概要

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，TF-IDF を用いて，新聞記事群から事物の関係情報を単語ネットワークとしてまとめたものを構築した．Doen ら [2] は，単語ネットワークを構築する際に，事物と無関係であるノードの削除を行った．しかし，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がある．

そこで本研究では，単語ネットワークを構築した後，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う．リンクに関係性を示す文字列を付与することで，構築した単語ネットワークから得られる情報がより詳細なものとなる．本研究の目的は，ノード同士の関係性を分かりやすくすることにより，単語ネットワークの利便性の向上を図ることである．

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文字列を付与することで，ノード同士の関係性への理解が深まるかを調査した．調査の結果，文字列を付与することで，関係が分かりづらいノード同士の関係性を確認した．また，MRR と 1 位正解率と 5 位正解率を用いて，リンクに付与する文字列が適切なものであるかの評価を行った．付与する文字列に，余分な部分や，関係性をさらに分かりやすくするには不十分な部分があっても正解とする基準とした場合，MRR を用いた評価では約 7 割，一位正解率を用いた評価では約 6 割，5 位正解率を用いた評価では約 9 割の性能を得ることができた．

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 [2] 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．しかし大竹らと 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がある．

そこで本研究では，新聞記事群データからノード同士の関係を表す文字列を抽出し，抽出した文字列を単語ネットワークのリンクに付与する手法を提案する．リンクに関係性を示す文字列を付与することで，構築した単語ネットワークから得られる情報がより詳細なものとなる．本研究の目的は，ノード同士の関係性を分かりやすくすることにより，単語ネットワークの利便性の向上を図ることである．

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情報がなく，関係性が分かりづらいという問題を解決するために，ノード同士の関係を単語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を付与する．

- リンクに文字列を付与することで，関係が分かりづらい単語同士の関係性を確認した．

- 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った．付与する文字列に，余分な部分や，関係性をさらに分かりやすくするには不十分な部分があっても正解とする基準で評価した場合，MRR を用いた評価では約 7 割，1 位正解率を用いた評価では約 6 割，5 位正解率を用いた評価では約 9 割の性能を得た．

- 文字列を抽出する際に，句読点を区切りとする手法と，句点を区切りとする手法の 2 つの手法で実験を行った．2 つの手法での実験結果を評価し，性能を比較した．その結果，2 単語間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評価は，句読点を区切りとする手法の方が良い性能となった．しかし，2 単語間の関係を示すものとして適切ではあるが，余分な部分があっても正解とする基準で評価した場合，句点を区切りとする手法の方が良い性能となった．また，2 単語間の関係を示すものとして適切ではあるが，余分な部分や，関係性をさらに分かりやすくするには不十分な部分があっても正解とする基準で評価した場合，両手法ともほぼ同等の性能という結果となった．

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，本研究の基本となるネットワーク構築の流れについて述べる．第 4 章では，リンクへの文字列付与の手法を提案する．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 [5] 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 [7] 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 [8] 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 [9] 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 [10] 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．

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 [3][4] 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 [2] 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．

