# 文脈の類似度に着目した通時的対象の抽出の検討

**A Study on Diachronic Object Extraction Focusing on Contextual Similarity**

田中 克明
Katsuaki TANAKA

埼玉工業大学 人間社会学部 情報社会学科
Department of Informational Society Studies, Faculty of Human and Social Studies,
Saitama Institute of Technology

概要

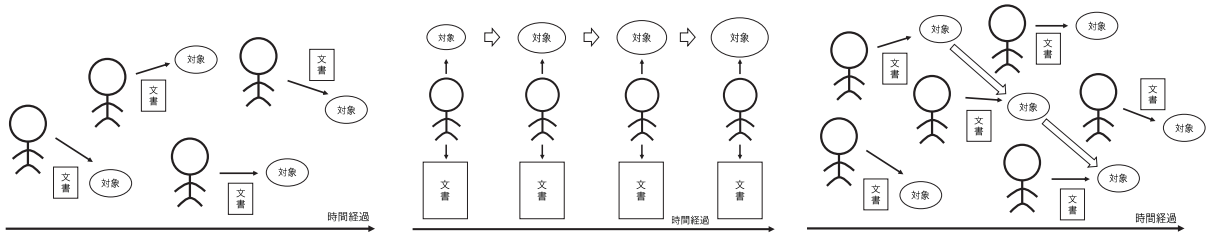
人間の活動を記録した文書集合には、試行錯誤の過程が記録されている。試行錯誤は、複数の関連する行為が、時間経過に沿って同一の対象へなされることにより行われる。そこで、この対象を通時的対象と名付け、対象の出現する文脈が類似している割合により、通時的対象を抽出する手法を提案する。提案手法を複数の文書集合に適用し、文書集合によって異なる評価結果を得た。

## 1 はじめに

人間のさまざまな活動を記録しそこから新たな知見を得るために、多くの文書が作成され蓄積されている。例えば、新たな人工物を設計する際には、要求を満たすためにどのような作業を行う必要があるか、過去の試行錯誤をこれらの蓄積された文書の集合から探し、参考にする。

例として、人工衛星の設計過程を観察して記録した文書集合として、作成時刻順に、「太陽電池の試験を行った (A)」「無線機を購入した (B)」「無線機の試験を行った (C)」「無線機の機種を決定した (D)」という文書が蓄積されている場合を考える。文書の分類を行うと、「太陽電池」「無線機」のように、どのような種類の機器を扱ったかを読み取ることができる。このような情報は、「どのような機材が必要か」の決定を行う際に役に立つ。一方、文書の作成時刻に着目すると、例えば、(B) と (C)、(D) の関係から、「無線機について、購入から試験の実施、機種決定へ」という作業の過程が読み取れる。このような時間の経過に沿った情報は、設計を進める際に、「ある対象についてどのように作業を進めるか」の決定を行う際に役立つと考えられる。

このように、文書集合には、(1) 記述対象の種類に関する情報（以下、多様性記述型、図 1(1)）、(2) 記述対象の変化に関する情報（変化記述型、図 1(2)）の大きく分けて 2 種類の情報が含まれており、一般的な文書集合ではこれらが混在している（図 1(3)）と考えられる [9]。




(1) 多様性記述型　(2) 変化記述型　(3) 多様性記述・変化記述型混在

図 1: 文書集合における文書と記述対象の関係

多様性記述型の文書集合から情報を得るためには、LDA [3] や Top2vec [1] などのトピックモデルを用いることができる。一方、変化記述型の文書集合から情報を得るためには、Dynamic Topic Models [2] や Dynamic Embedded Topic Model [4] を利用することが考えられるが、時間経過に沿った俯瞰的な単語の分布を取得するこれらの手法では、文書集合においてどの記述対象の変化に着すべきかの示唆を得るためには不十分である。

