修 士 論 文 の 和 文 要 旨

研究科・専攻	大学院　　情報理工学研究科　　総合情報学専攻　博士前期課程		
氏　　名	鷹栖　弘明	学籍番号	1330040
論 文 題 目	文節の係り受け関係を用いた観点に基づく意見クラスタリング		

要　　旨

Web 上には，様々なトピックに関する意見が存在し，トピックに関する意見には様々な観点のものが混在している．例えば,「原発」というトピックに関する意見には安全性やエネルギー，健康といった観点の意見が混在している．意見をこのような観点ごとに分類することで，観点ごとに意見を容易に把握・比較でき，新たな観点の意見を発見する手がかりにもなる．意見を観点ごとに分類する研究は少なく，分類する観点を予め設定しているものや，観点の差異を考慮していない手法がほとんどである．そこで本研究では，予め観点を設定せずに，文脈情報，とりわけ名詞と動詞の係り受け関係を考慮して意見集合に適した観点を自動的に特定・分類するクラスタリング手法を提案する．

本研究で提案する意見クラスタリング手法では,「意見の観点の違いは名詞と動詞の係り受け関係の違いに反映される」という仮定のもと，文節の係り受け関係から名詞Nと動詞Vのペア$\langle N,V\rangle$を抽出し，これをクラスタリングに利用する．具体的には，各意見から得られた文節の係り受け関係をもとに名詞とそれが係る動詞のペア$\langle N,V\rangle$を抽出する．そして，日本語 WordNet と潜在意味インデキシングを用いて計算した名詞Nどうしの類似度と動詞Vどうしの類似度から抽出した$\langle N,V\rangle$間の類似度を計算するが，特に，名詞Nどうしの類似度が高くなるほど動詞Vどうしの類似度が$\langle N,V\rangle$間の類似度に大きく影響を与えるように計算する．最終的に意見どうしの類似度を$\langle N,V\rangle$間の類似度から計算し，Ward 法による階層型クラスタリングを行う．

評価実験では，意見集合に対して人手による観点に基づいた分類と提案手法および従来のクラスタリング手法による分類がどの程度近いかということを指標として分類性能を調べた．実験の結果，提案手法では従来手法より高い分類性能となり，提案手法が有用であることが示された．

平成 26 年度　電気通信大学大学院
情報理工学研究科　総合情報学専攻　修士論文

文節の係り受け関係を用いた観点に基づく意見クラスタリング

提出年月日：　平成 27 年 1 月 30 日
提出者：　学籍番号 1330040
氏名 鷹栖 弘明
コース：　経営情報学コース
指導教員：　内海 彰　教授
尾内 理紀夫　教授

目次

1 はじめに

Web 上には様々な製品やサービスに関するレビューや，時事問題などに関する意見が存在している．このような Web 上のレビューや意見が「肯定的・否定的なものなのか」，「どういった評価項目・観点から述べられているのか」ということを知ることは，製品・サービスを利用する上でも，時事問題について自身の意見の幅を広げる上でも非常に有用である．しかし，amazon[*1]や楽天[*2]などに代表されるショッピングサイトにおける商品レビューの充実や，Twitter[*3]や Facebook[*4]などに代表されるマイクロブログサービスや SNS（Social Networking Service）による情報発信の容易さから，Web 上には膨大な量のレビューや意見が存在しており，すべてに目を通して，得たい情報を探し出すのは多大なる労力が必要である．

そこで，Web 上に存在するレビューや意見を抽出・整理することでユーザにとって有用な情報を探し出す「意見マイニング」や「センチメント分析」の研究が多く行われている．

これらの研究の多くは，あらかじめ評価項目が明確に決まっている製品やサービスに関する意見・レビューを対象として，それらの評価項目や極性（positive/negative）に基づく意見分類・要約を行っている [Pang 02, Turney 02, Hu 04, Liu 05]．最近では，評価項目が明確ではない時事問題などに関する意見に対しても，賛成・反対のようないくつかの立場に分類・要約する研究が行われている [Oh 09, Paul 10, Scholz 12, Trabelsi 14]．

しかし，あるトピックに関する意見集合には賛成・反対のような立場とは別に，様々な観点を示す意見が混在している．例えば「原発（問題）」というトピックに関する意見には，「安全性」や「エネルギー」，「健康」といった様々な観点を示す意見が含まれている．そのため，特定のトピックに関する意見を自動的に観点ごとに分類することで，観点ごとの意見を容易に把握・比較することができ，今まで気付かなかった新たな観点を発見する手がかりにもなる．

このような観点に基づく意見分類の研究は，今までほとんど行われていない．Wikipedia の外部情報を利用してあらかじめ用意した観点ごとに意見を分類する研究 [横本 11] は行われているが，観点をあらかじめ決めるのが困難な場合も多い上に，外部情報に依存した観点が設定されてしまうという問題点もある．また，あらかじめ観点を用意することなく，意見を観点に基づいて分類（クラスタリング）する研究 [Luo 09, 鷹栖 13] も行われているが，観点の性質を活かしきれ

[*1] http://www.amazon.co.jp/
[*2] http://www.rakuten.co.jp/
[*3] http://twitter.com/
[*4] http://www.facebook.com

ていないという問題点がある．

そこで，本研究ではあらかじめ観点を用意せず，観点の差異を考慮して意見集合を観点ごとにクラスタリングすることを目的とする．観点に基づく意見のクラスタリングには，意見どうしの類似度を計算する必要がある．言語表現間の一般的な類似度計算には，BoW(Bag of Words) [*5]アプローチに基づく TF-IDF 値[*6]などを用いたベクトル空間モデルが用いられるが，共通語を多く含む意見どうしが同じ観点を示すものであるとは限らない．また，ある意見中の TF-IDF 値が高い単語が，その意見の観点を示すとは限らない．そこで本研究では，文脈情報，とりわけ名詞と動詞の係り受け関係を考慮して，意見を観点ごとにクラスタリングする手法を提案する．

[*5] 単語の並びなどを考慮せず，文書中で単語が出現した頻度のみを考慮するモデル
[*6] 特定の文書に多く出現する単語を重要度が高いとみなす手法

2 関連研究

意見マイニングの分野では，評価項目や評価軸が明確である製品やサービスなどを対象として，複数の意見やレビューをセンチメント（positive/negative）に基づいて分類する研究 [Pang 02, Turney 02, Liu 12] が多く行われている．これらの研究では，文中に出現した単語の頻度や，製品などの評価に用いられる特徴的な単語の有無などを分類に用いる素性としている．

本研究の対象でもある時事問題などに対する意見を分類する研究も近年行われている．Oh ら [Oh 09] は，政治問題について述べられたブログ記事を対象に，単語単位の n-gram や単語の共起を素性として，記事を「賛成」や「反対」というグループに分類している．Oh らと同様に意見を賛成や反対といったグループに分類するその他の研究 [Anand 11, Paul 10, Somasundaran 10, Scholz 12, Trabelsi 14] では，単語の極性や係り受け関係，助動詞（should や outght など）などを素性として分類に用いている．これらの研究では，意見を賛成・反対といった立場に分類することが目的であり，意見の観点に基づいた分類をしていない．また，これらの研究の中には単語の係り受け関係を分類に用いているものもあり，本研究においても単語の係り受け関係を用いるが，関連研究では，単語の品詞を考慮せずそのまま素性として用いている一方，本研究では，係り受け関係の中でも名詞と動詞の係り受け関係を考慮するという点において，これらの研究とは異なる．

意見を観点ごとに分類する研究としては，横本ら [横本 11] の研究がある．この研究では，ユーザの意見が述べられたブログ記事を対象として，Wikipedia の情報を用いて記事集合を観点ごとに分類する手法を提案している．分類に用いる観点として，トピックを表す話題語を含む Wikipedia 記事集合を取得し，それらの記事タイトルの中で分類対象のブログ記事に多く出現するものを用いている．しかし，この手法では，Wikipedia の記事タイトルに出現しない観点を設定することができず，それが原因で対象のブログ記事集合の分類に適した観点集合を設定できない可能性がある．さらに，ブログ記事中に含まれる単語が直接観点になるとは限らないという問題点もある．

本研究と同様に，意見をクラスタリングする研究としては，Luo ら [Luo 09] と鷹栖ら [鷹栖 13] がある．Luo らは，文中に含まれる単語や句の出現頻度をもとに TF-PDF 値という TF-IDF 値を改良した重み付け手法を用いて，意見が述べられた Web ページ集合に対してクラスタリングを行っている．最終的にはクラスタリングによって得られた各クラスタを観点とみなして，クラスタごとに意見の賛成・反対を求めることによって，トピックに関する意見の特徴を探ることを目的としている．しかし，TF-PDF 値自体は観点の特徴や差異を考慮した重み付け手法ではない上に，クラスタリング手法自体も一般的な BoW に基づいた手法に過ぎない．また，鷹栖らは，

Twitter 上に存在する意見（意見ツイート）を対象として，ユーザがつぶやいた意見ツイートの周りにある意見に関連したツイートを利用して，意見ツイートのクラスタリングを行っている．鷹栖らの研究では観点に基づいたクラスタリングを目指しているものの，観点の性質を活かしきれていないという問題点がある．以上のことから，観点の性質を活かした意見のクラスタリング手法を提案した研究は今までに行われていない．

一般的な文書クラスタリングの研究において，単語どうしの結びつき（共起語）を考慮した文書クラスタリング手法が，小熊ら [小熊 05] や村上ら [村上 07] によって提案されている．本研究においても，名詞と動詞という単語の結びつきを考慮するが，これらの研究では，1 文書内に共起する単語の情報を利用しているだけであり，係り受け関係などの文脈情報までは利用していない．

係り受け関係を用いてクラスタリングを行う研究では，類義語発見のために単語をクラスタリングすることを目的とした研究 [真野 08, 風間 09] が多く，文書をクラスタリングすることを目的とした研究は行われていない．

3 要素技術

本章では，本研究および評価実験で用いる要素技術について述べる．

3.1 形態素解析

日本語における形態素解析とは，与えられた語句または文を形態素（単語）に分割することである．形態素とは，意味を持つ最小単位の語のことを指す．

小さい燃料から電力を作ることができるので賛成です。

例えば，以上のような文に対して形態素解析を行うと，以下のように分割される．

小さい/燃料/から/電力/を/作る/こと/が/できる/の/で/賛成/です/。

（スラッシュ/は区切り文字である）

本研究では，形態素解析に MeCab(ver.0.996)[*7]を利用する．MeCab では分割した単語に品詞情報などが付加される．例えば，先ほどの例文を MeCab にかけると，以下のような結果が得られる．なお，解析の辞書には UniDic(ver.2.1.2)[*8]を用いる．

```
小さい   形容詞, 一般,*,*, 形容詞, 連体形-一般, 小さい, 小さい, チイサイ, チーサイ
燃料     名詞, 普通名詞, 一般,*,*,*, 燃料, 燃料, ネンリョウ, ネンリョー
から     助詞, 格助詞,*,*,*,*, から, から, カラ, カラ
電力     名詞, 普通名詞, 一般,*,*,*, 電力, 電力, デンリョク, デンリョク
を       助詞, 格助詞,*,*,*,*, を, を, ヲ, オ
作る     動詞, 一般,*,*, 五段-ラ行, 連体形-一般, 作る, 作る, ツクル, ツクル
こと     名詞, 普通名詞, 一般,*,*,*, 事, こと, コト, コト
が       助詞, 格助詞,*,*,*,*, が, が, ガ, ガ
できる   動詞, 非自立可能,*,*, 上一段-カ行, 連体形-一般, 出来る, できる, デキル, デキル
の       助詞, 準体助詞,*,*,*,*, の, の, ノ, ノ
で       助動詞,*,*,*, 助動詞-ダ, 連用形-一般, だ, で, ダ, デ
賛成     名詞, 普通名詞, サ変可能,*,*,*, 賛成, 賛成, サンセイ, サンセー
です     助動詞,*,*,*, 助動詞-デス, 終止形-一般, です, です, デス, デス
。       補助記号, 句点,*,*,*,*,。,。,*,*
```

[*7] http://mecab.sourceforge.net/
[*8] http://sourceforge.jp/projects/unidic/

3.2 構文解析（係り受け解析）

日本語における構文解析（係り受け解析）とは，文中の文節の係り受け構造を発見することである．

本研究では，構文解析に CaboCha(ver.0.67)[*9]を利用する．前節で挙げた例文を CaboCha にかけると，以下のような係り受け構造が得られる．なお，解析の品詞体系モデルには MeCab と同様に UniDic を用いる．

```
小さい
└燃料から
  ｜電気を
  └┴作る
      └ことが
        └できるので
          └賛成です。
```

CaboCha では，以下のように係り受け構造と MeCab による形態素解析の結果を合わせて出力することができ，本研究では，この出力された情報を利用する．

```
* 0 1D 0/0 1.056797
小さい 形容詞, 一般,*,*, 形容詞, 連体形-一般, チイサイ, 小さい, 小さい, チーサイ, 小さい, チーサイ, 和,*,*,*,*
* 1 3D 0/1 1.057218
燃料 名詞, 普通名詞, 一般,*,*,*, ネンリョウ, 燃料, 燃料, ネンリョー, 燃料, ネンリョー, 漢,*,*,*,*
から 助詞, 格助詞,*,*,*,*, カラ, から, から, カラ, から, カラ, 和,*,*,*,*
* 2 3D 0/1 2.633533
電力 名詞, 普通名詞, 一般,*,*,*, デンリョク, 電力, 電力, デンリョク, 電力, デンリョク, 漢,*,*,*,*
を 助詞, 格助詞,*,*,*,*, ヲ, を, を, オ, を, オ, 和,*,*,*,*
* 3 4D 0/0 1.894854
作る 動詞, 一般,*,*, 五段-ラ行, 連体形-一般, ツクル, 作る, 作る, ツクル, 作る, ツクル, 和, ツ濁, 基本形,*,*
* 4 5D 0/1 1.954887
こと 名詞, 普通名詞, 一般,*,*,*, コト, 事, こと, コト, こと, コト, 和, コ濁, 基本形,*,*
が 助詞, 格助詞,*,*,*,*, ガ, が, が, ガ, が, ガ, 和,*,*,*,*
* 5 6D 0/2 1.954887
できる 動詞, 非自立可能,*,*, 上一段-カ行, 連体形-一般, デキル, 出来る, できる, デキル, できる, デキル, 和,*,*,*,*
の 助詞, 準体助詞,*,*,*,*, ノ, の, の, ノ, の, ノ, 和,*,*,*,*
で 助動詞,*,*,*, 助動詞-ダ, 連用形-一般, ダ, だ, で, デ, だ, ダ, 和,*,*,*,*
* 6 -1D 0/1 0.000000
賛成 名詞, 普通名詞, サ変可能,*,*,*, サンセイ, 賛成, 賛成, サンセー, 賛成, サンセー, 漢,*,*,*,*
です 助動詞,*,*,*, 助動詞-デス, 終止形-一般, デス, です, です, デス, です, デス, 和,*,*,*,*
。 補助記号, 句点,*,*,*,*,,。 ,。 ,,。 ,, 記号,*,*,*,*
```

[*9] http://code.google.com/p/cabocha/

3.3 日本語 WordNet

日本語 WordNet[*10]とは，日本語の概念辞書である．英語の WordNet [*11]がベースとなっており，WordNet には図 3.1 のように単語間の上位・下位概念関係が階層構造で記述されている．日本語 WordNet では，日本語の単語が英語 WordNet 内の英単語と紐づけて記述されており，構造としては英語 WordNet も日本語 WordNet も全く同じである．

個々の概念はそれぞれ「synset」という単位で表現され，それらが他の synset と結びつく形になっている．この階層構造（層の深さや各 synset に属する単語数など）を用いることで，単語間の概念類似度を計算することができる．

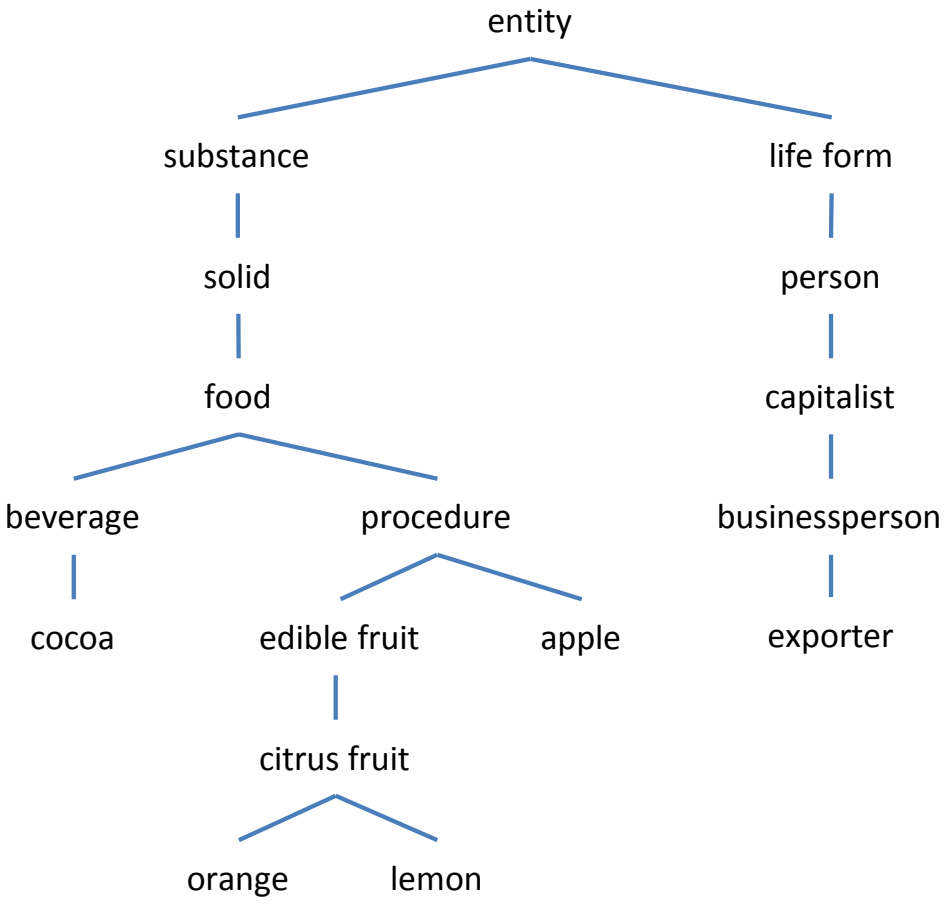

図 3.1 WordNet の階層構造例

[*10] http://nlpwww.nict.go.jp/wn-ja/
[*11] http://wordnet.princeton.edu/

3.4　潜在意味インデキシング

潜在意味インデキシング（Latent Semantic Indexing; LSI）[Deerwester 90] は，文書検索において頻繁に用いられる手法で，高次元の単語・文書行列を低次元の空間（行列）へ射影することにより，検索の精度や速度を改善することができると報告されている．

3.4.1　単語・文書行列と類似度

今，以下のような単語・文書行列 M があるとする．

$$M = \begin{array}{c} \\ w_1 \\ \vdots \\ w_i \\ \vdots \\ w_m \end{array} \begin{array}{c} \begin{array}{ccccc} d_1 & \cdots & d_j & \cdots & d_n \end{array} \\ \begin{pmatrix} f_{11} & \cdots & f_{1j} & \cdots & f_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ f_{i1} & \cdots & f_{ij} & \cdots & f_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ f_{m1} & \cdots & f_{mj} & \cdots & f_{mn} \end{pmatrix} \end{array}$$

なお，f_{ij} は単語 w_i が文書 d_j に出現した頻度を示す．

行列 M から，単語 w_i の特徴ベクトル $\boldsymbol{w}_i$ と文書 d_j の特徴ベクトル $\boldsymbol{d}_j$ は，それぞれ以下のように表現される．

$$\boldsymbol{w}_i = (f_{i1}, \cdots, f_{ij}, \cdots, f_{in})$$
$$\boldsymbol{d}_j = (f_{1j}, \cdots, f_{ij}, \cdots, f_{mj})^T$$

この特徴ベクトルを用いて単語どうしや文書どうしの類似度を計算することができる．例えば，次の 4 つの文書から，「**車**」と「**自動車**」という 2 つの単語の類似度を計算することを考える．

1. **大学には車で行きます。**
2. **大学には自動車で行きます。**
3. **大学には自転車で行きます。**
4. **調布には自転車で行きます。**

まず，これら 4 つの文書から自立語[*12]を抽出し，それらの出現頻度を保持するような単語・文書行列 M を得る．

[*12] 名詞や動詞など，それだけで意味を持つ単語

$$M = \begin{array}{c} \\ \textbf{大学} \\ \textbf{車} \\ \textbf{行く} \\ \textbf{自動車} \\ \textbf{自転車} \\ \textbf{調布} \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & d_2 & d_3 & d_4 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

このとき，「**車**」と「**自動車**」の特徴ベクトルは以下のように表される．（便宜的に「**車**」と「**自動車**」の特徴ベクトルをそれぞれ $\boldsymbol{w}_2, \boldsymbol{w}_4$ とする）

$$\boldsymbol{w}_2 = (1,\ 0,\ 0,\ 0)$$
$$\boldsymbol{w}_4 = (0,\ 1,\ 0,\ 0)$$

そして，コサイン類似度を用いると単語どうしの類似度は以下のように計算することができる．

$$\mathrm{sim}\,(\textbf{車}, \textbf{自動車}) = \cos(\boldsymbol{w}_2, \boldsymbol{w}_4) = \frac{\boldsymbol{w}_2 \cdot \boldsymbol{w}_4}{|\boldsymbol{w}_2||\boldsymbol{w}_4|} = \frac{0}{1 \times 1} = 0$$

しかし，上の計算結果からも分かる通り，このままでは 2 つの単語の類似度は 0 になってしまう．これは，2 つの単語が共通して出現する文書がないためである．この 2 つの単語以外にも，「**自動車**」と「**自転車**」なども同様に類似度が 0 になってしまう．

このような問題に対処する方法が，潜在意味インデキシング（LSI）である．LSI では，高次元の行列を低次元に次元圧縮し，単語の持つ意味や概念を考慮した意味空間を構築することで，類似度計算を行えるようにする．

3.4.2 次元圧縮

高次元空間の次元圧縮は，自然言語処理の分野だけでなく画像処理など多くの分野で用いられている．

特に自然言語処理の潜在意味インデキシング（LSI）における高次元空間の次元圧縮には，一般的に特異値分解（Singular Value Decomposition; SVD）が用いられる．

今，階数 r，$m \times n$ の行列 M に対する特異値分解は次のように定義される．

$$M = U\Sigma V^T \tag{3.1}$$

ここで，U は $m \times m$ の列直行行列，V は $n \times n$ の列直行行列である．また，Σ は式 (3.2) に示す

ような $m \times n$ の行列である．

$$\Sigma = \begin{pmatrix} S & O_{r,n-r} \\ O_{m-r,r} & O_{m-r,n-r} \end{pmatrix} \tag{3.2}$$

$$S = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_r), \quad (\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0) \tag{3.3}$$

行列 S は式 (3.3) のような対角行列であり，その要素 σ_i は行列 M の特異値と呼ばれる．

LSI による次元圧縮には 2 種類の方法がある．1 つは「もとの行列の次元数はそのままで階数を削減する」方法である．もう 1 つは「もとの行列の次元数自体を削減する」方法である．

階数の削減による次元圧縮

特異値分解により行列 M から得られた各行列（U, Σ, V）に対し，U と V の $k+1$ 列目（$k < r$）以降を削除した行列を U_k, V_k とし，Σ の $k+1$ 行目と $k+1$ 列目以降を削除した行列（式 (3.3) において対角要素，すなわち特異値を k 個まで取ったときの行列 S と同義）を Σ_k としたとき，式 (3.4) のように，これらの行列を掛けあわせることで行列 M を k 次元に近似することができる．

$$M \simeq M_k = U_k \Sigma_k V_k^T \tag{3.4}$$

近似された行列 M_k を「意味空間」と呼び，この行列の行（単語）ベクトルを見ることで，単語どうしの意味的な類似度を計算することができる．

次元数の削減による次元圧縮

式 (3.5) のように，行列 V の $k+1$ 列目（$k < r$）以降を削除した行列 V_k を使うことで，行列 M を k 次元に近似することができる．

$$M_k = M \left(V_k^T\right)^T = M V_k \tag{3.5}$$

また，式 (3.6) のように，行列 U の $k+1$ 列目（$k < r$）以降を削除した行列 U_k と，行列 Σ の特異値を k 個まで取った行列 Σ_k を掛けあわせることでも，k 次元に近似することができる．

$$M_k = U_k \Sigma_k \tag{3.6}$$

式 (3.5) と式 (3.6) における M_k は，それぞれ異なるものを示しているように見えるが，式 (3.4) を用いることで同じものだと分かる．（式 (3.7)）

$$
\begin{aligned}
MV_k &\simeq M_k V_k \\
&= \left(U_k \Sigma_k V_k^T\right) V_k \quad (\because \text{式 (3.4)}) \\
&= U_k \Sigma_k I_k \quad \left(\because V^T V = I \iff V_k^T V_k = I_k\right) \\
&= U_k \Sigma_k
\end{aligned} \tag{3.7}
$$

前節で例示した単語・文書行列 M を階数の削減により次元圧縮を行うことを考える．例えば，次元圧縮後の次元数 k を $k = 2$ としたとき，行列 M は次のように近似される．

$$
\begin{aligned}
M_2 &= U_2 \Sigma_2 V_2^T \\
&= \begin{pmatrix} -0.55237 & -0.44178 \\ -0.17714 & -0.27430 \\ -0.70034 & 0.07200 \\ -0.17714 & -0.27430 \\ -0.34604 & 0.62061 \\ -0.14797 & 0.51379 \end{pmatrix} \begin{pmatrix} 2.84104 & 0. \\ 0. & 1.53233 \end{pmatrix} \begin{pmatrix} -0.50328 & -0.42033 \\ -0.50328 & -0.42033 \\ -0.56273 & 0.16369 \\ -0.42039 & 0.78730 \end{pmatrix}^T \\
&= \begin{pmatrix} 1.07436 & 1.07436 & 0.77229 & 0.12675 \\ 0.42997 & 0.42997 & 0.21441 & -0.11934 \\ 0.95501 & 0.95501 & 1.13774 & 0.92332 \\ 0.42997 & 0.42997 & 0.21441 & -0.11934 \\ 0.09506 & 0.09506 & 0.70891 & 1.16202 \\ -0.11934 & -0.11934 & 0.36544 & 0.79657 \end{pmatrix}
\end{aligned}
$$

もとの行列 M と比べて，ゼロ要素がなくなったことが分かる．このとき，2 次元に次元圧縮した「**車**」と「**自動車**」の特徴ベクトル $\boldsymbol{w}_2^{(2)}, \boldsymbol{w}_4^{(2)}$ は以下のように表される．

$$
\begin{aligned}
\boldsymbol{w}_2^{(2)} &= (0.42997, 0.42997, 0.21441, -0.11934) \\
\boldsymbol{w}_4^{(2)} &= (0.42997, 0.42997, 0.21441, -0.11934)
\end{aligned}
$$

これらの特徴ベクトルから単語どうしの類似度 $\mathrm{sim}_{k=2}$ をコサイン類似度により計算すると，

$$
\begin{aligned}
\mathrm{sim}_{k=2}(\textbf{車}, \textbf{自動車}) &= \cos\left(\boldsymbol{w}_2^{(2)}, \boldsymbol{w}_4^{(2)}\right) \\
&= \frac{\boldsymbol{w}_2^{(2)} \cdot \boldsymbol{w}_4^{(2)}}{\left|\boldsymbol{w}_2^{(2)}\right| \left|\boldsymbol{w}_4^{(2)}\right|} = \frac{0.42997}{0.65572 \times 0.65572} = 1.0
\end{aligned}
$$

このように，次元圧縮を行うことで単語どうしの類似度をより適切に計算することができる．

ここまでは，単語どうしの類似度計算という観点から LSI の説明を述べてきたが，文書どうしの類似度も同様に計算することができる．

3.5　潜在的ディリクレ配分法

潜在的ディリクレ配分法（Latent Dirichlet Allocation; LDA）とは，Blei ら [Blei 03] によって提案されたトピックモデルであり，このモデルは「文書は複数の潜在的なトピックからなる単語で構成されている」という仮定に基づいている．

LDA では，文書におけるトピックの出現確率と，各トピックにおける単語の出現確率を多項分布で仮定し，仮定した多項分布にディリクレ分布を用いることで，トピックの推定を可能にしている．

LDA による文書のトピックモデルの生成過程は次の通りである．

1. 文書 d におけるトピックの出現確率分布（多項分布）θ_d を Dirichlet 事前分布から選択する．

$$\theta_d \sim \mathrm{Dirichlet}(\alpha)$$

   Dirichlet 分布とは，「ある n 個の事象について i 番目の事象が $\alpha_i - 1$ 回観測された場合に，その事象の生起確率が x_i である」ということを示した確率分布のことである．つまり，大まかに言ってしまえば，確率分布の確率分布である．

2. 文書 d に含まれる単語 w_i について

   (a) 多項分布 θ_d から単語 w_i のトピック z_i を選択する．

$$z_i \sim \mathrm{Mlutinomial}(\theta_d)$$

   (b) トピック z_i における単語の出現確率分布（多項分布）から，単語 w_i を選択する．

$$w_i \sim p\left(w_i | z_i, \beta\right)$$

なお，α は Dirichlet 事前分布のパラメータであり，β はトピックモデルのパラメータである．このとき，LDA のグラフィカルモデルは図 3.2 のようになる．

以上のことから，文書 d の生成確率は次のように表される．なお，N_d は文書 d に含まれる単語の数を示す．

$$p\left(d|\alpha,\beta\right) = \int p\left(\theta_d|\alpha\right)\left(\prod_{i=1}^{N_d}\sum_{z_i} p\left(z_i|\theta_d\right) p\left(w_i|z_i,\beta\right)\right) d\theta_d \tag{3.8}$$

