情報工学コース卒業研究報告

議事録集合からの特徴語抽出とその応用に関する研究

杉浦　広和

名古屋大学 工学部 電気電子・情報工学科

2009年2月

概要

これまでに経験のない仕事をする場合、その仕事において必要となる知識を持っていないことが多い。そのようなとき、必要な情報を効率よく取得するためには、まず何が必要な情報であるかを知ることが重要である。この問題は、ある分野における辞書を作成することにより解決することができる。辞書を作る際には見出し語や索引語を決めなければならないが、その作業を人手で行うことは労力が大きい。そこで、できるだけ人手による作業を少なくし、見出し語や索引語を抽出できる手法が必要となる。

そこで、そういった語の抽出を機械的に行う手法を考える。語の抽出は筆者の所属する研究室で行われるゼミを記録した議事録集合を対象として行う。見出し語となるような特徴的な語（特徴語）にはある共通の性質があると考えられる。たとえば、主語や目的語になりやすかったり、発表の資料に出現しやすい傾向にあると考えられる。本研究では、特徴語を抽出するために特徴語となり得るような語（特徴語候補）の選出と Support Vector Machine（SVM）と呼ばれる識別器を用いた分類を行う。識別器による分類を行うためには、一つ一つのデータに対してベクトルを付与する必要がある。このベクトルを特徴ベクトルと呼び、特徴ベクトルには先に挙げた特徴語の性質、主語や目的語になる、資料に出現する、といった異なる観点からなる特徴を定量化し、ベクトルの要素として設定する。特徴語抽出は次のような流れにより行う。まず議事録集合に含まれるテキストを形態素解析によって形態素に分割する。次に、複合語などの連接する形態素を連結し、特徴語の元となる語の集合を生成する。さらにいくつかの制約に基づいてフィルタリングし、特徴語候補を決定する。この特徴語候補のそれぞれに特徴ベクトルを設定し、SVM による分類の結果、特徴語の集合を抽出する。本論文では、その特徴ベクトルの設定と SVM による分類とその評価、及び抽出された特徴語の応用について述べる。分類の評価は、分類により抽出された語中に含まれる、実際の特徴語の割合により評価する。本研究における特徴語は、議事録に出現している語のため、時間情報と密接に結びついている。その情報を利用することで、話題の推移状況を知ることができる。

目 次

第1章　はじめに

今や世の中には大量の情報が溢れかえっている。人口の増加と通信手段の発達に伴い、一日あたりにやり取りされる情報量は近年、増加の一途をたどっている。

実際に、米国 VeriSign[1] の調査リポートでは、2008 年第 1 四半期でのドメイン登録数が全世界で 1 億 6,200 万件であった。これは、前年同期比で 26 %増、前の四半期（2007 年 10 月-12 月期）に比べ 6 %増に当たるとのことだった。1 つのドメインに複数のコンテンツが含まれることは自然であり、Web 上から情報を取得するためには 1 億 6,200 万以上のコンテンツを相手にしなければならないのである。そういった情報の洪水に飲まれないためには、必要な情報を取捨選択し、効率的な情報収集に努めなければいけない。

ユーザが情報を収集しようと考えたとき、いったい何をするのだろうか。先ほども述べたように、Web 上には大量の情報があふれている。これらすべてに目を通し、必要かどうかを判断することは現実的には不可能である。したがって、その時ユーザは検索という手段を取り、必要な情報に関係する Web ページを探し出すのである。検索により得られた Web ページは、ランク付けもされているため、ユーザは順に目を通していき、望みの情報を手に入れることができるのである。一般的に検索はクエリによって行われる。すなわち、ユーザは自らが調べたいことを理解し、調べたい事柄を検索のクエリとして設定しなければいけないのである。検索を行うことは、情報を選別することであり、必要な情報へと素早くたどりつくことに役立つ。しかし欠点として、ユーザが必要とする情報にたどり着くためのクエリを知っていなければいけない点がある。

多くの場合、ユーザは必要な情報を検索するためのクエリを知っている。しかし、研究室や会社の異動など、活動の場が変わった場合はその限りではない。知らなければいけないことがあるが、何を知ればいいのかがわからないという状態になるのだ。そうなると、検索により情報を集めることができなくなってしまう。なぜなら、検索のためのクエリを設定することができないからである。検索ができないとなると、人に聞くか、大量の情報を前に一つずつ要不要を確認して見ていかなければならなくなる。人に聞くことは、ある意味では最も確実で迅速な手段ではある。しかし、その情報が記憶に頼ったものである以上、忘れたり、思い込みによる齟齬があると考えられる。こういったときに、整理されたマニュアルの存在や、必要な知識が既にまとめられてあれば何も問題はない。ユーザは知っておかなければいけないと感じる部分に目を通せばよいのである。だが、それら

の作成にはコストがかかり、また、真に必要な情報であるかを事前に知ることが困難である。たとえば、携帯電話の説明書などは、近年の多機能化により、非常にページ数が多くなっている。しかし、ユーザが説明書の情報をすべて必要としているのは稀で、多くのユーザは自分に必要な部分、あるいは有用であると判断する部分だけを読むのである。

ユーザは大量の情報を前に、何かを知りたいという欲求がある。しかし、情報を集めるための検索を行うためのクエリを知らないとする。そうした場合、ユーザの欲求を叶えるためには、どのようなサポートが必要であるのか。当然のことながら、ユーザが知りたいと考えているのは、未知の情報についてである。では、未知の情報かどうかは誰が判断できるのか。無論それはユーザ自身に他ならない。ある情報がユーザにとって未知の情報であるかどうかを機械が判断することはできない。したがって、システムとしてサポートできるのは、ユーザに対して、どういった種類の情報があるのかを提示することである。そのためには、コンテンツの集合をいくつかの小さい集合に分類することが効果的である。ある情報がどのような内容であるかは、その情報中において特徴的な語を知ることで類推することができる。たとえば以下のような文書があったとする。

「関係者が集まり、討論・相談や決議をする会議という場において、スムーズに議論が進むということ、さらに、過去の議事内容を検索などの手段により参照して、再利用できるということは非常に有益である。「議事録」という形で議論内容を保存しておかないと、過去の議論は容易に忘れ去られ、何度も同じ議論を繰り返す恐れがある。さらに、過去の議論内容を参照したい場合も、あやふやな記憶から間違った内容を新たな議論の場に出してしまうと考えられる。これは、現在進行中の会議においても同様のことが言える。議論の途中で、現在までの議論の内容を確認できるということも同様に有益である。今までの議論を一度振り返ることで、人間の創造性を刺激することができる。しかしながら、オンラインの掲示板などによる会議とは異なり、オフラインの会議においては議論を再利用可能な形にするのは一般に困難である。」

この文書が何を意味しているのか、一見しただけで把握することは難しい。しかし、この文書中には「会議」「議事録」「オンライン」「オフライン」「再利用可能」などの、特徴的な語が含まれている。これらの語を見ることにより、この文書が「議事録」に関係していたり、「再利用可能」という話題についての話をしていることが類推できる。このように特徴的な語を抽出し、それらをまとめた辞書を作成することができれば、ユーザがこの辞書を調べることで、必要な情報にたどり着くことができると考えられる。

このようにして得られた特徴語は、分類を行う上でも役立つ。たとえば図1.1のようなコンテンツ集合を分類することを考える。

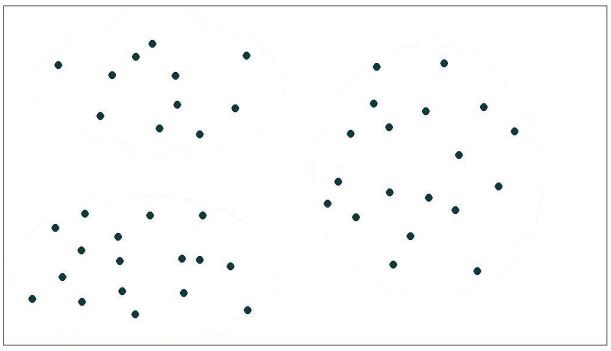

図 1.1: コンテンツ集合の二次元配置

図 1.1 は、二次元空間にコンテンツを配置した図であり、各点はひとつのコンテンツを表している。空間上の距離はコンテンツ間の類似度を表し、近いものほど類似しているとする。手がかりを用いない空間上の分布によるクラスタリングでは、距離の近いものを同じクラスとして分類するので、図 1.2 のようにクラス分けされる。

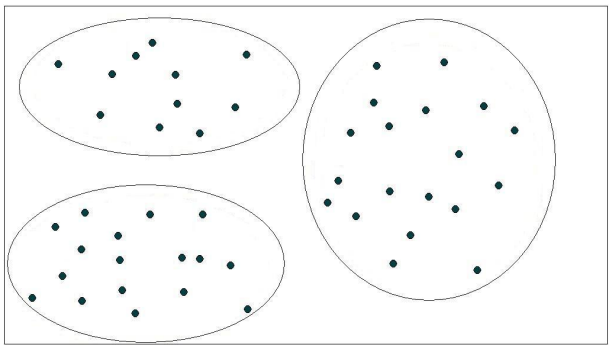

図 1.2: 距離による分類

図 1.2 のクラス分けは一見正しいように見える。しかし、いくつかの問題点がある。一つは、この方法では、一つのコンテンツは一つのクラスにしか分類されないことである。しかし、ひとつのコンテンツが持つ情報は必ずしも一つとは限らない。そのため、さまざまな観点での分類ができなければ、コンテンツの分類手法としては効果的と言えない。もう一つは、クラスタリングによる分類が正しいかどうかをユーザが判断することが難しい点である。「類似している」という考え方は、必ずしも推移的とは限らない。すなわち、A と B が似ていいるからと言って、B と C が似ているとは限らない。そのため、類似しているものを集めたクラ

スが、本当に正しいかを確かめるためには、クラス中のすべてのコンテンツ間の類似度を見なければいけない。

このように、何の手がかりも用いずに、コンテンツ同士を比較するやり方では分類に効果的とは言えない。これに対して、特徴語による分類を行った場合を考える。

特徴語を利用した分類では、どの特徴語に着目するかにより分類結果が変わる。そして一つのコンテンツは複数の特徴語を含むことができる。そのため、特徴語ごとのクラス、例えば「議事録」クラスや「再利用可能」クラスなど、を作れば一つのコンテンツが複数のクラスに分類されることになる。また、その特徴語で分類するべきかどうかの判断は、そのコンテンツと特徴語に関連があるか見ればよい。これは、そのクラス中のすべてのコンテンツ間の類似度を見るよりは手間が少なくて済む。

このように、特徴語を抽出することができれば、情報の収集に大きく役立つ。そこで、機械的な特徴語抽出の手法に関して考えてみる。機械的な処理がもっとも容易なコンテンツは、テキストである。画像や動画などでは、語を抽出するためには特別な処理が必要となるからである。そこで、本研究ではテキストから特徴語を抽出する。

筆者の所属する研究室では、毎週行われているゼミ中の発言を記録した議事録が存在する。この議事録は、ゼミ中の発言を書記が記録したもので、キータイプによりリアルタイムに計算機へと保存されている。発言の際には、発言者が専用の機器を用いて、発言の開始と終了を宣言する。そのため、発言間の明確な区切りが存在する。この一つの発言を一つのテキストとして考え、本研究ではこのテキスト中から特徴語抽出を行う。

特徴語を抽出するためには、特徴語に共通する性質を手掛かりとして用いる。特徴語は話題の中心となることが多く、発言の中では主語や目的語として使われるのではないかと考えられる。それは前後の語の品詞として助詞が出現しやすいと言い換えることができる。