## 2　通時的対象の抽出

### 2.1　通時的対象

変化記述型の文書集合には図 1(2)(3) のように、時間経過に沿った行為の対象となる、同一の対象が存在する。この対象のことを、以下、通時的対象と呼ぶ。通時的対象は、試行錯誤などの時間経過に沿った一連の行為の対象として行為主体が認識し、文書にも記述されるものである。一方、文書の蓄積過程において、行為の主体が認識しないうちに、偶然の結果として同一の対象についての記述が行われる可能性もある。この場合は、記述されている対象が同一であっても、通時的対象とは考えない。

すなわち、通時的対象とは、文書集合の作成者が、異なるタイミングで行われる複数の行為に共通する対象であると認識している対象であり、文書集合の作成者（複数でもかまわない）が、何らかの行為（観察も行為に含む）を連続して行い、文書集合中にその変化の過程が記述されたものである。

### 2.2　言及の類似による通時的対象の抽出

変化記述型の文書集合では、文書の作成者が、記述対象が以前と同じ対象であることを認識して記述を行う。文書の読み手の立場から考えても、対象についての記述がある程度類似していなければ、対象が同一であると認識することは難しい。そのため、文書集合に含まれる通時的対象への言及には類似性があると考えられる（図 2）。

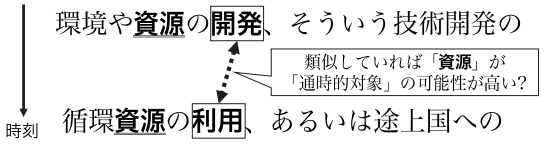



図 2: 類似する言及による通時的対象の抽出

これまでに筆者らは、対象への言及はその単語と共起する単語により行われるとした上で、通時的対象の候補となる文書中の単語 $w_i$（図 2 では「資源」）に対して、異なるタイミングで共起する異なる単語 $w_a$、$w_b$（図 2 では「開発」「利用」）の類似度を求め、全ての異なるタイミングの共起語間の類似度の平均を $w_i$ の言及類似率とし、言及類似率により通時的対象を求める研究を行ってきた [8][9]。言及類似率が高い単語と出現頻度が高い単語について、通時的対象であるかの比較評価を行い、言及類似率が高い単語は通時的対象である可能性が高く、言及類似率が通時的対象の抽出に有用と考えられることを確認している [8]。

## 3　文脈の類似による通時的対象の抽出

### 3.1　通時的対象と出現する文脈

言及類似率では、着目する単語と共起する単語1つ1つの類似度を用いて、通時的対象の抽出を試みた。これに対して、文の類似度を用いれば、着目する単語が出現する前後の単語列全体により、通時的対象であるかの判断を行えると考えられる（図3）。そこで、着目する単語の前後の単語列を文脈と考え、文脈の類似する割合（以下、文脈類似率）により、通時的対象の抽出を行う手法を検討する。

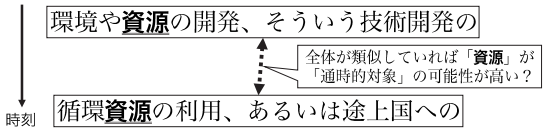



図 3: 類似する文脈による通時的対象の抽出

### 3.2　文脈類似率

文脈類似率を求める対象を表す単語を $w_i$ とする。まず、文書集合において $w_i$ が出現する周辺の文字列からなるテキスト $s_k$ を、$w_i$ の文脈として取り出す。言及類似率の計算では、前後5単語を言及と見なして計算対象としたため、同様の範囲を $s_k$ として取り出す。

次に、複数のテキスト $s_k$ 同士の類似度の計算を行うために、$s_k$ をベクトル化する。文のベクトル化には、文の分散表現を得ることを目的として BERT を改良した Sentence BERT [6] を利用した。実際の計算には、HuggingFace[1]にて公開されている日本語データを用いて学習済みの Sentence BERT のモデル [7] を用いた。

