# 差異に注目した複数文書融合手法

渡邊 拓也† 大野 成義†† 太田 学††† 片山 薫††† 石川 博†††

†，††，†††東京都立大学大学院工学研究科 〒192-0397 東京都八王子市南大沢 1-1
††職業能力開発総合大学校情報工学科 〒229-1196 神奈川県相模原市橋本台 4-1-1
E-mail:†jam@love.eei.metro-u.ac.jp, ††ohno@cs.uitec.ac.jp, †††{ohta, katayama, ishikawa}@eei.metro-u.ac.jp

**あらまし** インターネット上等には膨大な数の文書があり，同じ話題について書かれたものも多くある．同じ話題について書かれた多くの文書の中には興味のない情報が含まれていたり，重複があったりする．そのため，その中から求める情報を得るために多くの労力を費やしてしまう．そこで我々は，ある一つの文書に他の文書の差異情報を付加するシステムを開発した．文書の構文解析には茶筌，南瓜を用い，そこから意味的まとまりを抽出し，他の文書の同じ意味を表すと思われる部分に付加した．このシステムにより，他の文書で情報を補完しながら一つの文書を読む事や，一つの文書を読みながらより自分の興味に適合した文書を選択する事ができるようになる．

**キーワード** ユーザインタフェース，情報統合，知識処理

# A Distinction Emphasis Multi-document Fusion Technique

Takuya WATANABE† Shigeyoshi OHNO†† Manabu OHTA††† Kaoru KATAYAMA††† and Hiroshi ISHIKAWA†††

†，††，†††Graduate School of Engineering, Tokyo Metropolitan University, 1-1 Minami-Osawa, Hachioji-shi Tokyo, 192-0397 Japan
††Department of Information and Computer Science ,Polytechnic University, 4-1-1 Hashimotodai, Sagamihara, Kanagawa, 229-1196 Japan
E-mail:†jam@love.eei.metro-u.ac.jp, ††ohno@cs.uitec.ac.jp, †††{ohta, katayama, ishikawa}@eei.metro-u.ac.jp

**Abstract** There are huge number of documents on the Internet, not a few of which are written about the same topic. Uninteresting information or duplications are included in many documents written about the same subject. Therefore, it requires a great deal of labor to extract useful information from them. Then, we developed a system which adds to one document the difference information on the other documents. We use "Chasen" and "CaboCha" for syntactic analysis, extract semantic groups from their output, and add this group to parts of other documents which may mean the same. Using this system, users can read a document while complementing it with the information included in other documents, or choose more suitable documents while reading it.

**Keyword** User Interface, Information Integration, Knowledge Processing

## 1. はじめに

現在インターネット等を通じ，多くの文書データを容易に取得できる．しかし同時に情報が氾濫してしまっており，それらを効率的に理解するための手法は有用だと考えられる．

何かを調べる時に，関連文書を全て読むわけにはいかない．そこで我々は差異に注目した複数文書融合システムを開発した．本システムが各文書の差異を表現する事により，他の文書で情報を補完しながら特定の文書を読む事や，文書を効率的に取捨選択する事が可能になる．

同じ問題に対し，複数文書要約手法[3,4]が研究されている．しかし，「可読性が落ちる」，「知りたい情報とそうでもない情報の情報密度が変わらない」，「共通部分に注目し，一般的な情報しか表現されない．結果，差異は扱われない」という問題があった．

本システムに同じトピックについて書かれた複数の文書を入力すると，それぞれの文書に他の文書との比較結果を付加する．それにより，各文書の差異が表現される．

現在は比較的文書が整形されている新聞記事の融合を中心に実験を行っている．同じトピックについて書かれた複数の文書を取得する処理は本研究の対象外としている．しかし Google News[2]のような検索エンジンにより，同じトピックについて書かれた複数の新聞記事を自動的に取得する事が可能である．

本手法は文書から複数の意味的まとまりを抽出する．そしてユーザの要求に応じ，他の文書に含まれる同じ内容の意味的まとまりを比較しやすい形で提示する．

文献[1]では複数文書の同じ内容の部分をまとめ，一覧表示していた．しかし，元の文書よりも可読性が低くなってしまう可能性をはらんでいた．本手法では，ユーザは元の文書を読みながら必要に応じて付加情報を閲覧するので，可読性は維持されている．

