JAIST Repository
https://dspace.jaist.ac.jp/

Title	構文構造を用いた感情極性分類の精度向上
Author(s)	中山，貴樹
Citation	
Issue Date	2012-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/10421
Rights	
Description	Supervisor:東条敏，情報科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

修 士 論 文

構文構造を用いたテキスト感情極性分析の精度向上

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

中山　貴樹

2012 年 3 月

修 士 論 文

構文構造を用いたテキスト感情極性分析の精度向上

指導教官　東条　敏　教授

審査委員主査　東条　敏　教授
審査委員　白井　清昭　准教授
審査委員　島津　明　教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

1010045 中山　貴樹

提出年月: 2012 年 2 月

Copyright © 2012 by Yoshiki Nakayama

概 要

本研究では、ユーザによって書かれた文章を感情極性に基づいて自動的に分類するための手法の改善について提案を行う。近年、インターネット上で大量のユーザの書いた文章を取得することが容易になった。そして、この文章を有効に活用することで、文章中に書かれた意図を定量的に取り出すことが可能になる．その研究分野のひとつとして，テキスト感情極性分類がある．他方、近年の構文解析技術の発達により、構文構造を利用することで、Bag-of-Words のような表層的なモデルではとらえられない単語間の関係を比較的正確に解析できるようになっている。そこで我々は係り受け構造を利用し機械学習を行うことで，文の構造を考慮した 分類方法は提案されていない．

目 次

第 1 章　はじめに **1**
1.1　研究の目的と背景 1
1.2　本論の構成 2

第 2 章　関連研究 **3**
2.1　テキスト分類 3
2.1.1　感情極性分類 3
2.1.2　感情極性単語の発見 5
2.1.3　感情極性辞書 6
2.2　テキスト分類に対する機械学習的アプローチ 8
2.2.1　確率的分類器 8
2.2.2　決定木による分類器 8
2.2.3　サポートベクターマシン 9
2.3　構文構造 11
2.3.1　チャンキング 12

第 3 章　係り受け情報による学習データの構築 **16**
3.1　手法概要 17
3.2　手順 1　１つの単語で独立した特徴ベクトルの抽出 17
3.3　手順 2　N-gram によって分けられた単語での特徴ベクトルの抽出 18
3.4　手順 3　感情単語においての特徴ベクトルの抽出 19
3.5　提案手法　係り受けを考慮した特徴ベクトルの抽出 21

第 4 章　実験 **25**
4.1　語彙の獲得 25
4.2　実験の概要 27
4.3　実験結果 28
4.3.1　手法１の結果と考察 28
4.3.2　手順２の結果と考察 30
4.3.3　手順３の結果と考察 30
4.3.4　提案手法の結果と考察 34
4.3.5　係り受けによる単語の意味と極性の逆転回数 37

4.3.6　実験結果の比較 37

第 5 章　おわりに **39**

5.1　本研究のまとめ 39

5.2　今後の課題 40

図 目 次

2.1 決定木分類器 9
2.2 サポートベクターマシーン 11
2.3 トークンとチャンクの両方における分割とラベル付け 13

3.1 素性ベクトルの抽出 18
3.2 バイグラムでの抽出例 19
3.3 日本語評価極性辞書 20
3.4 係り受け情報 21
3.5 逆転 22
3.6 係り受け情報 23
3.7 素性ベクトルの抽出 24

4.1 素性ベクトルの抽出例１ 26
4.2 素性ベクトルの抽出例２ 26
4.3 素性ベクトルの抽出例３ 27
4.4 好意的なレビューの一例 27
4.5 非好意的なレビューの一例 28
4.6 unigram を用いたデータセット 29
4.7 bigram を用いたデータセット 31
4.8 trigram を用いたデータセット 31
4.9 unigram と感情極性単語を用いたデータセット 33
4.10 bigram と感情極性単語を用いたデータセット 33
4.11 trigram と感情極性単語を用いたデータセット 34
4.12 unigram と提案手法を用いた実験結果 35
4.13 bigram と提案手法を用いた実験結果 36
4.14 trigram と提案手法を用いた実験結果 36
4.15 各実験の比較結果 38

第1章　はじめに

1.1　研究の目的と背景

近年、インターネット上の大量のテキストが、様々な物事に関する情報を得るための重要な情報源となっている。テキストから著者の感情に関する情報を得るための技術として、テキスト感情極性分類と呼ばれる技術があり、企業による新製品の評判分析などに用いられている。感情情報処理技術とは，テキストを解析し包括的な調査をすることにより人の感性を客観的に評価することを目的とした技術である．その技術の１つとしてテキスト感情極性分類があり，これを用いて新製品のサーベイなど様々な場で用いられている．

感情極性分類は、人の手によって書かれたある対象について述べられている文章などを対象として、その文章の中に存在する言語表現が好意的，非好意的な極性であるかを自動的に判定する技術である．好意的または非好意的とは人が書いた文章を自動的に好意的または非好意的かに分類する為には，テキスト分類での Bag-of-Words モデルを使うのが一般的である．Bag-of-Words モデルとは，各単語が素性空間における一つの次元であると仮定する文書表現である．Bag-of-Words モデルを用いた感情極性分類の既存の手法は，文書中に出現する単語を素性とした機械学習によるアプローチの１つである．しかし、Bag-of-Words モデルでは、単語の表層的な感情極性を独立にとらえることしかできないため、高い精度を達成することは難しい．文書の中には，ほとんどの場合，好意的または非好意的な単語が複数含まれて書かれている．そのため，文書全体としての感情極性を正確に測ることが困難になる．さらに，逆説や否定などにより極性が反対になることもある．

複数の単語列による表現の感情極性は，例えば“ 美しくない ”という複数語では ”美しい ”だけでは好意的な極性だが，美しくない ”という全体を考慮した場合，非好意的な極性として捉える必要がある．また，文書全体を捉えるために，重要な要素になるものの１つとして接続詞がある．接続詞は，前後の文脈の関係を表す語で、品詞の一つである．接続詞によって，文書は前後の文で感情極性が変化を捉えることが，評価文分類で重要であるといえる．このような背景を踏まえ，本論文では係り受け構造を利用した感情極性分類を行う．

1.2　本論の構成

本論文の構成は以下の通りである．第２章でテキスト分類，構文構造の関連研究を述べる．第３章では実験に用いた手法と，構文構造から係り受け情報の抽出方法を述べる．第４章では実験結果について述べる．第５章ではまとめと今後の課題について述べる．

第2章　関連研究

2.1　テキスト分類

テキスト分類 [13] という技術は，与えられたテキストをあらかじめ人手で決定したいくつかのカテゴリに自動分類するタスクである．さらに,「機械可読」であった文書の表現形から「機械に理解可能」な文書の表現形に移行させる技術である．テキスト分類するためのタスクの１つとして，各文書を少数の概念またはキーワードでタグ付けすることである．タグ付けに使うことが出来る概念やキーワードのセットは，人手で事前に準備されており，そのセットの中で閉じていて，かつサイズも比較的小さい．キーワード間の階層構造も，事前に人手で用意されている．

テキストを分類する際に，文章をそのままの形では分類器を用いて直接処理することはできない．そのため，前処理の段階で文書を扱いやすい表現に変換する必要がある．本研究では，扱いやすい表現として素性ベクトルによって表現する．素性とは内部構造を持たない単純な要素であり，素性空間の一つの次元である.

一つの文書のすべての単語を素性として用いる方法を，Bag-of-words モデルという．さらに，素性空間の次元は基本的に二値の重みを与える方法を用いる．それは，素性に対応する語が出現していれば１，していなければ０という方法である.

そして，文章集合の異なる単語数は，文書の種類によって大きくも少なくもなる．大きな文書における単語の異なり数は膨大になりうる．これらの単語の大部分は分類タスクとは無関係であり，分類器の性能をおとすことなく取り除けるばかりか，取り除くことでノイズ除去による性能向上をもたらすこともある．無関係な単語を取り除くこの前処理の段階を素性選択と呼ぶ.

また，機械学習によるアプローチでは，あらかじめ分類済みの訓練用文書を用いてカテゴリの属性を学習させることで，分類器を自動的に構築する.

2.1.1　感情極性分類

感情極性分類は，評価情報を観点とした文書分類であり，ある評価文書が好意的極性か非好意的極性のいずれの極性をもつかを判定する課題である．評価文書分類を実現することは，次のような状況において有益な情報を提供する．例えば，ある商品に関する評価文書に関して，好意的極性をもつ評価文のみを集めることで，その商品の優れている点を

把握することができる．

評価表現の比率に基づく評価文書分類の手法は，評価文書中に存在する固有表現に着目し，好意的極性をもつ評価表現と非好意的極性をもつ評価表現の出現比率に従って，評価文全体の評価極性を求める．

Turney ら [4] は，評価極性値を求める語句に制限を付け，形容詞を含む句に対してのみ評価極性値を求めている. 彼らは，分類の手続きを３つのステップから構成した．

ステップ１ 評価文書に含まれる評価表現を評価極性値で抽出する

ステップ２ 抽出された評価表現の極性値の平均を求める

ステップ３ 平均極性値に応じて，文書全体の評価極性を決定する

評価実験は，自動車，銀行や映画に関するレビューを用いている．その実験の結果 74.39% の分類精度を得たと報告している．

また，文書の感情極性分類において，単語の出現パターンを用いて機械学習に基づく分類をするという試みもなされている．

Pang ら [1] は，教師あり機械学習に基づく文書分類が，評価文書分類にも有効であるかどうかを実験的に検証した．Unigram での語彙と感情極性単語や文章のサブカテゴリを素性ベクトルとし，それらの組み合わせる手法を提案している．そして，３種類の機械学習を用いて，それぞれの分類精度の比較をおこなっている．彼らの研究では，感情極性単語をあらかじめ固定しており，例えば“ *dazzing* ”や“ *suck* ”などである．これらのベクトル集合を用いて，ナイーブベイズや最大エントロピー法やサポートベクターマシンによる分類を行なった．実験結果により，サポートベクターマシンを用いて 82.9%という精度で分類できている．

また，松本ら [9] は，文章中の単語間の関係についての情報を系列パターンまたは部分依存木パターンと捉えることにより単語の出現パターンを抽出し，文書内に出現する単語の両方を分類に用いて分類した．系列パターンを捉える理由として，文中に出現する連続または非連続な単語列のパターンで，隣接していない２つ以上の単語間が複数存在した場合でも，抽出することができるという考えからである．部分依存木パターン捉える理由として，単語動詞の関係を親子関係として含むことで，単語の羅列である平文に比べ，単語間の関係の情報が整理されていると捉えることができるという考えからである．これらの考えを基に，映画のレビューデータセットを用いて実験を行ない，正解率が 85.6%を得たと報告している．

高村ら [7] は，複数語から成る評価表現のモデルおよびそれに基づいた分類手法を提案している．彼らの研究では「名詞＋述語」の場合を扱っている．そして名詞に対応する確率変数，述語に対応する確率変数を用意して，確率モデルを考えている．さらに複数語をとらえるために，隠れ変数をモデルに導入している．このモデルを用いて，感情極性という観点で類似した語がグループ化されるようなクラスタリングを実現し，感情極性分類を行なっている．

これらの研究では，複数語から素性ベクトルを用いることは，文書の感情極性分類において有用であることは示されているが，構文構造を用いた感情極性分類は言及されていない．

他にも，文書の階層に着目することで評価情報を得る研究として Pang ら [2] による手法が挙げられる．彼らは主観的な文章が文書全体の感情極性に有用であると考え，各文書の主観性を推定し文書の感情極性分類を行うという手法を提案している．そして，主観的な文章であるか客観的な文章であるかをグラフ理論を用いてコスト計算を行なっている．実験結果では，ナイーブベイズを用いた手法で 86.4%という精度で分類できたと報告している．

他には，好意的か非好意的かの２値で分類するのではなく，より細かい分類を行なっている研究もある．その１つとして，Pong ら [5] は，多値分類への対応として，分類ではなく回帰の考え方を導入することにより，２値よりも分類数の多い分類手法を提案している．彼らは metric labeling 法を適応することで，既存の分類器によって得られる結果を修正するしている． metric labeling 法は，高い類似度をもつ評価文書は，同等の評価点を持つという考えを取り込むことができ，評価文書間の類似度が高い時には，それらの評価文書間の差を小さくするような学習を行う．また，彼らは好意的極性文の比率という尺度に基づく類似度関数を提案している．この類似度関数は,「好意的文書の数」÷「文書の全文の数」としている．評価実験として，サポートベクターマシンと One-vs-Rest 法の組み合わせ，Support Vector Regression と metric labeling 法の３種類を比較している．実験結果は，３クラスに分類した場合 75%の分類精度，４クラスに分類した場合 65%の分類精度と報告している．