また、筆者の所属する研究室のゼミでは、必ずPowerPointによるスライドを資料として用いる。このスライド中には、発表に関する説明などが書かれているため、特徴語が多く出現することが考えられる。これは、特徴語であるならばスライドに多く出現するということでもある。

このような特徴語の性質を用いて、特徴語の抽出を行う。そこで、まず特徴語になりそうな語を特徴語候補として選び出し、その候補を二つに分類することを行う。候補の中から、前述の特徴語としての性質を持っている語、持っていない語の二つに分類し、特徴語の性質を持っている語を特徴語として抽出する。このような性質による分類を行うために、各性質を定量化し特徴ベクトルというものとして考える。

この特徴ベクトルには、前後の品詞による要素と、スライド中への出現数とい

う要素の二つの観点からなる情報が含まれている。このように、全く別の観点からなる情報を組み合わせることで、特徴語としての性質がより顕著に表れるのではないかと期待される。特徴ベクトルを用いた分類には、Support Vector Machine（SVM）[2] と呼ばれる識別器を用いる。SVM は高次元の特徴ベクトルを分類するのに適した識別器である。この SVM により、特徴語候補を特徴語と非特徴語に分類することで、特徴語抽出を行う。

では、そもそも特徴語とは何であろうか。特徴語と対になる言葉は一般語である。したがって、特徴語であるということは一般語ではないということである。そして、一般語とはその言葉通り、一般に使用され特別な説明を必要としない語である。ある分野において特徴語であるということは、その分野に関係の浅い人間に対して説明を要するということである。このことから、ある分野における特徴語とは、その分野における辞書の見出し語に相当するものと考えることができる。

本研究の目的は、議事録のテキストから特徴語候補を選出し、特徴ベクトルを用いることにより特徴語を抽出することである。このように議事録の発言テキストから特徴語を抽出することには大きな意義がある。それは、議論の最中に特徴的な語を入力することができないからである。一般に、何かを記録しながら議論を行うことは難しいとされている。したがって、その発表や発表中に行われた発言の特徴語を設定するのは、発表後となる。しかし、発表後に議事録を見直し、特徴語を選択し設定することはあまり効率的とは言えない。例えば、発表終了後に、各自自分の発言を見直し特徴語を設定することを考えると、発表中に多く発言するほど後の手間が増えることになる。このことにより、発言に対する負担が増えることは望ましくない。よって、発表後に機械的な手法により議事録から特徴語を抽出する必要がある。

また、議事録の発表はいくつかのプロジェクトに分けることができる。筆者の所属する研究室では、複数のプロジェクトが活動を行っているためである。各プロジェクトは全くの無関係ではないが、扱う対象が大きく異なり、それぞれのプロジェクトごとに特徴語となる語が違う。そこで本研究では、特徴語候補には各プロジェクトごとに別々の特徴ベクトルを設定し、そのプロジェクトごとに分類を行い特徴語を抽出する。

本論文では、議事録からの特徴語抽出の手法とその応用について述べる。本論文の構成は、第二章で特徴語候補の定義や特徴語の抽出方法について説明する。第三章では、実際に行った特徴語抽出とその応用について述べる。第四章では、議事録から得られた特徴語の評価と考察について述べる。第五章では、いくつかの関連研究について述べる。第六章では、まとめと今後の課題について述べる。

第2章　特徴ベクトルを利用した特徴語抽出

2.1　特徴語抽出の流れ

一般的な特徴語抽出にはTF-IDFが用いられる。TF-IDFは文書に対する語の重要度を数値化したものである。その算出式は以下のとおりである。

$$\text{TF-IDF} = tf \cdot idf$$

tf とはTerm Frequencyの略であり、その語の出現頻度を表す。一般的にはその語の文書中への出現回数を用いる。idf はInverse Document Frequencyの略で逆出現頻度と呼ばれ、一般には次の式を用いる。

$$idf = \log\left(\frac{D}{d}\right)$$

Dは全文書数を表し、dは語の出現する文書数を表す。TF-IDFが大きいほど、その文書に対してその語が強く結び付いていると考えられ、特徴語であると言える。しかし、TF-IDFによる手法では、文書から得られる特徴語の数が、文書量に大きく依存してしまう。長い文書であればそれだけ多くの特徴語が抽出でき、短い文書なら少ない特徴語しか抽出できない。だが、文書量と特徴語の数は必ずしも比例の関係であるとは限らない。短い文書に多くの話題が含まれている場合は特徴語の数は多くなるし、長い文書でも、同じことを繰り返し論じている場合では特徴語の数は少なくなる。そこで、文書単位ではなく、文書集合全体の中での語の特徴により特徴語を抽出する手法を考える。

そのために、ある語がどのように使われているのかということに着目した。特徴語となるような語は、ある文書中においての話題となるようなものが多い。それは、主語や目的語といったものとして出現しやすいということである。主語となる場合は、その語の後ろに「は」「が」といった助詞が来ることになり、目的語となる場合はその前に「を」「に」といった助詞が来ることになる。

また、本研究で対象としている議事録に記録されている発表では、資料としてPowerPointのスライドを利用している。特徴語となるような語は、こういった資料中に頻繁に出現することが予想される。スライド中に載せることのできる情報

には限りがあるため、発表者は重要と思われる情報を選出し、資料に載せるからである。したがって、資料中への出現回数によっても、特徴語であるかどうかの判断ができると考えられる。

そこで、このような特徴語が持つと考えられる性質を定量化し、特徴語抽出に利用する。本研究では、定量化した値をまとめて、ひとつのベクトルとして考える。このベクトルを、語の特徴を表すベクトルと考え、特徴ベクトルと呼ぶ。この特徴ベクトルを利用し、特徴語抽出を行う。特徴語を抽出するということは、特徴語とそうでない語の二つに分類することと同じである。本研究では、この分類のために Support Vector Machine（SVM）[2] と呼ばれる識別器を利用する。

また、こういった特徴語抽出を行う際には、まず文書中から特徴語の候補となる語を抽出する必要がある。そのためには文書を形態素解析し、語を取り出す必要がある。

本研究で提案する特徴語抽出は、図 2.1 のように行う。

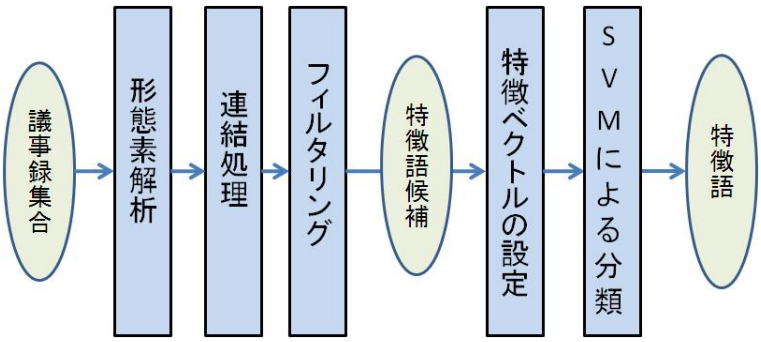

図 2.1: 特徴語抽出の流れ

本研究では議事録の集合を対象としている。その中に含まれるテキストを形態素解析する。分割された形態素は、複合語として連結できるものもあるため、これらを連結する。こうして特徴語候補の元となる語の集合から、いくつかの制約によるフィルタリングを行い、特徴語候補を抽出する。抽出された特徴語候補のそれぞれ特徴ベクトルを設定し、SVM による分類を行い、特徴語を抽出する。

以下では、特徴語候補の抽出と特徴語候補からの特徴語抽出とに分けて説明をする。

2.2 特徴語候補の抽出

2.2.1 形態素解析

日本語文は、英文と違い単語間の明確な区切りが存在しない。そのため、文書中から語を取り出すためには、形態素解析と呼ばれる処理を施さなければいけない。

形態素解析とは、テキストを形態素と呼ばれる単位に分割することを指す。この形態素は、厳密にいえば単語とは違った分割の単位ではあるが、おおよそ単語と同じようなものになる。そのため、形態素は品詞の情報を持つ。例えば以下のような文を形態素解析した場合、表 2.1 のような形態素に分割される。

「形態素解析により文書を分割する」

表 2.1: 形態素解析の結果例

形態素	名詞-一般
解析	名詞-サ変接続
により	助詞-格助詞-連語
文書	名詞-一般
を	助詞-格助詞-一般
分割	名詞-サ変接続
する	動詞-自立
。	記号-句点

このように、形態素解析による文書解析により、日本語文の中から語を取り出すことができる。

形態素解析のプログラムとして、Juman[3] や ChaSen[4] といったものがある。表 2.1 の形態素解析の結果は、ChaSen による解析の結果である。Juman と ChaSen では、形態素の品詞体系が異なるため、解析の結果が大きく変わる。本研究では、品詞の情報を多く用いるため、より詳細な品詞の分類が可能な ChaSen を用いて形態素解析を行う。

2.2.2 連結処理

形態素解析により得られた形態素は、語を構成する最小単位と考えることができる。したがって、この形態素の連結により語となることがある。

そこで、特徴語候補の元になる語の集合を作るために、形態素の連結処理を行う。連結して新たな語となるかどうかは、その形態素の品詞により判別する。表

2.1 の形態素の中から連接することができるものを考えると「形態素」と「解析」は連結により「形態素解析」になると考えられる。また「分割」と「する」を連結し「分割する」とすることもできると考えられる。この場合「形態素解析」は連結によリ語とする意味はあるが「分割する」に関してはその意味は薄いと考える。「形態素」「解析」はともに名詞である。このように名詞が連続する場合は、連結することにより新たな語となることがある。しかし「分類」と「する」は名詞と動詞の連接である。この場合は「分類」の動詞化であるため、特徴語として抽出する意義は薄い。

また、形態素解析は常に正確な分割を行うとは限らない。例えば、「議事録」という単語は「議事」と「録」に分割されてしまい、この二つの品詞も名詞となる。

そこで、本研究では連結処理を行うのは名詞が連接している場合とする。これにより、複合語などの二つ以上の形態素からなる語を特徴語の候補として選ぶことができるようになる。

2.2.3 フィルタリング

連結処理により、特徴語候補の元となる集合を取得した。この集合の中からいくつかの制約によるフィルタリングを行い、特徴語候補を抽出する。

まず、品詞の情報により語をふるい落とす。動詞や形容詞は、一般的な動作や形容を表すための語である。特徴的な動作や形容などは、名詞に対して「する」や「的」などを付け加えることで表す。このことから、動詞や形容詞などは特徴語として扱わないこととし、名詞を特徴語の対象とする。また、形態素解析により未知語として判別される形態素はカタカナやアルファベットからなる語であり、実際の文中で名詞として使われることもある。そのため、本研究では特徴語として抽出するのは名詞と未知語からなる語とし、特徴語候補の元となる集合から名詞、未知語以外の品詞を持つ形態素を除去する。

このようにして得られた集合にも、特徴語となりえない語は含まれている。形態素の品詞は、名詞にもいくつかの小分類が存在する。この小分類の中で、特徴語の先頭に来ることのないような項目が存在する。それは「接尾」「非自立」「代名詞」である。接尾の品詞情報を持つ形態素としては「的」や「化」といったものである。これらの形態素が語頭に来る語は、そもそも語としての条件を満たせていないと考えられる。非自立の品詞情報を持つ形態素は「こと」「もの」のように単独では意味をなさない語である。これらの形態素も、語頭に来ることはないと考えられる。代名詞に関しては、語頭となることはあり得る。しかし、代名詞を関する語は、「この事実」「あの分類」などのように、何かを指し示すときに使うものである。したがって、これらの語を特徴語と考えることはできない。こういった事実から、先頭の形態素として「接尾」「非自立」「代名詞」が来ている語は、特徴語候補には含めない。