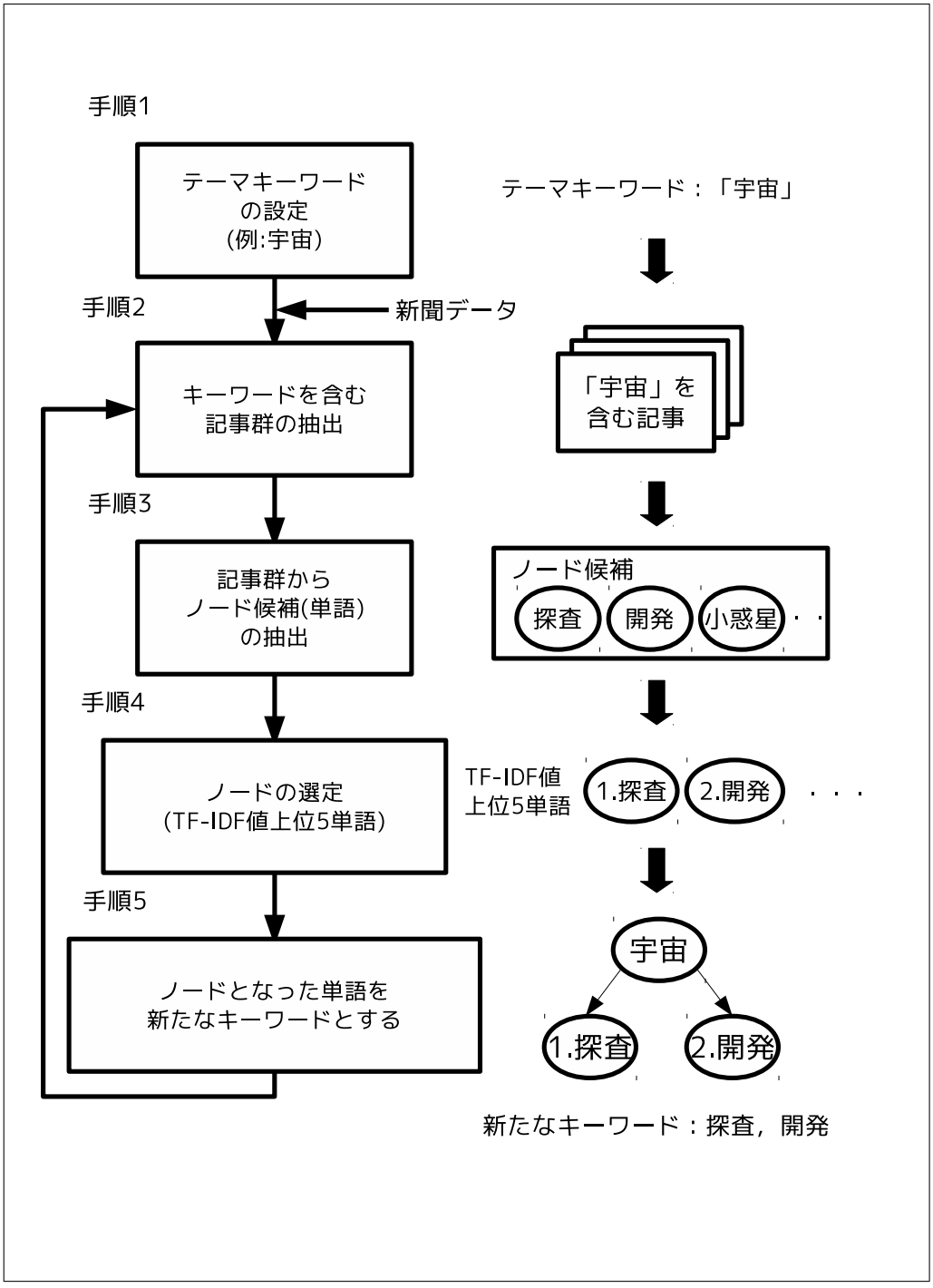

図 3.1: ネットワーク構築の流れ

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．

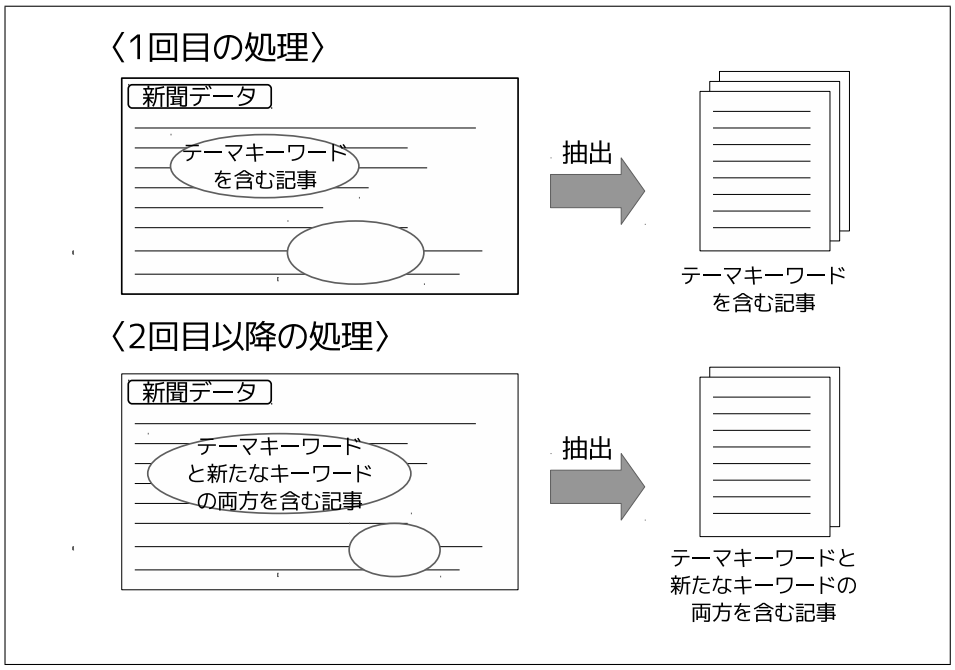

図 3.2: 記事の抽出

