平成20年度 情報工学コース卒業研究報告要旨

長尾 研究室	氏　　名	杉 浦　広 和
卒業研究題目	議事録集合からの特徴語抽出と その応用に関する研究	

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。語の抽出は私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。例えば、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かもしれない。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。特徴語抽出は図1のような流れで行う。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。そのため、図2のように時間軸と出現数を軸としたグラフに可視化することで、話題の推移状況を知ることができる。

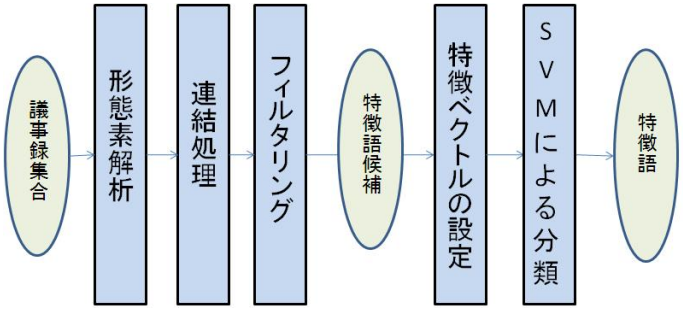

図1：特徴語抽出の流れ

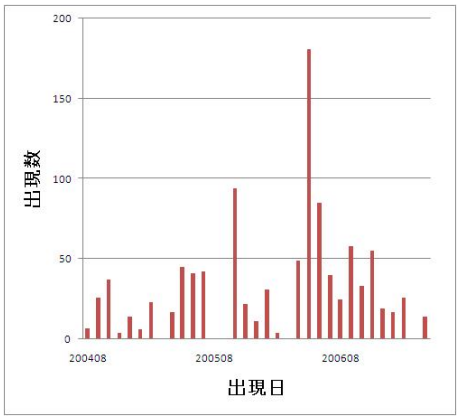

図2：「アノテーション」の
出現日と出現数のグラフ