2021年度（令和3年度）　卒業論文

単語ネットワークのリンクへの文字列付与における重心の利用

電気情報系学科 卒業論文検印	
学科長	

指導教員

村田真樹
村上仁一

鳥取大学工学部 電気情報系学科
自然言語処理研究室
B18T2003H 麻田 善次郎

概要

近年，電子テキストは増加し，大量の電子テキストから有用な情報を効率的に取り出す技術が求められている．大竹ら [1] は，言語テキスト処理技術を用いキーワードとなる単語を入力することで，電子テキストや新聞データ等のメディアから入力単語の概念にかかわる概要情報を抜き出し単語ネットワークを構築した．さらに窪 [2] の研究では，単語ネットワークのリンクにノード間の関係性を示す文字列を付与した．なお，この文字列は区切り方によって文や単語列になる．

しかし，関係性を示す文字列として，ノードの単語の間の文字列の内，出現頻度が高いものを付与しており，関係性をわかりやすくするには余分なものや不十分なものが付与されることがあった．

そこで本研究では，リンクに付与する文字列を選定する際，出現頻度の代わりに BERT や word2vec を用い，文字列の重心を利用する．そのようにすることで過不足ない要約を付与するように改良する．本研究の目的は，リンクに付与する文字列を選定する際，過不足ない要約を付与するようにし，単語ネットワークの利便性を向上させることである．

実際に「トヨタ」「宇宙」「ギリシャ」に関するネットワークを構築し，そのネットワークのノード間に付与する文字列を，重心を用いて選定した．選定した文字列を MRR と 1 位正解率と 5 位正解率を用いて，要約として適切なものであるかの評価を行った．

また，提案手法で得られた出力結果に対して，MRR を用いた評価，1 位正解率を用いた評価，5 位正解率を用いた評価を行い，その評価結果を従来手法の評価結果と比較した．5 位正解率を用いた評価において，2 単語の関係を示すものとして適切であるが余分な部分がある場合も正解とする基準で，従来手法が 68 %に対し，提案手法では 67 %の性能を得た．また，1 位正解率を用いた評価方法において，2 単語の関係を示すものとして適切な場合を正解とする基準で，従来手法では 17 %に対し，提案手法では 27 %の性能を得た．2 単語の関係を示すものとして適切であるが余分な部分がある場合も正解とする基準では，従来手法が 32 %に対し，提案手法では 40 %の性能を得た．提案手法は 5 位正解率を用いた評価においては従来手法と同等の性能であったが，1 位正解率を用いた評価においては性能の向上を確認できた．

目 次

表 目 次

図 目 次

第1章　はじめに

近年，インターネット上で様々な電子テキストが増加し，これらの電子テキストから有益な情報を取り出す技術が望まれている．大竹ら [1] は，電子テキストから特定のキーワードに基づく関係情報をネットワークとして抽出する方法を提案し,「地震」というキーワードに基づいて単語ネットワークの構築を行った．Doen ら [3] は，大竹らが構築したネットワークに関連のない事物のノードを含むことを確認し，それらのノードを削除を行った．窪 [2] は，大竹らと Doen らが構築したネットワークは，ノード同士の関係を示す情報がなく，関係性が分かりづらいという問題を確認し，ネットワークのリンクにノード同士の関係性を示す文字列の付与を行った．しかし，窪が付与した文字列は，ノードの単語の間の文字列の内，出現頻度が高いものを付与しており，関係性をわかりやすくするには余分なものや不十分なものが付与されることがあるという問題があった．

そこで本研究では，リンクに付与する文字列を選定する際，出現頻度の代わりに BERT や word2vec を用い，文字列の重心を利用する．そのようにすることで過不足ない要約を付与するように改良する．本研究の目的は，リンクに付与する文字列を選定する際，過不足ない要約を付与するようにし，単語ネットワークの利便性を向上させることである．

本研究の主張点を以下に示す．

- 単語ネットワークのノード同士の関係を示す文字列の付与において，文字列の選定に出現頻度ではなく重心を利用することで，過不足ない要約を付与するように改良する．

- 5 位正解率を用いた評価において，2 単語の関係を示すものとして適切であるが余分な部分がある場合も正解とする基準で，従来手法が 68 %に対し，提案手法では 67 %の性能を得た．また，1 位正解率を用いた評価方法において，2 単語の関係を示すものとして適切な場合を正解とする基準で，従来手法では 17 %に対し，提案手法では 27 %の性能を得た．2 単語の関係を示すものとして適切であるが余分な部分がある場合も正解とする基準では，従来手法が 32 %に対し，提案手法では 40 %の性能を得た．提案手法は 5 位正解率を用いた評価においては従来手法と同等の性能であったが，1 位正解率を用いた評価においては性能の向上を確認できた．

本論文の構成は以下の通りである．第 2 章では，本研究の関連研究を述べ，第 3 章では，ネットワーク構築とリンクへの文字列付与の流れについて述べる．第 4 章では，提案手法について説明する．第 5 章では，実験条件と実験結果や評価方法と評価結果を述べる．第 6 章では，結果の考察と今後の課題を述べる．第 7 章では，本論文のまとめを述べる．

第2章　関連研究

2.1　単語間の文字列を利用した関連研究

本研究は，単語間の文字列を用いて，単語間の関係情報を抽出しており，単語間の文字列を利用した研究に関連している．単語間の文字列を利用した関連研究を以下に示す．

村田ら [4] は，見出し語と辞書定義文を照合することにより，複合語の構成要素とその構成要素間の関係を示す表現を抽出した．例として,「アマチュア無線」という見出し語の定義文は,「アマチュアによる無線」となっていることから,「アマチュア」と「無線」の 2 つの構成要素と,「による」という構成要素間の関係を示す表現を抽出している．

村田ら [5] は，入力した単語対の間の文字列を利用して，入力した単語対の類似の単語対を自動抽出した．さらに，ユーザが入力した単語と同じ分野の用語を収集して可視化するシステムを開発した．例として,「赤色」と入力した場合,「朱色」や「紅色」といった類似した表現の用語を抽出している．

岡田ら [6] は，新聞記事群の文字列の出現頻度を用いて，テキストの分割単位となる文字列の自動取得を行った．例として,「という」「について」等のテキストの分割単位となる文字列を取得している．取得した文字列を用いて，テキストの分割を行うことで，通常の単語分割では細分化されてしまう複合名詞などを取得している．

2.2　要約の関連研究

本研究は，大量のテキストデータから単語間の関係を示す文字列を抽出している．この文字列は，大量のテキストデータからの要約とみなすことができるため，本研究は，要約の研究に関連する．要約の関連研究を以下に示す．

瀧川ら [7] は，入力文から名詞を抽出し，抽出した各名詞から名詞の共起語を取得している．取得した共起語を連想知識として用いることで，端的な要約を生成する手法を提案した．例として,「良い企業に内定をもらうために面接の練習を毎日行う」という入力文からは,「就職活動」という端的な要約を得ることができている．

西川ら [8] は，複数の文書から要約を作成する複数文書要約を，冗長性制約付きナップサック問題として捉えた．この問題に対し，ナップサック問題に基づく要約モデルに，冗長性を削減するための制約を加えることで，複数文書要約モデルを得ている．

森ら [9] は，複数の質問の答とその背景知識を一度に概観できる要約を生成する手法を提案している．複数の質問文を入力し，質問応答エンジンと語の出現分布を用いて，文の重要度の計算を行った．その結果，複数の質問文の答を含む要約文書を抽出している．

Liu[10] は，単一文書の要約の手法として抽出型の要約に着目した．抽出型要約のタスクに BERT[11] を使用するため，BERTSUM などの様々なモデルを設計し，従来の手法と比較を行った．

Rossiello ら [12] はテキストの要約抽出の手法としてセントロイドベースの抽出法を提案した．この手法では，文書の要約を行う際，その文書中の意味のある単語を選択し，出現頻度を基に順位付けを行う．順位が上位の単語のベクトルの合計として重心を求め，重心に近い文を文書中から選択し，要約とした．

2.3 関係情報を表すネットワークの関連研究

本研究は，単語の関係性を示すネットワークを構築しており，以下の研究に関連する．

松尾ら [13][14] は，Web 上の情報から，人間関係のネットワークを抽出した．その際に，抽出手法として，氏名の関係性の強さを知るために様々な指標を用いて実験を行った．

第3章　先行手法

3.1　ネットワーク構築の概要

新聞記事群のデータ（本論文では，新聞データと呼ぶ）から単語ネットワークを構築する．ネットワークの構築の手法は，大竹ら [1] の手法，テーマ限定抽出法，テーマ無関連削除法 [3] の 3 つの手法があるが，本研究ではテーマ限定抽出法を用いてネットワークの構築を行う．ネットワーク構築の流れを図 3.1 に示す．また，本章では，テーマ限定抽出法のみを説明する．