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」

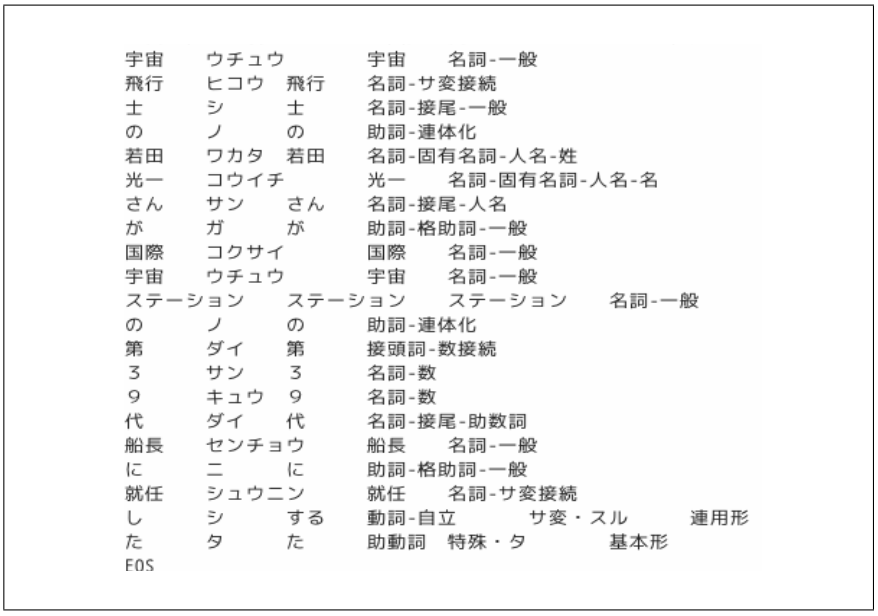

図 3.3: 形態素解析の出力例

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，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．