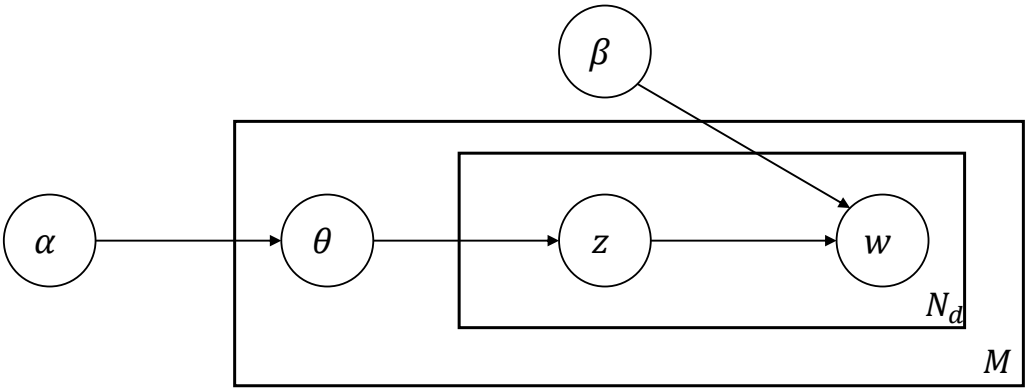

図 3.2　LDA のグラフィカルモデル

また，M 個の文書からなる文書集合 $\boldsymbol{D}$ の生成確率は式 (3.9) のようになる．

$$p\left(\boldsymbol{D}|\alpha,\beta\right)=\prod_{d=1}^{M}p\left(d|\alpha,\beta\right) \tag{3.9}$$

LDA では式 (3.10) のように，式 (3.9) の対数を取ったものを最大化するようなパラメータ $\theta_d, \boldsymbol{z}$ の推定を行い，トピックの選択（推定）を行う．

$$\begin{aligned}\log p\left(\boldsymbol{D}|\alpha,\beta\right)&=\log\left(\prod_{d=1}^{M}p\left(d|\alpha,\beta\right)\right)\\&=\sum_{d=1}^{M}\log p\left(d|\alpha,\beta\right)\end{aligned} \tag{3.10}$$

文書 d に含まれる単語集合 $\boldsymbol{w}$（$|\boldsymbol{w}|=N_d$）のトピック推定には $p\left(\theta_d,\boldsymbol{z}|\boldsymbol{w},\alpha,\beta\right)$ を求めることになる．しかし，式 (3.11),(3.12) のように，$p\left(\theta_d,\boldsymbol{z}|\boldsymbol{w},\alpha,\beta\right)$ の計算には，$p\left(\boldsymbol{w}|\alpha,\beta\right)$ を計算しなければならず，θ_d の積分や z_i の和を直接計算することはできない．

$$p\left(\theta_d,\boldsymbol{z}|\boldsymbol{w},\alpha,\beta\right)=\frac{p\left(\theta_d,\boldsymbol{z},\boldsymbol{w}|\alpha,\beta\right)}{p\left(\boldsymbol{w}|\alpha,\beta\right)} \tag{3.11}$$

$$p\left(\boldsymbol{w}|\alpha,\beta\right)=\int p(\theta_d|\alpha)\left(\prod_{i=1}^{N_d}\sum_{z_i}p(z_i|\theta_d)p(w_i|z_i,\beta)\right)d\theta_d \tag{3.12}$$

この問題に対処するため Blei らは変分ベイズ法を用いて $p\left(\theta_d,\boldsymbol{z}|\boldsymbol{w},\alpha,\beta\right)$ を別の確率分布に近似し，EM アルゴリズムを用いてパラメータを推定している．（詳細は文献 [Blei 03] を参照.）

本研究では，トピック推定における計算の簡略化を図るため，パラメータの推定に Collapsed Gibbs Sampling を用いる．また，トピックにおける単語の出現確率分布を予め Dirichlet 事前分布から選択する．そのため，先ほどの文書トピックモデルの生成過程における 2.(b) では，単語

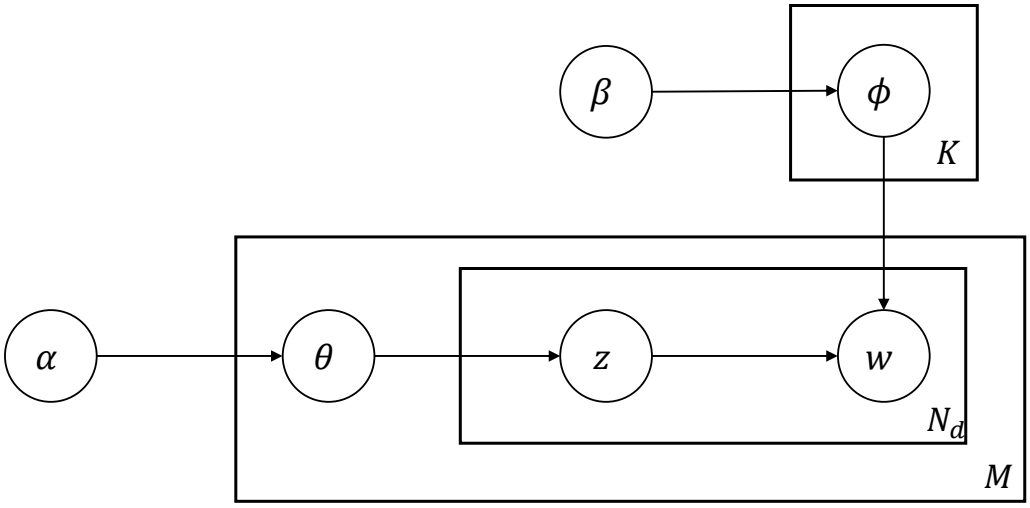

図 3.3　Smoothed LDA のグラフィカルモデル

w_i はトピック z_i における単語出現確率分布 ϕ_{z_i} から選択することになる．

$$\phi_{z_i} \sim \text{Dirichlet}(\beta)$$
$$w_i \sim \text{Multinomial}(\phi_{z_i})$$

このことから，パラメータ β も α と同様に Dirichlet 事前分布のパラメータとなり，このような確率分布の生成を行った LDA を特に Smoothed LDA と呼ぶ．（Smoothed LDA のグラフィカルモデルを図 3.3 に示す．なお，図中の K はトピック数を指す．）

3.5.1　Collapsed Gibbs Sampling

Collapsed Gibbs Sampling（CGS）は，直接計算が困難な確率分布の代わりにそれを近似するようなサンプル（データ）列を生成する手法である．具体的には，文書中の各単語に対して，予めランダムなトピックを割り当てておき，各単語に関してトピックを逐次更新していくという流れを取る．この更新を繰り返すことで，尤もらしい θ と ϕ を得ることができる．

ある文書 d 中の単語 w_i に対する CGS の更新式（近似式）は次の通りである．

$$p(z_i = t | w_i = m, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) \propto \frac{C_{d,t} + \alpha}{\sum_t C_{d,t} + K\alpha} \frac{C_{m,t} + \beta}{\sum_m C_{m,t} + V\beta} \tag{3.13}$$

なお，$\boldsymbol{z}_{-i}$ はトピック集合 $\boldsymbol{z}$ からトピック z_i を除いたもの，$\boldsymbol{w}_{-i}$ は単語集合 $\boldsymbol{w}$ から単語 w_i を除いたものを指す．また，$C_{d,j}$ は文書 d がトピック t に割り当てられた回数，$C_{m,t}$ は単語 m がトピック t に割り当てられた回数，V は全単語数を指す．

CGS による θ と ϕ の推定結果は次のようになる．なお，$\theta_{d,t}$ は文書 d におけるトピック t の生

起確率であり，$\phi_{m,t}$ はトピック t における単語 m の生起確率である．

$$\theta_{d,t} = \frac{C_{d,t} + \alpha}{\sum_t C_{d,t} + T\alpha} \tag{3.14}$$

$$\phi_{m,t} = \frac{C_{m,t} + \beta}{\sum_m C_{m,t} + V\beta} \tag{3.15}$$

LDA では，この推定されたパラメータ θ と ϕ を用いて，トピックに基づく文書クラスタリングや単語クラスタリングを行うことができる．

3.6 クラスタリング

クラスタリングとは，データ解析手法の 1 つであり，あるデータ集合を事前知識（予め与えられた分類基準など）なしに自動的に分類する教師なし機械学習手法のことである．

一般的に「分類」と呼ばれるものは，自然言語処理分野においては「分類（classification)」と「クラスタリング（clustering)」に分けられる．前者の「分類（classification)」は教師あり機械学習手法と呼ばれ，予め与えられた教師（正解）データをもとに分類基準を決め，その分類基準を用いて教師データとは別のデータを分類するというものである．代表的な教師あり機械学習手法として，決定木や Naive Bayes，SVM（Support Vector Machine)，k-近傍法などがある．一方，後者の「クラスタリング（clustering)」は，教師なし機械学習手法と呼ばれ，教師データなしに自動的に分類基準を決めていき，データ集合を任意のグループに分割する（クラスタを生成する）というものである．クラスタリング手法は，非階層型クラスタリング手法と階層型クラスタリング手法の大きく 2 つに分けられ，それぞれの代表的なものとして k-means 法や Ward 法などがある．

「分類」では予め分類項目が設定されていることから，どのデータがどの項目に分類されたかが分かりやすいが，逆に分類項目を限定してしまったり教師データの作成コストが非常に高いというデメリットがある．一方で「クラスタリング」では，教師データが必要ないことや，分類項目を限定することがないため，柔軟な分類を行うことができる．「クラスタリング」では，生成されたクラスタがどのような特徴を表しているか分からないというデメリットもあるが，本研究では「予め観点を用意することなく，意見を観点ごとに分類する」という目的から「クラスタリング」を行う．

本研究では，クラスタリングの中でも階層型クラスタリング手法を利用する．なお，階層型クラスタリング手法のアルゴリズムは以下の通りである．

1. 各要素を，それぞれ要素数 1 のクラスタとする．
2. クラスタどうしのクラスタ間距離を求める．
3. クラスタ間距離の最も小さいクラスタどうしを併合する．
4. クラスタリングの終了条件を満たしていれば終了する．満たしていなければ，手順 2 に戻る．

手順 2 のクラスタ間距離の計算には様々な手法が提案されており，本研究ではその中でも Ward 法を用いる．

4 提案手法

4.1 提案手法の構想

従来用いられてきた Bag of Words（BoW）アプローチに基づくクラスタリング手法は，トピックを表す内容語（話題語）の共通性・類似性に基づいた文書（クラスタ）間の類似度を計算するため，結果としてクラスタリング対象の文書集合に含まれるトピックを基準とした分類が行われる．したがって，本研究で対象とするような，ある特定のトピック（時事問題）に関する意見集合に対して BoW アプローチに基づくクラスタリングを行うと，その結果はそのトピックのサブトピック（トピックの下位概念）に基づく分類になりやすいと考えることができる．

一方，あるトピックに関する意見の観点そのものを表す語句は，意見中には陽に出現しにくいと考えられる．例えば，原発（問題）に対する意見の観点として，「安全性」や「エネルギー政策」,「健康への影響」などが考えられる．これらは，原発（原子力発電所）の下位概念というよりは，原発を議論するにあたっての視点であり，意見中に陽に述べられることは多くはない．よって，BoW アプローチに基づくクラスタリング手法をそのまま意見集合に適用したとしても，これらの観点に基づく分類が行われる可能性は低い．

本研究では，このような観点の違いは，内容語（名詞）そのものの違いではなく，その使われ方に反映されていると考える．より具体的には，名詞と動詞の係り受け関係，すなわち述語・項構造の違いに反映されると仮定する．例えば，表 4.1 は原発（問題）に関する観点の異なる 2 つの意見の例を示している．これらの 2 つの意見は「**燃料**」という共通の名詞を含んでいるが，それぞれ「**作る**」と「**消費する**」という異なる動詞に係ることから，燃料のどのような側面が述べられているかが異なっている．「発電技術」という観点からはエネルギー源としての燃料が話題となっているのに対して，「発電コスト」という観点からは燃料の消費が話題となっており，このような

表 4.1 トピック「原発」に対する観点の異なる意見の例

観点	意見
発電技術	**原発の稼働には賛成です。原子力発電だと小さい 燃料 から電力を作ることができるなんて知りませんでした。少資源の日本にとっては 消費 の少ない原発の方がいい気がします。**
発電コスト	**火力発電だと電力の生成に多くの 燃料 を 消費する ことになります。円安で輸入費もかかりますし、原発も選択肢の 1 つだと思います。**

観点の違いが「**燃料**」という名詞の用いられ方，すなわち「**燃料**」が係る動詞の違いに反映されていると言える．そこで，文節の係り受け関係から名詞 N と動詞 V の名詞・動詞ペア $\langle N, V \rangle$ を抽出し，名詞 N どうしの類似度ではなく名詞・動詞ペアどうしの類似度に基づいて意見どうしの類似度を計算することで，観点の差異を考慮したクラスタリングが実現できると考える．

さらに，名詞と動詞の係り受け関係を用いることによって，サ変可能名詞[*13]のような単語が文中でどの品詞で用いられているかを考慮した類似度計算が可能となる．例えば，表 4.1 の 2 つの意見には「**消費**」という語が含まれているが，この語は前者では名詞の，後者では動詞の機能を担っている．しかし，一般的に BoW アプローチに基づく文書間類似度の計算手法では，このような単語が文中で名詞として機能しているのか，動詞として機能しているのかを区別せずに利用している．提案手法では，名詞と動詞の係り受け関係を考慮することで，単語が文中でどの品詞で用いられているかを同定し，名詞 N どうし，動詞 V どうしの類似度を適切に計算できることが期待できる．

[*13] 「消費」や「開発」のように名詞の直後に動詞の「する」が付くことで動詞化するもの

4.2 概要

本研究で対象とする意見は，1 つ以上の文から構成される短い文章である．提案手法では，ある特定のトピック（時事問題）に関する意見の集合に対して，1 つの意見に単一の観点が付与されると仮定して，排他的なクラスタリングを行う．

本研究で提案するクラスタリング手法の手順を以下に示す．

1. クラスタリングの対象となる意見集合の各意見に対して，そこに含まれるすべての名詞・動詞ペアを抽出する．
2. 各意見 o_i をそこに含まれる名詞・動詞ペア集合 $\boldsymbol{P}(o_i)$ で表現し，名詞・動詞ペア集合間の類似度として，意見どうしの類似度を計算する．
3. 手順 2 で計算される意見どうしの類似度を用いて，Ward 法による階層型クラスタリングを行う．

以降の 4.3 節と 4.4 節では，それぞれ手順 1 と手順 2 の詳細を述べる．

4.3　名詞・動詞ペアの抽出

各意見に対して，係り受け解析を行い，文節の係り受け関係から動詞の機能を担う語の抽出と名詞の機能を担う語の抽出を行い，名詞・動詞ペア $\langle N, V \rangle$ を抽出する．

4.3.1　動詞 V の抽出

原則として，ある文節中で形態素解析により動詞と判断されたものをそのまま動詞として抽出するが，文節中に非自立語扱いの動詞（「する」や「ある」など，それ自体で意味を持たない動詞）しか存在しない場合は，その文節に出現する名詞を動詞として抽出する．ただし，名詞が出現しない場合には，非自立語扱いの動詞をそのまま動詞として抽出する．

例えば，「**代替エネルギーを開発する**」という文では，「**開発する**」という文節で「**する**」という非自立語扱いの動詞が存在することから，名詞である「**開発**」が動詞として抽出される．

4.3.2　名詞 N の抽出

4.3.1 節で抽出された動詞を含む文節に係る文節 P_i に含まれる名詞 N を抽出して，名詞・動詞ペアを生成する．

ただし，自立語 W を含む文節 P_j が文節 P_i に係るとき，W が以下に示す条件を満たせば，名詞 N の修飾語とみなして複合名詞 $\langle W, N \rangle$ を抽出する．

■自立語 W が名詞の場合：

W を含む文節 P_j が，助動詞または助詞の「の」を伴って文節 P_i に係るとき，W を修飾語とする．

■自立語 W が動詞の場合：

W を含む文節 P_j の終端が W であるとき，W を修飾語とする．

■自立語 W が形容詞の場合：

そのまま，W を修飾語とする．

例えば，「**これからの自然の脅威に備える。**」という文からは，図 4.1 のような係り受け構造が得られる．なお，形態素解析上，名詞と判定された単語を N() で，動詞と判定された単語を V() で囲っている．

このとき，名詞「**自然**」が助詞「**の**」を伴って名詞「**脅威**」に係ることから，〈 **自然**, **脅威** 〉という複合名詞が抽出される．また，名詞「**脅威**」は動詞「**備える**」に係ることから，最終的に

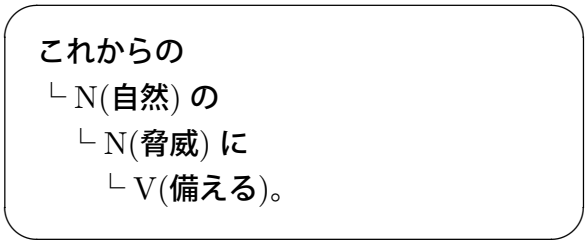

図 4.1　係り受け構造の例 1

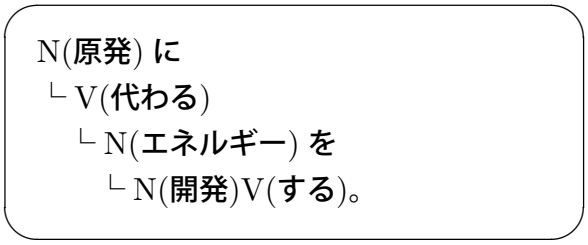

図 4.2　係り受け構造の例 2

⟨⟨ **自然**, **脅威** ⟩, **備える** ⟩ という名詞・動詞ペアが抽出される．

また，「**原発に代わるエネルギーを開発する。**」という文からは，図 4.2 のような係り受け構造が得られる．このとき，まず，名詞「**原発**」が動詞「**代わる**」に係ることから，⟨ **原発**, **代わる** ⟩ という名詞・動詞ペアが抽出される．また，動詞「**代わる**」はそれ自身が文節となり，名詞「**エネルギー**」に係ることから，⟨ **代わる**, **エネルギー** ⟩ という複合名詞が抽出される．加えて，名詞「**エネルギー**」が係る文節には名詞「**開発**」と動詞「**する**」が含まれるが，4.3.1 節で述べたように，「**する**」は非自立語扱いの動詞であるので，「**開発**」が動詞として抽出され，⟨⟨ **代わる**, **エネルギー** ⟩, **開発** ⟩ という名詞・動詞ペアも抽出される．

4.4 意見間の類似度の計算

意見 o_x, o_y に含まれる名詞・動詞ペアの集合をそれぞれ $\boldsymbol{P}_x = \{\langle N_i, V_i\rangle_i^x\}, \boldsymbol{P}_y = \{\langle N_j, V_j\rangle_j^y\}$ とし，意見 o_x, o_y の類似度 $\mathrm{sim}_o(o_x, o_y)$ を式 (4.1) で定義する．

$$\mathrm{sim}_o(o_x, o_y) = \frac{\mathrm{nvSim}_x + \mathrm{nvSim}_y}{|\boldsymbol{P}_x| + |\boldsymbol{P}_y|} \tag{4.1}$$

$$\mathrm{nvSim}_x = \sum_{i=1}^{|\boldsymbol{P}_x|} \max_j \left[\mathrm{sim}_{nv}\left(\langle N_i, V_i\rangle_i^x, \langle N_j, V_j\rangle_j^y\right)\right]$$

$$\mathrm{nvSim}_y = \sum_{j=1}^{|\boldsymbol{P}_y|} \max_i \left[\mathrm{sim}_{nv}\left(\langle N_i, V_i\rangle_i^x, \langle N_j, V_j\rangle_j^y\right)\right]$$

上式において，$\mathrm{sim}_{nv}\left(\langle N_i, V_i\rangle_i^x, \langle N_j, V_j\rangle_j^y\right)$ は 2 つの名詞・動詞ペア $\langle N_i, V_i\rangle_i^x$ と $\langle N_j, V_j\rangle_j^y$ の類似度を表している．したがって，nvSim_x は，意見 o_x の各名詞・動詞ペア $\langle N_i, V_i\rangle_i^x$ に対する意見 o_y の名詞・動詞ペア集合 $\boldsymbol{P}_y$ との最大類似度の和であり，nvSim_y は逆に，意見 o_y の各名詞・動詞ペア $\langle N_j, V_j\rangle_j^y$ に対する意見 o_x の名詞・動詞ペア集合 $\boldsymbol{P}_x$ との最大類似度の和である．

以降の節では，式 (4.1) の計算に必要な名詞・動詞ペア間の類似度 sim_{nv} の計算方法について述べる．なお，sim_{nv} は名詞どうしや動詞どうしの類似度を用いて計算するため，まず 4.5 節で単語間の類似度の計算方法について述べた後に，4.6 節で名詞・動詞ペア間の類似度の計算方法について述べる．

4.5 単語間の類似度計算

単語どうしの類似度計算には，日本語 WordNet を用いた類似度と潜在意味インデキシング（以下，LSI）[Deerwester 90] により構築した意味空間を用いた類似度を利用する．

単語 w_i, w_j の日本語 WordNet を用いた類似度を jwn_w，LSI を用いた類似度を lsi_w としたとき，w_i と w_j の類似度 $\mathrm{sim}_w(w_i, w_j)$ を式 (4.2) で定義する．

$$\mathrm{sim}_w(w_i, w_j) = \alpha \times \mathrm{jwn}_w(w_i, w_j) + (1 - \alpha) \times \mathrm{lsi}_w(w_i, w_j) \tag{4.2}$$

なお，α $(0 \leq \alpha \leq 1)$ は，jwn_w と lsi_w のどちらの類似度の影響を強くするかを示すパラメータであり，その値が大きいほど日本語 WordNet を用いた類似度を重視することになる．ただし，w_i, w_j のどちらかが日本語 WordNet に存在しない場合は，$\alpha = 0$ とする．

4.5.1 日本語 WordNet を用いた類似度

日本語 WordNet を用いた単語間の概念類似度は，Resnik の手法 [Resnik 95] を用いて計算する．Resnik の手法では，単語 w_i と w_j の概念類似度 jwn_w を式 (4.3) のように定義している．

$$\mathrm{jwn}_w(w_i, w_j) = \max_{\substack{c_k \in S_1(w_i) \\ c_l \in S_1(w_j)}} [\mathrm{sim}_c(c_k, c_l)] \tag{4.3}$$

$S_1(w_i), S_1(w_j)$ は，それぞれ単語 w_i, w_j を含む概念（synset）の集合を指す．このとき，概念 c_k と c_l の類似度 $\mathrm{sim}_c(c_l, c_k)$ は式 (4.4) より計算される．

$$\mathrm{sim}_c(c_k, c_l) = \max_{c \in S_2(c_k, c_l)} [-\log p(c)] \tag{4.4}$$

式 (4.4) における $S_2(c_k, c_l)$ は，概念 c_k と c_l に共通する上位概念の集合を指す．なお $p(c)$ は，概念 c のすべての下位概念の数を全概念数で割った値を求める関数である．ここでの全概念数とは，WordNet に登録されている全概念数 117659 のことである．

ただし，式 (4.4) のままでは類似度の最大値が 1 にならないことから，本研究では全概念数 N で正規化した式 (4.5) を用いる．

$$\mathrm{sim}_c(c_k, c_l) = \max_{c \in S_2(c_k, c_l)} \left[-\frac{\log p(c)}{\log N} \right] \tag{4.5}$$

4.5.2 LSI を用いた類似度

単語 w_i と w_j の LSI により構築した意味空間を用いた類似度 lsi_w をコサイン類似度を利用して式 (4.6) のように定義する．

$$\begin{aligned}\mathrm{lsi}_w(w_i, w_j) &= \frac{1 + \cos\left(\boldsymbol{u}_i^{(d)}, \boldsymbol{u}^{(d)_j}\right)}{2} \\ &= \frac{1}{2}\left(1 + \frac{\boldsymbol{u}_i^{(d)} \cdot \boldsymbol{u}_j^{(d)}}{\left|\boldsymbol{u}_i^{(d)}\right| \times \left|\boldsymbol{u}_i^{(d)}\right|}\right)\end{aligned} \tag{4.6}$$

$\boldsymbol{u}_i^{(d)}, \boldsymbol{u}_j^{(d)}$ は，クラスタリングの対象とするすべての意見に含まれる自立語の出現頻度を要素とした単語・文書行列に対して，特異値分解を用いて行列の次元数を d に次元圧縮を施した後の単語 w_i, w_j の特徴ベクトルを示している．また，コサイン類似度は 2 つのベクトルがなす角度のコサインを求めることに相当するので，その値が取る範囲は-1〜1 となる．そのため，式 (4.6) では単語間の類似度が 0〜1 の範囲の値を取るようにスケール調整を行っている．

4.6 名詞・動詞ペア間の類似度計算

2 つの名詞・動詞ペア $\langle N_i, V_i\rangle_i, \langle N_j, V_j\rangle_j$ 間の類似度 sim_{nv} を名詞 N_i, N_j の類似度 sim_n と動詞 V_i, V_j の類似度 sim_v から式 (4.7) で定義する．

$$\mathrm{sim}_{nv}\left(\langle N_i, V_i\rangle_i, \langle N_j, V_j\rangle_j\right) = \mathrm{sim}_n + \left((1-\lambda) + \lambda(\mathrm{sim}_n)^2\right) \times \mathrm{sim}_v \tag{4.7}$$

式 (4.7) は，sim_n と係数付きの sim_v の和を取る形になっている．sim_{nv} の計算式を式 (4.7) とした理由は，sim_n が小さければ sim_v の大小に関わらず 2 つの名詞・動詞ペア $\langle N, V\rangle$ が異なる内容を表す可能性が高く，sim_{nv} を小さくする必要があると考えたからである．そのため，sim_n が大きくなるほど sim_v が sim_{nv} に与える影響が大きくなるように，sim_v の係数が設定されている．λ はその影響度合いを示すパラメータであり，その値が大きくなるほど sim_n と sim_v がより連動した sim_{nv} が計算される．

図 4.3 は $\lambda = 2/3$ における式 (4.7) を表したグラフであり，先述した通り，sim_n が大きくなるほど sim_v が sim_{nv} に与える影響が大きくなっている（グラフの傾きが大きくなっている）．

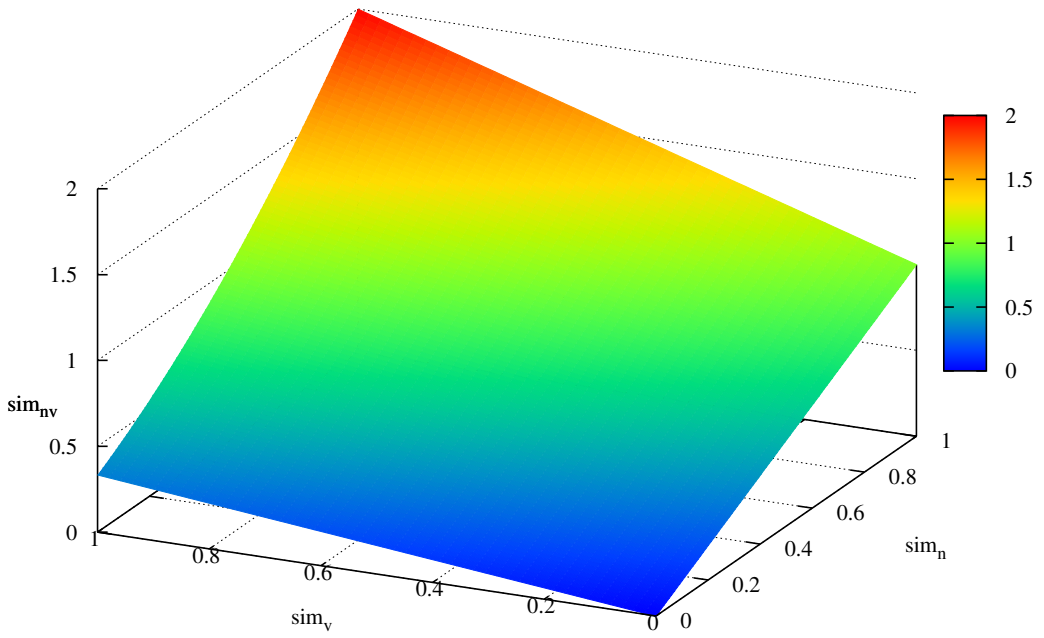

図 4.3 $\lambda = 2/3$ における式 (4.7) の 3 次元グラフ

4.6.1 名詞 N どうしの類似度

名詞 N_i, N_j 間の類似度 sim_n は，N_i, N_j それぞれが単一の名詞である場合と修飾語を含む複合名詞である場合とで計算方法が異なる．

■ **N_i と N_j が単一名詞の場合：** 4.5 節で定義した式 (4.2) で計算する．

■ **N_i と N_j の片方のみが複合名詞の場合：**

N_i が複合名詞 $\langle N_{i,1}, N_{i,2}\rangle$（すなわち，$N_j$ は単一名詞）とすると，式 (4.8) のように N_i に含まれる修飾語 $N_{i,1}$ と N_j 間，被修飾語（主辞名詞）$N_{i,2}$ と N_j 間の類似度を式 (4.2) で計算し，パラメータ β $(0 \leq \beta \leq 1)$ を用いて和を取る．

$$\mathrm{sim}_n(N_i, N_j) = \beta \times \mathrm{sim}_w(N_{i,1}, N_j) + (1-\beta) \times \mathrm{sim}_w(N_{i,2}, N_j) \tag{4.8}$$

■ **N_i と N_j の両方が複合名詞の場合：**

複合名詞 $N_i = \langle N_{i,1}, N_{i,2}\rangle$ と $N_j = \langle N_{j,1}, N_{j,2}\rangle$ に対して，式 (4.9) のように，両複合名詞に含まれる修飾語どうし，被修飾語どうしの類似度を式 (4.2) で計算し，式 (4.8) と同じパラメータ β を用いて和を取る．

$$\mathrm{sim}_n(N_i, N_j) = \beta \times \mathrm{sim}_w(N_{i,1}, N_{j,1}) + (1-\beta) \times \mathrm{sim}_w(N_{i,2}, N_{j,2}) \tag{4.9}$$

なお，式 (4.8) と式 (4.9) に共通するパラメータ β は，修飾語に基づく類似度が全体の類似度に与える影響の度合いを示しており，その値が大きいほど修飾語による類似度の影響が強くなる．したがって，$\beta = 0$ とすると，修飾語を無視した主辞名詞のみの類似度を求めることになる．

4.6.2 動詞 V どうしの類似度

動詞 V_i, V_j 間の類似度 sim_v は，式 (4.2) を用いて計算する．

4.7 クラスタリング

意見のクラスタリングには，階層型クラスタリング手法である Ward 法を用いる．なお，初期状態（各クラスタが各意見にあたる場合）のクラスタ間距離は，意見どうしの非類似度（距離）となる．例えば，初期クラスタ C_x, C_y のクラスタ間距離は，意見 o_x, o_y の非類似度に相当するので，以下のように計算される．

$$d(C_x, C_y) = d(o_x, o_y) = 2 - \mathrm{sim}_o(o_x, o_y) \tag{4.10}$$

ここで，$2 - \mathrm{sim}_o(o_x, o_y)$ とした理由は，意見どうしの最大類似度が 2（正確には名詞・動詞ペアどうしの最大類似度が 2）だからであり，最大類似度からその類似度を引くことで非類似度となる．

4.7.1 Ward 法

任意のクラスタ C_p と C_q の距離 $d(C_p, C_q)$ は，Ward 法では以下のように定義される．

$$d(C_p, C_q) = E(C_p \cup C_q) - E(C_p) - E(C_q) \tag{4.11}$$

なお，$E(C_i)$ は次を満たす関数であり，$\boldsymbol{c}_i$ はクラスタ C_i の重心ベクトルを指す．

$$E(C_i) = \sum_{\boldsymbol{x} \in C_i} (\boldsymbol{x} - \boldsymbol{c}_i)^2 \tag{4.12}$$

このように，クラスタ間の距離を計算する場合は，各クラスタ（または初期状態の各要素）がベクトル空間で表現されている必要がある．しかし，提案手法においては各意見をベクトル空間で表現することができないため，式 (4.11) を用いてクラスタ間の距離を計算することができない．そこで，提案手法では Lance-Williams の更新式 [Lance 67] を用いてクラスタ間の距離を計算する．

あるクラスタ C_p がクラスタ C_{pa}, C_{pb} が併合してできたものであるとき，クラスタ C_p と C_q のクラスタ間距離 $d(C_p, C_q)$ は Lance-Williams の更新式により以下のように計算される．

$$d(C_p, C_q) = \frac{n_{pa} + n_q}{n_p + n_q} d(C_{pa}, C_q) + \frac{n_{pb} + n_q}{n_p + n_q} d(C_{pb}, C_q) - \frac{n_q}{n_p + n_q} d(C_{pa}, C_{pb}) \tag{4.13}$$

なお，n_i はクラスタ C_i に含まれる要素数である．

Lance-Williams の更新式を用いることで，各意見をベクトル空間で表現しなくとも，あらかじめすべての意見どうしの距離を計算しておくことで直接クラスタリングすることができる．

5　評価実験

5.1　実験材料

評価実験には，あるトピックに関する意見を紹介する Web サイトや，ニュース記事やコラム（エッセイ）に対してユーザがコメントができる Web サイト[*14] に掲載されている意見を用い，実験者が予め選択した表 5.1 に示す 4 つのトピックに関する意見の中から 40 件ずつランダムに取得した.