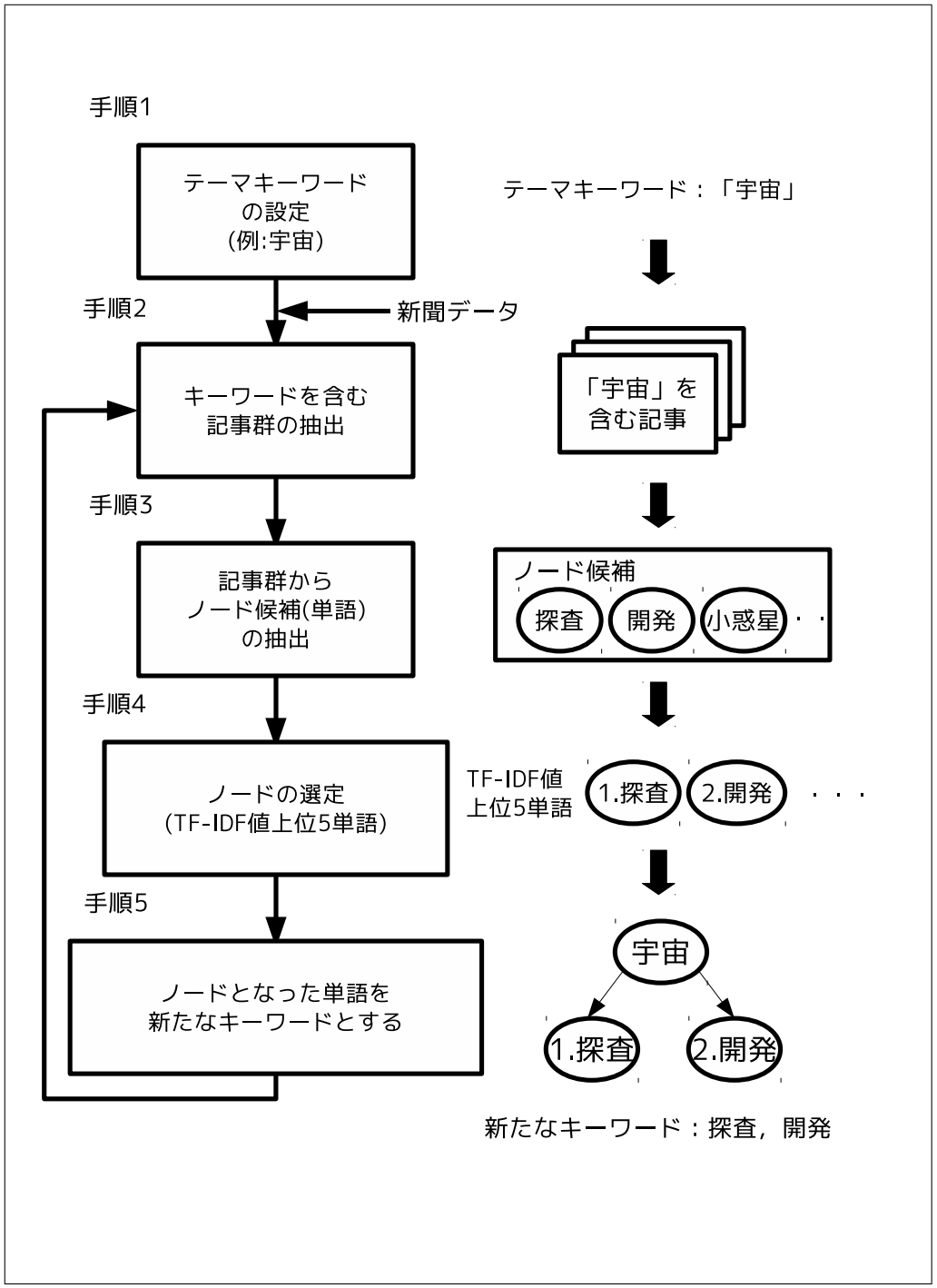

図 3.1: ネットワーク構築の流れ

3.2　テーマキーワードの設定

構築したいネットワークの主となる概念を，テーマキーワードとして設定する．例としては,「トヨタ」「宇宙」「ギリシャ」等の単語である．

3.3　キーワードを含む記事の抽出

新聞データから，キーワードを含む記事の抽出を行う．記事の抽出方法を図 3.2 に示す．

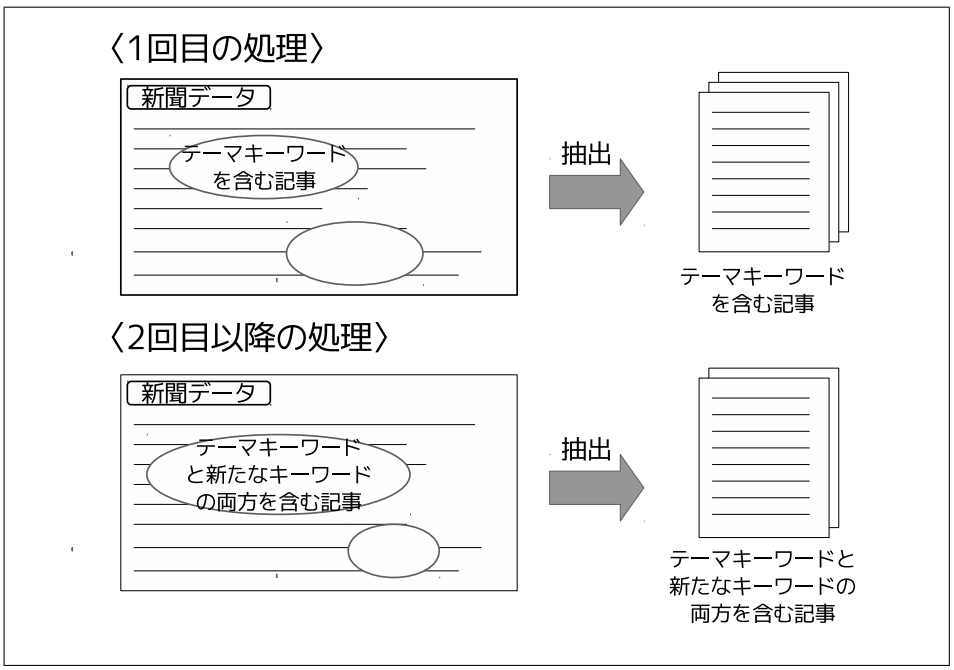

図 3.2: 記事の抽出

1 回目の記事の抽出は，テーマキーワードを含む記事とする．2 回目以降の記事の抽出は，テーマキーワードと，次のノードとなった新たなキーワードの，2 つのキーワードを含む記事とする．

3.4　記事の形態素解析

抽出された記事に対して，形態素解析を用いて，名詞を取り出す．

形態素解析とは，テキストを形態素と呼ばれる単位に分割することである．形態素は，厳密には単語とは違った分割の単位だが，おおよそ単語と同じようなものになり，品詞の情報を持つものである．形態素解析結果の例を図 3.3 に示す．

入力 :「宇宙飛行士の若田光一さんが国際宇宙ステーションの第３９代船長に就任した」

```
宇宙    ウチュウ        宇宙    名詞-一般
飛行    ヒコウ  飛行    名詞-サ変接続
士      シ      士      名詞-接尾-一般
の      ノ      の      助詞-連体化
若田    ワカタ  若田    名詞-固有名詞-人名-姓
光一    コウイチ        光一    名詞-固有名詞-人名-名
さん    サン    さん    名詞-接尾-人名
が      ガ      が      助詞-格助詞-一般
国際    コクサイ        国際    名詞-一般
宇宙    ウチュウ        宇宙    名詞-一般
ステーション    ステーション    ステーション    名詞-一般
の      ノ      の      助詞-連体化
第      ダイ    第      接頭詞-数接続
３      サン    ３      名詞-数
９      キュウ  ９      名詞-数
代      ダイ    代      名詞-接尾-助数詞
船長    センチョウ      船長    名詞-一般
に      ニ      に      助詞-格助詞-一般
就任    シュウニン      就任    名詞-サ変接続
し      シ      する    動詞-自立       サ変・スル      連用形
た      タ      た      助動詞  特殊・タ        基本形
EOS
```

図 3.3: 形態素解析の出力例

図 3.3 のように，形態素解析を行うことで，品詞の情報を持った単語に分割する．本研究では，記事の形態素解析に ChaSen を用いる．また，形態素解析を用いて名詞を取り出す際に，一文字，ひらがなのみ，数字のみの単語を除外する．

