び替え，上位単語のみを利用した特徴量を用いて識別を行う．2 組のデータ列を $(x_i, y_i)(i = 1, 2, ...n)$ とすると，相関係数は式 5.1 で表される．$\bar{x}, \bar{y}$ はそれぞれのデータの平均である．

$$\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \tag{5.1}$$

5.1　結果

One-Hot 表現では，相関係数の絶対値の高い順に 1 単語から 1000 単語（次元）までを使用した場合の結果を図 5.2 に示す．結果として，One-Hot 表現を用いた特徴量においては相関係数を用いて単語を選出した場合，79 次元および 80 次元にて 57.1%を示し識別率の向上が見られた．

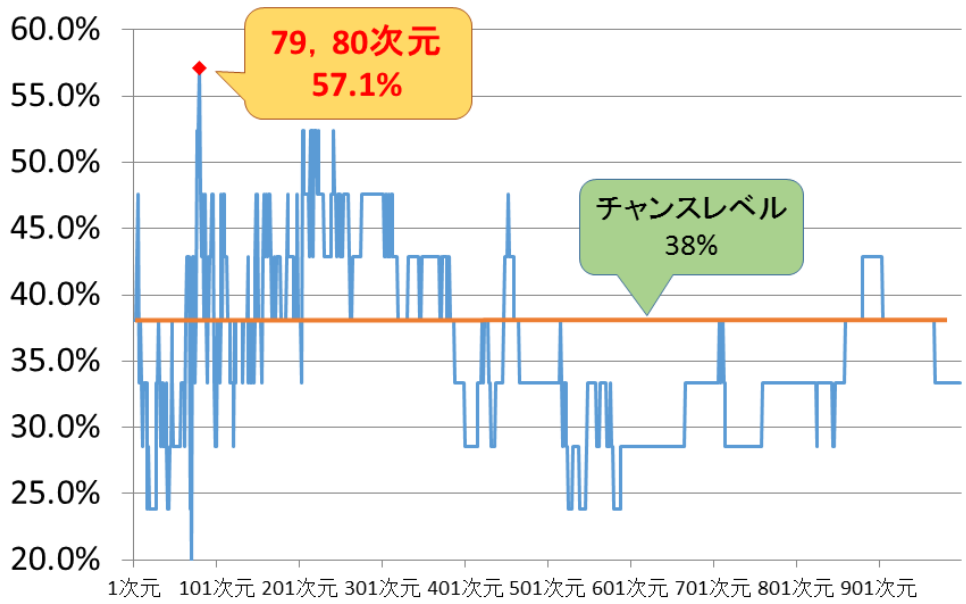

図 5.2　相関係数を用いた高次元特徴量での識別率

Word2Vec による低次元特徴量を用いる場合は，One-Hot 表現で最も識別率の高かった単語数を使用する．また，記事中の単語の重複を認める場合と認めない場合での差を観察した結果を図 5.3 に示す．今回は，One-Hot 表現での結果から 80 単語を使用し，10 次元から 100 次元までを 10 次元ごとに測定した．結果として，記事中の単語の重複を認めない場合では，識別率のチャンスレベルを越えることがなかったが，重複を認めた場合では，次元数 60 のとき識別率 47.6%を示し識別率の向上が見られた．

