通時的対象抽出への文書分類の適用に関する検討

A Study of Applying Document Classification to Diachronic Object Extraction

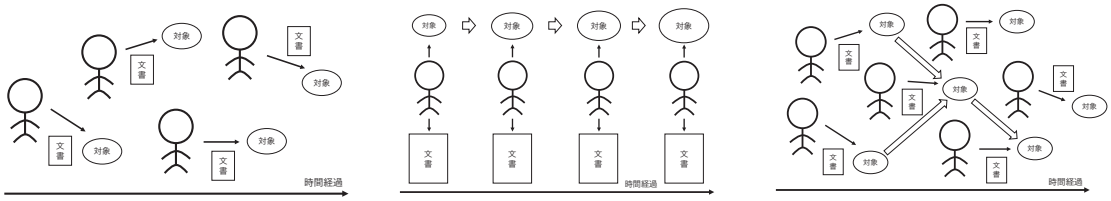
田 中　克 明*
Katsuaki TANAKA

1．はじめに

人間の活動に伴い、試行錯誤の過程などを記録した様々な文書が作成され蓄積されている。また、計算機を用いて、キーワードを指定して検索を行う、大規模言語モデルを通して文書内容に関する質問を行い回答を得るなどの方法により、文書集合を理解するための支援を得ることもできる。しかし、キーワードや質問など、「何に着目して文書集合を理解しようとするか」という情報は、人間から計算機に与える必要がある。

文書集合には様々な異なる内容を記した多様性記述型集合（図1-a）と、ある対象への一連の行為を記した変化記述型集合（図1-b）があり、実際の文書集合ではこれらが混在（図1-c）していると考えられる。変化記述型の集合には人間が行った試行錯誤など時間経過に沿った変化が含まれることから、変化記述型集合を得ることが文書集合の理解において重要である。そこで筆者らは、変化記述型集合の記述対象である、通時的対象を文書理解の要として抽出する手法の検討を進めてきた。これまでに、通時的対象への言及は関連する行為であるため類似した表現になると仮定し、言及が類似する度合いをもとにして通時的対象を抽出する手法 [1] [2]、対象の候補が出現する前後の文脈が類似する度合いをもとにして通時的対象を抽出する手法 [3]を提案した。

これらの手法では、通時的対象の候補となる単語に対して、文書集合全体での出現状況をもとに通時的対象の抽出を行う。そのため、単語が複数種類の異なる対象を表す場合や、複数の変化期記述型集合の対象、すなわち目的が異なる試行錯誤の対象となっている場合に、それらを区別することができなかった。そこで本論文では、通時的対象への言及を含めた文書断片をクラスタリングによりあらかじめ分類しておき、分類されたクラスタごとに通時的対象の抽出を行う手法を提案する。



a．多様性記述型　　b．変化記述型　　c．多様性記述・変化記述混在型

図1：多様性記述型・変化記述型・混在型の文書集合

*埼玉工業大学人間社会学部情報社会学科

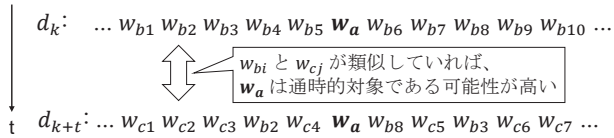


図2：通時的対象の抽出

2．通時的対象と複数の変化記述型集合

本研究では、前述したように文書集合内に記述された時間の経過に沿った一連の行為（変化記述型集合）の対象を通時的対象と名付け、通時的対象の候補を文書中の単語から抽出することを目指す。

例えば、以下のような文a～fがあるとする。

a. 桜のつぼみが付いた。
b. 桜の花が咲いた。
c. 桜の花の絵を描いた。
d. 桜の木が緑色になった。
e. 桜が紅葉した。
f. 桜の葉をお茶にした。

a→eは、時間の経過に沿って桜に関する観察や絵を描くという行為が、時間の経過に沿って行われており、桜を通時的対象と考えることができる。また、a～eのような記述の集合を変化記述型集合と呼ぶことにする。一方、f は a～e全体とは直接関係ないが、桜の葉が成長したものを茶の材料にしたことから、d→f（eは含まない）という変化記述型集合と考えられる。このように、同一の対象である「桜」に対して、「四季を通して観察する」「葉を茶にする」という2つの異なる変化記述型集合が存在し得る。