3.5　ノード候補の抽出

形態素解析を行った後，ノード候補となる単語の抽出を行う．

3.4 節の図 3.3 を例とすると,「宇宙」「飛行」「若田」「光一」「国際」「ステーション」「船長」「就任」といった単語がノード候補として抽出される．

3.6　ノード候補の選定

TF-IDF を用いて，抽出されたノード候補の中から，実際にノードに用いる単語を選定する．TF-IDF 値の上位 5 単語をキーワードと関係性の強い単語とする．

TF-IDF について説明する．TF-IDF は抽出した記事内におけるノード候補となっている単語の重要度を表す．TF-IDF は以下の式 3.1 で算出される．

$$TF\text{-}IDF = tf_t * \log \frac{N}{df_t} \tag{3.1}$$

tf_t はキーワードを含む記事群での単語 t(ノード候補) の出現回数，df_t は全記事での単語 t の出現記事数とし，N は新聞データの全記事数とする．この式からどの記事にも現れるような重要度の低い単語については低い重みを，他の記事にあまり現れないような貴重な単語には高い重みを与えることになる．

3.7　ネットワークの拡大

3.6 節で得た TF-IDF 値の上位 5 単語を，キーワードから繋がる，次のノードとする．次のノードを新たなキーワードとして設定し，3.3 節の 2 回目以降の処理に戻る．3.3 節から本節までの処理を繰り返すことにより，単語ネットワークを構築していく．構築したネットワークの例を図 3.4 に示す．図 3.4 は，テーマキーワードを「宇宙」とし，毎日新聞 2014 年度から構築したネットワークである．

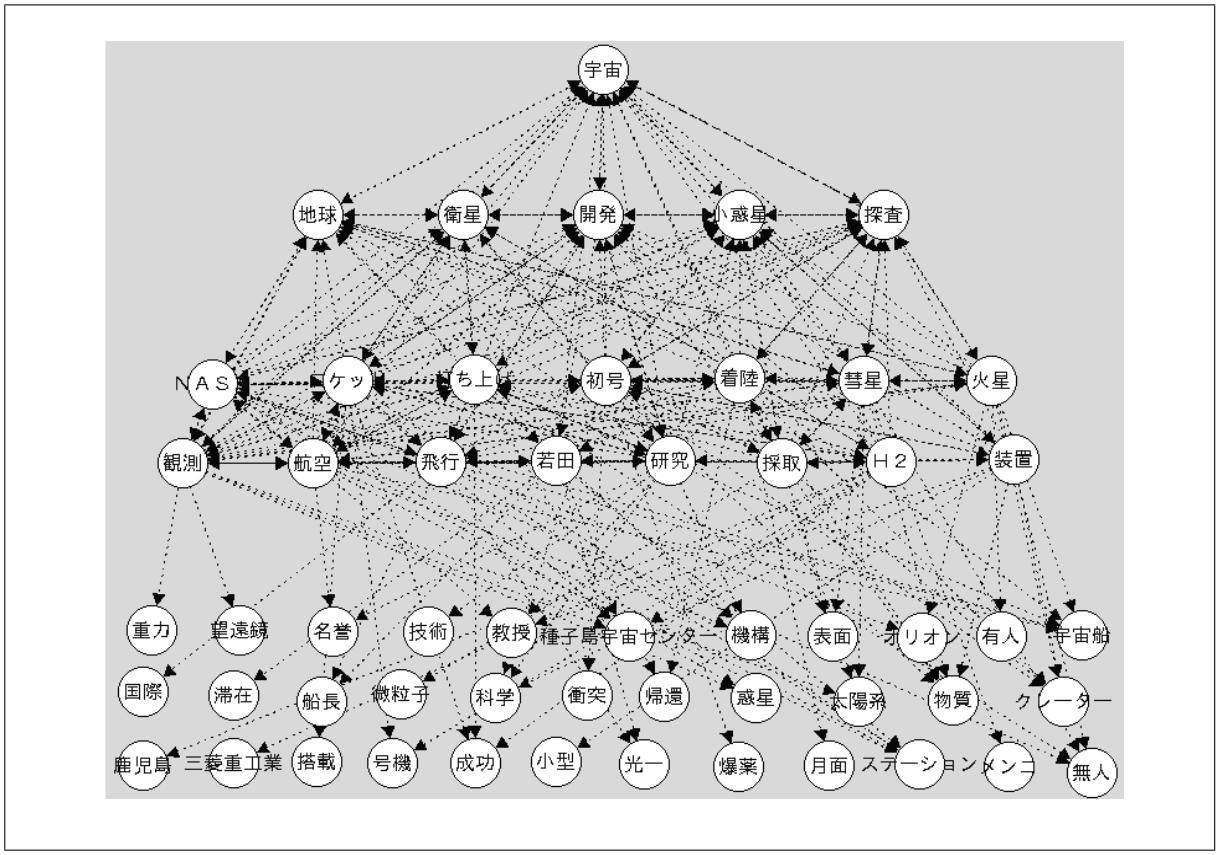

図 3.4: 構築したネットワークの例

3.8　リンクに付与する文字列の選定

単語ネットワークのノード間の関係性を分かりやすくするため，リンクに単語同士の関係性を示す文字列の付与を行う．入力を新聞データと，3.7 節の図 3.4 の「宇宙」「探査」のような単語対データとし，出力をリンクに付与する文字列とする．付与する文字列の選定の手法を以下に示す．図 3.5 は，単語ネットワークのリンクへの文字列付与の例である．

1. 新聞データから，2 単語の間の文字列 (文字列 A と呼ぶ) を抽出する．
2. 2 単語と文字列 A の連接したものを含み，句読点で区切られた文字列 (文字列 B と呼ぶ) を抽出する．文字列 A と文字列 B の抽出例を表 3.1 に示す．

表 3.1: 文字列 A と文字列 B の抽出例

単語対	文字列 A	文字列 B	元の文字列
「ギリシャ」「国債」	の	中国は財政再建に取り組むギリシャの国債を購入し	中国は財政再建に取り組むギリシャの国債を購入し、ユーロ防衛に協力する姿勢を示すなど欧州への影響力を拡大している。
「トヨタ」「水素」	自動車は	トヨタ自動車は水素で動く燃料電池車を２０１４年度に国内で販売と発表	トヨタ自動車は水素で動く燃料電池車を２０１４年度に国内で販売と発表。市販は世界初となる見通し

3. 文字列 B の中で，最も優先度が高い文字列 (出現頻度が高いものや，文字長が短いものを優先度が高い文字列とする．これを文字列 C と呼ぶ) を取得する．これを各文字列 A に対して行う．
4. 3 において取得した文字列 C のうち，優先度が最も高い文字列を選定する．
5. 選定した文字列をリンクに付与する．

優先度の式は以下の 3 つのうちのいずれかを用いる．3.2 式は，文字列の出現頻度が高いものを優先する式であり，3.3 式は，文字列の文字長が短いものを優先する式である．3.4 式は，割り算で優先度を求める式である．以降，式 3.2 を「頻度大」，式 3.3 を「文字長小」，式 3.4 を「割り算」と表記する．

$$\text{優先度} = (\text{頻度} * 10000) - \text{文字長} \tag{3.2}$$

$$\text{優先度} = -(\text{文字長} * 10000) + \text{頻度} \tag{3.3}$$

$$\text{優先度} = \frac{\text{頻度}}{\text{文字長}} \tag{3.4}$$

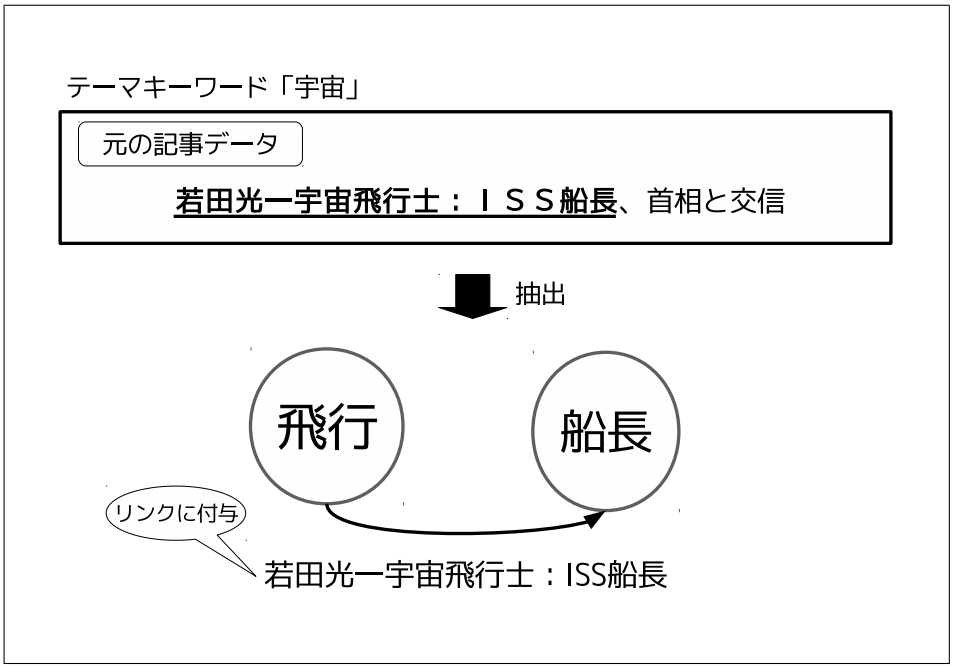

図 3.5: 単語ネットワークのリンクへの文字列付与の例

第4章　提案手法

本章では，本研究の提案手法について説明する．4.1 節では，文字列の重心を算出するために用いた BERT[11] について，4.2 節では， word2vec[15] について記述する．4.3 節では，提案手法について記述する．