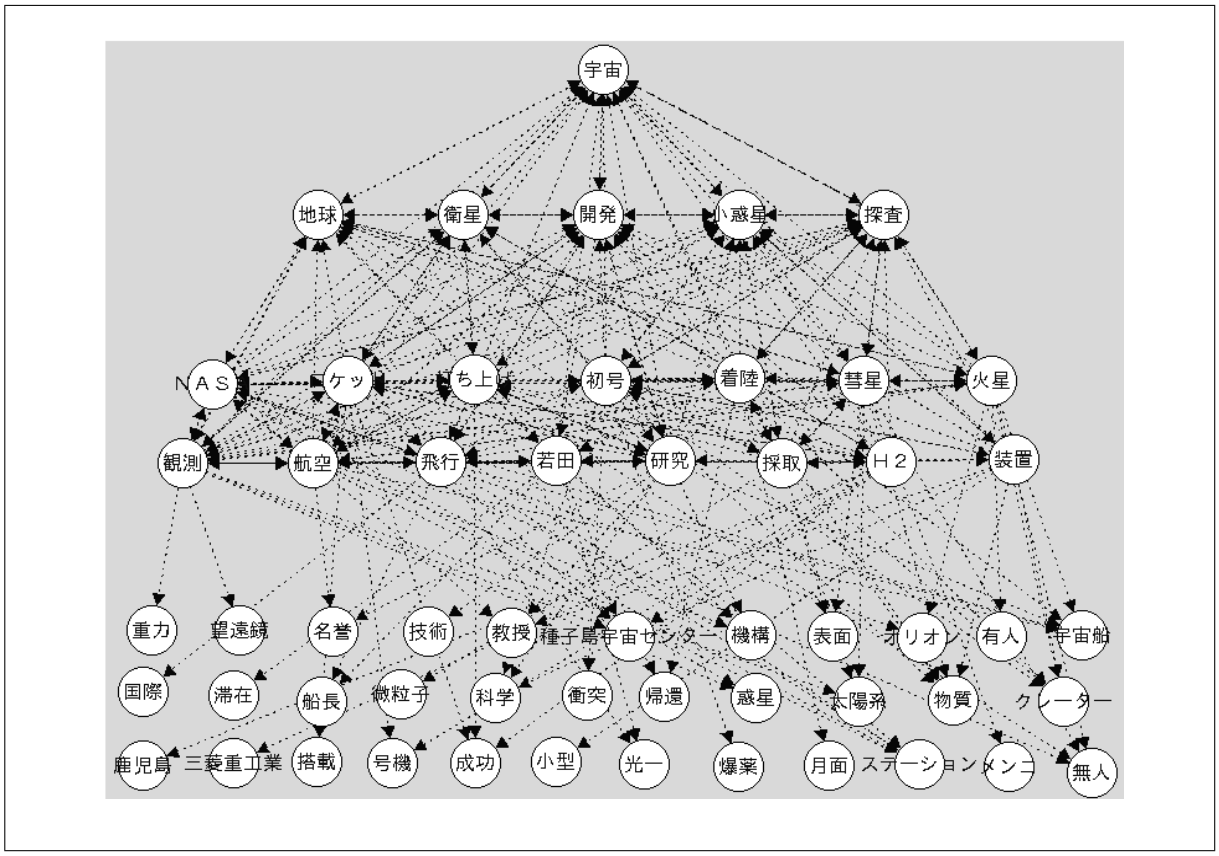

図 3.4: 構築したネットワークの例

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　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 4.1 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 4.1 に示す．

表 4.1: 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．4.1 式は，文字列の出現頻度が高いものを優先する式であり，4.2 式は，文字列の文字長が短いものを優先する式である．4.3 式は，割り算で優先度を求める式である．

$$\text{優先度} = (\text{頻度} * 10000) - \text{文字長} \tag{4.1}$$

$$\text{優先度} = -(\text{文字長} * 10000) + \text{頻度} \tag{4.2}$$

$$\text{優先度} = \frac{\text{頻度}}{\text{文字長}} \tag{4.3}$$

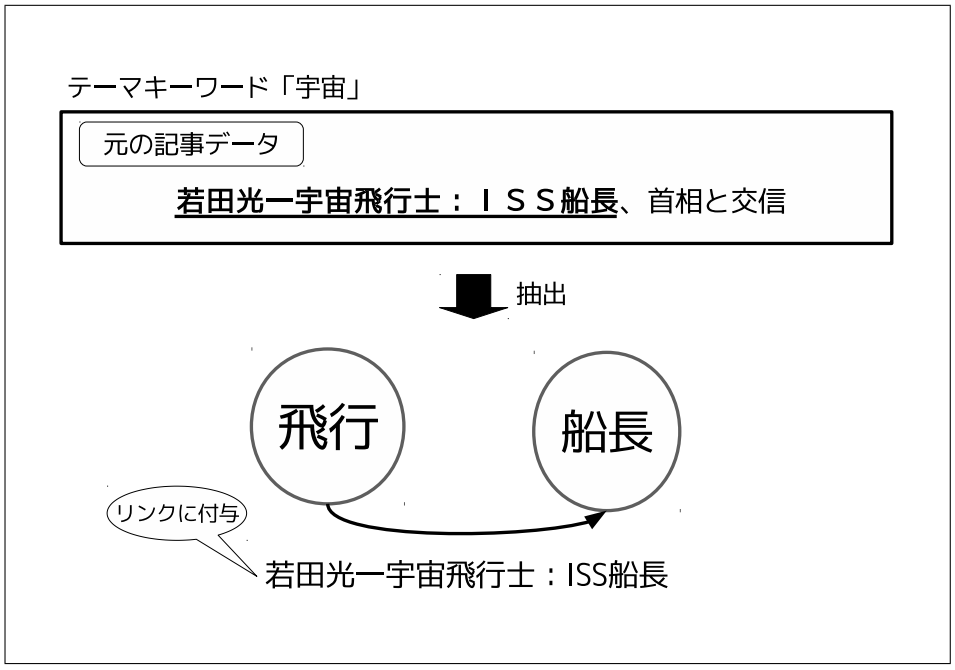

図 4.1: 単語ネットワークのリンクへの文字列付与の例

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.