着目単語 $w_i$ が出現する2つのテキスト $s_a$、$s_b$ に対し、Sentence BERT により求めたベクトル表現を $\overrightarrow{s_a}$、$\overrightarrow{s_b}$ とする。このとき、$\overrightarrow{s_a}$ と $\overrightarrow{s_b}$ の類似度 $\mathrm{sim}(s_a, s_b)$ を、コサイン類似度もとに以下のように定義する。

$$\mathrm{sim}(s_a, s_b) = \begin{cases} 0 & (s_a = s_b \text{または} t(s_a) = t(s_b)) \\ \dfrac{\overrightarrow{s_a} \cdot \overrightarrow{s_b}}{|\overrightarrow{s_a}||\overrightarrow{s_b}|} & (s_a \neq s_b) \end{cases} \tag{1}$$

対象を含む同一のテキストは、対象について同じ内容を述べており、通時的対象について記しているものではないと考えられるため、類似度が0となるようにした。また、同一時刻に作成されたテキストも、異なるタイミングに行われた行為ではないため、通時的対象であるかの判断には役立たないと考え、類似度を0とした。なお、$t(s_k)$ は、テキスト $s_k$ が含まれる文書の作成時刻を表す。

このように定義した $\mathrm{sim}(s_a, s_b)$ を用いて、文脈類似率 $\mathrm{SCR}(w_i)$ を、以下により求めた。

$$\mathrm{SCR}(w_i) = \frac{\log(n)}{m} \sum_{a=1}^{n} \sum_{b=a+1}^{n} \mathrm{sim}(s_a, s_b) \frac{|t(s_a) - t(s_b)|}{\sigma} \tag{2}$$

なお、着目する単語 $w_i$ の文書集合中の出現回数を $n$、$\mathrm{sim}(s_a, s_b) \neq 0$ であった回数を $m$、単語 $w_i$ が出現する文書の作成時刻の標準偏差を $\sigma$ とする。また、出現回数が2、3回程度と少ない単語で

[1] https://huggingface.co/

言及している文脈が類似する場合、類似度の平均が大きくなる。そのため、出現回数が大きい単語の言及類似率がある程度大きくなるよう、類似度の平均に対して単語の出現回数に基づく $\log(n)$ を乗じた値を文脈類似率とした。

# 4 実験と考察

## 4.1 対象とする文書集合

表 1 に示す 3 種類の文書集合を対象に、文脈類似率を求めた。

1 つめの小型人工衛星設計議事録は、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[2] 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・実験記録などである。

| 小型人工衛星設計議事録 | |
|---|---|
| 期間 | 2000 年 1 月 5 日〜2002 年 12 月 12 日 |
| 文書数 | 398 |
| 異なり単語数 | 7877 |

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

| ツイート集合（**2014** 年前半の「人工知能」検索結果） | |
|---|---|
| 期間 | 2013 年 12 月 25 日〜2014 年 6 月 6 日 |
| 文書数 | 43862（収集データの $\frac{1}{3}$） |
| 異なり単語数 | 22251 |

表 1: 対象文書集合の概要

2 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[3]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。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」部分のみを文書集合に含めた。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。

3 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文書集合とした。

## 4.2 実験結果

各文書集合を対象に、文脈類似率を求めた。文脈類似率上位 10 位までの単語について、文書集合中での単語の出現回数（$n$）、文脈類似率（SCR）、後述する関連性評価値 $rel_i$ を表 2 に示す。地球環境部会議事録で最上位の「オマスエタノール」は、文書中には「バイオマスエタノール」として出現しており、形態素解析器に同単語が登録されていなかったため得られた表記である。

結果を確認するために、文脈類似率が上位 10 位の単語それぞれについて、単語とその周辺の記述を元の文書の内容を確認し、各単語が通時的対象であるか、各単語が出現する箇所の記述を確認し、Discounted Cumulative Gain（以下、DCG）[5] を求めるための関連性評価値 $rel_i$ を付与した。

