Sent2vecを用いた文書の特徴ベクトルと抽出的要約の生成手法

Kadoki, Tomu / 門木, 斗夢

(出版者 / Publisher)
法政大学大学院情報科学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編
(巻 / Volume)
16
(開始ページ / Start Page)
1
(終了ページ / End Page)
6
(発行年 / Year)
2021-03-24
(URL)
https://doi.org/10.15002/00023864

Sent2vec を用いた文書の特徴ベクトルと抽出的要約の生成手法
Generating document feature vector and extractive summary using sent2vec

門木　斗夢
Tomu Kadoki
法政大学情報科学研究科情報科学専攻
E-mail: tomu.kadoki.6e@cis.k.hosei.ac.jp

Abstract

In this paper, we propose a feature vector to express semantic contents of a document and proposes a method to generate the semantic summary of the document using the feature vector. First, a document is represented in a numerical matrix that is built up from a set of distributed representations of sentences, such as sent2vec. Next, singular value decomposition is applied to the matrix to generate three feature matrixes of the document. Then, we properly analyze a column vector, called singular vector, to choose feature vectors for expressing semantic contents of the document. The semantic summary is generated by each component of the feature vector, or by calculating distances between semantic content vector and sentences in the document. As a result of experiment, we found the feature vector represented the main and sub topics of the document. Furthermore, as a result of the experiment with a dataset of text summarization called DUC2002, we found the generating semantic summary shared 45% of the words in single document summarization and 38% of the words in multiple document summarization against the manually created summary.

1. まえがき

近年の情報化社会では，大量のテキスト情報がインターネットを介して交換されている．情報化が進むにつれ，情報サイトやブログだけでなく，テレビや新聞で報道されていたニュースや，SNS など，そのテキスト情報の数や種類は増加の傾向にある．人々は，これらのテキストから情報を取得し，自分で情報を整理し，行動をしている．しかし，テキスト情報が膨大に増えていく中で，人々が得ることができる情報も増え続け，必要な情報が埋もれてしまうことや，情報を取得・整理することに対する負担も大きくなっている．このような状況において，テキストの内容を意味的に分析し，分類や検索に役立てる技術が求められている．大量のテキスト情報をコンピュータが意味的に分析し，分類や検索，要約などのタスクを自動的に行うことで，人々の負担を大幅に減らすことができるだけでなく，より有用な情報を得ることにもつながる．このような研究の例として，単語や文の意味的特徴を低次元ベクトルとして表す分散表現などの研究が進められている．古典的な TF-IDF などの疎なベクトルではなく，分散表現の密なベクトルで文や単語を表現することで，より意味的な内容を表現できるだけでなく，コンピュータの負担も減らすことができる．

本研究では，文の分散表現を拡張し，文の集合としての文書を対象に，文書の意味的なベクトルを生成する手法を検討する．文の分散表現の Sent2Vec[1]を用いることで，文書を行列表現へと変換し，特異値分解を行うことで，文書の意味的な特徴ベクトルを生成する．また，その特徴ベクトルを利用し，教師なしで抽出型の文書要約を行うタスクについて，従来の手法とは異なる手法を提案し，従来の手法との比較や，その特徴などを実験する．

2. 関連研究

2.1. Sent2Vec

Sent2Vec[1]は，文の分散表現モデルの一つであり，各文に対し，文の意味的な特徴を表すベクトル表現を与える技術である．古典的な TF-IDF などでは，単語の出現頻度に基づいたベクトルで単語や文を表現するため，各ベクトル表現の次元数は非常に大きくなり，各ベクトル表現は疎なベクトルとなってしまう．一方で，分散表現では，各単語や文を低次元で固定次元のベクトルで表現し，それらのベクトルは非常に密となる．よって，コンピュータの負担を減らすことができるだけでなく，文や単語の意味的特徴を各ベクトル表現でより表現できる．Sent2Vec[1]では，文の構成単語と単語の n-gram によって文を表現する．そして，ある文からその文中の単語や n-gram を推定するというタスクを大量の文を用いてニューラルネットワークで学習することで，各単語と単語の n-gram の特徴が重み行列に学習され，それぞれのベクトルが特徴ベクトルとなる．この学習された特徴ベクトルを用いて，一文中に現れるすべての単語と n-gram の特徴ベクトルを加算し，平均ベクトルを計算することで，文のベクトル表現を生成する．単語や n-gram の特徴ベクトルの学習過程においては，ニューラルネットワークを用いることで，大量の文から文脈を学習する必要があるが，文のベクトル表現を生成する過程においては，単純な平均ベクトルの計算によって生成することができるため，生成についての計算量を軽減できている点が特徴である．