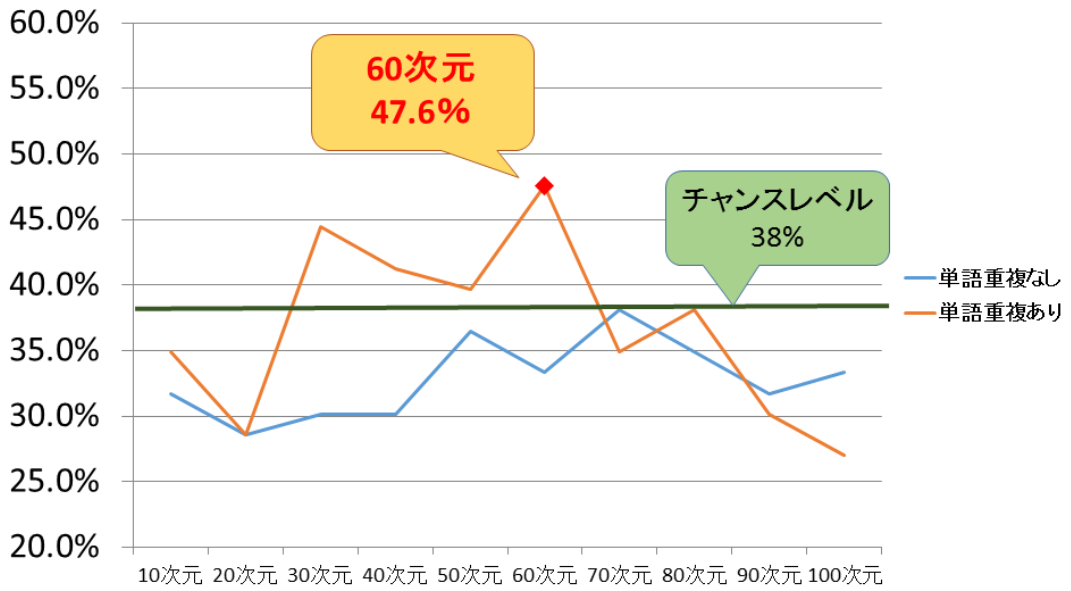

図 5.3　相関係数を用いた低次元特徴量での識別率

5.2　考察

本実験から，記事中の単語の数により識別精度に差が出るのではないかと考える．単語数が多い場合においては，低次元特徴量で計算するデータの合計値が増加することが考えられる．そのため，特徴量設計の段階で，教師ラベルと関係する層が出来ているのではないかと考える．また，相関係数を用いた実験結果から単語の出現頻度を考慮する必要もあると考えられる．記事中の単語の出現頻度を考慮しないことは情報量の欠落につながるため，語彙数を限定した特徴設計を行う場合においては，出現頻度を考慮することでベクトルの合計値に変化が表れ，識別がしやすくなるため，単語の出現頻度を考慮した特徴設計が必要になると考える．

新聞記事からの株価予測に関してはサポートベクターマシン内でのハイパーパラメータの設定に交差確認法を用いているため，訓練時ハイパーパラメータ設定には未来のデータから

過去のデータを識別して学習を行っていることになっていることから，厳密には未知のデータの識別になっていない．そのため，新聞記事からの株価予測においては，交差確認法によるバリデーションセットの選択ではなく，1 月から 10 月までの記事データを訓練データとし，11 月の記事データを訓練時のバリデーションセットとして用いて学習器の訓練を行い，12 月の記事データを未知のデータとして識別することが正確な識別となると考える．

第 6 章

まとめ

本研究では，ニューラルネットワークモデルを使用する Word2Vec を用いた低次元特徴量によるテキストデータからの識別の有効性について検証を行った．結果として，次元数の削減により次元の呪いが回避でき学習時間において優位な差が見られ，SPAM メールの識別においては次元数 24 のとき識別率 95.2%と従来の One-Hot 表現を用いた識別よりも識別率の向上が見られた．そのことから，Word2Vec を用いた低次元特徴量はテキストデータからの識別において有効であると結論づけられる．ただし，データセットによっては識別率に変動が見られ，従来の One-Hot 表現よりも識別率が低下する場合もあった．これは，Word2Vec 実行時パラメータに関しての詳細な検討をしておらず，デフォルトのパラメータを使用していることも原因であると考える．そのため，Word2Vec による単語の低次元表現を行う際に，対象とするデータセットに対しての最適なパラメータ設定が今後の課題として挙げられる．