このとき、通時的対象の抽出を文集合a～f全体から行うのではなく、(a, b, c)と(d, e, f)の2つに分類した上で行うことにより、桜の花に関する行為と桜の葉に関する行為を分けて扱うことができる。そこで本稿では、文書集合をあらかじめ分類し、分類結果を対象として通時的対象の抽出を行うことで、同一の記述対象への複数の変化記述型集合を、別々の集合へ分割することを目指す。

3．提案手法

前節で述べたように、本稿では、通時的対象の抽出対象となる文を分類した後、分類結果を対象に通時的対象の抽出処理を行う手法を提案する。分類と抽出の処理は以下の流れで行う。

⑴　文書断片の作成とクラスタリング

通時的対象の抽出は、図2に示すように候補となる単語（w_aとする）の前後の単語列（w_bあるいはw_cとする）の類似度により行う。そこでまず、文書集合中のすべての文書についてMeCab [4]を用いて形態素解析を行い、名詞および未知語と判定された単語を候補となる単語として得る。次に、候補となる単語のすべての出現箇所について、その前後N語の範囲からなる文書断片を生成する。

続いて、文書断片から抽出処理を行う前にこれらの分類を行う。すべての文書断片についてSentence BERT [5] [6]を用いて埋め込み表現を求め、HDBSCAN [7] [8]により埋め込み表現のクラスタリングを行う。この結果得られたクラスタを1つの文書集合とみなし、クラスタごとに文書断片からの通時的対象の抽出を行う。

⑵　クラスタ内の言及類似率による通時的対象の抽出

通時的対象となる単語の候補の抽出には、言及類似率による手法 [2]を用い、言及類似率が高い単語を通時的対象の候補とする。ここでは、前節で行ったクラスタリングの結果を用いて、同一クラスタ内の文書断片を対象に以下の手順で言及類似率の計算を行う。

まず、文書断片を生成する元としたw_a以外のすべての単語w_{bi}(i=1,...,m)と、w_{bi}が出現する時刻$t(w_{bi})$を求める。文書断片はw_aの前後N単語から

生成されしたため、$\{w_{bi}\}$は、クラスタ内においてw_aと前後N語の範囲で共起するすべての単語を示す。

次に単語間の類似度をもとに、以下の式によりw_aに対する言及類似率SMR(w_a)を求める。

$$\mathrm{SMR}(w_a) = \frac{\sum_{i=1}^{m}\left(\max\{\mathrm{sim}(w_{bi}, w_{bj}); j \in i+1, \ldots, m\}\right)}{m}$$

単語間の類似度sim(w_{bj}, w_{bk})は、単語wに対して日本語Wikipediaの文書をもとに学習済みのword2vec [9]モデルであるWikipedia Entity Vector [10]から得た埋め込み表現$\vec{w}$と、wが出現する時刻$t(w)$の集合の四分位範囲により、以下のように定義した。

$$\mathrm{sim}(\overrightarrow{w_{bi}}, \overrightarrow{w_{bj}}) = \begin{cases} \dfrac{\overrightarrow{w_{bi}} \cdot \overrightarrow{w_{bj}}}{|w_{bi}||w_{bj}|} & \{t(w_{bi})\}と\{t(w_{bj})\}の第2・3四分位範囲が重ならないとき \\ 0 & \{t(w_{bi})\}と\{t(w_{bj})\}の第2・3四分位範囲が重なるとき \end{cases}$$

4．実験と評価

複数の文書集合に提案手法を適用し、通時的対象の抽出を行った。

⑴　対象とする文書集合

文脈類似率による通時的対象の抽出 [3]と同様に、表2に示す3つの文書集合を対象とし、クラスタリングとクラスタ内の言及類似率の計算を行った。

1つ目の文書集合であるCubeSat XI-IV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 [11]の設計プロジェクトにともない作成された、議事録・マニュアル・実験記録などである。

