企業事業文書に対するグラフ構造化と Graph Convolutional Networks を用いた特徴語抽出の提案

GRAPH STRUCTURING FOR CORPORATE BUSINESS DOCUMENTS AND FEATURE WORD EXTRACTION FROM THE GRAPH USING GRAPH CONVOLUTIONAL NETWORKS

松浦遼
Ryo MATSUURA
指導教員　藤井章博

法政大学大学院理工学研究科応用情報工学専攻修士課程

This study proposes a graph structuring and feature word extraction method for business information documents extracted from corporate websites for corporate and market trend research. The graphs created from business documents consist of a hierarchical structure graph reflecting the relationship between businesses and companies, and a dependency graph using the dependency relations of business documents. For feature word extraction, R-GCNs were used for the dependency graph, and the F value was 0.33 for the business document graph consisting of a single company, and 0.37 for the business document graph consisting of multiple companies in a single industry.

Key Words *:business documents, Graph Convolutional Networks, GCNs*

１．はじめに

企業が自社の活動や決算，自社に関するニュースなど会社についての様々な情報を発信する上で，企業のウェブは非常に重要な役割を果たしている．そのため，企業ウェブページにはその企業の特徴が表されていると考えられる．

本研究では，企業ウェブページから事業に関する文書の取得，分析を行い，その文書から企業事業文書グラフの作成とその企業が行っている事業を表す特徴語抽出を目的とする．特徴語抽出では企業事業文書中に出現する単語の係り受け関係を用いてグラフ化し，係り受け関係を用いた特徴語抽出手法について検討する．

２．企業事業文書

（１）データセット

本研究において企業事業文書とは，企業が Web サイト上で事業を示唆するワードを用いて紹介しているテキストや，そのリンク先のテキストとする．

企業事業文書として，文部科学省科学技術・学術制作研究所 NISTEP が提供している NISTEP 企業名辞書 ver.2020_2 [2]に記載のある企業から，松浦ら[3]の手法を用いて 1568 社より，22251 文書を抽出した．

（２）企業事業文書における特徴語

本研究における企業事業文書の特徴語は大きく 2 種類に分けられる．特徴語とその例を表 1 に示す．

表 1　企業事業文書の特徴語

その事業が提供しているシステムや機能 （例）　〇〇を実現する～～
その事業を通じて解決する課題 （例）　〇〇を改善する～～

３．提案手法

（１）グラフを用いた特徴語抽出

本研究における企業事業文書の特徴語はその事業の生産，販売，サービスに関する単語であるため，特定の単語の近くに出現することが多いと考えられる．例えば”提供”,“支援”などである．そのため，各単語の関係性をグラフで表すことができれば，特定の単語の近傍や，特定の関係性を用いることで，特徴語を推測できると考える．本研究では単語同士の関係性を文章における単語の係り受け関係を用いて表現する．

（２）企業事業文書グラフ

本研究における企業事業文書グラフは，企業事業文書のもつ階層構造を表した階層構造グラフと企業事業文書中の単語の係り受け関係を表した係り受けグラフの２つ

に分けられる.

a）ノードの性質

本研究では，企業事業文書に対して 4 種類のノードを用いてグラフ構造化を行う．各ノードとそのノードが持つ属性（プロパティ）を表２に示す.

表２　ノードと各ノードのプロパティ

ノード名	プロパティ
company_node	company_name, company_id
business_node	company_id, business_name, business_id
sentence_node	business_id, sentence_id
word_node	word_id

company_node は属性として企業名と企業 ID 持ち，business_node は属性として事業名と事業 ID に加えてどの企業の事業かを表す企業 ID を持つ. sentence_node は事業 ID とその企業事業文書において何番目の文章になるかを示す文章 ID を持つ．また，business_node はその企業から抽出した企業事業文書を表すため，その企業の企業事業文書だけ存在し，sentence_node は各企業事業文書に存在する文章の数だけ存在する.