最後に、ある語の一部である語を除去する。例えば「議事録」という語は形態素解析により「議事」と「録」に分割されるが、それぞれ単体で使用されることはほとんどない。そういった場合、「議事」や「録」が特徴語とはならず「議事録」が特徴語となるはずである。このように、特定の語と連結することが多い語は、単独で特徴語となることはないと考える。そこで、ある語 s の出現数の 90%以上の割合で w を含む語 x の出現があった場合、w は x の一部であると考える。この場合 w は特徴語候補から除外する。

したがって、本研究で特徴語候補となるのは以下の制約条件を満たす語である。

- 構成する形態素は名詞か未知語である
- 先頭の形態素は「接尾」「非自立」「代名詞」の形態素情報を持たない
- 特定の語に頻繁に（90%以上の割合で）含まれることがない語

2.3 特徴語候補からの特徴語抽出

2.3.1 特徴ベクトルの設定

特徴ベクトルは、特徴語とそうでないものがそれぞれ判別できるようなものを設定しなければいけない。そこで、特徴語の言語的性質を考える。特徴語となるような語は、その文書中での話題となり、主語や目的語になることが多い。これは、説明や解説を行うためには主語や目的語になる必要があるからである。主語となる場合はその後ろに「が」「は」といった助詞が出現し、目的語となる場合はその前後に「を」「に」といった助詞が来ることが多い。

そこで、特徴ベクトルの要素として前後の品詞の出現割合を利用する。特に着目するべきは助詞の出現割合である。既に述べたとおり、助詞が前後に来る場合は特徴語となることが期待されるからである。そこで、特徴語候補の前後に助詞が出現する割合とそれ以外の割合を、特徴ベクトルの要素として考える。しかし、助詞の中にもさらに細かい品詞分けがなされている。特徴語が助詞と連接することが多いことは経験的に理解できるが、そもそも助詞と連接することが多いのは、特徴語以外の名詞でも同じである。そこで、助詞以下の細かい情報を別の特徴ベクトルの要素として考え、より詳細な情報を利用する。助詞はさらに、格助詞, 引用, 連語, 係助詞, 終助詞, 接続助詞, 特殊, 副詞化, 副助詞, 副助詞／並立助詞／終助詞, 並立助詞, 連体化の 10 個に分類されている。しかしこの中で、「特殊」の出現数は非常に少ないので要素からは除外する。また、前後に記号が来る場合のことを考えてみる。ここでいう記号とは句点や読点、各種括弧などのことである。これが前後に来るということは、そこで意味的な切れ目になるということである。よって、前後に記号が来る割合も特徴ベクトルの要素として考える。

しかし、前後の品詞による要素だけでは、文書に出現する回数に大きく依存してしまう。文書量に依存しない分類を行うために、別の観点からなる要素を用いる必要がある。本研究においては、議事録という特殊な文書から特徴語を抽出する。そこで、議事録から得られる特徴を特徴ベクトルの要素として取り入れる。本研究において対象としている議事録は、ゼミの発表を記録したものである。筆者の所属するの研究室では、ゼミの発表にはすべて PowerPoint の資料を用意する。そこで、この PowerPoint のスライドを利用して特徴ベクトルの要素を設定する。

スライドに載せることのできる情報には限りがある。そのため、発表者は必要となる情報だけをスライド中に記述するはずである。したがって、スライド中に出現する語は特徴語となりやすいと考えられる。そこで、特徴ベクトルとして、スライド中に出現するかどうかを要素として与える。さらに、出現する位置にも注目する。一般的に、スライドの一枚目には発表のタイトルを記述する。すなわち、一枚目のスライド中に出現する語は発表のタイトルに出現することになる。発表のタイトルは文章量的には少ないが、そこに出現するということは特徴語としての性質として重要と考えることができる。また、一枚一枚のスライドには、タイトルが付けられている。スライドのタイトルは、そのスライドの内容を端的に表すものをつけるため、ここに出現する語も、重要な語であると考えられる。そこで、特徴ベクトルとして、発表のタイトル、スライドのタイトル、それ以外、の三つに出現する回数を別の要素として利用する。さらに、出現する場所の文章量と重要度を考え、発表のタイトルは 100 倍、スライドのタイトルは 10 倍する。

このようにして各特徴語候補に対して特徴ベクトルを設定していく。

2.3.2　SVM による分類

各特徴語候補が持つ特徴ベクトルを、利用して特徴語を抽出する。そのために、特徴語候補を特徴語クラスと非特徴語クラスに分類することを考える。このようなクラス分類のために、SVM と呼ばれる識別器を利用する。

分類を行うための識別器には様々な種類がある。K-means 法 [5] やウォード法 [5]、多層パーセプトロン [6] などがある。これの識別器には、大きく分けて二つの種類がある。教師なしクラスタリングと教師ありクラスタリングである。

教師なしクラスタリングは、特徴ベクトルの分布からクラス分けを行う手法である。K-means 法やウォード法などがこれにあたる。この手法では、それぞれのクラスの分布が偏っている場合はうまくいくが、そうではない場合はうまくいかない。今回の特徴ベクトルで、その分布が偏っている保証がないため、この手法を適用することは難しい。

もう一つの教師ありクラスタリングは、分類を行う前に学習データを与える手法である。この手法は学習データを参考に、分離のためのパラメータを決定し分類を行う。しかし、教師ありクラスタリングにはいくつかの既知の問題がある。一

つは次元の呪い [6] と呼ばれる問題であり、もう一つは認識率の問題である。次元の呪いとは、高次元の特徴ベクトルを用いたときにおこる問題で、次元数の増加に伴い識別に時間がかかったり、どの特徴ベクトルも「似ていない」となってしまうことである。すなわち、望ましい分類の結果が得られなくなってしまうのである。認識率に関しては、多くの識別器では分離のためのパラメータに対して初期値を設定しなければならない。この初期値をどのように取るかにより、認識率に差が出てしまのである。

しかし、SVM では高次元のベクトルに対応し、高い識別率による分類が可能となっている。SVM は教師データから最適分離超平面を求め、未知データの分類を行う。SVM の識別関数は以下の式により表わされる。

$$f(x) = sign(g(x))\, g(x) = w^t x + b$$

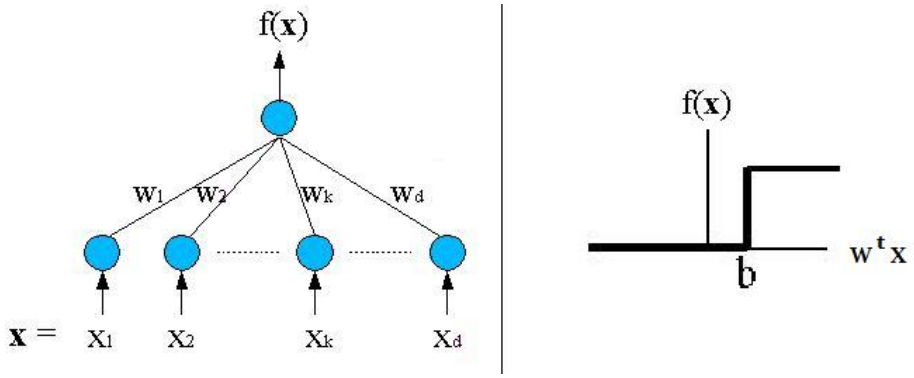

図 2.2: （左図）SVM のモデル（右図）識別関数の出力図

識別器として知られているものに、k-means 法やニューラルネットワークなどがある。それらの識別器との最大の違いはマージン最大化 [2] にある。SVM では、教師データの中から名前の由来にもなっているサポートベクター（Support Vector）を選び出す。これは、分離超平面に最も近いデータである。サポートベクターと分離超平面との距離をマージンと呼び、このマージンを最大化するところに SVM の特徴がある。

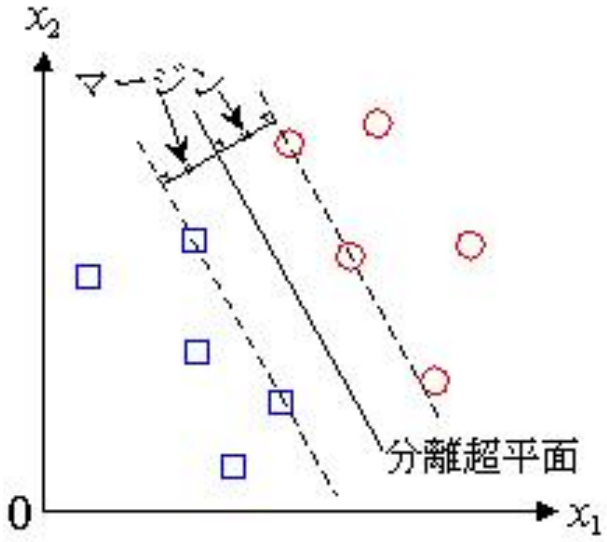

図 2.3: マージン最大化

これは、分離超平面の取り方に、明確なルールがあるということである。例えば、バックプロパゲーション学習 [6] では、与えられた教師データに関してのみ, 学習結果を保証し、特徴空間上の学習データが与えられていない領域に関しては、学習結果は初期値に依存する。その結果、初期値によっては汎化能力が低くなってしまい、認識率が低くなってしまうことがある。しかし、SVM ではマージン最大化という明確な基準が存在する。その基準は「分離超平面に最も近いデータとの距離が最も大きい」という、直感に合うものである。したがって、汎化能力が高くなり、認識率の向上につながる。

また、もう一つの特徴として、次に述べるカーネルトリック [2] が挙げられる。カーネルトリックとは、SVM を線形分離不可能な問題にも対応させるための仕組みである。線形分離不可能な問題を解くためには、元の特徴空間を線形分離可能な別の特徴空間に非線形写像することを行う。一般に、線形分離可能性は、特徴空間の次元が大きくなるほど高くなるため、より高次元への写像が行われる。したがって、写像後の新しい特徴空間での分類の際に、高次元の演算が行われ、その計算に膨大な時間がかかってしまう。しかし、SVM では目的関数や識別関数が、入力パターンの内積のみに依存した形になっているので、内積の計算さえできればよい。そこで、写像関数 ϕ による非線形写像後の空間における二つの要素 $\phi(x_1)$ と $\phi(x_2)$ の内積が $\phi(x_1)^t\phi(x_2) = K(x_1, x_2)$ のように計算できるのであれば、実際に $\phi(x_1)$ などを計算する代わりに、$K(x_1, x_2)$ を利用し、最適な識別関数を求めることができる。ここで、K のことをカーネルと呼び、カーネルを利用した非線形への対応手法をカーネルトリックと呼ぶ。

マージン最大化とカーネルトリックの二つの特徴から、SVM は高い汎化能力と線形分離不可能な問題への対応を実現している。本研究においては、多くの特徴語候補から特徴語の分類を行うため、汎化能力の高さは重要であり、また、線形分離可能であるかを事前に知ることが難しいため、SVM による分類が有効であると考えられる。そのため、特徴語抽出のための分類には SVM を用いる。

特徴語の抽出は人手で作成した学習データにより SVM に学習させ、そのパラメータを用いて、特徴語候補を特徴語クラスと非特徴語クラスに分類することにより行う。

第3章　議事録からの特徴語抽出とその応用

3.1　議事録からの特徴語抽出

3.1.1　対象とした議事録について

本研究で対象とした議事録は、筆者の所属する研究室で毎週行われるゼミの様子を記録したものである。一つの議事録は、一つのXML形式のファイルとして記録されている。この議事録ファイル中には、書記によって書きとめられた発表中の発言が記述されている。以下に、議事録ファイル中に記述されている発言の例を示す。

```
<statement id="statement5" type="follow" traceID="trace4"
 animationID="animation4_6">
  <speaker>
    <person refer="nagao"/>
  </speaker>
  <link from="statement4" type="default"/>
  <timecode>
    <start time="14:13:56"/>
    <end time="14:16:17"/>
    <update time="14:16:20"/>
  </timecode>
  <text>何も考えずにキーワードを打ちこむケースと，一覧から選ぶケースと何が違うのか．
        検索ボタンを押す前に違いがわかる．タグ一覧から選択することがどれくらい
        合致しているのか？それは根本的な問題ではないか．多くのタグクラウドは，
        タグ一覧のみだして，クリックするのみ．文字を打ち込む場合はない．
        その考えが正しくない可能性がある．妥当だと思う．
        世の中がそういう風に検索してくれない．
  </text>
</statement>
```