謝辞

本研究を進めるにあたり，ご指導して頂きました高知工科大学 情報学群 吉田真一准教授に心より感謝申し上げます．吉田先生には，研究室配属からお世話して頂き，本研究の内容のみならず．サーバやネットワークの知識やプログラムコマンドなど，様々なことを教えて頂きました．時には，夜遅くまで吉田先生の経験談や研究室の先輩のお話，またネットワークの技術や少しコアなお話など，時間を忘れる位にいつも楽しかったです．本研究を進めるにあたっては，内容理解だけではなくプログラム上でのサポートや新しい実験機材を提供してくださったおかげで，研究をより進めることができたと感じます．また，大学説明会での研究室代表やファジィ学会での合宿に参加する機会を与えてくださったりなど，貴重な経験をさせて頂けて吉田研究室に配属して本当に良かったと感じております．就職してからも，お会いする機会がありましたら，宜しくお願い致します．

副査を心よく引き受けてくださった高知工科大学 情報学群 岩田誠教授ならびに福本昌弘教授に心より感謝申し上げます．研究に対しての的確な助言および文書の添削等のおかげで良いものにすることができました．岩田教授には，輪講を通して技術に対する思考力を鍛えて頂きました．福本教授には，研究室の行事を通して気さくに話しかけてくださり，また自分の考えに対して助言を下さったりと数多く助けて頂きました．両教授には，重ねて深く感謝申し上げます．

同研究室の松尾氏，塩見氏，西本氏，前原氏，田中氏には大変お世話になりました．修士 2 年生の松尾氏には，本研究をはじめとする機械学習に関する知識や，TA として授業のレポートに対する知識や技術の内容など挙げればきりがないほどご指導を頂きました．また，私が 4 年になってからはご飯にも誘って頂き，その際にも自身の研究の内容を理解してご指導頂いたりなど大変ためになりました．よく恐い印象を持たれやすいとのことでしたが，とても気さくで個人的な印象としましては，松尾氏の学部時代のお話や自身の経験のお話は大変面白かったです．次にお会いする時には，再び新しい体験談を聞かせてください．同学年

の塩見氏には，研究が同じ分野ということからプログラムの解読や LaTex での関数などを教えて頂きました．塩見氏は，すごく天然で心配になる点が多くありましたが，その明るいキャラクターのおかげで研究室が賑やかでした．次に再会できる日はいつになるかは分かりませんが，そのキャラクターでまた笑わせてくれる日を楽しみにしています．西本氏には，いつも気さくに話し相手になって頂き，輪講の際の数学式や英語など様々な場面で助けて頂きました．研究室での活動においても，一番頼りになる存在であり，夏合宿では車を出して頂いたりと本当に助かりました．飲み会での酔った西本氏はいつもよりもテンションが高く面白かったです．また，高知に戻ってきた時には仲良くしてくれると嬉しいです．前原氏にも，大変お世話になりました．同学年では唯一の女性ということも有り大変な思いをしていたかもしれませんが，前原氏のデザインのセンスや元気な姿を見る度に，改めて関心していました．発表前などには弱音を吐く私に対して，川上さんなら大丈夫と声をかけてくれたのは嬉しかったです．また，塩見氏との言い争いは見ていて楽しく，いつも笑わせてもらいました．田中氏は，研究室配属の時に大変お世話になりました．私の思いつきの行動に対して付き合ってくれたり，席が隣ということから雑談をしたりなど大変楽しかったです．田中氏が同時に配属されたことで，楽しく研究室活動を始められたと感じます．残念ながら 4 年では，あまり話をすることが出来ませんでしたが，また一緒にゲームや雑談をできる機会を楽しみにしています．

同研究室の後輩の皆様には，研究室でのイベントの幹事などお疲れ様でした．夜遅くまで作業を行ったり，何かに向かって集中しているときの皆様は本当によくやっていたと思います．皆様が吉田研究室に入ってきてくれたおかげで，研究室内の雰囲気も明るくなり楽しく研究室での活動を行うことができました．来年は，次の後輩に対して指導できるよう就職活動および大学院進学に向けて頑張ってください．

そして，私を大学まで通わせてくれた家族に対しては，本当に心より感謝申し上げます．手間のかかる子供ではありましたが，今度は私が支えられるように頑張りたいと思いますので，これからも宜しくお願い致します．

最後に，大学 4 年間で出会った友人や先輩方並びに後輩には改めて御礼申し上げます．

参考文献

[1] 山口 裕輝, “ テキストマイニングによる株価予測に適した機械学習,” 高知工科大学学士学位論文, 2012.