意見は 1 文以上から構成されるもので，4 つのトピックの意見全体における 1 意見あたりの平均文数は 4.49 文であった.

表 5.1　実験に用いた意見のトピックと各トピックにおける平均文字数・文数

トピック	平均文字数	平均文数
原発	132	4.02
TPP	180	5.28
STAP 細胞	148	4.48
人口問題	133	4.18
全体平均	148	4.49

[*14] http://blogos.com

5.2 実験手順

評価実験の手順は，以下の通りである．

1. 各トピックの意見集合に対して，3 人の被験者により人手でそれぞれ観点ごとに意見がまとまるように分類を行ってもらうことで，各トピック 3 種類ずつ正解データを用意した．
   (a) まず，各意見ごとに，その意見が示す観点を列挙（付与）してもらった．この際，複数の観点を示すと判断された意見については，観点を複数付与してもらった．
   (b) グループ間で意見が重複しないよう，似た観点を示す意見ごとにグループを作ってもらい，最終的にそのグループとして尤もらしい観点を決めてもらった．なお，複数の観点を示す意見については，被験者の判断により，その意見に最もふさわしい（その意見で最も主張したいと思われる）観点を採用し，適宜グループを作ってもらった．
2. 各トピックごとに，人手による分類結果と同じ観点の数で，意見集合に対して提案手法を用いてクラスタリングを行った．（クラスタリングの終了条件を「クラスタ数が人手による分類結果と同じ観点の数になったとき」に設定した．）
3. 人手により生成された観点のグループ（以降，正解クラスタ群と呼ぶ）と提案手法により生成されたクラスタ群の近さを評価指標として，クラスタリング精度を計算した．（評価指標については次節で説明する．）

5.3 評価指標

提案手法により生成されたクラスタ群と人手により生成された正解クラスタ群がどの程度近いかの指標として，再現率と適合率からなる F 値を用いて評価を行った．F 値の計算は折原ら [折原 08] と同様に，2 つのクラスタ群で F 値の総和が最大になるようなクラスタの組み合わせを決定して計算した．

提案手法により生成されたクラスタ群を $\boldsymbol{S} = \{S_1, \cdots, S_c\}$（$c$ はクラスタ数である），人手により生成された正解クラスタ群を $\boldsymbol{L} = \{L_1, \cdots, L_c\}$ としたとき，クラスタ S_i に含まれる意見の数を s_i，クラスタ L_j に含まれる意見の数を l_j，S_i と L_j の両方に含まれる意見の数を n_{ij} とする．このとき，任意のクラスタ S_i と L_j との F 値 $F(S_i, L_j)$ は，再現率 $R(S_i, L_j)$，適合率 $P(S_i, L_j)$ から以下のように求まる．

$$R(S_i, L_j) = \frac{n_{ij}}{l_i} \tag{5.1}$$

$$P(S_i, L_j) = \frac{n_{ij}}{s_j} \tag{5.2}$$

$$F(S_i, L_j) = \frac{2 \times R(S_i, L_j) \times P(S_i, L_j)}{R(S_i, L_j) + P(S_i, L_j)} \tag{5.3}$$

再現率 R は完全性を評価するための尺度であり，クラスタ L_j に含まれる意見の中でクラスタ S_i にも含まれる意見の割合を示す．適合率 P は正確性を評価するための尺度であり，クラスタ S_i に含まれる意見の中でクラスタ L_j にも含まれる意見の割合を示す．また，F 値は再現率と適合率の調和平均である．

例えば，図 5.1 のように，7 つの意見が 3 つのクラスタ（観点）に分けられたとき，提案手法により生成されたクラスタ群 $\boldsymbol{S}$ と正解クラスタ群 $\boldsymbol{L}$ の各クラスタ間の F 値は表 5.2 のようになる．

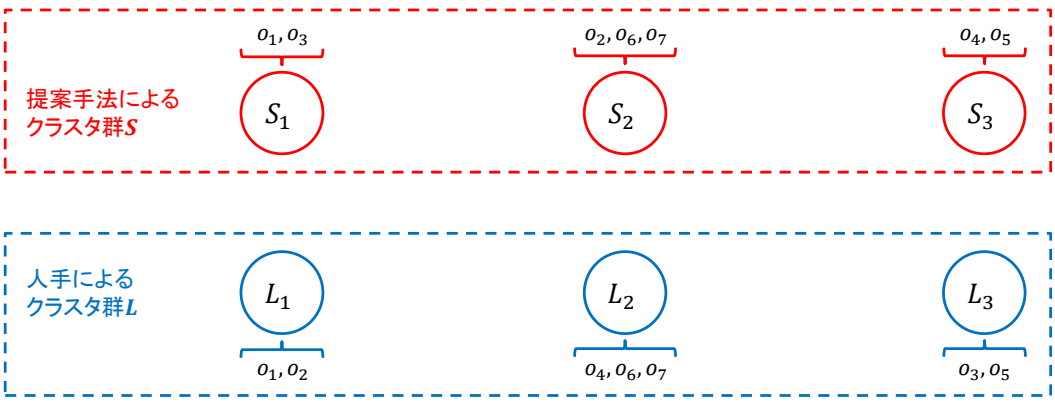

図 5.1　o_1〜o_7 の 7 つの意見のクラスタリング例

表 5.2　図 5.1 における各クラスタ間の F 値

		$\boldsymbol{L}$		
		L_1	L_2	L_3
$\boldsymbol{S}$	S_1	0.50	0.00	0.50
	S_2	0.40	0.67	0.00
	S_3	0.00	0.40	0.50

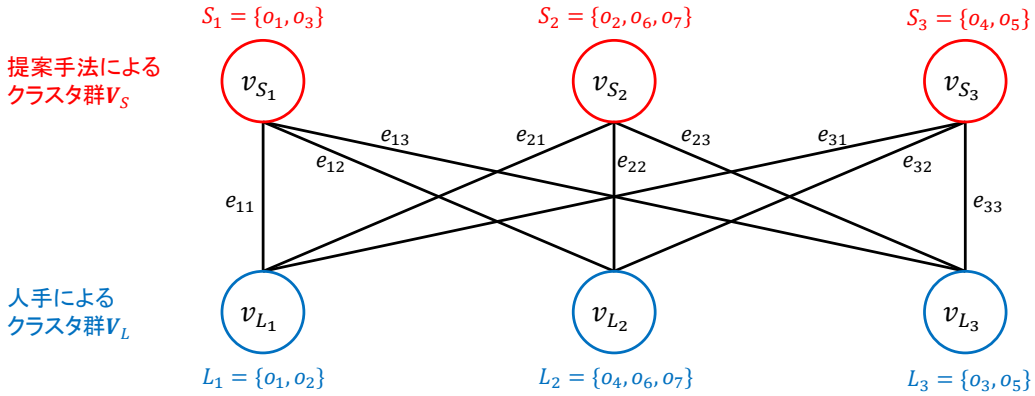

図 5.2　完全 2 部グラフ $K_{|S|,|L|}$ の例

次に，提案手法により生成されたクラスタ群 $\boldsymbol{S}$ と正解クラスタ群 $\boldsymbol{L}$ をそれぞれ 2 つの頂点集合 $\boldsymbol{V}_S, \boldsymbol{V}_L$ とし，それぞれの頂点 v_{S_i}, v_{L_j} をすべて結んだ完全 2 部グラフ[*15] $K_{|S|,|L|}$（図 5.2）を得る．なお，$\boldsymbol{E}$ はそれぞれの頂点を結んだエッジ e_{ij} の集合である．

$$\boldsymbol{V}_S = \{v_{S_1}, v_{S_2}, \cdots, v_{S_c}\} \tag{5.4}$$

$$\boldsymbol{V}_L = \{v_{L_1}, v_{L_2}, \cdots, v_{L_c}\} \tag{5.5}$$

$$\boldsymbol{E} = \{(v_{S_i}, v_{L_j}) | v_{S_i} \in \boldsymbol{V_S}, v_{L_j} \in \boldsymbol{V_L}\} \tag{5.6}$$

このとき各頂点は，それぞれのクラスタ群に含まれるクラスタに対応される．任意の頂点 v_{S_i} と v_{L_j}（クラスタ S_i と L_j）を結ぶ辺の重み $W(v_{S_i}, v_{L_j})$ は式 (5.7) のように，クラスタ間の F 値に全体の意見数 n（図 5.1 で示した例にならえば $n = 7$）のうち正解クラスタ L_j に含まれる意見数 l_j の割合を掛けて計算する．

$$W(v_{S_i}, v_{L_j}) = \frac{l_j}{n} F(S_i, L_j) \tag{5.7}$$

[*15] グラフ理論における 2 部グラフにおいて，片方の集合に属する各頂点から別の集合に属するすべての頂点に辺が伸びているものを特に完全 2 部グラフという．

表 5.3　図 5.2 におけるエッジの重み

		$\boldsymbol{V_L}$		
		v_{L_1}	v_{L_2}	v_{L_3}
$\boldsymbol{V_S}$	v_{S_1}	$e_{11} = 0.14$	$e_{12} = 0.00$	$e_{13} = 0.14$
	v_{S_2}	$e_{21} = 0.11$	$e_{22} = 0.29$	$e_{23} = 0.00$
	v_{S_3}	$e_{31} = 0.00$	$e_{32} = 0.17$	$e_{33} = 0.14$

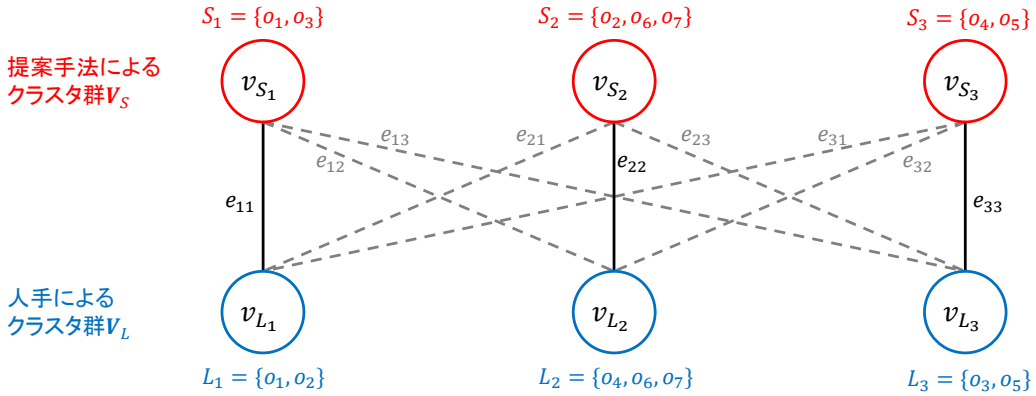

図 5.3　最大マッチング問題から得られるクラスタの組み合わせ

以上から得られた完全 2 部グラフの重み付き最大マッチング問題を解くことで，F 値の総和が最大になる組み合わせを決定し，そのときの F 値の平均を最終的な評価値とした．

図 5.1 のような例のもとでは，頂点（クラスタ）間のエッジの重みは表 5.3 のようになることから，完全 2 部グラフの重み付き最大マッチング問題から得られるクラスタの組み合わせは図 5.3（黒の実線で繋がっているクラスタどうしの組み合わせ）のようになる．このとき，組み合わせをもとに表 5.2 から計算した F 値の総和は 1.67 であるので，その平均である最終的な評価値としての F 値は 0.56 となる．

5.4 比較手法

本研究で提案したクラスタリング手法が既存の手法に比べ，どの程度の性能を示すか調査するために比較手法を用意した．

意見をクラスタリングする既存手法としては，Luo ら [Luo 09] と鷹栖ら [鷹栖 13] の手法が挙げられる．しかし，Luo らの手法は観点の差異や特徴を考慮したものではなく，得られたクラスタに含まれる特徴（肯定的・否定的など）を分析することに焦点をおいているため比較手法からは除外した．（文献中の単語への重み付け方法，クラスタリングの流れ等の説明が不明瞭であったことも除外理由の 1 つである．）また，鷹栖らの手法は，Twitter 上に存在する意見ツイートを対象としたクラスタリングであるため，本研究の評価実験で扱う意見集合に適用することができない．

そのため，意見集合に特化したものではない従来の文書クラスタリング手法になるが，比較手法として，LSI 法，LDA 法，MVSC 法の 3 つを用意した．

5.4.1 LSI 法

LSI 法は，一般的によく用いられるクラスタリング手法である．具体的には，意見に含まれる自立語の出現頻度を素性とした単語・文書行列に対して次元圧縮を行い，得られた文書（意見）の特徴ベクトルを用いてクラスタリングする方法である．意見どうしの類似度はコサイン類似度により計算し，クラスタリングには提案手法と同様に階層型クラスタリング手法である Ward 法を用いた．

5.4.2 LDA 法

LDA 法は，LDA（Latent Direchlet Allocation; 潜在的ディリクレ配分法）[Blei 03] を用いてクラスタリングする方法である．

LDA は，「1 つの文書には複数のトピックからなる単語が含まれる」という仮定をもとにしたモデルであるが，これを文書単位ではなく意見単位で考えると，「1 つの意見には複数の観点からなる単語が含まれる」という仮定をもとにしたモデルとしてみなすこともできる．そのため，本研究では LDA を用いたクラスタリングを比較手法の 1 つとして用意した．

LDA では最終的に各文書におけるトピック生起確率分布（どのトピックがどの程度の割合で含まれているかという分布）が推定される．これを先ほどの仮定から，各意見における観点の生起確率分布とみなして，最も生起確率が高い（最も含まれる割合が高い）観点をその意見の観点として採用することでクラスタリングを行った．

5.4.3 MVSC 法

MVSC 法は，Nguyen ら [Nguyen 12] によって提案された非階層型クラスタリング手法である．

あるクラスタ C_r に含まれる文書 d_i, d_j の正規化された（ベクトルの大きさを 1 とした）特徴ベクトルを $\boldsymbol{x}_i, \boldsymbol{x}_j$ としたとき，そのコサイン類似度は一般的に次のように原点を中心とした 2 つのベクトルの成す角のコサインから計算される．

$$\mathrm{sim}(d_i, d_j) = \cos(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{|\boldsymbol{x}_i||\boldsymbol{x}_j|} = \boldsymbol{x}_i \cdot \boldsymbol{x}_j \tag{5.8}$$

一方で Nguyen らは，文書 $d_i, d_j \in C_r$ 間の類似度をクラスタ C_r 以外のクラスタに属する文書 d_h の特徴ベクトル $\boldsymbol{x}_h$ を中心とした 2 つのベクトル $\boldsymbol{x}_i, \boldsymbol{x}_j$ が成す角からコサイン類似度を計算している．

$$\begin{aligned}
\mathrm{sim}(d_i, d_j | d_i, d_j \in C_r) &= \frac{1}{n - n_r} \sum_{\boldsymbol{x}_h \in \bar{C}_r} \cos(\boldsymbol{x}_i - \boldsymbol{x}_h, \boldsymbol{x}_j - \boldsymbol{x}_h) \\
&= \frac{1}{n - n_r} \sum_{\boldsymbol{x}_h \in \bar{C}_r} (\boldsymbol{x}_i - \boldsymbol{x}_h) \cdot (\boldsymbol{x}_j - \boldsymbol{x}_h) \\
&= \frac{1}{n - n_r} \sum_{\boldsymbol{x}_h \in \bar{C}_r} \boldsymbol{x}_i \cdot \boldsymbol{x}_j - (\boldsymbol{x}_i + \boldsymbol{x}_j) \cdot \boldsymbol{x}_h + \boldsymbol{x}_h \cdot \boldsymbol{x}_h
\end{aligned} \tag{5.9}$$

なお，n は全文書数，n_r はクラスタ C_r に含まれる文書数を指す．また，特徴ベクトルは全て正規化されたものであり，以降の説明においても同様である．

Nguyen らは，この方法により計算した類似度を用いてクラスタリングを行っている．手順としては，まずクラスタ C_r に含まれる全ての文書どうしで類似度の総和 I_r を計算する．

$$\begin{aligned}
I_r &= \sum_{d_i, d_j \in C_r} \mathrm{sim}(d_i, d_j) \\
&= \sum_{\boldsymbol{x}_i, \boldsymbol{x}_j \in C_r} \frac{1}{n - n_r} \sum_{\boldsymbol{x}_h \in \bar{C}_r} \boldsymbol{x}_i \cdot \boldsymbol{x}_j - (\boldsymbol{x}_i + \boldsymbol{x}_j) \cdot \boldsymbol{x}_h + \boldsymbol{x}_h \cdot \boldsymbol{x}_h \\
&= \frac{1}{n - n_r} \sum_{\boldsymbol{x}_i, \boldsymbol{x}_j \in C_r} \sum_{\boldsymbol{x}_h \in \bar{C}_r} \{\boldsymbol{x}_i \cdot \boldsymbol{x}_j - (\boldsymbol{x}_i + \boldsymbol{x}_j) \cdot \boldsymbol{x}_h + \boldsymbol{x}_h \cdot \boldsymbol{x}_h\}
\end{aligned} \tag{5.10}$$

また，全文書の特徴ベクトルの総和を D とし，クラスタ C_r に含まれる文書の特徴ベクトルの総和を D_r とすることで，以下のような関係式を得ることができる．

$$\sum_{\boldsymbol{x}_i \in C_r} \boldsymbol{x}_i = \sum_{\boldsymbol{x}_j \in C_r} \boldsymbol{x}_j = D_r \tag{5.11}$$

$$\sum_{r} D_r = D \tag{5.12}$$

$$\sum_{\boldsymbol{x}_h \in \bar{C_r}} \boldsymbol{x}_h = D - D_r, \quad |\boldsymbol{x}_h|^2 = 1 \tag{5.13}$$

これより，式 (5.10) は以下のように変形することができる．

$$\begin{aligned} I_r &= \frac{1}{n-n_r} \left\{ (n-n_r) \sum_{\boldsymbol{x}_i, \boldsymbol{x}_j} \boldsymbol{x}_i \cdot \boldsymbol{x}_j - 2n_r \times \sum_{\boldsymbol{x}_i} \sum_{\boldsymbol{x}_h} \boldsymbol{x}_i \cdot \boldsymbol{x}_h + n_r^2 (n-n_r) \right\} \\ &= D_r \cdot D_r - \frac{2n_r}{n-n_r} D_r \cdot (D - D_r) + n_r^2 \\ &= \frac{n+n_r}{n-n_r} |D_r|^2 - \frac{2n_r}{n-n_r} D_r \cdot D + n_r^2 \\ &= \frac{n+n_r}{n-n_r} |D_r|^2 + \left(1 - \frac{n+n_r}{n-n_r} \right) D_r \cdot D + n_r^2 \\ &= I_r(n_r, D_r) \end{aligned} \tag{5.14}$$

次に，任意のクラスタ C_p に含まれる文書 d_s が，C_p 以外の任意のクラスタ C_q に移動したときの I_p, I_q の変化量 $\Delta I_p, \Delta I_q$ をそれぞれ計算する．

$$\Delta I_p = |I_p(n_p - 1, D_p - \boldsymbol{x}_s) - I_p(n_p, D_p)| \tag{5.15}$$

$$\Delta I_q = |I_p(n_q + 1, D_q + \boldsymbol{x}_s) - I_q(n_q, D_q)| \tag{5.16}$$

このとき，クラスタ C_p を除く全てのクラスタの中で最も変化量が大きい C_q を決定し，ΔI_q が ΔI_p より大きければ，文書 d_s をクラスタ C_q へ移動する．つまり，あるクラスタに含まれる文書を別のクラスタに移動することを考えるとき，コスト（類似度）が大きくなるような移動先のクラスタを決定していくという手順を取る．これをすべての文書に対して，移動する文書がなくなるまでクラスタリングを続ける．

MVSC 法を比較手法とした理由は，従来の文書間類似度の計算方法とは異なる方法をとっており，k-means 法など従来のクラスタリング手法より良い精度が得らたとの報告があったからである．

5.5 パラメータについて

提案手法では単語間の類似度計算（式 (4.2)）や名詞・動詞ペアどうしの類似度計算（式 (4.7)），名詞どうしの類似度計算（式 (4.8), 式 (4.9)）においてパラメータ α, β, λ を用いていることから，評価実験では，これらのパラメータの値を 0〜1 の 0.1 刻みで変化させてクラスタリングを行った．また，提案手法と比較手法の LSI 法では意味空間の構築に次元圧縮を用いていることから，意味空間の次元数についても提案手法・LSI 法ともに 5〜35 の 5 刻みで変化させてクラスタリングを行った．

パラメータの組み合わせを変化させて分類した結果に対する性能評価（F 値の算出）としては，パラメータの組み合わせごとに F 値を計算し，最も高い F 値を採用することが考えられる．しかし，そのままではパラメータ依存の結果に陥ってしまう（パラメータを増やし，調整することでいくらでも精度が上がってしまう）ため，複数のパラメータを用いた分類結果に対して性能評価を行う際は，一般的に交差検定[*16]により F 値を計算する．本研究における評価実験では，各トピックごとに 3 人ずつ正解データを作ってもらった（各トピックごとに 3 種類の正解データが存在する）ことから，Leave-one-out 交差検定により F 値を計算した．

5.5.1 Leave-one-out 交差検定

Leave-one-out 交差検定とは，n 個の正解データがあったとき $n-1$ 個のデータを学習（訓練）データとして，残り 1 個のデータをテストデータとしてパラメータの学習（決定）と分類を別々に行う検定のことである．本研究に沿った手順は以下の通りである．

1. 3 人の人手により生成された正解クラスタ群（$\boldsymbol{C}_x, \boldsymbol{C}_y, \boldsymbol{C}_z$）のうち，2 つの正解クラスタ群を選択する．（ここでは例として $\boldsymbol{C}_x, \boldsymbol{C}_y$ を選択する）
2. パラメータの組み合わせを変化させて得たクラスタ群と手順 1 で選択した正解クラスタ群 $\boldsymbol{C}_x, \boldsymbol{C}_y$ それぞれとで F 値（F_x, F_y）を計算し，F_x, F_y の平均が最も高かったときのパラメータの組み合わせを P とする．
3. 手順 2 で得たパラメータの組み合わせ P でクラスタリングを行い，その結果得られたクラスタ群と手順 1 で選択されなかった正解クラスタ群（$\boldsymbol{C}_z$）との間で F 値（F_z）を計算し，この F_z を正解クラスタ群 $\boldsymbol{C}_z$ に対する最終的な F 値とする．

[*16] 複数の正解データで性能が最大になるときのパラメータを学習し，その学習したパラメータを用いて別の正解データに対する性能を求める方法．

4. 正解クラスタ群 $\boldsymbol{C}_x, \boldsymbol{C}_y$ に対する最終的な F 値を手順 1〜3 と同様に計算する．
5. すべてのトピック（原発，TPP，STAP 細胞，人口問題）について手順 1〜4 を行う．

5.6 実験結果

以上の評価実験手順・指標の設定のもと，実験を行った結果を表 5.4 に示す．なお，表中の k はクラスタ（観点）数を指し，太字の数値は各行で最も高かった F 値を指す．「パラメータの最適値」列は，Leave-one-out 交差検定により学習されたパラメータであり，d_p, d_c はそれぞれ提案手法と LSI 法における意味空間の次元数を指す．また，提案手法において交差検定を行わずに，各正解クラスタ群に対して F 値が最大となったときのその値とパラメータを表 5.5 に示す．

なお，比較手法の LDA 法では，ディリクレ事前分布を生成するためのハイパーパラメータ α, β をそれぞれ 0.5 に設定し，観点の数に相当するトピック数のパラメータ k を人手による分類と同じ観点の数に設定した．パラメータ（各意見における観点の生起確率分布）の推定には Collapsed Gibbs Sampling を用い，Sampling の回数は 500 回とした．ただし，Sampling の内部処理で乱数を発生させているため，クラスタリングを 50 回行い，最終的な F 値はその 50 回の平均とした．また，MVSC 法においても，クラスタの初期状態を設定するために乱数を発生させているため，クラスタリングを 50 回行い，最終的な F 値はその 50 回の平均とした．

実験結果より，すべての正解クラスタ群に対して提案手法において最も良い精度を得ることができた．このことから，提案手法が有用であると言える．

表 5.4　クラスタリングの評価実験結果（交差検定）

トピック	k	F 値				パラメータの最適値	
		提案手法	LSI	LDA	MVSC	提案手法 $(\alpha, \beta, \lambda, d_p)$	LSI d_c
原発	7	**0.406**	0.326	0.302	0.239	$0.6, 0.6, 0.9, 25$	25
	9	**0.424**	0.339	0.227	0.299	$0.9, 0.8, 0.9, 20$	10
	12	**0.435**	0.294	0.237	0.321	$0.8, 0.7, 0.8, 15$	25
平均		**0.421**	0.320	0.255	0.286	-	
TPP	9	**0.468**	0.352	0.320	0.246	$0.3, 0.6, 0.7, 20$	35
	10	**0.525**	0.368	0.336	0.286	$0.1, 0.7, 0.8, 15$	35
	12	**0.422**	0.345	0.332	0.238	$0.3, 0.7, 0.8, 20$	5
平均		**0.472**	0.355	0.329	0.257	-	
STAP 細胞	10	**0.539**	0.380	0.370	0.308	$0.4, 0.5, 0.7, 15$	35
	$11_{(1)}$	**0.520**	0.450	0.352	0.290	$0.4, 0.7, 0.9, 15$	35
	$11_{(2)}$	**0.526**	0.432	0.370	0.293	$0.3, 0.8, 0.7, 15$	35
平均		**0.528**	0.421	0.364	0.297	-	
人口問題	8	**0.405**	0.358	0.350	0.239	$0.4, 0.5, 0.7, 10$	35
	10	**0.481**	0.362	0.306	0.246	$0.4, 0.7, 0.7, 15$	35
	11	**0.427**	0.344	0.328	0.267	$0.2, 0.8, 0.6, 10$	35
平均		**0.438**	0.355	0.328	0.251	-	

表 5.5　クラスタリングの評価実験結果（最大 F 値）

トピック	k	最大 F 値	パラメータ			
			α	β	λ	d_p
原発	7	0.535	0.8	0.7	0.9	15
	9	0.534	0.7	0.7	0.8	30
	12	0.544	0.6	0.9	0.9	15
平均		0.538	-			
TPP	9	0.548	0.2	0.8	0.7	15
	10	0.529	0.3	0.5	0.6	20
	12	0.550	0.3	0.6	0.8	35
平均		0.542	-			
STAP 細胞	10	0.586	0.6	0.5	0.9	15
	$11_{(1)}$	0.555	0.5	0.4	0.6	15
	$11_{(2)}$	0.636	0.4	0.7	0.9	15
平均		0.592	-			
人口問題	8	0.534	0.2	0.6	0.5	25
	10	0.548	0.3	0.6	0.7	25
	11	0.532	0.1	0.8	0.6	25
平均		0.538	-			

6 考察

6.1 有用性の評価

提案手法において，意見から抽出した名詞・動詞ペア $\langle N, V \rangle$ をクラスタリングに利用することで，クラスタリング性能にどのような影響を与えるかを調べるために，以下の 2 つの条件におけるクラスタリング性能を求め，評価実験で得られた提案手法の性能と比較した．

条件 1 名詞・動詞ペア $\langle N, V \rangle$ の名詞 N のみを利用して類似度を計算する．（すなわち，式 (4.7) において $\mathrm{sim}_v = 0$ とする．）

条件 2 条件 1 に加えて，名詞・動詞ペアの（複合）名詞 N を動詞との係り受け関係を考慮せずに形態素解析上で名詞と解析されたすべての語をもとにして抽出する．

条件 1 におけるクラスタリング性能を求めることで，4.1 節で述べた「観点の違いが，名詞が係る動詞の違いに反映されている」という考えが間違いでないか確かめることができる．また，条件 2 におけるクラスタリング性能を求めることで，4.1 節で述べた「単語が文中でどの品詞で用いられているかを同定することは，単語間の類似度計算に有用である」という考えが間違いでないか確かめることができる．