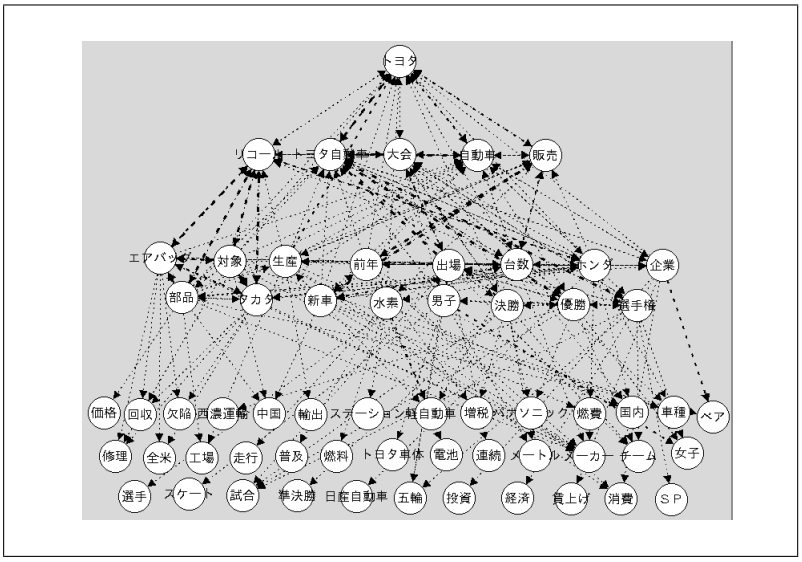

図 5.1: 「トヨタ」のネットワーク図

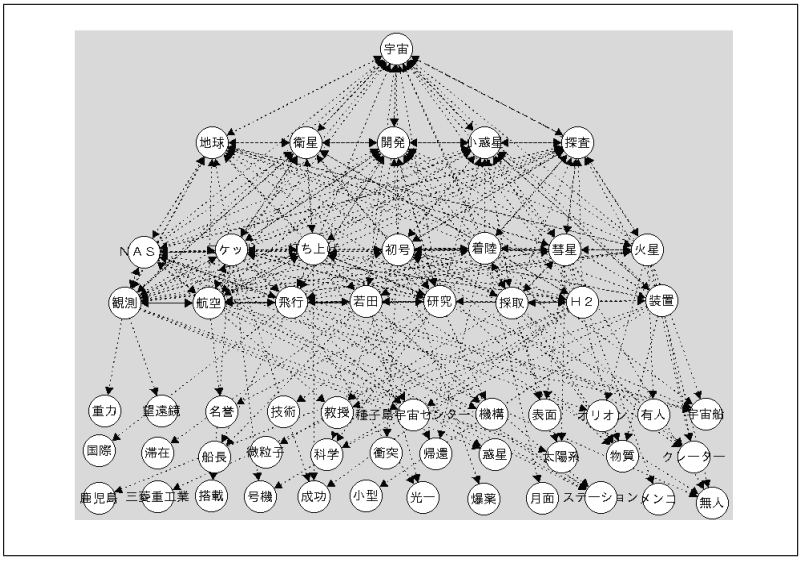

図 5.2: 「宇宙」のネットワーク図

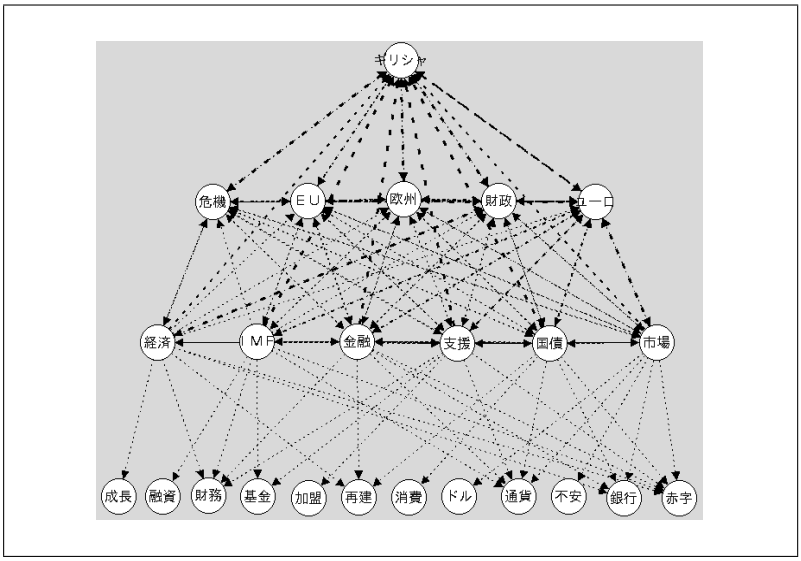

図 5.3: 「ギリシャ」のネットワーク図

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 4.1 節の手法の出力結果を優先度の上位 5 つまでの文字列とし，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，優先度の式を，式 4.1，式 4.2，式 4.3 の 3 パターンの 4 段階評価を行う．以降，本章では式 4.1 を「頻度大」，式 4.2 を「文字長小」，式 4.3 を「割り算」と表記する．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　句読点での実験結果