2.1.2 感情極性単語の発見

テキスト分類や情報抽出の技術を応用して，意見の記述されている範囲で，その意見が好意的に記述されているか，それとも非好意的な記述かの判断することは解決すべき問題の一つである．

飯田ら [6] は，意見を対象，属性，評価値の３つ組みとして定義し，文章からその３つの組みを抽出する手法を提案している．この枠組では，要約処理を４つのサブタスクに分解して考えている．

1. 属性，属性値，主観評価となりうる表現の収集
2. テキスト無いの属性と属性値の対の抽出
3. 抽出した対の好意的または非好意的判定
4. 判定した好意的または非好意的の値を用いた意見の要約である

彼らの研究では，意見性を持った＜属性，評価値＞の対を対象文章中から抽出することを目的としている．ただし，対の抽出には，次の２つの問題が混在している．

1. ＜属性，評価値＞の対を同定する問題

2. 同定した対が意見性を持つか否かを判定する問題

この問題設定に対して，提案手法では，対象文章内に存在する評価値となりうる候補と対になる属性を同定し，同定した対を用いて評価値候補が意見性を持つか否かを判定する，という２段階の処理を行い，最終的に＜属性，評価値＞のペアを抽出している．候補の抽出は，属性辞書と評価値辞書を用いて，まず対象文章内から評価値辞書の項目に該当する評価値候補を探し，その評価値候補に対してある範囲内に存在する属性候補を網羅的に抽出する．次に，評価値候補と対になる属性の同定の問題において，評価値候補とそれと対になる可能性のある属性候補集合が与えられた場合に，属性候補集合のそれぞれの要素と評価値候補が対になるか否かの２値分類問題としている．

評価実験は，商品に関するレビューを用いている．このレビューに対し，茶筅と Cabocha を用いて，形態素解析と係り受け解析を行なっている．学習器にはサポートベクターマシーンを用いた．その実験の結果 78.3%の抽出精度を得たと報告している．

2.1.3 感情極性辞書

文書内に存在する感情の発見は，様々な応用可能性を有する重要な課題である．テキストの感情極性分析における重要な資源の一つとして，単語の感情極性をまとめたコーパスがある．単語の感情極性を自動的に抽出し，正確な辞書を構築することが解決すべき課題の一つである． 小林ら [10] は，好意的や非好意的評価に必要な辞書的知識の集合を用いて，これを国語辞典の語釈文からブートストラップ的に獲得する手法を提案している．彼らの研究では，辞書的知識をコーパスからブートストラップ的に獲得することを提案している．彼らの研究では，ブートストラップ的知識獲得として，次の手がかりを利用している．

1. “ 嬉しい ”は好意的,“ 悲しい ”は非好意的などの感情表現

2. 「Xするので，嬉しい」から「Xする」は好意的が推定できるなど，接続関係と好意的または非好意的の推移の関係

3. 「〜することができる」は好意的,「天気が悪い」は非好意的などの辞書的知識

この手がかりを使用して，次のような手順を用いて辞書を作成した．

好意的または非好意的の等式の生成

任意の言語単位の言語表現に対して，好意的か非好意的かの値を返す関数を定義する．その関数を $pn(Exp)p, n, a, e$ とする．p:文脈に関わらず好意的，n:文脈に関わらず非好意的，α:好意的か非好意的かが文脈に依存して決まる，e:好意的でも非好意的でもない．そして,「見出し語の好意的または非好意的の評価値と語釈文の好意

的または非好意的の評価値は原則として一致する」という仮定の基，見出し語と語釈文のペアを生成していく．

好意的または非好意的の等式の書き換え

好意的または非好意的の書き換えは，好意的または非好意的辞書と評価規則を用いて行なっている．

好意的または非好意的仮説の生成

見出し語の好意的または非好意的かが既知の場合，そこから語釈文を構成する言語表現が好意的か否かを推測する．また，見出し語の好意的または非好意的かが未知であり，語釈文の好意的または非好意的かが既知であるならば，語釈文から見出し語の好意的か非好意的かを判断する．

好意的または非好意的辞書への登録

好意的または非好意的辞書に登録される

実験結果により 84.1%の適合率で好意的非好意的な辞書を作成できたと結論づけている．

東山ら [8] は，好意的や非好意的といった知識が感情極性分類には有用な情報になりうると考えた．そして，名詞の評価極性を記述した辞書の構築を行なっている．名詞に限定した点に関しては，形容詞や動詞に比べ，名詞表現の異なりが膨大になるため，人手での作成には多大なコストがかかるためである．彼らの研究では，自態の好意的または非好意的判定にとっての名詞の重要性を調べ，好意的または非好意的に関する述語の選択選好性，教師あり学習を利用して名詞の評価極性を獲得する手法を提案している．彼らは，名詞の評価極性知識の獲得の為にまず次のような基本的アイディアに着目している. 要素の評価極性についてセンシティブな選択選好を持つものが相当数存在する．例えば,“ 心がける ”のヲ格には“ 安全運転 ”や“ 心配り ”のような好意的な好意を示す名詞が出現しやすい．また“ 防ぐ ”のヲ格には“ 火事 ”や“ 事故 ”のような非好意的な出来事を表す名詞が出現しやすい．この性質を利用し，次の手順で実験を行なっている．

1. コーパスから「名詞＋格助詞＋述語」の３つの組みを網羅的に抽出する．
2. １から得られる共起行列から格名詞の共起ベクトルを生成する際に，各名詞について，それと共起する「格助詞＋述語」の組みのうち，名詞との自己相互情報量が０より大きく，設定した閾値以下の頻度となる一般的すぎない述語パターンのみを選択する．
3. 学習を行うにあたって十分な数のポジティブな名詞およびネガティブな名詞を訓練事例とする．

実験方法は，サポートベクターマシンを利用しポジティブとそれ以外，ネガティブとそれ以外に訓練事例を分割し，それぞれについて学習分類を行う．このような手法を用い

て分類した結果，述語の選択選好性という観点に基づく仮定は名詞の評価極性に効果的に働くということを確認している．

2.2 テキスト分類に対する機械学習的アプローチ

機械学習によるアプローチとは，あらかじめ分類済みのトレーニング用文書の集合を用いることにより，カテゴリの性質を学習させ，分類器を自動的に構築する．このプロセスを教師あり学習と言う．機械学習をテキスト分類に適応するためには，次のような点を考慮する必要がある．

1. 各文書を分類するカテゴリの集合を決める
2. 各カテゴリに対しトレーニング用集合が必要である
3. 各文書を表現する素性集合を決める必要がある
4. 分類に用いるアルゴリズムを決める必要がある

2.2.1 確率的分類器

確率的分類器 [3] は，文書 d がカテゴリ c に属する確率 P（c | d）と捉え，この確率をベイズの定理を適応して計算する．

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

$P(c|d)$ を計算するためには，文書 d の構造について何らかの仮定を置く必要がある．文書の素性ベクトルによる表現を d＝（w_1，w_2，$\cdots$）とするとき，素性間の条件付き独立性を仮定すると，$p(c|d)$ は単語ごとの条件付き確率の積に分解され、

$$P(d|c) = \prod_i P(w_i|c)$$

が成り立つ．
この仮定から得られる分類器は，ナイーブベイズ分類器と呼ばれる．

2.2.2 決定木による分類器

決定木分類器は木構造に着目して，内部の接点がある素性に対応し，１つの接点から伸びる枝はその素性の重みに関するテストに対応し，葉は分類のカテゴリに対応している．決定木は木の根を起点にして，その文書によって条件が満たされる枝を順番に葉に達

するまで下りていくことによって文書を分類する．文書は，その葉に対応するカテゴリに分類される．

決定木分類の多くは二値の文書素性を用いるため，木は二分木になる．次に決定木分類器の図を示す.

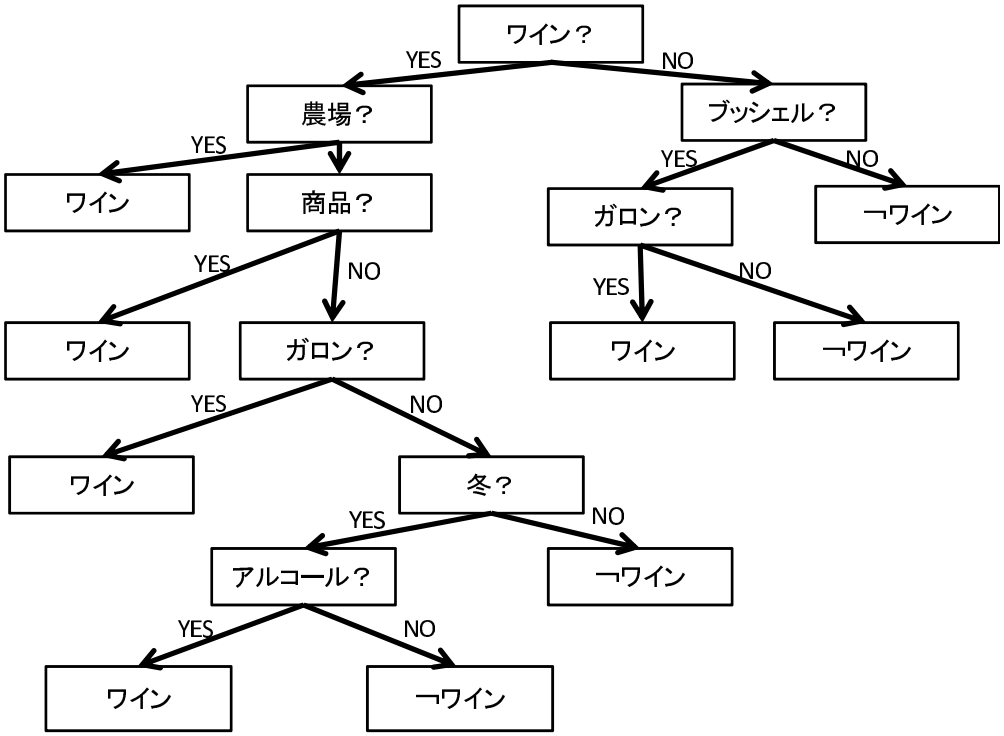

図 2.1: 決定木分類器

一般的には，木は以下のように再帰的に構築される．各ステップで素性 f を一つ選び，トレーニング用集合を２つの部分集合に分割する．１つは f を含むもので，もう１つは f を含まないものである．これを繰り返し，ある部分集合内に１つのカテゴリに属する文書だけが残ったら，その時点で葉を生成する．

2.2.3 サポートベクターマシン

サポートベクターマシンでのクラス分類の目的は，高次元特徴空間でうまく分離する超平面を，計算量的に効率良く学習する方法として提供することである．サポートベクターマシンの最も単純なモデルは，最大マージンクラス分類器は，特徴空間で線形に分離可能なデータにのみ機能する．特徴空間を妥当に選択し，この特徴空間において，最大マージン超平面を発見することがサポートベクターマシンの戦略の一つである．マージンとは，超平面と正例と負例の点集合の中でその超平面に最も近い点との距離である．

そのため，サポートベクターマシンを実装するには，凸最適化問題を利用する，まず，(λw，λb）とスケール変化したとしても，超平面（w，b）は変化しない．よって，線形クラス分類器の定義は，根本的に自由度が内在する．この場合，幾何マージンではなく，関数出力として推定されるマージンが変化する．この関数出力としてのマージンを関数マージンと呼ぶ．幾何マージンを最適化するには，関数マージンを１に固定し，次に重みベクトルの大きさを最小化すればよい．そして，幾何マージンを計算するために w を正規化する．幾何マージン γ は

$$\gamma = \frac{1}{2}\left(\langle \frac{w}{\|w\|_2} \cdot x^+ \rangle - \langle \frac{w}{\|w\|_2} \cdot x^- \rangle\right) = \frac{1}{\|w\|_2}$$