例えば，

火力発電だと電力の生成に多くの燃料を消費することになります。

このような文からは，条件 1 と条件 2 において次のような名詞・動詞ペア $\langle N, V \rangle$ が抽出される．

条件 1 ⟨⟨ **電力**, **生成** ⟩, **消費** ⟩, ⟨⟨ **多く**, **燃料** ⟩, **消費** ⟩

条件 2 ⟨⟨ **火力**, **電力** ⟩, None⟩, ⟨⟨ **発電**, **電力** ⟩, None⟩, ⟨⟨ **電力**, **生成** ⟩, None⟩, ⟨⟨ **生成**, **消費** ⟩, None⟩, ⟨⟨ **多く**, **燃料** ⟩, None⟩, ⟨⟨ **燃料**, **消費** ⟩, None⟩

なお，条件 2 では，動詞との係り受け関係を考慮していないことから，動詞 V を「None」としている．このような各条件で抽出された名詞・動詞ペアを用いて類似度計算を行い，クラスタリングの性能を求めた．

また，提案手法における複合名詞を利用した類似度計算の方法がクラスタリング性能に与える影響を調べるために，以下の条件との比較も行った．

条件 3　名詞・動詞ペア $\langle N, V \rangle$ の（複合）名詞 N において，修飾語の情報を考慮せず単一名詞のみを利用する．（すなわち，式 (4.8),(4.9) において $\beta = 0$ とする．）

以上の 3 つの条件下における実験結果を図 6.1 に示す．なお，図中の各トピックごとの F 値は，3 つの正解クラスタ群に対する F 値のマクロ平均であり，各条件におけるクラスタリング性能の評価手順・指標は，評価実験と同じである．また，図 6.1 における詳細な F 値や Leave-one-out 交差検定によるパラメータの最適値は付録 A の表 A.1 に掲載している．

6.1.1　名詞・動詞ペアの利用について

図 6.1 の提案手法と条件 1 を比較すると，すべてのトピックにおいて提案手法の方が高い F 値を取っていた．つまり，動詞どうしの類似度を考慮したときの方がより観点に基づいたクラスタリングができていることを示しており，名詞・動詞ペア $\langle N, V \rangle$ を用いた類似度計算が有効であると言える．

トピック全体として，名詞・動詞ペアを利用することで精度の向上が見られたことから，単一の名詞だけでは観点の差異を考慮できない意見が存在していたと考えられる．例えば，

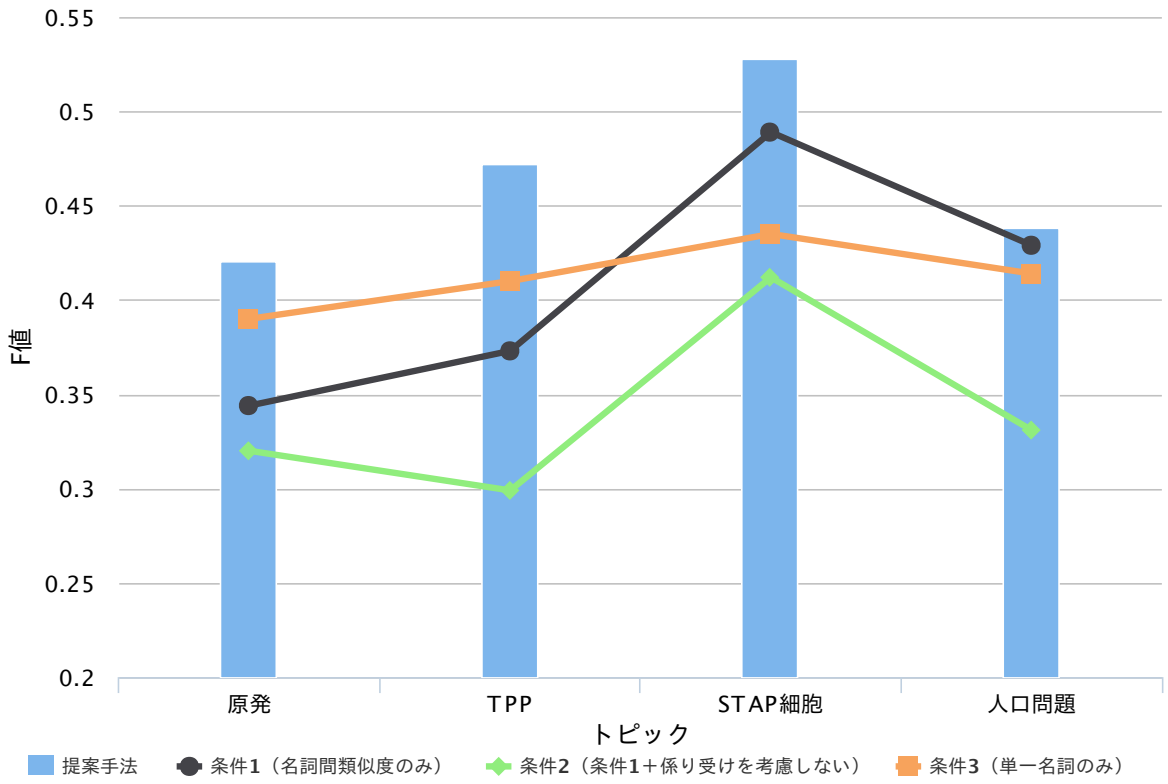

図 6.1　条件 1〜3 における実験結果

- **近年では、安い人件費のために製造業が 海外 に拠点を移転し、日本の産業界は空洞化を招いています。TPP に参加しなかったとすれば、企業そのものが日本から去ってしまうことを招きます。**
- **日本はこれまで 海外 と競争してこなかったのです。労働力も例外ではありません。今後のグローバル社会の中で日本人が活躍していくためには、外国人との競争は避けられないのです。**

このような 2 つの意見からはそれぞれ以下のような名詞・動詞ペア $\langle N, V \rangle$ が抽出される．

- ⟨ **製造**, **移転** ⟩，⟨**海外**, **移転** ⟩，⟨ **拠点**, **移転** ⟩，⟨ **産業**, **招く** ⟩，⟨ **空洞**, **招く** ⟩，⟨ **企業**, **去る** ⟩
- ⟨**海外**, **競争** ⟩，⟨ **労働**, **例外** ⟩，⟨⟨ **外国**, **競争** ⟩, **避ける** ⟩

このとき，2 つの意見にはそれぞれ「**海外**」という共通の名詞や「**産業**」と「**労働**」といった意味の近い名詞などが含まれていることから，一般的なクラスタリング手法では同じクラスタ（観点）の意見であると判定されてしまう．しかし，提案手法では，「**海外**」という共通の名詞を含む意見どうしでも名詞とペアになる動詞が「**移転**」と「**競争**」とで異なっており，他の $\langle N, V \rangle$ 間でも類似度が低くなることから，それぞれ異なる観点の意見であると判定できていた．

一方で，トピック「人口問題」では，名詞・動詞ペアを用いることで（提案手法と条件 1 を比較すると）若干の精度の向上は見られたが，大きな差は見られなかった．その原因としては，抽出された名詞・動詞ペアの動詞に非自立語扱いの単語（「ある」「なる」「いう」「対する」など）が多く含まれていたことが挙げられる．このような動詞は観点の違いを反映する情報としては不十分であることから，非自立語扱いの動詞が抽出された場合の名詞・動詞ペアどうしの類似度計算方法を改善する必要があると考えられる．

また，提案手法と条件 2 を比較すると，すべてのトピックにおいて提案手法で大きく精度が向上していることから，動詞との係り受けを用いて抽出した名詞の情報，つまり間接的に動詞の情報がクラスタリングに有用であるといえる．

条件 1 と条件 2 を比較すると，すべてのトピックにおいて条件 1 の方が高い F 値を取っていた．このことから，単純に形態素解析で名詞と判断されたものより，係り受け関係から名詞の機能を担っていると判断されたものを名詞として利用する方が適切に類似度を計算できると言える．特に，トピック「STAP 細胞」では，名詞・動詞ペア $\langle N, V \rangle$ を用いることで，従来の名詞（形態素解析で名詞と判断されたもの）でノイズとなるようなものが除去された形になったために，大きく精度が向上したと考えられる．例えば，

> **小保方氏の論文について，明確に不正疑惑があって，それに対して小保方氏は一部公正な生データが出せなくて疑惑が解消できなくて，不正認定された．それだけの事です．不正はあったから適切に処**

分する．STAP 細胞があるかどうかは理研が調査する．アカデミア研究者以外の人間が把握すべき事実はそれで十分．

この意見は「研究の不正」という観点が人手により付与されたものであるが，意見中には観点候補にはなりえないような名詞（「**理研**」や「**調査**」，「**把握**」など）が含まれており，一般的なクラスタリング手法では意見どうしの類似度におけるノイズになってしまう．しかし，この意見から抽出される名詞・動詞ペア $\langle N, V \rangle$ は

$\langle$ **不正**, **ある** $\rangle$，$\langle$ **疑惑**, **ある** $\rangle$，$\langle\langle$ **公正**, **生** $\rangle$, **出せる** $\rangle$，$\langle\langle$ **公正**, **データ** $\rangle$, **出せる** $\rangle$，$\langle$ **疑惑**, **解消** $\rangle$，$\langle$ **細胞**, **ある** $\rangle$，$\langle\langle$ **アカデミア**, **人間** $\rangle$, **把握** $\rangle$，$\langle\langle$ **研究**, **人間** $\rangle$, **把握** $\rangle$

このようになっており，意見中の観点候補にはなりえないような名詞がある程度除去されている．

名詞・動詞ペアどうしの類似度計算（式 (4.7)）において，パラメータ λ を用いていたが，評価実験の結果（表 5.4,5.5）ではその値は 0.5 以上とやや大きい値となっていた．つまり，式 (4.7) における動詞どうしの類似度 sim_v の係数が名詞どうしの類似度 sim_n に大きく依存していたということになる．仮に，λ の値が小さい（極端に言えば $\lambda = 0$）ときのことを考えると，名詞・動詞ペアどうしの類似度は sim_n と sim_v の単純な和となって名詞と動詞の情報を別々に見ることになり，4.6 節で述べた「sim_n が小さければ sim_v の大小に関わらず名詞・動詞ペアは異なる内容のものを示す」という考えと矛盾してしまう．しかし，実際には λ の値は小さくても 0.5 と大きくなっていたことから，提案手法におけるこの考えは正しかったと言える．

6.1.2　複合名詞の利用について

図 6.1 の提案手法と条件 3 を比較すると，すべてのトピックにおいて提案手法の方が高い F 値を取っていることが分かる．このことから，単一の名詞だけでなく修飾語を考慮した複合名詞をクラスタリングに利用することは有効だと言える．

特に，トピック「TPP」や「STAP 細胞」では，複合名詞を利用することで精度の改善が大きく見られた．実際，表 5.4,5.5 のトピック「TPP」「STAP 細胞」においても，パラメータ β の値は若干のバラツキはあるものの 0.7 前後と，修飾語の情報に重きが置かれた結果になっていた．つまり修飾語に観点を特徴づけるような語が多く含まれていたことが精度の改善に繋がったと考えられる．

表 5.4,5.5 におけるトピック「原発」や「人口問題」もパラメータ β の値が 0.5〜0.9 と修飾語に基づく類似度に依存した結果になっているものの，図 6.1 を見ると複合名詞を利用する場合と単一

名詞のみを利用する場合とで精度にあまり差がないことが分かる．このようになった理由として，修飾語・被修飾語の両方に観点を特徴づけるような語が含まれていることが挙げられる．例えば，

- **これからの太陽光発電パネルのような代替案がしっかりと着手するまでは政治判断で発電所稼働をストップする事が出来ないのが現状です。**
- **原子力発電をせずに日本全国の電力がまかなえるのなら原子力発電はやめるべきですが、現状むずかしいので代わりの案や代替エネルギーが出来るまでは原子力発電は必要だと思います。**

これらの意見からは

- ⟨⟨ **太陽**, **代替** ⟩, **着手** ⟩，⟨⟨ **太陽**, **案** ⟩, **着手** ⟩，⟨⟨ **発電**, **代替** ⟩, **着手** ⟩，⟨⟨ **発電**, **案** ⟩, **着手** ⟩，
  ⟨⟨ **パネル**, **代替** ⟩, **着手** ⟩，⟨⟨ **パネル**, **案** ⟩, **着手** ⟩，⟨ **政治**, **ストップ** ⟩，⟨ **判断**, **ストップ** ⟩，
  ⟨ **発電**, **ストップ** ⟩，⟨ **稼働**, **ストップ** ⟩，⟨ **原子**, **ストップ** ⟩
- ⟨ **原子**, **する** ⟩，⟨ **発電**, **する** ⟩，⟨⟨ **やめる**, **必要** ⟩, **思う** ⟩，⟨⟨ **全国**, **電力** ⟩, **まかなえる** ⟩，
  ⟨ **原子**, **やめる** ⟩，⟨ **発電**, **やめる** ⟩，⟨ **現状**, **出来る** ⟩，⟨⟨ **出来る**, **必要** ⟩, **思う** ⟩，
  ⟨⟨ **代わり**, **代替** ⟩, **出来る** ⟩，⟨⟨ **代わり**, **エネルギー** ⟩, **出来る** ⟩，⟨ **代替**, **出来る** ⟩，
  ⟨ **エネルギー**, **出来る** ⟩，⟨ **必要**, **思う** ⟩

このような名詞・動詞ペア $\langle N, V \rangle$ がそれぞれ抽出される．この 2 つの意見は，ともに「代替エネルギー」という観点が人手により付与されており，前者の意見から抽出された $\langle N, V \rangle$ には非修飾語に「**代替**」という単語が入っている．また，後者の意見から抽出された $\langle N, V \rangle$ には修飾語に「**代わり**」，被修飾語に「**代替**」，単一名詞にも「**代替**」という単語が入っている．

このとき，修飾語の情報に重きが置かれた場合，後者の意見から抽出された $\langle N, V \rangle$ には修飾語・単一名詞のいずれにも「**代替**」や「**代わり**」という単語があることから類似度の計算に支障はないが，前者の意見から抽出された $\langle N, V \rangle$ には被修飾語である「**代替**」の情報が小さく見られてしまうため，類似度の計算に不都合が生じる．しかし，このように修飾語の情報に重きが置かれた場合でも被修飾語の情報のみを利用した場合と近い精度が得られた（類似度の計算ができていた）要因として，4.5 節で述べた LSI により構築した意味空間を用いた単語どうしの類似度計算が効いていると考えている．意味空間に含まれる単語は，その単語と同じ意見に出現した別の単語の情報を含むという特性がある．つまり，先ほどの例で言えば，「**太陽**」や「**着手**」「**発電**」といった修飾語には被修飾語である「**代替**」の情報も潜在的に含まれていることになる．そのため，修飾語の情報に重きが置かれた場合でも「**代替**」という語の意味を潜在的に含んだ語により適切に類似度が計算できたと考えられる．

また，条件 3 におけるトピック「TPP」の F 値を取ったときのパラメータ（表 A.1 参照）を見ると，すべて $\lambda = 0$ となっていた．つまり，名詞・動詞ペアどうしの類似度計算（式 (4.7)）において名詞どうしの類似度 sim_n と動詞どうしの類似度 sim_v が独立して利用されたということになる．これは，トピック「TPP」では修飾語に動詞を利用することが比較的有用である[*17]ことが関係していると思われる．

提案手法では，複合名詞の修飾語に原則として自立語を用いており，その中には動詞および動詞の機能を担っていると判定された名詞も含まれていることから，条件 3 のように $\beta = 0$ として修飾語の情報を利用しない場合には，これらの修飾語としての動詞の情報は無視されてしまう．しかし，$\lambda = 0$ のときの名詞・動詞ペアどうしの類似度 sim_{nv} は，以下のように名詞どうしの類似度 sim_n と動詞どうしの類似度 sim_v の和となり，これはある意味で複合名詞どうしの類似度と似たものを表すことになる．

$$\begin{aligned}\mathrm{sim}_{nv} &= \mathrm{sim}_n + \left((1-0) + 0 \times (\mathrm{sim}_n)^2\right) \times \mathrm{sim}_v \\ &= \mathrm{sim}_n + \mathrm{sim}_v\end{aligned}$$

例えば，複合名詞どうしの類似度計算において 2 つの修飾語が動詞だとすると，そのような複合名詞どうしの類似度は修飾語である動詞どうしの類似度と被修飾語である名詞どうしの類似度との和になり，形式上は上式と同じになる．つまり，条件 3 におけるトピック「TPP」では，修飾語の情報が使えない代わりに $\lambda = 0$ とすることで動詞どうしの類似度を修飾語（としての動詞）どうしの類似度を計算するようにパラメータが調整されたと考えられる．

[*17] 後述の 6.4 節で動詞のみを修飾語にする場合で，提案手法の次点に良い精度となった．

6.2 日本語 WordNet・LSI を用いた単語間類似度について

表 5.4 においては，正解データによって多少のばらつきはあるものの，全体的にパラメータ α の値が 0 や 1 に偏るということはなかった．つまり，日本語 WordNet 用いた類似度と LSI により構築した意味空間を用いた類似度をうまく組み合わせることで単語どうしの類似度が適切に計算されたと考えられる．

トピック「原発」では，表 5.4,5.5 ともに α の値は全体的に高くなり，日本語 WordNet を用いた類似度に重きを置いた結果になっていた．このような結果となった背景として，人手により付与された観点の性質が関係していると思われる．例えば，トピック「原発」においては，意見中の単語が直接観点に結びつくというよりは，「危険性」ならば「安全」，「兵器」ならば「平和」といったように単語の概念（連想）的関係が観点に結びつく傾向があった．また，前節でも述べたように，「原発」では LSI を用いた類似度も有用であることから，概念辞書である日本語 WordNet の情報をメインにしつつ，LSI の情報を控えめに利用するような α の値となったと考えられる．

トピック「TPP」「人口問題」においては，人手で付与された観点を見ると，意見に出現した単語が直接観点に結びつく傾向があった．そのため，表 5.4,5.5 におけるパラメータ α の値が小さめに，つまり文中の単語を用いて構築した意味空間から単語どうしの類似度を計算する LSI を中心に利用することで精度の向上に繋がったと考えられる．また，「TPP」については，提案手法により名詞と判定された単語のうち日本語 WordNet に存在していなかった単語が 20 %程度あったことから，4.5 節で述べた「2 つの単語のどちらかが日本語 WordNet に存在しない場合は，LSI を用いた類似度のみを利用する」傾向が強く，それがパラメータ α にも反映されたと考えられる．

6.3 エラー分析

比較手法より高い精度を得られた提案手法であるが，中には誤った観点クラスタに意見が属してしまっていることも見られた．そのような正しくクラスタリングができなかった主な原因は以下の 3 点が考えられる．

- 意見が示す観点とは関係のない名詞・動詞ペア間で類似度が大きくなってしまった．
- 異なる観点を特徴づけるような名詞・動詞ペアが複数抽出されてしまった．
- 意見が示す観点を特徴づけるような名詞・動詞ペアが抽出されなかった．

6.3.1 名詞・動詞ペア間の類似度計算について

誤ったクラスタリングが行われた主な原因の 1 つとして，意見が示す観点とは関係のない名詞・動詞ペア間で類似度が大きくなってしまったことが挙げられる．

特に，名詞・動詞ペア間の類似度の中でも名詞どうしの類似度計算において，2 つの名詞が意見の観点とは関係ないものであるのに関わらず類似度が高くなってしまっていたという事例が多く見られた．これは，提案手法の名詞・動詞ペア間の類似度計算における「名詞どうしの類似度が小さければ動詞どうしの類似度の大小に関わらず 2 つの名詞・動詞ペアが異なる内容を表す可能性が高い」という考えから起因していると考えられる．つまり，名詞どうしの類似度が大きければ大きいほど名詞・動詞ペア間の類似度は大きくなりやすくなる．そのため，不当な名詞どうしの類似度が大きくなってしまうと，そのまま名詞・動詞ペア間の類似度が大きくなり，異なる意見を示す意見どうしが同じ観点クラスタに属してしまうことになる．

このような問題に対処するためには，名詞・動詞ペアの抽出方法と名詞・動詞ペア間の類似度計算方法のさらなる改善が必要である．また，抽出した名詞・動詞ペア（または名詞，動詞それぞれ）に重み付けを施す必要もあると考えられる．

また，単語間の類似度が不当に大きくなってしまったのは，日本語 WordNet や LSI の特性も影響していると考えられる．例えば，日本語 WordNet を用いた類似度では，「発生」や「発展」などの単語が，どの単語とも類似度が大きくなりやすくなっていた．また，LSI により構築した意味空間では 6.1.2 節で述べたように，同じ意見中に出現した単語の情報が潜在的に含まれる．しかし，意見の観点とは関係のない単語の情報が作用してしまうと，不当な単語どうしで類似度が大きくなってしまう．さらに，LSI を用いた類似度では，式 (4.6) のようにコサイン類似度が 0〜1

の値を取るようにスケールを調整しているが，コサイン類似度自体がマイナス値を取ることが少なく，LSI を用いた類似度が 0.5 以上の値を取りやすくなっていた．そのため，意見の観点とは関係のない名詞（または名詞・動詞ペア）に対しては類似度が小さくなるような重み付けを施すことや，スケール調整をせずにコサイン類似度が 0 未満のときは強制的に類似度を 0 にすることが必要である．

6.3.2 名詞・動詞ペアの抽出について

前節に加え，誤ったクラスタリングが行われた主な原因として，異なる観点を特徴づけるような名詞・動詞ペアが複数抽出されてしまったことや，意見が示す観点を特徴づけるような名詞・動詞ペアが抽出されなかったことも挙げられる．

例えば，次の意見には「安全性」や「代替案」「雇用」などの観点が人手により付与された．

> **原子力発電には、条件付きで反対です。理由としましては、事故が発生した際のリスクが高すぎる点です。現在起きている福島第一原子力発電の事故での被害や農作物等の影響が大きいからです。ただ条件付きとした点については、原子力発電に替わる安定した電力の供給方法や現在原子力発電所で働いている人などの雇用の確保などが必要だからです。**

この意見から抽出された名詞・動詞ペア $\langle N, V \rangle$ は次のようになっており，人手により付与されたそれぞれの観点を特徴づけるような $\langle N, V \rangle$ であった．

〈 **理由**, **する** 〉，〈 **事故**, **発生** 〉，〈 **リスク**, **すぎる** 〉，〈 **条件**, **する** 〉，〈 **原子**, **替わる** 〉，
〈 **発電**, **替わる** 〉，〈 **原子**, **働く** 〉，〈 **発電**, **働く** 〉

しかし，5.2 節でも述べたように，正解データを作成してもらった際，複数の観点を示す意見については，被験者の判断により，その意見に最もふさわしい（その意見で最も主張したいと思われる）観点を採用して観点ごとに分けてもらっていた．つまり，提案手法では 1 つの意見に 1 つの観点が含まれるという想定のもと，排他的クラスタリングを行っていることから，正しく $\langle N, V \rangle$ が抽出できていたとしても意図しない意見どうしが同じ観点クラスタに属してしまうことがある．上で挙げた意見の例で言えば，人手では最終的に「代替案」という観点クラスタに割り振られたが，提案手法では 〈 **リスク**, **すぎる** 〉 という $\langle N, V \rangle$ が大きく作用して「安全性」という観点クラスタに割り振られてしまっていた．

そのため，提案手法において，複数のクラスタ（観点）に属することを許すような非排他的なクラスタリングに適した類似度の計算方法の考案が必要であると考えられる．

また，次のような意見では，それぞれ「日米関係・外交」「保身」という観点が人手で付与され，

- **経済上のメリット、デメリットについてを議論することは必要なことです。しかし、TPP に参加する、しない、というのは日本にとってどうか、を考えた場合デメリットのほうが大きい。それでも私は、最終的には参加せざるを得ない状況になると考えています。これは上記でアメリカとの関係が非常に悪くなるという推察からです。そして日本が本当に考えなければならないのは、そのデメリットに関して、本当に把握しきれるかどうかです。きちんと把握し、日本の国会で議論し、防衛策、対応策を立てるべきです。**
- **ＳＴＡＰの存在有無は別として、今回の理研の対応は自らのずさんな管理と研究体制が、一人の研究員から露呈してしまって、結論ありきで慌ててトカゲのしっぽとして切った、という印象しかありませんね。組織が疲労しているのか腐敗レベルまで達してるかは分かりませんが、責任ある立場の人他の入れ替えは必要でしょう。しっぽ切って済まされる問題ではありません。**

抽出された名詞・動詞ペア $\langle N, V \rangle$ は，それぞれ次のようになっていた．

- ⟨ **デメリット**, **つく** ⟩, ⟨ **最終**, **参加** ⟩, ⟨⟨ **参加**, **状況** ⟩, **なる** ⟩, ⟨ **上記**, **なる** ⟩, ⟨ **関係**, **なる** ⟩, ⟨ **本当**, **考える** ⟩, ⟨ **デメリット**, **関する** ⟩, ⟨ **本当**, **把握** ⟩, ⟨ **国会**, **議論** ⟩, ⟨ **対応**, **立てる** ⟩, ⟨ **策**, **立てる** ⟩
- ⟨ **存在**, **ある** ⟩, ⟨ **有無**, **ある** ⟩, ⟨ **別**, **する** ⟩, ⟨ **対応**, **する** ⟩, ⟨⟨ **いう**, **印象** ⟩, **ある** ⟩, ⟨ **研究**, **露呈** ⟩, ⟨ **体制**, **露呈** ⟩, ⟨ **結論**, **ある** ⟩, ⟨⟨ **トカゲ**, **しっぽ** ⟩, **する** ⟩, ⟨ **組織**, **疲労** ⟩, ⟨ **腐敗**, **達する** ⟩, ⟨ **レベル**, **達する** ⟩, ⟨ **責任**, **ある** ⟩, ⟨ **しっぽ**, **切る** ⟩

人手で付与された観点と抽出された $\langle N, V \rangle$ を見比べると，観点を特徴づけるような $\langle N, V \rangle$ が含まれていない，または抽出できていないことが分かる．例えば，1 つ目の意見では，「日米関係・外交」といった観点が人手で付与されていることから，⟨⟨ **アメリカ**, **関係** ⟩, **悪くなる** ⟩ といったような $\langle N, V \rangle$ が抽出されることが望ましい．しかし，「**アメリカ**」という固有名詞は名詞から省いており，「**悪い**」という形容詞も修飾語としてみなされてしまうため，期待するような $\langle N, V \rangle$ を抽出することができない．また，2 つ目の意見では，「**トカゲのしっぽとして切った**」という文が「保身」という観点につながるが，この文から得られる $\langle N, V \rangle$ の情報だけでは「トカゲのしっぽ切り」と「保身」を結びつけることができない（類似度を計算することができない）．

このように，意見中から観点を特徴づける名詞・動詞ペア $\langle N, V \rangle$ が抽出できていないことや，抽出した $\langle N, V \rangle$ だけでは特徴づけることが困難な観点が人手により付与されているものがあったことも正しくクラスタリングができなかった原因の 1 つだと考えられる．そのため，$\langle N, V \rangle$ の抽出方法の改善や名詞または動詞の抽出条件を変える（緩くする）必要がある．

6.4 修飾語の種類について

提案手法では，名詞・動詞ペアの複合名詞における修飾語に自立語（名詞・動詞・形容詞）を用いている．そこで，修飾語の品詞によるクラスタリング性能の変化を調べるため，以下の 4 つの条件におけるクラスタリング性能を求め，評価実験で得られた提案手法の性能と比較した．

条件 a 名詞・動詞ペア $\langle N, V \rangle$ の（複合）名詞 N において，修飾語の情報を考慮せず単一名詞のみを利用する．（すなわち，式 (4.8),(4.9) において $\beta = 0$ とする．）

条件 b 名詞・動詞ペア $\langle N, V \rangle$ の（複合）名詞 N において，修飾語に名詞のみを利用する．

条件 c 名詞・動詞ペア $\langle N, V \rangle$ の（複合）名詞 N において，修飾語に形容詞のみを利用する．

条件 d 名詞・動詞ペア $\langle N, V \rangle$ の（複合）名詞 N において，修飾語に動詞のみを利用する．

以上の 4 つの条件下における実験結果を図 6.2 に示す．なお，図中の各トピックごとの F 値は，3 つの正解クラスタ群に対する F 値のマクロ平均であり，各条件におけるクラスタリング性能の評価手順・指標は，評価実験と同じである．また，図 6.2 における詳細な F 値や Leave-one-out 交差検定によるパラメータの最適値は付録 B の表 B.1 に掲載している．

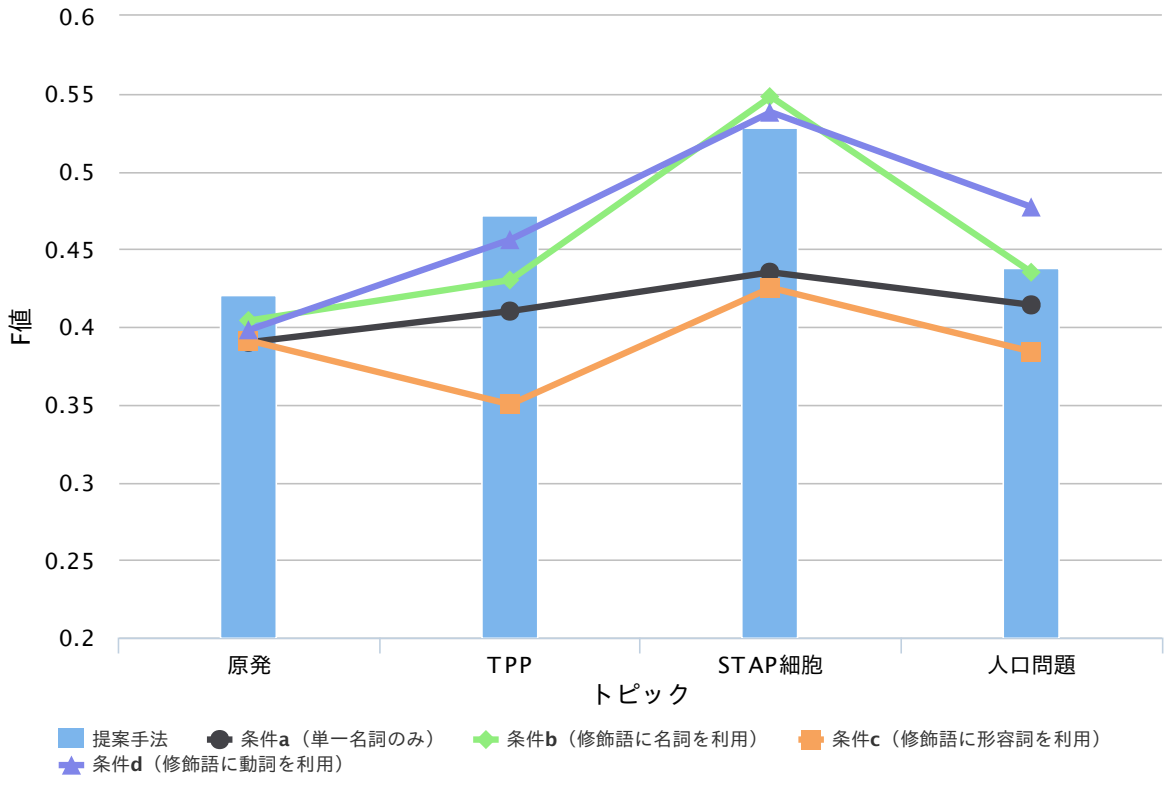

図 6.2 修飾語の種類による F 値の変化

図 6.2 より，トピック「STAP 細胞」では条件 b（名詞のみを修飾語に利用する）ときで最も良い性能を示し，トピック「人口問題」では条件 d（動詞のみを修飾語に利用する）ときで最も良い精度を示したが，他のトピックでは提案手法で，つまり自立語すべてを修飾語に利用することでクラスタリング性能の向上につながったことが示された．

