The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

# 共起語の類似度と時刻分布を利用した文書集合からの変化記述の対象抽出の試み

Extract Object of Changes from Documents using Similarities of Co-occurrence Word and its Time Distribution

田中 克明 *1
Katsuaki TANAKA

*1埼玉工業大学人間社会学部
Faculty of Human and Social Studies, Saitama Institute of Technology

In this paper, we proposed document sets consists of two types, a diversity description type that records actions to different objects and a change description type that recordings actions to objects that could be regarded as the same and attempted to extract objects of a change description documents. We assumed that co-occurrence words of a word indicating the object would have high similarities and words having different appearance time distributions in change description documents. As a result, we confirmed that it is possible to extract objects of a change description type documents.

## 1. はじめに

人間は、さまざまなことを文書として残している。文書の種類によらず、文書は「人間」が文書の記述の「対象」に何らかの「行為」(「観察」も行為とする) を行いその内容を記したものであること、また、文書数は時間の経過に沿って増えていくことが、共通する。

文書記述の例として、植物を観察して文書に記録する場合を考えると、「ツユクサが咲いていた (A)」「アジサイがもうすぐ咲きそうだ (B)」「アジサイが色づいた (C)」「アジサイの花に見えるの部分は萼だ (D)」といった文書が考えられる。(A)(B)(C)(D) のうち、「ツユクサ」と「アジサイ」という別々の植物について記した (A)(B) は、図 1 のように、異なる対象への行為を記述した文書の集合 (以下、多様性記述型) である。これに対し、「アジサイ」について別の時刻に記した (B)(C) は、図 2 のように、同一とみなせる対象への異なる時刻の行為を記した文書の集合 (以下、変化記述型) である。なお、(B)(C) に対する (D) のように、同一とみなせる対象について記述でも、時間の経過が意味をなさない場合、多様性記述型であると考える。

変化記述型の文書集合では、同一とみなせる対象への異なる時刻の行為を記述するため、文書集合から時間の経過を読み取ることができる。一方、多様性記述型の文書集合からは、時間の経過を読み取ることができない。すなわち、本稿で扱う「同一とみなせる対象」とは、「何かが変化したということは、なにか変化しないものがあることが必須である。」[溝口 12] とされる「なにか変化しないもの」のことである。

次に、植物を観察して記された文書集合をもとに、新たに植物を育てることにしたとしよう。多様性記述型の文書集合からは、どのような種類の植物があるかの情報を得ることができ、育てる植物の種類を決めることに役立つ。変化記述型文書集合からは、どのように植物が育っていくかの情報を得ることができ、時間を追って植物を育てる途中で行う行動の決定に役立つ。このように、多様性記述型の文書集合と変化記述型の文書集合からは、得られる情報の性質が異なる。*1

連絡先: 田中克明, 埼玉工業大学人間社会学部情報社会学科, 〒369-0293 埼玉県深谷市普済寺 1690, jsai2019@katsuaki-tanaka.net

*1 多様性記述型の文書集合から得られた情報を変換し、同一の対象への行為を蓄積し変化記述型の結果を生み出すことが「知的な振る舞い」であると考えられるが、本稿では取り扱わない。

一般的に、文書をひとつの集合として捉えると、「(A)(B)(D) または (B)(C)」のように、同一の対象に関する記述か否かの明確な分類のもとに文書が記録されていることは少なく、図 3 のように多様性記述型と変化記述型が混在している。そこで本稿では、文書集合から変化記述型となっている部分集合を見つけることを目的とし、変記述型の文書集合の核となる「同一とみなせる対象」の抽出手法を提案する。

## 2. 関連研究

文書集合に記述されている内容を把握するための手法として、LDA[Blei 03] に代表されるトピックモデルが挙げられる。トピックモデルにより、文書に含まれる複数の特徴的な記述内容の確率分布を「トピック」として得ることができる。さらに、Dynamic Topic Models [Blei 06] などにより、文書が作成された時間の経過に沿ってトピックを抽出することが可能である。しかし、これらの手法では、トピックが、特徴が似た多様性記述型の内容を表すものか、変化記述型の内容を表すものかの判断は、トピックの内容を確認する人間が行う必要がある。

また、時間の経過に沿って文書を処理する手法として、ニュース記事や SNS などの文書集合から出来事 (イベント) に関する記述を時系列に追って抽出する研究がなされている [Wanner 14]。時系列文書を扱う点に置いて本研究と類似するが、社会性が大きい出来事の抽出を扱っており、抽出結果の利用者と対象文書の間に相応の共通理解が存在することを前提とする。