[2] 中井 淳人, “ 株価の時系列変化の予測のための特徴選択,” 高知工科大学学士学位論文, 2013.

[3] 奥村 順哉, “ディープラーニングによる経済記事テキストデータを用いた株価予測,” 高知工科大学学士学位論文, 2014.

[4] 進藤 智則, “ディープラーニングは万能か,”pp.44-52, 日経エレクトロニクス 2015 年 6 月号, 2015.

[5] “Word2Vec,” https://code.google.com/archive/p/word2vec/ (2015.9.1)

[6] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, “Efficient Estimation of Word Representations in Vector Space,” Cornell University Library arXiv.org, arXiv:1301.3781v3 [cs.CL], 2013.

[7] 西尾泰和, “word2vec による自然言語処理,” O’ReillyJapan, 2014.

[8] Nekki Cristianini, John Shawe-Taylor 著, 大北 剛 訳, “サポートベクターマシン入門,”p9, 共立出版株式会社, 2005.

[9] 平井 有三, “はじめてのパターン認識,” 森北出版株式会社, 2012.

[10] 加藤 和平, 大島 考範, 二宮 崇, “Word2Vec と深層学習を用いた大規模相伴分析,” 言語処理学会 第 21 回年次大会, 2015.

[11] 岩井 美樹, 二宮 崇, “word2vec に基づく熟語項構造の分布表現獲得,” 言語処理学会 第 21 回年次大会, 2015.

付録 A

識別結果詳細

本実験での 1 次元 50 次元までの各データの識別率を以下に示す.

表 A.1: 1 次元から 50 次元までの識別率

次元数	株価予測	SPAM メール	レビュー予測
1 次元	38.0%	61.7%	52.4%
2 次元	38.0%	86.5%	48.0%
3 次元	38.0%	94.0%	56.1%
4 次元	38.0%	94.0%	56.1%
5 次元	38.0%	95.3%	55.5%
6 次元	38.0%	95.1%	49.5%
7 次元	31.0%	95.4%	54.2%
8 次元	38.0%	95.4%	53.0%
9 次元	40.0%	94.4%	57.0%
10 次元	37.0%	95.6%	55.4%
11 次元	34.0%	95.9%	58.9%
12 次元	33.0%	94.3%	56.7%
13 次元	30.0%	95.4%	59.0%
14 次元	30.0%	95.6%	54.8%
15 次元	32.0%	95.6%	56.1%

表 A.1: 1 次元から 50 次元までの識別率

次元数	株価予測	SPAM メール	レビュー予測
16 次元	31.0%	95.3%	57.0%
17 次元	32.0%	94.8%	59.2%
18 次元	35.0%	96.0%	57.5%
19 次元	32.0%	95.8%	58.1%
20 次元	29.0%	96.0%	56.8%
21 次元	30.0%	96.1%	56.7%
22 次元	30.0%	96.2%	59.7%
23 次元	29.0%	96.3%	56.6%
24 次元	30.0%	96.0%	55.8%
25 次元	28.0%	96.3%	55.5%
26 次元	28.0%	96.0%	56.1%
27 次元	30.0%	96.3%	54.9%
28 次元	29.0%	96.0%	57.2%
29 次元	28.0%	96.4%	56.3%
30 次元	30.0%	96.5%	56.9%
31 次元	28.0%	96.5%	60.3%
32 次元	29.0%	96.7%	56.4%
33 次元	28.0%	96.5%	55.5%
34 次元	30.0%	96.3%	55.8%
35 次元	26.0%	96.7%	59.0%
36 次元	30.0%	96.1%	56.7%
37 次元	29.0%	96.4%	60.5%
38 次元	30.0%	96.2%	57.2%