表1：文書集合の概要

CubeSat XI-IV議事録	
期間	2000/1/5～2002/12/12
文書数	580
異なり単語数	7879
地球環境部会議事録	
期間	2001/02/16～2012/10/24
文書数	5910（発言ごと）
異なり単語数	12991
「人工知能」を含むツイート集合	
期間	2013/12/25～2014/6/6
文書数	43862
異なり単語数	22251

2つ目の地球環境部会議事録は、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 [12]の議事録である。この議事録は会議ごとに作成され、日時、出席者、議事次第、配布資料一覧、議事の詳細からなり、各会議でほぼ同じ形式を持つ。議論の内容は「議事」セクションに記載されているため、文書集合にはこの部分のみを含めた。また、会議での各発言は独立した趣旨を持つと考えられるため、1つの発言を1つの文書として扱い、83の議事録から5910の発言を異なる文書として文書集合を作成した。

3つ目の「人工知能」を含むツイートからなる文書集合は、2013年12月から2014年6月にTwitter（現X）から収集した「人工知能」を含むツイートである。この期間には、2014年1月に発行された人工知能学会会誌の表紙が多くの注目を集め、多数のツイートが投稿された。収集したツイートの約3分の1から、リツイートと既存のツイートと冒頭20文字が一致するツイートを除外したものを処理対象の文書集合とした。

表2：関連性評価値

評価値	評価概要
3	評価語を対象とした変化記述がなされている
2	評価語が複数の対象を表し、複数の変化記述がある
1	評価語が複数の対象を表し、変化記述と多様性記述が混在する
0	評価語が記述の対象ではない、または変化記述がない

表3：通時的対象の抽出結果と評価

	CubeSat XI-IV 議事録			地球環境部会議事録			「人工知能」を含むツイート		
	単語	SMR	評価	単語	SMR	評価	単語	SMR	評価
1	ピン	0.5463	1	自体	0.5605	0	人工知能	0.5589	0
2	条件	0.537	1	GDP	0.5585	1	人	0.5547	0
3	素子	0.5349	1	考え	0.5580	0	女性	0.5543	0
4	範囲	0.5325	2	制度	0.5576	1	人間	0.5512	1
5	CUBE	0.5322	3	中国	0.5559	3	ジェン	0.5498	1
6	他	0.5308	0	そのもの	0.5536	0	表紙	0.5495	2
7	事項	0.5298	1	アメリカ	0.5527	0	自分	0.5493	0
8	DCDC	0.5278	3	別	0.5524	0	人間	0.5471	1
9	外部	0.5276	1	エネルギ	0.5519	1	アンドロイド	0.5468	1
10	センサ	0.5272	2	農業	0.5518	1	感情	0.5467	1

表4：文書集合と手法ごとの精度(P)とDCG10

	CubeSat VI-IV 議事録		地球環境部会議事録		「人工知能」を含むツイート	
手法	P	DCG_{10}	P	DCG_{10}	P	DCG_{10}
言及類似率（クラスタリングあり）	0.9	7.197	0.5	3.409	0.6	2.654
言及類似率（クラスタリングなし）	0.8	9.891	0.9	12.7	0.7	5.068

⑵　結果

言及類似率の計算を、N=5、hdbscanによるクラスタリング時の最小クラスタ要素数を20として行った。表3に、各文書集合から3章に述べた手法によりクラスタリング後に得られた言及類似率（SMR）の上位10単語を示す。また、各単語が通時的対象であるかを確認するために、各単語の言及類似率の計算を行ったクラスタに含まれる文書断片の記述を確認し、通時的対象であるか否かの関連性評価値を表2に基づいて付与した。この際、単語によってはクラスタ内の文書断片数が多数となるため、文書断片の出現期間を5つに分割し、各期間からランダムに4つずつ文書断片を選択することにより合計20個の文書断片の確認を行った。クラスタ内の文書断片数が20を下回る場合には、クラスタ内のすべての文書断片を確認した。

