自己組織化マップを用いた
記述型アンケートの内容分類の検討

はじめに

選択型アンケート

機械による集計作業

機械による集計作業
のため分析作業が容易

記述型アンケート

人手による集計作業

・アンケート回答に全て目を通し分析
を行うため、作業時間がかかる
・分析者によって分析結果が異なる

集計作業が
困難

アンケートの集計作業の自動化が求められている

はじめに

集計作業の補助として、文の分類の自動化を行う

仮定：同一の単語により構成された文は内容も同一である

“毎日かかさず勉強をする”

“勉強を毎日2時間以上する”

文中に出現する単語により文を内容ごとに分類

違う単語が同義として使用されているケースも多々存在する

勉強を毎日2時間以上する

毎日休まず学習する

単語クラスタリングを行い、同義の単語のクラスタを作成

提案手法の流れ

アンケート文章

遅刻や欠席をしない
資格取得のため勉強をする

単語（標準形）

遅刻 / や / 欠席 / を / し / ない
資格 / 取得 / の / ため / 勉強 / を / する

単語クラスタ

{勉強　学習　・・・}
{ 取得　得る　・・・}

文クラスタ

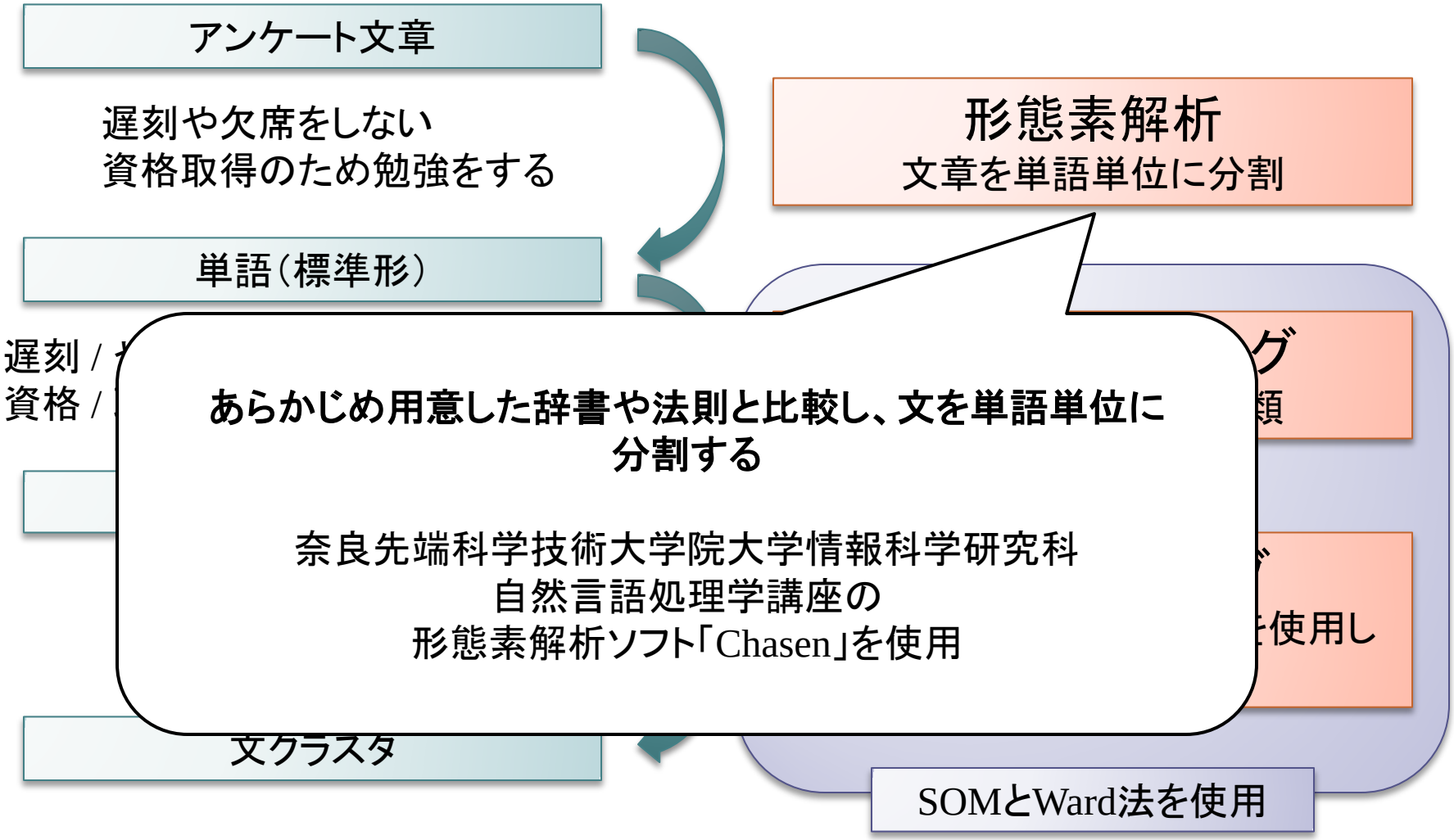
形態素解析
文章を単語単位に分割

単語クラスタリング
出現確率で単語を分類

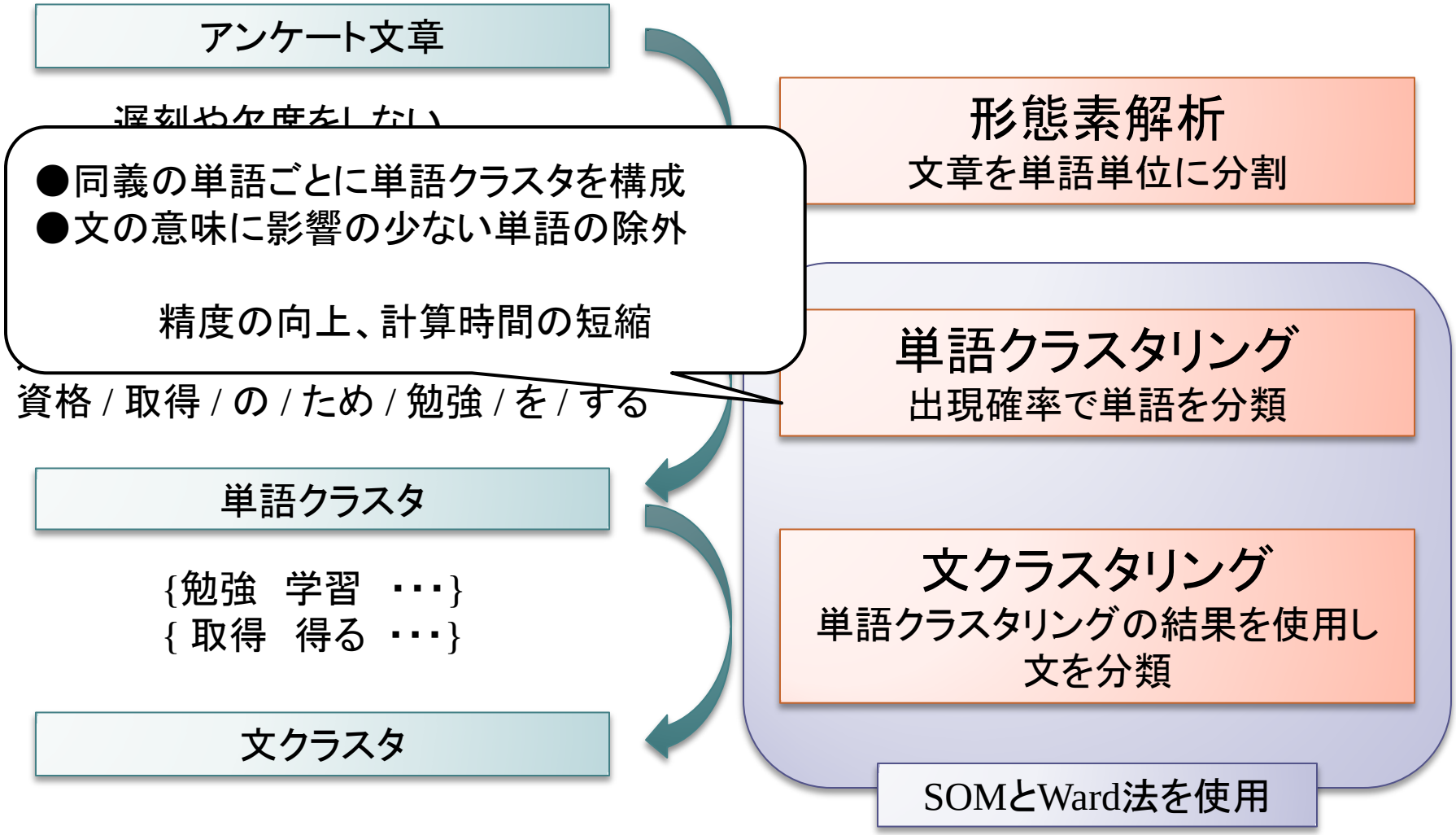
文クラスタリング
単語クラスタリングの結果を使用し
文を分類

SOMとWard法を使用

提案手法の流れ


アンケート文章
遅刻や欠席をしない
資格取得のため勉強をする
単語（標準形）
遅刻 /
資格 /
文クラスタ
形態素解析
文章を単語単位に分割
グ
類
使用し
SOMとWard法を使用
あらかじめ用意した辞書や法則と比較し、文を単語単位に分割する
奈良先端科学技術大学院大学情報科学研究科
自然言語処理学講座の
形態素解析ソフト「Chasen」を使用

提案手法の流れ


アンケート文章
遅刻や欠席をしない
●同義の単語ごとに単語クラスタを構成
●文の意味に影響の少ない単語の除外
精度の向上、計算時間の短縮
資格 / 取得 / の / ため / 勉強 / を / する
単語クラスタ
{勉強　学習　・・・}
{ 取得　得る　・・・}
文クラスタ
形態素解析
文章を単語単位に分割
単語クラスタリング
出現確率で単語を分類
文クラスタリング
単語クラスタリングの結果を使用し