以降，第２章では関連研究，第３章では提案手法，第４章では実験について述べ,第５章でまとめを行う．

## 2. 関連研究

### 2.1. 複数文書要約手法

本システムは特定の文書に他の文書との比較結果を付加するものだが，複数文書要約手法は複数の文書を一つの文書にまとめる手法である．

代表的なものに newsblaster[3]があり，自然言語処理により意味解析を行っている．文書群を（ｉ）よく似た文書，（ii）特定の人物に対する伝記，（iii）その他の３つに分類し，それぞれ異なる手法で要約を生成する．我々の提案手法が対象としている「（ｉ）よく似た文書」に絞り，処理の流れを以下に示す．

（１）文のトピックを特定

（２）同一トピックの文をグルーピング

（３）文から「単語を頂点，助詞や時制を頂点の属性とする有向グラフ」を生成

（４）有向グラフを既存の自然言語自動生成システムに入力し，自然言語を生成

英語の自然言語処理により意味解析を行うので日本語には適用できない．文書全体の要旨だけを知りたい時には有効であるが，詳しく知りたい部分があった時にはいくつかの原文をそのまま読むしかない．要約文の各部分の出所も代表的なものを一つあげているにすぎない．

本手法は詳しく知りたい部分だけいろいろな文書から情報を集める事が可能である．それらの出所も全て提示するので，自分にあった文書を取捨選択する事が可能である．

また，文書間の差異に注目した MMR[4]という要約手法がある．しかし，文書の部分毎の差異に注目し，詳しく知りたい部分だけ「他の文書の情報をつまみ食い」する事はできない．

### 2.2. 複数文書クラスタリング手法

Google News や RSS Reader により，大量の文書の取得が可能である．両者ともキーワードによる絞り込み・文書生成時間による並び替えが可能である．特に Google News は文書のトピックによるクラスタリングも可能である．

同じトピックについて書かれた複数の文書を取得する処理は本研究の対象外としている．しかしこれらを用いる事により，同じトピックについて書かれた複数の文書を自動的に取得する事が可能となる．

## 3. 提案手法

### 3.1. 概要

本手法は文書を形態素まで分解し，意味的なまとまり（意味情報と呼ぶ）を抽出し，ユーザの要求に応じて同じ内容の意味情報を比較しやすい形で提示する．

提案手法の流れと内部データのデータ構造を以下に示す．

**STEP1:分解**

文書を文に，文を文節に，文節を形態素に分解する．

**STEP2:再結合**

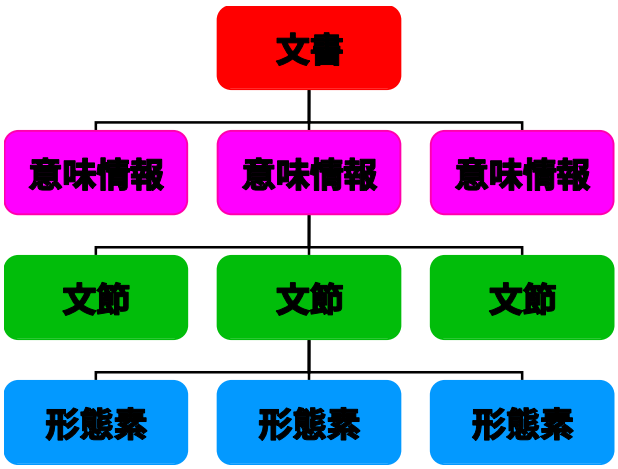
意味情報を抽出する．各文節の意味的役割を求める．

**STEP3:比較**

各意味情報毎に，同じ内容を表している意味情報を求める．

**STEP4:提示**

ユーザが情報を比較しやすいように文節を整列し直して提示する．




文節の例：「魔法戦隊は」

形態素の例：「魔法」「戦隊」「は」

**図１．内部データ構造**

### 3.2. 分解

#### 3.2.1. 解析ソフトによる分解

茶筌[5]を用いて文を形態素に分解し，南瓜[6]を用いて形態素を文節毎にまとめる．茶筌により形態素の品詞情報を得る事ができ，南瓜により文節毎の主たる形態素や，係り受け関係を得る事ができる．

### 3.2.2. 文節をまとめなおす

ユーザが情報を比較しやすくなるように，提示の段階で文節の並び順を変更する．文節と文節の間に他の文節を挿入すると間違った意味を表現してしまう所は，他の文節を挿入しないようにする必要がある．そのため，ここでそのような文節の並びを一つの文節としてまとめてしまう．以下に例を示す．