第 4 章の提案手法の出力結果の例を表 5.7，表 5.5，表 5.6 に示す．

表 5.5: ネットワーク「トヨタ」，単語対「タカタ」「修理」の出力例

優先度の式	出力結果	評価結果
頻度大	自動車部品大手タカタ製エアバッグのリコール（回収・無償修理）問題で	○
文字長小	タカタ製エアバッグ：４７４万台修理を	△
割り算	自動車部品大手タカタ製エアバッグのリコール（回収・無償修理）問題で	○

表 5.6: ネットワーク「宇宙」，単語対「ロケット」「小惑星」の出力例

優先度の式	出力結果	評価結果
頻度大	小惑星探査機「はやぶさ２」を載せた主力ロケットＨ２Ａ２６号機を打ち上げた	○
文字長小	Ｈ２Ａロケットで打ち上げられる小惑星探査機「はやぶさ２」	○
割り算	小惑星探査機「はやぶさ２」を載せた主力ロケットＨ２Ａ２６号機を打ち上げた	○

表 5.7: ネットワーク「ギリシャ」，単語対「支援」「ＥＵ」の出力例

優先度の式	出力結果	評価結果
頻度大	ＥＵはギリシャ支援によってユーロ防衛の決意を示し	□
文字長小	ＥＵ支援環境整う	×
割り算	ＥＵ：ギリシャ支援	○

5.6　句点での実験結果

第 4 章の 4.1 節で述べたリンクに付与する文字列の選定の手法で，手法の一部を変え，文字列の出力を行った．手法の変更点は,「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を抽出する」の句読点で区切る部分を，句点のみで区切るように変更した．実験条件は，5.1 節と同様の条件である．変更した手法の出力結果の例を表 5.10，表 5.8，表 5.9 に示す.

表 5.8: ネットワーク「トヨタ」，単語対「タカタ」「修理」の出力例 (句点)

優先度の式	出力結果	評価結果
頻度大	自動車部品大手タカタ製エアバッグのリコール（回収・無償修理）問題で、米下院エネルギー・商業委員会は３日、上院に続いて公聴会を開いた	□
文字長小	タカタ製エアバッグ：４７４万台修理を	△
割り算	米運輸省の道路交通安全局は１８日、欠陥が見つかった自動車部品大手タカタ製エアバッグのリコール（回収・無償修理）の対象地域を全米に拡大するようホンダなど自動車メーカーに指示したと発表した	□

表 5.9: ネットワーク「宇宙」，単語対「ロケット」「小惑星」の出力例 (句点)

優先度の式	出力結果	評価結果
頻度大	三菱重工業と宇宙航空研究開発機構（ＪＡＸＡ）は３日午後、小惑星探査機「はやぶさ２」を載せた主力ロケットＨ２Ａ２６号機を打ち上げた	□
文字長小	今年１２月にＨ２Ａロケットで打ち上げられ、小惑星１９９９ＪＵ３に到着するのは１８年６月	×
割り算	三菱重工業と宇宙航空研究開発機構（ＪＡＸＡ）は３日午後、小惑星探査機「はやぶさ２」を載せた主力ロケットＨ２Ａ２６号機を打ち上げた	□

表 5.10: ネットワーク「ギリシャ」，単語対「支援」「ＥＵ」の出力例 (句点)

優先度の式	出力結果	評価結果
頻度大	ＥＵはギリシャ支援によってユーロ防衛の決意を示し、危機が他のユーロ加盟国に飛び火する事態の回避を目指す	□
文字長小	ＥＵ支援環境整う	×
割り算	ＥＵ：ギリシャ支援合意	○

5.7　句読点での評価結果

本節は，句読点を区切りとした手法で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優先度の上位 5 つを取り出し，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．

5.7.1　「トヨタ」の評価結果

「トヨタ」の MRR を用いた評価結果を表 5.11 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: 「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

表 5.12: 「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

表 5.13: 「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

5.7.2 「宇宙」の評価結果

「宇宙」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: 「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

