計測自動制御学会東北支部 第 302 回研究集会 (2016.6.24)
資料番号 302-7

震災体験談についてのテキストマイニング

Text mining on Great East Japan earthquake experiences

○熊野雄大 *，武山　泰 **

○ Yu'dai Kumano*, Yasushi Takeyama**

*八戸工業大学 大学院 工学研究科 電子電気・情報工学専攻,
**八戸工業大学 工学部 システム情報工学科

*Doctor of Engineering Program in Electronic, Electrical and Information Engineering,
Hachinohe Institute of Technology,
**Department of System and Information Engineering, Faculty of Engineering,
Hachinohe Institute of Technology

キーワード： テキストマイニング (text mining)，東日本大震災 (Great East Japan earthquake)

連絡先： 〒 031-8501　八戸市大字妙字大開 88–1 八戸工業大学 工学部 システム情報工学科
武山　泰，Tel.: (0178)25-8097，Fax.: (0178)25-1691，Email: takeyama@hi-tech.ac.jp

1. はじめに

近年，情報通信技術の発展にともない，インターネット上などで大量の情報が行き交っている．多くの数値情報なども蓄積されるようになってきていて，これらの一部は「ビッグデータ」と呼ばれ，分析・活用され始めている．

一方，インターネット上の SNS などで多くのテキストも行き交っており，テキストマイニングと呼ばれるデータ分析の対象となっている．例えば，NHK の TV 番組においては,「つぶやきビッグデータ」として，Twitter の tweet の解析結果が用いられていた．

本研究においては，東日本大震災の際の体験談として，収集され，アーカイブされているテキストを対象にテキストマイニングを試みた．

2. テキストマイニング

テキストマイニングとは，文字列を対象とするデータマイニングのことで，通常のテキストをデータとして，単語や文節で区切り，それらの出現の頻度や，共出現の相関などを解析することで有用な情報を取り出すテキストデータの分析方法である．

テキストデータの多くは形式が定まっておらず，また日本語は英語などとは違って単語間にスペースが挟まれず，また，膠着語であり，文法の揺らぎが大きいことなどから，従来，形態素解析が困難とされてきた．近年，自然言語処理の発展により実用的な水準での分析が可能となってきている．

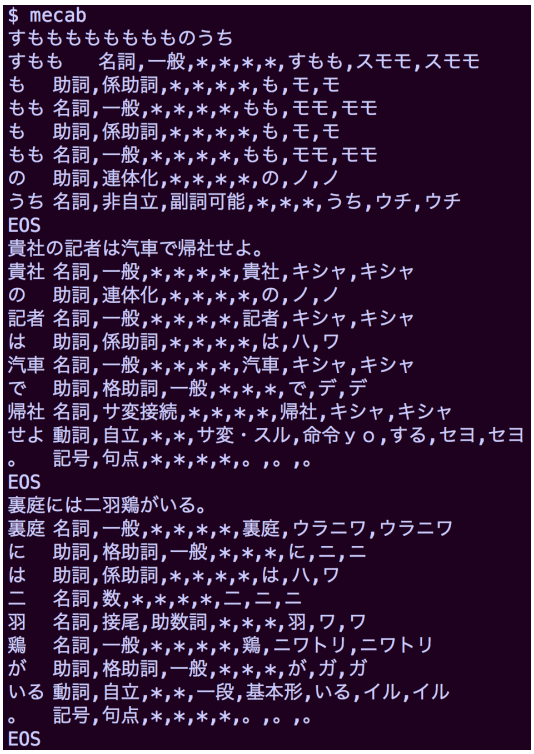

Fig. 1　形態素解析の例．Example of morphological analysis using MeCab.