文を分類
SOMとWard法を使用

提案手法の流れ

アンケート文章

遅刻や欠席をしない
資格取得のため勉強をする

単語（標準形）

遅刻 / や / 欠席 / を / し / ない
資格 / 取得 / の / ため / 勉強 / を / する

単語クラスタ

{勉強　学習　・・・}
{ 取得　得る　・・・}

文クラスタ

形態素解析
文章を単語単位に分割

単語クラスタリング
出現確率で単語を分類

文クラスタリング
単語クラスタリングの結果を使用し
文を分類

SOMとWard法を使用

文クラスタリング

文中に出現する単語から文の内容ごとに分類を行う

文中に含まれる単語クラスタから
出現単語ベクトル X_S を決定する

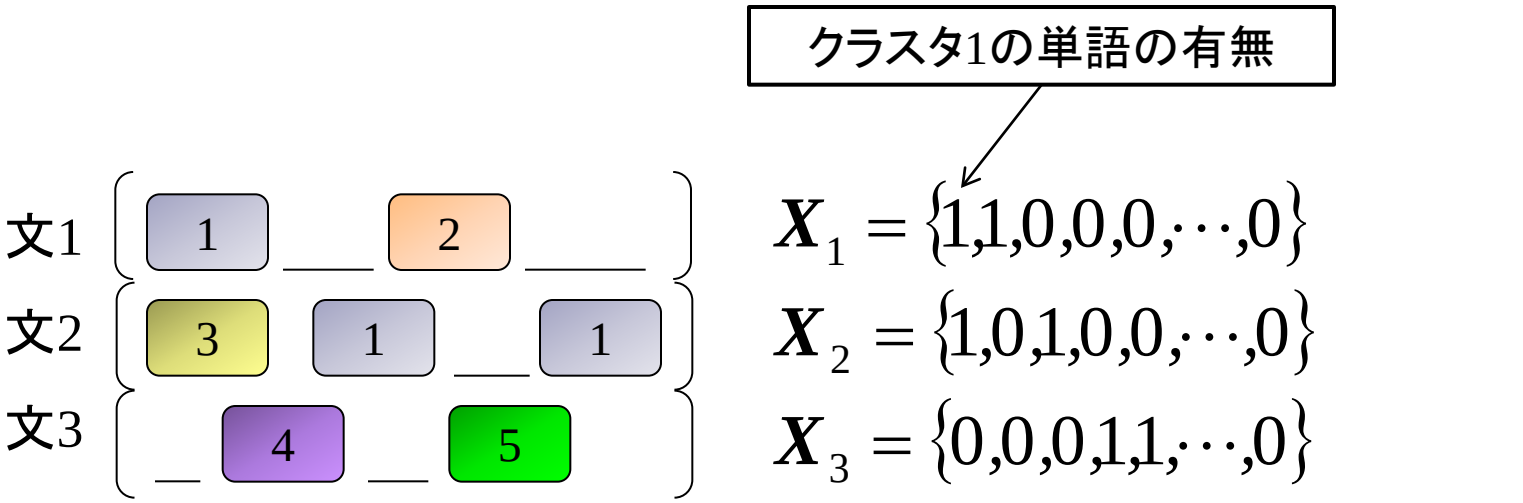
$$X_s = \left\{ x_1^s, x_2^s, x_3^s \cdots x_n^s \right\}$$

$$x_n^s = \begin{cases} 1 \cdots \text{単語クラスタnが文s内に存在する場合} \\ 0 \cdots \text{単語クラスタnが文s内に存在しない場合} \end{cases}$$