表 5.15: 「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

表 5.16: 「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

5.7.3 「ギリシャ」の評価

「ギリシャ」の MRR を用いた評価結果を表 5.17 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: 「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

表 5.18: 「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

表 5.19: 「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

5.7.4　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 5.20 に，1 位正解率を用いた評価結果を表 5.21 に，5 位正解率を用いた評価結果を表 5.22 に示す．

表 5.20: 「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

表 5.21: 「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

表 5.22: 「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

5.8 句点での評価結果

5.6 で述べた，提案手法の一部を変更した手法の出力結果を評価する．実験条件は，5.1 節と同様の条件である．評価方法は，5.2 節，5.3 節，5.4 節と同様の評価方法で行う．評価する単語対は，5.7 節と全て同じ単語対を用いる．

5.8.1 「トヨタ」の評価結果 (句点)

「トヨタ」の MRR を用いた評価結果を表 5.23 に，1 位正解率を用いた評価結果を表 5.24 に，5 位正解率を用いた評価結果を表 5.25 に示す．

表 5.23: 「トヨタ」の MRR を用いた評価結果 (句点)

	○	○□	○□△
頻度大	0.08	0.35	0.66
文字長小	0.21	0.21	0.45
割り算	0.17	0.21	0.55

表 5.24: 「トヨタ」の 1 位正解率を用いた評価結果 (句点)

	○	○□	○□△
頻度大	0.05	0.30	0.55
文字長小	0.20	0.20	0.35
割り算	0.10	0.15	0.40

表 5.25: 「トヨタ」の 5 位正解率を用いた評価結果 (句点)

	○	○□	○□△
頻度大	0.10	0.45	0.85
文字長小	0.25	0.25	0.60
割り算	0.30	0.35	0.80

5.8.2 「宇宙」の評価結果 (句点)

「宇宙」の MRR を用いた評価結果を表 5.26 に，1 位正解率を用いた評価結果を表 5.27 に，5 位正解率を用いた評価結果を表 5.28 に示す.

表 5.26: 「宇宙」の MRR を用いた評価結果 (句点)

	○	○□	○□△
頻度大	0.18	0.65	0.66
文字長小	0.29	0.47	0.54
割り算	0.24	0.65	0.73

表 5.27: 「宇宙」の 1 位正解率を用いた評価結果 (句点)

	○	○□	○□△
頻度大	0.10	0.50	0.50
文字長小	0.25	0.30	0.40
割り算	0.15	0.55	0.65

表 5.28: 「宇宙」の 5 位正解率を用いた評価結果 (句点)

	○	○□	○□△
頻度大	0.35	0.85	0.90
文字長小	0.35	0.65	0.70
割り算	0.40	0.85	0.90

5.8.3 「ギリシャ」の評価 (句点)

「ギリシャ」の MRR を用いた評価結果を表 5.29 に，1 位正解率を用いた評価結果を表 5.30 に，5 位正解率を用いた評価結果を表 5.31 に示す.

表 5.29: 「ギリシャ」の MRR を用いた評価結果 (句点)

	○	○□	○□△
頻度大	0.16	0.53	0.69
文字長小	0.36	0.39	0.74
割り算	0.40	0.48	0.72

表 5.30: 「ギリシャ」の 1 位正解率を用いた評価結果 (句点)

	○	○□	○□△
頻度大	0.10	0.30	0.45
文字長小	0.30	0.30	0.60
割り算	0.25	0.25	0.50

表 5.31: 「ギリシャ」の 5 位正解率を用いた評価結果 (句点)

	○	○□	○□△
頻度大	0.25	0.85	1.00
文字長小	0.50	0.55	0.95
割り算	0.60	0.80	1.00

5.8.4　3 つのネットワークを合わせた場合の評価 (句点)

「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 5.32 に，1 位正解率を用いた評価結果を表 5.33 に，5 位正解率を用いた評価結果を表 5.34 に示す．

表 5.32: 「トヨタ」「宇宙」「ギリシャ」の MRR を用いた評価結果 (句点)

	○	○□	○□△
頻度大	0.14	0.51	0.67
文字長小	0.29	0.36	0.58
割り算	0.27	0.45	0.67

表 5.33: 「トヨタ」「宇宙」「ギリシャ」の 1 位正解率を用いた評価結果 (句点)

