

2.2 テキストデータの数値化と景気動向と相関の強いキーワードの抽出

2.1 - 1 で述べた有意表現を抽出するテキストマイニングのプロセスを 2.2 - 1 ~ 2.2 - 7 の各項にて詳述する。

2.2 - 1 分析対象の絞り込み

景気ウォッチャー調査には、「景気の現状に対する判断理由等」と「景気の先行きに対する判断理由」の2種類がある。前者は、現状の景気についての意見を集めたものであり、表 2.2 - 1 に示す情報が含まれる。景気の状態判断の列は「良くなっている」「やや良くなっている」「変わらない」「やや悪くなっている」「悪くなっている」のいずれかが記録されており、その理由がコメントとして記録されている。後者は、景気の先行きに関する調査で表 2.2 - 1 に示す情報が含まれる。景気の先行き判断の列には、今後景気が「良くなる」「やや良くなる」「変わらない」「やや悪くなる」「悪くなる」のいずれかが記録されており、同じくその理由がコメントとして記録されている。

有意表現の抽出は、景気判断との相関性が高い単語を得るために、景気が良い時と悪い時の一方に偏って出現するものを抽出することが望ましいと考えられる。そこで、本プロセスでは、図 2.2 - 1 に示すように、現状の景気判断において「良くなっている」「やや良くなっている」または先行きの景気判断において「良くなる」「やや良くなる」と回答しているコメントのみを「景気が良いコメント」と定義し、「景気が悪いコメント」についても同様に定義する。この定義に従う分類を行い、テキストマイニングの対象とするコメント集合を得る。なお、景気状況が「変わらない」と記録されているものについては有意表現の抽出対象からは除外している。また、主だった回答等が存在しない理由により具体的な記述がないコメントは、事前に除去済みである。

本テキストマイニングでは、2010 年 1 月から 2014 年 12 月までの 5 年分のコメントを分析対象としている。図 2.2 - 1 には各工程により得られるコメント数を付記している。結果として半分程度のコメントについては分析対象から除外されているが、景気が良い／悪いのそれぞれで数万のコメントが得られているため、全体傾向を見ることができるコメント数が得られていると考えている。

表 2 . 2 - 1 景気の現状に対する判断理由等

報告月	景気の現状判断	地域	追加説明及び 具体的状況の説明
2011/3	良くなっている	北海道	エコカー補助金が・・・
2011/3	悪くなっている	南関東	東日本大震災の影響で・・・
：	：	：	：

表 2 . 2 - 1 景気の先行きに対する判断理由

報告月	景気の先行き判断	地域	景気の先行きに対する 判断理由
2014/3	悪くなる	東北	きっと駆け込み需要の反動で・・・
2014/3	やや悪くなる	沖縄	消費税増税が・・・
：	：	：	：