（1）「主役の」＋「細川茂樹が」
→「主役の細川茂樹が」
（2）「仮面ライダー響鬼」＋「(ひびき)」
→「仮面ライダー響鬼（ひびき)」

## 3.3. 再結合

### 3.3.1. 意味情報を抽出

意味的にまとまった文節群を意味情報としてまとめる．１文に含まれる全ての文節を操作するまで以下の２つの操作を繰り返し，意味情報を抽出する．

（1）文節を走査して主たる形態素が動詞である文節又は，be 動詞の補語となる文節を探す
（2）主たる形態素が動詞である文節に直接的・間接的に係っている文節をまとめる

**例：**図２の文節群から意味情報を生成する．ここで，矢印は係り受け関係を表し，主たる形態素は下線で示した．

まず「知事は」,「長野市民が」・・・と，文節を始めから走査して主たる形態素が動詞である文節又は，be 動詞の補語となる文節を探す．すると「求めた」の主たる形態素が動詞だとわかる．

次に「求めた」に係っている文節をまとめる．「長野市民が」,「泰阜村選管に」,「登録取り消しを」が「求めた」に係っている．さらに「長野市民が」,「泰阜村選管に」,「登録取り消しを」に係っている文節はないので，これら４文節をまとめて一つの意味情報とする．

意味情報が一つ生成されたので，「求めた」の次の文節から再走査する．「第三者訴訟参加人。」が be 動詞の補語になっていると考えられる．

「第三者訴訟参加人。」に係っている文節をまとめる．「知事は、」,「同訴訟の」が「第三者訴訟参加人。」に係っている．さらに「求めた」が「同訴訟の」に係っているので，「求めた」は「第三者訴訟参加人。」に間接的に係っていると考えられる．よって「求めた」もこの意味情報内にまとめる．

ただし，「求めた」より意味情報を生成している．よって，「求めた」より生成した意味情報が「同訴訟の」に係っていると考える．

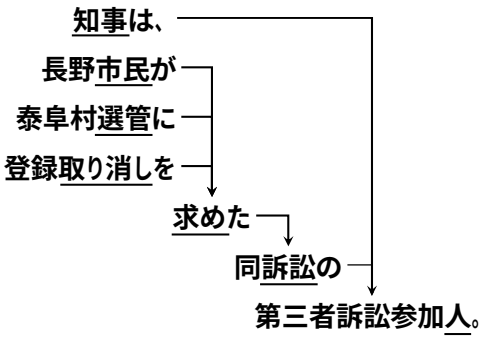



図２．文節群の例

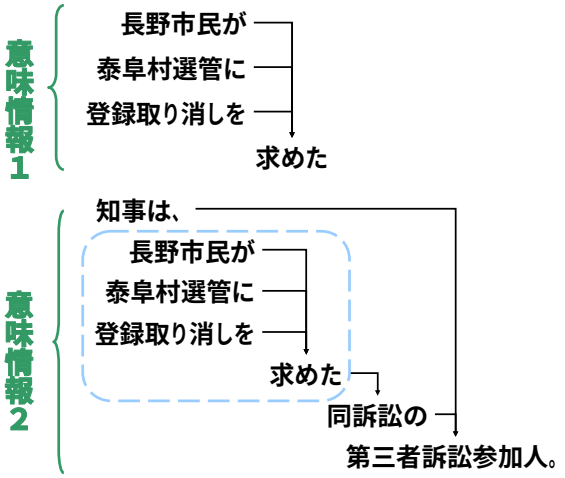



図３．生成された意味情報

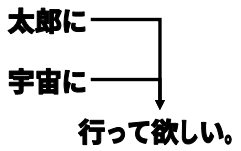



図４．意味情報の例

結果，図３の２つの意味情報が生成される．ここで，他の意味情報内に含まれる意味情報を点線で囲って示した．

### 3.3.2. 文節の役割を求める

意味情報内の各文節の意味的役割を求める．これは，各意味情報の内容の一致具合を計るためや，提示の段階で同じ役割を近くに配置して意味情報を比較しやすくするためである．

