単語制約を用いた概念ネットワークの改良

本田 涼太 [*1]　村田 真樹 [*2]　馬 青 [*3]
[*1] 鳥取大学 工学部 電気情報系学科
[*2] 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
[*3] 龍谷大学 先端理工学部 数理・情報科学課程
[*1,*2]{b17t2096a@edu.,murata@}tottori-u.ac.jp
[*3]qma@math.ryukoku.ac.jp

1　はじめに

近年，インターネットの普及等により電子テキストが増加している．これら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．そこで言語テキスト処理技術を用いテーマ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概念ネットワークの研究が進められた．

これまでの研究で，概念ネットワークの構築に際して，大竹ら [1] は，TF-IDF 法を用いて概念ネットワークの構築手法を提案した．また，土遠ら [2] は，概念ネットワークに出現した単語にテーマキーワードと無関係な単語があることに着目し，これら無関係な単語を出現させないために，「テーマ限定抽出法」を提案した．

しかし，これまでの研究では，関連する単語を概念ネットワークとして表示する際に，単に TF-IDF 値が大きい単語を取り出して，ネットワークを構築しているため，よく似た内容の単語であっても離れて出現することがあった．

そこで本研究では，この概念ネットワークの構築において，Word2vec[3] を用いてある単語から発展するネットワークの単語を同種の単語に制約する．そのようにすることでネットワークをより見やすくするように改良する．本研究の目的は，ネットワークの構築において出現する単語を同種の単語に制約し，より見やすいネットワークを構築することである．単語制約を行いよく似た単語を近くに配置することで見やすくしたネットワークの一例を図 1 に示す．図 1 では，「ビデオカメラ」と「レンズ」，「パソコン」と「スマートフォン」という似た意味の単語が近くに配置されて，見やすくなっている．

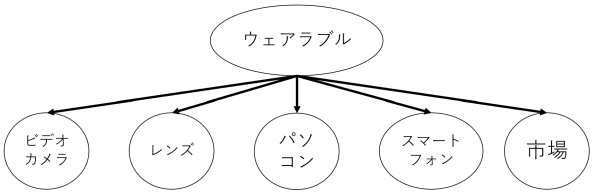


図 1　単語制約したネットワークの一例

2　先行手法

本節では，これまでの概念ネットワークに関する研究において用いられた手法について述べる．

2.1　ネットワーク構築の手順

まず, 大竹らが提案したネットワークの構築手法を述べる．

手順 1　構築したいネットワークの主となる単語をテーマキーワードとして設定する．
手順 2　キーワードとなる単語を含んだ記事群を抽出し，その記事群から形態素解析を用いて名詞のみを抽出する．その際，一文字，ひらがなのみ，数字のみの単語を除外する．
手順 3　手順 2 で抽出された単語の出現頻度を調べ，上位 100 単語をノード候補とする．
手順 4　得られたノード候補の中から，TF-IDF 法を用い，値の大きな上位 5 単語をネットワークのノードとして選定する．TF-IDF 法については 2.2 節にて述べる．
手順 5　手順 2 から手順 4 を繰り返して概念ネットワークを拡張する．

2.2 TF-IDF 法

ネットワークの構築において，ノードを選定する際に利用した TF-IDF 法について述べる．この節では入力データの電子テキストを新聞データとして説明する．TF とは単語頻度 (Term Frequency) のことであり，入力データにおいて，単語 t が出現した頻度のことをいう．また，DF は文書頻度 (Document Frequency) のことであり，単語 t がある記事群 A において出現した記事の数のことをいう．N を記事群 A の総記事数として，TF-IDF 法を用いたノードの選定式を式 (1) に示す．

$$w = tf * log(\frac{N}{df}) \tag{1}$$

2.3 テーマ限定抽出法

土遠らは，大竹らの構築したネットワークにテーマ限定抽出法を導入した．テーマ限定抽出法とは，2.1 節の手順 2 において，記事を抽出する際に，テーマキーワードと現在のキーワードの両方を含む記事を抽出するようにしたものである．そうすることにより，テーマキーワードと関連性のない，または，関連性が薄いと考えられる単語が取り出されにくくなる．

3 提案手法

本節では，本研究の提案手法である，単語制約を用いてネットワークを構築するという点について述べる．

3.1 Word2vec を用いた単語制約

単語制約には，Google 社が開発した Word2vec 内にある「単語のクラスタリング」を利用する．

単語のクラスタリングとは，Word2vec にテキストデータを学習させ，単語をベクトル化し，そのベクトルのコサイン類似度を求め類似度の高い単語をまとめて単語のクラスタを作り，各クラスタにクラスタ番号を割り当てるものである．このクラスタ番号が一致している単語群を似た意味を持つ単語とする．

3.2 単語制約を用いたネットワーク構築の手法

本研究では，2.1 節で述べた概念ネットワーク構築の手法に，2.3 節で述べたテーマ限定抽出法と，3.1 節で述べた単語制約の手法を加えてネットワークの構築を行う．この構築の手法を以下に示す．

