共起語の類似度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 based on Similarity of Co-occurring Words

田 中 克 明*
Katsuaki TANAKA

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。日々の作業メモや会議の議事録、SNSへの書き込みまで、その種類は多岐にわたる。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」（ただ観察したことも行為とする）を行いその内容を記したものであること、また、文書の数は時間の経過に沿って増えていくことが、共通する。

ここで、植物を観察して文書に記録する場合を想定する。このとき、「スイセンが咲いていた(A)」「梅のつぼみが大きくなってきた(B)」「梅の花が咲いた(C)」といった内容が記述された文書を、例として考えることができる。(A)(B)(C)のうち、「スイセン」と「梅」という別々の植物について記した(A)と(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、「梅（厳密に同じ樹かどうかはわからないが）」について記した(B)と(C)は、図２のように、同一とみなせる対象への行為を記した文書の集合（以下、変化記述型）である。変化記述型の文書集合では、同一の対象への異なる時刻の行為を記述するため、時間の経過が文書集合から読み取れる。一方、多様性記述型では、対象が同一かを意識しないため、時間の経過を文書から読み取れない可能性がある。

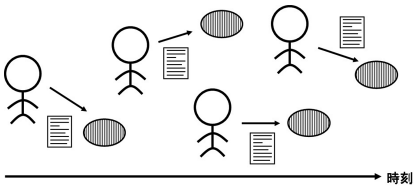

図１：異なる対象への記述からなる文書集合
（多様性記述型）

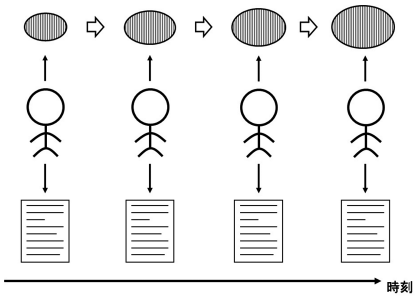

図２：関連する対象への記述からなる文書集合
（変化記述型）

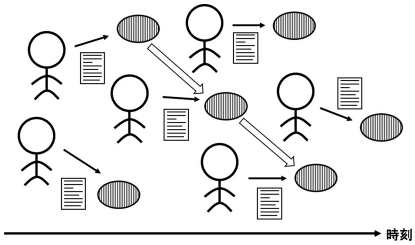

図３：一般的な文書集合
（多様性記述型と変化記述型が混在）

次に、植物を観察して記された文書集合をもと

＊　埼玉工業大学人間社会学部情報社会学科

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。しかし、一般的に文書をひとつの集合として捉えると、「(A)(B)または(B)(C)」のように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。

本研究では、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「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［1］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［2］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、特徴が似た多様性型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［3］。変化記述型文書集合のように時間経過に沿った変化に着目する点は本研究に近いが、社会性が大きい出来事の抽出を扱っており、得られた情報を扱う人間が、それらの出来事を知っていることを前提とする。これに対し本研究では、文書集合の内部での変化記述型の対象を抽出し、読み手が新しい知識を得る支援を行うこと目的とする。

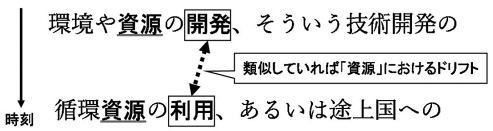

図4：ドリフトであるかの判断例

その他、変化記述型の文書集合の作成・管理を意図したシステムとして、オントロジーを中心として人工物の設計・運用中を行うシステム［4］があげられる。本研究における「対象」および「対象への行為」を、人工物の構造と対応して構造化されたオントロジーとして記述し、オントロジーに沿って文書を記録・管理を行う。

さらに、オントロジーの構築においては、「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。」［5］として、本研究の変化記述型の文書集合における「対象」と同様の概念が必要とされている。

3．提案手法

⑴　対象への言及内容の「ドリフト率」

文書集合において同一の記述対象が出現する際に、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ある程度認識して文書を作成する。あるいは、対象についてある程度類似した記述を行っていなければ、同一の対象と人間が認識することは難しい。そのため、対象への言及内容はある程度の「ずれ（ドリフト）」の範疇に収まると考えられる。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は認識せ

表１：対象文書集合の概要

CubeSat XI-IV設計議事録	
期間	2000/1/5～2002/12/12
文書数	580
異なり単語数	7879
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

ずに文書を作成する。そのため、同じ対象について記した文書でも、文書が異なると対象への言及内容が、ずれよりも、大きく異なったものになると考えられる。

そこで本研究では、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と同時に出現する単語の内容が、別の文書でも類似していれば、対象への言及がドリフトの範囲に収まっていると考え（図４）、その割合を「ドリフト率」として求める。ドリフト率が低く言及内容がばらばらの単語は「多様性記述型文書の対象」であり、一方、ドリフト率が高く言及内容に類似性がある単語は、変記述型の文書集合の「対象」であると考えられる。

⑵　ドリフト率の計算

ドリフト率は、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 との類似度を計算する。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、類似する単語 w_c が存在する割合を求め、これを単語 w_a のドリフト率とする。