提案手法で定義している文節の役割を以下に示す．各役割は格文法に若干の役割を付加したものである．

（1）主格　　（2）対象格
（3）道具格　　（4）場所格
（5）源泉格（始点）　　（6）目標格（終点）

（７）時間格　　（８）原因
（９）述語節　　（１０）接続節
（１１）（英語で言う所の）be 動詞の補語
（１２）その他（形容詞節（句）・副詞節（句））

主たる形態素の品詞と文節が含む助詞の組み合わせにより，文節毎に各役割のスコアを求める．スコアは０から１の実数である．候補となる役割が一つしかない文節は，その役割のスコアが１になる．

また，意味情報内に含まれている他の意味情報は，形容詞句・副詞句（その他）と考える．

候補となる役割が二つ以上ある文節は，検索エンジンを用いて各役割に対応するスコアを求める．

**例：**図４の意味情報の各文節の役割を求める．

「行って欲しい。」は主たる形態素が動詞なので述語節だとわかる．

「太郎に」と「宇宙に」は共に主たる形態素が名詞であり，助詞は「に」である．よって，主たる形態素の品詞と助詞からは一意に役割を決められない．

主たる形態素が名詞で助詞が「に」の時，考えられる役割は対象格・場所格・源泉格・目標格・時間格・原因の６種類である．

「太郎に」と「宇宙に」のどちらも時間を表す名詞を含まないので時間格ではない．ここで時間を表す名詞とは，南瓜又は我々が時間を表す名詞に指定したものである．南瓜が指定した名詞に「日」などがあり，我々が指定した名詞に「明日」などがある．

検索エンジンを用いて残りの５種類の役割それぞれのスコアを求める.この操作では例えば,「～のために」で使われる事が多い（検索ヒット数が高い）名詞は原因のスコアが高くなる．

名詞の後ろにつくフレーズと役割の対応関係を図５に示す．

まず，

「主たる形態素＋フレーズ」の検索ヒット数
÷フレーズ単独の検索ヒット数

を求める.「太郎」の意味を求める時には，

「太郎に対して」の検索ヒット数
÷「に対して」の検索ヒット数

を求める．この値をαと定義する．

「太郎」の意味を求める時，各フレーズに対して求めたαを表１に示す．

「場所各又は目標格」を表すαは３種類あるので，このうち値が最も高いものを「場所格又は目標格」の値とする．ここで，各役割のαを表２に示す．

「に対して」　→対象格
「の方に」
「に向かって」　→場所格又は目標格
「へと」
「から」　→源泉格
「のために」　→原因

**図５．フレーズと役割の対応関係**

**表１．各フレーズのα**

| フレーズ | α（単位：$10^{-5}$） |
|---|---|
| に対して | 29.29 |
| の方に | 1.52 |
| に向かって | 18.99 |
| へと | 4.12 |
| の方から | 1.49 |
| のために | 7.38 |

**表２．各役割のα**

| 役割 | α（単位：$10^{-5}$） |
|---|---|
| 対象格 | 29.29 |
| 場所格又は目標格 | 18.99 |
| 源泉格 | 1.49 |
| 原因 | 7.38 |

**表３．各役割のスコア**

| 役割 | スコア | |
|---|---|---|
| | 「太郎に」 | 「宇宙に」 |
| 対象格 | 0.51 | 0.23 |
| 場所格又は目標格 | 0.33 | 0.69 |
| 源泉格 | 0.03 | 0.01 |
| 原因 | 0.13 | 0.07 |

各役割を代表するαの和が１になるように調整したものを，各役割のスコアとする.「宇宙」に対しても同様にしてスコアが計算できる.「太郎に」と「宇宙に」の各スコアを表３に示す.

表３のスコアから各文節の役割を求める．役割は（１）「その他」を除く同じ役割が複数の文節に属さないようにする（２）スコアの総和が最も高くなるようにする，という規則に基づき求める.

上記の規則により,「太郎に」は対象格,「宇宙に」は場所格又は目標格となる.

「宇宙に」は場所格なのか目標格なのかわからない.この時は「宇宙に」が直接的又は間接的に係っている述語節により役割を特定する.