条件 c（形容詞のみを修飾語に利用する）における性能は，提案手法に比べて全体的に低くなっていたが，これは，もとより意見中に出現する形容詞が少なかったためだと考えられる．また，条件 c における F 値を取ったときのパラメータ β[*18]の値は正解データの半分以上が 0 となっていたことから，名詞・動詞ペアどうしの類似度計算において修飾語としての形容詞の情報はあまり影響がないと推測される．

トピック「STAP 細胞」においては，実際に抽出された名詞・動詞ペアを分析したところ，修飾語となった名詞がその意見の観点を直接示しているということが他のトピックより多く見られた．例えば，「組織体制」という観点が人手で付与された意見では，「組織」や「体制」といった名詞の修飾語がそのまま含まれていた．また，「STAP 細胞」では，条件 d の動詞のみを修飾語に利用する場合でも提案手法より（若干ではあるが）高い性能を示していた．

トピック「人口問題」では，条件 d の動詞のみを修飾語に利用するときで最も良い精度を示した．しかし，6.1.1 節で述べたように，「人口問題」では非自立語扱いの動詞が多く抽出されていたことを踏まえると，観点を特徴づける情報でもないような単語が修飾語になったとしても提案手法より精度が高くなるとは考えにくい．そこで，「人口問題」において実際に抽出された名詞・動詞ペアを分析したところ，修飾語には非自立語扱いの動詞はほとんど含まれていなかった．つまり，「ある」や「なる」といった非自立語扱いの動詞は文末に出現することが多いために名詞を修飾する（名詞を含む文節に係る）ことがほとんどなく，逆に観点の差異を反映するような動詞だけが修飾語として抽出されたことが精度の向上に繋がったと考えられる．

先に述べたように，修飾語としての形容詞の情報は名詞・動詞ペア間の類似度計算にあまり影響を与えないと考えると，トピック「STAP 細胞」や「人口問題」では名詞・形容詞・動詞を修飾語としている提案手法においても条件 b,d と同等の性能が示されてもおかしくないはずである．しかし，実際には条件 b,d に比べて提案手法の方が性能が低くなっている．これは，複数の品詞を修飾語に用いると，異なる品詞の修飾語どうしが干渉し合って意図しない修飾語どうしで不当に類似度が高くなってしまったことが原因だと考えられる．提案手法では「消費」や「開発」といったサ変可能名詞[*19]のような単語が文中でどの品詞で用いられているかを同定しているが，複

[*18] β の値が小さいほど修飾語の影響が小さくなる
[*19] 名詞の直後に動詞の「する」が付くことで動詞化するもの

数の品詞を修飾語に用いると，修飾語どうしの類似度計算において品詞の違いが考慮されない場合が出てくる．例えば，2 つの意見に「消費」という単語が出現したとき，片方の意見では名詞として，他方の意見では動詞としての機能を担うものと同定されたとしても，名詞と動詞を修飾語に用いると，同定された品詞に関わらず「消費」という単語どうしの類似度が計算されてしまう．そのため，修飾語どうしの類似度計算において，修飾語の品詞を考慮するよう計算方法を考案する必要がある．

一方，トピック「原発」においては，利用する修飾語の品詞による精度の変化が小さい結果となった．また，このトピックでは，修飾語を利用しない場合（条件 a）においても提案手法との精度に差はあまり見られなかった．この要因としては，6.1.2 節で述べたように，修飾語だけでなく被修飾語（単一名詞を含む）にも観点を特徴づけるような語が含まれていたことが挙げられる．

6.5　正解クラスタ群について

手法の評価をするにあたり，正解データ（クラスタ群）を作った作成者の間でクラスタ内の意見がどの程度重複しているかを調べる必要がある．仮に，作成者の間でバラバラな観点のクラスタが生成され，評価実験で高い精度が得られたとしても，それは個々の作成者に依存した結果，つまり一般性を欠いた結果とみなされる可能性がある．

評価（作成）者間のデータの一致度を見る統計量としては κ 統計量があるが，これは予め決めたカテゴリに分類した際に評価者間の一致度合いを見るものであるため，本研究のように評価者によってカテゴリ数（観点の数）が異なる場合には利用できない．

そこで，本研究ではクラメールの連関係数を利用してその一致度合いを調べた．クラメールの連関係数は分割表[*20]の行と列の項目の関連度合いを測る指標 [金 09] であり，この係数 V は分割表から求めた χ^2 値をもとに以下のように計算される．

$$V = \sqrt{\frac{\chi^2}{N(k-1)}} \quad (0 \leq V \leq 1) \tag{6.1}$$

N は分割表における総度数，k は分割表の行数と列数の小さい方を指す．

例えば，図 6.3 のように作成者 A と作成者 B によりそれぞれクラスタ群 $\boldsymbol{A}, \boldsymbol{B}$ が作成されたとき，分割表は表 6.1 のようになる．なお，表 6.1 のセル A_1B_1 の値はクラスタ A_1 とクラスタ B_1 に共通して属する要素（意見）の数を示している．

この分割表から各セルの期待値を計算すると，表 6.2 のようになる．なお，各セルの期待値は，

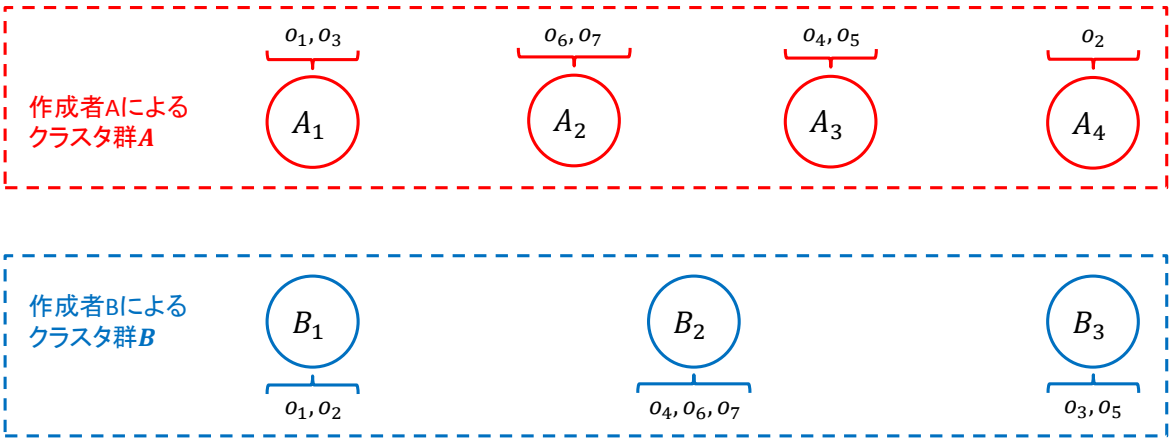

図 6.3　作成者 A と作成者 B により作成されたクラスタ群の例

[*20] 任意の行カテゴリ r と列カテゴリ c にともに属するデータの数を記録した表

表 6.1　図 6.3 における分割表

		$\boldsymbol{A}$				行合計
		A_1	A_2	A_3	A_4	
$\boldsymbol{B}$	B_1	1	0	0	1	2
	B_2	0	2	1	0	3
	B_3	1	0	1	0	2
列合計		2	2	2	1	7

表 6.2　表 6.1 の期待値表

		$\boldsymbol{A}$				行合計
		A_1	A_2	A_3	A_4	
$\boldsymbol{B}$	B_1	4/7	4/7	4/7	2/7	2
	B_2	6/7	6/7	6/7	3/7	3
	B_3	4/7	4/7	4/7	2/7	2
列合計		2	2	2	1	7

表 6.3　各トピックごとの作成者ペア間の連関係数 V

トピック	ペア	連関係数 V
原発	A − B	0.707
	A − C	0.735
	B − C	0.789
平均		0.744
TPP	A − B	0.762
	A − C	0.730
	B − C	0.628
平均		0.707
STAP 細胞	A − B	0.713
	A − C	0.696
	B − C	0.773
平均		0.727
人口問題	A − B	0.617
	A − C	0.706
	B − C	0.698
平均		0.674

該当行の合計と該当列の合計の積を総度数で割った値である．

このとき，表 6.1 および表 6.2 のセル A_1B_1 の値をそれぞれ $R(A_1, B_1), E(A_1, B_1)$ とすると，χ^2 値は次のように求まる．

$$\chi^2(\boldsymbol{A}, \boldsymbol{B}) = \sum_{i=1}^{|\boldsymbol{A}|} \sum_{j=1}^{|\boldsymbol{B}|} \frac{\left(R(A_i, B_j) - E(A_i, B_j)\right)^2}{E(A_i, B_j)} \simeq 7.58$$

なお，$|\boldsymbol{A}|, |\boldsymbol{B}|$ はそれぞれクラスタ群 $\boldsymbol{A}, \boldsymbol{B}$ のクラスタ数を指す．

以上より，図 6.3 における作成者 A と作成者 B の連関係数 V は式 (6.1) から，以下のように求まる．

$$V(A, B) = \sqrt{\frac{7.58}{7 \times (3-1)}} \simeq 0.736$$

この連関係数の最大値は 1 であることから，図 6.3 で示した例においては，作成者 A により作成されたクラスタ群と作成者 B により作成されたクラスタ群は比較的一致していると言える．

以上から，本研究における正解データに対して計算した連関係数 V を表 6.3 に示す．なお，表中の「ペア」列は 3 名の作成者 A〜C から 2 名を選んだときのペアを指す．

いずれのトピックおよびペアでも連関係数がおよそ 0.7 前後であったことから，作成者間でクラスタリング結果にばらつきはあまりないと考えられる．

7 ツイートへの応用

本章では，マイクロブログサービスの 1 つである Twitter に存在する意見（以下，意見ツイート）に対して，本研究における提案手法を適用した観点に基づくクラスタリング手法（以下，本手法）について述べる．

本研究で提案した名詞・動詞ペアを用いることに変わりはないが，Twitter では 1 ツイートあたり 140 文字の字数制限があるため，意見ツイートどうしの類似度を計算するには情報量が少なく，適切にクラスタリングすることができないという問題点がある．そこで，意見ツイートどうしの類似度を適切に計算するのに十分な情報を得るために，意見ツイートに関連するユーザのツイートを考慮し，意見の観点に基づいてクラスタリングを行う．

7.1　マイクロブログサービス

マイクロブログサービスの代表的なサービスとして，本章でも扱う Twitter がある．これらのマイクロブログサービスを従来のブログサービス（Yahoo ブログ[*21]など）と比較したとき，大きく分けて以下の 3 点の違いがある．

- リアルタイム性
- 字数制限
- follow（購読機能）

まず，マイクロブログサービスの特徴の 1 つとして，高いリアルタイム性が挙げられる．従来のブログサービスでは長い文章量を持つ記事が投稿されることがあるため，ユーザによって投稿時間に差が生じる場合がある．しかし，マイクロブログでは 1 日に何度も書き込まれることが一般的であり，あるイベントやテレビ番組を見ている間にユーザが実況として逐次「つぶやき（ツイート）」を投稿することが多い．

また，従来のブログサービスでは，文字数制限は特に設けられておらず，ユーザが書けるだけ記事を書くことができる．しかし，マイクロブログサービスでは，文字数制限（Twitter の場合は 140 字）が設けられており，ある出来事に対して一回の投稿で全てを記述することが困難となっている．そのため，従来のブログサービスでは一回の投稿で済んでいた内容が，複数のツイートによって構成されることがある．

Twitter では，他のユーザを follow することでそのユーザのツイートを購読することができる．これにより，ユーザは他のユーザのツイートをリツイート[*22]することができる．また，他のユーザ ID を参照（「**@ユーザ ID**」と表記される）することで，他のユーザと簡単にやりとり（リプライ）することができる．

[*21] http://blogs.yahoo.co.jp/
[*22] E メールの転送にあたる機能

7.2 関連研究

2 章の関連研究でも述べたように，Twitter 等を問わず，Web 上の意見を観点に基づいてクラスタリング手法を提案している研究は行われていない．

Twitter を対象とした観点に基づく分類は行われていないが，Twitter を対象とした意見や評判の自動分類（センチメント分析）の研究 [橋本 11, Jiang 11] は近年行われている．

橋本ら [橋本 11] は，意見ツイート中に用いられる評価表現の違いをセンチメント分析により抽出し，喜びや悲しみなど 10 種類の感情極性を用いて意見ツイートを感情ごとに分類している．しかし，これはセンチメント分析による感情表現しか見ていないため，意見の内容から観点ごとに分類する本手法とは異なる．

また，Jiang ら [Jiang 11] は，製品の評判情報などを対象として，評判文（以下，評判ツイート）を肯定/否定/中立の 3 値に分類している．Jiang らは，分類を行う際に，評判ツイートの周りに存在する評判ツイートに関連したリプライやリツイートを考慮している．これから述べるクラスタリング手法においてもユーザの意見ツイートの周りに存在するツイートを考慮するが，肯定/否定/中立の分類が目的ではない．

7.3　意見ツイートのクラスタリング手法

本手法では，ある特定のトピック（時事問題）に関する意見ツイートの集合について，1 つの意見ツイートに単一の観点が付与されると仮定し，排他的なクラスタリングを行う．

本章で提案する意見ツイートのクラスタリング手法の手順を以下に示す．

1. あるユーザ x_i がつぶやいた意見ツイート o_i の周りに存在するユーザ x_i のツイート集合 A_i から，意見に関連するツイート集合（以下，関連ツイート集合）R_i を抽出する．
2. 各意見ツイート o_i および，各関連ツイート集合 R_i への係り受け解析で得られた文節の係り受け関係から，それぞれ名詞・動詞ペアを抽出する．
3. 各意見ツイート o_i とその関連ツイート集合 R_i をそれぞれに含まれる名詞・動詞ペア集合 $\boldsymbol{P}(o_i), \boldsymbol{P}(R_i)$ で表現し，意見ツイート間の類似度を名詞・動詞ペア集合間の類似度として計算する．
4. 手順 3 で計算した意見ツイート間の類似度を用いて，Ward 法による階層型クラスタリングを行う．

7.3.1 ツイートへの前処理

ツイートには URL やリプライ（返信）用の表記など雑多な情報が含まれている．そのため，関連ツイートの抽出や名詞・動詞ペアの抽出の前に，ツイートに対して以下のような前処理を行う．

1. 句点に相当する文を区切る文字（全角スペースや「！？♪…」といった記号など）をすべて句点「。」に変換する．
2. URL やリプライ表記（**@ユーザ ID**）の表記を削除する．
3. 1,2 を適用したツイートに対して Unicode 正規化[*23]を行う．

ツイートにはつぶやいたユーザ自身が自分のツイートに対してハッシュタグ（**#タグ名**）というタグを付けることができ，ツイートにもそのタグ文字列が含まれる．一般的にツイートへの処理においてはハッシュタグの情報は削除されるが，本手法では意見ツイートに付けられたハッシュタグは 1 つの意見表明として捉えることができると考え，これを残すことにする．

また，ツイートの特徴として引用リツイート[*24]があり，鷹栖ら [鷹栖 13] の研究においては引用ツイートの情報を削除していたが，本手法では引用されたツイートにも観点を特徴づける情報が含まれると考え，残すことにする．

[*23] 全角の英数字を半角に統一したり，半角のカタカナを全角に統一すること
[*24] 他のユーザがつぶやいたツイートを引用した上でつぶやくもの

7.3.2 関連ツイートの抽出

Twitter では，文字数の制限や投稿のしやすさから，あるトピックに対するツイートが連続して投稿されることがある．つまり，複数回の投稿により 1 つの意見が構成される場合があるため，意見に関連する複数の（投稿）ツイートを抽出し，クラスタリングに利用する．

関連ツイートの抽出には，鷹栖ら [鷹栖 13] と同様に，意見ツイート o_i とその周りに存在するツイート $a_{ij} \in A_i$ 間の 3 種類の類似度を利用する．

sim_b： 文字 bigram 単位の意味空間上でベクトル表現された o_i と a_{ij} のコサイン類似度

sim_m： 形態素単位の意味空間上でベクトル表現された o_i と a_{ij} のコサイン類似度

sim_t： o_i と a_{ij} の投稿時間による時間類似度

時間類似度は，戸田ら [戸田 07] の手法を利用する．戸田らは，「文書間のタイムスタンプが離れるごとに，一定の割合で文書内容の類似度が減少する」という仮定に基づいて時間類似度を式 (7.1) のように定義している．

$$\mathrm{sim}(t_P, t_Q) = T_0 \times \exp\left(-\frac{0.693}{t_{1/2}}|t_P - t_Q|\right) \tag{7.1}$$

t_P, t_Q はそれぞれ単位を日とした記事 P, Q のタイムスタンプ，T_0 は $t_P = t_Q$ のときの重み，$t_{1/2}$ は時間類似度が 50% になるタイムスタンプの差（半減期）を示すパラメータである．複数のツイートから構成される意見は，まとまった時間内にツイートされると考えられるので，本手法においても戸田らの手法を利用する．なお，本手法ではタイムスタンプの単位を秒とする．

先に挙げた 3 種類の類似度の重みとして，パラメータ μ, ν, ρ $(\mu + \nu + \rho = 1)$ を用いて o_i と a_{ij} の類似度 $\mathrm{sim}(o_i, a_{ij})$ を式 (7.2) のように定義し，類似度が閾値 T を超えた a_{ij} を関連ツイートとして抽出する．

$$\mathrm{sim}(o_i, a_{ij}) = \mu \times \mathrm{sim}_b(o_i, a_{ij}) + \nu \times \mathrm{sim}_m(o_i, a_{ij}) + \rho \times \mathrm{sim}_t(o_i, a_{ij}) \tag{7.2}$$

7.3.3 名詞・動詞ペアの抽出

意見ツイート集合およびそれらの関連ツイート集合に対して係り受け解析を行い，得られた文節の係り受け関係から名詞 N とそれが係る動詞 V のペア（名詞・動詞ペア）$\langle N, V \rangle$ を抽出する．詳細な抽出方法は 4.3 節と同様である．

なお，ツイートは文字数が少なく，文法が不完全という特徴があるため，ツイートによっては係り受け解析が正しくできず，名詞・動詞ペアを抽出できない可能性がある．そのため，1 ツイート中で名詞・動詞ペアを抽出できなかった場合は，動詞との係り受け関係を考慮せずに名詞のみを抽出し，4.6 節で述べた名詞・動詞ペア間の類似度計算に用いる．このような場合，動詞の情報を利用することができないため，動詞どうしの類似度は 0 とする．

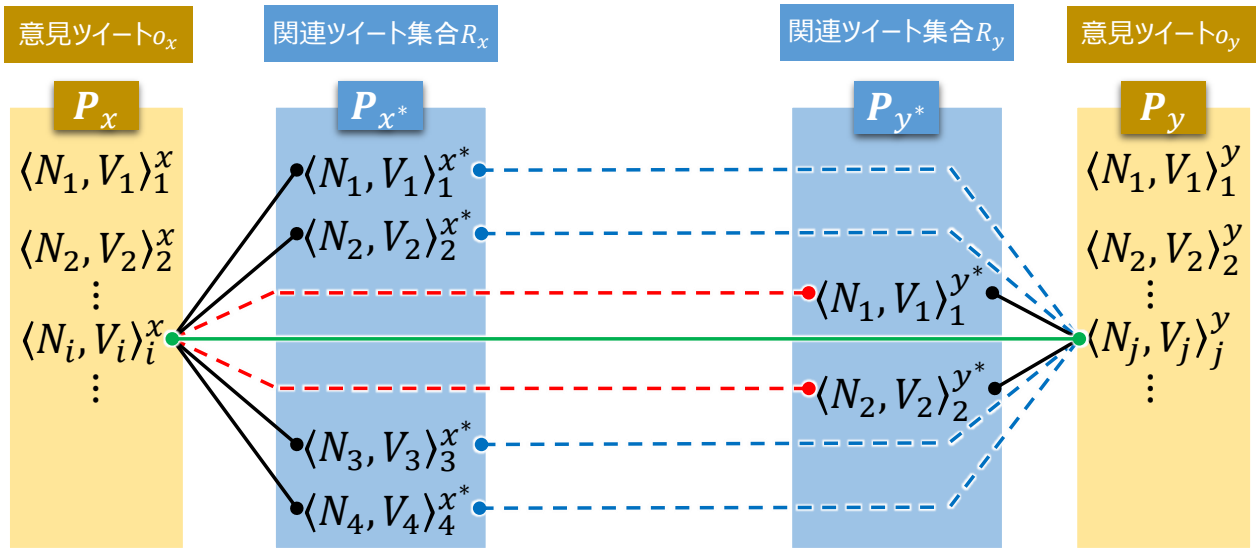

図 7.1　名詞・動詞ペア間の関係図

7.3.4　意見ツイートどうしの類似度計算

意見ツイート o_x, o_y から抽出した名詞・動詞ペアの集合をそれぞれ $\boldsymbol{P}_x = \{\langle N_i, V_i\rangle_i^x\}, \boldsymbol{P}_y = \{\langle N_j, V_j\rangle_j^y\}$ とし，o_x, o_y の関連ツイート集合 R_x, R_y から抽出した名詞・動詞ペアの集合をそれぞれ $\boldsymbol{P}_{x^*} = \{\langle N_k, V_k\rangle_k^{x^*}\}, \boldsymbol{P}_{y^*} = \{\langle N_l, V_l\rangle_l^{y^*}\}$ とする．

今，関連ツイート集合 R_x は意見ツイート o_x に関連するものであることから，$\langle N_k, V_k\rangle_k^{x^*}$ は o_x が示す観点を特徴づける材料であると仮定すると，$\langle N_j, V_j\rangle_j^y$ と $\langle N_k, V_k\rangle_k^{x^*}$ の類似度が高いとき，$\langle N_k, V_k\rangle_k^{x^*}$ は意見ツイート o_y が示す観点を特徴づける材料でもあると言える．つまり，2 つの意見ツイートが示す観点の材料が同じであれば，その意見ツイートどうしは観点が似ていると考えることができる．

以上のことを示す意見ツイートおよび関連ツイート集合の名詞・動詞ペア間の関係を図 7.1 に示す．なお，図 7.1 では例として $|\boldsymbol{P}_{x^*}| = 4, |\boldsymbol{P}_{y^*}| = 2$ としている．

図 7.1 の点線・実線はそれぞれ名詞・動詞ペア間の類似度を示している．赤と青の点線は，片方の意見ツイートに含まれる $\langle N, V\rangle$ と他方の意見ツイートの関連ツイート集合に含まれる $\langle N, V\rangle$ との類似度を示しており，この類似度が高くなるほど 2 つの意見ツイートが示す観点の材料が同じだと言える．なお，黒の実線は，関連ツイート集合に含まれる名詞・動詞ペアが意見ツイートの示す観点を特徴づけるものであることを指す．意見ツイート o_x, o_y から抽出した任意の名詞・動詞ペア $\langle N_i, V_i\rangle_i^x, \langle N_j, V_j\rangle_j^y$ 間の最終的な類似度は，赤の点線における最大類似度と青の点線における最大類似度，緑の実線で示される類似度の平均とする．

以上のことから，意見ツイートどうしの類似度 $\mathrm{sim}_o\left(o_x, o_y\right)$ を式 (7.3) のように定義する．

$$
\begin{aligned}
\mathrm{sim}_o\left(o_x, o_y\right) &= \frac{\mathrm{nvSim}_x + \mathrm{nvSim}_y}{|\boldsymbol{P}_x| + |\boldsymbol{P}_y|} \qquad (7.3)\\
\mathrm{nvSim}_x &= \sum_{i=1}^{|\boldsymbol{P}_x|} \max_j \left[M\left(\langle N_i, V_i\rangle_i^x, \langle N_j, V_j\rangle_j^y\right)\right]\\
\mathrm{nvSim}_y &= \sum_{j=1}^{|\boldsymbol{P}_y|} \max_i \left[M\left(\langle N_i, V_i\rangle_i^x, \langle N_j, V_j\rangle_j^y\right)\right]
\end{aligned}
$$

nvSim_x は意見ツイート o_x の各名詞・動詞ペア $\langle N_i, V_i\rangle_i^x$ に対する意見ツイート o_y の $\boldsymbol{P}_y$ との最大類似度の和である．nvSim_y は逆に，意見ツイート o_y の各名詞・動詞ペア $\langle N_j, V_j\rangle_j^y$ に対する意見ツイート o_x の $\boldsymbol{P}_x$ との最大類似度の和である．なお，意見ツイート間の名詞・動詞ペアどうしの類似度は式 (7.4) を満たす関数 M により計算される．

$$
\begin{aligned}
M\left(\langle N_i, V_i\rangle_i^x, \langle N_j, V_j\rangle_j^y\right) &= \frac{m_0 + m_1 + m_2}{1 + f(\boldsymbol{P}_{x^*}) + f(\boldsymbol{P}_{y^*})} \qquad (7.4)\\
m_0 &= \mathrm{sim}_{nv}\left(\langle N_i, V_i\rangle_i^x, \langle N_j, V_j\rangle_j^y\right)\\
m_1 &= \begin{cases} \max_l \left[\mathrm{sim}_{nv}\left(\langle N_i, V_i\rangle_i^x, \langle N_l, V_l\rangle_l^{y^*}\right)\right] & (\boldsymbol{P}_{y^*} \neq \varnothing)\\ 0 & (otherwise) \end{cases}\\
m_2 &= \begin{cases} \max_k \left[\mathrm{sim}_{nv}\left(\langle N_j, V_j\rangle_j^y, \langle N_k, V_k\rangle_k^{x^*}\right)\right] & (\boldsymbol{P}_{x^*} \neq \varnothing)\\ 0 & (otherwise) \end{cases}\\
f(\boldsymbol{P}) &= \begin{cases} 0 & (\boldsymbol{P} = \varnothing)\\ 1 & (otherwise) \end{cases}
\end{aligned}
$$

m_0 は緑の実線が示す類似度を，m_1, m_2 はそれぞれ赤の点線における最大類似度と青の点線における最大類似度を表している．関数 M は m_0, m_1, m_2 の平均を返すが，関連ツイート集合が空の場合は，m_1, m_2 を平均から除くようにしている．

7.4　評価実験

評価実験には，鷹栖ら [鷹栖 13] の研究で用いた「原発」「衆議院選挙」「尖閣諸島」の 3 つのトピックに関する意見ツイートを利用した．また，各トピックにおいて 2 名ずつ個別に観点ごとに意見ツイートを分類してもらい，各トピックごとに 2 種類のクラスタリングの正解データを用意した．この評価実験においても，人手による分類と同じ観点の数で提案手法によるクラスタリングを行った．なお，評価指標については，5.3 節と同様である．

7.4.1　比較手法

比較手法には，鷹栖ら [鷹栖 13] の手法と LSI 法を用意した．

鷹栖らの手法では，LSI により構築した形態素単位の意味空間を利用して，各意見ツイートおよび各関連ツイートの特徴ベクトルを生成する．最終的には意見ツイートとその関連ツイート集合の特徴ベクトルの重心ベクトルを意見ツイートの特徴ベクトルとして意見ツイート間のコサイン類似度を計算し，Ward 法による階層型クラスタリングを行う．意味空間はすべての意見ツイート O と関連ツイート集合全体 $\boldsymbol{R}$ に含まれる単語の出現頻度を要素とする単語・文書行列に対して次元圧縮を施して構築する．なお，関連ツイート $r_{ij} \in R_i$ については，行列の要素として，r_{ij} に含まれる単語の出現頻度に r_{ij} と意見ツイート o_i の類似度を掛けた値を用いる．また，鷹栖らと同様に各意見ツイートの関連ツイート集合は，人手で関連ツイートと判定されたものを用いた．

LSI 法は，関連ツイート集合の情報を利用せず，LSI により構築した意味空間から生成した意見ツイートのみの特徴ベクトルを用いて意見ツイート間のコサイン類似度を計算し，Ward 法による階層型クラスタリングを行う方法である．意味空間は各意見ツイートに含まれる単語の出現頻度を要素とする単語・文書行列に対して次元圧縮を施して構築した．

7.4.2 実験結果

関連ツイート集合については，鷹栖ら [鷹栖 13] における関連ツイート抽出の評価実験において，抽出性能が最も高くなったときの関連ツイート集合をそのまま用いた．

本手法については，パラメータ α, β をそれぞれ 0〜1 の 0.1 刻みで変化させてクラスタリングを行った．また，本手法・比較手法ともに LSI により構築した意味空間の次元数を 5 から全意見ツイート数（ただし，鷹栖らの手法では意見ツイートと関連ツイートの総数）までの 5 刻みで変化させた．名詞・動詞ペアどうしの類似度計算（式 (4.7)）におけるパラメータ λ については，その値を変動させずに $\lambda = 2/3$[*25] と固定した．

なお，各トピックにおいて F 値は正解クラスタ群ごとに計算するが，先述したように本手法や比較手法ではパラメータや意味空間の次元数を変化させてクラスタリングを行っていることから，交差検定を用いて F 値を計算した．交差検定による F 値の計算方法は，5.5.1 節と同様である．

以上の実験結果を表 7.1 に示す．なお，表中の k はクラスタ（観点）数を指し，太字の数値は各行で最も高かった F 値を指す．また，「パラメータの最適値」列は，交差検定により学習されたパラメータであり，パラメータ d_p, d_{c1}, d_{c2} は，それぞれ本手法，鷹栖らの手法，LSI 法における意味空間の次元数を指す．