## 3. 提案手法

### 3.1 変化記述型文書集合と単語への言及の類似率

文書において同一の対象を記述する際に、変化記述型 (図 2) の文書集合では、以前と同じ対象であることを認識して記述を行う。逆に文書を読む場合においても、対象についてある程度類似した表現が行われていなければ、人間が同一の対象と認識することが難しくなる。そのため、対象への言及内容にある程度の類似があると考えられる (図 4)。一方、多様性記述型の文書集合 (図 1) では、以前の対象と同じかを意識せずに

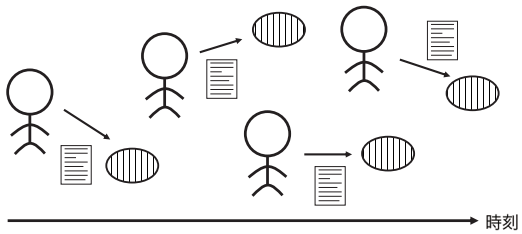


図 1: 異なる対象への記述からなる文書集合（多様性記述型）

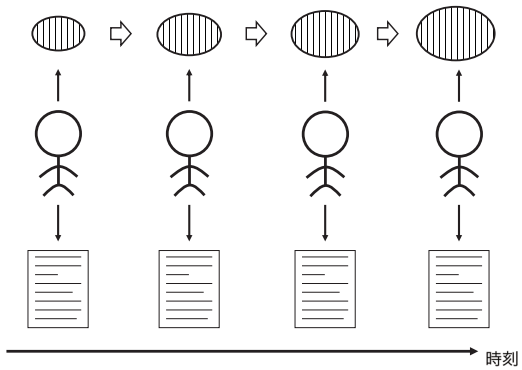


図 2: 関連する対象への記述からなる文書集合（変化記述型）

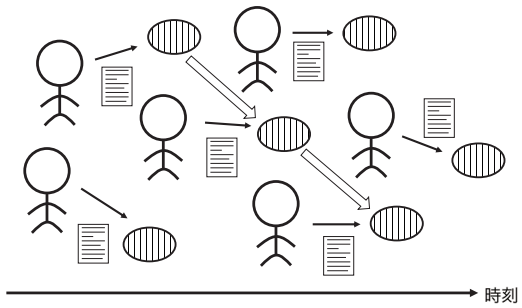


図 3: 一般的な文書集合（多様性記述型と変化記述型が混在）

記述を行うため、同じ対象を扱う文書において、対象への言及内容が類似するか否かは、一概に判断できない。

本稿では、対象が 1 つの単語で表現される状態を仮定し、ある対象を表現する単語への言及は、その単語と共起する単語により行われるとする。異なる時刻に記された文書において、対象を表現する単語に対して共起する単語が類似していることが、多様性記述型の文書集合を示すとなると考え、単語への言及の類似率（以下、言及類似率）を求める。すなわち、言及類似率が高い単語は、「同一とみなせる対象」であり、変化記述型の文書集合であるものと考える。

### 3.2 言及類似率の計算

言及類似率は、以下のように計算した。まず、着目単語 $w_a$ を定め、ある文書 $D_i$ において単語 $w_a$ と共起する単語のひとつを単語 $w_b$ とする。次に、文書 $D_i$ とは異なる文書 $D_j$ において単語 $w_a$ と共起する単語 $w_c$ に対して、単語 $w_b$ と類似するかの判定を行う。さらに、単語 $w_b$ と単語 $w_c$ が文書集合の中の異なる時刻で使われているかの判定を行う。これを繰り返し、単語 $w_a$ と共起するすべての単語 $w_b$ について、その他の文書で単語 $w_a$ と共起する単語 $w_c$ が、異なる時刻で類似する割合を、言及類似率として求める。

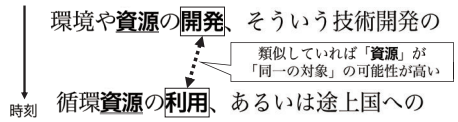


図 4: 同一の対象への記述判断例

表 1: 対象文書集合の概要

| 環境省中央環境審議会地球環境部会議事録 | |
|---|---|
| 期間 | 2001/2/16〜2012/10/24 |
| 文書数 | 5910（発言ごと） |
| 異なり単語数 | 12991 |

| 「人工知能」を含むツイート | |
|---|---|
| 期間 | 2013/12/25〜2014/6/6 |
| 文書数 | 43862（収集データの $\frac{1}{3}$） |
| 異なり単語数 | 22251 |