「～に」という文節が場所格になるのは「ある」「いる」「住む」などの一部の述語節のみである．よって,「～に」が係っている述語節の主たる形態素がこれらの場合は「場所格」に，その他の場合は「目標格」となる.「宇宙に」は「行って欲しい」に係っているので,

目標格となる．

EDR 概念辞書を始めとする各種辞書を用いた方が正確な判定が行えると思われるが，未知後に対応するために現在の手法を取っている．

### 3.4. 比較

各文書から抽出された意味情報を比較し，同じ内容の意味情報を求める．比較は意味情報の全ての組み合わせ毎に行う．

接続節や形容詞節（句），副詞節（句）は付加的要素が強く，意味の中心にはならないと考える．そこで，残りの役割の一致率により意味情報を比較する．

現在は一致率が０．６以上の組み合わせを同じ内容としている．

**例：**図６の２つの意味情報を比較する．図６は各行が意味情報を，スペースが文節の区切りを表す．各文節の役割は括弧で示してある．

同じ役割の文節同士を比較すると，主格と述語は等しく，目標格は等しくないと判定される．

意味情報Ａは３文節中２文節が一致し，意味情報Ｂは４文節中２文節が一致する．そこでそれぞれの一致率を２／３，２／４と考える．２つの意味情報の内どちらかの一致率が０．６を越えた場合，２つの意味情報は同じ内容だと判定する．よって，図６のペアは同じ内容だと判定される．

ここで，文節の一致判定について説明する．文節の一致判定では形態素の一致率を用いて一致判定を行う．意味情報の一致判定と違う事は，形態素には役割がない事である．また，比較する文節がお互いに他方が持たない固有名詞や数字を含む時は，例外的に不一致とした．

### 3.5. 提示

#### 3.5.1. 概要

本システムの出力を図９に示す．出力は JavaScript を用いた HTML である．この例は４つの新聞記事を入力したものである．

画面上部にはある一つの新聞記事が表示されている．青文字で示しているリンクをクリックすると，その文節を含む意味情報と，その意味情報と同じ意味を表す他の記事の意味情報が画面下部に併記される．

併記された他の記事の意味情報をクリックすると，画面上部に表示されている記事がクリックした意味情報が属する記事に変わる．

ユーザは併記された意味情報を見て情報を補完する事や，自分にあった表現をしている文書を選択する事ができる．

また，併記された意味情報にマウスカーソルを合わせると，その意味情報の前後の文節が表示される．画面上部に表示されている新聞記事以外は全文が表示されないが，この機能により様々な情報の「つまみ食い」，文書の取捨選択ができる．