[2]https://www.space.t.u-tokyo.ac.jp/nlab/project.html#s3
[3]https://www.env.go.jp/council/06earth/yoshi06.html

| | 小型人工衛星設計議事録 | | | | 地球環境部会議事録 | | | | ツイート集合 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 単語 | $n$ | SCR | $rel_i$ | 単語 | $n$ | SCR | $rel_i$ | 単語 | $n$ | SCR | $rel_i$ |
| 1 | SSG | 21 | 8.5830 | 3 | オマスエタノール | 13 | 5.1163 | 3 | CH | 9165 | 10.391 | 0 |
| 2 | コンピュータ | 29 | 4.6982 | 3 | 資料 | 4742 | 4.1006 | 0 | NAVER | 9198 | 10.342 | 0 |
| 3 | CCM | 10 | 4.4618 | 3 | エネルギ | 5428 | 4.0728 | 2 | 気持ち | 9987 | 10.171 | 0 |
| 4 | DOWN | 13 | 4.4343 | 1 | 清貧 | 33 | 4.0704 | 0 | 家事 | 10060 | 10.154 | 0 |
| 5 | スペクトル | 17 | 4.3864 | 2 | 議事 | 562 | 4.0047 | 0 | 男 | 10280 | 9.8477 | 0 |
| 6 | CCN | 22 | 3.8884 | 3 | 日本 | 4522 | 3.9421 | 3 | まとめ | 9771 | 9.5587 | 0 |
| 7 | 氷 | 18 | 3.8599 | 0 | 目標 | 4353 | 3.9056 | 2 | 女性 | 14206 | 8.2942 | 0 |
| 8 | PCH | 11 | 3.8336 | 0 | 原子力 | 1110 | 3.8359 | 3 | ロボット | 13380 | 7.9655 | 0 |
| 9 | GATE | 18 | 3.7599 | 2 | 環境 | 5896 | 3.8016 | 2 | 人工知能学会 | 14784 | 7.8154 | 0 |
| 10 | アンテナ | 1082 | 3.7129 | 2 | 省エネ | 1547 | 3.7927 | 1 | 表紙 | 17389 | 7.2577 | 2 |

表 2: 文脈類似率（SCR）上位 10 語と評価

$\mathrm{DCG}_k$ 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。

$$\mathrm{DCG}_k = rel_1 + \sum_{i=2}^{k} \frac{rel_i}{\log_2 i} \tag{3}$$

本実験では、文脈類似率の大きい単語が通時的対象であるかの評価を行うことを目的とし、関連性評価値を表 3 のように定めた。

| 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: 通時的対象かを判断するために用いた関連性評価値

出現数が多い単語では、文書集合中での出現箇所が数千以上にのぼる。そのため、評価対象の各単語について、出現箇所をランダムに 20 か所選択し、記述を確認することにより評価値を定めた。この際、通時的対象は、図 1(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 文書ずつを選択して記述の確認対象とし、関連性評価を行った。

また、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出来ることから、表 2 において $rel_i \geq 1$ である単語の割合を、精度として求めた。

表 4 に、各文書集合における、文脈類似率、言及類似率と出現頻度がそれぞれ上位 10 語の単語について、精度（P）と $\mathrm{DCG}_{10}$ の値を示す。

| | 小型人工衛星設計議事録 | | 地球環境部会議事録 | | ツイート集合 | |
|---|---|---|---|---|---|---|
| 手法 | P | $\mathrm{DCG}_{10}$ | P | $\mathrm{DCG}_{10}$ | P | $\mathrm{DCG}_{10}$ |
| 文脈類似率 | 0.8 | 11.46 | 0.7 | 9.067 | 0.1 | 0.6021 |
| 言及類似率 | 0.8 | 9.891 | 0.9 | 12.70 | 0.7 | 5.068 |
| 出現頻度 | 0.8 | 9.548 | 0.7 | 7.517 | 0.2 | 3.000 |