本研究において word_node が表す単語とは，文節ごとに存在する主辞となる単語とする．例えば「当社は，クラウドコンピューティングや人工知能，データ分析を用いてサービスを提供しています．」という文章であれば，”当社”，”クラウドコンピューティング”，”人工知能”，”データ分析”，”用い”，”サービス”，”提供”が word_node として生成される．これは企業事業文書における特徴語が複合名詞で表される場合が多く，単純な形態素で表した場合，特徴語を表現することが難しいためである．また，日本語の性質として助詞を始めとする特定の単語は高確率で文章に存在する. そのためグラフ化，特徴語推定を行う際にノイズになると考え，文節ごとの主辞単語のみを word_node として用いた.

b）企業事業文書階層構造グラフ

本研究における階層構造グラフは有向グラフで表される.

company_node と business_node の関係性を図 1，business_node と sentence_node の関係性を図 2，sentence_node と word_node の関係性を図 3 に示す.

なお, グラフの可視化にはグラフデータベース Neo4j[4]を用いた．赤ノードが company_node，緑ノードが business_node，橙ノードが sentence_node，青ノードが word_node をそれぞれ表す.

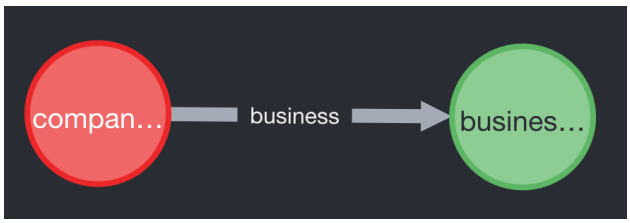

図 1　company_node と business_node の関係性

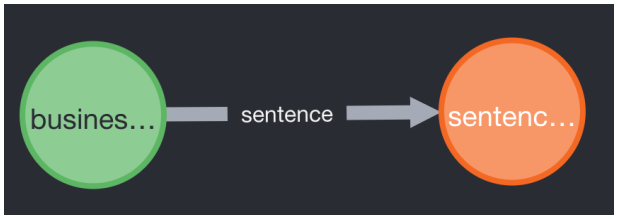

図 2　business_node と sentence_node の関係性

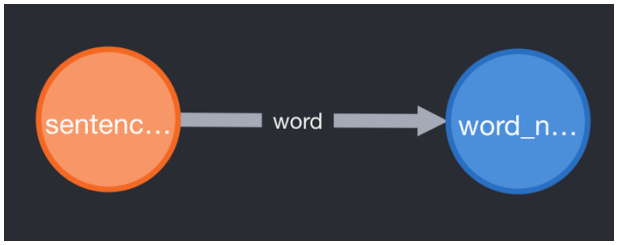

図 3　sentence_node と word_node の関係性

company_node i をC_i，C_iが行っている business_node j を$B_j^{C_i}$とする．C_iと$B_j^{C_i}$は business エッジを用いて関係を持つ．また，$B_j^{C_i}$のk番目の sentence_node を$S_k^{B_j^{C_i}}$とすると，$B_j^{C_i}$と$S_k^{B_j^{C_i}}$は sentence エッジを用いた関係を持ち，$S_k^{B_j^{C_i}}$内で出現する word_node l をW_lとすると，$S_k^{B_j^{C_i}}$とW_lは word エッジを用いた関係を持つ.

上記の関係性を持った階層構造グラフを，株式会社 ACCESS[5]の企業事業文書に対して適応した例を図 4 に示す.