	○	○□	○□△
頻度大	0.08	0.36	0.50
文字長小	0.25	0.27	0.45
割り算	0.17	0.32	0.52

表 5.34: 「トヨタ」「宇宙」「ギリシャ」の 5 位正解率を用いた評価結果 (句点)

	○	○□	○□△
頻度大	0.23	0.72	0.92
文字長小	0.37	0.48	0.75
割り算	0.43	0.67	0.90

第6章　考察

6.1　リンクへの文字列付与の考察

リンクへの文字列の付与を行うことで，2 単語がどのような関係を持っているのかが分かりづらい単語同士の関係性を得ることができた．しかし，単語対によっては，リンクに付与する文字列が取得できない場合があった．原因としては，提案手法で文字列を抽出する際に，2 単語の間の文字列に句読点が含まれる場合は，その文字列を省いているため，取得する文字列がないのではないかと考えられる．特に，記事数が少ない単語対に対して起こりうる可能性が高く，今後の課題として，手法の改良を行わなければならない．

6.2　優先度の式の考察

句読点を区切りとする手法で，文字列を取得する際に用いた優先度の式について考察をする．優先度の式は，文字列の出現頻度を重視する式，文字列の文字長が短いものを重視する式，出現頻度と文字長の割り算で優先度を求める式の 3 つの式を用いた．

まずは，文字長が短いものを重視した式について考察を行う．文字長を重視した式は，全ての評価方法と評価基準で，一番低い性能であることがわかった．原因として，抽出する文字列の短いものを重視しているので，2 単語間の関係性を示す情報の量が少なくなるためと考えられる．

次に，出現頻度を重視する式と割り算で優先度を求める式について考察を行う．出力の文字列が，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，2 つの式はほぼ同等の性能であることがわかった．しかし，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や，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，僅かながら出現頻度を重視した式が性能が良いという結果となった．

さらに，2 つの式は，1 位正解率での値が同等なのにも関らず，5 位正解率では頻度を

重視した式の方が，良いという結果がでている．つまり，頻度を重視した式では，優先度の上位 5 つの中に正解があることが多いということである．これより，出現頻度を重視することによって，2 単語の関係性を示すものとして適切な文字列が取得しやすくなると考えられる．

6.3 句点と句読点の考察

文字列を抽出する際に，句点で区切るのか句読点で区切るのかを変更し，評価を行った．その評価結果について考察する．出力の文字列が，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，句点を区切りとしても，句読点を区切りとしても性能はほぼ変わらなかった．しかし，2 単語間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，句読点で区切る方が良いという結果となり，関係を示すものとして適切であるが，余分な部分を含む場合を正解とする基準では，句点の方が良いと結果となった．この原因として，句点で区切ることによって，出力される文字列が句読点の場合よりも長くなり，余分な部分を含む文字列が多く抽出されたからなの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 [2] 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．しかし大竹らと 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があった．

そこで本研究では，新聞記事群データからノード同士の関係を表す文字列を抽出し，抽出した文字列を単語ネットワークのリンクに付与する手法を提案した．

その結果，リンクに文字列を付与することで，関係が分かりづらい単語同士の関係性を確認した．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った．付与する文字列に，余分な部分や，関係性をさらに分かりやすくするには不十分な部分があっても正解とする基準で評価した場合，MRR を用いた評価では約 7 割，1 位正解率を用いた評価では約 6 割，5 位正解率を用いた評価では約 9 割の性能を得た．

文字列を抽出する際に，句読点を区切りとする手法と，句点を区切りとする手法の 2 つの手法での実験も行った．2 つの手法での実験結果を評価し，性能を比較した．

その結果，2 単語間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評価は，句読点を区切りとする手法の方が良い性能となった．しかし，2 単語間の関係を示すものとして適切ではあるが，余分な部分があっても正解とする基準で評価した場合，句点を区切りとする手法の方が良い性能となった．また，2 単語間の関係を示すものとして適切ではあるが，余分な部分や，関係性をさらに分かりやすくするには不十分な部分があっても正解とする基準で評価した場合，両手法ともほぼ同等の性能という結果となった．今後は，評価した単語対数が少ないので，別の単語ネットワークの単語対も評価していきたいと考えている．

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知能情報工学科計算機工学講座 C 研究室の村田真樹教授，村上仁一准教授そして計算機工学講座 C 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] 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.

[3] 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.

[4] 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.

[5] 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.

[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] 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.

[8] 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.

[9] 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.

[10] 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.