4.1　BERT

BERT とは，Bidirectional Encoder Representations from Trans-former の略で，「Trans-former による双方向のエンコード表現」と訳され，2018 年 10 月に Google の Jacob Devlin らの論文 [11] で発表された自然言語処理モデルである．従来の自然言語処理では，大量のラベルのついたデータを用意させ，処理を行うことで課題に取り組む．しかし従来の手法に対し，BERT は事前学習でラベルのないデータをはじめに大量に処理を行う．その後，ファインチューニングで少量のラベルの付いたデータを使用させることで課題に対応させる．事前学習済みの BERT をファインチューニングし，文頭の CLS と呼ばれるトークンのベクトルや，CLS トークン，SEP トークンを除いた各トークンのベクトルの平均を用いることで入力文のベクトルを求めることができる．

4.2　Word2vec

Word2vec は Google 社の Tomas Mikolov ら [15] によって提案されたニューラルネットワーク (Skip-gram) の手法である．Skip-gram は，文脈を利用して与えられた単語と与えられた単語の周辺に出現する単語を予測できるように，単語ベクトルの学習を行うモデルである．意味的に関連が強い単語はベクトルが近くなる．

4.3　リンクに付与する文字列選定の提案手法

本研究では，3.8 節の手順 3 において優先度を求める際，頻度を用いるのではなく文字列をベクトル化し，その重心を利用する．リンク間に付与する文字列の付与手順を以下に示す．また，n 個の記事から文字列 B を抽出し，文ベクトルの重心を求める手順を図示したものを図 4.1 に示す．

手順１ BERT と word2vec の 2 種類の方法を用いて，文字列 B をベクトルで表現する．word2vec を用いる際には，Mecab を用いて文字列 B を単語ごとに分かち書きし，それらの単語を全てベクトルに変換し，その平均をとって文字列 B の文ベクトルを得る．

手順２ ベクトル表現された文字列 B の重心を求める．重心を求める際，同一の文字列を含めて算出する方法と同一の文字列は 1 度しか計算に含めない方法の 2 種類の方法で算出を行う．

手順３ 求めた重心との類似度が高い順に文字列 B の順位付けを行い，最も重心に近い文字列 B を要約として付与する．

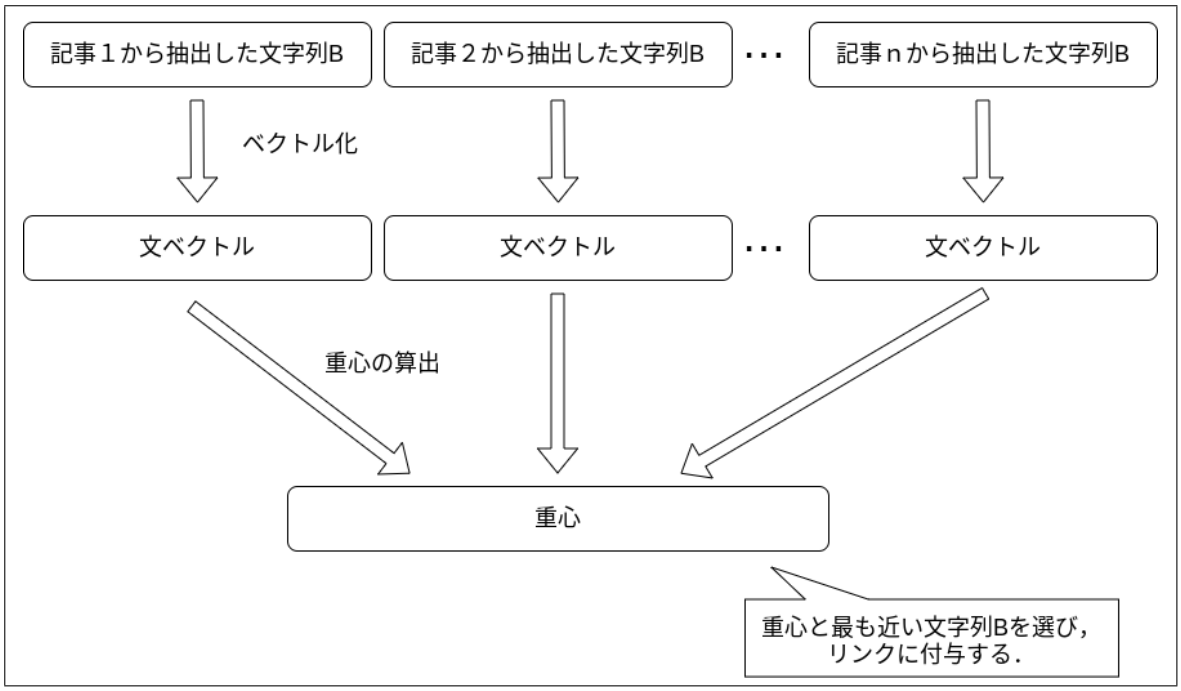

図 4.1: 重心の算出

以降，word2vec を用い，同一の文字列を含めて算出する方法を「word2vec(重複あり)」，word2vec を用い，同一の文字列は 1 度しか計算に含めない方法を「word2vec(重複なし)」，BERT を用い，同一の文字列を含めて算出する方法を「BERT(重複あり)」，BERT を用い，同一の文字列は 1 度しか計算に含めない方法を「BERT(重複なし)」と表記する．

第5章　実験

5.1　実験条件

本実験では，テーマキーワードを「トヨタ」「宇宙」「ギリシャ」の3つとし，ネットワークを構築する.「トヨタ」は191単語対で構成されたネットワーク,「宇宙」は228単語対で構成されたネットワーク,「ギリシャ」は99単語対で構成されたネットワークとなっている．実験データには,「トヨタ」「宇宙」のネットワークを構築する際に，毎日新聞2014年度の1年分の記事102,547記事を用いる.「ギリシャ」のネットワークを構築する際に，毎日新聞2010年度の1年分の記事92,807記事を用いる.「トヨタ」のネットワークを図5.1,「宇宙」のネットワークを図5.2,「ギリシャ」のネットワークを図5.3に示す．本実験で用いるネットワークは，図5.1，図5.2，図5.3のように，4段階層のネットワークとなっている．