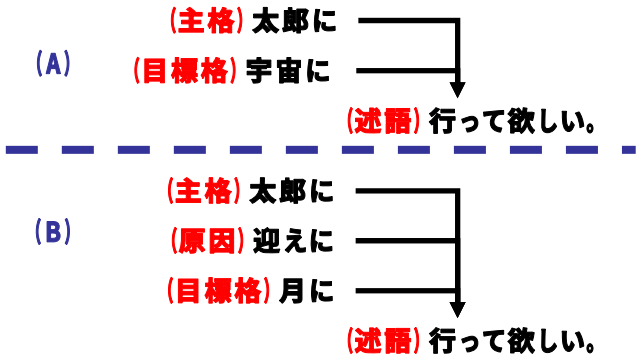



図６．比較判定に用いる例

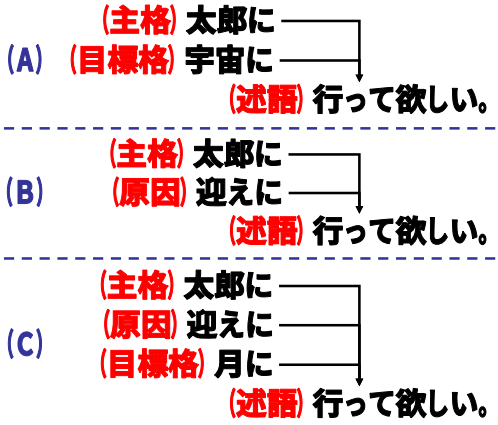



図７．比較される意味情報

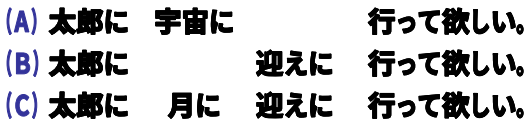



図８．併記された意味情報

リンクの存在する（文字が青くなっている）文節は，その文節を含む意味情報と同じ意味を表す意味情報が存在する文節である．

#### 3.5.2. 意味情報の比較結果の提示

具体例を用いて意味情報の比較方法を説明する．

**例：**図７の３つの意味情報を併記する．各文節の役割は括弧で示してあり，意味情報Ａは画面上部に表示されている記事から生成されたものとする．

意味情報を併記した結果を図８に示す．

意味情報の併記は，画面上部に表示されている記事から生成された意味情報（この場合意味情報Ａ）を中心に行われる．

図８を表としてとらえると，各行が各意味情報を示しており，各列が各役割を示している．

まず，中心となる意味情報Ａをそのままの語順で配置する．

次に，意味情報Ａに含まれる役割を他の意味情報が持っていれば，対応する列に配置する．この場合意味情報Ｂ・Ｃの「太郎に」，「行って欲しい。」と意味情報Ｃの「月に」が配置される事になる．

最後に，意味情報Ｂ・Ｃの「迎えに」が残っているので配置する．「迎えに」は述語の前に配置されているという情報を用い，「行って欲しい。」という述語の直前に配置する．


図９．本システムの出力

(a)

| | | | | | |
|---|---|---|---|---|---|
| 長野県の田中康夫知事が | 長野市から | 泰阜（やすおか）村に | | 住民票を | 移し、 |
| 「知事は | | | 地域住民の目線で自治を考えるため | 住民票を | 移した。 |

(b)

| | | |
|---|---|---|
| 田中知事は | 「泰阜村に迷惑をかけるために村民になったのではないが、信念に基づいて与えられた権利を行使する」と | 話した。 |
| 田中知事は | 「居住の自由を選択したいという私自身の考えに基づいて上告を決めた。<br>単身赴任など多くの人が複数の住所を持つ時代だ。最高裁の判断を仰ぎたい」と | 話した。 |

(c)

| | | | | |
|---|---|---|---|---|
| | 「第三者訴訟参加人」として | 訴訟に | | 参加していた |
| 田中知事 | 「訴訟結果により権利を害される第三者」として、 | この訴訟に | 行政事件訴訟法に基づき、 | 参加し |

(d)

| | | | |
|---|---|---|---|
| 学力、スポーツともに | トップの有名私立高校に | “代打教師”として | 就任した |
| | | 私立東新学院に代打教師として | 就任した |
| 学力、スポーツともに | トップの有名私立高校に | “代打教師”として | 就任した |

図１０．実験結果

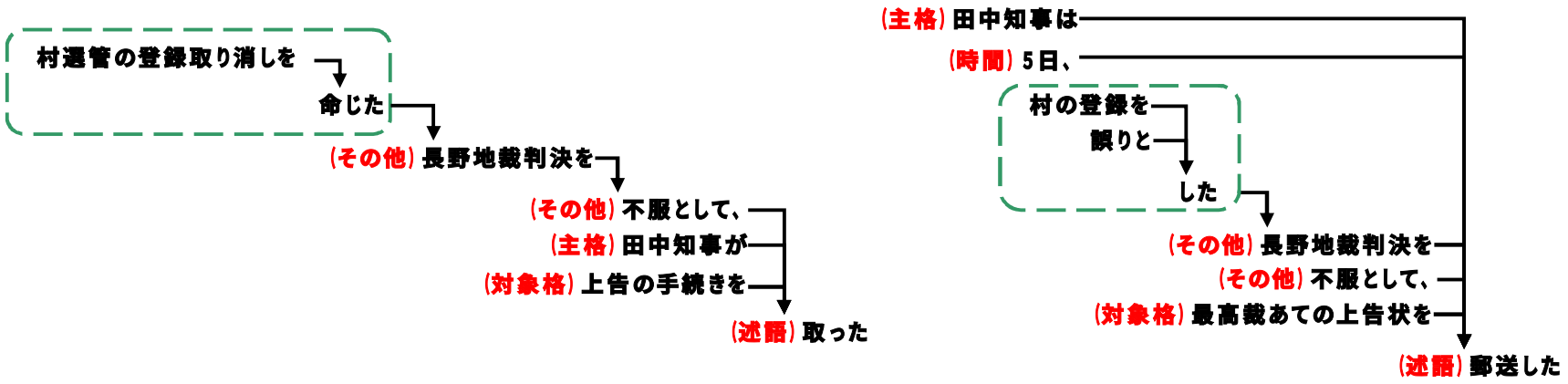



図１１．併記されなかった意味情報の例

## 4. 実験新聞記事の融合

### 4.1. 新聞記事の融合

本システムに，長野県の田中知事の住民票が二重登録されている問題について報じた４つの記事を入力した．併記された意味情報を図１０(a)～ (c)に示す．図１０では併記された意味情報を表で表した．表の各行が各意味情報に，各列が各役割に対応している．