図 3.1: 議事録XML中の発言部分（一部）

このstatementタグ内に、発言に関する情報が含まれている。この中で、書記により記録された発言の内容を表すテキストは、textタグに含まれている文章であ

る。本研究では、この議事録ファイル中の text タグの中身を特徴語抽出のために利用する。

また、議事録ファイルには発表に関する情報も含まれている。

```
<presentation uri="http://www3.nagao.nuie.nagoya-u.ac.jp/
 minute/slide/masuda_080208"/>
<theme>
  <project>VA</project>
  <topic>VA</topic>
  <type>RR</type>
</theme>
```

図 3.2: 議事録 XML 中の発表情報

presentation タグにより、この議事録を一意に表すアドレスが記述されている。また、この発表がどのプロジェクトのものであるのかも theme タグ中の project タグにより記述されている。

これらの情報は、議事録の XML ファイルを解析することで容易に取得することができる。本研究では、特徴語抽出の前段階として、これらの情報をあらかじめ XML ファイルから抽出し、データベースへと保存している。

3.1.2　特徴語候補の抽出

第二章で定義した手法に基づき、実際に議事録の書記テキストから特徴語を抽出する。筆者の所属する研究室では、毎週ゼミを行い、その中での発言をゼミ中に書記が記録を取っている。書記の記録はすべて計算機に保存されている。このファイルの中から、発言内容に該当するテキストを取り出す。対象とした議事録は、2003 年 5 月 14 日から 2008 年 6 月 18 日までの 488 議事録で、取り出した総発言数は 19,722 発言だった。取り出したテキストは ChaSen を用いて形態素解析により形態素に分割し、データベースへと保存する。データベースには PostgreSQL を利用した。形態素解析の結果、総形態素数はのべ 573,539（異なり数 15,348）だった。ChaSen による形態素解析の結果、各形態素には以下のようなデータが与えられる。

- surface(文書中における実際の表記)
- basic_string(surface の活用のない基本形)
- cform(活用の種類)

- pos1〜pos4(品詞情報。1,2,3,4 の順に階層が深くなる)
- start(文章中の開始位置)
- end(文章中の終了位置)
- length(この形態素の長さ)
- reading(この形態素の読み)
- pronunciation(この形態素の実際の発音。「ドウサ」などは「ドーサ」になる)

surface は文章中に出現している表記そのものであり、これを連接することで元の文書を再現できる。basic_string は surface が活用されていた場合は、その基本形を表す。名詞などの活用のない形態素では surface と同じになる。cform には活用があった場合、その活用の種類を表す。pos1 から pos4 までは形態素の品詞情報である。start, end はテキスト中での形態素の出現位置である。start は 0 から始まり、end の文字を含まない。例えば、「形態素解析」の中から start=0, end=3 を取り出した場合「形態素」となる。length は形態素の長さであり end と start の差に等しい。reading, pronunciation はそれぞれ読みと発音である。

このデータに特徴語候補の抽出のためにいくつかのデータを加える。本研究で対象としている議事録は、その発言の一つ一つにユニークな URL を持っている。これは、筆者の所属する研究室の議事録が Web 上に公開され、発言の一つ一つにアクセス可能なためである。この URL を target として形態素のデータに付け加える。また、それぞれの発表はいくつかのプロジェクトに分けられる。そこで、その形態素が出現した発表のプロジェクト情報も付け加える。

以上により与えられるデータはデータベースに保存し、特徴語抽出の際に利用する。

このデータを用いて、まずは特徴語候補を抽出する。候補の抽出はデータベースに対し、クエリ（SQL 文）を与えることで実現する。

特徴語候補の条件は以下のようになる。

- 形態素の連接からなる
- 構成する形態素は名詞か未知語である
- 先頭の形態素は非自立、接尾、代名詞の品詞情報を持たない
- 90%以上の割合で同じ語に含まれる語は除く

名詞・未知語の品詞情報は pos1 に、非自立・接尾・代名詞の品詞情報は pos2 に入っているため、それらを特徴語候補の条件に合うように制約をつけたクエリに

より特徴語候補を抽出する。クエリを設定する際には、形態素の連接数を決める必要がある。そして、この連接数の最大値は 9 とした。これは、名詞の形態素が 10 個以上連接することがないことを確認したためである。クエリにより取得する情報は、連接するそれぞれの形態素の surface、出現する target、target 内での出現回数である。これらの情報は、後の分類のための特徴ベクトルを設定するときに必要となる情報も含んでいる。こうして取得した情報も、やはりデータベースに保存する。特徴語候補のテーブルは以下のような情報を持つ。

- word(特徴語候補)
- morpheme(特徴語候補を構成する形態素)
- target(出現する target)
- count(target 中の出現回数)
- g(大域的頻度)
- project(出現する発言が行われた発表のプロジェクト)

word は特徴語候補を表し、morpheme はその語を構成する形態素を表す。例えば、「形態」「素」「解析」の三つの形態素の連接からなる特徴語候補の場合、word=「形態素解析」、morpheme=「形態 素 解析」となる。target は特徴語候補の出現する発言の URL を表す。count は、この特徴語の発言中における出現数である。g は大域的頻度を表す double 値で、TF-IDF の算出に用いる。議事録集合から抽出した特徴語候補は 18,115 語となった。

大域的頻度とはその後がどれだけ少ない文書に出現しているか、ということを表す値で、一般的には IDF（Inverse Document Frequency：逆出現頻度）を用いる。本研究においても IDF を用い、その算出式は log([総文書数] / [語の出現する文書数]) である。この値が小さいと、多くの文書に出現することになり、特徴語ではなく、一般語であると考えられる。ここで問題となるのは、大域的頻度を求めるための文書集合である。本研究においては、対象を研究室のゼミの議事録としているため、語の出現に関して偏っていると考えられる。すると、この議事録を対象として大域的頻度を求めた場合、研究室特有の語が大域的頻度の値として小さくなり、一般語となってしまう。例えば、筆者の所属する研究室では、アノテーションを利用した研究を行っており、アノテーションという語は重要な意味を持っている。したがって、多くのゼミや発言の中にアノテーションという単語が出現する。そのため、議事録の集合を大域的頻度の算出に用いると、アノテーションという語の大域的頻度は低くなってしまい、「状況」「場合」などといった一般語と同じような値になってしまう。

そこで、大域的頻度を求める際には外部の情報を利用する。大域的頻度の値の信頼性を上げるためには、広範な話題を含む大量の文書集合であることが望ましい。そこで、本研究では情報の偏りが少ないと思われるWebページを利用して大域的頻度の算出を試みた。Webページに記載する情報は、ページ作者が自由に選ぶことができ、またあらゆる分野の人間が作者となりうる。そのため、話題の偏りは比較的少ないと考えられる。そこで、語をクエリとしたWeb検索のヒット数を語の出現する文書数と考えて、大域的頻度を求める。Web検索エンジンとしてYahoo検索を利用する。Yahooでは開発者向けにYahoo API[7]が配布されており、その中に検索APIが存在し、これを利用する。このAPIを利用するとYahoo検索エンジンによる検索の結果を取得することができ、検索にヒットしたページ数を取得することができる。そのようにして大域的頻度を求め、データベースに保存する。

ここまでの手順により、特徴語候補を抽出することができた。しかし、特徴語抽出を行うためには、特徴ベクトルを設定する必要があり、そのための情報がまだ不足している。そこで、特徴ベクトルに用いる情報を収集する。

3.1.3 特徴ベクトルに用いる情報の収集

まずは、特徴語候補の前後の品詞情報をデータベースに保存する。特徴語となる語は主語や目的語となることが多い。つまりは前後の品詞として助詞の出現する割合が多くなると考えられる。そこで、特徴語候補の前後の品詞情報をデータベースに入れ、特徴ベクトルに利用するのである。特徴語候補と同じように、形態素情報が保存されているデータベースから、特徴語候補とその前後の品詞情報を取得する。特徴語候補を取得するためのクエリを、その前後に連接する形態素の品詞情報を取得するように修正し、求めるデータを取得する。ただし、文頭や文末の特徴語候補は、前後いずれかの連接がなく形態素情報が存在しないため、このようなクエリでは前後どちらかの形態素が存在しない語の前後の品詞情報を取得することができない。そのため、文頭の特徴語、文末の特徴語、それぞれの前や後の品詞情報を別のクエリにより取得する。このようにして得られた連接する品詞情報を入れるデータベースは以下のとおりである。

- word(特徴語候補)
- morpheme(特徴語候補の形態素情報)
- front_pos(wordの前に出現する品詞)
- back_pos(wordの後ろに出現する品詞)
- count(出現する回数)

- target(出現する発言の URL)
- project(出現する発言が行われた発表のプロジェクト)

word, morpheme, target, project は特徴語候補のデータベースと同じ情報である。front_pos, back_pos は、それぞれ前と後ろの品詞情報であり、pos1, pos2, pos3, pos4 の連結となっている。ただし、文頭の特徴語候補の front_pos は「文頭」、文末の特徴語候補の back_pos は「文末」となる。count は word, front_pos, back_pos の target 中の出現回数である。

次に、スライド中の特徴語候補をデータベースへと保存する。特徴語となるような語は、スライド中に頻繁に出現すると考えられるため、特徴ベクトルにスライド内への出現回数を利用する。そこで、まずスライド中のテキストを形態素解析し、データベースへと保存する。そして議事録からの特徴語候補と同じように、クエリによりスライド中の特徴語候補を抽出しデータベースへ保存する。スライド中の形態素や特徴語候補を保存するデータベースは、それぞれ議事録中の形態素や特徴語候補を保存するものとほとんど同じ形式であるが、ひとつ情報を追加する。それは、出現する場所の情報である。

一枚のスライドには、タイトル部と本文が存在する。また、一枚目のスライドはタイトルスライドとして用いられ、発表のタイトルが記述される。本研究ではこのような別の場所での出現回数を、それぞれ個別に求めるため、語の出現する場所を表す情報をデータベースに加える。また、スライド中の特徴語候補は議事録の特徴語候補の出現があるかどうかを判別するために用い、TF-IDF によるランク付けを行わないため大域的頻度 g は不要となる。よって、スライド中の特徴語候補のデータベースは以下のようになる。

- word(特徴語候補)
- morpheme(特徴語候補を構成する形態素)
- target(出現する target)
- count(target 中の出現回数)
- location(スライド中の出現場所)
- project(出現する発言が行われた発表のプロジェクト)

word, morpheme, target, count, project は議事録の特徴語候補と同じであるが、g の代わりに location を保存する。出現場所は、発表タイトル、スライドタイトル、それ以外で分類する。さらに、出現場所の違いによる、文書量と重要度の違いから、保存する値は、発表タイトルの場合では出現回数$\times$ 100、スライドタイトルの