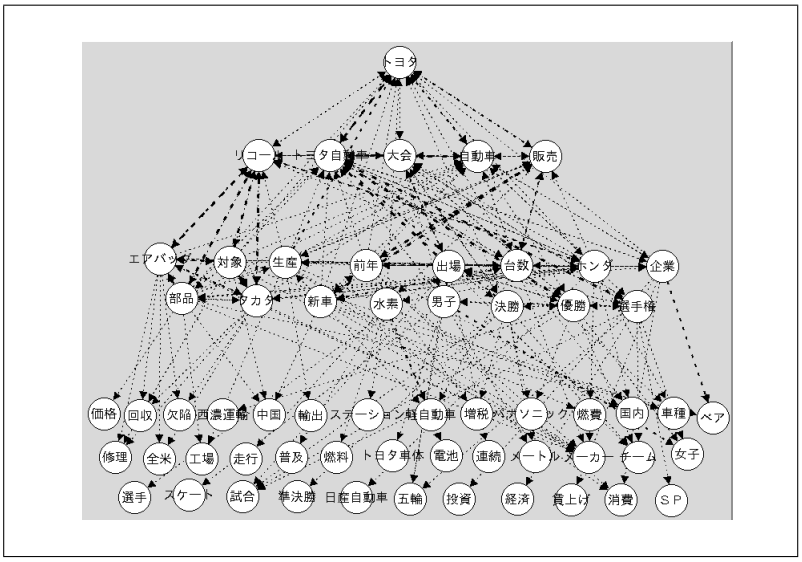

図 5.1: 「トヨタ」のネットワーク図

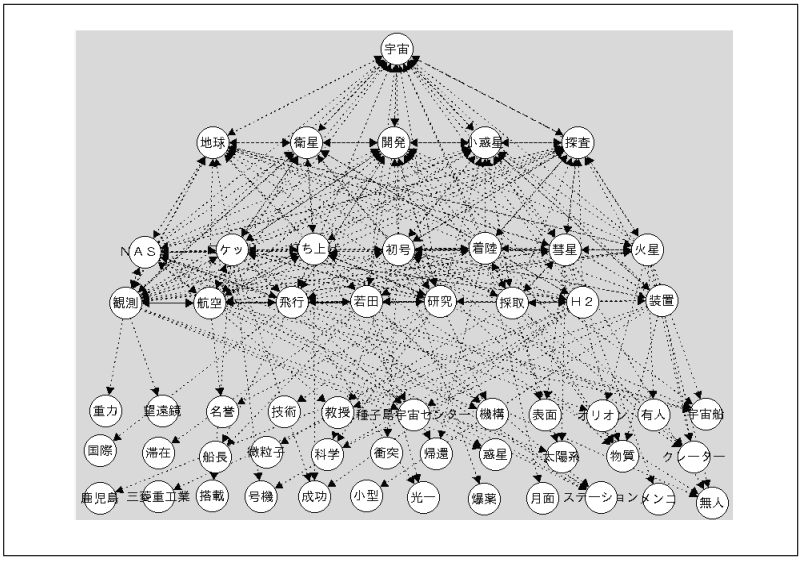

図 5.2: 「宇宙」のネットワーク図

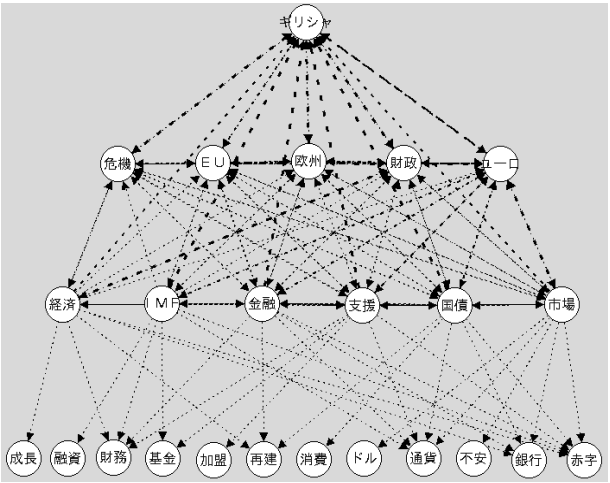

図 5.3: 「ギリシャ」のネットワーク図

5.2　人手による4段階評価の方法

ネットワークのリンクに付与する文字列が適切であるかの評価を行う．

評価データには,「トヨタ」「宇宙」「ギリシャ」のネットワークごとにランダムで取り出した，20 単語対を用いる．すなわち，合計 60 単語対を用いる．また，評価する際の参考データとして，各単語対を含む記事をランダムで 10 記事ずつ抽出したものを用いる．各単語対に対して，第 4 章で述べた 4.3 節の手法の出力結果を重心ベクトルとの類似度が高い順にランク付けした上位 5 つまでの文字列を，抽出した記事を参考に，○□△×の 4 段階の評価を人手で行う．また，重心ベクトルを算出する手法を，word2vec(重複なし) と word2vec(重複あり) と BERT(重複なし) と BERT(重複あり) の 4 パターンの 4 段階評価を行う．単語対を「飛行」「船長」とした場合を例として，○の評価基準と評価例を表 5.1，□の評価基準と評価例を表 5.2，△の評価基準と評価例を表 5.3，×の評価基準と評価例を表 5.4 に示す．

表 5.1: ○の評価基準と評価例

評価基準	2 つの単語間の関係を示すものとして適切である場合
評価例	若田光一宇宙飛行士：ISS 船長

表 5.2: □の評価基準と評価例

評価基準	2 つの単語間の関係を示すものとして適切であるが余分な部分がある場合
評価例	日本人初の船長を務めた若田光一宇宙飛行士 (50) は 14 日午前 7 時 58 分

表 5.3: △の評価基準と評価例

評価基準	2 つの単語間の関係を示すものとして適切であるが，さらに関係を分かりやすくするには，不十分な部分がある場合
評価例	日本人宇宙飛行士の船長が誕生している

表 5.4: ×の評価基準と評価例

評価基準	2 つの単語間の関係を示すものとして不適切である場合
評価例	後任の船長となった米航空宇宙局（ＮＡＳＡ）のスティーブン・スワンソン飛行士は「コウイチのリーダーシップは素晴らしかった」とたたえ

表 5.1 を○と評価した理由は,「若田光一宇宙飛行士が ISS の船長となった」という意味ととれる文字列が,「飛行」「船長」の単語間の関係性を無駄なく適切に表していると判断したからである．表 5.2 を□と評価した理由は，2 単語間の関係性を示すものとしては適切だが,「14 日午前 7 時 58 分」という余分な部分があると判断したからである．表 5.3 を△と評価した理由は，2 単語間の関係性は適切に示しているのだが，さらに関係性を分かりやすくするためには，人名等の情報がなく，不十分な部分があると判断したからである．表 5.4 を×と評価した理由は，参考データから決定した正解の情報とは違ったため，関係を示すものとしては不適切だと判断したからである．

5.3 MRR を用いた評価方法

4 段階評価を行った後，優先度の上位 5 つの出力のうち正解がどの順位に当てはまるかを利用して，MRR(Mean Reciprocal Rank) で評価する．MRR とは，以下の式 5.1 で表される評価値である．

$$MRR = \frac{\sum_{i=1}^{N} 1/r_i}{N} \tag{5.1}$$

N は評価する対象の総数，r_i は評価対象 i がもつ最も高い正解の順位である．今回は，優先度の上位 5 つを出力としているため，$1 \leq r_i \leq 5$ となる．本研究では，○のみを正解とする基準，○と□を正解とする基準，○と□と△を正解とする基準の 3 パターンの基準で評価を行う．

5.4 n位正解率を用いた評価方法

4段階評価を行った後，優先度の上位5つの出力のうち正解がどの順位に当てはまるかを利用して，n位正解率を用いて評価する．n位正解率とは，優先度の上位n個の候補において，1位からn位のいずれかに正解が含まれる場合，1の得点を与え，その合計を問題数で割った値のことである．本研究では，1位正解率と5位正解率を用いる．また，MRRと同様に，○のみを正解とする基準，○と□を正解とする基準，○と□と△を正解とする基準の3パターンの基準で評価を行う．

5.5 実験結果

第4章の提案手法の出力結果の例として，ネットワーク「トヨタ」，単語対「企業」「投資」の出力例を表5.5，ネットワーク「宇宙」，単語対「ロケット」「衛星」の出力例を表5.6，ネットワーク「ギリシャ」，単語対「経済」「ＥＵ」の出力例を表5.7に示す．

表 5.5: ネットワーク「トヨタ」，単語対「企業」「投資」の出力例

手法	出力結果	評価結果
BERT(重複なし)	もっとも政情への不安から投資先の見直しを迫られる企業が増える可能性もある	△
word2vec(重複なし)	大口投資家が投資先企業の監視を強める行動原則「スチュワードシップ・コード」を導入する動きが拡大	○
BERT(重複あり)	企業が先行きへの自信を深めることで投資や雇用・賃金の改善が一段と進み	△
word2vec(重複あり)	企業の投資意欲の強さが浮き彫りになった	×

表 5.6: ネットワーク「宇宙」，単語対「ロケット」「衛星」の出力例

手法	出力結果	評価結果
BERT(重複なし)	政府の情報収集衛星を搭載したH2Aロケット27号機を	△
word2vec(重複なし)	気象衛星「ひまわり8号」を搭載した国産ロケット「H2A」25号機を打ち上げた	○
BERT(重複あり)	政府の情報収集衛星を搭載したH2Aロケット27号機を	△
word2vec(重複あり)	人工衛星やロケットの残骸など	×

表 5.7: ネットワーク「ギリシャ」，単語対「経済」「ＥＵ」の出力例

手法	出力結果	評価結果
BERT(重複なし)	一部欧州紙は最近「ＥＵがスペインに対し経済支援の可能性」と報じたが	×
word2vec(重複なし)	ＥＵ最大の経済国ドイツ抜きにはユーロ防衛は不可能な上	×
BERT(重複あり)	EU 加盟国のうち経済・財政の「優等生」がユーロを導入しているという建前から	×
word2vec(重複あり)	ＥＵ最大の経済国ドイツ抜きにはユーロ防衛は不可能な上	×