また、この関連性評価値を用いて、Discounted Cumulative Gain（以下、DCG）[13]を求めた。DCGは、文書検索システムの評価などに用いられる値であり、結果として得られたうちの上位k個の関連性評価値を用い、以下の式により求めることができる。

$$\mathrm{DCG}_k = rel_1 + \sum_{i=2}^{k} \frac{rel_i}{\log_2 i}$$

さらに、関連性評価値が 1 以上であれば、変化記述型の文書集合を探す手がかりとすることが可能であることから、表3において評価値が1以上である単語の割合を精度として求めた。求めたDCG10と精度（P）の値を表4に、クラスタリングによる分類を行わずに抽出を行った結果 [2]とあわせて示す。

5．考察

⑴　実体を示さない単語の抽出

表4の言及類似率上位10語による通時的対象の抽出精度、DCG10を確認すると、CubeSat XI-IV議事録においてわずかに精度が向上した以外、どちらの値も事前にクラスタリングを行わない既存手法の方が良い値となった。特に地球環境部会議事録では値が大きく低下している。表3にて言及

表5：地球環境部会議事録における「自体」の出現例

日付	クラスタ内の文書断片における出現例（一部）
2003/1/29	影響の調査等を行っております。基本的には、油**自体**の量はそれほど多くなったものですから、
2003/9/25	幹事会決定なども踏まえ、このペーパー**自体**、タイトルのところに京都メカニズム活用連絡
2004/4/2	目途がまったく立たないということ。その目標**自体**が例えばRPS法の今議論がありましたように、
2004/11/9	資料の３ページ目にございますが、作業**自体**はそれほど大きなものではございませんで、
2005/6/29	をしないとなかなか日本の環境政策**自体**も進まないという点があると思います。ここでは
2006/6/2	排泄ということですが、これは活動量**自体**がだんだんと減っているということもありまして、

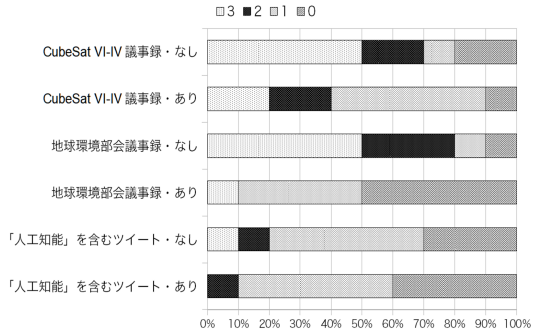


図3：関連性評価値の割合
（なし：クラスタリングなし、あり：クラスタリングあり）

類似率が上位の単語を確認すると､「自体」「考え」「そのもの」「別」など、実体を表さずに実体を指示する単語が抽出されている。例として、地球環境部会議事録において言及類似率が1位となった「自体」が抽出されたクラスタ内の文書断片の一部を、表5に示す。「自体」が何を示すかは断片によって異なるが､「多くなか った」「めどが立たない」「大きくない」「進まない」「減っている」などが「自体」とともに出現しており､「自体」への言及の類似度は高い。すなわち、本手法の目指した結果は得られており、手法そのものが通時的対象手法としては適切ではないことに起因すると考えられる。

これらの単語は既存手法 [2]では上位には現れなかった。これは、文書集合全体から求めた言及類似率に対し、クラスタリングによりw_aを含む文書断片のうち類似したものがクラスタを形成した結果、クラスタ内での言及類似率が高くなったものと考えられる。

提案手法では通時的対象の候補として、形態素解析により得られた品詞が名詞または未知語である単語を言及類似率の計算対象としている。よりよい抽出結果を得るためには、固有表現抽出などと組み合わせ、実体を表す単語に限定するなどの手法が考えられる。