2.2. 自動文書要約

自動文書要約には，いくつかの種類が存在する．まず，要約の対象となる文書の数である．ある一つの文書から

Supervisor: Prof. Satoru Fujita

要約を生成する手法を単一文書要約，複数の文書から要約を生成する手法を複数文書要約という．次に，要約を生成する手法についてである．文書内の文の関係性や重要性に基づき，元となる文書から重要な文を抽出する抽象型要約と，文書内の意味をコンピュータが解釈し，新たな要約文を複合的に生成する抽象型要約である．例えば，抽出型要約を行う手法の一つとして，潜在意味解析(LSA)を用いる手法では，文と単語の関係性を示す行列として文書を表現し，その行列に対して特異値分解を行うことで，その文書内のトピックにおける各文の影響度を示す特徴ベクトルとその重要度を計算する．その後，それらを用いて適切な文を選択することで要約を生成する．この時，関係性を示す行列には，TF-IDF を用いる手法や，単語の分散表現を用いる手法[2]などが提案されている．また，グラフ構造を用いて抽出型要約を行う手法も提案されている．この手法では文書中の各文をノードとし，2 つの文の間の類似度を各枝の重みとしたグラフを構築する．そして，構築したグラフ上で重要なノードを計算することで要約となる抽出する．TextRank[3]では，2 つの文に共通する単語によって類似度が計算され，そのグラフ上で PageRank アルゴリズムを実行することで，重要なノードで決定される．また，TF-IDF を拡張した，BM25 と呼ばれる指標に TextRank の類似度を置き換えて計算する手法も研究されている[4].

3. 提案手法

3.1. 文書の特徴ベクトル

本研究では，文書を文の集合として扱い，その構成要素である文のベクトル表現から，文書の特徴ベクトルを計算する手法を提案する．

まず，本手法では，文書を文単位に分割し，構成要素の文をベクトル表現$\mathbf{s}_j$に変換する．この時，本手法では，文をベクトル表現へと変換する手法として，文の分散表現の Sent2Vec を用いる．したがって，ベクトル表現は文の意味的特徴を示し，すべてのベクトル表現の次元数は Sent2Vec のモデルで設定された次元数で固定される．次にこのベクトル表現$\mathbf{s}_j$を文書内の時系列順に並べることで，式(1)のように，文書の行列表現Dを得る．

$$\mathrm{D} = (\mathrm{s}_1\ \mathrm{s}_2 \dots \mathrm{s}_n)^{\mathrm{T}} \tag{1}$$

この時，全てのベクトル表現は同じ次元数となるため，文書の行列表現Dの大きさは，(文の数)×(Sent2Vec のベクトル表現の次元数)となる．また，Sent2Vec のベクトル表現は各文の意味的特徴を示すことから，この行列表現Dは，文書の各文の意味的特徴によって構成され，文書の文レベルの意味的特徴を示すと考えられる．

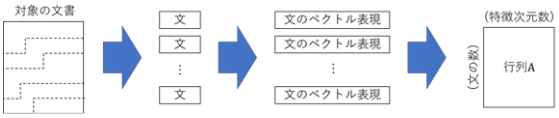


図 1.文書から行列表現を得る手法.

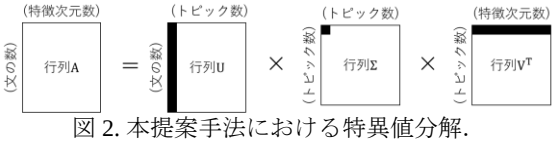


図 2. 本提案手法における特異値分解.