となり，結果としてでてきたクラス分類器の関数マージンとなる．

そして，この問題を対応する双対問題に変換させる．主問題のラグランジアンは

$$L(w, b, \alpha) = \frac{1}{2}\langle w \cdot w \rangle - \sum_{i=1}^{l} \alpha_i \lfloor y_i(\langle w \cdot x_i \rangle + b) - 1 \rfloor$$

なお，$\alpha i > 0$ をラグランジュ乗数とする．

対応する双対問題を得るために，

$$\frac{\partial L(x, b, \alpha)}{\partial w} = w - \sum_{i=1}^{l} y_i \alpha_i x_i = 0$$

$$\frac{\partial L(w, b, \alpha)}{\partial b} = \sum_{i=1}^{l} y_i \alpha_i = 0$$

という関係をえる．その後主問題に再度代入し

$$\begin{aligned} L(w, b, \alpha) &= \frac{1}{2}\langle w \cdot w \rangle - \sum_{i=1}^{l} \alpha \lfloor y_i(\langle w \cdot x_i \rangle + b) - 1 \rfloor \\ &= \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j \langle x_i \cdot x_j \rangle + \sum_{i=1}^{l} \alpha_i \\ &= \sum_{i=1}^{l} \alpha_i - \frac{1}{2} y_i y_j \alpha_i \alpha_j \langle x_i \cdot x_j \rangle \end{aligned}$$

となる．

よって，サポートベクトルは次の図のようになる．

本研究では，サポートベクターマシーンを文章の好意的または非好意的の分類問題に用いる．

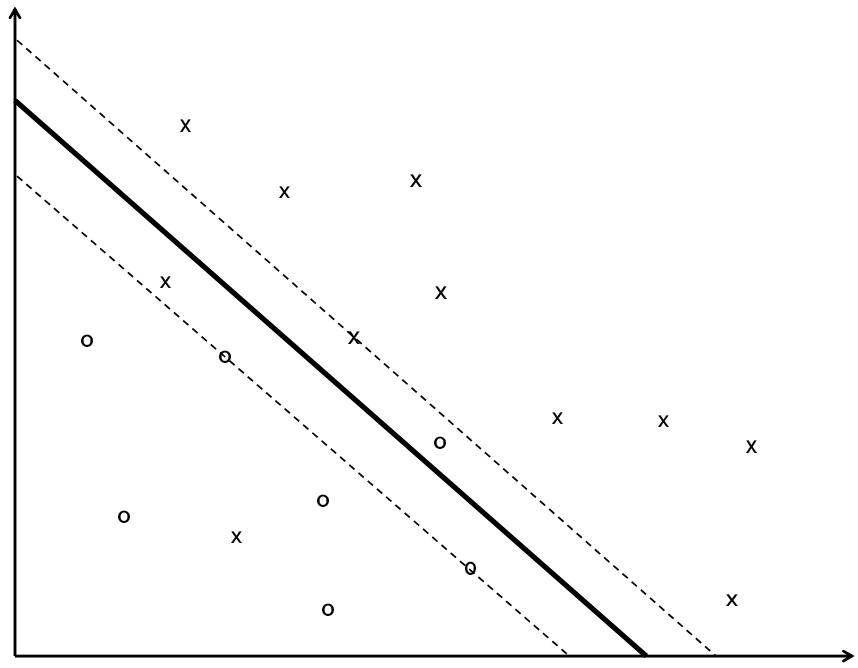

図 2.2: サポートベクターマシーン

2.3 構文構造

構文解析 [13] は，ある解析器が採用した文法理論にしたがって，文の完全な構文解析を行う．おおまかに分類すると，文法理論は句構造文法と依存文法に分けられる．

句構造文法とは，再帰的に構築された句から文の文法的な構造を記述する．句とは，構文的にグループ化された要素を言う．句構造文法においては，名詞句，動詞句，前置詞句，形容詞句，そして節を区別して考える．それぞれの句は，文法規則にしたがって空の句もしくはそれより小さな句や単語から構成される．

一方，依存文法は文の構成要素を独立した言語単位とはみなさず，単語と単語の直接的な関係に焦点を当てている．文の依存構造解析の典型的なものは，単語をノードとして，特定の関係をエッジとするようなラベル付きタグから成る．

また，浅い構文解析がある．これは，解析の深さを追求しない代わりに処理速度と健全さの問題を解決する方法である．浅い構文解析は簡単で曖昧さのない部分的な解析結果だけを生成する．

形態素 [14] とは，文章の形態の階層に関して，最下位の形態を呼ぶ．形態素の分類は，事態素の種別に対応する．文を構成する形態素は，いかなる言語であっても一定の規則に従って配列される．形態素の配列に規則性がある．形態素列として文が事例を表現するのであれば，形態素が意味として表示するのは事態の部分であり，形態素の意味を結合する

ことにより，文が表示する事例が構成されるという仮定が存在する．

自然言語で記述されたテキストに限らず，複雑な事象のデータ処理をして，分析する上では，いくつかの共通する技術がある．具体的には，分類手法，クラスタリング，確率モデル，ルールベースシステムなどである．複雑なシステムの挙動には何らかの確率的なプロセスとしてモデル化するとうまくいくことが非常に多い．確率モデルは，予測値を１つに決定してしまう断定的なモデルよりも多くの場合において良い精度を出すことが多い．それらの確率モデルが自然言語テキストから意味を抽出する様々なタスクで特に有用であることが明らかになってきている．そのなかで確率文脈自由文法というアプローチがある．確率文脈自由文法は，５つの要素からなる G＝（T, N, S, R, P で表現される．ただし，T は終端記号の集合，N は非終端記号の集合，S は非終端開始記号，R は文法規則の集合，$P : R \to \lfloor 0, 1 \rfloor$ はその確率である．文法規則は次の形である

$$n \to s_1 s_2 \cdots s_k$$

ただし，n は非終端記号，s_i はトークンまたは他の非終端記号である．見ての通り確率的文脈自由文法は，通常の文脈自由文法に関数 P を追加したものである．

一般的に，文法の非終端記号は意味のある言語的概念と対応している．文法には，同じ文字列がいくつもの違った規則適応の系列で生成されうるという意味で曖昧さが存在する．例えば“ like apples ”という文字列は，前置詞と名詞という組み合わせという解析も，動詞と名詞の組み合わせという解釈も取ることができる．非確率的文法では，別々の構文解析木を比較する方法がなく，そのためある入力列に対して得られる唯一の情報は文法的か否かのみである．構文解析木とは，構文解析の結果を表した木のような図のことである．これに対して確率的文脈自由文法では，各構文解析はそれぞれ確率を持つため，確率が高いものを選択するという形で選択するという形で曖昧性を解決し，最も良い構文解析を見つけることができる．

2.3.1 チャンキング

浅い構文解析や固有表現認識を行うための基礎的な技術のひとつとしてチャンキングがある．これは図 2.3 に示すように，複数のトークンで構成されたテキスト断片に分割し，ラベル付けすることを言う小さい囲みは単語レベルのトークン化と品詞タグ付けを意味する．大きい囲みはより上位のチャンキングを意味する．それぞれの大きい囲みは，チャンクと呼ばれる．

工藤ら [11] は，チャンキングの段階活用による係り受け解析モデルを提案する．彼らの研究で提案する係り受け解析は機械学習アルゴリズムに依存せず，サポートベクターマシンを用いる．そして，チャンキングの段階活用による構文解析のアルゴリズムは次のようになっている．

係り受け解析は，あらかじめ文節にまとめられ属性付けされた文節列 $b_1, b_2, \cdots, b_m$ を

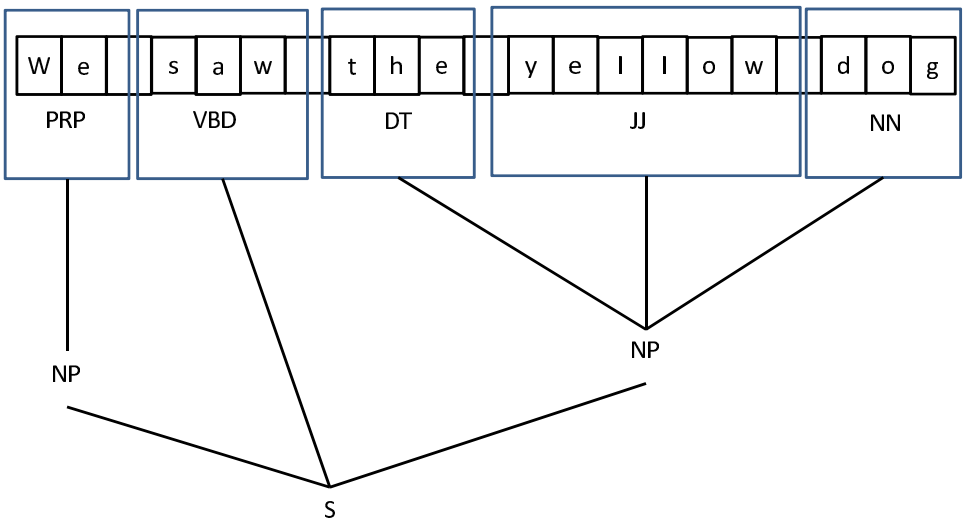

図 2.3: トークンとチャンクの両方における分割とラベル付け

B，係り受けパターン列 $Dep(1), Dep(2), \cdots, Dep(m-1)$ をと定義する．これ以降， D は次の制約を満たすものと仮定する．

1. 文末を除き，各文節はその文節の後方側に必ず１つの掛り先をもつ
2. 係り受け関係は交差しない

統計的係り受け解析とは，上記の制約を満たし，入力された文字列に対する条件付き確率を最大にするパターン列を定義することである．

条件付き確率は，たとえば最大エントロピー法を用いることで計算することができる。Xを観測列の確率変数，Yをラベル列の確率変数とし，Yのすべての要素 Y_i はラベルの有限集合 L を値域と仮定する．変数XとYは同時分布だが，条件付き確率は $p(\mathrm{X})$ を明示的にモデル化することなしに条件付きモデル $p(Y|X)$ を構築する．チャンキングの段階活用による，構文解析のアルゴリズムは次のようになる．

1. 入力文節すべてに対し，係り受けが未定という意味のタグを付与する
2. 文末の文節を除く意味のタグが付与された文節に対し，直後の文末に係るか推定する．係る場合は，削除を意味するタグを付与，後ろから２番目の文節は無条件に削除タグを付与する

3. 意味のタグの直後にある全ての削除タグおよびその文節を削除する

4. 残った文節が一つの場合は，終了，それ以外は２．に戻る

マージンではなく，関数出力として推定されるマージンが変化する．この関数出力としてのマージンを関数マージンと呼ぶ．幾何マージンを最適化するには，関数マージンを１に固定し，次に重みベクトルの大きさを最小化すればよい．実験結果より 89.29%の係り受け正解率で解析モデルができたと報告している．

工藤ら [12] は，半構造化テキストの機械処理，分類において一般には任意のアイテム集合といった素朴な素性を用いることは，テキストが内包する構造を無視していることにほかならない．分類に有効な部分構造が事前に分からない事が多く，人でだけでは有用な素性を取りこぼす可能性がある．理想的な手法は，半構造化テキストから有効な部分構造を自動的に抽出し，学習，分類を行う手法である．この考えに基づき，半構造化テキストのための学習，分類アルゴリズムを提案している．彼らの研究の特徴は次の３つである．

1. テキストの構造を考慮した学習，分類が可能

2. 学習の素性に，構造の部分構造を用いるが，そのサイズに制限を設けず，全部分木を素性候補とする

3. 分析を容易にするために，分類に必要な最小限の部分構造を学習アルゴリズムが自動的に選択する

彼らは，木に対するいくつかの定義，表記を次のようについて言及している．ラベル付き順序木は，すべてのノードに一意のラベルが付与されており，兄弟関係に順序が与えられる木である．単語の並びは，ラベル付き順序木の特殊形であるラベル付き系列でも表記可能である．t と u をラベル付き順序木としている．ある木 t と u がマッチするとは，t と u のノード間に，１対１の写像関数が定義できる．ラベル付き順序木の分類問題は，L 個のトレーニングデータ $\mathrm{T} = \{\langle x_i, y_i \rangle\}_{i=1}^{L}$ から，分類関数 f（x）： $\mathrm{X} \rightarrow \{\pm 1\}$ を導出することである．通常の学習，分類問題と異なる点は，事例 x_i が，数値ベクトルで表現されるのではなく，ラベル付き順序木として表現される点である．　適応なルールを発見するために，彼らは高速なアルゴリズムを提案している．
提案手法は，以下の３つから構成される．