文クラスタリング

文中に出現する単語から文の内容ごとに分類を行う

文中に含まれる単語クラスタから
出現単語ベクトル $\boldsymbol{X}_S$ を決定する



出現単語ベクトルを入力として
SOMとWard法によるクラスタリングを行う

実験1

アンケート「人間力」
鳥取大学生320名に対して「今年度の目標」について尋ねるアンケート

構成された文クラスタ例

・単位を落とさない
・全ての単位を取る
・1年時落としたものを回収
・なるべく多くの単位取　　etc

・同義の文のみで構成されている
・その意味をもつ文全てがそのクラスタに属す
30クラスタ中7クラスタ　　597文中140文

・電気主任技術者の資格を取る
・技術を身につける
・サークルの全国大会に出場
・サークルとの両立　　etc

2種類の意味の文により構成されている
30クラスタ中20クラスタ　　597文中317文

・目覚ましを使う
・ダキョウしないこと　　etc

他に似た文が存在しない文により構成
（特殊な言い回し　誤字脱字　ひらがな）
30クラスタ中1クラスタ　　579文中140文

検討

提案手法の問題点1

回答文の大多数を同義の文が占めているとき
その文が結果に与える影響が大きくなり過ぎる（場合がある）

出現回数の多い文の学習量が多くなるため
周囲の文も同一のクラスタに分類されてしまう

改善策

学習前に同一文の除外を行う
同一文：お互いに出現単語ベクトルの等しい文

$$\boldsymbol{X}_1 = \{1,1,0,0,0,0\}$$
$$\boldsymbol{X}_2 = \{1,0,1,0,0,0\}$$
$$\boldsymbol{X}_3 = \{1,1,0,0,0,0\}$$
$$\vdots$$

検討

提案手法の問題点1

回答文の大多数を同義の文が占めているとき
その文が結果に与える影響が大きくなり過ぎる（場合がある）

出現回数の多い文の学習量が多くなるため
周囲の文も同一のクラスタに分類されてしまう

改善策

学習前に同一文の除外を行う
同一文：お互いに出現単語ベクトルの等しい文

$$\boldsymbol{X}_1 = \{1,1,0,0,0,0\}$$
$$\boldsymbol{X}_2 = \{1,0,1,0,0,0\}$$
$$\boldsymbol{X}_3 = \{1,1,0,0,0,0\}$$
$$\vdots$$

同一ベクトルは
1つを残して除外

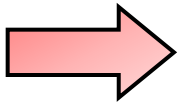
実験2

同一文の除外を行わない場合と行った場合の比較

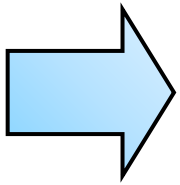
アンケート「人間力」
鳥取大学生320名に対して「今年度の目標」について尋ねるアンケート