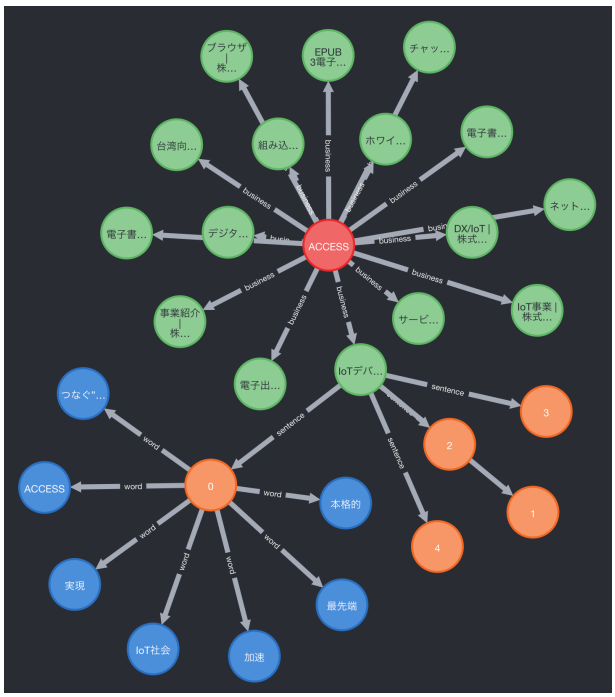

図 4　株式会社 ACCESS の企業事業文書を用いた企業事業文書階層構造グラフ

c）単語係り受け関係グラフ

企業事業文書に対する形態素解析と word_node 同士の係り受けの解析には，Megagon Labs が提供する日本語自然言語処理オープンソースライブラリ GINZA[6][7]を用いた．GINZA は Universal Dependencies[8]に基づく依存構造解析を行う．

本研究における係り受け関係グラフは有向グラフで表現される．係り受け関係の表現に用いるエッジは Universal Dependencies に記載されている Universal Dependency Relations によって区別される．

前述の例文「当社は，クラウドコンピューティングや人工知能，データ分析を用いてサービスを提供しています．」に対する文節主辞単語における係り受け関係を図 5 に，係り受け関係グラフに適応した例を図 6 に示す．

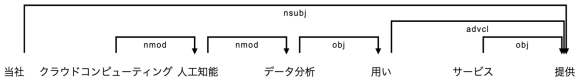

図 5　例文における文節主辞単語の係り受け関係

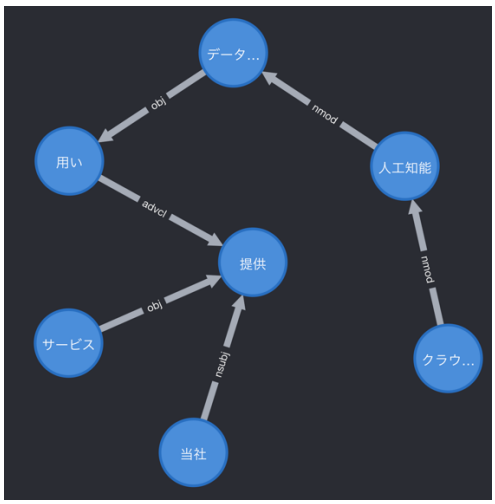

図 6　例文を用いた係り受け関係グラフ

図 5, 図 6 の矢印に付随する英単語は係り受け関係の種類を表している．本研究においてグラフ作成で用いた係り受け関係一覧とその個数を表 3 に示す．係り受け種類の表記は Universal Dependencies と同じである．

表 3　係り受け関係の種類とその個数

係り受け種類	個数	係り受け種類	個数
nmod	15886	compound	213
obl	7449	amod	78
acl	7165	dislocated	65
obj	6459	csubj	41
advcl	3480	ccomp	26
nsubj	3336	nummod	4
advmod	733	discourse	4
cc	633	punct	2
dep	619	case	2
det	326	aux	2

本研究ではグラフを 2 種類作成した．

一つは企業ごとに企業事業文書に出現する単語集合を分けて，単一の企業に出現する単語と係り受け関係のみで係り受けグラフを作成する．本論文ではこれを company_graph と呼ぶ．

もう一つは業種ごとに企業事業文書に出現する単語集合を分けて，その業種に属する企業事業文書に出現する単語と係り受け関係のみで係り受けグラフを作成する．本論文ではこれを industry_graph と呼ぶ．