表 4: 文書集合ごとの精度 (P) と $\mathrm{DCG}_{10}$

### 4.3 考察

#### 4.3.1 文書集合ごとの傾向の検討

表 4 を確認すると、まず、ツイート集合にからの通時的対象の抽出が上手く行えていないことが分かる。これは、文書集合中に同一の記事のタイトルを含む（2ch 書き込みのまとめ、NAVER まとめなど）ツイートが多いためであると考えられる。記事タイトルのみのツイートであれば、同一のテキストの類似度を 0 とする $\mathrm{sim}(s_a, s_b)$ の定義により類似度が小さくなるが、記事タイトル何らかの追記がなされると同一のテキストではなくなるため、$\mathrm{sim}(s_a, s_b)$ の値が大きくなる。このため、「CH」「NAVER」など記事のタイトルに含まれる単語の文脈類似率が高くなったと考えられる。これを解決するためには、1. 記事タイトルのみが比較対象となるように、文脈として取り出すテキストを短くする、2. $\mathrm{sim}(s_a, s_b)$ が一定以上の値のものが多数ある場合には、文脈類似率の計算対象から除外するなどの方法が考えられる。

一方、小型人工衛星設計議事録における文脈類似率による通時的対象の抽出は、$\mathrm{DCG}_{10}$ の値が大きく、比較的良い結果が得られていることがわかる。表 2 の $n$（単語の出現回数）を見ると、比較的出現回数が少ない単語が抽出されている。これらも通時的対象ではあるが、より出現回数が多い候補も得ることもできるよう、例えば出現回数が一定以上の単語についてのみ文脈類似率を求めるなど、手法の検討が必要である。

地球環境部会議事録では、文脈類似率に基づく上位語を言及類似率による実験結果 [8] と比較すると、単語の出現回数 $n$ が大きい単語については同様の傾向が見られる一方、$n$ が小さい通時的対象となる単語も抽出できている点が異なると言える。小型人工衛星設計議事録では $n$ が小さい語が主として抽出されたことなど踏まえ、文脈類似率の性質、計算手法について、さらなる検討が必要であろう。