手順 1 Word2vec の単語のクラスタリング機能を用いて，単語をクラスタ番号ごとにまとめる．

手順 2 2.1 節の手順 2 で得られたノード候補の単語の中ですでに出現している単語を除き，各単語の TF-IDF 値を計算する．

手順 3 TF-IDF 値を計算した後に TF-IDF 値の大きい順に単語を並べ，各単語のクラスタ番号を取り出し，クラスタ番号ごとに単語をまとめる．(図 2)

手順 4 クラスタ番号が同じ単語ごとに TF-IDF 値を計算し，TF-IDF 値が大きい順にクラスタ番号を並べる．

手順 5 手順 4 で求めた TF-IDF 値が上位 5 位までのクラスタ番号を持つ単語を抜き出し，上位のクラスタ番号に所属する単語から順に 5 個までネットワークに表示させる．

このうち，手順 4 で述べた TF-IDF 値の計算方法として，TF-IDF 合計値法と TF-IDF 最大値法の 2 通りを提案する．TF-IDF 合計値法は，クラスタ番号が同じ単語ごとにそれらの単語の TF-IDF 値を足し，その合計値上位 5 位までのクラスタ番号に所属する単語をネットワークに表示させる方法である．TF-IDF 最大値法は，クラスタ番号が同じ単語ごとにそれらの単語の TF-IDF 値の最大の値を探し，その最大値上位 5 位までのクラスタ番号に所属する単語をネットワークに表示させる方法である．

また，これら 2 つの計算方法の計算の一例を図 3 に示す．

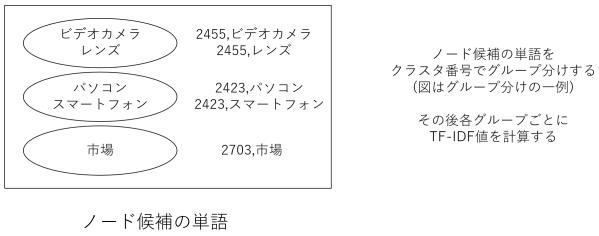


図 2 手順 3 の概要

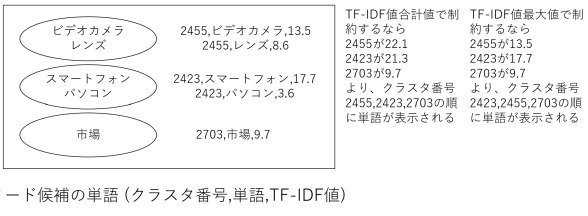


図 3 TF-IDF 値の計算例

4 実験と評価

4.1 実験データ

本実験では，2.1 節で述べた従来手法と，3.2 節で述べた 2 つの提案手法の合わせて 3 つの手法で概念ネットワークを構築した．まず，単語制約を行うために，Word2vec に学習させるデータとして，毎日新聞 12 年分 (2007〜2018) のデータ (1,166,761 記事) を使用した．単語のクラスタリングを行う際，クラスタ数は 5,000 とした．ネットワークを構築する際のデータとして 2018 年の毎日新聞の記事 (88,032 記事) を使用した．また，テーマキーワードとして 20 個の単語を選定した．

4.2 実験結果と評価基準

構築したネットワークについて次の 2 つの観点から評価した．

4.2.1 有用性

ネットワークに出現した単語がテーマキーワードの概念を理解するのに役に立つかという観点で評価した．具体的な評価点を以下に示す．

- 出現した単語について，あまり知らなかった事柄を知れた場合や，キーワードの概念を知るうえで役に立つと判断した場合．
- それぞれが知っている単語であっても，ノードのリンクによって新たな情報が得られる場合，意外な関係性である場合．

以上の 2 点に該当する箇所を，それぞれのテーマキーワードで計測した結果を表 1 に示す．

4.2.2 見やすさ

ネットワークの見やすい部分があるかという観点で評価した．具体的な評価点を以下に示す．

- 似た意味の単語が並んで出現していることにより，見やすくなっている場合．情報がまとまっていると考えられる場合．
- 似た意味の単語が並んでいることにより，知らなかった単語でも，web 検索等をして並んでいる単語が同じような意味であると判断した場合．

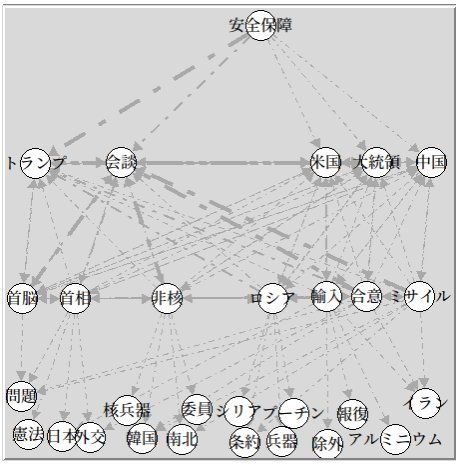
以上の 2 点に該当する箇所を，それぞれのテーマキーワードで計測した結果を表 2 に示す．