5.6 評価結果

本節は，提案手法の評価について述べる．まず,「トヨタ」「宇宙」「ギリシャ」のネットワークから，ランダムに 20 単語対ずつを取り出した．そして，各単語対に対して，提案手法を用い，出力結果である付与する文字列の重心ベクトルを算出し，重心ベクトルに類似度が高い順にランク付けを行った．類似度が高い上位 5 つの文字列を取り出し，MRR を用いた評価，1 位正解率を用いた評価，5 位正解率を用いた評価の 3 つ評価を行った．実験条件は，5.1 節で述べた通りである．評価方法は，5.2 節，5.3 節，5.4 節で述べた方法で行う．

5.6.1 「トヨタ」の評価結果

「トヨタ」の MRR を用いた評価結果を表 5.8 に，1 位正解率を用いた評価結果を表 5.9 に，5 位正解率を用いた評価結果を表 5.10 に示す．

表 5.8: 「トヨタ」の MRR を用いた評価結果

	○	○□	○□△
頻度大	0.20	0.26	0.66
文字長小	0.13	0.25	0.45
割り算	0.23	0.29	0.69
BERT(重複なし)	0.17	0.26	0.70
word2vec(重複なし)	0.24	0.34	0.62
BERT(重複あり)	0.20	0.29	0.77
word2vec(重複あり)	0.25	0.34	0.65

表 5.9: 「トヨタ」の 1 位正解率を用いた評価結果

	○	○□	○□△
頻度大	0.15	0.20	0.60
文字長小	0.10	0.10	0.25
割り算	0.20	0.25	0.65
BERT(重複なし)	0.05	0.15	0.55
word2vec(重複なし)	0.20	0.30	0.50
BERT(重複あり)	0.10	0.15	0.70
word2vec(重複あり)	0.20	0.25	0.55

表 5.10: 「トヨタ」の 5 位正解率を用いた評価結果

	○	○□	○□△
頻度大	0.30	0.40	0.80
文字長小	0.15	0.35	0.75
割り算	0.25	0.35	0.75
BERT(重複なし)	0.45	0.50	0.85
word2vec(重複なし)	0.35	0.45	0.80
BERT(重複あり)	0.30	0.40	0.85
word2vec(重複あり)	0.40	0.45	0.80

5.6.2 「宇宙」の評価結果

「宇宙」の MRR を用いた評価結果を表 5.11 に，1 位正解率を用いた評価結果を表 5.12 に，5 位正解率を用いた評価結果を表 5.13 に示す．

表 5.11: 「宇宙」の MRR を用いた評価結果

	○	○□	○□△
頻度大	0.26	0.61	0.75
文字長小	0.27	0.43	0.59
割り算	0.29	0.60	0.74
BERT(重複なし)	0.24	0.40	0.68
word2vec(重複なし)	0.27	0.50	0.76
BERT(重複あり)	0.27	0.52	0.77
word2vec(重複あり)	0.25	0.58	0.79

表 5.12: 「宇宙」の 1 位正解率を用いた評価結果

	○	○□	○□△
頻度大	0.15	0.50	0.65
文字長小	0.20	0.30	0.45
割り算	0.20	0.50	0.65
BERT(重複なし)	0.15	0.30	0.50
word2vec(重複なし)	0.20	0.40	0.70
BERT(重複あり)	0.20	0.45	0.70
word2vec(重複あり)	0.15	0.50	0.70

表 5.13: 「宇宙」の 5 位正解率を用いた評価結果

	○	○□	○□△
頻度大	0.45	0.80	0.90
文字長小	0.35	0.65	0.80
割り算	0.45	0.80	0.90
BERT(重複なし)	0.40	0.55	0.85
word2vec(重複なし)	0.40	0.65	0.85

5.6.3 「ギリシャ」の評価

「ギリシャ」の MRR を用いた評価結果を表 5.14 に，1 位正解率を用いた評価結果を表 5.15 に，5 位正解率を用いた評価結果を表 5.16 に示す.

表 5.14: 「ギリシャ」の MRR を用いた評価結果

	○	○□	○□△
頻度大	0.43	0.46	0.64
文字長小	0.24	0.26	0.56
割り算	0.33	0.36	0.61
BERT(重複なし)	0.38	0.54	0.77
word2vec(重複なし)	0.56	0.66	0.84
BERT(重複あり)	0.46	0.50	0.65
word2vec(重複あり)	0.39	0.44	0.55

表 5.15: 「ギリシャ」の 1 位正解率を用いた評価結果

	○	○□	○□△
頻度大	0.20	0.25	0.40
文字長小	0.05	0.05	0.35
割り算	0.15	0.15	0.35
BERT(重複なし)	0.25	0.40	0.65
word2vec(重複なし)	0.40	0.50	0.75
BERT(重複あり)	0.30	0.35	0.50
word2vec(重複あり)	0.25	0.30	0.35

表 5.16: 「ギリシャ」の 5 位正解率を用いた評価結果

	○	○□	○□△
頻度大	0.85	0.85	1.00
文字長小	0.60	0.70	0.95
割り算	0.70	0.75	1.00
BERT(重複なし)	0.60	0.75	0.90
word2vec(重複なし)	0.80	0.90	0.95
BERT(重複あり)	0.65	0.75	0.85
word2vec(重複あり)	0.60	0.65	0.90

5.6.4　3つのネットワークを合わせた場合の評価

「トヨタ」「宇宙」「ギリシャ」の 20 単語対ずつをまとめ，60 単語対での評価を行った．MRR を用いた評価結果を表 5.17 に，1 位正解率を用いた評価結果を表 5.18 に，5 位正解率を用いた評価結果を表 5.19 に示す．

表 5.17: 「トヨタ」「宇宙」「ギリシャ」の MRR を用いた評価結果

	○	○□	○□△
頻度大	0.30	0.44	0.68
文字長小	0.21	0.31	0.53
割り算	0.28	0.41	0.68
BERT(重複なし)	0.26	0.40	0.71
word2vec(重複なし)	0.36	0.50	0.74
BERT(重複あり)	0.31	0.44	0.73
word2vec(重複あり)	0.29	0.45	0.67

表 5.18: 「トヨタ」「宇宙」「ギリシャ」の 1 位正解率を用いた評価結果

	○	○□	○□△
頻度大	0.17	0.32	0.55
文字長小	0.12	0.15	0.35
割り算	0.18	0.30	0.55
BERT(重複なし)	0.15	0.28	0.57
word2vec(重複なし)	0.27	0.40	0.65
BERT(重複あり)	0.20	0.31	0.63
word2vec(重複あり)	0.20	0.35	0.53

表 5.19: 「トヨタ」「宇宙」「ギリシャ」の 5 位正解率を用いた評価結果

	○	○□	○□△
頻度大	0.53	0.68	0.90
文字長小	0.37	0.57	0.83
割り算	0.47	0.63	0.88
BERT(重複なし)	0.48	0.60	0.87
word2vec(重複なし)	0.52	0.67	0.87
BERT(重複あり)	0.45	0.60	0.87
word2vec(重複あり)	0.47	0.60	0.88

評価の結果，5 位正解率を用いた評価において，2 単語の関係を示すものとして適切であるが余分な部分がある場合も正解とする基準で，従来手法が 68 %に対し，word2vec（重複なし）では 67 %の性能を得た．また，1 位正解率を用いた評価方法において，2 単語の関係を示すものとして適切な場合を正解とする基準で，従来手法では 17 %に対し，word2vec(重複なし) では 27 %の性能を得た．2 単語の関係を示すものとして適切であるが余分な部分がある場合も正解とする基準では，従来手法が 32 %に対し，word2vec(重複なし) では 40 %の性能を得た．提案手法は 5 位正解率を用いた評価においては従来手法と同等の性能であったが，1 位正解率を用いた評価においては性能の向上を確認できた．

5.6.5 有意差検定

各手法で有意差を調べるために片側検定の t 検定を行った．３つの従来手法と４つの提案手法の性能を，MRR を用いた評価結果，1 位正解率を用いた評価結果，5 位正解率を用いた評価結果でそれぞれ比較した.「トヨタ」,「宇宙」,「ギリシャ」の３つのネットワークを合わせた場合の評価結果の計 60 個のデータを用いた．ここで，有意水準は５％である．MRR を用いた場合の，2 単語の関係を示すものとして適切な場合を正解とする基準での検定結果 (p 値) を表 5.20 に示す．MRR を用いた場合の，適切であるが余分な部分がある場合も正解とする基準での検定結果 (p 値) を表 5.21 に示す．MRR を用いた場合の，2 単語間の関係を示すものとして適切であるが，余分な部分やさらに関係を分かりやすくするには不十分な部分がある場合を正解とする基準での検定結果 (p 値) を表 5.22 に示す． 1 位正解率を用いた場合の，2 単語の関係を示すものとして適切な場合を正解とする基準での検定結果 (p 値) を表 5.23 に示す． 1 位正解率を用いた場合の，適切であるが余分な部分がある場合も正解とする基準での検定結果 (p 値) を表 5.24 に示す． 1 位正解率を用いた場合の，2 単語間の関係を示すものとして適切であるが，余分な部分やさらに関係を分かりやすくするには不十分な部分がある場合を正解とする基準での検定結果 (p 値) を表 5.25 に示す．5 位正解率を用いた場合の，2 単語の関係を示す