#### 4.3.2 単語ごとの傾向の検討

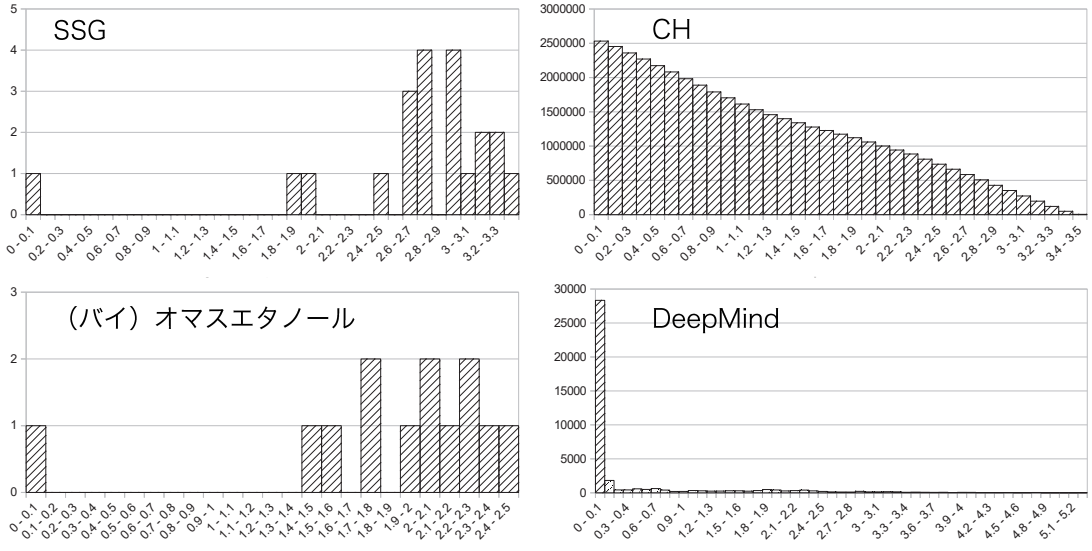



図 4: $\mathrm{sim}(s_a, s_b)\frac{|t(s_a)-t(s_b)|}{\sigma}$ の分布

図 4 に、表 2 で言及率が最上位であった単語について、式 (2) の

$$\sum_{a=1}^{n}\sum_{b=a+1}^{n}\mathrm{sim}(s_a,s_b)\frac{|t(s_a)-t(s_b)|}{\sigma} \tag{4}$$

における $\mathrm{sim}(s_a,s_b)\frac{|t(s_a)-t(s_b)|}{\sigma}$（ただし $\mathrm{sim}(s_a,s_b)\neq 0$ の場合）の分布を示す。同時に、ツイート集合に含まれる単語「DeepMind」に関する分布も示す。

「SSG」「(バイ) オマスエタノール」は、出現回数はどちらも 10〜20 回程度と少ないが、類似する文脈において出現しており、式 4 の計算結果も、複数の大きな値を示していることを確認できる。「CH」は 2ch まとめブログのタイトルとそれへの言及として出現している。そのためか、その他の出現回数が多い単語の分布では指数関数的な凹形状を示す 0.4〜2.6 付近において、直線的な分布を示しており、出現回数が多い。その結果、これらを平均した文脈類似率が高い値を示したものと考えられる。

「DeepMind」は、ツイート集合において「Google が DeepMind を買収した」という報道への言及として出現している。報道記事のタイトルの引用などを行わずに、作成者自身が記した文からなるツイートがほとんどであり、また、報道への言及を行うのみで継続した記述対象とはなっていない。このように、典型的な多様性記述型の対象を表す単語の場合には、相互の類似度が低い文脈中に多く出現すると考えられる。

## 5 おわりに

文書集合において、その蓄積の過程における試行錯誤の中心であったと考えられる通時的対象の抽出を、対象とみなす単語の前後の文字列の類似度により求められる文脈類似率により行う手法を提案し、複数の性質の異なる文書集合を対象に実験を行った。また、実験結果に対する考察を行った。

文脈類似率を用いることにより、既存手法の言及類似率と比較して、文書集合中で比較的出現回数が少ない通時的対象を見つけることができ、また、典型的な多様性記述型の対象を見つけられる可能性もあることがわかった。一方、類似した文脈で出現するものの、ただ言及されているだけであり試行錯誤の中心ではない対象があり、文脈類似率を用いた抽出では、これらが候補の上位となってしまう可能性があることを確認した。

また、異なるタイミングにおける記述であることを、式 2 では比較するテキストの作成時刻の差と着目対象の記述の作成時刻の標準偏差により重み付けした。4.3 節では検討を行わなかったが、着目対象の出現回数と合わせ、作成時刻のばらつきと記述対象が通時的対象であるかの関連について、さらなる検討を行う予定である。

## 参考文献

[1] Dimo Angelov. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020.

[2] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120, 2006.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.

[4] Adji B Dieng, Francisco JR Ruiz, and David M Blei. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*, 2019.

[5] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, Vol. 20, No. 4, pp. 422–446, 2002.

[6] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[7] Isao Sonobe. sonoisa/sentence-bert-base-ja-mean-tokens-v2 · hugging face. https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2. Accessed: 2023/2/5.

[8] Katsuaki Tanaka and Koichi Hori. Finding diachronic objects of drifting descriptions by similar mentions. In *Proc. of Pacific Rim Knowledge Acquisition Workshop*, pp. 32–43, 2019.

[9] 田中克明. 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. 第 33 回人工知能学会全国大会論文集, 2019.