場合では出現回数× 10、それ以外の場合では出現回数、とする。この数値は経験的に決定した。

以上により取得した品詞情報とスライド情報により、各特徴語候補の特徴ベクトルを設定する準備が整った。

3.1.4 特徴ベクトルの設定

このようにして得られた情報を元に特徴語候補に特徴ベクトルを設定していく。特徴ベクトルに用いる要素は、前後に出現する品詞の割合と、スライド中への出現情報である。

本研究では、プロジェクトごとに特徴語を抽出するため、特徴語候補に対する特徴ベクトルの設定はプロジェクトごとに行う。したがって、同じ特徴語候補でも、プロジェクトにより特徴ベクトルが変化する。

特徴ベクトルの設定は、データベースにクエリを与えることで行う。品詞情報とスライド情報は、別々のデータベースに存在するため、それぞれに対してクエリを与える必要がある。このクエリには、どのプロジェクトの特徴ベクトルを取得するかを指定し、目的のプロジェクトに関する情報を集める。このようにして収集した情報は、各プロジェクトごとに特徴ベクトルを保存するデータベースを用意し、そこに保存する。

特徴ベクトルには前後に出現する品詞の割合を利用している。こういった統計情報は、サンプルが少ないと、データとしての信頼性が低くなってしまう。例えば、出現数が一回しかない語の前後に出現する品詞は、100%の出現率となってしまう。こういった情報はノイズになることが予想され、分類の精度を低くする恐れがある。そのため、各プロジェクトにおいて一度しか出現しない語は分類には利用しない。

このデータベースには各特徴語候補の TF-IDF 値も保存する。TF-IDF 値は、文書中への出現頻度を利用するため、同じ語でもプロジェクトごとに違う値が出てくるためである。また、このデータベースに保存することで、ここから特徴ベクトルを取得する際に、TF-IDF によるソートが行えるようにもなる。

3.1.5 特徴語候補からの特徴語抽出

ここまでで、各プロジェクトの特徴語候補に対して特徴ベクトルを設定した。次に、これらの特徴ベクトルを利用し、特徴語候補から特徴語を見つけ出すために、SVM による分類を行う。

本研究では、各プロジェクトごとに分類を行うが、プロジェクトによっては議事録のデータ数が少なく、うまく分類できないと予想されるプロジェクトもあっ

た。本手法では、特徴ベクトルによる分類を行うが、その要素として統計量を用いている。そのため、データ数が少ない場合ではうまくいかない。そこで、実際の分類を行うのは、データ数が上位の3つのプロジェクトとした。このプロジェクトはそれぞれ「AT（Attentive Townvehicle）」「DM（Discussion Mining）」「VA（Video Annotation）」と呼ばれるプロジェクトである。本研究では、これら3つのプロジェクトからの特徴語抽出を行った。

SVMは学習データを必要とするため、まずはそのデータを集める必要がある。学習データとは、正しく分類されている正解となるデータのことであり、機械的に取得することはできない。そこで、各プロジェクトのメンバーに、そのデータを作成してもらった。データの作成には、各プロジェクト2名ずつにいくつかの特徴語候補を見てもらい、それらの語を特徴語となるかどうかを判断してもらった。本研究では、特徴語の基準は「プロジェクトに関係のある語で、非プロジェクトメンバーに説明を要する語」とし、この定義に基づいて特徴語の判断をしてもらたった。そして、2人ともが特徴語、あるいは特徴語ではないと判断したものを学習データとして利用した。各プロジェクトごと間での分類の差をなくすため、学習データ数はすべて、特徴語として50語、特徴語でない語として50語、合計100語とした。

SVMではその実行の際に、カーネル関数[2]と呼ばれるものを指定する必要がある。カーネル関数とは、SVMが線形分離不可能な問題に対応するために用いる関数で、多項式カーネル関数やシグモイド関数と呼ばれるものがある。この関数は二つの特徴ベクトルx_1, x_2の内積を求める関数であり$K(x_1, x_2)$として表すことができる。カーネル関数の取り方により分類に差が出るが、一般に最適なカーネルを求めることは難しいそこで、本研究では最も適用範囲が広いと言われるガウシアン型カーネルを用いる。ガウシアン型カーネルはカーネル関数$K(x_1, x_2)$を次のように定義する。

$$K(x_1, x_2) = exp\left(-\frac{|x_1 - x_2|}{2\sigma^2}\right)$$

σは任意の実数である。この関数値は、入力x_1, x_2間の元の特徴空間での距離が大きいほど小さい値をとる。カーネル関数は二つの入力の写像後の特徴空間における内積として定義されている。したがって、カーネル関数の値が小さいということは、写像後の空間での内積が小さいということになる。そのため、ガウシアン型カーネルを用いることで、元の特徴空間において孤立している点が、写像後の特徴空間上で原点近くに配置されるような写像関数による分類をおこなうことができる。

ガウシアン型カーネルを利用する際には、パラメータσを設定しなければいけない。しかし、最適な値を決定する方法が確立されていないため、本研究では、学習データにを与えた後の分類の結果を確認しながら値を調整した。得られた特徴

語候補のうち実際の特徴語となりうる語はそれほど多くはないと考えられる。そこで、特徴語として分類された語の数が特徴語候補の 20%程度になるようにパラメータ値を調整した。

SVM による分類は、svm_light[8] というソフトウェアを用いた。このソフトウェアは、大量のデータに対して高速に動作する。svm_light は学習用モジュールと識別用モジュールから構成されている。学習用モジュール svm_learn.exe は以下のようなパラメータにより呼び出される。

svm_learn [options] examle_file model_file

[options] では様々なオプションが利用できるが、本研究において利用したものは次のものである。

表 3.1: svm_learn で利用したオプション

オプション	値	解説
-t	2	カーネルの選択。2 はガウシアン型カーネルを表す
-g	結果により調整	ガウシアン型カーネルのパラメータを設定
-c	100000	誤識別率とマージンの比重

-t オプションにより、用いるカーネル関数をガウシアン型カーネルとし、-g オプションによりそのパラメータを指定する。このパラメータ値は識別後の結果を見ながら決定した。-c オプションには大きな値を設定することで、学習データの誤識別率を下げることができる。example_file は学習用データを記述したファイルを指定する。そのフォーマットは以下のとおりである。

[line] .=. [target] [feature]:[value] [feature]:[value] ...[feature]:[value]

この一行が一つのデータを表している。[target] は、このデータが正例、負例のどちらであるかを表す。1 なら正例を、-1 なら負例となる。[feature]:[value] では特徴ベクトルの要素の値を記述する。[feature] で特徴ベクトルの何番目の要素なのかを指定し、[value] により値を指定する。[feature] は順に値が増えるように記述しなければいけない。model_file は出力ファイルであり、学習結果から得られた SVM の各種パラメータが記述されている。このファイルは識別用モジュールの実行の際に読み込まれ、学習結果による分類を実現する。識別用モジュールは以下のようなパラメータにより呼び出される。

svm_classify [options] examle_file model_file output_file

本研究において、識別用モジュールのオプションは使用せず、すべてデフォルトの値で実行した。example_file は識別用のデータである。ファイルのフォーマットは、学習用データとほとんど同じであるが、学習用のデータは [target] の値として 0 を設定する。model_file には学習用モジュールにより出力されたものを指定する。

model_file から識別関数を決定し、識別用のデータを分類する。output_file には識別結果を出力するファイルを指定する。一行に一つのデータの識別関数の値が記述される。このファイルだけでは、どの語がどちらに分類されたかがわからないが、output_file のデータの順番は、example_file に記述した特徴ベクトルの順番と同じであるため、出力後に両方のファイルを比べ、結果と語を一致させる。output_file に記述された値の正負により、正のクラスに分類されたか、負のクラスに分類されたかを知ることができる。学習データとして、特徴語となるものを正のクラスとして与えたので、正のクラスに分類されている特徴語候補が、本手法により抽出できた特徴語となる。この分類結果は、学習の際に与えたパラメータに依存している。パラメータを変更することにより、分類の偏りが変化する。本研究においては、特徴語候補のうち 20 %程度が特徴語になるという推論の元、特徴語に分類された語数が、特徴語候補数の 20 %程度になるようにした。学習の段階では、パラメータの影響を知ることができないため、適当な値から開始し、特徴語に分類された語数を確認しながらパラメータを変更していった。各プロジェクトのパラメータを決定し、実際に分類を行った結果を表 3.1 に示す。

表 3.2: 各プロジェクトにおける分類結果

プロジェクト	A	B	特徴語候補数	特徴語に分類された語（割合）
AT	100	36	1869	346（18.5%）
DM	100	100	2459	570（23.2%）
VA	100	42	1938	424（21.9%）

A:学習の際に与えたデータ数
B:実際に学習に用いられたデータ数

SVM による分類では、特徴ベクトルを分離するための分離超平面を求めるために、学習データのすべてを用いているわけではない。これは、マージン最大化と呼ばれる SVM の特徴で、分離超平面に最も近いデータのみを利用して分離超平面を決定している。そのため、与えた学習データ数と、実際に学習に利用されたデータ数が違う場合がある。実際に使われたデータ数が少ない場合は、特徴ベクトルを分離するために分離超平面から遠いデータが多いということである。これは、特徴ベクトルがどれほど偏っているかということと関連しており、実際に使用するデータが少ない場合、学習データは、各クラスごとに偏っていると言え、使用したデータが多い場合には、学習データが密集していると言える。したがって、DM プロジェクトは実際に使用した学習データ数が多いため、特徴ベクトルが密集していたことがわかる。

このように行われた分類により、特徴語抽出がなされたかを評価する必要がある。第四章では、この分類結果が妥当なものであるかどうかを評価する。

3.2 議事録から得られた特徴語の応用

3.2.1 辞書の作成補助

本研究では、議事録から得られた特徴語を辞書の作成に利用することを考えている。本研究で抽出した特徴語は、辞書の見出し語としての役割を持つ語である。人手で辞書を作成するためには大きな労力が必要となり、特に見出し語や索引語となる語を選ぶ際には必要と思われる語を議事録や論文などを見ながら一つづつ見つけていかなければいけない。しかし、提案手法では、学習データとして与える 100 語を選ぶだけでよいため、全てを人手で作業することに比べ必要な労力は少なくなる。辞書の見出し語を作成することができれば、その語に対する説明を書き加えることで、辞書を作成することができる。この説明を書き加える際にも、特徴語は有効に働く。

本研究で抽出した特徴語は、議事録の中に出現する語である。したがって、その語の説明を書く際には、議事録が手掛かりとなる。しかし、単純にその語が出現する議事録や発言を取り出すだけでは、膨大な量が出てきてしまう。また、語が表す意味や、説明として必要な情報などは、時間とともに変化することが考えられる。

そこで、議事録の時間情報を利用して、これらの問題の解決を図る。本研究では、議事録に出現する語の中から特徴語を抽出した。議事録は、その発表が行われた日付と関連付けられているため、その中から抽出した特徴語にも同様に日付の関連付けが可能となる。

語に対する日付の関連付けから、話題の推移を知ることができる。日付と出現数の関連をグラフに可視化することで、その語が集中して使われている時期を知ることができる。

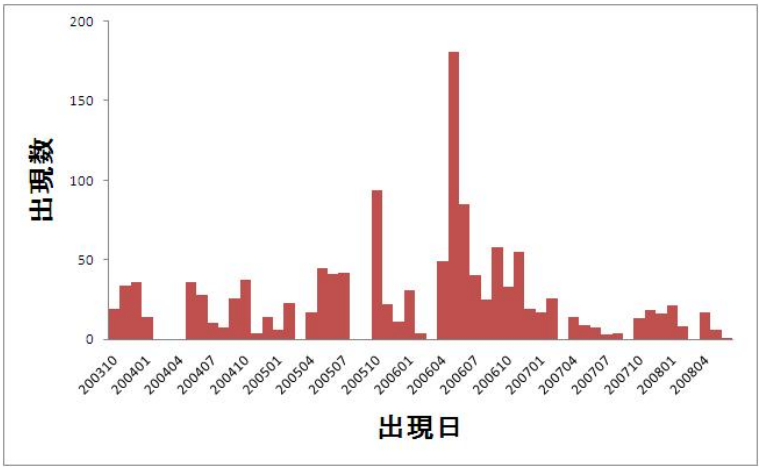

図 3.3: 「アノテーション」の出現日と出現数のグラフ

図 3.1 は「アノテーション」という語の出現日と出現数の関係を表したグラフである。ここでは、一か月を単位として、その月での出現数を表示している。図 3.1 から、「アノテーション」という語は、2006 年 4 月から 2006 年 7 月にかけての出現数が多くなっている。したがって、この時期においてアノテーションに関する活発な議論が行われていたと推測される。