処理対象とする文書からの単語の取得は、MeCab[*2] による形態素解析の結果を用い、形態素解析後に名詞として得られた単語を選択することにより行った。また、単語 $w_a$ と 単語 $w_b$ が共起するとは、形態素解析により得られた名詞の単語列において、$w_a$ と $w_b$ が 5 単語以内あることを指すものとした。

単語が類似するかの判定は、日本語版 Wikipedia に含まれる単語を Word2vec[Mikolov 13] によりベクトル化した Wikipedia Entity Vectors[*3] から、全単語を 300 次元のベクトルとして学習済みのモデルを用い、単語を表現するベクトルのコサイン類似度

$$sim(\overrightarrow{w_b}, \overrightarrow{w_c}) = \frac{\overrightarrow{w_b} \cdot \overrightarrow{w_c}}{|\overrightarrow{w_b}||\overrightarrow{w_c}|} \quad (1)$$

を求め、$sim(\overrightarrow{w_b}, \overrightarrow{w_c}) \geq 0.5$ となる単語 $w_b$、$w_c$ を、類似するものとした。なお、処理対象の文書集合に存在するが、Wikipedia Entity Vectors に存在しない単語は、言及類似率の計算対象外とした。

単語 $w_b$ と単語 $w_c$ が異なる時刻で使われているかの判定は、単語 $w_a$ と共起して単語 $w_b$ が出現する時刻、同様に単語 $w_a$ と共起して単語 $w_c$ が出現する時刻の分布に着目し、相互の第 2 四分位範囲から第 3 四分位範囲に重なりがなければ、異なる時刻で使われているものとした。

## 4. 実験

### 4.1 対象とする文書集合

表 1 に示す 2 つの文書集合を対象とし、提案手法を適用した。1 つめは、日本の環境政策に関する諮問機関である環境省中央環境審議会のうち、地球温暖化に関する内容を中心に扱う地球環境部会の議事録である。議事録は、日時・出席者・議事次第・配布資料一覧・議事から構成され、会議 1 回ごとにほぼ同様の形式で記述されている。会議へは委員としてほぼ決まったメンバーが出席している他、外部のゲストが出席し話題の提供を行っていることが多い。会議の進行は、外部のゲストによる話題提供の後、その内容について委員が質疑と議論を行う、という形となっている。議論の内容は議事録の「議事」に記述されていたことから、「議事」部分のみを分析の対象とした。

∗2 http://taku910.github.io/mecab/

∗3 https://github.com/singletongue/WikiEntVec

表 2: 地球環境部会議事録 言及類似率

| | 単語 | 言及類似率 | 単語 | 出現率 |
|---|---|---|---|---|
| 1 | 法律 | 0.8638 | 委員 | 0.0101 |
| 2 | 事業 | 0.8554 | 環境 | 0.0095 |
| 3 | 大気 | 0.8511 | エネルギ | 0.0088 |
| 4 | 温室 | 0.8477 | 日本 | 0.0073 |
| 5 | 状況 | 0.8474 | 資料 | 0.0072 |
| 6 | 効率 | 0.8468 | 目標 | 0.0065 |
| 7 | 目的 | 0.8465 | 地球 | 0.0047 |
| 8 | 部分 | 0.8458 | 技術 | 0.0047 |
| 9 | 公共 | 0.8455 | 部会 | 0.0044 |
| 10 | 観点 | 0.8448 | 制度 | 0.0044 |

表 3: 「人工知能」を含むツイート 言及類似率

| | 単語 | 言及類似率 | 単語 | 出現率 |
|---|---|---|---|---|
| 1 | ロボット | 0.8519 | 人工知能 | 0.0684 |
| 2 | 人間 | 0.8474 | 表紙 | 0.0375 |
| 3 | 自分 | 0.8469 | 人工知能学会 | 0.0320 |
| 4 | 人 | 0.8407 | 女性 | 0.0309 |
| 5 | 学会 | 0.8266 | ロボット | 0.0292 |
| 6 | 人類 | 0.8254 | 男 | 0.0225 |
| 7 | 機械 | 0.8219 | 家事 | 0.0220 |
| 8 | 人工知能学会 | 0.8177 | 気持ち | 0.0218 |
| 9 | 表紙 | 0.8141 | まとめ | 0.0214 |
| 10 | 世界 | 0.8137 | NAVER | 0.0201 |