表 A.1: 1 次元から 50 次元までの識別率

次元数	株価予測	SPAM メール	レビュー予測
39 次元	32.0%	96.2%	58.9%
40 次元	29.0%	96.4%	54.9%
41 次元	34.0%	96.4%	56.7%
42 次元	31.0%	96.8%	56.8%
43 次元	32.0%	96.2%	56.6%
44 次元	32.0%	96.8%	59.3%
45 次元	29.0%	96.4%	58.3%
46 次元	29.0%	95.8%	57.8%
47 次元	31.0%	96.3%	56.0%
48 次元	32.0%	96.5%	56.5%
49 次元	29.0%	96.4%	57.8%
50 次元	30.0%	96.2%	57.0%

付録 B

選出単語

相関に基づく単語選出によって選出された単語で最も識別率の高い上位 1000 単語を以下に示す．Word2Vec での相関処理では，上位 80 単語のみを利用して特徴量を計算した．

表 B.1　選出単語一覧

ホット	余暇	ＭＩＴ	並	東北大学	占拠	画質
使用	唯一	絆	ちりばめる	快適	盛り	再現
女流	地層	雄一郎	主席	痩せる	ボーダーレス	課長
車中	缶詰	光男	飛び込み	学歴	見送り	駆使
小川	退去	安原	分社	隆司	早める	容器
竜王	完璧	高層	鵬	地裁	海面	綱渡り
太刀打ち	遠慮ない	まき	つきまとう	押しつぶす	敏弘	含める
トータル	安川	越	引け	見聞き	造反	愚か
見なす	像	僅差	豊田合成	魅せる	あれこれ	筆
わし	下向き	魔よけ	クジラ	抜け道	設備	済み
分断	蘇州	協和	生誕	ボトルネック	歩	蓋
つかう	加賀谷	建て替える	頂上	取っ手	ヵ年	合点
冠婚葬祭	新進	フリー	散発	恩	衰退	楢原
孝行	乗り継ぐ	スタートライン	広東	智絵	手伝い	主将
長野	変革期	話しかける	ダーバン	独力	徳之島	コーティング
丹念	由香	桃子	落胆	雨期	一昔	おふくろ

用具	振り返る	執筆	補う	出品	華道	愛知製鋼
半日	松戸	そごう	就職	白酒	モニタリング	東武ストア
ＣＡＴＶ	入力	札幌	心臓	運ぶ	新鮮	脂肪
フリル	有益	速まる	東京理科大	メニュー	秋口	逆襲
電光	大豊	再訪	授賞	われわれ	光熱	マック
食い違い	製作所	産出	物資	気質	メンツ	ダラス
いたす	研	兼松	尺度	昌夫	フロー	移民
今どき	大洗	はき出す	首尾	会談	ハンガリー	後ろ向き
イギリス	燃料	考慮	おやじ	女王	白黒	お祝い
懇親	俊夫	国男	柏木	排せつ	争い	凍結
担い手	丼	貧しい	重み	教頭	煙突	舌鼓
戻せる	富田林	水口	アイルランド	燃焼	日比野	カナダ
長谷川	目新しい	請求	患う	幾何	択捉島	しんぽ
放り込む	外回り	しょく	容姿	カンパラ	そっけ	隆太
凍り付く	あんた	党紀	みと	神体	献上	杉並
遊園	物体	村越	買い付ける	成し遂げる	ヘクタール	テラス
かたどる	東洋	矢口	症	チャンピオンズ	職場	飛島
豪	貼り紙	帝政	室長	作成	美女	外す
高め	人民大会堂	そっぽ	両方	多	石室	インフラ
市庁舎	ユニオン	景気動向指数	柴山	バージョン	格納	方程式
コロラド	和樹	最高検	コンサルタント	荒木	悪影響	感情
宮田	八郎	アブドラ	不全	貧困	勇	細い
上物	覇	観艦式	同書	指揮	休む	西方
ミルボン	杉野	中興	応え	強める	水野	ドライバー
背任	メガビット	光一	攻撃	フラ	護岸	ファイナル