（３）R-GCNs を用いた特徴語抽出

a）Relational Graph convolutional Networks(R-GCNs)

Relational Graph convolutional Networks(R-GCNs)[9]は，グラフ形式のデータに対して，用いられるグラフ畳み込みニューラネットワーク(GCNs)の一つである．

R-GCNs はノードの特徴量を抽出し，各ノードの畳み込みを自身と周囲のノードを用いて行われる．この時，各ノ

ードの畳み込みが関係ごとに行われ，最終的に各ノードに対して周囲のノードと関係性を考慮した特徴量が得られる．

b）word_node の特徴量

本研究では word_node の特徴量として，単語分散表現を用いた．単語分散表現は 1568 社，22251 文書から fasttext[10]を用いて取得した．学習は skip-gram 法を用い，分散表現の次元数は 100 とした．

また，学習の際は形態素で行い，複合名詞など，複数の単語が含まれる word_node に対しては含まれている単語の分散表現の合計値をそのノードの特徴量とした．

４． 実験

企業事業文書データセットから，総務省が定めている日本標準産業分類のうち情報通信業に分類される 42 社，280 の企業事業文書に対して，各単語に特徴語ラベル付けを行い，係り受け関係グラフの作成を行なった．

company_graph と industry_graph それぞれのノード数とエッジ数，特徴語ノード数を表 4 に示す．

表 4　各グラフのノード数 エッジ数 特徴語ノード数

	ノード数	エッジ数	特徴語ノード数
company_graph	29690	46523	2033
industry_graph	17617	46523	1780

学習において，industry_graph は全ノード中 8 割を学習データ，2 割をテストデータとし，company_graph では 42 社の 8 割に当たる 33 社のグラフを学習データ，残りの 11 社のグラフをテストデータとした．学習パラメータとして，学習率を 0.005，中間層の次元を 100，epoch 数を 400 とした．

また，従来の手法として，SVM，MLP を用いて同様の実験を行い，特徴語抽出の精度を比較した．

５． 結果

company_graph, industry_graph 対する R-GCNs と SVM, MLP それぞれの再現率，適合率，F 値を表 5 に示す．R-GCNs と MLP は 10 回実験を行い，その平均値を示す．また R-GCN（company_graph）に対する PR 曲線を図 7 に，R-GCN(industry_graph)に対する PR 曲線を図 8 に示す．

表 5　　各手法に対する再現率，適合率，F 値

	再現率	適合率	F値
R-GCN(company_graph)	0.35±0.02	0.32±0.04	0.33±0.03
R-GCN(industry_graph)	0.40±0.03	0.34±0.03	0.37±0.03
SVM	0.06	0.72	0.11
MLP	0.28	0.32	0.30