この行列表現Dに対して，式(2)のように特異値分解を行い，U，Σ，V の 3 つの行列の積形式に分解する．

$$\mathrm{D} = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}} \tag{2}$$

この時，UとVは直交行列，Σは対角行列である．特異値分解の対象となる文書の行列表現Dを m×n の行列とした場合，行列Uの各列ベクトル$\mathbf{u}_i$は左特異ベクトルと呼ばれ，行列Dに関する m 次元の正規直行基底である．また，行列Vの各列ベクトル$\mathbf{v}_i$は右特異ベクトルと呼ばれ，行列Dに関する n 次元の正規直交基底である．そして，行列Σの対角成分σ_iは，対応する特異ベクトルの行列Dにおける重要度となり，各対角成分を特異値と呼ぶ．

LSA や本手法では，これらの特異ベクトルが文書のトピックを示す．本手法での右特異ベクトル$\mathbf{v}_i$の要素数は，Sent2Vec のベクトル表現の次元数と同数であることから，右特異ベクトル$\mathbf{v}_i$は，Sent2Vec のモデルと同じ空間にあって，文書のトピックを意味的にマップした特徴ベクトルであると考えることができる．一方で，左特異ベクトル$\mathbf{u}_i$の要素数は，文書の文の数と同数であるため，左特異ベクトル$\mathbf{u}_i$は，各トピックに関する各文の影響度を表す．この時，各特異値は，文書の各トピックや対応する特異ベクトルの重要度を示すスカラ量となる．

さらに，文書全体を意味的に支配するトピックの特徴ベクトル$\mathbf{v}_i$は，それに対応するσ_iが大きくなると考えられるため，特異値分解をした結果に対して，対応するσ_iが降順となるように，右特異ベクトル$\mathbf{v}_i$や左特異ベクトル$\mathbf{u}_i$を入れ替えてソートする．また，特異値分解の性質上，$\mathbf{v}_i$は文書を構成する多くの文に対して逆向きのベクトルとなることがあるため，右特異ベクトル$\mathbf{v}_i$がそれぞれ逆向きのベクトルかを判定し，逆向きのベクトルである場合には，その右特異ベクトル$\mathbf{v}_i$と対応する左特異ベクトル$\mathbf{u}_i$の全要素の符号を反転させる．逆向きのベクトルを判定する手法については，文書の各文のベクトル表現$\mathbf{s}_j$との距離によって判定する平均距離法と，対応する$\mathbf{u}_i$の要素によって判定する成分法の二つの手法を提案する．まず，平均距離法では，右特異ベクトル$\mathbf{v}_i$とその逆ベクトル$-\mathbf{v}_i$について，文書の文のベクトル表現$\mathbf{s}_j$との距離をそれぞれ計算し，すべての文における平均距離をそれぞれ計算する．そして，式(3)のように，平均距離が近いベクトルを正しいベクトルとして判定する．

$$\begin{cases} \sum_k distance(v_i, s_k) < \sum_k distance(-v_i, s_k) \ \rightarrow \ u_i, v_i \\ \sum_k distance(v_i, s_k) > \sum_k distance(-v_i, s_k) \ \rightarrow \ -u_i, -v_i \end{cases} \tag{3}$$

この時，distance は距離関数であり，本研究の実験ではマンハッタン距離を用いる．平均距離法では，文書のトピックを正確に反映できるが，実行時間を多く必要とする．

一方，成分法では，対応する左特異ベクトル$\mathrm{u_i}$の要素の半分以上が負の値となる場合には，逆方向のベクトルである可能性が高いと考え，式(4)のように，逆方向のベクトルを正しいベクトルとして判定する．

$$\begin{cases} \left(\sum_k sgn(u_{ik})\right) > 0 \quad \rightarrow \quad u_i, v_i \\ \left(\sum_k sgn(u_{ik})\right) < 0 \quad \rightarrow \quad -u_i, -v_i \end{cases} \tag{4}$$

この時，sgn は符号関数である．成分法は，平均距離法よりも実行時間が短くなるというメリットがある．

以上の処理を行った結果，Vの第1列特異ベクトルになった$\mathrm{v_i}$について，文書の中心的なトピックを示す，中心トピックベクトルとして定める．Vの第2列以降は，サブトピックを表現する特徴ベクトルとなる．