2.1　形態素解析

形態素解析 (morphological analysis) とは，文法的な情報が注記されない自然言語のテキストデータから，対象言語の文法と，辞書と呼ばれる単語の品詞等の情報にもとづいて，形態素 (morpheme, 言語において意味を持つ最小単位) に分割し，それぞれの形態素の品詞などを判別する作業のことである．

自然言語処理分野において主要なテーマのひとつであり，かな漢字変換や機械翻訳などの応用面からも重要である．

2.1.1　MeCab

MeCab は，京都大学情報科学研究科と日本電信電話株式会社コミュニケーション科学基礎研究所の共同研究ユニットプロジェクトを通じて開発されたオープンソースの形態素解析エンジンである．

Fig. 1 に MeCab を使用した形態素解析の例を示す．

「mecab」のコマンドで起動した後に，日本語のテキストを入力すると，形態素に分割され，形態素ごとに，形態素の表層形，品詞, 品詞細分類 1, 品詞細分類 2, 品詞細分類 3, 活用型, 活用形, 原形, 読み, 発音，の順に表示する．

本研究では解析に R を用いることから，形態素解析には RMeCab のパッケージを用いる．

2.2　分析に用いたテキスト

分析に用いたのは，青森震災アーカイブで公開されている震災体験談である．

青森震災アーカイブは，八戸市，三沢市，おいらせ町，階上町の 4 市町が運営し，八戸市庁のサーバで公開されている．

震災体験談の収集は，平成 25 年度に行われ，体験談の収集の一部，および，インタビュー音源からのテープ起しについて，八戸工業大学が実施した．

3.　震災体験談についてのテキストマイニング

3.1　N-gram

N-gram とは，テキスト内のある言語単位（文字や形態素，品詞など）が一般に N 言語単位，隣接して用いられる言語単位の共起関係を表し，文書の特徴を表すものと考えられる．

3.2　共起の指標

RMeCab パッケージに実装されている collocate() 関数は，第 1 引数にファイル名，第 2 引数 node にノード，第 3 引数 span に前後の語数を指定する．

今回は第 1 引数に震災テキスト，第 2 引数 node には震災テキストの中から出現数の多かった名

Table 1　bigram 分析結果．N-gram analysis result.

bigram	Freq	品詞限定	Freq	名詞のみ	Freq
て-いる	7351	人-たち	541	人-たち	541
する-て	6075	お客-さん	506	お客-さん	506
と-思う	3895	こと-ない	494	避難-所	468
ん-けど	3265	避難-所	468	の-ん	430
が-ある	3194	ない-ん	433	の-の	390
に-なる	2943	安否-確認	377	安否-確認	377
て-くる	2683	次-日	374	次-日	374
の-は	2531	懐中-電灯	353	こと-ん	353
の-ほう	2103	の-ん	348	懐中-電灯	353
に-は	2057	ん-それ	302	ん-それ	307
なる-て	1993	の-の	295	一-人	293
の-が	1961	一-人	293	反射-式	291
か-な	1919	反射-式	291	震災-後	287
ん-よ	1819	震災-後	287	ん-の	283
に-行く	1727	被害-ない	286	式-ストーブ	249
を-する	1685	こと-ん	265	3-日	221
な-と	1664	ん-の	259	発電-機	200
する-てる	1508	式-ストーブ	247	よう-感じ	190
の-か	1452	いい-の	241	電気-復旧	190
よう-に	1395	3-日	221	よう-気	186
に-いる	1377	ない-の	219	津波-の	186
よ-ね	1366	の-ない	212	子供-たち	185
て-も	1313	電気-ない	204	一-回	179
って-いう	1271	発電-機	200	地震-とき	176

詞「contact(連絡)」,「damage（被害）」,「earthquake(地震)」,「electrical（電気）」,「evacuation（避難）」,「information（情報）」,「man(人)」,「tsunami（津波）」,「work（仕事）」の９つ．第3引数の span はデフォルトのまま３で行った．

collocate() 関数によってノードと共起するタームの頻度が算出されれば，次に共起語の頻度が有意に大きいかを調べることができる．ここで有意とは，二つのタームが特定の範囲内に共起した回数が，偶然では考えられないほど大きいことを意味する．有意かどうかを統計的に判定する基準として，コーパス言語学ではT値やMI値といった指標が用いられる．