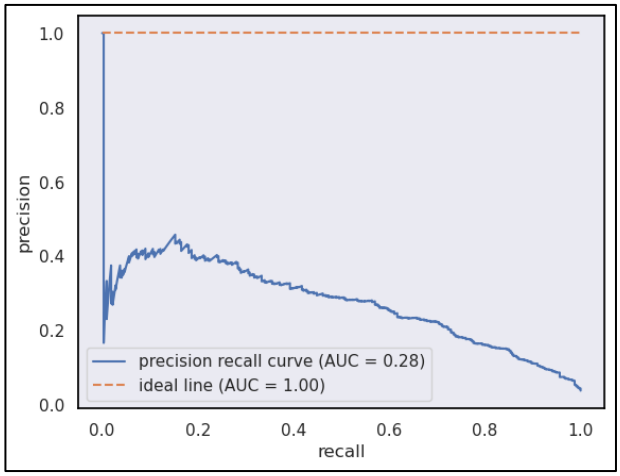

図 7　　R-GCN(company_graph)に対する PR 曲線

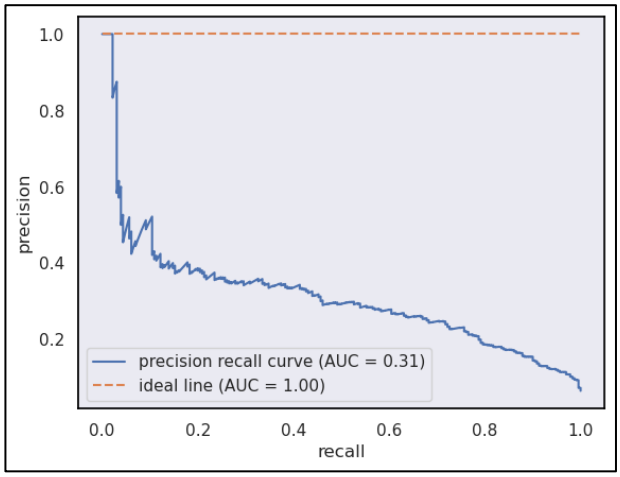

図 8　　R-GCN(industry_graph)に対する PR 曲線

R-GCN(company_graph)に比べて R-GCN(industry_graph)がより高い精度となった．また，どちらのグラフにおいても再現率が適合率より精度が高かった．また，company_graph, industry_graph どちらにおいても従来の手法より高い精度となった．

６． 考察

compay_graph に比べて industry_graph の方が精度が高かった理由として，グラフの密度が考えられる

company_graph での 42 社の企業事業文書グラフのノード数とエッジ数の平均はそれぞれ 706.9 と 1107.7 だった．R-GCNs は周囲のノードの特徴量と，関係性の種類を各ノードの特徴量として畳み込むため，Industry_graph に比べノード数やエッジ数が少ない企業単体の企業事業文書グラフでは，特徴語の特徴量獲得がより困難であったと考えられる．

Industry_graph において，偽陽性に分類されたノードと２つ以上隣接するノードの部分グラフを図 9，テストデータの特徴語ノードと隣接するノードの部分グラフを図 10 に示す．

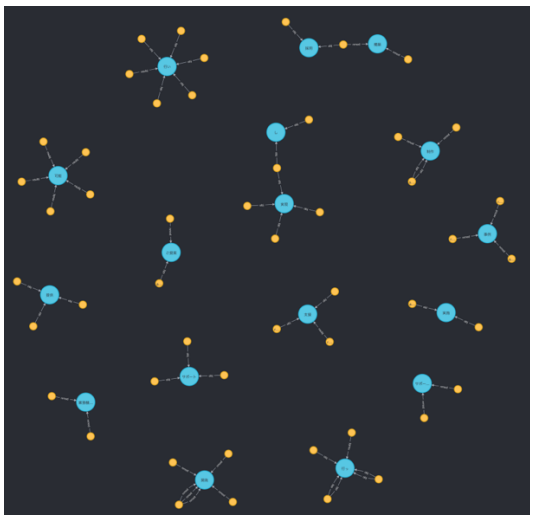

図 9　　偽陽性に分類されたノードの部分グラフ
(黄色：偽陽性ノード)

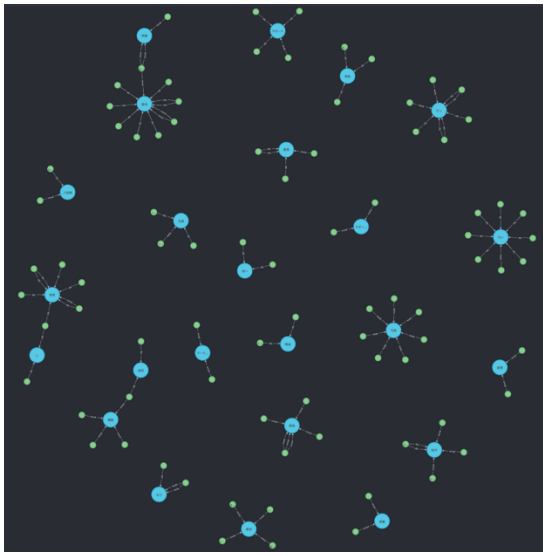

図 10　　テストデータの特徴語ノード部分グラフ
（緑色：特徴語ノード）

特徴語ノード 4 つ以上が集中しているノードの隣接特徴ノード数と，偽陽性に分類された部分グラフにおける隣接する偽陽性ノードの数を表 6 に示す.