また、議事は会議における各個人の発言として記述されており、発言ごとに趣旨が異なると考えられることから、1 つの発言を 1 つの文書とみなし、83 の議事録から得られた 5910 発言を異なる文書として扱った。

2 つめは、Twitter より 2013 年 12 月〜2014 年 6 月に収集した「人工知能」を含むツイートである。この時期には、2014 年 1 月刊行の人工知能学会会誌の表紙が話題となり、数多くのツイートがなされた。収集したツイートの全体の約 $\frac{1}{3}$ から、公式リツイートや、URL などを取り除いたものを処理対象とした。なお、筆者らはこのデータを用いたトピック遷移分析システムの提案を行っている [田中 14]。

### 4.2 計算結果

得られた言及類似率上位 10 語と、比較のために出現率上位 10 語を、表 2、3 に示す。なお、上位 10 語（表 2、表 3）には、言及類似率の計算により得られた結果から、文書の「対象」となる可能性が低いと考えられる、「名詞-形容動詞語幹」（「必要」など「〜ない」となる単語）や「名詞-サ変接続」（「実験」「試験」など「〜する」となる単語）と形態素解析器により品詞付けされた単語を除いたものを示した。

### 4.3 上位語の評価

表 2、表 3 に示した言及類似率の上位 10 語について、各単語への言及内容について、単語とその周辺の記述を元の文書の内容を確認し、「同一とみなせる対象」であるか、評価を行った。出現頻度が高い単語では、文書集合の中で各単語が出現する箇所は数千以上にのぼるため、各単語について出現する箇所をランダムに 20 か所選択し、確認を行った。この際、変化記述型文書集合の対象は、図 3 のように時間経過に沿って出現すると考えられるため、文書集合を作成時刻順に 5 つのグループに分け、各グループから 4 文書ずつを選択した。評価の結果を表 4、5 に示す。

表 4: 上位 10 語の評価（地球環境部会議事録）

| 言及類似率 | | 出現率 | |
|---|---|---|---|
| 単語 | 対象 | 単語 | 対象 |
| 法律 | ○ | 委員 | × |
| 事業 | × | 環境 | × |
| 大気 | ○ | エネルギ | × |
| 温室 | ○ | 日本 | ○ |
| 状況 | × | 資料 | × |
| 効率 | × | 目標 | ○ |
| 目的 | ○ | 地球 | ○ |
| 部分 | × | 技術 | × |
| 公共 | ○ | 部会 | × |
| 観点 | × | 制度 | ○ |

表 5: 上位 10 語の評価（「人工知能」を含むツイート）

| 言及類似率 | | 出現率 | |
|---|---|---|---|
| 単語 | 対象 | 単語 | 対象 |
| ロボット | × | 人工知能 | ○ |
| 人間 | ○ | 表紙 | ○ |
| 自分 | × | 人工知能学会 | × |
| 人 | × | 女性 | × |
| 学会 | ○ | ロボット | × |
| 人類 | ○ | 男 | × |
| 機械 | ○ | 家事 | × |
| 人工知能学会 | × | 気持ち | × |
| 表紙 | ○ | まとめ | × |
| 世界 | × | NAVER | × |

単語とそれに対する周辺の記述を評価するにあたり、まず、文書集合の記述自身への言及は、「同一とみなせる対象ではない」と判断した。地球環境部会議事録の「委員」「環境」「資料」「部会」「地球」「部会」が該当する。また、同一内容の文書（ツイート）により多数出現する単語も、同様に判断した。「女性」「ロボット」「男」「まとめ」などが該当する。

### 4.4 考察

変化記述型の文書の「対象」と考えられる単語の割合は、地球環境部会議事録、「人工知能」を含むツイートともに、言及類似率を用いたほうが文書中で「対象」である単語の割合が高く、目標とした「対象」の抽出が行えていると言えそうである。

文書における実際の記述を確認すると、単語が複数の意味に使われている場合があり、この場合は、同一の対象とはみなせないと判断した。例えば、地球環境部会議事録における「事業」は、「売電事業者」「温室効果ガスを排出する事業者」などのように、複数の種類の事業（者）を表していた。同様に、ツイートにおける「人」は、「〜と発言した人」「人型」「人の意識」などさまざまな意味で用いられていた。