3.2.1　T 値

T 値とは，統計解析で平均値の差の検定を行う場合などに使われる指標である．したがって，母集団の分布が正規分布となっていることを前提とする．ところが，単語は文法や書き手の意図に制約されているので，正規分布を仮定したT値は，言語データを分析する手段としては適切ではないとする考え方がある．しかし，コーパス言語学ではタームの共起関係の有無を調べる指標とし広く使用されている．T 値を計算する式は以下である．

$$\text{T 値} = \frac{\text{実数値} - \text{期待値}}{\text{実数値の平方根}} \qquad (1)$$

ここで分母の“ 実数値の平方根 ”は，ノードと共起語に関する標準偏差の近似値を表すT値の評価であるが，統計学ではT値の絶対値が2を超えるかどうかを簡易的な目安にすることがあるが，コーパス言語学では 1.65 以上であれば，二つのタームの共起は偶然ではないと考える．

3.2.2　MI 値

一方，MI 値は情報科学の相互情報量にもとづく指標である．相互情報量とは，ある記号が出現することが，別の特定の記号の出現を予測させる度合いを意味する．コーパス言語学では二つのタームの独立性を図る指標として使われている．この指標が大きい場合，二つのタームは独立ではなく，一方のタームが出現していればその共起語が現れる可能性が高いことになる．MI 値はコーパス言語学で以下の式で定義されている．なお対数の底は２である．

$$\text{MI 値} = \log_2 \frac{\text{共起関係}}{\text{共起語の期待値}} \qquad (2)$$

“ 共起語の期待値 ”は，その共起語のテキスト全体での頻度をテキストの総トークン数で割り，その値にスパン内の総数後を乗じた値である．

MI 値は低頻度語を強調する傾向があるため，テキスト全体での出現頻度は低くとも，専門語のようにテキストを特徴付けるタームを抽出するのに役立つ，MI 値が 1.58 を超える場合，二つのタームの間に共起関係がある．ただし，低頻度語を強調するため，極端に頻度が少ないタームを MI 値で評価するのは好ましくない．一般には，T 値の方がバランスの取れた指標として用いられる．

いずれにせよ，T 値も MI 値も，正規分布あるいはタームのランダム性を仮定している指標なので，その数値の大小を厳密に比較するものではなく，大まかな目安と考えるべきである．

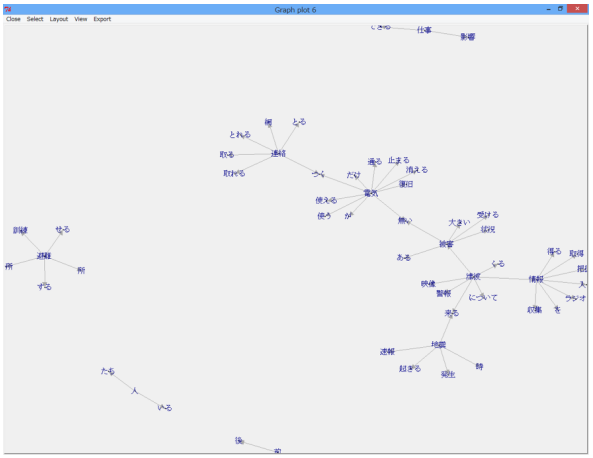

Fig. 2　collScores() で After の値が T ≧ 1.65 AND MI ≧ 1.65 AND Total ≧ 2 のもの（50 個表示）

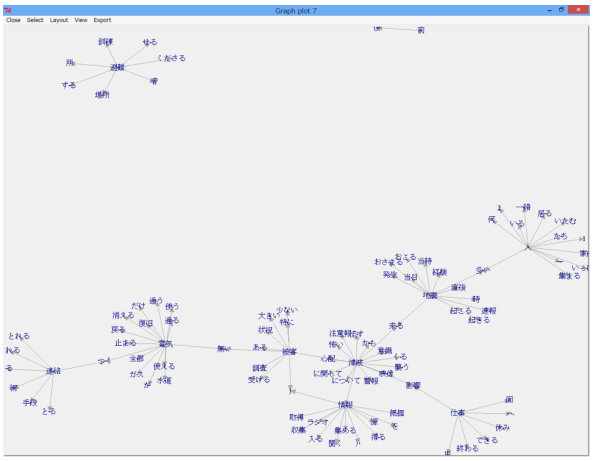

Fig. 3　collScores() で After の値が T ≧ 1.65 AND MI ≧ 1.65 AND Total ≧ 2 のもの（100 個表示）