すべての正解データにおいて本手法で最も高い F 値が得られたことから，関連ツイート集合と名詞・動詞ペアを用いた本手法は，意見ツイートを対象とした観点に基づくクラスタリング手法として有用であると言える．

[*25] 予備実験で $\lambda = 1/3, 1/2, 2/3$ と変化させた結果，2/3 のときに最も良い精度を示したことから，本実験でも $\lambda = 2/3$ とした．なお，本研究における評価実験（表 5.4,5.5）においても λ は 0.5〜0.9 の値を取っていることから，$\lambda = 2/3 \simeq 0.7$ は比較的妥当な値だと考えられる．

表 7.1　意見ツイートへのクラスタリングの評価実験結果

トピック	k	F 値			パラメータの最適値		
		本手法	鷹栖ら [鷹栖 13]	LSI	提案手法 (α, β, d_p)	鷹栖ら (d_{c1})	LSI (d_{c2})
原発	7	**0.406**	0.259	0.217	$0.4, 0.8, 20$	30	5
	9	**0.397**	0.334	0.232	$0.5, 0.6, 20$	10	20
平均		**0.401**	0.297	0.225	-		
衆議院選挙	$5_{(1)}$	**0.853**	0.710	0.710	$0.5, 0.6, 5$	15	10
	$5_{(2)}$	**0.867**	0.750	0.711	$0.5, 0.6, 5$	30	10
平均		**0.860**	0.730	0.711	-		
尖閣諸島	10	**0.522**	0.466	0.441	$0.4, 0.8, 5$	5	10
	11	**0.507**	0.412	0.427	$0.2, 0.9, 15$	5	5
平均		**0.515**	0.439	0.434	-		

7.5 考察

7.5.1 関連ツイートと名詞・動詞ペアの有用性

本手法において，関連ツイートや名詞・動詞ペアの利用がクラスタリング性能にどのような影響を与えるかを調べるために，以下の 3 つの条件におけるクラスタリング性能を求め，評価実験で得られた本手法の性能と比較した．

条件 1 関連ツイートの情報を利用せずに類似度を計算する．（すなわち，式 (7.4) において m_0 のみを計算する．）

条件 2 名詞・動詞ペアの名詞のみを利用して類似度を計算する．（すなわち，式 (4.7) において $\mathrm{sim}_v = 0$ とする．）

条件 3 条件 2 に加えて，名詞・動詞ペアの（複合）名詞を，動詞との係り受け関係を考慮せずに形態素解析上で名詞と解析されたすべての語をもとにして抽出する．

特に，条件 2 における性能を求めることで，動詞どうしの類似度がクラスタリング性能に与える影響を調べることができる．また，条件 3 で抽出された（複合）名詞には文中で動詞の機能を担う語（「開発する」の「開発」などのサ変可能名詞[*26]）が含まれることから，条件 2 と性能を比較することで，動詞との係り受け関係を考慮した名詞の抽出がクラスタリング性能に与える影響を調べることができる．

以上の各条件におけるクラスタリング性能を図 7.2 に示す．

図中の本手法（青い棒グラフ）と条件 1 を比較すると，トピック「衆議院選挙」では同じ精度となったが，他のトピックでは本手法で最も良い精度が得られたことから，関連ツイートの情報を利用することは有用であると言える．本手法と条件 2 の比較においても，トピック「衆議院選挙」では同じ精度となったが，他のトピックでは本手法で最も良い精度が得られたことから，動詞どうしの類似度を考慮することで，より観点に基づいたクラスタリングができることが示された．条件 1 と条件 2 を比較すると，トピック「衆議院選挙」では同じ精度となったが，他のトピックでは条件 1 の方が精度が良いことから，本手法の精度向上には関連ツイートの利用が大きく作用したと考えられる．また，条件 2 と条件 3 を比較すると，すべてのトピックにおいて条件 2 の方が精度が良いことから，動詞との係り受け関係を用いて抽出した名詞をクラスタリングに利用することは有用であると言える．

[*26] 名詞に動詞「する」が付属することで動詞化するもの

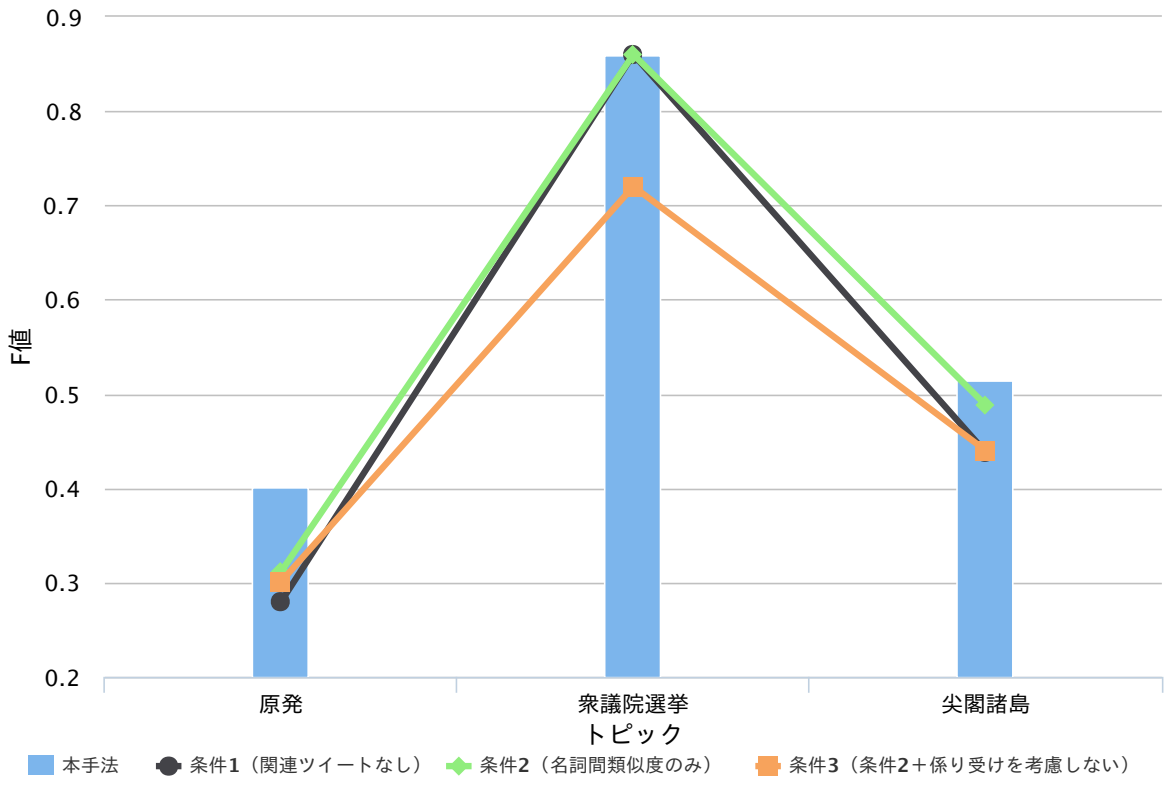

図 7.2　条件 1〜3 における実験結果

トピック「衆議院選挙」では，本手法と条件 1,2 で同じ精度となり，比較手法においても他のトピックに比べて高い精度が得られた．このことから，「衆議院選挙」では，クラスタリングするには意見ツイートに含まれる情報（名詞・動詞ペア）だけで十分であり，意見ツイートやその関連ツイート集合に含まれる名詞が直接意見の観点を示すことが多いという特徴があると考えられる．実際，鷹栖ら [鷹栖 13] の名詞を利用したクラスタへのラベリング*27における評価実験では，「衆議院選挙」で最も良い精度が得られていた．

トピック「衆議院選挙」では，ツイートに含まれる名詞が直接意見の観点を示すことが多いという特徴があるものの，本手法と条件 3 を比較すると，本手法の方が精度が良いことから，動詞との係り受け関係を用いて抽出した名詞の情報，つまり間接的に動詞の情報がクラスタリングに有用であると言える．

実験結果とは関係ないが，本実験で用いたトピック「衆議院選挙」のツイートは，2012 年に行われた衆議院選挙に対するツイートであるため，昨年（2014 年）に行われた衆議院選挙に対するツイートではない．これらの 2012 年と 2014 年における衆議院選挙に対する意見の観点を比較し，世論の政治に対する観点の変化などを分析することで興味深い結果が得られる可能性がある．

*27 各クラスタを 1 つの文書とみなして，クラスタごとに名詞の TF-IDF 値を計算し，その上位 3 件を出力している．

7.5.2 エラー分析

正しくクラスタリングができなかった主な原因として，意見とは関係のない関連ツイート集合が抽出されたことが挙げられる．関連ツイート集合については，鷹栖ら [鷹栖 13] における関連ツイート抽出の評価実験において，抽出性能が最も高くなったときの関連ツイート集合をそのまま用いているため，意見とは関係のない関連ツイートが少なからず含まれていた可能性がある．つまり，意見とは関係のない名詞・動詞ペアが類似度計算に利用されてしまったために，異なる観点の意見どうしで類似度が高くなってしまったと考えられる．

また，1 つの意見に異なる観点を示す名詞・動詞ペア $\langle N, V \rangle$ が含まれていたために，誤ったクラスタに意見が属してしまったことも原因の 1 つだと考えられる．例えば，ある意見に「安全性」という観点を示す $\langle N, V \rangle$ と「政治」という観点を示す $\langle N, V \rangle$ が含まれているとき，正解データではその意見が「安全性」という観点を示すクラスタに属していたとしても，排他的なクラスタリングを行うと「政治」という観点を示すクラスタに属してしまうことがある．そのため，非排他的なクラスタリングに適するような類似度の計算方法を考案する必要がある．

一方で，意見の観点を示すような名詞・動詞ペアが抽出できていないツイートも見られた．ツイートは文が短く，文法が不完全という特徴があることから，係り受け解析が正しくできず，ツイートから名詞・動詞ペアが抽出できない場合がある．そこで本手法では，1 つのツイートから名詞・動詞ペアが抽出されなければ，動詞との係り受け関係を考慮せずに名詞のみを抽出し，これを名詞・動詞ペアどうしの類似度計算に利用している．1 つのツイートは，複数の文からなる場合もあることから，名詞・動詞ペアの抽出自体は文単位で行っているが，文によっては意見の観点とは全く関係のない名詞・動詞ペアが抽出されてしまうことがある．そのため，意見の観点を特徴づけるような語を含む文からは名詞・動詞ペアが抽出できず，意見の観点とは関係のない語を含む文からペアが抽出されてしまうと，1 つのツイートとしては不適切なペアが抽出されてしまうことになる．

例えば，以下の意見ツイートは 3 つの文から成り立っている．

> **全く滅茶苦茶ですよね。尖閣諸島は明らかに日本の領土です。良く纏まった文書がありますので参考にして下さい。**

また，この意見ツイートから抽出される名詞・動詞ペア $\langle N, V \rangle$ は以下の 2 つである．

> ⟨⟨ **纏まる**, **文書** ⟩, **ある** ⟩, ⟨ **参考**, **する** ⟩

しかし，これらの抽出された名詞・動詞ペアは，3 文目からしか抽出されていない．この意見が仮に「領土の主張」という観点を示すのであれば，2 つの目の文から名詞・動詞ペア（実際には，2 つ目の文には動詞が含まれないので名詞のみ）が抽出されるのが望ましい．もし 3 つ目の文からも名詞・動詞ペアが抽出されなければ，ツイートに含まれる名詞だけの情報が抽出されるが，実際には 3 つ目の文からは名詞・動詞ペアが抽出されているために，「ツイート」として持つ情報は上記の 2 つのペアだけになってしまう．つまり，適切な類似度計算ができなくなってしまう．

特に，トピック「尖閣諸島」ではこのような事例が他のトピックに比べて多く，このこともクラスタリングが正しくできなかった原因の 1 つだと思われる．対処策としては，ペア抽出の有無の判定をツイート単位ではなく文単位にすることや，ペア（動詞）抽出の条件を緩める方法が考えられる．

8 おわりに

本研究では，従来の文書クラスタリング手法に用いられる Bag of Words や TF-IDF 値を用いずに，文節の係り受け関係から抽出した名詞・動詞ペア $\langle N, V \rangle$ を用いることで，意見集合を観点に基づいてクラスタリングする手法を提案した．

評価実験より，すべてのトピックにおいて従来の文書クラスタリング手法より高い精度を得ることができ，提案手法の有用性を確認することができた．また，動詞との係り受け関係を用いて，名詞の文中における役割（品詞）を同定することや修飾語を用いた複合名詞の利用も類似度計算に有用であることが分かった．

Twitter における意見ツイート集合に対しても，意見に関連するツイートを利用し，提案手法を適用してクラスタリングを行うことで高い精度を得られたことから，提案手法の汎用性を確認することができた．

今後の課題として，名詞・動詞ペア $\langle N, V \rangle$ の抽出方法の改善や $\langle N, V \rangle$ どうしのより優れた類似度計算方法，複数のクラスタに属することを許容する非排他的クラスタリング手法に適した類似度計算方法の考案が挙げられる．また，本研究では，生成されたクラスタがどのような観点を示す意見集合なのか分かりづらいことから，クラスタへのラベリング手法の考案も今後の課題である．

参考文献

[Anand 11] Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., and Minor, M.: Cats Rule and Dogs Drool!: Classifying Stance in Online Debate, in *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pp. 1–9 (2011)

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent Dirichlet Allocation, *The Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[横本 11] 横本 大輔, 林 東權, 牧田 健作, 宇津呂 武仁, 河田 容英, 福原 知宏, 神門 典子, 吉岡 真治, 中川 裕志, 清田 陽司：特定トピックに関するブログ記事集合の観点分類における Wikipedia の利用, 第 3 回データ工学と情報マネジメントに関するフォーラム論文集, DEIM '11（A4-3）(2011)

[Deerwester 90] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R.: Indexing by latent semantic analysis, *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, Vol. 41, No. 6, pp. 391–407 (1990)

[鷹栖 13] 鷹栖 弘明, 小林 聡, 内海 彰：Twitter における観点に基づいた意見文クラスタリング, 言語処理学会 第 19 回年次大会発表論文集 A4-3, pp. 252–255 (2013)

[折原 08] 折原 大, 内海 彰：HTML タグを用いた Web ページのクラスタリング手法, 情報処理学会論文誌, Vol. 49, No. 8, pp. 2910–2921 (2008)

[戸田 07] 戸田 浩之, 北川 博之, 藤村 考, 片岡 良治：時間的近さを考慮した話題構造マイニング, 電子情報通信学会 第 18 回データ工学ワークショップ（DEWS2007） 論文集 L6-4 (2007)

[Hu 04] Hu, M. and Liu, B.: Mining and Summarizing Customer Reviews, in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pp. 168–177 (2004)

[Jiang 11] Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T.: Target-dependent Twitter Sentiment Classification, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 151–160 (2011)

[小熊 05] 小熊 淳一, 内海 彰：語の共起情報を用いた文書クラスタリング, 人工知能学会 第 19 回全国大会論文集 2E1-01 (2005)

[風間 09] 風間 淳一, Saeger, S. D., 鳥澤 健太郎, 村田 真樹：係り受けの確率的クラスタリングを用いた大規模類義語リストの作成, 言語処理学会 第 15 回年次大会発表論文集 C1-6, pp. 84–87 (2009)

[橋本 11] 橋本 和幸, 中川 博之, 田原 康之, 大須賀 昭彦：センチメント分析とトピック抽出によるマイクロブログからの評判傾向抽出, 電子情報通信学会論文誌. D, 情報・システム, Vol. 94, No. 11, pp. 1762–1772 (2011)

[真野 08] 真野 光平, 竹内 孔一：項関係にある名詞との共起を考慮した動詞のクラスタリング, 言語処理学会 第 14 回年次大会発表論文集 B5-1, pp. 1033–1036 (2008)

[村上 07] 村上 浩司, 橋本 泰一, 乾 孝司, 内海 和夫, 石川 正道：共起語に基づいた階層型文書クラスタリング手法, 情報処理学会研究報告. 情報学基礎研究会報告, Vol. 2007, No. 54, pp. 13–20 (2007)

[Lance 67] Lance, G. N. and Williams, W. T.: A general theory of classificatory sorting strategies 1. Hierarchical systems (1967)

[Liu 05] Liu, B., Hu, M., and Cheng, J.: Opinion Observer: Analyzing and Comparing Opinions on the Web, in *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pp. 342–351 (2005)

[Liu 12] Liu, B. and Zhang, L.: A survey of opinion mining and sentiment analysis, in Aggarwal, C. C. and Zhai, C. eds., *Mining Text Data*, pp. 415–463, Springer (2012)

[Luo 09] Luo, Y., Lin, G., and Fu, Y.: Finer Granularity Clustering for Opinion Mining, in *Proceedings of the 2009 Second International Symposium on Computational Intelligence and Design - Volume 01*, ISCID '09, pp. 282–285 (2009)

[金 09] 金 明哲：テキストデータの統計科学入門, 岩波書店 (2009)

[Nguyen 12] Nguyen, D. T., Chen, L., and Chan, C. K.: Clustering with Multiviewpoint-Based Similarity Measure, *IEEE Trans. on Knowl. and Data Eng.*, Vol. 24, No. 6, pp. 988–1001 (2012)

[Oh 09] Oh, A., Lee, H., and Kim, Y.: User Evaluation of a System for Classifying and Displaying Political Viewpoints of Weblogs, in *Proceedings of the Third International ICWSM Conference*, ICWSM '09, pp. 68–71 (2009)

[Pang 02] Pang, B., Lee, L., and Vaithyanathan, S.: Thumbs Up?: Sentiment Classification Using Machine Learning Techniques, in *Proceedings of the ACL-02 Conference on Empirical*

Methods in Natural Language Processing - Volume 10, EMNLP '02, pp. 79–86 (2002)

[Paul 10] Paul, M. J., Zhai, C., and Girju, R.: Summarizing Contrastive Viewpoints in Opinionated Text, in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pp. 66–76 (2010)

[Resnik 95] Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy, in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI '95, pp. 448–453 (1995)

[Scholz 12] Scholz, T. and Conrad, S.: Integrating viewpoints into newspaper opinion mining for a media response analysis, in *Proceedings of KONVENS 2012*, pp. 30–38 (2012)

[Somasundaran 10] Somasundaran, S. and Wiebe, J.: Recognizing Stances in Ideological Online Debates, in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pp. 116–124 (2010)

[Trabelsi 14] Trabelsi, A. and Zaiane, O. R.: Finding Arguing Expressions of Divergent Viewpoints in Online Debates, in *Proceedings of the 5th Workshop on Language Analysis for Social Media*, LASM '14, pp. 35–43 (2014)

[Turney 02] Turney, P. D.: 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*, ACL '02, pp. 417–424 (2002)

謝辞

学部時代から本研究（観点に基づく意見クラスタリング手法の考案）を行うにあたり，3 年間多大なるご指導ご鞭撻を賜りました内海彰教授に深く御礼申し上げます．並びに，多くのアドバイス・議論をして頂いた内海研究室の皆様，評価実験に協力してくださった関係者の皆様にも深く御礼申し上げます．

付録 A　図 6.1 における F 値とパラメータ

表 A.1　図 6.1 における F 値とパラメータ

トピック	k	条件 1		条件 2		条件 3	
		パラメータ (α, β, d_p)	F 値	パラメータ (α, β, d_p)	F 値	パラメータ (α, λ, d_p)	F 値
原発	7	0.0,0.1,30	0.341	0.9,0.2,10	0.329	0.1,0.9,25	0.406
	9	0.8,0.5,15	0.332	0.6,0.9,10	0.306	0.0,0.6,25	0.395
	12	0.8,0.6,15	0.359	0.6,0.5,10	0.324	0.2,0.9,25	0.369
平均		-	0.344	-	0.320	-	0.390
TPP	9	0.5,0.5,30	0.374	0.1,0.4,35	0.247	0.6,0.0,5	0.385
	10	0.7,0.4,15	0.348	0.1,0.5,5	0.329	0.7,0.0,5	0.427
	12	0.4,0.5,35	0.396	0.5,0.5,10	0.322	0.7,0.0,5	0.418
平均		-	0.373	-	0.299	-	0.410
STAP 細胞	10	0.7,0.3,20	0.470	0.1,0.8,20	0.407	0.6,0.9,20	0.448
	$11_{(1)}$	0.7,0.3,20	0.521	0.3,0.8,15	0.428	0.4,0.9,10	0.417
	$11_{(2)}$	0.0,1.0,10	0.476	0.3,0.6,15	0.400	0.5,0.6,15	0.441
平均		-	0.489	-	0.412	-	0.435
人口問題	8	0.1,1.0,10	0.360	0.3,0.0,10	0.291	0.0,0.1,10	0.419
	10	0.5,0.8,15	0.480	0.0,0.2,5	0.343	0.0,0.5,20	0.411
	11	0.5,0.8,15	0.447	0.5,0.5,10	0.360	0.2,0.3,25	0.413
平均		-	0.429	-	0.331	-	0.414

表中の k は，クラスタ（観点）数を指しており，各条件は次の通りである．

条件 1　名詞・動詞ペア $\langle N, V \rangle$ の名詞 N のみを利用して類似度を計算する．（すなわち，式 (4.7) において $\mathrm{sim}_v = 0$ とする．）

条件 2　条件 1 に加えて，名詞・動詞ペアの（複合）名詞 N を動詞との係り受け関係を考慮せずに形態素解析上で名詞と解析されたすべての語をもとにして抽出する．

条件 3　名詞・動詞ペアの（複合）名詞 N において，修飾語の情報を考慮せず単一名詞のみを利用する．（すなわち，式 (4.8),(4.9) において $\beta = 0$ とする．）

なお，条件 1,2 において，$\mathrm{sim}_v = 0$ のときはパラメータ λ の影響が無視されるため，λ は表に掲載していない．

付録 B　図 6.2 における F 値とパラメータ

表 B.1　図 6.2 における F 値とパラメータ

トピック	k	条件 b		条件 c		条件 d	
		パラメータ $(\alpha, \beta, \lambda, d_p)$	F 値	パラメータ $(\alpha, \beta, \lambda, d_p)$	F 値	パラメータ $(\alpha, \beta, \lambda, d_p)$	F 値
原発	7	0.1,0.0,0.9,25	0.394	0.1,0.0,0.8,25	0.408	0.7,0.6,0.7,25	0.365
	9	0.0,0.1,0.9,15	0.426	0.2,0.3,0.8,20	0.375	0.5,0.5,0.9,20	0.424
	12	0.2,0.1,0.8,25	0.392	0.0,0.2,0.9,15	0.389	0.5,0.9,0.9,20	0.404
平均		-	0.404	-	0.391	-	0.398
TPP	9	0.3,0.5,0.7,20	0.436	0.2,0.0,0.6,10	0.356	0.2,0.8,0.7,10	0.477
	10	0.1,0.6,0.8,25	0.383	0.1,0.0,0.5,25	0.353	0.0,0.8,0.7,25	0.435
	12	0.3,0.9,0.8,25	0.471	0.4,0.4,0.6,15	0.340	0.3,0.9,0.9,35	0.455
平均		-	0.430	-	0.350	-	0.456
STAP 細胞	10	0.8,0.5,0.7,20	0.548	0.2,0.4,0.7,15	0.449	0.4,0.5,0.9,15	0.534
	$11_{(1)}$	0.8,0.6,0.7,20	0.554	0.3,0.2,0.9,15	0.424	0.4,0.5,0.9,15	0.538
	$11_{(2)}$	0.8,0.5,0.7,20	0.541	0.1,0.0,0.8,15	0.402	0.4,0.5,0.8,15	0.542
平均		-	0.548	-	0.425	-	0.538
人口問題	8	0.5,0.7,0.6,10	0.436	0.4,0.0,0.9,25	0.341	0.4,0.4,0.7,10	0.486
	10	0.3,0.4,0.7,25	0.422	0.3,0.0,0.9,20	0.398	0.4,0.4,0.8,20	0.501
	11	0.4,0.2,0.8,20	0.448	0.5,0.1,0.9,15	0.413	0.4,0.4,0.8,15	0.444
平均		-	0.435	-	0.384	-	0.477

表中の k は，クラスタ（観点）数を指しており，各条件は次の通りである．

条件 b　名詞・動詞ペア $\langle N, V \rangle$ の（複合）名詞 N において，修飾語に名詞のみを利用する．

条件 c　名詞・動詞ペア $\langle N, V \rangle$ の（複合）名詞 N において，修飾語に形容詞のみを利用する．

条件 d　名詞・動詞ペア $\langle N, V \rangle$ の（複合）名詞 N において，修飾語に動詞のみを利用する．

なお，条件 a は表 A.1 の条件 3（名詞・動詞ペアの（複合）名詞 N において，修飾語の情報を考慮せず単一名詞のみを利用する）と同じであるため，掲載は割愛した．

付録 C　評価実験に用いた意見のサンプル

C.1　トピック「原発」

- 安全対策を根本から見直し、二重三重の対策をした上である事を条件に賛成です。少なくとも、日本の環境で原子力にかわるエネルギーが安定して供給出来るまでは原子力に頼らざるを得ないでしょう。それならば安全対策をきちんとして、新エネルギーが供給されるまでの間でも原子力での発電を望みます。
- 私は原子力発電所が動いてくれることには反対ではありません。その発電所を作る場所が問題だと思います。危険な状態になったとしても、住んでいる人に影響が出ないような、極端なことを言えば、砂漠の真ん中とかに作ればいいと思います。
- エネルギーを大量に使うということは、人間が、文化が進展していっている証拠なのではないでしょうか。そして、必要なエネルギーを効率よくつくれるに越したことはありません。どんなことにも危険性はつきものです。原子力発電所にはその危険性をはるかに超えるメリットがあるように思います。
- 石油に頼らないことがメリットだと考えられます。石油が高騰したときに、ものの値段も高くなる、電気などの光熱費も高くなるのでは、一気に経済が不安定になるでしょう。そのときに電気くらい石油ではないものに頼りたいと思います。
- 危険性がよく議論されますが、発電するときの危険性は原子力発電に限ったことではないように思います。そのほかの発電が普通にされているのに、原子力発電だけどうしてこれだけ反対されるのでしょうか。危険が起こったときの大きさは大きいかもしれませんが、とても起こりにくい安全な発電方法でもあると思います。
- 原子力発電をせずに日本全国の電力がまかなえるのなら原子力発電はやめるべきですが、現状むずかしいので代わりの案や代替エネルギーが出来るまでは原子力発電は必要だと思います。
- 原子力発電をやめて火力発電を増やすことは、地球温暖化問題の世界的な潮流に全く逆行することになります。そして、まだ再生可能エネルギーでは必要となる電力を十分に賄うことはできません。原子力発電をやめている間は火力発電に頼るほかなくなってしまうのです。今、有力視されている天然ガスにおいても、石油や石炭に比べて二酸化炭素が出るのが少ないとは言いますが、かなり出てしまうのは事実としてあります。今こそ、原子力発電のさらなるイノベーションが必要なのであると考えます。
- 発電コストが安い、二酸化炭素を発生させないという点で賛成ですが、日本のように地震の多い国では福島の原子力発電所の様な事故が起きて、大惨事を起こす可能性があり後々莫大なコストがかかることもあるため、地震のない場所での原子力発電なら賛成します。
- 原子力発電がなければ、日本の電力が賄えないと思います。2011 年の夏に計画停電が行われ、人々の生活が不自由になってしまいました。原子力を反対する人もいますが、今の電力消費生活から離脱することは出来ないのです。もし、原子力に変わるエネルギーが開発されなければ、今の日本は変わることができないでしょう。

- 私は中立の立場です。原子力はインフラの一部になっているため、そう簡単には停止出来ないと思いますし、また、原子力の代替となる発電システムの整備が進まないと、代替も難しいと考えます。
- もろ手を挙げての賛成ではありませんが、もはや致し方ないと言うのが正直な気持ちです。原子力発電の危険性を認識しながらも、事故は起こり得ないと思っていました。日本の技術力を信じていたのです。電気の恩恵にあやかり過ぎた事を、痛切に反省しました。節電というよりは、むしろ今までが無駄使いだったと誰もが知ったはずです。私たち一消費者は、この浪費を正すべく、生活の見直しを早急に計りました。では、政府や電力会社はどうでしょうか。福島での事故後のような、緊迫感は感じられません。企業は、利潤を追求するだけでよいのでしょうか。日本の経済を失墜させない為にも、責任ある製品作りをして頂きたいのです。原子力発電に依存し過ぎない、代替えエネルギーの開発こそ急務です。その時こそ、条件付きで反対と言える時です。ですので今はまだ推進派です。原子力発電事故は、人類に与えられた教訓と思えるのです。
- 反対の理由は、電力を安定的に供給できないからです。幾ら災害が合っても、火力発電所であれば、比較的短い時間で復旧可能ですが、原子力発電所ではそうは行きません。安全面など点検事項が多すぎて、復旧に年単位の時間がかかる発電方法など、このスピードが要求される時代に、事業として成り立たない事は自明です。
- 原子力発電所の稼働の中止をする場合には、太陽光発電パネルを具体的に国家レベルで仕上げてそれが稼働した時に初めて原子力発電の中止に賛成が出来ますが、そうでもない現状の定まらない時期に原子力が危ないから反対だと鼻息を荒くしても意味があまりないと考えています。
- 理由はコストが高いからです。安いと言われてきたのは、無理な想定を重ねて、モデルの計算で算出した結果に過ぎないことが、現在ではばれてしまっています。立命館大学の大島堅一氏など様々な人が実績値で費用を産出していますが、その結果はコストが高い発電ということを示しています。
- 世界へ技術のアピールができる、原子力発電による経済効果があることも必要で理解はできます。しかし、地球全体が生存していくのに「危険性」が高すぎる原子力発電では意味がないと思います。
- 原子力発電で使い終わった燃料って、確か地下に埋められるんじゃなかったでしょうか。聞いた話なのでよくわからないですが、もし埋められるとかであれば、地下水とか将来にわたってのことが心配です。
- 私は反対です。原子力発電は他の発電に比べてとびっきり不安で心配だからです。この不安と心配を解消させてくださる根拠を示されたら、多くの人が納得されるのではないでしょうか。それができないとなると、やっぱり危険なのではないでしょうか。