一方、ツイートにおける「人間」は、「人間の価値観」「人間の感情」「人間のコメント」のように、ほぼ同じ内容を表現していることから、同一の対象であると判断した。しかし、記述には時間に沿った変化は含まれておらず、「人間」を含む記述が、ツイートからなら文書集合において変化記述型の対象となっているとは言いがたい。

具体的な事物は、ほぼ対象として記述されていた。例えば、

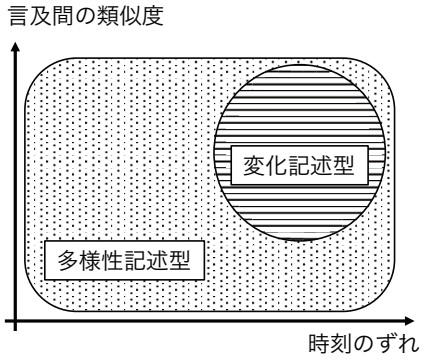


図 5: 多様性記述型・変化記述型文書集合の分布例

「法律」は「法律の施行」「法律の適用可能生」「法律大系全般」、「学会」は「学会誌にふさわしいか悩んだ」「学会誌の表紙について公式見解」「学会誌の特集がいい返しに」など、時間経過に沿った記述が行われており、変化記述型の対象となっている。

## 5. 課題

本稿における研究の課題とそれらついての検討を述べる。

変化記述型の文書集合には、「異なる時刻において共起語が類似する単語が存在する」ことを仮定して実験を行った。しかし、多様性記述型の文書集合でも、別の時刻の同一の対象について以前とは無関係に記述を行うことがあることから、得られた結果は、両方の記述形式が混ざったものとなる。両形式の文書集合を、対象への言及内容の類似度と記述時刻のずれを軸として整理すると、図 5 のように重なる。「同一とみなせる対象」を文書集合から見つけるためには、この重なりを区別する手法が必要である。

本稿では、「対象がある単語で表される」としたが、同一の単語が異なる意味で使われることが多々ある。これに対応するために、対象単語を含む記述の全てから文書をランダムに選択して評価を行うのではなく、評価の対象とする単語への言及類似度が高い部分ごとに分割することで、変化記述型の文書集合を選択できるか、検討を行う。

また、実験結果を確認するために、言及類似率上位の単語と出現頻率が上位の単語との比較を行ったが、単語 $w_a$ とそれぞれ異なる文書で共起する単語 $w_b$、$w_c$ は、単語の共起関係をネットワークと考えた際、単語 $w_a$ を媒介として接続されことから、今後、媒介中心性と言及類似率の比較を行いたいと考えている。

次に、提案手法における単語間の類似度の計算のために Wikipedia Entity Vectors として配布されている Wikipedia の記述内容をもとにした単語のベクトルデータを用いた。このため、実験対象の文書集合には出現するが Wikipedia Entity Vectors に含まれない単語の類似度計算は行えない。計算対象とする文書集合も含めて単語のベクトル表現を求めることにより、文書中のすべての単語について、類似度計算を行うことができる。また、類似度計算が正確に行えているかを比較、評価することも必要である。

## 6. おわりに

本稿では、文書集合には多様性記述型と変化記述型の 2 通りが考えられることを述べ、変化記述型の記述の「対象」を抽出するため、言及類似率を提案した。提案手法を 2 つの形式が異なる文書集合に適用し、結果を確認した。今後、課題として述べた点を改良しつつ、性質の異なる文書集合への本手法への適用を試みる。

## 謝辞

本研究は JSPS 科研費 JP16K00702 の助成を受けたものである。

## 参考文献

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1002 (2003)

[Blei 06] Blei, D. M. and Lafferty, J. D.: Dynamic Topic Models, in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120 (2006)

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient Eestimation of Word Representations in Vector Space, in *Proceedings of International Conference on Learning Representations 2013* (2013)

[Wanner 14] Wanner, F., Stoffel, A., Jckle, D., Kwon, B. C., Weiler, A., and Keim, D. A.: State-of-the-Art Report of Visual Analysis for Event Detection in Text Data Streams, in *EuroVis - STARs*, pp. 125–139 (2014)

[溝口 12] 溝口 理一郎：オントロジー工学の理論と実践, オーム社 (2012)

[田中 14] 田中 克明：Twitter におけるトピック遷移分析システムの提案, 第 7 回インタラクティブ情報アクセスと可視化マイニング研究会予稿集, pp. 22–27 (2014)