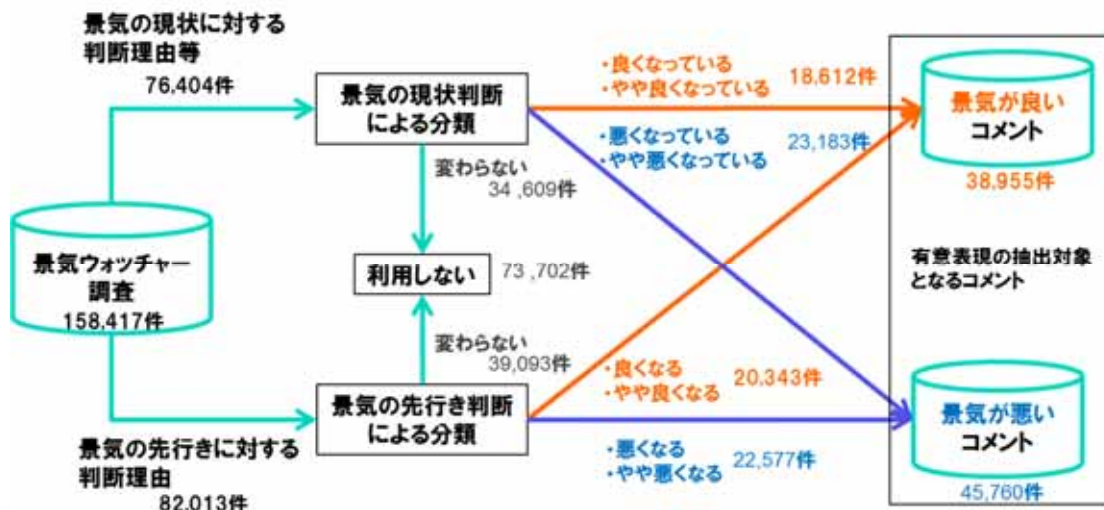


図 2 . 2 - 1 有意表現の抽出対象となるコメントの絞り込み方法

2.2.2 コメントからの単語の抽出

本ステップでは、前段のステップで絞り込んだコメント集合に対して、形態素解析エンジンを用いてコメント中の文を単語に分割する。これにより、各コメントに含まれる単語の表層形、品詞、原形¹を特定することができる。今回、単語分割を行う形態素解析エンジンはNECの研究所で開発した日本語言語解析エンジン JAna²を用いている。

例えば、「外出頻度が落ちたこともあり、灯油の取扱量も10%程度増加した。」という文は表2.2.2のように分割される。

表2.2.2 形態素解析結果の例

表層形	品詞	原形
外出	サ変	外出
頻度	名詞	頻度
が	助詞	が
落ち	動詞	落ちる
たこともあり	助動詞	たこともある
、	記号	、
灯油	名詞	灯油
の	助詞	の
取扱量	名詞	取扱量
も	助詞	も
10	数詞	10
%	単位	%
程度	助数詞	程度
増加	サ変	増加
し	サ変語尾	する
た	助動詞	た
。	記号	。

¹ 原形とは、動詞など活用により語尾が変化する単語の元の表現形式である。例えば、「景気が上向いている」と「景気が上向く」があった場合、前者は「景気／が／上向い／て／いる」、後者は「景気／が／上向く」と分割される。この時、表層形は「上向い」と「上向く」で異なる表現となるが、原形に変換することにより両者を「上向く」という同じ表現として扱うことができるようになる。

² JAna は、NEC が開発した日本語の形態素解析エンジンであり、130 万語規模の大規模辞書を持つことが特徴である。JAna は、NEC の機械翻訳、テキストマイニング、通訳等の製品に組み込まれて利用された実績をもつ。

2.2.3 単語の選択

本ステップでは、形態素解析の結果に対して品詞でフィルタリングを行い有意表現の候補になりうる単語を選択する。景気の良し悪しとは直接関係がない助詞等の付属語を除き、名詞や動詞等の自立語のみを選択する。こうすることで、有意表現の抽出結果が理解しやすくなると考えられる。図2.2.3は今回用いたフィルターの条件と単語が選択される様子の例である。図2.2.3では、黒枠で示される原形のうち、網掛けされていない単語が有意表現の候補として選択される。



図2.2.3 単語の選択方法

2.2.4 コメントから特徴表現の抽出

本ステップは、最初の2ステップで得られる景気の良いコメントと景気の悪いコメントの形態素解析結果を用いて、景気の良い時と悪い時の一方に偏って出現する表現を特徴表現として抽出する。抽出にあたっては、NECの特徴表現抽出ツール³を用いている。このツールの概要を図2.2.4に示す。このツールは、単語の並びから複合名詞やフレーズを特徴表現候補として抽出する機能と、その中で特徴表現らしさを特徴量スコアとして算出する機能から構成されている。特徴表現候補は、単語の並びとしての自然さを評価する指標を導入することで特定される。特徴量スコアは、特徴表現の候補がそれぞれ景気が良いコメントと景気が悪いコメントに何件含まれているかを求め、その独立性を評価するカイ二乗値として算出する。算出方法については資料1を参照されたい。