打者	賃	文武両道	四捨五入	間際	書き	取り込み
歓喜	中２	決裁	にぎやか	呼び出す	水産庁	堀江
裏金	チーズ	ドット	暦年	貢献	本社	信介
一部	製油	過ぎる	不測	触れ込み	ランダム	セリ
上司	尾山	喉頭	形づくる	栗本	察す	取り決める
けんじる	整地	国共	出し切れる	笛吹	双竜	温存
教	言い残す	うなずける	勧業	のぶ子	雷門	和枝
治彦	レーヨン	かわいそう	さぼる	奔放	バミューダ	来社
湖畔	アレックス	とく	シリーズ	表彰状	哲司	波風
焦げつく	師傅	晴れる	地理	自社	せり	忘れ
高品	あずかる	かなた	同案	物販	閉鎖	収れん
返納	帝国データバンク	チャット	盗み出す	思想	柔らかい	ＬＴ
ジャンル	横浜銀行	偏見	防戦	エアロ	政財界	森内
良しあし	入ろう	弾丸	港	戦国	脆弱	無責任
あす	メンタル	フュージョン	粗末	勧	南城	ベトナム
ドクター	意中	抵当	インスタント	焼酎	膝	ナンバー
引き離す	古都	管内	カビ	社長	狩野	核分裂
浩二	落札	若い	ファンド	短	寛治	酸味
リハビリ	鉱業	捕鯨	取り出し	安保理	京子	好ましい
割る	スタンフォード大	芸妓	自慢	かん	真央	大垣
曇らせる	格闘	興す	送受信	間口	大塚	別府
売れる	バランスシート	借り換え	混入	大発会	しごく	製氷
成人の日	敬吾	手近	貴文	添う	紆余曲折	片腕
譲り合う	書院	東札幌	経典	社葬	卓哉	さとる
瓜	夕闇	長子	備え付け	原題	亡国	危なげない

邦題	エフ・ディ・シィ・プロダクツ	満月	武川	暴騰	甲高い	トリック
押出	暁子	設け	拘留	八丈島	目移り	テレビジョン
はけ	手料理	遣う	業主	融解	大濠公園	露見
落着	かぼちゃ	宝飾	暗闇	測れる	高揚	当たり前
よぎる	たどり着ける	ウガンダ	鉄器	コ ー ディネート	素っ気	東松島
改定	ブラック	ありがたい	ガラス	決めつける	かき	日本ケンタッキー・フライド・チキン
信奉	彩色	国名	刑罰	同局	追分	龍谷大
抜粋	増長	酸性	憲治	自公民	引き合う	溶ける
淳一	許可	ＧＤＰ	誘致	使途	アトピー	東洋紡
読解	銀色	ふるう	あい	楽しみ	南極	野沢
ジョブズ	値動き	トランク	交わす	ウエスコ	貝印	書く
寿樹	オリ	げいこ	土踏まず	各駅	区民	人情味
英夫	陰陽	でっち上げ	安城	めくれる	近付く	善意
安易	停車	器具	下支え	日本製鋼所	相乗り	無罪
真保	パイプライン	クラブ	小杉	山木	紛失	効用
付属	重なる	水田	採択	古田	久門	ウォール街
一義的	変心	邦訳	蔚山	明く	敏い	政則
そろい踏み	米村	尾畑	古河機械金属	意外	冷やす	決選
無配	時事	合間	仕える	ヒーター	九大	英徳
中神	媛	救世主	隊	特	議定	堅
スコッチ	希	メーンバンク	偉い	モーグル	引き出し	バレル
都立	隠れ	勤	ガイダンス	在籍	略式	トレー
出頭	土俵	破綻	あっせん	もらえる	試料	点