単語を求めるために、処理対象とする文書はMeCab［６］を用いて形態素解析を行う。形態素解析後に名詞として得られた単語をドリフト率計算の対象とした。また、単語 w_a と w_b が共起するとは、形態素解析の結果として得られた単語列において、単語 w_a と w_b が５単語以内あることを指すものとした。

単語間の類似度は、日本語版Wikipediaに含まれる単語をWord2vec［７］によりベクトル化したWikipedia Entity Vectors［８］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

$$\mathrm{cosinesim}(\vec{w}_a, \vec{w}_b) = \frac{\vec{w}_a \cdot \vec{w}_b}{|w_a||w_b|}$$

により求めた。類似度が別途定めるしきい値以上の単語 w_a , w_b を、類似した単語とした。

４．実験

⑴　対象とする文書集合

２つの文書集合を対象とし、提案手法を適用した。１つ目は、小型人工衛星CubeSat XI-IV［９］の設計会議議事録（以下、CubeSat議事録）、２つ目は環境省中央環境審議会地球環境部会［10］の議事録（以下、地球環境部会議事録）である。各文書集合の概要を表１に示す。

CubeSat議事録は、東京大学大学院工学系研究科航空宇宙工学専攻中須賀研究室にて行われた小型人工衛星CubeSat XI-IVの設計・運用プロジェクトに関連して作成された、議事録・マニュアル・実験記録などからなる。同プロジェクトは、衛星の機能ごとに複数のチームに別れて進められ、議事録もチームごとに作成されている。そのため、各議事録において、日付、タイトル、文書作成者名は統一された書式で記されているが、会議内容部分は、少しずつ異なる形式で記述されている。

地球環境部会議事録は、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

表２：上位20語（CubeSat議事録）
グレーアウトは異なる上位20語にも出現する単語

	ドリフト率	単語	出現率	単語
1	0.8794	基板	0.009701	XI
2	0.8773	XI	0.009201	電源
3	0.8761	熱	0.008207	月
4	0.8725	Ⅱ	0.007960	アンテナ
5	0.8725	太陽電池	0.007335	基板
6	0.8723	回路	0.006646	データ
7	0.8706	方法	0.006220	太陽電池
8	0.8683	アンテナ	0.006036	温度
9	0.8679	Ⅲ	0.005947	電圧
10	0.8678	状況	0.005642	地上
11	0.8649	電池	0.005631	電子
12	0.8648	温度	0.005584	電池
13	0.8646	カメラ	0.005137	カメラ
14	0.8640	軌道	0.004606	構造
15	0.8623	構造	0.004527	回路
16	0.8619	衛星	0.004233	電力
17	0.8614	真空	0.003975	環境
18	0.8564	データ	0.003796	CW
19	0.8559	ネットワーク	0.003754	OBC
20	0.8539	電源	0.003702	電流

表３：上位20語（地球環境部会議事録）
グレーアウトは異なる上位20語にも出現する単語

	ドリフト率	単語	出現率	単語
1	0.9376	地球	0.009766	委員
2	0.9331	温室	0.009366	環境
3	0.9318	環境	0.009078	エネルギ
4	0.9308	目標	0.007254	日本
5	0.9297	政府	0.006816	資料
6	0.9291	制度	0.006591	目標
7	0.9275	基本	0.004765	技術
8	0.9273	国際	0.004656	地球
9	0.9266	取組	0.004472	制度
10	0.9264	公共	0.004310	部会
11	0.9259	事業	0.004307	社会
12	0.9255	国内	0.004284	経済
13	0.9255	効果	0.004238	部分
14	0.9254	国民	0.004189	効果
15	0.9250	日本	0.004013	国際
16	0.9242	京都	0.003897	基本
17	0.9242	状況	0.003731	産業
18	0.9241	国	0.003597	政策
19	0.9238	中央	0.003487	京都
20	0.9236	税	0.003456	CO

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た。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。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。

⑵　計算結果

CubeSat議事録のドリフト率上位20語を表２、地球環境部会議事録のドリフト率上位20語を表３に示す。比較のために、出現率上位20語も記した。ドリフト率上位20語のうち背景をグレーにしたものは、出現率上位20語にも含まれる単語、出現率上位20語のうち背景をグレーにしたものは、ドリフト率上位20語にも含まれる単語である。上位20語（表２、表３）には、ドリフト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。

ドリフト率の計算における各単語が類似してい

表４：上位20語の評価（CubeSat議事録）

ドリフト率上位20語		出現率上位20語	
単語	対象	単語	対象
基板	○	XI	○
XI	○	電源	○
熱	○	月	×
Ⅱ	○	アンテナ	○
太陽電池	○	基板	○
回路	○	データ	×
方法	×	太陽電池	○
アンテナ	○	温度	○
Ⅲ	○	電圧	×
状況	×	地上	×
電池	○	電子	×
温度	○	電池	○
カメラ	○	カメラ	○
軌道	○	構造	×
構造	×	回路	○
衛星	○	電力	×
真空	○	環境	×
データ	×	CW	○
ネットワーク	○	OBC	○
電源	×	電流	×
「対象」割合	75%		55%