(1) 木の集合から，部分分木を完全に重複なく列挙するための正準的な検索空間を定義する

(2) （１）で定義した検索空間を深さ優先探索し，最大のゲインを与える部分木を発見する

(3) （１）,（２）だけでは，全部分木をしらみつぶしに列挙していることと変わりない．そこでゲインの上限値を見積もり，探索空間を枝刈りする

評価実験は，２種類のタスクで行なっている．

ＰＨＳレビュー分類

ＰＨＳユーザに，良い点または悪い点を区別して投稿するように指示した掲示板のデータ

文のモダリティ分類

被験者１名が，９９年毎日新聞の社説からランダムに選んだ６０記事に，分単位のモダリティ付与したデータ

その結果，提案手法の分類精度は，従来の手法に比べおよそ４００倍高速になったと報告されている．

第3章　係り受け情報による学習データの構築

本研究の手法は，感情極性分類に構文構造の係り受け情報を利用する．構文構造の係り受け情報とは，単語と単語の直接的な関係に着目した文の構造である．

従来の研究では，文書中に存在する単語の有無のみに着目することで文書が保有する情報を抽出していた．その情報を用いて感情極性分類が行われていた．しかし Bag-of-words 手法に依存している場合，各単語の出現の有無のみを考慮するため，句全体の感情極性に着目することは困難である．Bag-of-word 手法とは，各単語が素性空間における一つの次元であると仮定する手法である．また，素性空間とは，内部構造を持たない単純な素性の集合空間である．

感情単語を用いた分類の研究は，語彙ネットワークを利用する方法や文脈一貫性を用いることで感情単語を抽出する．その抽出した単語を活用することで，感情極性分類が行われている．しかし，係り受け構造まで考慮した研究はない．そこで，本研究は，構文構造の係り受け情報に着目することで文書の感情極性分類の精度向上を目指す．

本研究で構文構造の係り受け情報を用いるために，Cabocha を用いて構文解析を行う．Cabocha とは，サポートベクターマシーンに基づく日本語係り受け解析器である．これを用いることで，構文構造から単語と単語の関係が抽出可能である．それらの関係の中から，着目する単語として係り受け先の単語が日本語感情極性辞書に存在する単語に限定する．日本語感情極性辞書とは，約 8 千 5 百表現の評価極性を持つ名詞に対して評価極性情報を付与し，人手によるチェック済みのデータである．これらの単語と，係り受け元として接続詞の中で逆説の意味を持つものがある場合に感情極性が反対になる可能性があるとする．例えば，文書全体が好意的な文として書かれており，その一文に“ あの絵は汚いけど珍しい ”という文があったとす.“ 汚い ”は非好意的な単語であるが，文書全体としては好意的に書かれているので，非好意的な単語は全体の感情極性と異なる．そこで“ 汚い ”と“ けど ”で係り受け構造に着目することで，逆説の係り受け先に存在する感情単語が全体の感情極性と異なる極性を持つことが分かる．

この特徴量を，サポートベクターマシーンの学習データとして活用することで感情極性をより正確に捉えることができると考える．そして，学習モデルを用いることにより，各文書の感情極性の逆転の有無が分かる．感情極性単語の極性が逆転したことにより，文書に存在する感情極性の単語数が変動する．これにより，正確な感情極性の単語数を調べることができる．これらの結果を，感情極性分類に用いることで，より正確な文書分類が可

能になると考える．

3.1　手法概要

本章では本研究の手法について述べる．感情極性分類に必要な素性ベクトルを文書から獲得し，ポジティブまたはネガティブかを示す文章なのかをテキスト分類によって獲得する．

本研究の特徴として，構文構造から係り受け情報を抽出し，学習データに有用に用いることである．係り受け情報に着目する理由は，各構造に存在する単語が接続詞や否定などによって，個々の単語の保持する感情極性が反対になる場合が多くあり，そのため正確な感情極性を知ることは簡単には分からない．そこで各構造での感情極性に着目することが重要であるからである本研究で提案する手法として，次の４種類の特徴ベクトルを抽出する操作を施す.

手順１ １つの単語で独立した素性ベクトルの抽出

手順２ N-gram によって分けた単語での素性ベクトルの抽出

手順３感情単語の素性ベクトルの抽出

提案手法　係り受けを考慮した素性ベクトルの抽出

3.2　手順１　１つの単語で独立した特徴ベクトルの抽出

文章中の全単語を個々の素性ベクトルとして抽出する．１つの文章中に存在する個々の単語を単独のベクトルとする．
個々の単語を素性ベクトルとし，一つ一つの文章を一つの特徴量とする．そうすることにより，各文章の内容を定量化することと同義であると考える．
特徴量を表現するために，次のステップを行う．

Step1 わかち書きされたデータを作成する

Step2 全データから単語辞書を作成する

Step3 各データにおける単語が作成した辞書に存在するか調べ取り出す

Step1 として，わかち書きされたデータを作成する．そのために，茶筅を用いて行う．茶筅とは，入力文を単語単位に分割し品詞を付与する形態素解析器である．わかち書きとは，語句の区切りに空白を挟んで記述する書き方である．日本語は基本的に語で区切りをせず書かれていため，個々の単語の素性ベクトルを抽出しにくい．わかち書きをすることで，単語の素性ベクトルが抽出しやすくなる．

次に，全使用データから全単語を含む文字列辞書を作成する．次に，各データから作成した辞書に含まれている単語の素性ベクトルを抽出する．
素性ベクトルの抽出方法を図 3.1 に示す．

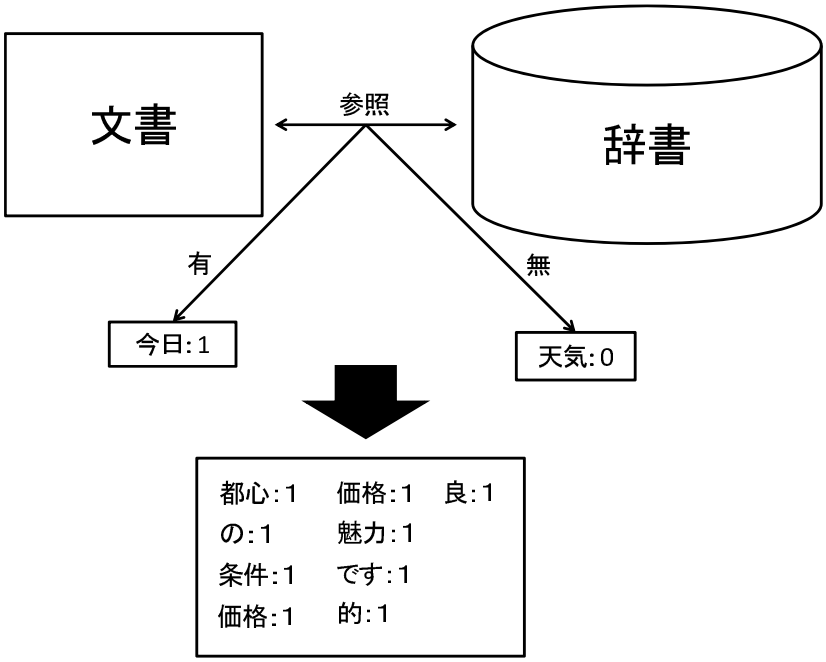

図 3.1: 素性ベクトルの抽出

3.3 手順2 N-gram によって分けられた単語での特徴ベクトルの抽出

手順１を使用しても，各文章の内容を捉えることは難しい．それは，日本語の文章中には，複数語によって意味を持つ語彙が多数存在する．例えば“ 魅力的 ”などがある．これは，Unigram のままでは“ 魅力 ”と“ 的 ”に分けられてしまう．そのため，操作１で抽出した単語では意味を示すものではない文字列になって抽出されている可能性がある．
そこで，単語を抽出方法として N-gram を用いる．N-gram とは，隣り合った文字列または単語の組み合わせを作成する．１単語を元に抽出する方法を unigram という．これは手順１と同じである．２単語の並びを抽出する方法を bigram，３文字の並びを抽出する方法を trigram という．N-gram で抽出することで，文章中に存在する複数語をカバーすることが可能になると考える．unigram と bigram や trigram を組み合わせることにより，文章中に存在する語を網羅できる可能性が高くなると考える. そこで，対象とする文章に対して，N-gram の辞書を作成し，対象の文章から素性ベクトルを抽出することで，文書

分類の精度が良くなる可能性がある．

本研究で使用する N-gram の例を図 3.2 に示す．

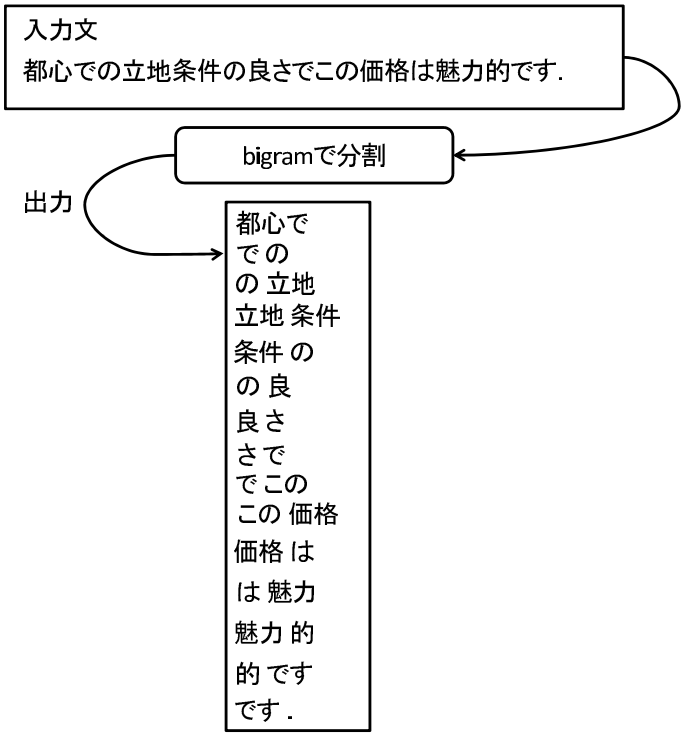

図 3.2: バイグラムでの抽出例

3.4 手順 3 感情単語においての特徴ベクトルの抽出

文章中に存在する感情単語辞書を用いた特徴量の抽出．感情極性分類をする際に，ある特定の単語に対する評価極性を用いることで分類精度につながる．本研究における感情単語の設定は，日本語評価極性辞書を使用する．この辞書は，東山らが作成した評価極性を持つ名詞，約 8 千 5 百表現に対して評価極性情報を付与し，人手によるチェック済みのデータである．辞書の中身の一例を図 3.3 に示す．

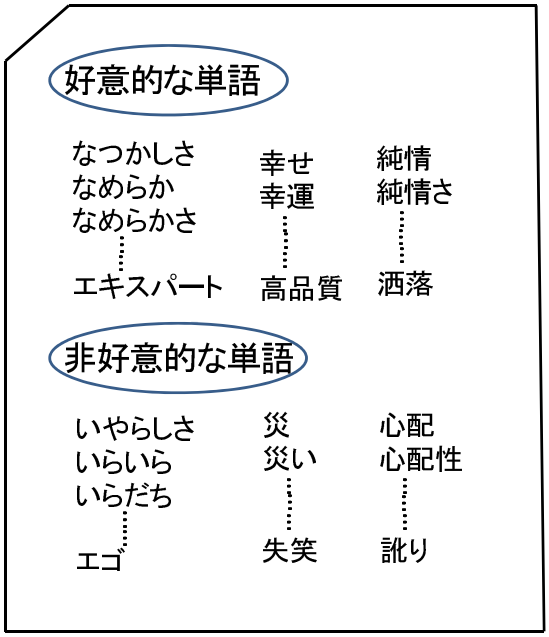

図 3.3: 日本語評価極性辞書

そこで，対象とする文章の中に，日本語評価極性辞書に含まれる名詞の有無を調べる．文章中に含まれる好意的単語または非好意的単語の総数を数え上げる．機械学習をする際に，両方の単語の出現回数を用いることで，その文章が好意的であるか，非好意的であるかを判断することが可能である．一つの文章中に，好意的単語と非好意的単語のどちらか片方だけが出現することは極めて稀である．本研究で使用する文章を用いて，図 3.4 に 2 種類の感情極性単語の出現割合を示す．この結果から，否定的な文書でも好意的な単語が使われる割合が高い．それは，人の手によってレビューを書くときは，否定的な文書においても，否定的な意味をもつ単語は使わず肯定的な意味をもつ単語を使うことで，否定的な文書を構築していると考える．