3.2. 要約文の抽出

本研究では，中心トピックベクトルに関して，トピック距離法とトピック選択法の 2 種類の抽出要約手法を提案する．まず，トピック距離法では，中心トピックベクトル$\mathrm{v_1}$と，文書中の各文の Sent2Vec のベクトル表現$\mathrm{s_j}$との距離を計算する．そして，この距離をスコアとし，スコアが小さい順に文をソートし，上位の文を要約文として抽出する．一方で，トピック選択法では，中心トピックベクトル$\mathrm{v_1}$の各文に対する影響度を示す$\mathrm{u_1}$を用いる．$\mathrm{u_1}$の各要素は，中心トピックに対する各文の貢献度を示すと考えられることから，各成分の値をスコアとし，スコアが大きい順に文をソートし，上位の文を要約文として抽出する．この時，どちらの手法についても，追加のアノテーションデータを必要とせず，教師なしで実行することができる．トピック選択法では，文書のトピックに対する各文の貢献度を用いるため，LSA を用いた既存の抽出型要約手法と同様であるが，トピック距離法は，Sent2Vec の空間における文書のトピックの位置を直接示す特徴ベクトルとの距離を用いているため，LSA を用いた既存の抽出型要約手法とは大きく異なる．

4. 実験準備

4.1. 使用するデータセット

本研究では，DUC2002 と呼ばれるデータセットを実験に用いる．DUC2002 は，自動文書要約の分野で有名なデータセットの一つであり，ニュース記事の文書が収録されている．DUC2002 には，59 個のテーマに関する複数の文書が存在し，各テーマには 10 個ほどの文書が存在している．この時，テーマとしては，単一の自然災害，単一のイベント，ある種類の複数のイベント，単一の個人の経歴に関する内容が記されている．各文書には，100 単語ほどの抽象要約が一つずつ存在し，各テーマには 10 単語，50 単語，100 単語，200 単語ほどの抽象要約と，200 単語，400 単語ほどの抽出要約が複数存在している．これらの要約は，人手によって作成されているため，DUC2002 は，単一文書要約や複数文書要約における，人間が作成する要約を比較対象とし，システムが作成した要約を評価するためのデータセットである．DUC2002 に対する前処理を行い，単一文書要約のデータセットとして利用する際には，1 種類の文書が必ず 1 回だけ出現するようにし，複数文書要約のデータセットとして利用する際には，1 種類のテーマに対し，4 種類の抽象要約と 2 種類の抽出要約が 1 セットだけ存在するように処理を行った．また，複数の文書を一つの文書として結合し，より大きな文書に対する単一文書要約のデータセットとして利用する．これらのことから，本研究で用いる DUC2002 は，他の研究で用いられている完全なデータセットとは少し異なったデータセットとなっている．

4.2. 要約の評価指標

本研究では，作成した要約の評価指標として，自動文書要約の分野で広く利用されている，Rouge と呼ばれる指標[5]を用いる．Rouge はシステムが作成した要約と，人手で作成した要約の間の単語的な一致度を評価する．この Rouge にはいくつかの種類があり，最も有名な Rouge-N では，システムが作成した要約と，人手で作成した要約の間の n-gram における recall を計算する[5]．Rouge-N の recall の計算式は式(5)の通りである[5].

$$\text{Rouge-N} = \frac{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count(gram_n)} \tag{5}$$

この時，$ReferenceSummaries$はシステムが作成した要約の集合を示し，$gram_n$は対象の要約における n-gram の集合を示す．また，$Count(gram_n)$は，n-gram の集合の大きさを示し，$Count_{match}(gram_n)$は，システムが作成した要約と，人手で作成した要約の間で共起する n-gram の数を示す．Rouge-N において，n-gram を unigram としたものを Rouge-1，bigram としたものを Rouge-2 といい，それ以降も n-gram の内容に応じて Rouge-N は計算することができる．Rouge の提案論文[5]では，recall の計算式のみが記載されているが，Rouge を提案した著者が公開している Perl スクリプトでは，Precision，Recall，F 値の 3 種類が計算することができる．この Rouge-1 や Rouge-2 における要約のスコアと，人手で評価した要約のスコアにおいては，ピアソンの積率相関係数によって関係性が確認されており[5]，TextRank[3]や単語の分散表現を用いる LSA の手法[2]などでも利用されていることから，本研究では，Rouge-1 の Precision，Recall，F 値の 3 種類を，要約を評価するための指標として用いる．

4.3. 実験で比較する手法

本研究の実験で比較する抽出型要約の手法は以下の通りである．これらの手法は，単一文書要約を想定しており，教師なしで実行できるため，追加のアノテーションデータを必要としないモデルである．

[Text-Document]：Bannani-Smire らの手法[6]を応用した手法である．Bannani-Smire らの手法では，文書の Sent2Vec のベクトル表現と各フレーズの Sent2Vec のベクトル表現との距離によってキーフレーズを抽出していたが，この手法では，対象をフレーズではなく，文に対して拡張す