このように、語が集中的に出現している時期を知ることは、辞書の説明に利用可能な議事録を知る手ががりとなる。ある語が集中的に表れている時期には、その語の説明や、関連する議論が多く行われていると考えられる。そこで、その時期の議事録中の対象とする語を含む発言をピックアップすることで、目的の発言を見つけることができると考えられる。あるいは、そこからさらに細かく情報を選別してもよい。発言を行った発言者により、その内容の意味は少し変わる。一般に、発表者以外の発言は、発表者に対する質問や意見が主な内容であることが多く、発表者の発言は、質問に対する解答や自らの意見を述べる場合が多い。

このように、グラフ上での突出した部分が一つであるなら、比較的話は簡単である。しかし図 3.2 のような場合は少し考える必要がある。

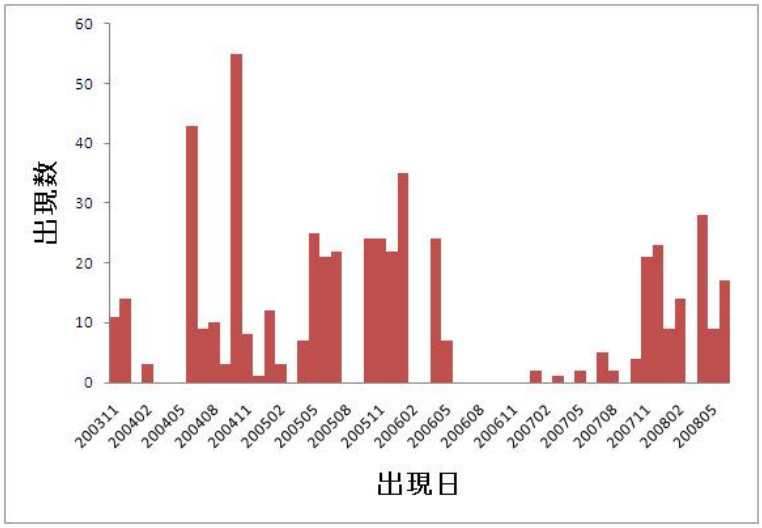

図 3.4: 「AT」の出現日と出現数のグラフ

図 3.2 は筆者の所属する研究室で開発している AT と呼ばれる乗り物の出現日と出現数のグラフである。この AT はこれまでに何度も改良を重ねているため、さまざまなバージョンの AT が存在する。したがって、同じ AT という語でも、その意味するところは必ずしも同じとは限らない。図 3.2 からわかるように、AT の出現数は、ある特定の期間に集中しているわけではない。グラフを見ると、2004 年 5 月から 2005 年 2 月まで、2005 年 5 月から 2006 年 5 月まで、2007 年 11 月から現在までの、三つの山があるように取れる。

このように山が複数ある場合には、それぞれの山で語の意味が変わっている可能性がある。山が複数あるということは、その語に関する議論が一度落ち着いた後に、再度議論が始まっているからと考えることができる。そういった場合、以前とは異なる視点で議論が行われている可能性が高く、その語の意味が変化することがある。こうした意味の変遷を辞書から知ることができれば、過去の議事録を見る際にも、語の意味を正確に把握した状態で閲覧することができる。

ある期間に集中的に出現しているような語は、その部分に着目することで辞書としての説明を記述する際に役立つ議事録や発言への手掛かりを得ることができる。しかし、出現が集中せず、大体において同じ頻度で出現する語もある。こういった語で、出現数が平均的に高い語は、そのプロジェクトの活動において前提となる語だと考えられ、最も重要な語の一つだと言える。

このように、時間と出現数の関係を利用することで、特徴語の性質についての手掛かりを得ることができる。議事録以外の文書、例えば論文や Web 上のテキストなどでは、そのコンテンツが作成された時間情報が、必ずしもその中に出現する語と結びつかない。

例えば、数人の研究者が同時期に、ある共通の話題を論文で取り上げようとして、執筆を始めたとする。しかし、それを書き上げるまでにかかる時間は、人それぞれであるため、論文の完成時間がそろうことはない。そうすると、その中に出現する語を論文の作成日と関連付けることはできない。かといって、その語をいつ使い始めたのかをわざわざ記録するのは手間がかかる。

Web上のテキストは、論文に比べれば素早く書きあげることができるため、関連付けができるように思える。しかし、アイデアを温めてからテキストを書く人は少なからず存在する。また、ある話題が落ち着いたときに、改めて振り返る目的で文書を書くこともある。

このように、一般のテキストではその文書中の語に時間情報を対応付けることは困難である。しかし、議事録では、会議中の発言のリアルタイムに記録しているため、その時間情報と内容との関連は確かなものとなっている。そのため、議事録の時系列順に、語の出現数を並べた場合の偏りは、情報として信頼できるものである。したがって、この情報を用いることで特徴語のさらなる応用が可能になると考えられる。

3.2.2 特徴ベクトルの再利用

本研究では特徴語候補に対して特徴ベクトルを設定することで特徴語を抽出した。この特徴ベクトルを再利用することを考える。

特徴語の中には、意味的に近いものが存在する。たとえば本研究では、特徴語を辞書の見出し語として定義した。そのため、「特徴語」と「見出し語」は共通の意味を持っていると考えられる。意味的なつながりを持った語は、文書中の使われ方においても、共通する性質があると思われる。

そこで、本研究で利用した特徴ベクトルのうち、前後の品詞情報だけを用いて何らかのクラスタリングを行うことで、意味的に近い語をクラスタリングできると考えられる。このクラスタリングにより、類義語や関連語を見つけることができると期待される。

また、意味的なつながりがある場合は、共起される語にも共通の傾向があると考えられる。特に、特徴語同士の共起は、語の性質を表す上で重要な情報となる。特徴語という一般的ではない語が共起されているということは、そのつながりが強いものであるからと推測される。そこで、特徴語との共起の傾向を調べることで、特徴語を分類することができると考えられる。

このように抽出した特徴語を利用することで、辞書の作成を円滑に行うことができる。そして、このようにして作成された辞書を用いることで、自らに必要な知識に、容易にたどり着くことができるようになる。

3.2.3 入力支援

本研究では、議事録の中から特徴語を抽出した。したがって、この特徴語が再び議事録に出現する可能性は十分に考えられる。

議事録のテキストは、書記が会議中にキータイプにより入力している。そのため、タイプミスや表記の揺れなどが存在する。例えば、「PowerPoint」と入力するべきところで「PwerPoint」と入力をミスしたり、「クエリー」「クエリ」のようにまったく同じ語の表記に違いが出ることがある。

これらのミスを回避するために、本手法で抽出した特徴語を利用する。書記がテキストを入力する際に、特徴語の中からインクリメンタル検索を行い、表記のミスや揺れを防ぐことができると考えられる。先ほどの例でいえば「P」を入力した時点で「PowerPoint」「Presentation」などの語を書記に提示する。そうすれば、書記はこの候補の中から語を選び、共通の表記が可能となる。

これは、書記だけではなく、論文の作成にも応用できる。論文中の語に表記ミスや揺れがないかを調べるのは大変な労力となる。そのため、論文の作成の時点でそういったミスをなくすことができるとよい。そこで、書記と同じように特徴語のインクリメンタル検索を行い、候補中から選択することで統一した表記にすることができる。

第4章　評価・考察

4.1　評価方法

提案手法では、特徴語候補を特徴語クラスと非特徴語クラスに分類している。この分類を評価するために、各クラスに含まれる実際の特徴語の割合を調べる。分類がうまくいっているならば、特徴語クラスに含まれる特徴語の割合は高く、非特徴語に含まれる特徴語の割合は低くなっているはずである。しかし、特徴語クラスと非特徴語クラスの割合を比較するだけでは、分類が妥当であるということはできるが、特徴語抽出としての有効性が示せない。

そこで、TF-IDF による特徴語抽出との比較を行う。TF-IDF による特徴語抽出は、各特徴語候補の TF-IDF 値を求め、降順にソートしその上位何語かを特徴語とするものである。すべての特徴語候補、分類により特徴語クラスに分類された語、非特徴語クラスに分類された語、それぞれの TF-IDF 上位 200 語中に実際の特徴語が含まれている割合を比較する。提案手法による特徴語抽出が有効であるなら、特徴語クラス、すべての特徴語候補、非特徴語クラス、の順に特徴語を含む割合が大きいはずである。

実際の特徴語かどうかは、学習データとの一貫性を保つために各プロジェクトのメンバーに判断していただいた。特徴語かどうかは、学習データの作成と同じく、各プロジェクト二名に「プロジェクトに関係のある語で、非プロジェクトメンバーに説明を要する語」を選んでもらい、二人の共通部分を特徴語と判断する。

また提案手法では特徴ベクトルを利用しているが、その要素に前後の品詞情報とスライド中の出現数の二つの観点からの要素を採用した。このように全く別の観点から特徴ベクトルを構築したことの有効性も評価する。そこで、前後の品詞情報、スライド情報、それぞれ単独で特徴ベクトルとした分類との、実際の特徴語を含む割合の比較を行う。また、単独の情報による分類と組み合わせた分類との抽出できた語の傾向についても分析する。

4.2　TF-IDF との比較による評価

各プロジェクトごとの特徴語候補の語数、特徴語クラスの語数、非特徴語クラスの語数と、それぞれ上位 200 語が含む特徴語の割合を表 4.1 に示す。

表 4.1: 各クラスの語数と上位 200 語の特徴語を含む割合

	A	B	C	D	E	F	G	H
AT	346	71	35.5%	1523	35	17.5%	45	22.5%
DM	570	88	44.0%	1889	38	19.0%	50	25.0%
VA	424	73	36.5%	1514	28	14.0%	39	19.5%

A:特徴語クラスに分類された語数
B:特徴語クラスの上位 200 語中の特徴語の語数
C:特徴語クラスの上位 200 語中の特徴語の割合
D:非特徴語クラスに分類された語数
E:非特徴語クラスの上位 200 語中の特徴語の語数
F:非特徴語クラスの上位 200 語中の特徴語の割合
G:全特徴語候補の上位 200 語中の特徴語の語数
H:全特徴語候補の上位 200 語中の特徴語の割合

表 4.1 の C、F、H が上位 200 語中の特徴語の割合である。これを比較すると、すべてのプロジェクトにおいて特徴語クラスの割合が最も高くなっている。また、特徴語クラスの割合は全特徴語候補の割合よりも高く、非特徴語のクラスの割合は全特徴語候補の割合よりも低くなっている。次に、上位 200 語までの特徴語を含む割合の推移を図 4.1 に示す。