RMeCab パッケージで T 値と MI 値を求める，そのためには collocate() 関数の出力であるオブジェクトを第 1 引数として，collocate() 関数で指定したノードを第 2 引数 node に，同じく第 3 引数 span に前後の語数を指定して collScores() 関数を実行した．

この実行結果から共起関係をグラフで表すことにし，ネットワーク分析に着目した．ネットワーク分析では，ネットワークマップ，グラフの構造に関する指標と統計量などを用いる．ネットワーク分析のフリーツールとしては Graphviz, Pajek, NetDraw,D PClus などがある．R には，ネットワーク分析のパッケージ sna，network, graph, igraph, inetowork などがある．

今回は，グラフ操作が便利である igraph を用いた．collScores() 関数で After 列かつ T 値が 1.65 以上かつ MI 値が 1.58 以上かつ頻度が 2 回以上のものを抽出しグラフにした．

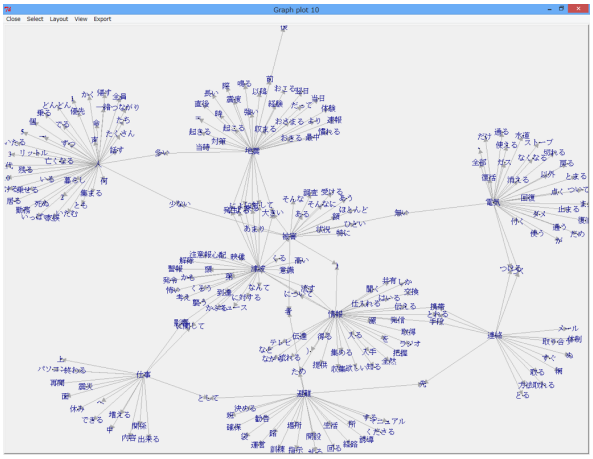

Fig. 4　collScores() で After の値が T ≧ 1.65 AND MI ≧ 1.65 AND Total ≧ 2 のもの（250 個表示）

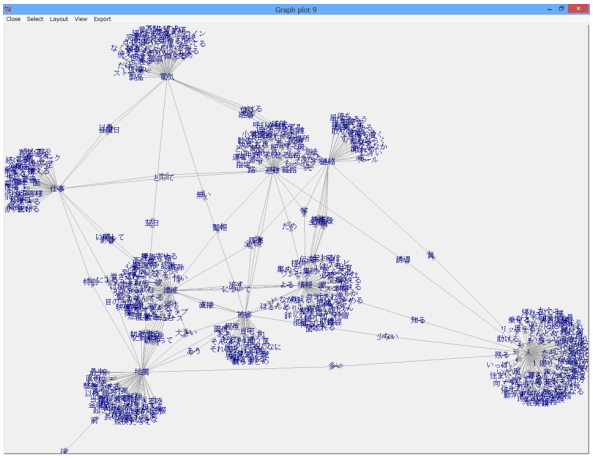

Fig. 5　collScores() で After の値が T ≧ 1.65 AND MI ≧ 1.65 AND Total ≧ 2 のもの（517 個表示）

4.　おわりに

語のネットワークマップとは，基本的には，文，あるいはテキストの中で用いられた語をノードとし，同時に用いられた場合は，語と語を線 (辺として) でリンクしたグラフである．共起パターンに前後の関係がある場合は有向グラフ，そう

ではない場合は無向グラフを用いる．また共起パターンの多少に関する重みを線の太さや長さで示す工夫が行われている．言語学においては，コーパスから共起パターンを抽出する．

図からは詳細な被害はわかりづらい．例えば「情報」→「仕入れる」となっているが肝心などこから仕入れているのか,「津波」→「第」第何波のことかわからない．バイグラムだけでなくトライグラムまで調べてみる必要があると感じた．

参考文献

1) 石田基広 (著) : R によるテキストマイニング入門，森北出版 (2008)
2) Jin Mingzhe: 統計的テキスト解析 (6) 語のネットワーク分析, https://www1.doshisha.ac.jp/ mjin/R/61/61.html (2016 取得)
3) Kazuhiro Takemoto: R+igraph, https://sites.google.com/site/kztakemoto/r-seminar-on-igraph—supplementary-information (2016 取得)