る．トピック距離法と手法的に類似しているが，特徴ベクトルを生成する手法ではないため，サブトピックという概念が存在しない．

[TF-IDF]：LSA を用いた抽出型要約の手法であり，TF-IDF によって文書を表現した行列を用いる，基本的な手法である．特異値分解の結果から重要文を選択する手法については，最も中心的なトピックのベクトルのみを用いる．また，成分法を用いることで，逆ベクトルかを判定し，処理をしている．トピック選択法と類似しているが，文書を表現した行列が異なる．

[AI-Sabahi らの手法] [2]：LSA を用いた抽出型要約の手法であり，単語の分散表現によって文書を表現した行列を用いる手法である．特異値分解の結果から重要文を選択する手法については，AI-Sabahi らの手法通り，全てのトピックのベクトルを用いる．トピック選択法と類似しているが，文書を表現した行列が異なる．

[TextRank] [3]：グラフ構造を用いた抽出型要約の手法であり，2 つの文に共通する単語に基づいて，2 つの文の間の類似性を表現し，PageRank アルゴリズムで重要な文を決定している．グラフ構造を用いるため，本研究の抽出型要約の手法とは大きく異なる．

[TextRank+BM25] [4]：グラフ構造を用いた抽出型要約の手法であり，TextRank[3]の 2 つの文の間の類似性に BM25 と呼ばれる指標を用いている．グラフ構造を用いるため，本研究の抽出型要約の手法とは大きく異なる．

[Lead-3]：DUC2002 の作成者によって，ベースラインとして提案されている手法であり，文書の最初の数文を取り出し，要約文とする．複数文書要約の場合には，最後の文書の最初の数文を取り出す．

[Coverage-Lead-3]：DUC2002 の作成者によって，複数文書要約のベースラインとして提案されている手法であり，各文書の最初の数文を取り出し，要約文とする．

5. 実験結果

5.1. Sent2Vec の意味的表現力

Sent2Vec の文のベクトル表現は，その文の意味的な特徴を示すため，同じような意味を示す文の距離は小さく，異なる意味を持つ文との距離は大きくなるはずである．したがって，DUC2002 のデータセットでは，各文書がテーマによって分類されているため，同じテーマの文書や文との距離は小さく，異なるテーマの文書や文との距離は小さくなるはずである．ゆえに，各テーマの全ての文書に対し，同じテーマの全ての文書との平均距離と標準偏差，異なるテーマの全ての文書との平均距離と標準偏差を計算した．また，文についても同様に，それぞれ平均距離と標準偏差を計算した．この時，距離関数にはマンハッタン距離を用いた．

表 1. DUC2002 におけるテーマ間の距離.

	文書		文	
	平均	標準偏差	平均	標準偏差
同じテーマ	9.41	3.83	24.00	2.83
異なるテーマ	16.69	2.00	25.51	1.70

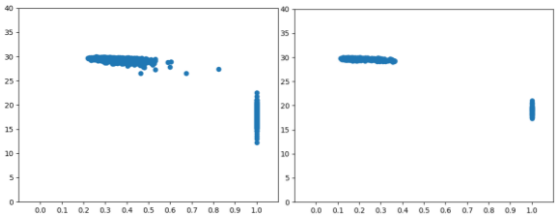
以上の距離比較により，同じテーマの場合，文書と文の両方で距離が近くなることが分かった．すなわち，同じような内容の文や文書の場合，Sent2Vec のベクトル表現は，距離が近くなるということである．文書単位で同じテーマの場合には，文書の内容が大体同じとなるため，距離が全体的に近くなるが，細かい内容が異なるため，標準偏差は一定以上あると考えられる．一方，異なるテーマの場合には，文書の内容がそもそも異なるため，距離が全体的に遠くなるが，距離が遠いという点では同じなので，標準偏差は小さくなると考えられる．文単位では，文書よりも内容が細かくなり，異なる内容が増えるため，同じテーマの場合にも距離が全体的に遠くなり，異なるテーマの場合にはより距離が遠くなると考えられる．よって，本研究で用いる Sent2Vec は，文や文書の特徴や意味的内容を一定以上示していることが分かった．