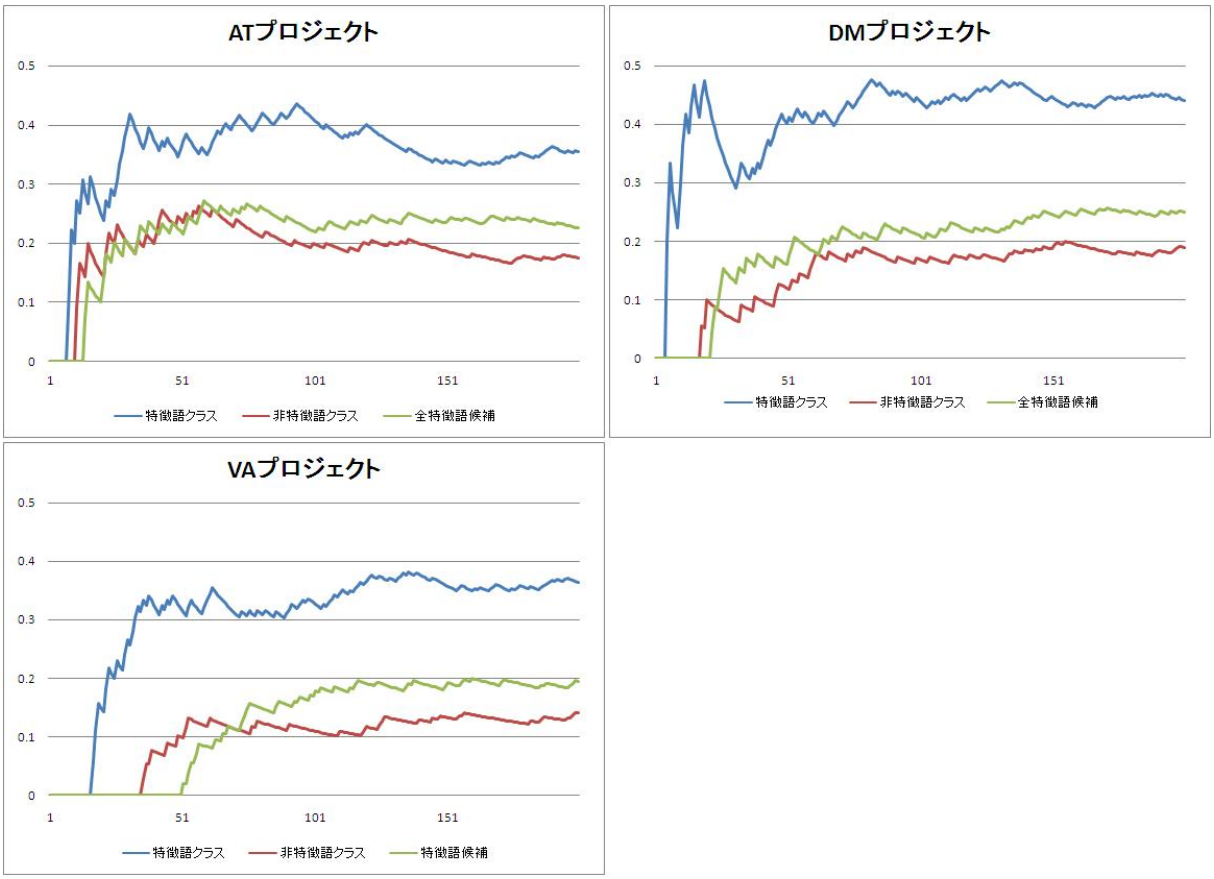

図 4.1: 各プロジェクトの特徴語を含む割合の推移

図 4.1 のグラフは、横軸が TF-IDF による順位を、縦軸が横軸の示す順位までに含まれる特徴語の割合を示している。青色のグラフが特徴語クラスを、赤色のグラフが非特徴語クラスを、緑色のグラフが TF-IDF のグラフをそれぞれ表している。図 4.1 からわかるとおり、どのプロジェクトにおいても上位 200 語の特徴語を含む割合は、常に特徴語クラスが最大となっている。また、それぞれの分類のグラフは TF-IDF の順位が下がるにつれ変化が少なくなっている。したがって、上位 200 語以降の語についても、同様の割合で特徴語を含むと予想される。このことから、提案手法による特徴語抽出は、TF-IDF による手法に比べ高い精度で特徴語抽出が行えていると言える。

4.3　特徴ベクトルの評価

次に、特徴ベクトルについて評価する。既に述べたように、本研究では、特徴ベクトルとして前後の品詞情報とスライド中の出現数を利用している。この二つの観点からなる要素を組み合わせたことにより、分類にどのような影響を与えているかを調べる。前後の品詞情報からなる要素をのみを利用した特徴ベクトル、スライド中の出現数のみを要素として利用した特徴ベクトル、二つの要素をまとめた特徴ベクトルの、三つの特徴ベクトルによる分類結果を分析する。分析の対象は、TF-IDF 上位 200 語とする。まず、各特徴ベクトルによる分類の結果を表 4.2 に示す。

表 4.2: 各特徴ベクトルによる分類結果

プロジェクト	用いた情報	A	B	C
AT	品詞情報	352	63	31.5%
	スライド情報	271	63	31.5%
	品詞情報＋スライド情報	346	71	35.5%
DM	品詞情報	543	85	42.5%
	スライド情報	566	66	33.0%
	品詞情報＋スライド情報	570	88	44.0%
VA	品詞情報	435	69	34.5%
	スライド情報	432	55	27.5%
	品詞情報＋スライド情報	424	73	36.5%

A:特徴語クラスの語数
B:特徴語クラス上位 200 語中の特徴語の語数
C:特徴語クラス上位 200 語中の特徴語の割合

表 4.2 から、情報を組み合わせることにより、抽出できる特徴語の数は増えていることがわかる。しかし、情報を組み合わせて抽出した語が、単独の情報で抽出できる語をすべて含んでいるとは限らない。情報を組み合わせた結果、単独の情報では抽出できていた語が落ちていることもある。それとは逆に、単独では抽出できなかった語が、情報を組み合わせることで抽出できるようになることもある。各プロジェクトにおける、特徴ベクトルに用いる情報の違いによる抽出語の差を表 4.3 に示す。

表 4.3 からわかるように、情報を組み合わせることで落ちてしまっている語は多くある。このように情報を組み合わせることによって落ちてしまった語はどのような語であるのか。まず品詞情報単独では抽出できていた語について調べた。AT プロジェクトでは 38 語が組み合わせることによって落ちている。これは、特徴ベ

表 4.3: 特徴ベクトルによる抽出語の違い

	A	B	C	D
AT	38	48	25	39
DM	35	38	35	39
VA	30	31	24	33

A:品詞情報では抽出できたが組み合わせることで抽出できなくなった語数
B:品詞情報では抽出できなかったが組み合わせることで抽出できた語数
C:スライド情報では抽出できたが組み合わせることで抽出できなくなった語数
D:スライド情報では抽出できなかったが組み合わせることで抽出できた語数

クトルにスライド情報を追加したことが原因であると考えられるので、これらの語のスライド情報を調べた。これらの組み合わせにより落ちた語は、スライド中への出現回数が極端に少なかった。スライドに載せることができる情報は限られているため、細かい説明などは口頭で行うことが多い。そういったスライドに出現しにくい特徴語は、品詞情報だけを用いることで抽出できるが、スライド情報を追加することで、それがノイズとなり抽出できなくなったと考えられる。各プロジェクトの特徴語、非特徴語のスライドへの平均出現回数との比較を表 4.4 に示す。

表 4.4: スライド情報の比較

プロジェクト	分類	A	B	C
AT	特徴語	117	49	35
	非特徴語	44	22	15
	品詞情報 [1]	5	4	7
DM	特徴語	176	65	50
	非特徴語	87	58	43
	品詞情報 [1]	303	47	30
VA	特徴語	124	56	47
	非特徴語	40	26	21
	品詞情報 [1]	0	7	16

A:発表タイトルへの出現回数× 100 の平均
B:スライドタイトルへの出現回数× 10 の平均
C:発表タイトル、スライドタイトル以外への出現回数の平均
[1]:品詞情報だけで抽出できるが組み合わせると抽出できない語

非特徴語は特徴語に比べ、スライド中への出現回数が少ない。そのため、スラ

イド中への出現回数が少ない語は、非特徴語として分類されやすいと考えられる。

次に、スライド情報単独では抽出できていたが、品詞情報を追加することで抽出できなくなった語について調べた。これらの語は、品詞情報がノイズとなった可能性がある。例えば、出現回数が非常に少ない場合では、極端な偏りができてしまい、品詞の情報が特徴語としての性質を満たさないと考えられる。あるいは、そもそも品詞情報では特徴語としての性質がない語の可能性もある。

本研究では特徴語が主語や目的語になるという前提で、その前後の品詞情報を用いて特徴語抽出を行っているが、それが有効に働かない特徴語も存在する。いろいろな語と組み合わさることが多い語などは、その前後に助詞が来ることは少なくなる。例えば「構造」という特徴語はその後ろに助詞以外が来る割合は50%近くに上る。これは「構造化」「構造的」のように、その後ろに名詞の品詞が来る割合が多いためである。こういった接尾語が連結する語は、前後の品詞情報が特徴語としての性質を満たさなくなることが多い。

第5章　関連研究

本研究のように、テキスト中から何らかの情報を抽出しようとすることをテキストマイニングと呼ぶ。テキストマイニングによる研究には、文書の自動分類や[9][10]、特徴語抽出 [11]、単語の意味抽出 [12] などがある。その中から本研究に関連するいくつかの研究について述べる。

5.1　PAI（Priming Activation Indexing）

文書中から抽出する特徴語にはさまざまなものが考えられる。その中で、論文中から著者の主張を特徴語として抽出する研究がおこなわれている。この研究では、語の活性度という点に着目して特徴語抽出を行っている [13]。

人が文書の内容を理解できるのは、文書中の語が脳に記憶され、それを解釈するからである。脳の解釈の方法について詳細は未知であるが、記憶のメカニズムは徐々に解明されつつある。記憶には、ある語が想起（活性化）されるとその語に関連する語も想起（活性化）されるというプライミング効果があり、その想起の早さはその想起頻度に依存することが確認されている。

このプライミング効果は文書の理解の作用にも深く関連していると考えられている。それは、文書を読み進めるにつれ話題が読者の頭の中に展開され、それに伴い記憶が活性化されていく中で文脈を理解し、内容を把握していると考えられるからである。著者もそれに対応して、ある主張を文書で行う場合に、著読者に理解してもらえるように文章を構成していくことが自然である。まずは研究の背景を述べ、そこから話題を膨らまし著者の話題へと導いていき、それを理解するための基礎を十分作り上げた上で主張を展開するのだと考えられる。このような著者の展開に誘導されるようにして読者の記憶に強く残る語、すなわち強く活性化される語を抽出するための手法が PAI である。

PAI では文書を意味的なまとまり（セグメント）に分割し、一つ一つのセグメントごとに語のネットワークを形成する。このネットワークのリンク構造に活性伝搬と呼ばれる手法を適用することで、各語の活性値と呼ばれるものを算出する。この活性値の高い語が、著者の主張したい語であると考えられる。

PAI による特徴語抽出は、論文のような長さのある文書を対象として、著者の主張したい語を特徴語として抽出することが目的となっている。だが、本研究で

対象としている文書は議事録中の発言であるため、一つ一つの文書が短く PAI を適用することができない。しかし、語の活性度という着目点は、本研究においても見習うべき点であると考えられる。

5.2　特徴ベクトルを用いた語の意味の弁別

本研究では、最終目標として辞書の作成を挙げた。そのためには、語の意味を知る必要があるが、これを機械的に抽出することは難しい。そこで、特徴ベクトルを用いたクラスタリングにより語の意味を弁別する研究が行われている [14]。

九岡らの研究では、本研究と同じように語に対して特徴ベクトルを設定し、k-means 法によるクラスタリングを行い、語の意味を弁別している。この研究では、連接ベクトル、文脈ベクトル、連想ベクトル、トピックベクトルという 4 つの特徴ベクトルを定義し、それぞれの特徴ベクトルでの分類を行っていた。本研究とは違い、これらのベクトルは語に対して与えられるのではなく、文書中に出現している個々の出現語に対して特徴ベクトルを設定している。そして、個々の出現語を分類することで語の意味を弁別している。