徴	館山	演技	溶接	撮像	常	イルミネーション
死に体	うめく	理容	わかる	翻す	考え直す	毎年
ＪＲ西日本	座り込む	口出し	絶え間	使い捨て	受け付ける	大なた
軍配	慢性	小中学生	稲盛	歩調	正幸	慶
坂井	政春	能美防災	橋口	麻原	近鉄百貨店	柘植
溶岩	国鉄	寛容	脳裏	強者	申しあげる	風圧
延び	空転	ディン	マークシート	みつる	原島	眼前
ひろあき	りかん	布製	浅草	焼き物	土壌	求刑
メガネ	東大	打って出る	全社	所沢	厳戒	衛
いたる	緻密	ゴム	ジョージア	簡単	革命	エルニーニョ
用材	看取	人肌	無尽蔵	金髪	粗大	老若
なすりつける	たかい	天下分け目	石飛	すする	とおい	湿地
不整脈	崇高	武蔵野美術大学	比類	アトラス	貴彦	併任
めげる	金網	片りん	洗髪	撃ち合う	万葉	泰成
あつみ	金納	やぶさか	兼備	幻滅	青天井	育める
舌触り	目張り	燃やせる	串焼き	議運	全数	いっさい
みちお	トーンダウン	火気	はじき返す	紀尾井	重病	上関原発
せわ	ふとい	映り	旦那	暴れ	メラニン	死角
鉄骨	ログ	物質	渡り歩く	パラマウントベッド	チョコレート	習い事
主宰	覆る	向上	バックス	添加	三条	支所
喜べる	未踏	大将	語源	旅行	ニッポン	財務
スピーディー	完走	鷺	組み替える	友好	タイプ	繰る
食費	将軍家	官房	教訓	北陸電力	市町村	パース
煩雑	花街	佐藤	遺産	省内	陽性	忠則
これら	滞る	グルメ	耕す	関心事	古語	弘光

甲羅	専修大学	火ダネ	満ち足りる	回覧	陣地	誘導体
芝崎	弘明	津田沼	海風	活用度	老眼	洗い場
倉荷	でんぱ	常石	企て	帰京	テルアビブ	フランス
封	雅雄	朝鮮人民軍	聯合	ＦＧ	黒船	ＷＢＣ
銘じる	刺す	将軍	様相	号機	身長	功罪
仮説	差し引き	原子	弁護	圧巻	注水	延長
伊丹	カシオ計算機	マネジャー	建株	ネック	体育	下水
動議	思いつき	ケチャップ	ピカ	おこす	実装	抑える
坪井	バウンド	命運	智夫	喜朗	栽培	純度
真紀	ジョーンズ	尾鷲	主張	全体	融	浪江
神秘	磨き	明治維新	卓越	基板	丸い	三菱
申	橘	精子	勝ち取る	しなやか	ぶす	ばね
スモール	裏腹	コカ・コーラ	邦	不法	トヨタ	中西
感触	ナッシュ	丸順	鉄心	教室	大間	祐
和成	洗い流す	深沢	壁画	すき	立ち会い	見過ごす
寸断	真っ先	稲野	初旬	敏夫	散らす	のぞみ
大倉山	クチマ	鯉	小野田	三豊	病変	理想科学工業
但馬	大塚商会	アワード	辻井	帝京大学	朝市	紀生
除数	カンゾウ	北海道銀行	多難	姿勢	メンバー	ヒラリー
大林	ジャマイカ	願う	無駄遣い	懲役	募債	ヒビ
オンワード	宿願	病巣	顔合わせ	制憲	強がる	パスカル
成熟	ペ	決行	ソング	カモ	デュアル	智彦
北越	塗装	メッセージ	不向き	開封	ダボス	スパイク
巻き込む	故事	福永	各自	換気	隣る	節電
コープ	桜田	重労働	クアラルンプール	禁煙	北海道電力	阪神

出光興産	ＮＩＨ	全部	一真	糸島	望遠鏡	ヤマ
演算	プラズマ	慕う	上場	ラジオ	致死	ニジマス
宴席	スターン	船便	すがすがしい	ステロイド	深海魚	導入
アセアン	フローリング	保管	シスコ	最低限	彫刻	