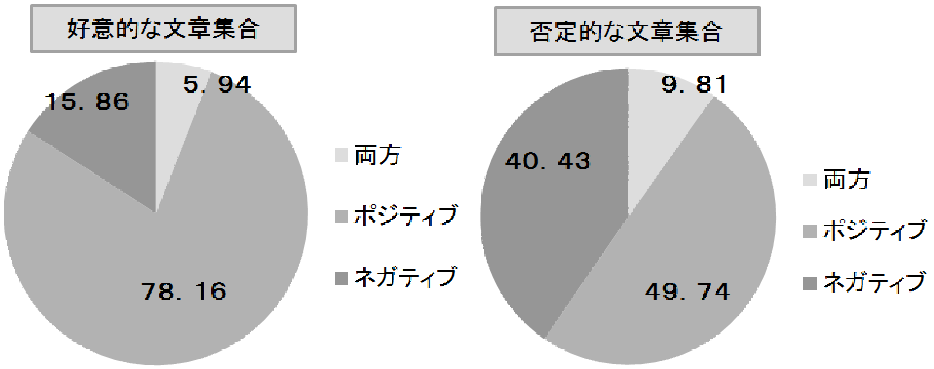

図 3.4: 係り受け情報

3.5　提案手法　係り受けを考慮した特徴ベクトルの抽出

文書を感情極性分類する際に，手順１～３を使用して獲得したベクトルを用いることで，精度が向上する期待は出来る．しかし，人が文章を書く上で，感情極性単語が保持している感情極性が，文章中で変化することがある．変化する要因の１つとして，文章中の構文構造に着目する．構文構造の単語同士の係り受け関係によって，感情極性が入れ替わることがある．例えば,“ お気に入りのホテルだが今回は部屋が非常に暑かった． ”という文がある．この文書を書いた本人は，全体として非好意的な文章として書かれている．“ お気に入り ”という単語は，単語自体の極性は好意的な極性を持つ．しかし，接続詞の“ だが ”が“ お気に入り ”に係ることにより,“ お気に入り ”という単語が非好意的な極性を持つ単語として作用すれば，文章の極性がより非好意的な極性に偏るので，分類精度の向上につながると考える．この逆転の考え方を図 3.5 に示す．

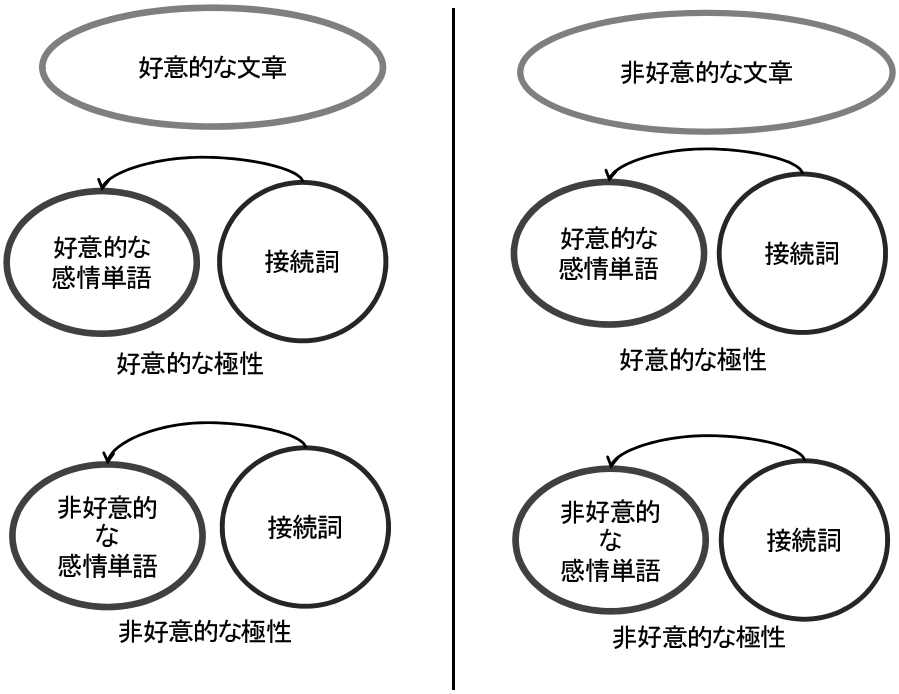

図 3.5: 逆転

これらの例が示すように，文章中には感情極性単語と接続詞などの関係によって，感情極性の変化が起こることが度々ある．そのため，分類精度の向上を目指す要因の１つとして，これらの感情極性の変化を捉えることは必要な要素と言える．　そこで，本研究は，構文構造による係り受け関係と感情極性単語の関係を考慮した学習モデルを提案する．接続詞や否定語が感情極性単語と係り受け関係にある場合，その感情極性単語の持つ感情極性と文章の意図として使われる感情極性が異なるという点に着目し，その変化を考慮した学習を行う．この学習を用いて得られた結果が，実際の各文章で使われている感情極性を示していると考える．これを用いて，文章の持つ感情極性単語の総数を変化させることで，より正確に分類することが可能であると考える．

始めに，係り受け関係を考慮した学習モデルを作成する．学習モデルを作成するために必要な構文構造は，Ｃａｂｏｃｈａを使用する．Ｃａｂｏｃｈａはサポートベクターマシーンを基礎とした日本語構文解析器である．文書の１文１文にＣａｂｏｃｈａを用いることで構文構造を取り出す．この構文構造の情報を用いて，係り受け関係の情報を素性ベクトルの形として抽出する．例を図 3.6 に示す．

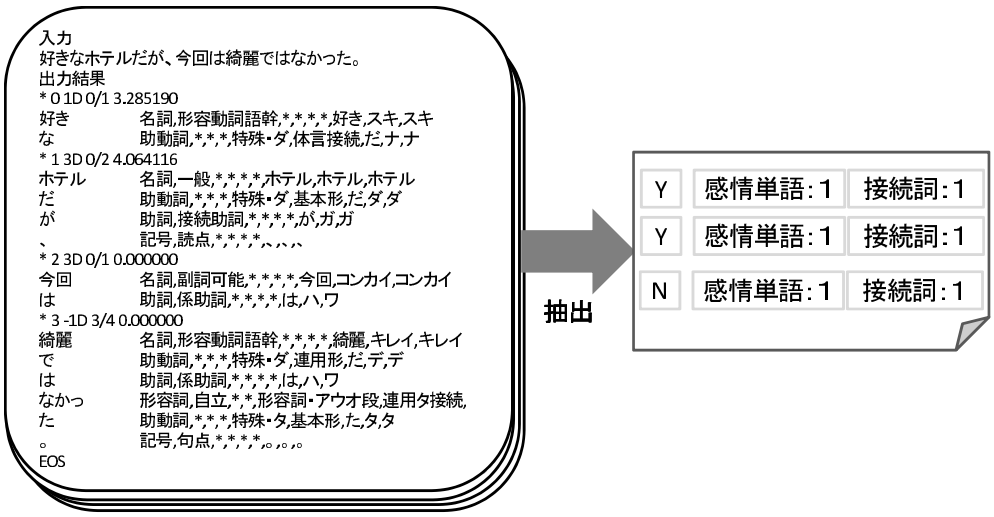

図 3.6: 係り受け情報

この素性ベクトル集合を用いて，サポートベクターマシンにより学習させることで，感情極性の逆転の有無を学習したモデルを構築する．この学習モデルを，感情極性分類を行う文書の１文１文をテストデータと見なすことで，各文で感情単語の意味と極性が反対になったかが確認することが出来る．そして，各文書の感情極性単語の意味と極性が逆転した総数を調べる．この総数を使用し，手順３で得た感情極性単語の総数に変化を与える．例えば，好意的な文章のなかで，非好意的な感情極性単語が係り受け関係により反対になっていることが出力された場合，好意的な感情極性単語の総数は増加され，非好意的な単語の総数は減少する．これを，感情極性分類を行う文書全てに用いる．これによって得られた素性ベクトルの結果を，サポートベクターマシーンを用いて学習する．この学習結果を使用し，文書の感情極性分類を行う．

提案手法は実際には次の手順を行う．

1. Ｍｅｃａｂを用いて，使用する文書データを単語ごとにわかち書きする
2. １の結果の文書を，１文１文に分ける
3. １文１文に対して，Ｃａｂｏｃｈａを用いて構文解析を行う
4. 構文解析の結果から，係り受け関係の情報を抽出する

5. 抽出したデータを用いてサポートベクターマシーンで学習する
6. 感情極性分類を行うデータも同様にＣａｂｏｃｈａを用いて構文解析を行う
7. ７の情報から，係り受け関係の情報を抽出する
8. ８で作成したデータを，６で作成した学習モデルを使用する
9. その結果から得られた，各文章単位での逆転した総数を数える
10. 逆転した総数の結果を用いて，手順３で得た感情極性単語の総数に影響を与える
11. 影響を受けたデータを用いてサポートベクターマシーンで学習をする
12. テストデータを使用して，分類を行う

手順の概要を次の図に示す.

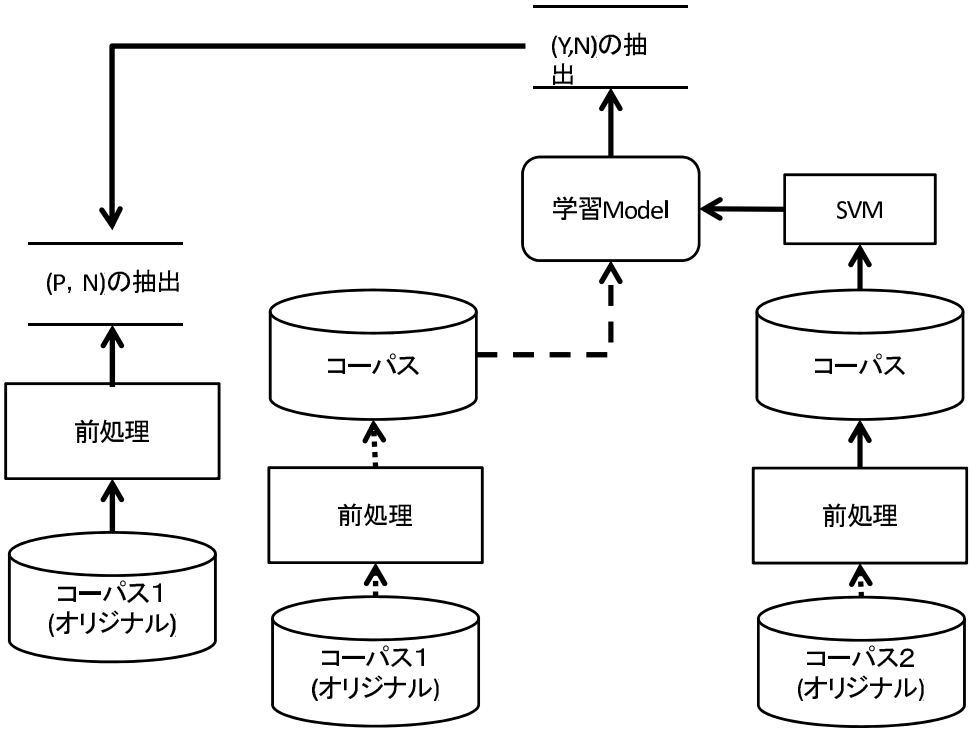

図 3.7: 素性ベクトルの抽出

第4章　実験

今回，提案手法の有用性を検証するため，実験を行った．

1. 語彙の獲得
2. 係り受け情報の獲得
3. 感情極性分類

この実験の詳細を次に記す．

4.1　語彙の獲得

感情極性単語の辞書を用意する．本研究では，日本語評価極性辞書 [15] を参照して辞書を用意した．日本語評価極性辞書は用言を中心に収集した評価表現約 5 千件のリストを一部改編し，人手で評価極性情報を付与したデータと，評価極性を持つ名詞，約 8 千 5 百表現に対して評価極性情報を付与した，人手によるチェック済みのデータである．本研究では，日本語評価極性辞書の名詞のみを使用している．今回の実験では，出来るだけ多くの感情極性単語の数を調べるため，すべての日本語極辞書の名詞を使用している．　素性ベクトルの表現方法を順に述べる．