図１０(a)の比較結果を見ると，「知事が泰阜村に住民票を移した」という情報に「地域住民の目線で自治を考えるため」という情報が付加されている事がわかる．

図１０(b)の比較結果を見ると，知事の話した２つの台詞がまとめられている事がわかる．

図１０(c)の比較結果を見ると，「第三者訴訟参加人」という言葉が「訴訟結果により権利を害される第三者」という意味を持っている事がわかる．また，「行政事件訴訟法に基づき」訴訟に参加しているという情報も付加されている．

さらに，図１０(c)の比較結果から，下（左）の意味情報を含む文書はよりわかりやすく書かれていると推測でき，上（右）の意味情報を含む文書はより簡潔に書かれていると推測できる．そうしてユーザは，一つの文書を読みながら自分の興味に適合した文書を選択する事ができる．

この文書群からは図１１に示した２つの意味情報が生成された．各文節の役割は括弧で示してある．２つの意味情報はほぼ同じ内容だと考えられるが，主格しか一致しないので併記されなかった．

### 4.2. 映画のあらすじの融合

本システムに，「代打教師　秋葉、真剣です！」のあらすじ５つを入力した．併記された意味情報を図１０(d)に示す．

図１０(d)の比較結果を見ると，「トップの有名私立高校に」という文節と「私立東新学院に」という文節が異なる列に配置されている．これは文節の役割を求めた際，「トップの有名私立高校に」という文節を目標格に，「私立東新学院に」という文節を対象格に判定してしまったためである．

## 5. まとめ

本稿では，氾濫している情報を効率的に理解するため，ある一つの文書に他の類似文書の差異情報を付加するシステムを開発した．

従来研究されてきた複数文書要約手法は，可読性低下，ユーザのニーズに無関係な情報密度，共通部分のみに注目して差異情報を不使用という問題があった．

本システムはこれらの問題を解決し，他の文書で情報を補完しながら一つの文書を読む事や，一つの文書を読みながらより自分の興味に適合した文書を選択する事を可能にする．

今後の課題を以下に示す．

・役割判定の洗練

現在は役割を判定する文節の情報だけを使って役割を判定している．今後，文節の係る動詞の情報も用いて役割を判定したい．

・意味情報一致判定の洗練

現在の手法は違う内容の意味情報ペアを同じ内容だと判定する事は少ないが，同じ内容を表す意味情報ペアを違う内容だと判定してしまう例が多い．より正確に意味情報の一致判定を行いたい．

・意味情報の近さによるフィルタリング

現在はかなり近い内容の意味情報も，少し遠い内容の意味情報も区別なく表示してしまっている．ユーザの要求に応じて併記する意味情報の「近さ」をフィルタリングしたい．

・長い意味情報の表示方法

長い意味情報を併記するとわかりやすさに欠けてしまう．そこで長い意味情報は簡潔にまとめて表示し，ユーザの要求に応じて長く表示するようにしたい．

・類義語の判定

現在の形態素一致判定は「同じ」と「違う」という区別しかない．類義語判定によりグレーゾーンも扱いたい．

## 謝　辞

本研究の一部は，文部科学省科学研究費補助金特定領域研究(2)(課題番号:16016273)，および(独)情報通信研究機構からの受託研究費(研究題目:文書融合技術と複数文書の重要な差異部分の抽出技術に関する研究)による．

## 文　献

[1] 渡邊拓也，太田学，片山薫，石川博，”格文法を用いた複数文書融合手法，”DBWS2004, July 2004.

[2] GoogleNews:http://news.google.co.jp/ Accessed 2005.

[3] Newsblaster: http://www1.cs.columbia.edu/nlp/newsblaster/ Accessed 2005.

[4] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st International Conference on Research and Development in Information Retrieval, pages 335--336, 1998.

[5] 茶筌:http://chasen.naist.jp/hiki/ChaSen/ Accessed 2005.

[6] 南瓜:http://chasen.org/~taku/software/cabocha/ Accessed 2005.