C.2 トピック「TPP」

- 議論するなら、TPP で日本は損をするけれども、その代わりに何が手に入るのかってとこが問題なんじゃないの？
- TPP でのデメリットとして、よく農業が取り上げられますけど、こちらにあるように、オバマ大統領にとって利益があるというのはあまり聞かなかった話なので勉強になります。郵政民営化から今回の TPP までが、アメリカが日本にある国債を自由にするための計画だったという話は本当なのでしょうか？
- それに対して、日本の各党の党議員個人での意見は明確に発言しているが、党としての回答は示していない事に問題がある。また現在の第一党である民主党も、党首である野田氏と各省長も意見がバラバラのまま、野田氏は参加を表明する事を先日メディアへ表明した。これも問題あり。やはりメリット、デメリット面の両方があるので、もっと国民の意見を聞いて、慎重に判断すべきであるのではと思う。外交も大切だが、その前に確実に言える事は、まず日本国内の政治、福祉、震災対応、雇用などの充実が必要である。今回のどじょう野田内閣は、波風を起こさない、何もしない、首相として有名。ＴＰＰに関しても、何もしない、参加表明をメディアでしただけで、最終的に国民の意見も聞いてないので、現状は没になるがいいのではないだろうか。
- TPP に参加すると、労働者の行き来も自由化されるので、治安の面でも問題は出てくるでしょう。欧州で移民を受け入れた国はどこも問題多発していますし、治安が悪くなることはほぼ確実でしょう。事件が起こり、犯人が逮捕されればまだいいですが、日本は犯罪人引き渡し条約を結べていないので、国外に逃亡されると被害者や遺族は泣き寝入りするしかなくなる可能性があります。ちなみに、警察の予算は年々削減されていますから、期待できません。外国人だけでなく、日本人の貧富の差が開く事で全体的に犯罪が増え、女性やお年寄りのひとり歩きは危険になると考えられます。
- 日本の報道を見ていると、TPP 参加へのデメリットばかりを大々的に報道していると同時に野田総理への批判の声を顕著にしているように感じます。それは国民の声として TPP 参加への反対意見が多いのだと思っていいのでしょうか？
- 今の農家の平均年齢は６４・２歳だから未来のために賛成だと思う。しかもＴＰＰを逃すと経済が２～３パーセントおとろえるとのデータもあるので賛成です。
- まずは、公約を平気で反故にする幼稚な政党に日本の根幹をゆるがすような政策はとってもらいたくないです。アメリカに擦り寄って得をするのは資産家資本家の金持ち連中であって平民貧乏人はますます貧乏になるって寸法ですね。50 年後日本の人口も 1 億割って 8 千万になり 60 歳以上は 4 割だそうですね。日本は 51 年後に 51 番目の州になっているかもですね。
- 近海の海底資源の開発も失う恐れがあると耳にしたのですが。デメリットの方がはっきりしているのに対し、メリットは不透明すぎで参加に対して魅力を感じません。
- 今 TPP 参加への是非についての論文をまとめているところです。TPP 参加による直接的なデメリットもサイト様のおっしゃる通りだと思うのですが、参加表明によるアメリカの日本を手に入れた確信というのが、医療や経済構造への口出しへと繋がるのですよね。医療問題、雇用問題、農業問題様々を考慮して政府には NO と言う勇気を見せてもらいたいと思いました。

- 殴り合いで勝つのはもはや、不可能である。というのはメリット、デメリットの議論からはっきりしたはずです。私は加入したくないという敗北主義などという気はさらさらありません。負けるなら喧嘩をしない、というのは至極当然の民意でしょう。国内法に優先する国際条約、だから議論して対応するんですよ。私は「とりあえず交渉に」などとは言っていません。最終的に入らなければならなくなるって考えてるんですよ。入ったら負けるなんてのは分かってますよ。でも負け方ってものがあるでしょう。
- まあ人民元と直接取引とかやりだした政府だからな。それで日米安保とか馬鹿じゃね？マッチポンプＴＰＰ交渉も普天間も、何もしなければいいものをわざわざこじらせていくのは、日本を真面目に滅亡させたいんだろう。解散しないのも、外交が内閣専決だからあと 1 年好き放題やるためだろ。
- 今のＴＰＰには絶対に手を出してはいけません。皆さんの意見を聞いて安心しました。安倍政権でもＴＰＰ推進派の筆頭を起用したし、ヒヤヒヤしているのでもっとＴＰＰに対する議論を展開し、民意を政治に反映させるべき。
- 財投は郵便貯金の上限額が 2000 万になる預金額でこれが 3000 万になる頃にはあの政調会長の郵政に対する小泉郵政の抜本改革が問われるでしょうね。つまりは郵政の問題点が露呈されることになるでしょうね
- 競争する事自体を否定はできませんが、そもそも競争とは力が大凡拮抗している者同士が争うからこそ成り立つと思うのですよね。決して日本企業が圧倒的に弱いと言うわけではありません。しかし、力量差の大きい者の争い (弱肉強食) を「競争」と呼んでいいのか些か疑問です。TPP 推進派はじめ新自由主義者の方々はその辺りに思慮が及んでいるのかと思います。
- 今回参加しなかったとしてもＴＰＰ自体は残り国内外から手を変え品を変え参加しろという声はあがり続けるでしょうね。それにＴＰＰに入ったとして目に見える影響は数年から 10 年ぐらいと経ってからわかるのでしょう。先の事だと思ってしまったり、ＴＰＰは実は広い範囲で影響があるのでなかなか実感が湧きにくい。だからこそＴＰＰをよく知らなければならないですね。
- ＴＰＰに参加して崩壊するのは、本来、とうの昔に崩壊すべき農業者だけで本当に日本にとって有益な真剣な専業農家にとっては崩壊どころかチャンスでしょう。全国の多くの農業者は兼業で、収入の大半を副業で得ているので失業はあり得ません。反対している面々を見れば、異様に過保護にされた業種ばかり。医師会がＴＰＰに反対していると聞いただけで、ＴＰＰの有益性がわかるというものです。医師不足と言われているのに、自己利益のために医師を増やそうとしない悪人たち外圧でしか崩すことのできない既得権の大きな壁をうちこわす絶好の機会です。
- 経済圏ができるというのは、国境がなくなるということで、ほかの国は皆そういうことを考えている。自分たちに有利なルールを作らないといけないのに、枠組みができあがってから参加しようとしてもだめだ。参加しないと日本が衰退すると思う。

C.3 トピック「STAP 細胞」

- ＳＴＡＰ細胞の是非はこの再別問題で、理研という組織自体に問題があるのは明白でしょう。かりにＳＴＡＰ細胞が非であったとして、それはその個人が勝手にしたことだ、組織は無関係である、組織に泥を塗った責任で首だ。かりにＳＴＡＰ細胞が是であったなら、それは組織の上にあってこその成果だ。
- 目下のところ、小保方氏の味方は三木弁護士です。彼女への弁護は、単に個人の弁護というよりも日本の闊達な研究風土への弁護という意味合いがあると思います。このまま放っておけば、またまた優秀な研究者が日本へ恨みを抱いて流出する傾向を強めてしまうでしょう。三木弁護士の役割はとても大きいと思います。
- 研究者の中には非常識人や発達障害者的人物の割合が一般社界より多少多いかも知れません。でもそのような人達が研究者として生き残れるのは、本業で人並み以上の成果を上げているからです。小保方氏の場合、これまでに公になっている情報から判断して、有能な研究者である証拠は全く存在せず、逆の証拠ならあまた存在する状況です。このような人を切ることは税金で賄われている理研の責務と言える。
- 理研は、判断を放棄して、世間の目から一目散に逃げる選択をしたように思う。禍根が残ることでダメージはより深くなると思うが、目先の事態の沈静化を期待したのだろう。
- 論文として発表している以上、コツを教えられないが認めて欲しいなんて通用しないのです。iPS が認められたのは世界中で再現ができたからです。「コツは教えられない」なんてことを山中先生は言いませんでした。
- 日本が地獄でアメリカが天国であるような意見が多いが強い違和感を持つ。小保方氏側の弁護士からの情報では、小保方氏に対しては複数のオファーがあるとのことである。もしもこれが本当だとすれば、地獄のような日本を捨てて天国のような国に行き、そこで頑張って貰いたいものである。
- STAP 細胞が将来何兆円にもなる利益をもたらす可能性は否定できない。だが、もしもその可能性の確率が百万分の１であったとすれば、利益の期待値は１千万円に満たないことになる。可能性としては如何に巨額であろうとも、その可能性の確率を議論しなければ空論となる。
- STAP 細胞の有無と論文の不備不正は別問題で、理研が再調査不要としたのは後者ですね。今回、理研は新たな事実として、小保方氏が過去にも Nature や science に同じ内容を投稿、不備を指摘されていたことを明かしているとおり、単純なミスという言い分は苦しいです。
- 今回の騒動は、ビジネス・特許視点で議論されるケースは少ないが、この視点が最も重要であると思われる。小保方氏へのオファーは複数あるらしいが、ヴァカンティ氏又はその周辺へ行くのではないかと推測します。もしそうなれば、ヴァカンティ氏の視点で観ると、この騒動は想定内であって逆に高笑いしている可能性も有る。この騒動で誰がメリットを享受出来たか？という疑問を持ちつつ、特許の行方を注視するべきでしょう。
- 理研と小保方氏が争っているのはＳＴＡＰ細胞の有無ではなく、小保方氏の行為に研究不正があったか否かです。小保方氏は、何度も与えられた機会に際して、自己の行為に不正がなかったことを適切に説明すればよかっただけで、バカンティ教授は関係ありません。

- ＳＴＡＰの存在有無は別として、今回の理研の対応は自らのずさんな管理と研究体制が、一人の研究員から露呈してしまって、結論ありきで慌ててトカゲのしっぽとして切った、という印象しかありませんね。組織が疲労しているのか腐敗レベルまで達してるかは分かりませんが、責任ある立場の人他の入れ替えは必要でしょう。しっぽ切って済まされる問題ではありません。
- 単純に小保方氏以外の人があの実験を行っていたならば、こんなことにはなっていなかったと思っています。というか、本当に実験をしたのかも今となっては怪しいですが・・・。ノートも、若山先生の証拠提出もありましたし、そろそろいいんじゃないでしょうか。iPS 細胞があるんですから。
- 小保方氏の論文について，明確に不正疑惑があって，それに対して小保方氏は一部公正な生データが出せなくて疑惑が解消できなくて，不正認定された．それだけの事です．不正はあったから適切に処分する．STAP 細胞があるかどうかは理研が調査する．アカデミア研究者以外の人間が把握すべき事実はそれで十分．
- 彼女は研究者として STAP 細胞をやりたいのなら，こんな法律遊びやってないで，Dr. Vacanti なりのところにさっさと行って，研究を再開してしまえばいいのです．STAP 細胞に関しては，ちゃんと検証されて，論文が出たらまたお話ししましょうという認識で十分です．
- 別に STAP 現象があるとかないとかの話じゃないですよ。論文の内容に不正があることは明白なので再調査は不要と言うことですね。STAP 細胞はあってもなくても論文の不正の判定の内容には無関係なんですが。この間違えたと言われる、電気泳動の写真ですが訂正後の写真と論文では、この論文が STAP 現象が起きた論証になっていないようですよ。もし STAP 現象が本物なら、一度撤回して必要な再実験をし、ストーリーを再構成して論文を正しくしなさいと言うことでしょう。
- 青色発光ダイオードは中村氏が言っていたように、日本では科学者が報われる土壌がないのだと思う。個人的には、小保方氏には日本でＳＴＡＰ細胞を証明してほしいという気持ちもあるが、ＳＴＡＰ細胞があるということを証明してほしいことへの思いの方が強い。日本で難しいのであれば、海外で成功して理研や文科省の鼻を明かしてやってほしいものですね。
- この分野での、「特許認定」と「論文」についての関係がわからないのですが、ある程度高名な認定機関に論文が受理されるのも「特許認定」の条件になっているのでしょうかね？単純に考えれば、論文とは関係なく、提示した特許申請内容に基づき、認定機関（特許庁）にて実現が確認されれば問題無いと思うのですが、どうもどの方の記事でも関係性があるように書かれているので不思議に思っています。

C.4 トピック「人口問題」

- 日本人 1 人１人の経済的な発展に主眼をおいた場合、人口下落の食い止めはあまり関係のない政策ということになる。それならば科学技術予算の拡大や公教育の充実など、単位当たりの生産率向上に寄与する政策を進めることにリソースを割いた方がよっぽど役に立つだろう。
- 未だかつて移民で成功した国はありません。移民を受け入れた国はどこの国でも、自国の低所得者層の仕事を移民が奪い、その二者で深刻な対立が起きるとともに低所得者層の失業率が上昇するという問題を抱えています。もちろん治安は悪化しますので警察の力は当然に足りなくなり、民間警備会社が必要となって、警備産業を中心に GDP は上昇するかもしれません。
- 人口を維持するというのは「生産年齢人口を維持する」という目的があるわけですが、そこに目を向けないまま「人口が減ってもいいじゃないか」と言うのは、たぶんこの問題の本質がわかっていないのでしょう。まあ人口はわかりやすい数値なので、そこに目が向きがちなのはわかるんですけどね……。
- 婚活や出生成功率、また離婚率の数値もどうでしょ。統計で出ている限りは、取り組まなきゃね。何らかの政策的具体策が、婚姻や子供を持つ家庭にあるべき数値を打ち出すんではないかとかね。
- 仮に只今この瞬間から出生率が劇的に向上したとしても、この先 20 年は生産年齢人口は減り続ける。そして、右肩上がりの時代に作られたこの国の社会制度はその重みに耐えられるようには出来ていない。社会保障制度改革は出生率の向上策と併せ、早急に取り組む必要があると思います。
- 人口を維持すべきか、減らしたほうがいいのかという論点も大事ですが、今は単に所得が少なすぎるために子供を生むに生めないという状況になってるだけですよね。これは不幸な話ではないかと。先に所得を安定させる方向で考えてはどうかでしょうか。
- 資源が問題なので人口減にしようというのなら、産業革命前のテクノロジー水準の暮らしに戻すとか、超低資源消費技術を確立するとか、そういう方向をもっと頑張らないといけませんね。
- 若者の婚姻率は収入や余暇時間の長さに比例しますから、若者の給料を上げ、休みを増やすことが一番の少子化対策だ、という、極めて当たり前の結論になります。ゆとりだなんだと言われる最近の若者ですが、可処分所得は低くなり、労働時間は長くなっています。ですから、労働基準法の順守に対する意識がかつてないほど高まるのは仕方のないことだと思います。
- ほとんどの自治体は消滅するでしょうね。一刻も早く都市部へ人口を移動させるべきです。都市部の公共交通網は本当に優れています。公的部門が効率的に動くためには、過疎地は必要ない。そう断言できます。
- 海外在住の方の話を聞くと本当に日本は改善余地がたくさんあるんだなって思えますね。特に公的部門が。田舎は若者に冷たいですよ。仕事無いですし。サービス業がほぼないです。老人人口もいつ急減するか、わかりませんので介護関係の仕事に将来性はありません。
- 何らかの形で日本を大改造しなければならないタイミングが近づいているのですね。それには当然、大きな投資が必要になり、それは需要を生み、経済を活性化させてくれるかもしてません。ここで重要なのは、国土・自治体の改造を前向きに行う事でしょう。どうせ維持できないものを、維持させることにお金をつぎ込んではいけません。そんなことをしたら、砂漠に水を撒くようなものです。

- 資本主義経済社会に内在する現象でしょう。我国でも GDP の 7 割がサービス部門ですからね。サービス産業は人が居て初めて成り立つと思いますがね。だから都市へ人口が集中する。一次産業製品の価格が安すぎるのでしょうね。極論ですが 10 倍ぐらいにすれば良いでしょうが。
- 日本の製造業の生産性はけっして高くないけど、非製造業もふくめ低成長分野への投資を続けるだけじゃなくてより有望な新興企業への投資に振り向けるようにすれば、人口減でも成長を達成できる余地はまだまだ日本にはあると見るのが自然だね。そのためにはやはり雇用のスムーズな移動が可能な社会が求められるわけです。
- 日本は人口の増加と成長を前提とした社会システムで成り立ってる。年金制度や年功序列体系もその一種。つまり人口ピラミッドがピラミッドの形をしているからこそ維持できるシステムなんだと思う。無限に成長し続ける事はできないから、これはもはや一種のねずみ講。社会システムの大改革をやるためにも、その時間を稼ぐためにも移民が必要。
- 少なくともいったん人口減、経済縮小に耐えて国を立て直すくらいの余裕はあるはず。これからも未来への借金を積み上げ続けて無理やり経済大国トップ 3 を維持するか、「実力」に見合った経済力を受け入れ、そこからじっくり復活するか。
- 機械化で補う事ができると思うので、人口減少はともかく、労働力の不足は、気にしなくてよいと思う。就職難の人余りは、相変わらずだし。
- 労働人口は減るかもしれませんが、仕事の数はもっと速く減ります。日本の失業率は低いですが、既に今、仕事数は少ないです。数年前に比べて少なくなっている印象です。介護は人手不足ですが、製造業は本当に減っていると思います。私が従事している職種は、事務専門職ですが、数年前と比べると求人が激減しました。
- 日本が現在の豊かさを保ち、年金など社会保障を維持可能なものとするためには、男女の区別なく全員が生涯労働するという環境が必要であることはほぼ間違いない。その上で、付加価値の高い産業へのシフトを進め、生産性を高めていく努力が必要になると考えられる。
- 人口力が今後も国家の財産であり続けるか、という問いには疑問符が付く。外国からの投資と、労働集約型産業によって経済成長を遂げてきた途上国が、次のステップで直面する問題が中所得国の罠であることは周知のことだ。国内総生産（GDP）がフローで算出されるため、人口大国は経済力が実態よりも強く見える一面がある。だが、人口力をバックに開発が進んだ国の中で、今後、順調に先進国型経済に移行できる国は、数えるほどしかないというのが大方の予測だ。

付録 D　人手により生成された正解クラスタ群のサンプル

D.1　トピック「原発」

観点：「電力行政」

- 私は原子力発電所が動いてくれることには反対ではありません。その発電所を作る場所が問題だと思います。危険な状態になったとしても、住んでいる人に影響が出ないような、極端なことを言えば、砂漠の真ん中とかに作ればいいと思います。
- 発電コストが安い、二酸化炭素を発生させないという点で賛成ですが、日本のように地震の多い国では福島の原子力発電所の様な事故が起きて、大惨事を起こす可能性があり後々莫大なコストがかかることもあるため、地震のない場所での原子力発電なら賛成します。
- 世界へ技術のアピールができる、原子力発電による経済効果があることも必要で理解はできます。しかし、地球全体が生存していくのに「危険性」が高すぎる原子力発電では意味がないと思います。
- 原子力発電で使い終わった燃料って、確か地下に埋められるんじゃなかったでしょうか。聞いた話なのでよくわからないですが、もし埋められるとかであれば、地下水とか将来にわたってのことが心配です。
- 私は反対です。原子力発電は他の発電に比べてとびっきり不安で心配だからです。この不安と心配を解消させてくださる根拠を示されたら、多くの人が納得されるのではないでしょうか。それができないとなると、やっぱり危険なのではないでしょうか。
- ちょっと思いつかないですが、たとえば無人島のような場所に原子力発電所をつくるということは無理なのでしょうか。人の数が少ないからといって、そこに住んでいる人を明らかな危険にさらすことには反対です。
- 原子力発電所というのは、やはり廃棄物の問題が解決できていません。例えて言うと、これが薬の場合で言う副作用があるまま使っているようなものです。薬の場合、明確に副作用がある場合おそらく誰も使わないでしょう。原子力発電はその副作用が解決できていないのに推進することは考えられません。したがって、反対です。
- 地球温暖化の危機が叫ばれ、化石燃料を燃やす事でその原因の二酸化炭素を増やすという事が知られて以来、原子力は二酸化炭素を排出しないクリーンな発電方法として注目されてきました。しかし、発電時には二酸化炭素を排出しなくても、燃料の採掘過程や発電所の製作過程で、膨大な二酸化炭素が放出される事は既に周知の事実です。その上、核分裂生成物というどうしようもなく危険な核のゴミまで生み出すのですから、とてもではありませんが「クリーン」などとは言えず、受け入れられる発電方法ではありません。
- 原子力発電には、条件付きで反対です。理由としましては、事故が発生した際のリスクが高すぎる点です。現在起きている福島第一原子力発電の事故での被害や農作物等の影響が大きいからです。ただ条件付きとした点については、原子力発電に替わる安定した電力の供給方法や現在原

子力発電所で働いている人などの雇用の確保などが必要だからです。

- 事故後の処理がいかに大変か毎日のニュースで知り、放射性物質におびえる毎日です。口にするものも選んで体質改善をしていたのに、さらに心配の種が増えてしまいました。地震の多い国で、狭い国土の日本には原子力発電反対です。
- 放射性廃棄物をどう処理するかがいまだにはっきりしていないため、賛成できません。どんどん増え続ける廃棄物を、日本の地下や海底に何万年も埋め続けるというのは安全上問題があります。廃棄物をつめたドラム缶から放射能が漏れているという話もよく聞くので、極めて危険であるように思います。

観点：「技術向上」

- 原子力発電はどんどん発展させるべきです。これから全世界で必要なものになってくると思います。そのときに、日本が一番効率がよく、安全性を備えた技術を持っておくべきだと思います。
- 原子力発電には反対です。この発電を使っていると、それに伴う研究が進むことは当然のこととなると思います。そしてその研究は、いずれ放棄したはずの兵器をつくるための技術となることも考えられます。

観点：「経済」

- 危険性の問題はあるかもしれませんが、現実的に考えると原子力発電所は必要だと思います。これがなくなると、火力発電に大部分を頼ることになって、石油の大量消費、それに伴う電気料金単価の大幅な値上げが予測されます。
- 石油に頼らないことがメリットだと考えられます。石油が高騰したときに、ものの値段も高くなる、電気などの光熱費も高くなるのでは、一気に経済が不安定になるでしょう。そのときに電気くらい石油ではないものに頼りたいと思います。
- 理由はコストが高いからです。安いと言われてきたのは、無理な想定を重ねて、モデルの計算で算出した結果に過ぎないことが、現在ではばれてしまっています。立命館大学の大島堅一氏など様々な人が実績値で費用を産出していますが、その結果はコストが高い発電ということを示しています。

観点：「当事者任せ」

- 日本の電力事情のために必要なのかもしれませんが、やはり態度の問題もあると思います。必要だからと一方的に言われても、そこに住む方々は納得されないでしょう。取引、駆け引きのような条件だけれはなく、きちんと気持ちをくみ取ることが大切だと思います。
- この賛成、反対を考えることは、何が正義で何が悪かを問うくらいに、本当は難しい問題だと思います。とりあえず、金銭面が絡まない人たちだけで真剣に話し合ってみてもらいたいと思います。

D.2 トピック「TPP」

観点：「治安」

- TPP に参加すると、労働者の行き来も自由化されるので、治安の面でも問題は出てくるでしょう。欧州で移民を受け入れた国はどこも問題多発していますし、治安が悪くなることはほぼ確実でしょう。事件が起こり、犯人が逮捕されればまだいいですが、日本は犯罪人引き渡し条約を結べていないので、国外に逃亡されると被害者や遺族は泣き寝入りするしかなくなる可能性があります。ちなみに、警察の予算は年々削減されていますから、期待できません。外国人だけでなく、日本人の貧富の差が開く事で全体的に犯罪が増え、女性やお年寄りのひとり歩きは危険になると考えられます。

観点：「医療」

- もし TPP に参加したら、日本の医療は平等でなくなるでしょう。アメリカは薬の価格が高いので、アメリカ国民は、民間の保険会社に高額な保険金を支払って医療費を捻出しています。TPP に参加して儲かるのは製薬会社と保険会社だけです。TPP に参加することで日本が誇る国民皆保険は崩壊し、お金がなければ医療が受けられずに死んでしまうような社会になりかねないのです。
- TPP に参加しなくても、国民皆保険維持は、国情から判断して無理があります。存続困難となっているので、混合診療の導入はやむを得ないのです。日本の医療で通用する外国人医師はそんなにいませんし、言葉の壁もあり、無条件で入ることはまず考えられません。ですが、日本の大学医学部卒業、国家試験合格相当の外国人医師が日本で働いてくれるのなら医師不足も解消できますし、労働賃金の安い発展途上国の参入で、介護などの労働力不足の解消にもなります。ある程度の医療の国際化は医療レベルを維持する意味でも必要なのです。国際化することで、医療機器は内外価格差が是正されることが期待されます。

観点：「保護主義」

- ＴＰＰに参加して崩壊するのは、本来、とうの昔に崩壊すべき農業者だけで本当に日本にとって有益な真剣な専業農家にとっては崩壊どころかチャンスでしょう。全国の多くの農業者は兼業で、収入の大半を副業で得ているので失業はあり得ません。反対している面々を見れば、異様に過保護にされた業種ばかり。医師会がＴＰＰに反対していると聞いただけで、ＴＰＰの有益性がわかるというものです。医師不足と言われているのに、自己利益のために医師を増やそうとしない悪人たち外圧でしか崩すことのできない既得権の大きな壁をうちこわす絶好の機会です。
- 既に多くの FTA を締結している日本が、TPP の議論すら否定的なのは理解出来ない。TPP の参加で「他産業の為に農業を犠牲にするのか」というが、国内景気の向上と農家所得の向上が連

動してきたことを思い出そう。保護主義が産業と農産物をダメにしてきた。

- TPP 参加だけで農業が良くなるとは思いませんが、拒絶する理由も逆に見当たらないってとこですかねぇ。もはや日本産が国際価値を為していない現実を農家にも国民にも受け入れてもらった上で、強い農業の再生に国家ぐるみに臨みたいものです。

観点：「政策」

- 殴り合いで勝つのはもはや、不可能である。というのはメリット、デメリットの議論からはっきりしたはずです。私は加入したくないという敗北主義などという気はさらさらありません。負けるなら喧嘩をしない、というのは至極当然の民意でしょう。国内法に優先する国際条約、だから議論して対応するんですよ。私は「とりあえず交渉に」などとは言っていません。最終的に入らなければならなくなるって考えてるんですよ。入ったら負けるなんてのは分かってますよ。でも負け方ってものがあるでしょう。
- 今のＴＰＰには絶対に手を出してはいけません。皆さんの意見を聞いて安心しました。安倍政権でもＴＰＰ推進派の筆頭を起用したし、ヒヤヒヤしているのでもっとＴＰＰに対する議論を展開し、民意を政治に反映させるべき。
- 今回参加しなかったとしてもＴＰＰ自体は残り国内外から手を変え品を変え参加しろという声はあがり続けるでしょうね。それにＴＰＰに入ったとして目に見える影響は数年から 10 年ぐらいと経ってからわかるのでしょう。先の事だと思ってしまったり、ＴＰＰは実は広い範囲で影響があるのでなかなか実感が湧きにくい。だからこそＴＰＰをよく知らなければならないですね。
- 経済圏ができるというのは、国境がなくなるということで、ほかの国は皆そういうことを考えている。自分たちに有利なルールを作らないといけないのに、枠組みができあがってから参加しようとしてもだめだ。参加しないと日本が衰退すると思う。

観点：「日本の衰退」

- 現状、日本は工業が盛んであり、国富のほとんどは工業による輸出産業が生みだしています。日本の GDP にしめる第一次産業（農業）の割合は 1.5 %にすぎません。近年、安いウォンを背景に韓国がどんどんシェアを伸ばしており、日本の工業は円高の影響もあって価格面で大きく差をあけられています。日本の製造業にとってこれは危機でしかありません。
- 経済圏ができるというのは、国境がなくなるということで、ほかの国は皆そういうことを考えている。自分たちに有利なルールを作らないといけないのに、枠組みができあがってから参加しようとしてもだめだ。参加しないと日本が衰退すると思う。

D.3 トピック「STAP 細胞」

観点：「法定での決着」

- 本題は、故意、悪意又は善意であったかは、最終的に法廷で決着した方が良いような気がします。でないと、同業者の醜い誹謗中傷がずっと続くと思いますね。科学界の倫理が、あたかも法律より上位にあるような幻想を抱いている研究者がいるのは欺瞞であり。それを知らしめる意味もあると思います。例えば、犯罪者扱いの発言は、これは法律の問題であり、業界ルールとは異なる。

観点：「研究者としての為人」

- 究者の中には非常識人や発達障害者的人物の割合が一般社界より多少多いかも知れません。でもそのような人達が研究者として生き残れるのは、本業で人並み以上の成果を上げているからです。小保方氏の場合、これまでに公になっている情報から判断して、有能な研究者である証拠は全く存在せず、逆の証拠ならあまた存在する状況です。このような人を切ることは税金で賄われている理研の責務と言える。
- 小保方さんは学位からしてコピペで取得したなんちゃって博士で、そのノリで妄想山盛り STAP 論文を発表した詐欺師か精神病患者です。これだけ研究者として不適格な証拠が揃った人間に税金を注ぎ込んで研究ゴッコをさせつづけるとしたらそれこそ世界の笑いものです。
- 私は小保方氏がボスの笹井氏の強引な結論誘導に即したデータを提供する様な事が有ったのではないかと疑ってます。そして、小保方氏は前から自分を引き揚げてくれる評価者に取り入って、彼らの望むデータを作る事に長けていたのではと思ってます。

観点：「メディア」

- 安倍晋三がわざわざ国会で小保方さんを絶賛して、メディアもノーベル賞とかアホみたいに持ち上げたのに、ネット住人達が論文と実験の両方のおかしな点を突きつけて、理研に 100 億ぶち込む安倍政権の成長戦略をぶち壊したんですよ。それでメディアは大騒ぎして、小保方さんを叩きまくって、自己正当化を図っている。マトモな論文や STAP 細胞の証拠を出せなかった小保方さんも大きな問題だが、ネット住人の尻馬に乗るしか出来ない、カスなメディアが自己批判の目をもっと持たなければならない。
- 組織内のイザコザも、ひょんなはずみにメディアが取り上げられれば、大衆の注視の的になり、それがまたメディアの取材攻勢を生んで、さらに人々の関心を生んで・・・というフィードバックによって、一転メディアにとっておいしいコンテンツになるわけだ。今は特に才覚がなくても、ひとたびメディアに取り上げられれば誰もが脚光を浴びられる時代。これからも次から次へと新しい「小保方さん」が出現するのだろう。
- メディアの影響というより見る側の倫理観だと思う．くだらなく、作為的且つ、メディアスクラムと分かっている．バラエティ情報番組を見るから、調子に乗って歯止めが利かなくなる．良く

も悪くも営利追求なので、NHK 以外は、ほぼ視聴率が結果になる.