1. 手順１, ２を用いた結果例

+1	241:1	2437:1	2438:1	...	2440:1
+1	147:1	152:1	221:1	...	910:1
+1	912:1	2440:1	2441:1	...	9551:1
+1	2440:1	2441:1	4303:1	...	4573:1
+1	2441:1	3269:1	18589:1	...	18597:1
+1	2441:1	6458:1	5655:1	...	18598:1
+1	1388:1	1555:1	2440:1	...	2441:1
+1	147:1	1103:1	1555:1	...	2440:1
+1	147:1	2440:1	4558:1	...	4560:1
-1	1607:1	2440:1	2441:1	...	4573:1
-1	2440:1	2441:1	3863:1	...	4231:1
-1	2461:1	8590:1	5144:1		
-1	1649:1	2435:1	2436:1	...	2440:1
-1	1715:1	2440:1	2441:1	...	3764:1
-1	1555:1	2319:1	2440:1	...	2441:1
-1	270:1	1472:1	1777:1	...	1791:1
-1	512:1	742:1	827:1	...	2440:1

感情極性　　単語集合

図 4.1: 素性ベクトルの抽出例１

2. 手順３を用いた結果例

+1	912:1	2440:1	2441:1	...	9551:1	617019:1	620370:2
+1	2440:1	2441:1	4303:1	...	4573:1	617019:0	620370:0
+1	2441:1	3269:1	18589:1	...	18597:1	617019:1	620370:1
+1	2441:1	6458:1	5655:1	...	18598:1	617019:0	620370:0
+1	1388:1	1555:1	2440:1	...	2441:1	617019:0	620370:6
+1	147:1	1103:1	1555:1	...	2440:1	617019:1	620370:0
+1	147:1	2440:1	4558:1	...	4560:1	617019:0	620370:0
-1	1607:1	2440:1	2441:1	...	4573:1	617019:0	620370:0
-1	2440:1	2441:1	3863:1	...	4231:1	617019:0	620370:1
-1	2461:1	8590:1	5144:1			617019:1	620370:2
-1	1649:1	2435:1	2436:1	...	2440:1	617019:0	620370:2
-1	1555:1	2319:1	2440:1	...	2441:1	617019:1	620370:1
-1	270:1	1472:1	1777:1	...	1791:1	617019:0	620370:0
-1	512:1	742:1	827:1	...	2440:1	617019:1	620370:0

感情極性　　単語集合　　感情極性単語総数

図 4.2: 素性ベクトルの抽出例２

3. 提案手法を用いた結果例

+1	912:1	2440:1	2441:1	...	9551:1	617019:3	620370:0
+1	2440:1	2441:1	4303:1	...	4573:1	617019:0	620370:0
+1	2441:1	3269:1	18589:1	...	18597:1	617019:1	620370:0
+1	2441:1	6458:1	5655:1	...	18598:1	617019:1	620370:0
+1	1388:1	1555:1	2440:1	...	2441:1	617019:1	620370:0
+1	147:1	1103:1	1555:1	...	2440:1	617019:1	620370:0
+1	147:1	2440:1	4558:1	...	4560:1	617019:0	620370:0
-1	1607:1	2440:1	2441:1	...	4573:1	617019:0	620370:0
-1	2440:1	2441:1	3863:1	...	4231:1	617019:0	620370:1
-1	2461:1	8590:1	5144:1			617019:0	620370:3
-1	1649:1	2435:1	2436:1	...	2440:1	617019:0	620370:2
-1	1555:1	2319:1	2440:1	...	2441:1	617019:0	620370:2
-1	270:1	1472:1	1777:1	...	1791:1	617019:0	620370:0
-1	512:1	742:1	827:1	...	2440:1	617019:0	620370:1

感情極性

単語集合

意味と極性の反転回数を考慮した感情極性単語総数

図 4.3: 素性ベクトルの抽出例３

4.2 実験の概要

各手順によって素性ベクトルを自動抽出するプログラムを実装した．また，係り受け解析を用いた逆転の有無の回数と感情極性総数の差分を自動的に計算し，素性ベクトルを抽出するプログラムを実装した． コーパスは，2010 年に公開された楽天トラベルの内，ユーザ評価及びユーザレビューを使用した．全レビューの内，最大 5,3 万件をトレーニングデータに，1,4 万件をテストデータに使用した．なお，２種類のデータは重複がなく，ランダムに取得した．また，構文解析に用いるデータとして，2 万件を使用した．そして，ユーザー評価を用いて，ユーザーレビューを星１から星３の文書と星４から星６の文書に分類した．データの内容の一部を図 4.3 と 4.4 に示す．

ウェルカムコーヒーはとてもよかったです。ロビーに行くたびにエスプレッソをいただきました。部屋も加湿器、空気清浄機が装備されていて快適でした。とてもよい出張となりました。次回も利用させていただきたいと思います。

図 4.4: 好意的なレビューの一例

ビジネスパックは要注意、夕食は無いので外で食べてから泊まるべし、土曜・日曜以外はホテル内のレストランは休みです。あまり早くホテルに入ると、外からの入浴客であふれてゆっくり入浴出来ないので、注意、サウナに入りたい場合は夜９時で終了なので、逆に早くチェックインして入浴を済ませてから夕食のために外出しなければいけません。浴室のロッカーは有料（１０円）です。２名以上なら食事を出してくれるので良いのですが、１人のビジネスパックはオプションで別途夕食を注文しても準備してもらえません。朝の入浴も６時３０分からなので要注意、夜は１１時で終了なので、外食する場合は飲み過ぎに注意、早めに帰らないと入浴できません。

図 4.5: 非好意的なレビューの一例

また，本研究では，素性ベクトルの組み合わせとして９種類を実験に使用する．この９種類を用いて，文書の感情極性分類を試みた．性能指標として，再現（recall），適合率（precision）とＦ値である．あるカテゴリに対する再現率は，そのカテゴリに本来属するすべての文書に対する，正しく分類された文書の割合として定義される．適合率は，分類器によってそのカテゴリに分類されたすべての文書に対する，正しく分類された文書の割合である．Ｆ値と呼ばれる値は，$\frac{2 \cdot RecallPrecision}{Recall+Precision}$ で定義され，再現率と適合率の調和平均として計算される．

4.3 実験結果

実験結果を，手順１，手順２，手順３，提案手法の順に述べる．

4.3.1 手法１の結果と考察

コーパスには楽天ユーザーレビューを使用した．トレーニングデータは，それぞれ 5.3 万件，2.7 万件，1.35 万件，0.7 万件，0.35 万件をランダムに取得した．また，テストデータは 1,4 万件をトレーニングデータと重複無くランダムに取得した． 図 4.5 と表 4.1 に，Unigram を抽出したデータだけを学習させた場合の実験結果を示す．精度は訓練事例の数を増加していくと精度は向上していく．これは，コーパスに出現する単語の総数が増加したため，分類精度が向上したと考えられる．

表 4.1: unigram を用いた実験結果

データ数	3500	7000	13500	27000	53000
Precision	49.77	49.87	49.77	50.01	50.12
Recall	96.42	96.62	96.85	97.83	98.62
F 値	65.65	65.79	65.75	66.19	66.46

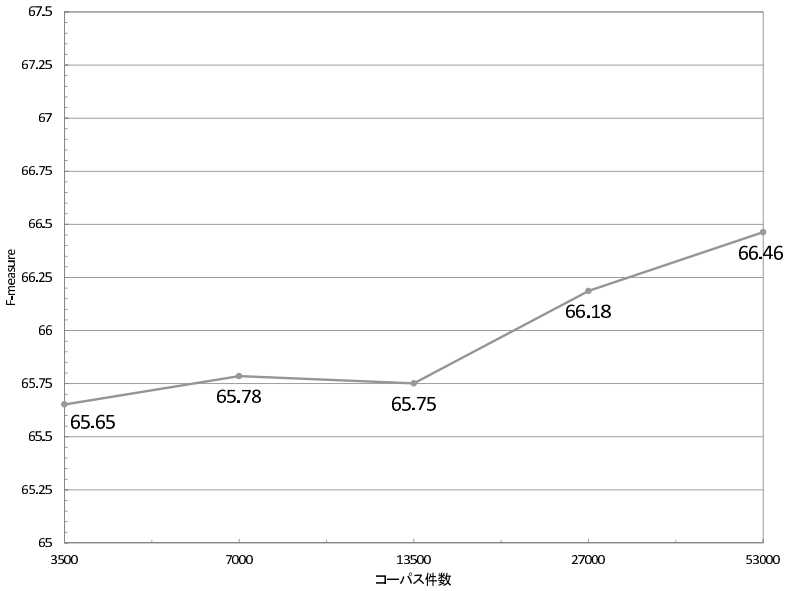

図 4.6: unigram を用いたデータセット

4.3.2 手順２の結果と考察

次に，手順２によって感情極性分類を行なった．コーパスには楽天ユーザーレビューを使用した．トレーニングデータは，それぞれ 5.3 万件，2.7 万件，1.35 万件，0.7 万件，0.35 万件をランダムに取得した．また，テストデータは 1.4 万件をトレーニングデータと重複無くランダムに取得した． 図 4.6 と表 4.2 に，bigram を抽出したデータだけを学習させた場合の実験結果を示す．精度は訓練事例の数を増加していくと精度は増加するものの，トレーニングデータの数に比べ精度の向上度合いは少なかった．図 4.7 と表 4.3 に，trigram を抽出したデータだけを学習させた場合の実験結果を示す．精度は訓練事例の数に比例して増加していく，しかし，bigram と同様に向上度合いは少なかった．これらは，unigram に比べると高い精度を示した．

表 4.2: bigram を用いた実験結果

データ数	3500	7000	13500	27000	53000
Precision	50.35	50.43	50.86	50.87	50.63
Recall	99.38	99.25	97.61	97.83	98.96
F 値	66.83	66.87	66.87	66.93	66.98

表 4.3: trigram を用いた実験結果

データ数	3500	7000	13500	27000	53000
Precision	50.40	50.28	50.30	50.28	50.26
Recall	98.25	98.80	99.28	99.47	99.55
F 値	66.62	66.64	66.77	66.79	66.79

4.3.3 手順３の結果と考察

次に，手順３によって感情極性分類を行なった．コーパスには楽天ユーザーレビューを使用した．トレーニングデータは，それぞれ 5.3 万件，2.7 万件，1.35 万件，0.7 万件，0.35 万件をランダムに取得した．また，テストデータは 1.4 万件をトレーニングデータと重複無くランダムに取得した．図 4.8 と表 4.4 に，unigram と感情極性単語総数を抽出したデータを学習させた場合の実験結果を示す．精度は，unigram のみを抽出したデータを学習させた場合に比べて，若干下がるものとなった．これは，unigram で抽出した素性ベクトルが多いため，感情極性単語総数が効果的に作用しなかったためと考えられる．図 4.9 と表 4.5 に，bigram と感情極性単語総数を抽出したデータを学習させた場合の実験結果を示す．精度は，bigram のみを抽出したデータを学習させた場合に比べ，精度は向上