5.2. 特異値と平均距離の関係

特異値σ_iの大きさと，それに対応するトピックベクトル$\mathbf{v}_i$の特徴について分析する．トピックベクトルと文書の各文との距離を計算し，平均や標準偏差を計算する．これにより，文書におけるトピックの重要度と，それに対応する特徴ベクトルの特徴を分析することができる．

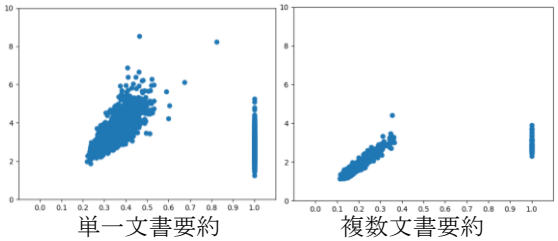
本実験では，距離を計算した後，2 種類のグラフを作成した．どちらのグラフでも，x 軸が特異値の大きさを示し，一つ目のグラフでは，y 軸はトピックベクトルと文書の各文との距離の平均を示し，二つ目のグラフでは，y 軸はその距離の標準偏差を示す．この時，単一文書要約のデータと複数文書要約のデータのそれぞれで実験を行い，それぞれの全ての結果を一つのグラフに表示している．その際，文書やテーマによって，特異値の最大値や傾向は異なるため，文書やテーマの特異値の最大値で各特異値を割ることで，文書やテーマのそれぞれで特異値を正規化している．すなわち，各グラフの x 軸は，中心トピックの重要度に対する比率を表している．また，全ての特異値をグラフにプロットすると，特徴が分かりにくいため，単一文書要約のデータでは 5 番目までの特異値やトピックベクトルで実験し，複数文書要約では，20 番目の特異値やトピックベクトルで実験を行う．

図 3 では，特異値が小さく，平均距離が大きいデータと，特異値が大きく，平均距離が小さいデータの 2 種類に分かれている．また，図 4 においても，特異値が大きいデータと，特異値が一定以下のデータの 2 種類に分かれている．この時，文書やテーマのそれぞれで特異値を正規化していることから，特異値が一番大きいデータは中心トピックベクトルである．したがって，中心トピックベクトルは，文書の各文と平均して距離が近いが，その標準偏差も一定以上あることが分かる．これは，中心トピックベクトルの周りにはより多くの文が存在し，その中心的な位置を中心トピックベクトルが示していると考えられる．一方で，中心トピックベクトル以外では，文書の各文と平均して距離が遠く，特異値が大きいほど平均距離が少しだけ近くなるものの，特異値の大きさはほとんど平均距離に影響していないことが分かる．しかし，特異値が大きいほど，距離の標準偏差は大きくなる



単一文書要約　　複数文書要約

図 3. 特異値の大きさと平均距離の関係.



単一文書要約　　複数文書要約

図 4. 特異値の大きさと距離の標準偏差の関係.

傾向にある．すなわち，より多くの文が周りに存在する中心トピックベクトルとは，少し離れた場所にそれ以外のトピックベクトルが存在し，その周りにもいくつかの文が存在すると考えられる．よって，中心トピックベクトルは，文書の中心的なトピックを示し，それ以外のトピックベクトルは文書の局部的なトピックを示しており，トピックの規模を特異値が示していることが分かった．

5.3. 単一文書要約

DUC2002 を用いて，抽出型要約によって 100 単語の単一文書要約を行った結果を表 2 に示す．この時，Sent2Vec は 700 次元の事前学習モデルを用い，逆ベクトルの判定には成分法を用いた．単一文書要約の実験では，Lead-3 が最も性能が高くなった．これは，DUC2002 がニュース記事に関する文書であったことで，最初の数文に文書の内容をまとめた文が存在し，それによって Lead-3 の性能が高くなった考えられる．一方で，Lead-3 を除いた場合には，TextRank+BM25 の性能が高くなった．それに続き，トピック選択法，トピック距離法，Text-Document の性能が高く，その性能はほとんど同じであるとわかった．このことから，単一文書要約においては，TextRank+BM25 には多少劣るものの，提案手法であるトピック距離法やトピック選択法の性能が示された．

表 2. 抽出型要約による単一文書要約の結果.

	Rouge-1		
	Recall	Precision	F-Value
トピック距離法	44.90	45.80	45.30
Text-Document	44.89	45.72	45.27
トピック選択法	44.90	45.83	45.31
TF-IDF	41.87	44.17	42.81
AI-Sabahi らの手法[2]	43.12	43.98	43.52
TextRank[3]	44.37	45.10	44.62
TextRank+BM25[4]	45.38	46.06	45.60
Lead-3	45.06	47.91	46.23

表 3. 抽出型要約による複数文書要約の結果.