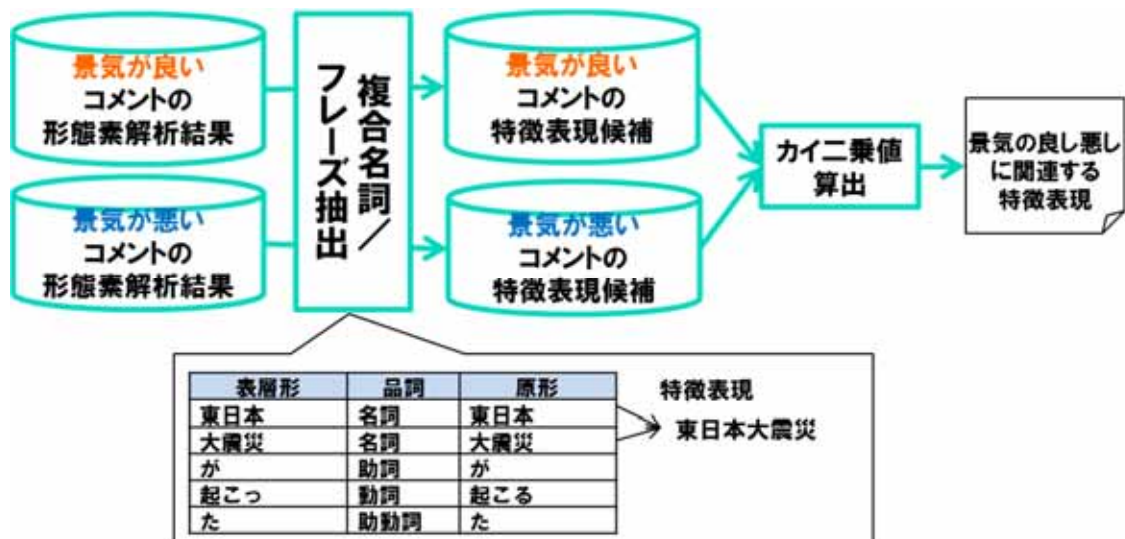


図2.2.4 特徴表現の抽出方法

実際に抽出した特徴表現の例を表2.2.4に示す。表中の「DF⁴ (POS)」は景気が良いコメントの数、「DF(NEG)」は景気が悪いコメントの数である。

³ これを利用することで、手作業による抽出に比べて辞書を構築するための時間を大幅に削減することが可能で、別タスクでの評価では約60%削減を確認している。

⁴ DF = Document Frequency。有意表現が登場するコメントの数。

表 2 . 2 - 4 特徴表現の例

景気が良い状態に含まれる特徴表現

No.	表現	DF (POS)	DF (NEG)
1	やや良く	2189	24
2	良くなる	3405	821
3	やや良くなる	1551	3
4	良くなっている	1203	97
5	増加傾向	911	118
6	増加傾向にある	583	37
7	やや良くなっている	382	6
8	動きが良い	401	26
9	増税前	1717	1034
10	傾向にある	1081	523
11	良くなってきている	270	13
12	消費税増税前	1604	979
13	好調に推移している	245	8
14	前の駆け込み	1428	869
15	前年を上回っている	240	17
16	良くなっていく	217	9
17	景気は良くなる	221	12
18	増税前の駆け込み	1294	791
19	増加が見込まれる	232	23
20	活発になる	179	14

景気が悪い状態に含まれる特徴表現

No.	表現	DF (POS)	DF (NEG)
1	悪くなる	61	2432
2	やや悪く	9	1576
3	やや悪くなる	0	1094
4	悪くなっている	24	1044
5	厳しい状況	153	1285
6	厳しくなる	34	616
7	東日本大震災	1588	3389
8	非常に厳しい	41	480
9	厳しい状況が続く	9	365
10	動きが悪く	20	392
11	やや悪くなっている	2	320
12	景気は悪く	11	340
13	非常に悪く	17	348
14	状況が続く	112	581
15	大震災の影響	465	1245
16	東日本大震災の影響	465	1244
17	動きが鈍く	22	342
18	少なくなっている	36	373
19	全くない	8	273
20	見通しは厳しい	1	249