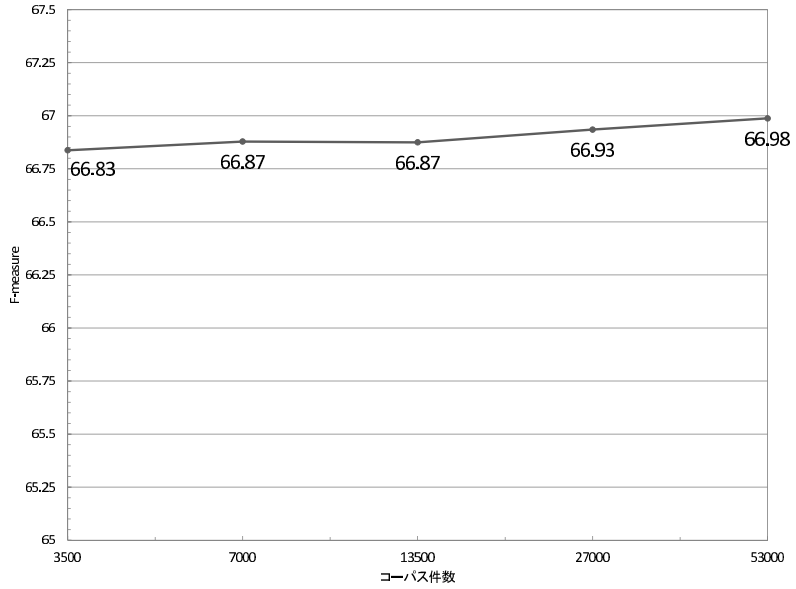

図 4.7: bigram を用いたデータセット

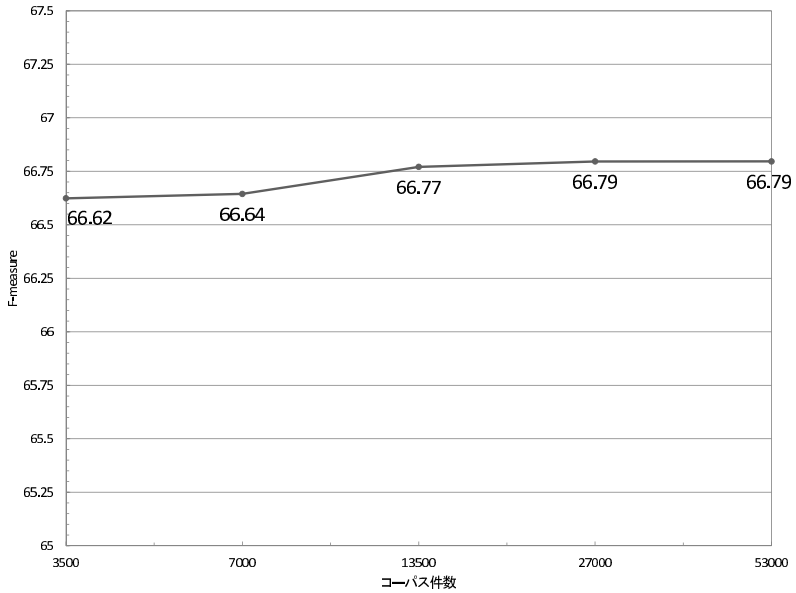

図 4.8: trigram を用いたデータセット

した．図 4.10 と表 4.6 に，trigram と感情極性単語総数を抽出したデータを学習させた場合の実験結果を示す．精度は，trigram のみを抽出したデータとあまり変わらなかった．

表 4.4: unigram と感情極性単語を用いた実験結果

データ数	3500	7000	13500	27000	53000
Precision	49.77	49.87	49.77	49.95	50.12
Recall	96.78	96.65	96.78	97.71	98.55
F 値	65.73	65.79	65.73	66.10	66.44

表 4.5: bigram と感情極性単語を用いた実験結果

データ数	3500	7000	13500	27000	53000
Precision	50.35	50.43	50.66	50.86	51.44
Recall	99.38	99.13	98.97	97.76	96.58
F 値	66.83	66.85	67.01	66.90	67.12

表 4.6: trigram と感情極性単語を用いた実験結果

データ数	3500	7000	13500	27000	53000
Precision	50.34	50.39	50.3	50.29	50.27
Recall	98.63	98.79	99.28	99.46	99.55
F 値	66.65	66.73	66.77	66.80	66.80

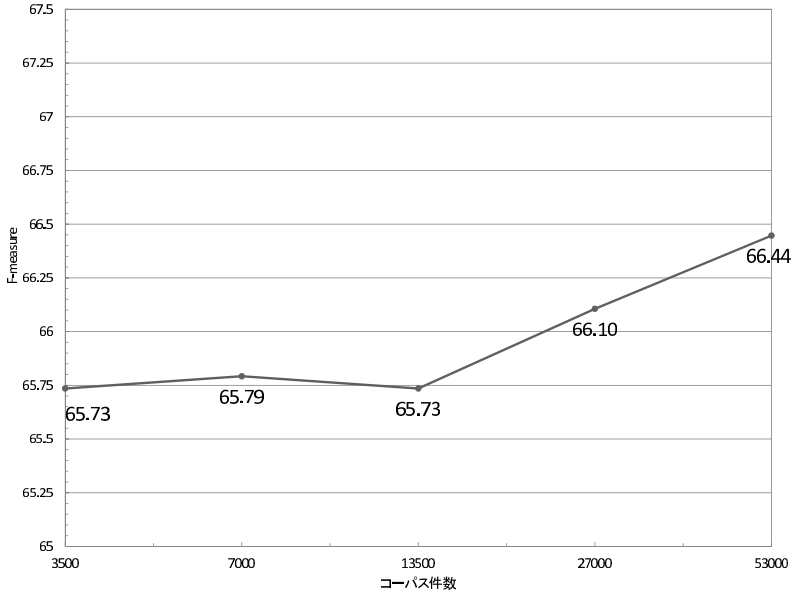

図 4.9: unigram と感情極性単語を用いたデータセット

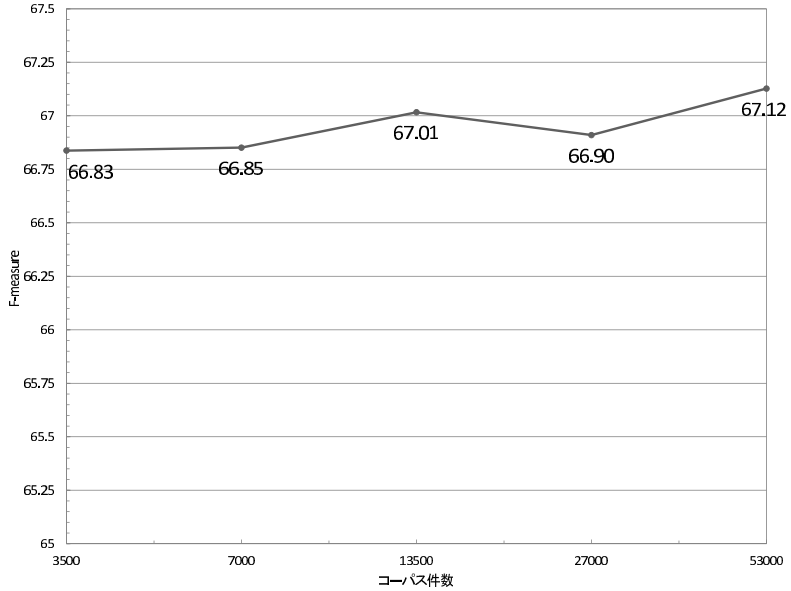

図 4.10: bigram と感情極性単語を用いたデータセット

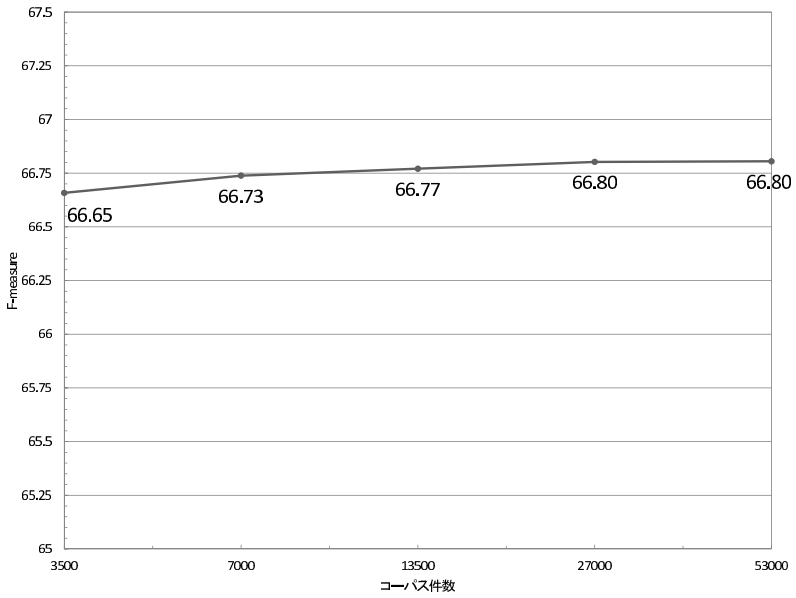

図 4.11: trigram と感情極性単語を用いたデータセット

4.3.4 提案手法の結果と考察

次に，提案手法を用いて感情極性分類を行なった．コーパスには楽天ユーザーレビューを使用した．コーパスには楽天ユーザーレビューを使用した．トレーニングデータは，それぞれ5.3万件，2.7万件，1.35万件，0.7万件，0.35万件をランダムに取得した．また，テストデータは1.4万件をトレーニングデータと重複無くランダムに取得した．図4.11と表4.7に，unigramと提案手法を用いて抽出したデータを学習させた場合の実験結果を示す．精度は，他のunigramと比べて高いものとなった．図4.12と表4.8に，bigramと提案手法を用いて抽出したデータを学習させた場合の実験結果を示す．精度は，他の実験に比べて一番高いものになった．図4.13と表4.9に，trigramと提案手法を用いて抽出したデータを学習させた場合の実験結果を示す．精度は，他のtrigramと比べ高いものとなった．

表 4.7: unigram と提案手法を用いた実験結果

データ数	3500	7000	13500	27000	53000
Precision	50.92	51.13	51.32	51.83	52.02
Recall	93.17	95.03	94.93	95.12	95.25
F 値	65.85	66.48	66.62	67.09	67.29

表 4.8: bigram と提案手法を用いた実験結果

データ数	3500	7000	13500	27000	53000
Precision	51.83	51.92	52.03	52.22	52.64
Recall	96.83	97.10	96.92	97.21	97.34
F 値	67.51	67.66	67.71	67.94	68.32

表 4.9: trigram と提案手法を用いた実験結果

データ数	3500	7000	13500	27000	53000
Precision	51.2	51.32	51.41	51.61	51.92
Recall	98.45	98.64	98.57	98.73	98.81
F 値	67.36	67.51	67.57	67.78	68.07