	Rouge-1		
	Recall	Precision	F-Value
トピック距離法	38.71	39.18	38.93
Text-Document	38.83	39.29	39.05
トピック選択法	38.72	39.19	38.95
TF-IDF	38.06	38.52	38.28
AI-Sabahi らの手法[2]	37.03	37.49	37.25
TextRank[3]	39.88	40.11	39.97
TextRank+BM25[4]	40.54	40.36	40.41
Lead-3	34.98	35.65	35.29
Coverage-Lead-3	39.12	41.64	40.10

5.4. 複数文書要約

DUC2002 を用いて，抽出型要約によって 200 単語の複数文書要約を行った結果を表 3 に示す． Sent2Vec の設定や逆ベクトルの判定法は単一文書要約と同様である．複数文書要約の実験では， Coverage-Lead-3 の性能が高くなった．これについても，DUC2002 がニュース記事に関する文書であることが関係していると考える．Coverage-Lead-3 を除いた場合には，TextRank+BM25 や TextRank の性能が高くなった．それに続き，トピック選択法，トピック距離法，Text-Document の性能が高く，その性能はほとんど同じことが分かった．よって，複数文書要約においても，TextRank+BM25 や TextRank には劣るが，提案手法のトピック距離法やトピック選択法の性能が示された．

6. 抽出型要約に関する考察

本研究で提案するトピック距離法とトピック選択法は，単一文書要約と複数文書要約の両方において，性能がほとんど同じであった．トピック選択法では，文書のトピックに対する各文の貢献度を示す特徴ベクトル$\mathbf{u}_i$を用いているが，トピック距離法では，Sent2Vec の空間における文書のトピックの位置を示す中心トピックベクトル$\mathbf{v}_i$を用いている．どちらの特徴ベクトルについても，同じ行列表現Dを特異値分解することで生成されるため，示している内容は同じである．よって，中心トピックベクトルとの距離を用いる抽出型要約の手法は，LSA を用いる従来の手法と同等の性能を持つといえる．また，そのトピック選択法は，単一文書要約と複数文書要約の両方において，TF-IDF や AI-Sabahi らの手法[2]よりも性能が高くなった．これらの 3 つの手法は，いずれも LSA を用いた従来の抽出型要約の手法である．TF-IDF は単語の出現頻度などによって行列表現を構成し，AI-Sabahi らの手法では，単語の分散表現を用いて行列表現を構成している．一方で，トピック選択法では，単語や文の特徴を学習した行列表現を構成している．文の特徴も利用できていることで，行列表現Dの性能が向上し，文書のトピックに対する各文の貢献度を示す特徴ベクトルの性能も向上したと考えられる．これらのことから，本研究で生成した中心トピックベクトルは，Sent2Vec の空間上で文書の意味的な内容を十分表現できており，それとの距離を用いるトピック距離法についても，LSA を用いた既存の抽出型要約よりも性能が高いといえる．

また，特定のベクトルとの距離を用いる手法であるトピック距離法と Text-Document については，多少の差はあるものの，ほとんど同じ性能となった．Sent2Vec では，単語や単語の n-gram の特徴ベクトルの平均ベクトルによって文のベクトル表現を生成する．よって，文書の Sent2Vec のベクトル表現では，より多く存在する中心的な特徴が強く表現される．一方で，トピック距離法で用いる中心トピックベクトルでは，特異値分解の性質上，文書の中心的な内容を表したものである．したがって，両手法の性能がほとんど同じになったと考えられる．しかし，本手法では，文書の局部的な内容を示すトピックベクトルも生成できるため，それらを用いることで，トピック距離法をより強化できると考えられる．

さらに，グラフ構造を用いる TextRank+BM25 については，提案手法であるトピック距離法やトピック選択法よりも性能が高くなった．TextRank では，トピックという区切りがないため，PageRank アルゴリズムで計算される重要度では，中心トピックの内容だけでなく，サブトピックの内容も加味できているため，それによって抽出型要約の性能が高いと考えられる．一方で，本手法では，中心トピックやサブトピックを示すトピックベクトルがそれぞれ存在しているが，トピック距離法やトピック選択法では中心トピックベクトルのみを用いている．なので，サブトピックの内容も加味する必要があると考える．