- 科学のなんたるかも知らずに、ただ話題戦だけを追ってセンセーショナルな報道だけを狙っているのでは、ほんとのメディアではありません。真実がなにかを追究するのがメディアであるべきです。この場合の真実とはＳＴＡＰ細胞が実際に作成できたのかどうかということです。論文作成のミスとかコピペとか画像の使い回しとかのプロセスの問題ばかりを追究してるだけではしょうがないのです。

観点：「利益」

- STAP 細胞が将来何兆円にもなる利益をもたらす可能性は否定できない。だが、もしもその可能性の確率が百万分の１であったとすれば、利益の期待値は１千万円に満たないことになる。可能性としては如何に巨額であろうとも、その可能性の確率を議論しなければ空論となる。
- 今回の騒動は、ビジネス・特許視点で議論されるケースは少ないが、この視点が最も重要であると思われる。小保方氏へのオファーは複数あるらしいが、ヴァカンティ氏又はその周辺へ行くのではないかと推測します。もしそうなれば、ヴァカンティ氏の視点で観ると、この騒動は想定内であって逆に高笑いしている可能性も有る。この騒動で誰がメリットを享受出来たか？という疑問を持ちつつ、特許の行方を注視するべきでしょう。
- この分野での、「特許認定」と「論文」についての関係がわからないのですが、ある程度高名な認定機関に論文が受理されるのも「特許認定」の条件になっているのでしょうかね？単純に考えれば、論文とは関係なく、提示した特許申請内容に基づき、認定機関（特許庁）にて実現が確認されれば問題無いと思うのですが、どうもどの方の記事でも関係性があるように書かれているので不思議に思っています。

観点：「問題の所在」

- STAP 細胞の有無と論文の不備不正は別問題で、理研が再調査不要としたのは後者ですね。今回、理研は新たな事実として、小保方氏が過去にも Nature や science に同じ内容を投稿、不備を指摘されていたことを明かしているとおり、単純なミスという言い分は苦しいです。
- 理研と小保方氏が争っているのはＳＴＡＰ細胞の有無ではなく、小保方氏の行為に研究不正があったか否かです。小保方氏は、何度も与えられた機会に際して、自己の行為に不正がなかったことを適切に説明すればよかっただけで、バカンティ教授は関係ありません。
- 別に STAP 現象があるとかないとかの話じゃないですよ。論文の内容に不正があることは明白なので再調査は不要と言うことですね。STAP 細胞はあってもなくても論文の不正の判定の内容には無関係なんですが。この間違えたと言われる、電気泳動の写真ですが訂正後の写真と論文では、この論文が STAP 現象が起きた論証になっていないようですよ。もし STAP 現象が本物なら、一度撤回して必要な再実験をし、ストーリーを再構成して論文を正しくしなさいと言うことでしょう。

D.4 トピック「人口問題」

観点：「自然任せ」

- この小さな国に 1 億 2000 万人以上が住んでいるんだから絶対数は多い。人口減少自体を憂う必要はない。適正数になれば、人口減少は自然に止まる。国内人口は増え続けるという前提がおかしいだけだ。

観点：「移民」

- 未だかつて移民で成功した国はありません。移民を受け入れた国はどこの国でも、自国の低所得者層の仕事を移民が奪い、その二者で深刻な対立が起きるとともに低所得者層の失業率が上昇するという問題を抱えています。もちろん治安は悪化しますので警察の力は当然に足りなくなり、民間警備会社が必要となって、警備産業を中心に GDP は上昇するかもしれません。
- 移民が入り込んで、生活保護受給者が増えれば、経済状況がさらに悪化して、もっと移民が必要だって話が出てくることでしょう。移民が必要だって主張は、悪化する可能性を無視していますから、無責任だと思います。

観点：「社会構造」

- 日本という国を超長期に渡り維持するのに、エネルギー問題や食料問題などを考えれば、１億人という人口は些か多すぎるのではないか？と思われる。そう捉えれば前向きに人口減少を捉え、それに応じた豊かな社会を目指す変革が考えられても良いのだが、現実は社会の変化を恐れているわけで上手く行くはずも無い。
- 何らかの形で日本を大改造しなければならないタイミングが近づいているのですね。それには当然、大きな投資が必要になり、それは需要を生み、経済を活性化させてくれるかもしてません。ここで重要なのは、国土・自治体の改造を前向きに行う事でしょう。どうせ維持できないものを、維持させることにお金をつぎ込んではいけません。そんなことをしたら、砂漠に水を撒くようなものです。
- 日本は人口の増加と成長を前提とした社会システムで成り立ってる。年金制度や年功序列体系もその一種。つまり人口ピラミッドがピラミッドの形をしているからこそ維持できるシステムなんだと思う。無限に成長し続ける事はできないから、これはもはや一種のねずみ講。社会システムの大改革をやるためにも、その時間を稼ぐためにも移民が必要。

観点：「労働形態」

- 子育て世代への子供手当等経済的援助は絶対必要ですね。しかし、自民党がやっていることは高齢者へのばらまき。これで、出生率が上がる訳がありません。民主党政権まえの自民党時代から出生率アップの有効的な手だてが講じられていません。問題先送り自民党政権では日本経済の復活はありません。子供手当がバラマキとされたのは、官僚にまったく旨味がないからで、マス

コミがこれを拡散し、バカな国民がのっかっただけ。

- 育児や介護は永遠に続くものではないのだから、その期間だけ短時間労働に切り替えて、子供の成長に合わせて労働時間を増やしたり介護が終わったらフルタイム勤務できるように夫婦ともにできる仕組みの方がいいと思う。
- 人口減少が経済や社会に与える影響は大きく、少子化対策に政府の予算をシフトさせることは意味のあることと考えられる。ただ、単純に少子化対策を行っただけでは、十分ではない可能性が高い。

観点：「広域連携」

- 人口減少・少子高齢化による財政悪化の観点から言えば、まずは簡素で効率的な広域連携をできる限り積み重ねていくと同時に、市町村内の移動完結率が低い問題もあり、狭域高密度に基礎自治体が多数存在する多摩地域の再編（合併）は、中長期的には不可避なのではないかと考えています。

付録 E　提案手法により生成されたクラスタ群のサンプル

E.1　トピック「原発」

クラスタ 1

- 石油に頼らないことがメリットだと考えられます。石油が高騰したときに、ものの値段も高くなる、電気などの光熱費も高くなるのでは、一気に経済が不安定になるでしょう。そのときに電気くらい石油ではないものに頼りたいと思います。
- 原子力発電には反対です。この発電を使っていると、それに伴う研究が進むことは当然のこととなると思います。そしてその研究は、いずれ放棄したはずの兵器をつくるための技術となることも考えられます。

クラスタ 2

- ここまで反対派がいるのに、簡単にやめることがないということは、つくる必要が確かにあるということだと思います。そしたら、賛成反対で論争するのではなく、つくることを前提に、どれだけ安全につくれるか、そばにいる人が不満がないようにできるかに重点を置くべきかもしれません。
- 日本の電力事情のために必要なのかもしれませんが、やはり態度の問題もあると思います。必要だからと一方的に言われても、そこに住む方々は納得されないでしょう。取引、駆け引きのような条件だけれはなく、きちんと気持ちをくみ取ることが大切だと思います。
- あのような事件が起きたことをきっかけに、賛成、反対の議論が色々なところで行われるようになってきたと思います。前提として、議論する全員が原子力発電所のそばに住んでいることを想定してみて欲しいと思います。そうすればどうにかしてでも反対したくなるでしょう。
- この賛成、反対を考えることは、何が正義で何が悪かを問うくらいに、本当は難しい問題だと思います。とりあえず、金銭面が絡まない人たちだけで真剣に話し合ってみてもらいたいと思います。

クラスタ 3

- 原子力発電なしで国家レベルの電力をまかなえるかどうかは、これからの太陽光発電パネルのような代替案がしっかりと着手するまでは政治判断で発電所稼働をストップする事が出来ないのが現状です。原子力稼働をストップする事は技術的に問題はないのですが、火力水力ではとても効率が悪くて電力供給のラニングコストが高すぎて維持できないからです。
- 簡単に反対される方が多いですが、原子力発電所をつくらなかった場合、代わりの電力をどうするという考えはなかなかあがってきません。危ないのは承知ですが電気が足りないという事態のデメリットのほうが大きいのではないでしょうか。
- 原子力が生命維持に危険性があるから反対だという早計な判断は出来ず、原子力発電の稼働に匹

敵する代替案が決まらないうちは原子力発電の運営のストップは出来ません。なぜなら、他の現存する発電手法はコストがべらぼうにかかるために現状の電力量と電力料金の維持が不可能だからです。

- 原子力発電所の稼働の中止をする場合には、太陽光発電パネルを具体的に国家レベルで仕上げてそれが稼働した時に初めて原子力発電の中止に賛成が出来ますが、そうでもない現状の定まらない時期に原子力が危ないから反対だと鼻息を荒くしても意味があまりないと考えています。

クラスタ 4

- 自分自身が原子力発電所の近くに住んでいると考えてみてください。とても安心した生活はできません。事故はおきないかもしれませんが、多少なりとも放射能が出ているのではないかと冷や冷やものです。
- 事故後の処理がいかに大変か毎日のニュースで知り、放射性物質におびえる毎日です。口にするものも選んで体質改善をしていたのに、さらに心配の種が増えてしまいました。地震の多い国で、狭い国土の日本には原子力発電反対です。
- 放射性廃棄物をどう処理するかがいまだにはっきりしていないため、賛成できません。どんどん増え続ける廃棄物を、日本の地下や海底に何万年も埋め続けるというのは安全上問題があります。廃棄物をつめたドラム缶から放射能が漏れているという話もよく聞くので、極めて危険であるように思います。

クラスタ 5

- 実際に、どれくらいの燃料でどれくらいの電気がつくれるのかを知れば、ほとんどの人が賛成にまわるのではないでしょうか。私もその中の一人で、あんなに小さなペレットから大量の電気ができるという事実には驚きました。
- 発電コストが安い、二酸化炭素を発生させないという点で賛成ですが、日本のように地震の多い国では福島の原子力発電所の様な事故が起きて、大惨事を起こす可能性があり後々莫大なコストがかかることもあるため、地震のない場所での原子力発電なら賛成します。
- 地球温暖化の危機が叫ばれ、化石燃料を燃やす事でその原因の二酸化炭素を増やすという事が知られて以来、原子力は二酸化炭素を排出しないクリーンな発電方法として注目されてきました。しかし、発電時には二酸化炭素を排出しなくても、燃料の採掘過程や発電所の製作過程で、膨大な二酸化炭素が放出される事は既に周知の事実です。その上、核分裂生成物というどうしようもなく危険な核のゴミまで生み出すのですから、とてもではありませんが「クリーン」などとは言えず、受け入れられる発電方法ではありません。

E.2 トピック「TPP」

クラスタ 1

- もし TPP に参加したら、日本の医療は平等でなくなるでしょう。アメリカは薬の価格が高いので、アメリカ国民は、民間の保険会社に高額な保険金を支払って医療費を捻出しています。TPP に参加して儲かるのは製薬会社と保険会社だけです。TPP に参加することで日本が誇る国民皆保険は崩壊し、お金がなければ医療が受けられずに死んでしまうような社会になりかねないのです。
- TPP に参加しなくても、国民皆保険維持は、国情から判断して無理があります。存続困難となっているので、混合診療の導入はやむを得ないのです。日本の医療で通用する外国人医師はそんなにいませんし、言葉の壁もあり、無条件で入ることはまず考えられません。ですが、日本の大学医学部卒業、国家試験合格相当の外国人医師が日本で働いてくれるのなら医師不足も解消できますし、労働賃金の安い発展途上国の参入で、介護などの労働力不足の解消にもなります。ある程度の医療の国際化は医療レベルを維持する意味でも必要なのです。国際化することで、医療機器は内外価格差が是正されることが期待されます。

クラスタ 2

- ＴＰＰに参加して崩壊するのは、本来、とうの昔に崩壊すべき農業者だけで本当に日本にとって有益な真剣な専業農家にとっては崩壊どころかチャンスでしょう。全国の多くの農業者は兼業で、収入の大半を副業で得ているので失業はあり得ません。反対している面々を見れば、異様に過保護にされた業種ばかり。医師会がＴＰＰに反対していると聞いただけで、ＴＰＰの有益性がわかるというものです。医師不足と言われているのに、自己利益のために医師を増やそうとしない悪人たち外圧でしか崩すことのできない既得権の大きな壁をうちこわす絶好の機会です。
- 既に多くの FTA を締結している日本が、TPP の議論すら否定的なのは理解出来ない。TPP の参加で「他産業の為に農業を犠牲にするのか」というが、国内景気の向上と農家所得の向上が連動してきたことを思い出そう。保護主義が産業と農産物をダメにしてきた。
- TPP 参加だけで農業が良くなるとは思いませんが、拒絶する理由も逆に見当たらないってとこですかねぇ。もはや日本産が国際価値を為していない現実を農家にも国民にも受け入れてもらった上で、強い農業の再生に国家ぐるみに臨みたいものです。
- 農業の就労人口は年々減っていますし、現在の農民の平均年齢は 65 歳と高齢化が進んでいます。日本の農業は先細りなのです。ついでに言えば日本の農産物は高く、競争力がないのは日本の農業自体の問題で海外の問題ではありません。逆にいえば関税によって守られていたので農業改革が行われず、競争力がないのです。農業改革を拒み、努力を惜しんでいるのに、TPP 参加で農業が滅びると言うのはただの責任の押し付けにすぎないのです。逆に TPP 参加は農業改

革のチャンスになるのです。

クラスタ 3

- 新しい視点や競争があるからこそ、品質・サービスも多様化し、消費者にとって選択肢が広がるのです。国を守ろうとするからこそ結果的に日本が取り残され、格差が生まれるのです。少子高齢化で日本は今後、労働力の確保が難しくなるでしょう。自由貿易の拡大が労働者の自由化に波及し、労働力の確保につながる TPP によって大勢の外国人労働者が日本にやってきます。まさに労働力もグローバル調達が可能になるのです。
- TPP で関税がなくなることで、輸入食料品が安く手に入る。消費者の選択肢は広がり、海外食品や加工品がスーパーに並ぶ場面が増えるでしょう。消費者は食費を安く抑えることができるのです。消費者にとって価格面での魅力は大きいでしょう。
- TPP に参加することで、食卓にのぼる食品の圧倒的なものが輸入品になり、国産品を見つけることは至難なことになります。米国など日本へ輸出している国からみると、米国産牛肉の輸入規制や農薬の残留基準などどうしても障壁となる日本のさまざまな規制がある。参加するにあたって、日本はその規制を緩和し基準を引き下げることをせざるおえないのです。安全基準が国際化していくことで食や品質の低下を招く可能性が高くなり、安心して食事ができる環境は、どんどん失われていくでしょう。
- 輸出商品の関税が撤廃もしくは減額されることで、海外の消費者に安く商品を提供できるようになり、大幅な売上増が見込めます。国際競争に勝ち抜くためには、価格競争力を付ける必要があり、関税撤廃による低価格化で、多くの輸出企業は販売数を増やすことができるでしょう。
- 日本に商品を輸入する際に、各業者は関税を支払っています。その関税費用の支払いがなくなれば、商品を安く売ることができます。消費者にとっても、輸入品を格安で購入できるようになります。

クラスタ 4

- TPP に参加すると、労働者の行き来も自由化されるので、治安の面でも問題は出てくるでしょう。欧州で移民を受け入れた国はどこも問題多発していますし、治安が悪くなることはほぼ確実でしょう。事件が起こり、犯人が逮捕されればまだいいですが、日本は犯罪人引き渡し条約を結べていないので、国外に逃亡されると被害者や遺族は泣き寝入りするしかなくなる可能性があります。ちなみに、警察の予算は年々削減されていますから、期待できません。外国人だけでなく、日本人の貧富の差が開く事で全体的に犯罪が増え、女性やお年寄りのひとり歩きは危険になると考えられます。
- 日本の規制で苦しい外国企業や会社経営層や株主には大きなメリット。安い労働賃金や関税撤廃で加盟発展途上国で殆ど製品をくみ上げ国内に持ち込み完成品に出来る。増収増益となれば財務省もホクホク。一般庶民は、賃金が下がり安全じゃないものを共用され挙句、福利厚生費が馬鹿ほど跳ね上がるだけ
- TPP に加盟すると労働力の移動も自由化され、加盟国からの労働者が日本にどんどんやってき

ます。そして、給料の安い外国人に職を奪われ、日本人の失業はどんどん増えます。職のある人も賃金は下落します。参加することで、340 万人の日本人が職を失うことになります。貧富の差が開き、低所得労働者の平均年収は 200 万円を切り、富裕層の収入は倍増します。そうなると、今以上に、生活保護者であふれて日本の社会保障は崩壊しかねない状態になるでしょう。

クラスタ 5

- 議論するなら、TPP で日本は損をするけれども、その代わりに何が手に入るのかってとこが問題なんじゃないの？
- 中曽根、小泉と日本の長期政権の特徴は、どれだけ米国の犬だったかで決まるようですが。民主党は権力維持のための TPP なのか？
- まあ人民元と直接取引とかやりだした政府だからな。それで日米安保とか馬鹿じゃね？マッチポンプＴＰＰ交渉も普天間も、何もしなければいいものをわざわざこじらせていくのは、日本を真面目に滅亡させたいんだろう。解散しないのも、外交が内閣専決だからあと 1 年好き放題やるためだろ。
- うる覚えで確証はないのですが、ＴＰＰには移民自由化が含まれていたと記憶しています。隣に反日を叫んでいる中国、韓国、北朝鮮、ロシアがいます。そして、ＴＰＰを推し進める民主党はかつて外国人参政権を提案しました。状況次第でＴＰＰにこれらの国が入ってくる可能性があり、億単位の移民を前に力で押し通される危険があるのではないでしょうか？最悪、日本の政治が乗っ取られる可能性も否定できないと思います。大企業がＴＰＰを推し進めたいのは、嘗てアメリカが黒人を使ってやったように、大量の安い労働力を国内へ入れたいからです。代わりに日本人の仕事は激減します。農業と工業の発展どうこうを言うなら、ＦＴＡを順番に結んでいけば良いだけ。これさえ韓国では大問題に発展しているのに、碌な情報も無しにこれ以上に危険なＴＰＰへ参加するのは自殺行為でしかありません。

E.3 トピック「STAP 細胞」

クラスタ 1

- 捏造と言い切れるわけではないが、真実だと信頼を得られるだけのデータを提示していただかないと・・・。なんせ毎年少なくとも６００億、今年は追加で１０００億税金から予算が出る機関ですので。

クラスタ 2

- 本題は、故意、悪意又は善意であったかは、最終的に法廷で決着した方が良いような気がします。でないと、同業者の醜い誹謗中傷がずっと続くと思いますね。科学界の倫理が、あたかも法律より上位にあるような幻想を抱いている研究者がいるのは欺瞞であり。それを知らしめる意味もあると思います。例えば、犯罪者扱いの発言は、これは法律の問題であり、業界ルールとは異なる。
- 小保方氏に倫理観や研究能力が欠如しているって指摘と、理研の業績としてどうなのかって話は、混同し無いほうが良いと思う。まずは、ＳＴＡＰの有無に対する認識を、副所長や山梨大の教授に宣言させるべきだと思う。この問題は、科学倫理だけでなく、利権の損得も絡んでいるから、見えるところだけ叩いて満足するのは危ないと思う。

クラスタ 3

- 安倍晋三がわざわざ国会で小保方さんを絶賛して、メディアもノーベル賞とかアホみたいに持ち上げたのに、ネット住人達が論文と実験の両方のおかしな点を突きつけて、理研に 100 億ぶち込む安倍政権の成長戦略をぶち壊したんですよ。それでメディアは大騒ぎして、小保方さんを叩きまくって、自己正当化を図っている。マトモな論文や STAP 細胞の証拠を出せなかった小保方さんも大きな問題だが、ネット住人の尻馬に乗るしか出来ない、カスなメディアが自己批判の目をもっと持たなければならない。
- 組織内のイザコザも、ひょんなはずみにメディアが取り上げられれば、大衆の注視の的になり、それがまたメディアの取材攻勢を生んで、さらに人々の関心を生んで・・・というフィードバックによって、一転メディアにとっておいしいコンテンツになるわけだ。今は特に才覚がなくても、ひとたびメディアに取り上げられれば誰もが脚光を浴びられる時代。これからも次から次へと新しい「小保方さん」が出現するのだろう。
- メディアの影響というより見る側の倫理観だと思う．くだらなく、作為的且つ、メディアスクラムと分かっている．バラエティ情報番組を見るから、調子に乗って歯止めが利かなくなる．良くも悪くも営利追求なので、NHK 以外は、ほぼ視聴率が結果になる．

クラスタ 4

- 日本が地獄でアメリカが天国であるような意見が多いが強い違和感を持つ。小保方氏側の弁護

士からの情報では、小保方氏に対しては複数のオファーがあるとのことである。もしもこれが本当だとすれば、地獄のような日本を捨てて天国のような国に行き、そこで頑張って貰いたいものである。

- STAP 細胞の有無と論文の不備不正は別問題で、理研が再調査不要としたのは後者ですね。今回、理研は新たな事実として、小保方氏が過去にも Nature や science に同じ内容を投稿、不備を指摘されていたことを明かしているとおり、単純なミスという言い分は苦しいです。
- 理研と小保方氏が争っているのはＳＴＡＰ細胞の有無ではなく、小保方氏の行為に研究不正があったか否かです。小保方氏は、何度も与えられた機会に際して、自己の行為に不正がなかったことを適切に説明すればよかっただけで、バカンティ教授は関係ありません。
- 小保方氏の論文について，明確に不正疑惑があって，それに対して小保方氏は一部公正な生データが出せなくて疑惑が解消できなくて，不正認定された．それだけの事です．不正はあったから適切に処分する．STAP 細胞があるかどうかは理研が調査する．アカデミア研究者以外の人間が把握すべき事実はそれで十分.
- 大学評価・学位授与機構教授の田中さんなので、小保方さん関連では STAP 以前の問題、すなわち直接的には早稲田の、一般的には大学・大学院の学位授与・卒業認定の評価が適切に行われてきたのか、という機構の本業に関わることについて踏み込んだことを書いてほしいなと思う。
- 私は小保方氏がボスの笹井氏の強引な結論誘導に即したデータを提供する様な事が有ったのではないかと疑ってます。そして、小保方氏は前から自分を引き揚げてくれる評価者に取り入って、彼らの望むデータを作る事に長けていたのではと思ってます。
- 今回、小保方女史が論文で顕そうとしたＳＴＡＰ現象は捏造。でも、ＳＴＡＰ現象を研究している人は小保方女史以外にもいるわけで、それはまた別の問題。小保方女史は会見での発言でもそうだが、これを意図的に一緒くたにしようとしている。小保方女史が不正を働いたからと言って、他の研究者の研究が否定されるものではない。

クラスタ 5

- 今回の騒動は、ビジネス・特許視点で議論されるケースは少ないが、この視点が最も重要であると思われる。小保方氏へのオファーは複数あるらしいが、ヴァカンティ氏又はその周辺へ行くのではないかと推測します。もしそうなれば、ヴァカンティ氏の視点で観ると、この騒動は想定内であって逆に高笑いしている可能性も有る。この騒動で誰がメリットを享受出来たか？という疑問を持ちつつ、特許の行方を注視するべきでしょう。
- この分野での、「特許認定」と「論文」についての関係がわからないのですが、ある程度高名な認定機関に論文が受理されるのも「特許認定」の条件になっているのでしょうかね？単純に考えれば、論文とは関係なく、提示した特許申請内容に基づき、認定機関（特許庁）にて実現が確認されれば問題無いと思うのですが、どうもどの方の記事でも関係性があるように書かれているので不思議に思っています。

E.4　トピック「人口問題」

クラスタ 1

- 海外在住の方の話を聞くと本当に日本は改善余地がたくさんあるんだなって思えますね。特に公的部門が。田舎は若者に冷たいですよ。仕事無いですし。サービス業がほぼないです。老人人口もいつ急減するか、わかりませんので介護関係の仕事に将来性はありません。
- 資本主義経済社会に内在する現象でしょう。我国でも GDP の 7 割がサービス部門ですからね。サービス産業は人が居て初めて成り立つと思いますがね。だから都市へ人口が集中する。一次産業製品の価格が安すぎるのでしょうね。極論ですが 10 倍ぐらいにすれば良いでしょうが。
- 少なくともいったん人口減、経済縮小に耐えて国を立て直すくらいの余裕はあるはず。これからも未来への借金を積み上げ続けて無理やり経済大国トップ 3 を維持するか、「実力」に見合った経済力を受け入れ、そこからじっくり復活するか。
- 機械化で補う事ができると思うので、人口減少はともかく、労働力の不足は、気にしなくてよいと思う。就職難の人余りは、相変わらずだし。

クラスタ 2

- 経済は、人口に合わせて縮小して良い。治安の良い社会で、暮らして行ければ、それで良い。国内で生産した分を消費して生活できる、孤立系にも戻れる状態を目指すべきだと思う。他国の富を吸い上げてまで、優雅な暮らしを目指す必要は無い。
- 人口減によって国内の需要が減れば消費、政府支出、投資は減少していくでしょう。しかし世界での需要が高まれば輸出によって GDP の成長率を伸ばすことはできる。日本にはエネルギーはほとんどないので製造業の貿易立国型、韓国のモデルが近いかもしれない。

クラスタ 3

- 日本という国を超長期に渡り維持するのに、エネルギー問題や食料問題などを考えれば、１億人という人口は些か多すぎるのではないか？と思われる。そう捉えれば前向きに人口減少を捉え、それに応じた豊かな社会を目指す変革が考えられても良いのだが、現実は社会の変化を恐れているわけで上手く行くはずも無い。
- 人口が減れば、エネルギーの消費量も減るし、人口密度も減って住みやすい環境になるかもしれない。二酸化炭素の削減を考えたら、人口減少はプラスだと思う。いっそのこと、環境保護のために人口 5000 万人を目標にしたらいいのでは？ 5000 万人で国としてやっていけるようにする方が、建設的な気がする。
- 人口を維持するというのは「生産年齢人口を維持する」という目的があるわけですが、そこに目を向けないまま「人口が減ってもいいじゃないか」と言うのは、たぶんこの問題の本質がわかっていないのでしょう。まあ人口はわかりやすい数値なので、そこに目が向きがちなのはわかるんですけどね……。

- 人口を維持すべきか、減らしたほうがいいのかという論点も大事ですが、今は単に所得が少なすぎるために子供を生むに生めないという状況になってるだけですよね。これは不幸な話ではないかと。先に所得を安定させる方向で考えてはどうかでしょうか。
- 資源が問題なので人口減にしようというのなら、産業革命前のテクノロジー水準の暮らしに戻すとか、超低資源消費技術を確立するとか、そういう方向をもっと頑張らないといけませんね。
- 人口と経済の関係は双方向性がある。出生率を高めるには若者が結婚して子どもを生んでもある程度の生活レベルを維持できる環境や経済水準が必要であるし、この記事が指摘する通り、国の経済レベルを高めるには労働人口増加が望まれる。
- 少子化は日本経済全てにリンクしている。これを、以前から私は主張していた。経済だけでは無く、年金・保険など社会福祉・社会保障に大きく影響する。拙速に移民を受け入れるには反対であり、他国の失敗を学ばねば成らぬ。故に、早期結婚と多産を促す必要がある。それは、政府主導で如何様にもできる。
- 問題は人口減少より、人口のボリュームゾーンである団塊世代って人達（60 台後半）が、ほぼリタイアし年金生活に入り購買力が減退した事でしょう。今年度の GDP 予想は１%位の UP と極めて低調なのは、全体の個人消費が低迷している事が原因であって、アベノミクスを続けようが転換しようが、全く関係ないです。

クラスタ 4

- やらなければいけないことはすでに明確で、「出生率を上げる」。どのようなやり方であってもこれが主軸であることは間違いないと思います。人口ピラミッドのくぼみの問題も大変ですが、社会福祉の見直しで対応の余地はあります。今後も新たな問題が出てくるでしょうが、「だからやらない」というわけにもいかず、それでも何とかしていかなければならない段階にきています。子どもたちの未来が豊かで幸多いものであることを願ってやみません。
- 日本人 1 人１人の経済的な発展に主眼をおいた場合、人口下落の食い止めはあまり関係のない政策ということになる。それならば科学技術予算の拡大や公教育の充実など、単位当たりの生産率向上に寄与する政策を進めることにリソースを割いた方がよっぽど役に立つだろう。
- 婚活や出生成功率、また離婚率の数値もどうでしょ。統計で出ている限りは、取り組まなきゃね。何らかの政策的具体策が、婚姻や子供を持つ家庭にあるべき数値を打ち出すんではないかとかね。
- 仮に只今この瞬間から出生率が劇的に向上したとしても、この先 20 年は生産年齢人口は減り続ける。そして、右肩上がりの時代に作られたこの国の社会制度はその重みに耐えられるようには出来ていない。社会保障制度改革は出生率の向上策と併せ、早急に取り組む必要があると思います。
- 子育て世代への子供手当等経済的援助は絶対必要ですね。しかし、自民党がやっていることは高齢者へのばらまき。これで、出生率が上がる訳がありません。民主党政権まえの自民党時代から出生率アップの有効的な手だてが講じられていません。問題先送り自民党政権では日本経済の

復活はありません。子供手当がバラマキとされたのは、官僚にまったく旨味がないからで、マスコミがこれを拡散し、バカな国民がのっかっただけ。

クラスタ 5

- 未だかつて移民で成功した国はありません。移民を受け入れた国はどこの国でも、自国の低所得者層の仕事を移民が奪い、その二者で深刻な対立が起きるとともに低所得者層の失業率が上昇するという問題を抱えています。もちろん治安は悪化しますので警察の力は当然に足りなくなり、民間警備会社が必要となって、警備産業を中心に GDP は上昇するかもしれません。
- 人口減少と共に産業の空洞化はより深刻になるものと予想されます。求められる社会モデルは、勝ち組へ餌に群がる鶏のように人が群れるモデルではなく、各産業分野が破綻しないよう人材を配置することにあり、これまでの日本モデルで非採算分野を切り捨てるやり方では、いずれ基礎インフラや食料を他国企業の良い様にされ、それは日本国と民族の滅亡を意味します。チベットやウィグルの現状は、明日の我が身に起こることですよ。
- 農業や漁業などの一次産業もその仕事自体がなくなることはないが、人口動態において流通は大きな影響を受けるし、今の一次産業において流通のことを抜きにして、その未来を考えることはできない。ある一定の人口規模がないと、サービス産業も成り立たないし、地域の規模自体がなくなると、夕張のように行政サービスも成り立たない。