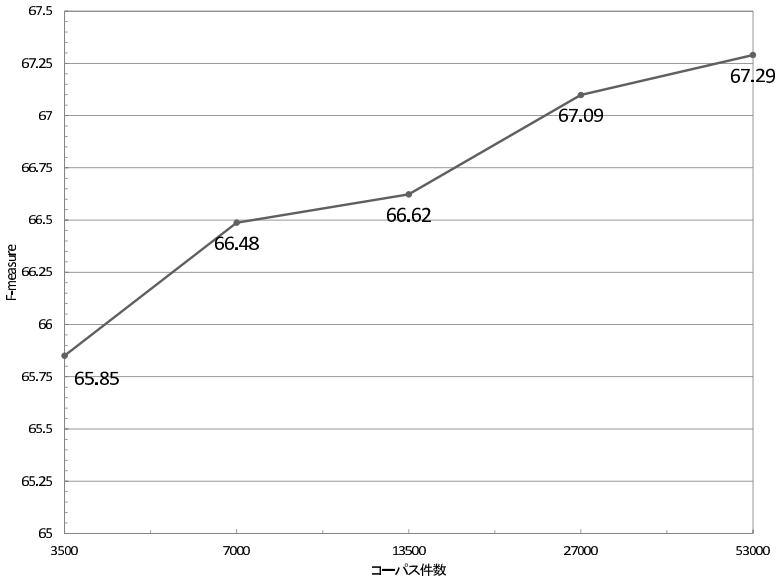

図 4.12: unigram と提案手法を用いた実験結果

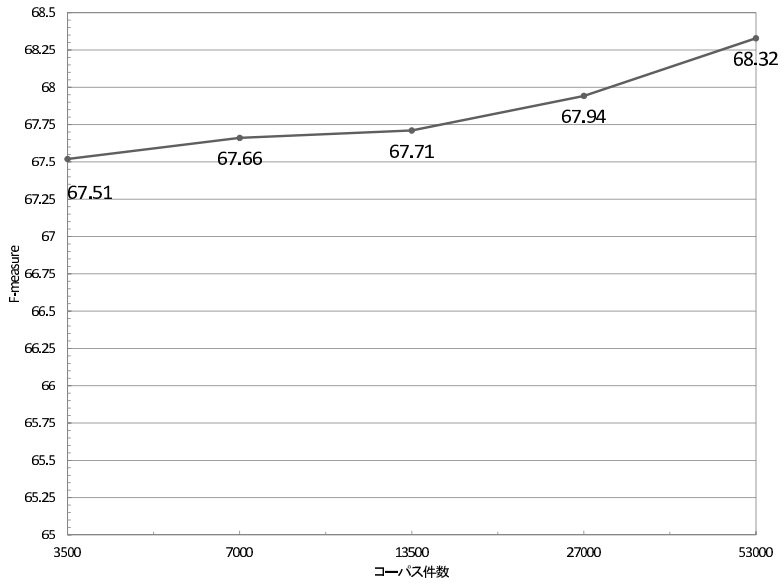

図 4.13: bigram と提案手法を用いた実験結果

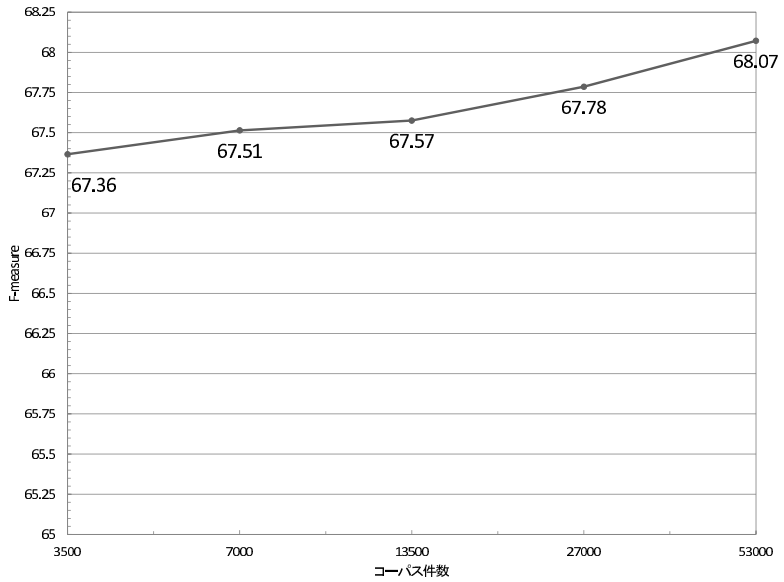

図 4.14: trigram と提案手法を用いた実験結果

4.3.5 係り受けによる単語の意味と極性の逆転回数

提案手法での構文構造から係り受け情報を取得し，サポートベクターマシーンを用いて学習モデルを構築した．その学習モデルを使用して，文書の感情極性分類に用いるコーパスの逆転回数を求めた．この回数を元に，実際の感情極性の総数を求めている．コーパスの数の差はあるが，ネガティブデータの逆転の比率が多い．これは，ネガティブとして書かれた文書には，感情極性単語の意味と極性が異なって使う傾向があると考えられる．逆転回数を表４．１０に示す．

表 4.10: 学習モデルを用いた逆転結果

データ数	１回	２回	３回	４回
ポジティブデータ (45000 データ)	2047	73	8	0
ネガティブデータ (8000 データ)	692	21	2	0

4.3.6 実験結果の比較

図 4.15 に，各実験結果ごとでＦ値が一番高い結果を示す．ただし，手順１と手順２は同様の手順と見なし，その中で一番良い精度の結果を示している．縦軸は感情極性分類の精度を，横軸は学習に使用した事例の数を示している．意味と極性が異なった回数を用いることで，好意的な感情極性の総数と非好意的な感情極性の総数の偏りが，手順３に比べ明確になるため，自然な結果であると思われる．この結果から，係り受け関係から感情極性総数の決定は適切に働いていることが分かる．

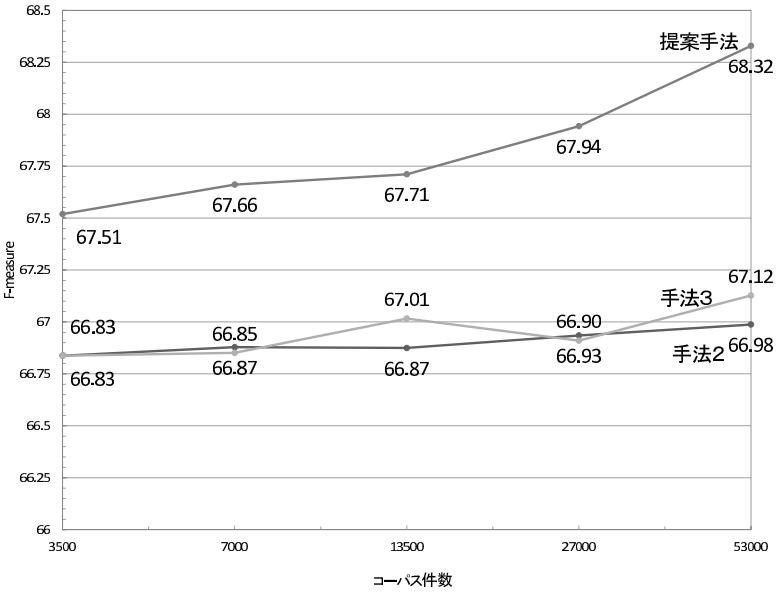

図 4.15: 各実験の比較結果

第5章　おわりに

本論文では，文書の感情極性分類のタスクにおいて，構文構造を特徴量として捉えることで，感情極性分類の分類精度を向上させる手法を提案した．本章では，本研究で得た知見と今後の課題について述べる．

5.1　本研究のまとめ

提案手法を以下にまとめる．

1. 素性ベクトルの作成
   (a) N-gram モデルを用いて，各文書の素性ベクトルを抽出を行なった．
   (b) 日本語評価極性辞書を使用し，各文書に存在する感情極性単語の総数を抽出を行なった．
   (c) 構文構造から係り受け情報を抽出を行なった．日本語評価極性辞書の感情極性単語と，それ以外の単語との係り受け関係を素性ベクトルとし，サポートベクターマシーンを用いて学習モデルを作成した．
   (d) 作成した学習モデルを用いて，各文章ごとの感情極性単語の意味と極性の反転回数を調べた．その回数を，感情極性単語の総数に反映させた．その結果を素性ベクトルとした．

実験の結果， Precision，Recall，F 値の評価基準において，各手法結果を比較したところ，提案手法と Bigram の組み合わせで素性ベクトルを作成したものが元も高い値であった．その原因を調査したところ，感情極性単語の意味と極性が反対の回数を用いて感情極性単語の総数を変動させた結果，好意的文章の好意的な感情極性単語の出現回数が増加し，非好意的な単語の回数が減少した．非好意的文章では，正反対のことが起きていた．つまり，好意的文章での好意的な感情極性単語の出現頻度が高くなり，非好意的文章では非好意的な感情極性単語の出現頻度が高くなった．これらの状態によって，分類精度が向上していた．また，感情極性単語の総数と提案手法を施した感情極性単語の総数では，文書と異なった極性の単語の出現回数が０回になりやすい．この点から，提案手法は感情極性分類の前処理として有用であると考えられる．

5.2 今後の課題

本研究では構文構造から情報を抽出することに重点を置いており，構文構造から有効であると思われる素性ベクトルを生成することができた．しかし，本研究では感情極性単語の発見というタスクには取り組んでおらず，提案する手法では新しい感情極性単語の発見ができない．本研究においても，感情極性単語の総数が好意的，非好意的ともに 0 というデータが少なくなかった．したがって，今後の課題としては，未知の感情極性単語を対象とする文書から発見することができるか，また新感情極性単語の感情極性を正確に付与できるかどうかを確認する必要がある．小林ら [10] や東山ら [8] の研究のように未知の感情極性単語かどうかを判断する方法を考案することが感情極性分類の精度向上における課題であると考えられる．また，飯田 [6] らは機械学習を用いて各語彙の感情極性の抽出を行なっており，この手法をあらかじめ対象の文章に用いることで，人手によって選ばれた感情極性単語を用いることをせずに，感情極性分類が行えると考える．

本研究では素性ベクトルを Bag-of-word モデルを用いて表現する際に，２値の重みを表現方法としている．しかし，素性ベクトルの表現方法は改善の余地があると考えられる．その中でも，訓練データにおける出現頻度の低い語彙の対応があげられる．これは対象の文書の表現を半教師学習の利用で対応できる可能性がある．

本研究でのテキストの感情極性分類精度をみると，感情極性単語と極性の逆転を学習させ，その結果を活用した場合の分類精度が一番向上している．その理由として，その係り受け関係と感情極性単語の学習結果を感情極性分類の感情単語総数に影響を与えていることが大きいと考えられる．しかし，本研究では，前処理の段階で数値を変更させる等の手間をかけている．その作業量を減らすためにも，係り受け関係と感情極性単語の関係を一つの特徴量として扱うことで，より扱いやすい特徴量と見なせる可能性がある．

本研究で使用したコーパスは，好意的文章の比率と非好意的文章の比率が非常に悪い．このため，非好意的文章の特徴量が正確に学習されていない可能性が捨て切れない．よって，ある程度大きなコーパスを使用する際における，分類させたいクラスのトレーニングデータの比率を調整させる必要があると考えられる．

また，本研究では構文構造の一部に着目することで，テキストの感情極性分類を行なった．しかし，文書全体の構文構造を考慮する方法を提案することで，より正確な感情極性単語の意味と極性の逆転などの関係を調べることが出来る可能性がある．

今回は感情極性分類を好意的または非好意的の２値に分類した．しかし，実際のレビューでは，どちらでもないという観点から書かれた文章が存在する可能性がある．どちらでもないという観点から書かれた文書には，感情極性単語が存在しない可能性もある．よって，より感情極性単語の影響に着目するためには，より細かい分類で実験を行なってみる必要性も考えられる．

謝辞

本研究を進めるにあたり，日頃から方針，内容について懇切丁寧にご指導下さいました鶴岡慶雅准教授に厚くお礼申し上げます．研究全般にわたり多くのご意見を下さいました東条敏教授に深く感謝申し上げます．また，東条研究室の皆様には，本研究に関する貴重なご支援をいただきました．この場を借りて感謝申しあげます．

参考文献

[1] Bo Pang, Lillian Lee and, Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceeding of the Conference on Empirical Methods in Natural Language Processing, pp76-86, 2002.

[2] Bo Pang, Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceeding of the 42th Annual Meeting of the Asscociation for Computational Linguistics, pp115-124, 2004.

[3] Domingos,P., Pazzani,M. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Journal of Machine Learning, Vol29, pp103-130,1997

[4] Peter D. Turne. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp417-424, 2002

[5] Bo Pang, Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp115-124, 2005

[6] 飯田 龍, 小林 のぞみ, 乾 健太郎, 松本 裕治, 立石 健二, 福島 俊一. 意見抽出を目的とした機械学習による属性-評価値対同定 情報処理学会自然言語処理研究会予稿集, pp.21-28. 2005

[7] 高村 大也, 乾 孝司, 奥村 学. 隠れ変数モデルによる複数語表現の感情極性分類 情報処理学会論文誌, Vol.47, pp3021–3031, 2006

[8] 東山 昌彦, 乾 健太郎, 松本 裕治. 述語の選択選好性に着目した名詞評価極性の獲得, 言語処理学会第 14 回年次大会論文集, pp584-587, 2008.

[9] 松本 翔太郎, 高村 大也, 奥村 学. 単語の系列および依存木を用いた評価文書の自動分類第 3 回情報科学技術フォーラム, pp213–214, 2004.

[10] 小林 のぞみ, 乾 孝司, 乾 健太郎. 語釈文を利用した「p/n 辞書」の作成 人工知能学会 言語音声理解と対話研究会, pp45-50, 2001.

[11] 工藤 拓, 松本 裕治. チャンキングの段階活用による係り受け解析, 情報処理学会, pp97-104. 2001

[12] 工藤 拓, 松本 裕治. 半構造化テキストの分類のためのブースティングアルゴリズム 情報処理学会研究報告, 知能と複雑系研究会, pp163-168, 2004.

[13] ローネン. フェルドマン, ジェイムズ. サンガー, (IBM 東京基礎研究所テキストマイニングハンドブック翻訳チーム 訳), テキストマイニングハンドブック, 東京電気大学出版局, 2010

[14] 町田 健. 言語構造論基礎論 勁草書房, 2011

[15] 日本語評価極性辞書http://cl.naist.jp/ inui/research/EM/sentiment-lexicon.html

[16] 乾孝司, 奥村学, テキストを対象とした評価情報の分析に関する研究動向 自然言語処理 Vol.13, Num.3, pp201-241