(2)　単一の変化記述型集合の割合

本論文では2章に述べたように、同一の通時的対象に対する変化記述集合が単一のものとするために、言及類似率を計算する前にクラスタリングを導入した。同一の通時的対象が単一の変化集合ではなく複数の変化記述集合とかかわる場合、表2に示した関連性評価の結果は「2」となる。そこで、クラスタリングを行わない既存手法 [2]と提案手法における関連性評価値の割合の比較を行った（図3)。地球環境部会議事録では「2」の割合が減少しており、本手法の効果を確認できる。一方、CubeSat XI-V議事録､「人工知能」を含むツイートでは「2」の割合に変化がなく､「1」または｢0｣の割合が増加した。とくに｢1｣に関しては、関連性評価値を付与する際には評価語が関わる変化記述対象が単一か複数かを分けていないが、評価の過程において、複数ではなく単一の対象を示す場合を比較的多く確認できた。そのため、事前のクラスタリングの効果はあったものと考えられる。

関連性評価値を付与する際には、4.2節にて述べたように、評価語を含む文書断片を時間経過に沿ってランダムに選択している。そのため、同一クラスタ内だが言及の類似度が比較的低い部分が

選ばれる場合もあり、関連性評価の判断に影響を与えることが考えられる。通時的対象の抽出目的は変化記述型集合を見つけることにあるため、評価手法を再検討し、言及の類似度が高い部分を優先的に抽出する、評価「1」において対象が単一か複数かを分割する、などの方法を新たに導入することが考えられる。

6．おわりに

本論文では、文書集合から変化記述型集合を見つける手掛かりとなる通時的対象について、単一の変化の連なりに対する集合を見つけることを目指した。そのために既存手法を改良し、クラスタリングを行った後にクラスタごとに言及類似率を計算し、通時的対象を抽出する手法の提案と評価を行った。その結果、クラスタリングが影響を与えることを確認できた。

単一の変化記述集合に関わる部分を見つけることはできたが、抽出結果の評価は全体として低下した。この結果を踏まえ、言及類似率の計算候補とする単語を固有表現に限定する、評価を行う際に元の文書を確認する際に言及の類似度が高い部分を抽出するなど､手法の改善点の検討を行った。今後、これらの検討に加え、文書集合の性質の違いも加味し、抽出手法の改良と、通時的対象の活用手法の検討を進める予定である。

参考文献

[1] 田中克明, “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,” 埼玉工業大学人間社会学部紀要, 第 17号, pp. 15-21, 2019.

[2] K. Tanaka , K. Hori, “Finding Diachronic Objects of Drifting Descriptions by Similar Mentions,” Proc. of 2019 Pacific Rim Knowledge Acquisition Workshop, pp. 32-43, 2019.

[3] 田中克明, “文脈の類似度に着目した通時的対象の抽出の検討,” 埼玉工業大学先端科学研究所アニュアルレポート, 第 21号, pp. 10-17, 2023.

[4] “MeCab: Yet Another Part-of-Speech and Morphological Analyzer,” http://taku910.github.io/mecab/, 2024年1月10日アクセス.

[5] N. Reimers, I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp. 3982-3992, 2019.

[6] I. Sonobe, “sonoisa/sentence-bert-base-ja-mean-tokens-v2 · Hugging Face,” https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2, 2024年1月10日アクセス.

[7] R. J. G. B. Campello, D. Moulavi , J. Sander, “Density-Based Clustering Based on Hierarchical Density Estimates,” Advances in Knowledge Discovery and Data Mining, pp. 160-172, 2013.

[8] L. McInnes, J. Healy , S. Astels, “The hdbscan Clustering Library,” https://hdbscan.readthedocs.io/, 2024年1月10日アクセス.

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

[10] “Wikipedia Entity Vectors,” https://github.com/singletongue/WikiEntVec, 2024年1月10日アクセス.

[11] “XI (人工衛星),” https://ja.wikipedia.org/wiki/XI_(人工衛星), 2024年1月10日アクセス.

[12] “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, 2024年1月10日アクセス.

[13] K. Järvelin , J. Kekäläinen, “Cumulated gain-based evaluation of IR techniques,” ACM Transactions on Information Systems, Vol.20, No.4, pp. 422-446, 2002.