表５：上位20語の評価（地球環境部会議事録）

ドリフト率上位20語		出現率上位20語	
単語	対象	単語	対象
地球	×	委員	×
温室	○	環境	×
環境	×	エネルギ	×
目標	○	日本	○
政府	○	資料	×
制度	○	目標	○
基本	×	技術	×
国際	○	地球	○
取組	×	制度	○
公共	○	部会	×
事業	×	社会	○
国内	○	経済	×
効果	○	部分	×
国民	○	効果	○
日本	○	国際	○
京都	○	基本	×
状況	×	産業	×
国	○	政策	○
中央	○	京都	○
税	○	CO	○
「対象」割合	70%		50%

るか否かの判定は、先に述べたコサイン類似度が0.5以上であるか否かによって行った。また、それぞれの議事録に存在するが、Wikipedia Entity Vectorsに存在しない単語（CubeSat議事録の883単語、地球環境部会議事録の845単語）は、ドリフト率の計算対象外とした。

⑶　上位語の評価

表２、表３に示したドリフト率の上位20語について、変化記述型の文書集合の対象となっているか評価を行うため、元の文書中の記述を調査し、各単語への言及内容を確認した。出現頻度が高い単語であることもあり、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、各単語が出現する箇所を20か所ずつ選択し、確認行った。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、表５に示す。

単語とその周辺の記述を評価するに当たり、まず、会議体自身への言及は、会議進行上の言及であることが多いため、「対象ではない」と判断することにした。CubeSat議事録の「構造」「電源」「地上」「電子」「環境」および地球環境部会議事録の「地球」「環境」「委員」「部会」が該当する。

⑷　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、CubeSat議事録、地球環境部会議事録

ともに、ドリフト率上位20語では約７割、出現率上位20語では約５割であった。ドリフト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ことがわかる。

文書における実際の記述を確認すると、具体的な事物は、ほぼ対象となっていた。例えばCubeSat議事録における「基板」は、衛星に搭載する基板を指し､基板のサイズ､基板周りへのパーツ配置、作成した基板のテスト、衛星搭載時の物理的干渉の確認、基板へのエポキシ塗布、一部基板の作り直し、動作チェック、発注ミスによる再発注などの記述がなされていた。

具体的な事物ではないCubeSat議事録の「熱」は、熱の制御、安定化、熱サイクル、熱膨張、熱解析、解析過程での動作検証、熱対策、実衛星での試験にともなう熱膨張と､「熱」をどのように扱うかの検討過程が記述がなされており､「対象」であると判断した。

５．課題

本研究の課題とそれに対する検討を述べる。

まず、単語への記述がドリフトであるかを判定する際に、対象と仮定した単語 w_a と共起する単語 w_b、w_c の選択を､「異なる文書」から行い、文書の生成時刻には着目しなかった。このため、例えば、単語 w_b、w_c が交互に単語 w_a と共起する場合でも、ドリフトであると判定している。変化記述型の対象であるためには、図２のように時間的な順序関係が重要であることから、文書の生成時刻のパターンを考慮して、ドリフトであるかの判定を行う必要があるだろう。

次に、本研究では、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。ドリフト率の計算対象とする文書集合から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ろう。

また、実験結果として示した表２、表３では、ドリフト率上位の単語に対し、出現頻率が上位の単語との比較を行った。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、媒介中心性とドリフト率の比較を行いたいと考えている。

６．おわりに

本論文では、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、ドリフト率の計算手法を提案した。提案手法を２つの形式が異なる会議の議事録の集合に適用し、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。

７．参考文献

［１］ D. M. Blei, A. Y. Ng , M. I. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, Vol.3, pp. 993-1022, 2003.

［２］ D. M. Bleii , J. D. Lafferty, “Dynamic Topic Models,” Proceedings of 23rd International Conference on Machine learning, pp. 113-120, 2006.

［3］ F. Wanner , et al., “State-of-the-art report of visual analysis for event detection in text data streams,” Computer Graphics Forum, Vol.33, No.3, pp. 1 -15, 2014.

［4］ 高藤淳, 來村徳信 , 溝口理一郎, “オントロジー工学に基づく技術知識統合管理システムの発展とビジネス展開,” 人工知能学会論文誌, Vol.26, No.5, pp. 547-558, 2011.

［5］ 溝口理一郎, オントロジー工学の理論と実践, オーム社, 2012.

［6］ “MeCab: Yet Another Part-of-Speech and Morphological Analyzer,” http://taku910.github.io/mecab/, 2018年12月27日アクセス.

［7］ T. Mikolov, K. Chen, G. Corrado , J. Dean, “Efficient estimation of word representations in vector space,” Proceedings of International Conference on Learning Representations 2013, 2013.

［8］ “Wikipedia Entity Vectors,” https://github.com/singletongue/WikiEntVec, 2018年12月27日アクセス.

［9］ “東京大学工学系研究科中須賀研究室 衛星プロジェクト,” https://www.space.t.u-tokyo.ac.jp/nlab/project.html#s3, 2018年12月27日アクセス.

［10］ “環境省中央環境審議会地球環境部会,” https://www.env.go.jp/council/06earth/yoshi06.html, 2018年12月27日アクセス.