本研究とは違い、語の意味に着目した特徴ベクトルを用いているが、その考え方には参考にすべき点が多い。また、意味による分類を行うことは、辞書の作成という点からいっても、見習うべき点がある。しかし、この手法は大量の文書情報を必要としてる。本研究で対象とした議事録は、この手法が適用できるだけの十分な量ではなかったため、この手法の適用は今後の課題である。。

5.3　語の共起関係を利用した特徴語抽出

特徴語の抽出に語の共起関係を利用する研究は多く行われている。その中に χ^2 検定の χ^2 値を用いた研究がある [15]。

松尾らの研究では、文書中において重要な意味を持つ語は、共起する語に何らかの偏りがあると考えた。そこで、頻出語の出現割合と、頻出語との共起割合の間にどのくらいの偏りがあるかを χ^2 検定により調べた。χ^2 検定では、統計量 χ^2 を求めることにより、二つの分布のずれを知ることができる。χ^2 値が大きければ、二つの分布のずれが大きく、特徴語であると言える。松尾らの研究では、単純な出現割合ではなく、文の長さを考慮に入れて χ^2 値を求めている。

本研究では議事録集合という話題の広い文書集合を対象としているため、頻出語との共起関係が必ずしも特徴語の性質として現れるとは限らない。したがって、この研究で用いられている手法を適用することはできないが、共起関係を特徴語抽出に役立てることは、本研究においても重要であると考えられる。

5.4 語のランキング手法

特徴語のランキング付けには、TF-IDF が用いられことが多い。しかし、TF-IDF では出現数が多い語が過剰に評価されてしまい、ランキングの上位に来てしまう。したがって、いくつかの語が連接する複合語などの出現頻度が低い語は、本来重要な語であるにもかかわらず下位にあることが多い。この問題に対し相澤の研究 [16] では、複数の観点に基づくランキング手法の算出法を提案している。

相澤の手法では、複合語を適切に評価するために構成する語の関係や、語を追加することで新たな語となるかなどを評価するための尺度を定義している。それは、結合度、出現度、前接度、後接度、文脈度、重要度と呼ばれる尺度であり、これらに基づいてランキングを行う。この手法を用いることで、多くの語が連結した複合語でも、重要な語であれば上位にランキングされるようになる。語を適切にランキングすることは、特徴語抽出に大いに役立つ。例えば、本研究の特徴語候補をランキングし、あるランク以下の語を足切りすることができれば、より精度の高い特徴語抽出ができると考えられる。

しかし、この手法を本研究に適用することは難しい。本研究で対象としている議事録は、発表中の発言を記録したテキストである。一般に発言に用いられる語は短い語が選ばれる傾向にある。特に、長い語がスライドなどの資料中に出現している場合は、指示語により表されることが多い。例えば、「オンラインビデオアノテーション」という語についての発言をする際、この語がスライド中に出現していると、「このアノテーション」として発言することが多い。したがって、議事録中に出現する複合語は、非常に少ないことが予想される。実際、本研究で抽出した特徴語候補 5033 語のうち複合語は 1049 語、そのうち 3 つ以上の語から成る複合語は 102 語しかなかった。このように複合語が少ない状況では、相澤の手法は有効とは言えない。

第6章　まとめと今後の課題

6.1　まとめ

今までに経験のない仕事を行う際には、多くの場合その仕事に必要な情報を持っていない。仕事を円滑に進めるためには、必要な情報を知るための辞書のようなものがあることが望ましい。ユーザは、その辞書中の項目を見て、必要だと感じたならば、その語に関係する議事録や資料などを閲覧し、情報を収集することができる。このような辞書を作る際には、その見出し語や索引語を抽出することが必要となる。しかし、そのような特徴的な語は大量に存在すると考えられ、人手で選定することは現実的ではない。そこで本研究では、すでに存在する議事録から、機械的に特徴語を抽出する手法を提案した。

特徴語を抽出するためには、その性質について考えなければいけない。本研究では、前後に来る形態素の品詞情報と、発表に利用されるスライドへの出現数に着目した。特徴語となるような語は、主語や目的語となることが多く、そのため、前後の品詞として助詞が多くなり、逆に非特徴語は助詞が前後に来ることは少ないと推測される。スライドへの出現に関しても、特徴語となるような語は、発表スライドへの出現が多くなると考えられる。

このように特徴語となる語は、前後の品詞情報やスライドへの出現に関して、何らかの偏りが見られと考え、それらを特徴ベクトルとして利用し、SVM を用いた分類により特徴語を抽出した。その結果は、TF-IDF による特徴語抽出よりも高い精度で特徴語を抽出することができた。前後の品詞情報とスライド情報を組み合わせた効果についても検証した。単独の情報を用いた場合に比べ、情報を組み合わせた場合のほうが抽出できた特徴語の割合は高くなっていた。このことから、特徴ベクトルに複数の観点からなる要素を用いることは有効であった。

6.2　今後の課題

提案手法による特徴語抽出では、抽出した特徴語のうち、最高でも 40%程度しか実際の特徴語を含んでいなかった。機械的に特徴語を抽出するためには、より高精度での特徴語抽出が必要となるため、この割合を上げるための工夫が必要である。

6.2.1　特徴ベクトルの構築

本論文において、二つの観点からなる要素を組み合わせたことにより特徴語の抽出割合が増加したことを述べた。そこで、より精度の高い特徴語抽出を行うため、特徴ベクトルの要素に別の観点からなる要素を加えることが考えられる。

例えば、発言中の共起関係などを利用することも考えられる。特徴語であるならば、共起する語にも何らかの偏りがあると考えられる。共起する語が偏るならば特徴語である可能性が高くなり、逆に多くの語と共起するなら非特徴語であると考えられる。他にも、頻出語との共起の割合や、特徴語候補同士での共起関係も利用できると考えられる。

こういった共起関係を定量化し、特徴ベクトルの要素として利用することができれば、特徴語抽出の精度が上がることが期待される。

6.2.2　情報の組み合わせによる抽出語の違い

本研究では、特徴ベクトルとして前後の品詞情報とスライド情報を利用した。情報を組み合わせることで抽出できた特徴語の割合は高くなったが、単独の情報で抽出できていた語が抽出できなくなることもあった。これは、もともとの情報に対して追加した情報がノイズとなってしまったと考えられる。例えば、品詞情報が特徴語としての偏りを持っていたとしてもスライド情報が非特徴語としての偏りを持っていた場合などである。

こういったことは、特徴ベクトルに用いる情報を増やすことで、より顕著になる可能性がある。この問題を解決するためには、特徴語についての詳細な分析が必要となる。例えば、品詞情報による抽出がうまくいった語には、ある傾向や類似点が見つかると考えられる。あるいは、プロジェクトの活動内容の差や、扱う問題の種類によっては、特徴語となる語の性質が顕著に表れる情報が違うことも予測される。

より多くのプロジェクトの特徴語に対しての調査を行うことで、それらの傾向を見つけることができると考えられる。その結果を特徴ベクトルに反映させたり、事前のフィルタリングに適用することで、特徴語抽出の精度を上げることができると考えられる。

6.2.3　分類の問題

本研究では特徴語候補の分類にSVMを利用した。これは、SVMが識別精度の高い識別器であったからである。

しかし、SVMは分類のために学習データを必要とする。学習データを選び出すことはどうしても人手の作業となってしまうためコストがかかる。さらには、分

類結果は学習データに依存するため、その選び方に注意を払わなければいけない。学習データを用いない分類手法としては、ウォード法や K-means 法などがある。これらの手法は、特徴ベクトルの分布の様子により分類を行っている。分類結果のクラスには、特徴ベクトルが類似しているものが集められているだけで、どれが特徴語のクラスとなるかはわからない。そのため、分類語に特徴語のクラスとするものを選ばなければいけない。

精度の高い特徴語抽出を行うためには、このようなさまざまな分類手法から最適な手法を選ぶ必要がある。最適な手法は、特徴ベクトルにより変わってくるため、その詳細な分析が必要となる。

参考文献

[1] VeriSign, http://www.verisign.com/

[2] 小野田崇. ”サポートベクターマシン (知の科学) ”, オーム社, 2007

[3] Juman, http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[4] Chasen, http://chasen-legacy.sourceforge.jp/

[5] 神嶌 敏弘, ”データマイニング分野のクラスタリング手法 (1) － クラスタリングを使ってみよう！ －”, 人工知能学会誌 vol.18 no.1 pp.59-65, 2003

[6] 石井 健一郎, 上田 修功, 前田 英作, 村瀬 洋. ”わかりやすいパターン認識”, オーム社, 1998

[7] Yahoo! デベロッパーネットワーク, http://developer.yahoo.co.jp/

[8] SVM-Light, http://svmlight.joachims.org/

[9] 鈴木誠, 平澤茂一. ”単語と N-gram の各カテゴリにおける出現頻度の比の和を用いたテキスト自動分類法”, 電気学会論文誌C（電子・情報・システム部門誌）Vol.129, No.1 (2009) pp.118-124, 2009

[10] 鈴木誠. ”カテゴリ間の単語頻度の差分を用いたテキストの自動分類”, 日本経営工学会論文誌 Vol.59, No.4 (2008/10) pp.355-363, 2008

[11] 斎藤和巳, 中野良平. ”ニューラルネットを用いたテキストの特徴語抽出”, 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 Vol.103, No.490 (20031201) pp.13-18, 2003

[12] 小田誠雄, 西村靖司, 小田まり子, 横田将生. ”市販電子化辞書からの自然言語の意味抽出”, 情報処理学会研究報告. 自然言語処理研究会報告 Vol.97, No.29(19970321) pp.21-26, 1997

[13] 松村真宏, 大澤幸生, 石塚満. ”語の活性度に基づくキーワード抽出法”, 人工知能学会論文誌, 17 巻 4 号 F, pp.398-406, 2002

[14] 九岡佑介, 白井清昭, 中村誠. ”複数の特徴ベクトルのクラスタリングに基づく単語の意味の弁別”, 言語処理学会第 14 回年次大会, 発表論文集, pp.572-575, 2008

[15] 松尾豊, 石塚満. ”語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム”, 人工知能学会論文誌 17 巻 3 号 D, pp.217-223, 2002

[16] 相澤彰子. ”テキストコーパスにおける特徴語抽出のための分析ツール”, 情報処理学会研究報告 情報学基礎研究会報告 Vol.2001, No.20(20010305) pp.113-120, 2001

謝辞

本研究を進めるにあたり、指導教員である長尾確教授には、研究に対する姿勢や心構えといった 基礎的な考え方から、 研究に関する貴重な御意見、論文執筆に関する御指導など、大変お世話になりました。 心より御礼申し上げます。

大平茂輝助教には、研究に関することから技術的なことまで幅広く御指導、御意見を頂き、大変お世話になりました。 心より御礼申し上げます。

土田貴裕さんには、研究に対する取り組み方や、基礎的な知識を身に付けるにあたり数々のご指導を頂きました。また、研究に行き詰まったときに的確な助言をいただくこともありました。ここに御礼申し上げます。

石戸谷顕太朗さん、増田智樹さん、尾崎宏樹さん、安田知加さん、森直史には、研究や実装に関することや、研究室生活における様々な面でお世話になりました。ありがとうございました。

長尾研究室秘書である鈴木美苗さんには、研究室生活や学生生活の様々な面でお世話になりました。 ありがとうございました。

最後に、影ながら見守っていただき、生活を支えていただいた両親にも最大限の感謝の気持ちをここに表します。 ありがとうございました。