表 6　　隣接正解ノード数と隣接偽陽性ノード数

単語	隣接正解ノード数	隣接偽陽性ノード数
提供	9	3
行い	8	6
可能	7	5
行っ	6	4
実現	6	4
開発	4	4
事例	4	3
制作	4	3
構築	4	2
サポート	4	3

表 5 より，複数の特徴語ノードが隣接するノードはその周囲のノードが特徴語として分類される可能性が高かった．また，隣接ノードが 1 つのみの特徴語ノードは，その多くが偽陰性に分類されており，十分に特徴量を得ることができなかったと考えられる．

本研究では隣接するノードの特徴量と係り受けの関係性のみに畳み込みを行い，特徴語ノードの特徴量を得たが，さらに周囲のノード特徴量の畳み込みや，係り受け関係に重みづけをしたグラフを用いることで精度向上が考えられる．

7．結論

本研究では，企業ウェブページから抽出した企業事業文書を用いた企業事業文書グラフの提案と，単語係り受け関係グラフからの特徴語抽出手法の検討を行った．単語係り受け関係グラフに対して R-GCBs を用いることで企業単体の単語係り受けグラフから F 値およそ 0.33，業種ごとの単語係り受けグラフから F 値およそ 0.37 を得た．また，企業ごとの企業事業文書における特徴語抽出，業種ごとの企業事業文書における特徴語抽出において，共に SVM，MLP よりも高い精度が得られた．

謝辞

本研究を進めるにあたり，熱心なご指導をいただきました法政大学理工学部応用情報工学科藤井章博教授に感謝いたします．研究活動を支えてくださった法政大学大学院理工学研究科平成 26 年度修士課程修了清水宏泰氏，法政大学大学院理工学研究科令和 3 年度修士課程修了田中直哉氏，藤井研究室の皆様に感謝いたします．また，修士課程修了まで支えてくださった両親にも感謝いたします．

参考文献

1) 福田悟志，難波英嗣，竹澤寿幸：論文と特許からの技術動向情報の抽出と可視化，情報処理学会論文誌，データベース，Vol.6，pp.12-29，2013.

2) 文部科学省科学技術・学術制作研究所，“NISTEP 企業名辞書 ver.2020_2”，https://nistep.repo.nii.ac.jp/，（参照 2023-02-05）.

3) 松浦遼，藤井章博：企業 Web サイトからの事業情報の抽出と業種別の類似事業推定，情報処理学会第 84 回全国大会講演論文集，pp.837-838，2022.

4) Neo4j，”Neo4j Graph Data Platform”，https://neo4j.com/，（参照：2023/02/05）.

5) 株式会社 ACCESS，”株式会社 ACCESS”，https://www.access-company.com/，（参照：2023/02/05）.

6) 松田寛，大村舞，浅原正幸：短単位品詞の用法曖昧性解決と依存関係ラベリングの同時学習，言語処理学会第 25 回年次大会発表論文集，pp.201-204，2019.

7) Megagon Labs，” GiNZA - Japanese NLP Library”，

https://megagonlabs.github.io/ginza/,（参照:2023/02/05）.
8）Universal Dependencies，” Universal Dependencies”, https://universaldependencies.org/,（参照 : 2023/02/05）.
9）Schlichtkrull Michael.，Kipf Thomas N.，Bloem Peter.， et al. : Modeling Relational Data with Graph Convolutional Networks，Lecture Notes in Computer Science，pp.593-607，2019．.
10）Bojanowski, P.， Grave, E.， Joulin, A.， et al.: Enriching Word Vectors with Subword Information，Transactions of the Association for Computational Linguistics ， Vol.5，pp.135-146，2017.