ものとして適切な場合を正解とする基準での検定結果 (p 値) を表 5.26 に示す．5 位正解率を用いた場合の，適切であるが余分な部分がある場合も正解とする基準での検定結果 (p 値) を表 5.27 に示す．5 位正解率を用いた場合の，2 単語間の関係を示すものとして適切であるが，余分な部分やさらに関係を分かりやすくするには不十分な部分がある場合を正解とする基準での検定結果 (p 値) を表 5.28 に示す．

表 5.20: 「トヨタ」「宇宙」「ギリシャ」の MRR の片側検定（○の評価基準）

	文字長小	割り算	BERT (重複なし)	word2vec (重複なし)	BERT (重複あり)	word2vec (重複あり)
頻度大	0.005	0.248	0.236	0.039	0.410	0.459
文字長小		0.021	0.114	0.002	0.014	0.025
割り算			0.351	0.042	0.308	0.363
BERT(重複なし)				0.036	0.144	0.285
word2vec(重複なし)					0.146	0.051
BERT(重複あり)						0.395

表 5.21: 「トヨタ」「宇宙」「ギリシャ」の MRR の片側検定（○□の評価基準）

	文字長小	割り算	BERT (重複なし)	word2vec (重複なし)	BERT (重複あり)	word2vec (重複あり)
頻度大	0.004	0.153	0.205	0.087	0.436	0.387
文字長小		0.011	0.060	0.001	0.017	0.005
割り算			0.394	0.044	0.343	0.184
BERT(重複なし)				0.012	0.212	0.179
word2vec(重複なし)					0.078	0.107
BERT(重複あり)						0.348

表 5.22: 「トヨタ」「宇宙」「ギリシャ」の MRR の片側検定（○□△の評価基準）

	文字長小	割り算	BERT (重複なし)	word2vec (重複なし)	BERT (重複あり)	word2vec (重複あり)
頻度大	0.001	0.419	0.288	0.092	0.134	0.219
文字長小		0.000	0.001	0.000	0.000	0.006
割り算			0.269	0.090	0.132	0.343
BERT(重複なし)				0.274	0.324	0.185
word2vec(重複なし)					0.421	0.034
BERT(重複あり)						0.038

表 5.23: 「トヨタ」「宇宙」「ギリシャ」の 1 位正解率の片側検定（○の評価基準）

	文字長小	割り算	BERT (重複なし)	word2vec (重複なし)	BERT (重複あり)	word2vec (重複あり)
頻度大	0.091	0.284	0.392	0.029	0.242	0.080
文字長小		0.022	0.266	0.029	0.048	0.029
割り算			0.299	0.084	0.383	0.329
BERT(重複なし)				0.045	0.185	0.222
word2vec(重複なし)					0.144	0.104
BERT(重複あり)						0.500

表 5.24: 「トヨタ」「宇宙」「ギリシャ」の 1 位正解率の片側検定（○□の評価基準）

	文字長小	割り算	BERT (重複なし)	word2vec (重複なし)	BERT (重複あり)	word2vec (重複あり)
頻度大	0.002	0.329	0.311	0.084	0.500	0.080
文字長小		0.001	0.022	0.000	0.003	0.000
割り算			0.410	0.067	0.392	0.130
BERT(重複なし)				0.035	0.284	0.175
word2vec(重複なし)					0.100	0.205
BERT(重複あり)						0.242

表 5.25: 「トヨタ」「宇宙」「ギリシャ」の 1 位正解率の片側検定（○□△の評価基準）

	文字長小	割り算	BERT (重複なし)	word2vec (重複なし)	BERT (重複あり)	word2vec (重複あり)
頻度大	0.003	0.500	0.410	0.055	0.066	0.284
文字長小		0.002	0.003	0.000	0.000	0.008
割り算			0.415	0.067	0.084	0.371
BERT(重複なし)				0.035	0.104	0.329
word2vec(重複なし)					0.392	0.026
BERT(重複あり)						0.029

表 5.26: 「トヨタ」「宇宙」「ギリシャ」の 5 位正解率の片側検定（○の評価基準）

	文字長小	割り算	BERT (重複なし)	word2vec (重複なし)	BERT (重複あり)	word2vec (重複あり)
頻度大	0.002	0.022	0.205	0.329	0.029	0.104
文字長小		0.041	0.054	0.009	0.083	0.067
割り算			0.405	0.130	0.383	0.500
BERT(重複なし)				0.311	0.242	0.405
word2vec(重複なし)					0.104	0.091
BERT(重複あり)						0.392

表 5.27: 「トヨタ」「宇宙」「ギリシャ」の 5 位正解率の片側検定（○□の評価基準）

	文字長小	割り算	BERT (重複なし)	word2vec (重複なし)	BERT (重複あり)	word2vec (重複あり)
頻度大	0.035	0.091	0.100	0.370	0.066	0.084
文字長小		0.126	0.299	0.067	0.299	0.320
割り算			0.311	0.284	0.299	0.299
BERT(重複なし)				0.104	0.500	0.500
word2vec(重複なし)					0.051	0.051
BERT(重複あり)						0.500

表 5.28: 「トヨタ」「宇宙」「ギリシャ」の 5 位正解率の片側検定（○□△の評価基準）

	文字長小	割り算	BERT (重複なし)	word2vec (重複なし)	BERT (重複あり)	word2vec (重複あり)
頻度大	0.022	0.161	0.209	0.161	0.209	0.329
文字長小		0.042	0.242	0.209	0.242	0.130
割り算			0.354	0.284	0.354	0.500
BERT(重複なし)				0.500	0.500	0.375
word2vec(重複なし)					0.500	0.284
BERT(重複あり)						0.354

検定の結果，word2vec(重複なし) は，MRR を用いた評価方法の場合，2 単語の関係を示すものとして適切な場合を正解とする基準では，頻度大，文字長小，割り算，BERT(重複なし) との間で有意差があった．適切であるが余分な部分がある場合も正解とする基準では，文字長小，割り算，BERT(重複なし) との間で有意差があった．2 単語間の関係を示すものとして適切であるが，余分な部分やさらに関係を分かりやすくするには不十分な部分がある場合を正解とする基準では，文字長小，word2vec(重複あり) との間で有意差があった．1 位正解率を用いた評価方法の場合，2 単語の関係を示すものとして適切な場合を正解とする基準では，頻度大，文字長小，BERT(重複なし) との間で有意差があった．適切であるが余分な部分がある場合も正解とする基準では，文字長小，BERT(重複なし) との間で有意差があった．2 単語間の関係を示すものとして適切であるが，余分な部分やさらに関係を分かりやすくするには不十分な部分がある場合を正解とする基準では，文字長小，BERT(重複なし)，word2vec(重複あり) との間で有意差があった．5 位正解率を用いた評価方法の場合，2 単語の関係を示すものとして適切な場合を正解とする基準では，文字長小，word2vec(重複あり) との間で有意差があった．適切であるが余分な部分がある場合も正解とする基準では，文字長小との間で有意差があった．BERT(重複なし) は，MRR を用いた評価方法の場合，2 単語間の関係を示すものとして適切であるが，余分な部分やさらに関係を分かりやすくするには不十分な部分がある場合を正解とする基準では，文字長小との間で有意差があった．1 位正解率を用いた評価方法の場合，適切であるが余分な部分がある場合も正解とする基準では，文字長小との間で有意差があった．2 単語間の関係を示すものとして適切であるが，余分な部分やさらに関係を分かりやすくするには不十分な部分がある場合を正解とする基準では，文字長小との間で有意差があった．word2vec(重複あり) は，MRR を用いた評価方法の場合，2 単語の関係を示すものとして適切な場合を正解とする基準では，文字長小との間で有意差があった．適切であるが余分な部分がある場合も正解とする基準では，文字長小との間で有意