実際に，抽出型要約によって抽出された文を表 4 に示す．1988 年に発生したハリケーンのギルバートに関するテーマを例とすると，トピック距離法では，表 4 の(1)のように，ハリケーンのギルバートの全体的な情報であり，主に風に関する文が抽出されていた．トピック距離法では，同じトピックの内容を示した特徴ベクトルを用いていることから，風という同じ内容の情報が抽出されたと考えられる，また，トピック選択法でもほとんど同じ文が抽出されている．さらに，TF-IDF では，表 4 の(2)のように，hurricane や inches などの特定の単語を含む文を抽出していた．また，TextRank+BM25 では，表 4 の(3)のように，ハリケーンのギルバートの全体的な情報であり，風や雨などの様々な内容を示す文がそれぞれ抽出されていた．これらのことから，本研究で生成したトピックベクトルは，従来の特徴ベクトルよりもトピックの内容を表現できているが，抽出型要約としての性能を向上させるためには，サブトピックの情報を活用する必要があると考える．実際に，サブトピックを活用する手法として，トピックの重要度を重みとし，中心トピックベクトルと，その次のサブトピックベクトルを平均したベクトルとの距離によって文を抽出する手法を考える．ハリケーンのギルバートの場合には，次のサブトピックベクトルのみとの距離を用いた場合には表 4 の(4)のように，人の被害に関する文を抽出されるが，トピックベクトルを平均したベクトルとの距離を用いることで，表 4 の(5)のように，ハリケーンの情報と人の被害に関する情報の両方に関係する文が抽出できる．このことからも，本研究で生成したトピックベクトルは， Sent2Vec の空間上で各トピックの内容をはっきりと示していると考えられる．

表 4. 抽出型要約の実験で抽出された文の例.

(1)	hurricane gilbert which has left nearly one in four jamaicans homeless slackened somewhat as it swirled over land but the storm was beginning to gain strength over open water as it moved toward the u.s. gulf coast with sustained winds of 120 mph
(2)	the hurricane center said gilbert at one point was the most intense storm on record in terms of barometric pressure which was measured tuesday at 26.13 inches breaking the 26.35 inches recorded for the 1935 hurricane that devastated the florida keys
(3)	hurricane gilbert swept toward jamaica yesterday with 100 mile an hour winds and officials issued warnings to residents on the southern coasts of the dominican republic haiti and cuba
(4)	mexican officials reported six deaths including two babies who drowned and at least seven people were injured but authorities were unable to reach many isolated villages
(5)	the cuban news agency prensa latina said 40000 people many of them foreign vacationers and students were evacuated tuesday from the isle of youth off the southwestern coast as the hurricane passed 200 miles to the south

7. むすび

本研究では，文書の中心トピックとサブトピックのそれぞれに対して，Sent2Vec の空間上での各トピックの位置を示す特徴ベクトルを生成する手法を提案した．また，その特徴ベクトルとの距離を用いることで，文書の要約文を抽出する手法を提案した．実験により，本研究の特徴ベクトルは，中心トピックやサブトピックの内容を十分に反映しており，それによって抽出される要約文は文書の全体的な情報を示していることが分かった．今後は，サブトピックや他の文書のトピックベクトルの内容を加味した抽出型要約の手法や，トピックベクトルの他のタスクへの転用などを検討する必要があると考える．

文　献

[1] M.Pagliardini, P.Gupta, M.Jaggi, “Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features”, NAACL 2018

[2] K.Al-Sabahi, Z.Zuping and Y.Kang, “Latent Semantic Analysis Approach for Document Summarization Based on Word Embeddings”, KSII Transactions on Internet and Information Systems, vol. 13, no. 1, pp. 254-276, 2019.

[3] R.Mihalcea, P.Tarau, “Textrank: Bringing order into text”, Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[4] F.Barrios, et al. “Variations of the similarity function of textrank for automated summarization”, arXiv preprint arXiv:1602.03606, 2016.

[5] C.Y.Lin, “Rouge: A package for automatic evaluation of summaries.”, Text summarization branches out. pp. 74-81, 2004.

[6] K.Bennani-Smires, et al. “Simple unsupervised keyphrase extraction using sentence embeddings”. arXiv preprint arXiv:1801.04470, 2018.