除外を行わない場合	除外を行う場合
文：597文 文クラスタ数：33	文：125文 文クラスタ数：59

健康管理の徹底
健康維持
健康第一に生きる
レポートをしっかりする
レポートを期日内提出
出たレポートはその週中に終わらせる
etc



健康管理の徹底
健康維持
健康第一に生きる
etc



レポートをしっかりする
レポートを期日内提出
etc

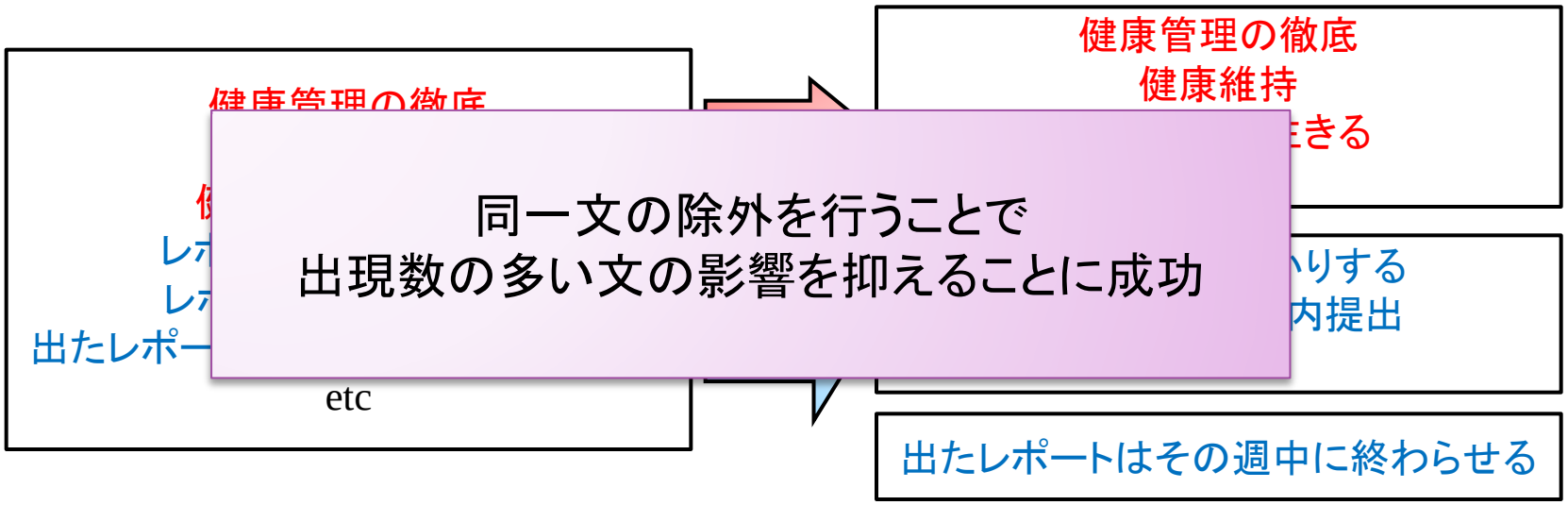
出たレポートはその週中に終わらせる

実験2

同一文の除外を行わない場合と行った場合の比較

アンケート「人間力」
鳥取大学生320名に対して「今年度の目標」について尋ねるアンケート

除外を行わない場合	除外を行う場合
文：597文 文クラスタ数：33	文：125文 文クラスタ数：59



同一文の除外を行うことで
出現数の多い文の影響を抑えることに成功

検討

提案手法の問題点2

回答文が長文であったり、回答者数が多いアンケートの場合
出現単語が膨大な数になり処理時間が増大する

改善策

単語クラスタリング時に文への影響の少ない単語を除外しておく

実験3

単語の除外を行わない場合と行った場合の比較

エレベータに対するアンケート　全540文

除外を行わない場合	除外を行う場合
文クラスタ数：64	文クラスタ数：64
同義の文のみで構成されたクラスタ 16クラスタ　104文（19.3%）	同義の文のみで構成されたクラスタ 13クラスタ　113文（20.9%）
2つの意味の文により構成されたクラスタ 25クラスタ　268文（49.6%）	2つの意味の文により構成されたクラスタ 25クラスタ　240文（44.5%）
上記以外のクラスタ 23クラスタ　168文（31.1%）	上記以外のクラスタ 26クラスタ　187文（34.6%）

実験3

単語の除外を行わない場合と行った場合の比較

エレベータに対するアンケート　全540文

除外を行わない場合	除外を行う場合
文クラスタ数：64	文クラスタ数：64
同義の文のみで構成されたクラスタ 16クラスタ　104文（19.3%）	同義の文のみで構成されたクラスタ 13クラスタ　113文（20.9%）
2つの意味の文により構成されたクラスタ 25クラスタ　268文（49.6%）	2つの意味の文により構成されたクラスタ 25クラスタ　240文（44.5%）
上記以外のクラスタ 23クラスタ　168文（31.1%）	上記以外のクラスタ 26クラスタ　187文（34.6%）

単語の除外を行う場合の方が処理時間の面で有利

おわりに

本研究では

文中での単語の出現頻度から文の内容分類を行う手法を提案した
提案手法の問題点に対して改善手法を提案し
実際のアンケートに適用して効果を検証した

結果として

提案した手法により問題点をかなり改善することに成功した
文の内容分類を精度についてはさらなる改善が必要である
集計補助には現段階でも有効な効果を得ることができる

今後は

内容分類にあたって単語の出現頻度に次ぐ新たな指標の取り入れなど
精度の向上のために手法の改善が必要である