差があった．2 単語間の関係を示すものとして適切であるが，余分な部分やさらに関係を分かりやすくするには不十分な部分がある場合を正解とする基準では，文字長小との間で有意差があった． 1 位正解率を用いた評価方法の場合，2 単語の関係を示すものとして適切な場合を正解とする基準では，文字長小との間で有意差があった．適切であるが余分な部分がある場合も正解とする基準では，文字長小との間で有意差があった．2 単語間の関係を示すものとして適切であるが，余分な部分やさらに関係を分かりやすくするには不十分な部分がある場合を正解とする基準では，文字長小との間で有意差があった．BERT(重複なし) は，MRR を用いた評価方法の場合，2 単語の関係を示すものとして適切な場合を正解とする基準では，文字長小との間で有意差があった．適切であるが余分な部分がある場合も正解とする基準では，文字長小との間で有意差があった．2 単語間の関係を示すものとして適切であるが，余分な部分やさらに関係を分かりやすくするには不十分な部分がある場合を正解とする基準では，文字長小，word2vec(重複あり) との間で有意差があった． 1 位正解率を用いた評価方法の場合，2 単語の関係を示すものとして適切な場合を正解とする基準では，文字長小との間で有意差があった．適切であるが余分な部分がある場合も正解とする基準では，文字長小との間で有意差があった．2 単語間の関係を示すものとして適切であるが，余分な部分やさらに関係を分かりやすくするには不十分な部分がある場合を正解とする基準では，文字長小，word2vec(重複あり) との間で有意差があった．

第6章　考察

リンクに付与する文字列の選定方法として，従来手法では，出現頻度と文字長と割り算の 3 つの手法があった．重心を用いた手法では word2vec(重複なし) と BERT(重複なし) と word2vec(重複あり) と BERT(重複あり) の 4 つの手法を提案した．word2vec を用いた手法と BERT を用いた手法についてそれぞれ考察を述べる．

6.1　word2vec を用いた手法の考察

word2vec(重複なし) は，MRR を用いた評価方法と 1 位正解率を用いた評価方法において，2 単語の関係を示すものとして適切な場合を正解とする基準と，2 単語間の関係を示すものとして適切であるが，余分な部分がある場合を正解とする基準と，2 単語間の関係を分かりやすくするには不十分な部分がある場合も正解とする基準の 3 つの基準の全ての基準で，3 つの従来手法よりも性能が高かった．5 位正解率を用いた評価方法においては，word2vec(重複なし) は文字長小と割り算より高い性能となった．原因として，頻度や文字長は単語の意味や文の内容といった情報を含んでいないのに対し，word2vec を用いた手法では，文字列をベクトル化することで単語の意味や関係性などの情報を含んでいるためではないかと考えられる．また、word2vec(重複あり) は word2vec(重複なし) と比べて，5 位正解率を用いた評価方法の，2 単語間の関係を示すものとして適切であるが，余分な部分やさらに関係を分かりやすくするには不十分な部分がある場合を正解とする基準を除く全ての基準で性能が低かった．原因として，重心の算出に同一の文字列を含めることで，頻度が大きい文字列の優先度が高くなり，頻度大に近い性能になるためではないかと考えられる．

6.2　BERT を用いた手法の考察

BERT(重複なし) は，MRR を用いた評価方法と 1 位正解率を用いた評価方法において，2 単語間の関係を分かりやすくするには不十分な部分がある場合も正解とする基準では，3 つの従来手法よりも性能が高かったが，2 単語の関係を示すものとして適切な場合を正解とする基準と，2 単語間の関係を示すものとして適切であるが，余分な部分がある場合を正解とする基準では，頻度大と割り算よりも性能が

低かった．また，word2vec(重複なし) と比較すると，全ての評価方法の全ての評価基準でBERT(重複なし) の方が性能が低かった．本研究では，BERT の Fine-tuning を行っておらず，要約抽出に適していないモデルであることが，BERT(重複なし) の性能が低い原因であると考えられる．また、BERT(重複あり) は BERT(重複なし) と比べて，MRR を用いた評価方法と 1 位正解率を用いた評価方法の，全ての基準で性能が高かった．原因として，重心の算出に同一の文字列を含めることで，頻度が大きい文字列の優先度が高くなり，頻度大に近い性能になるためではないかと考えられる．

第7章　おわりに

先行研究では，大竹ら [1] は，電子テキストから特定のキーワードに基づく関係情報をネットワークとして抽出する方法を提案し，「地震」というキーワードに基づいて単語ネットワークの構築を行った．Doen ら [3] は，大竹らが構築したネットワークに関連のない事物のノードを含むことを確認し，それらのノードを削除を行った．窪 [2] は，大竹らと Doen らが構築したネットワークは，ノード同士の関係を示す情報がなく，関係性が分かりづらいという問題を確認し，単語ネットワークのリンクにノード間の関係性を示す文字列を付与した．しかし，関係性を示す文字列として，ノードの単語の間の文字列の内，出現頻度が高いものを付与しており，関係性をわかりやすくするには余分なものや不十分なものが付与されることがあった．そこで本研究では，リンクに付与する文字列を選定する際，出現頻度の代わりに BERT や word2vec を用い，文字列の重心を利用する手法を提案した．また，提案手法で得られた出力結果に対して，MRR を用いた評価，1 位正解率を用いた評価，5 位正解率を用いた評価を行い，その評価結果を従来手法の評価結果と比較した．5 位正解率を用いた評価において，2 単語の関係を示すものとして適切であるが余分な部分がある場合も正解とする基準で，従来手法が 68 %に対し，提案手法では 67 %の性能を得た．また，1 位正解率を用いた評価方法において，2 単語の関係を示すものとして適切な場合を正解とする基準で，従来手法では 17 %に対し，提案手法では 27 %の性能を得た．2 単語の関係を示すものとして適切であるが余分な部分がある場合も正解とする基準では，従来手法が 32 %に対し，提案手法では 40 %の性能を得た．提案手法は 5 位正解率を用いた評価においては従来手法と同等の性能であったが，1 位正解率を用いた評価においては性能の向上を確認できた．

今後は，さらなる性能向上のため，BERT の Fine-tuning を行うなど，提案手法の改良を検討したい．

謝辞

最後に，1 年間の間，研究を進めるに当たり，本研究のご指導を頂きました鳥取大学工学部電気情報系学科自然言語処理研究室の村田真樹教授，村上仁一准教授そして自然言語処理研究室の皆様に深く感謝するとともに心から御礼申し上げます．また，参考にさせていただいた論文の著者の方々に対して深く感謝申し上げます．

参考文献

[1] 大竹竜太, 村田真樹, 徳久雅人. 大規模テキストデータを用いた社会構造ネットワークモデルの自動抽出. 言語処理学会第 19 回年次大会発表論文集, pp. 798–801, 2013.

[2] 窪雄平. テキスト処理に基づく概念ネットワークの構築におけるリンクへの文字列付与. 言語処理学会第 16 回年次大会発表論文集, pp. 119–133, 2016.

[3] Y. Doen, M. Murata, R. Otake, and M. Tokuhisa. Construction of concept network from large numbers of texts for information examination using tf-idf and deletion of unrelated words. *SCIS&ISIS 2014*, pp. 1108–1113, 2014.

[4] 村田真樹, 内山将夫, 井佐原均. 辞書定義文を用いた複合語分割-語構成情報の抽出と考察-. 言語処理学会第 6 回年次大会発表論文集, pp. 411–414, 2000.

[5] 村田真樹, 馬青, 白土保, 井佐原均. 用語抽出用評価データの作成とその利用. 言語処理学会第 10 回年次大会併設ワークショップ「固有表現と専門用語」発表論文集, pp. 9–12, 2004.

[6] 岡田正平, 山本和英. 文字列の出現頻度情報を用いた分かち書き単位の自動取得. 言語処理学会第 19 回年次大会発表論文集, pp. 422–425, 2013.

[7] 瀧川和樹, 村田真樹, 土田正明, De Saeger Stijn, 山本和英, 鳥澤健太郎. 連想知識を用いた端的な要約の生成. 言語処理学会第 16 回年次大会発表論文集, pp. 298–301, 2010.

[8] 西川仁, 平尾努, 牧野俊朗, 松尾義博, 松本裕治. 冗長性制約付きナップサック問題に基づく複数文書要約モデル. 自然言語処理,Vol20,No4, pp. 585–612, 2013.

[9] 森辰則, 野澤正憲, 浅田義昭. 質問応答エンジンを利用した複数文書要約手法. 言語処理学会第 10 回年次大会発表論文集, pp. 289–292, 2004.

[10] Yang Liu. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*, 2019.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pp. 12–21, 2017.

[13] 松尾豊, 友部博教, 橋田浩一, 中島秀之, 石塚満. Web 上の情報からの人間関係ネットワークの抽出. 人口知能学会論文誌, Vol. 20, No. 1, pp. 46–56, 2005.

[14] 松尾豊, 友部博教, 橋田浩一, 石塚満. Web からの人間関係ネットワークの抽出と情報支援. 人口知能学会第 17 回全国大会講演論文集, pp. 1–4, 2003.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, p. 3111–3119, 2013.