表 1 各テーマキーワードと役に立つ単語の数

テーマキーワード	従来手法	tfidf 値 合計値法	tfidf 値 最大値法
5G	6	5	1
がん	2	1	4
イギリス	1	0	1
オリンピック	1	2	1
パソコン	3	1	3
ロボット	3	4	4
安全保障	3	2	2
遺跡	6	5	5
宇宙	4	9	9
映画	4	9	6
感染症	2	3	5
京都	0	0	0
銀河	5	5	5
産業構造	5	3	3
寺院	2	3	2
世界遺産	8	6	4
石油	1	1	1
台風	1	2	2
独立	5	2	2
廃線	1	2	1
平均値	3.15	3.25	3.05

表 2 各テーマキーワードと似た意味の単語が並んで、見やすくなっている部分の数

テーマキーワード	従来手法	tfidf 値 合計値法	tfidf 値 最大値法
5G	1	2	0
がん	0	0	0
イギリス	1	5	5
オリンピック	1	3	2
パソコン	0	1	1
ロボット	0	3	2
安全保障	1	5	4
遺跡	1	5	5
宇宙	2	4	2
映画	0	1	1
感染症	1	3	3
京都	0	0	2
銀河	1	1	1
産業構造	1	4	1
寺院	3	6	4
世界遺産	1	3	3
石油	1	6	7
台風	1	2	1
独立	4	5	6
廃線	3	3	1
平均値	1.15	3.05	2.55



役に立つ単語:非核，憲法，ロシアとシリア
見やすくなっている部分:首脳と首相

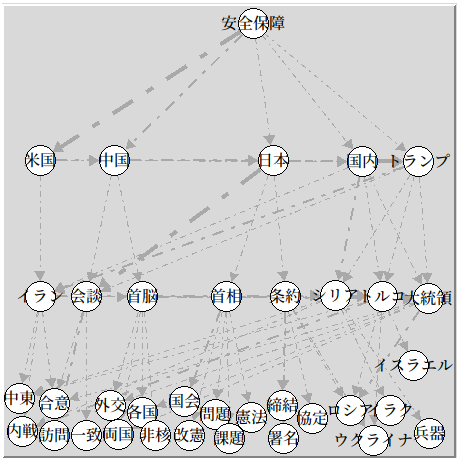
図 4　従来手法によるネットワーク

表 1 より，役に立つ単語の数の平均は，従来手法の 3.15 個に対して，TF-IDF 合計値法が 3.25 個と上回った．しかし，TF-IDF 最大値法は 3.05 個と下回った．また，表 2 より，見やすい部分の平均は，従来手法が 1.15 個であるのに対して，TF-IDF 合計値法が 3.05 個，TF-IDF 最大値法が 2.55 個といずれも上回った．この結果より，TF-IDF 合計値法が有用性，見やすさの観点からもっともよい方法であると考える．

また，構築したネットワークの一例として，テーマキーワードを「安全保障」として構築したネットワークを以下に示す．従来手法によるネットワークを図 4 に，TF-IDF 合計値法を用いたネットワークを図 5 に，TF-IDF 最大値法を用いたネットワークを図 6 にそれぞれ示す．また，それぞれの図の下にそのネットワークにおいて役に立つ単語と，似た意味の単語が並び見やすくなっている部分を列挙している．

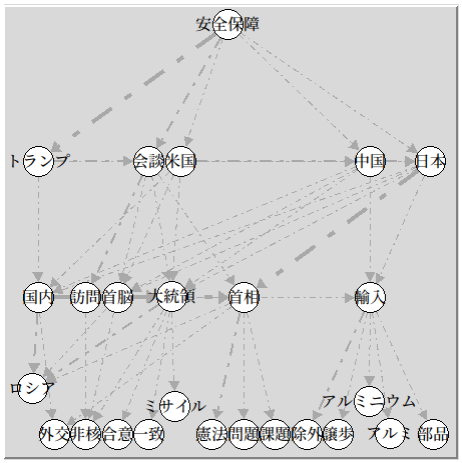
5　おわりに

本研究では，概念ネットワーク構築の際に Word2vec を利用した単語制約を用いて，似た意味の単語を近くに配置し，ネットワークを見やすくすることを目的とした．役に立つ単語の平均は，従来手法が 3.15 個で TF-IDF 合計値法が 3.25 個，TF-IDF 最大値法が 3.05 個と，従来手法と比べても情報量が減少することを抑え，見やすい部分の平均は，従来手法が 1.15 個に対して，TF-IDF 合計値法が 3.05 個，TF-IDF 最大値法が 2.55 個と，似た意味の単語が並んで見やすくなっている部分は増えた．また，提案



役に立つ単語:非核，憲法
見やすくなっている部分:米国と中国と日本，首脳と首相，シリアとトルコ，締結と署名，ロシアとウクライナ

図 5　TF-IDF 合計値法を用いたネットワーク



役に立つ単語:非核，憲法
見やすくなっている部分:米国と中国と日本，首脳と大統領と首相，合意と一致，問題と課題

図 6　TF-IDF 最大値法を用いたネットワーク

手法のなかでは，TF-IDF 合計値法を用いたネットワークのほうが，より見やすいネットワークを構築していることが分かった．

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]土遠雄大. 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無関連ノードの扱い. 鳥取大学卒業研究発表会論文, 2013.

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