2.2.5 同義語を統一化する処理

本ステップは、選択した単語と抽出した特徴表現において同義語が含まれる場合、それらを同じ表現に統一する作業を実施する。同義語の特定には同義語辞書を用いる。同義語辞書の作成は、NECの同義語辞書作成支援ツールを用いて行う。本ツールの処理概要を図2.2.5に示す。このツールでは、対象となる文書集合（今回の場合はコメントの集合）から複数の同義性の基準を用いて、多様な同義語の候補を生成することができる。同義性には、省略語（例：エアー／サスペンション→エアサス）、訳語一致（例：亀裂→クラック←ひび割れ）、文字列類似（例：ギアボックス＝ギヤボックス）、多段階生成（例：エンジン停止→engine stall→エンジnstール→エンスト）など様々なケースがあり、これらを組み合わせた総合的な基準で同義語の候補を取り出している。こうした同義語の候補には間違いもあるため、最終的には人によるチェックを行っている。

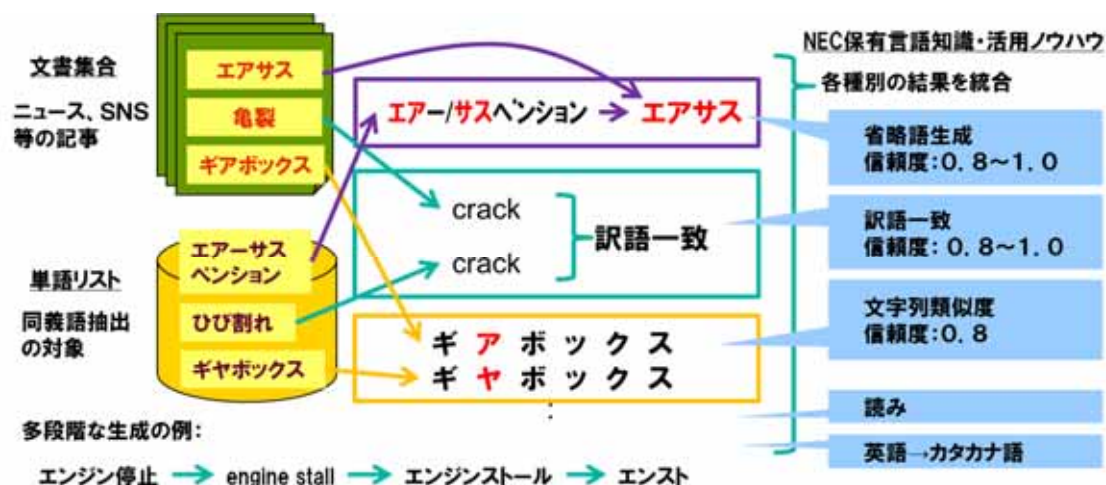


図2.2.5 同義語辞書作成支援ツールの処理概要

図2.2-5 同義語の統一化処理の概要を示す。この図は、「値段」と「価格」と「プライス」、「問い合わせ」と「問合わせ」が同義語として辞書化されており、「価格」と「プライス」を「値段」、「問合わせ」を「問い合わせ」に変換する処理を表している。

テキストマイニングにおいては、たまたまある単語や特徴表現が景気の良いコメントないし悪いコメントの一方に偏って含まれることがある。この現象は、特に母集合となるコメント数が少ない場合や、出現回数が少ない単語や特徴表現に起こりやすい。前者は取得するコメント数を増やすことが解決となる。後者は、同義語の統一化や後述する抽象化によりある程度解消することができる。図2.2-5 において「プライス」は、その単語だけを見ると景気の良いコメントだけにしか出てこないが、値段に関する同義語は景気が悪いコメントにも出現しやすいものと捉えられる。



図2.2-5 同義語の統一化処理

2.2.6 単語と特徴表現を抽象化する処理

本ステップでは、選択された単語と抽出された特徴表現の抽象化を行う。本抽象化においても前項で述べた少ない出現回数による弊害を避けることができる。例えば、図2.2-6では、各単語だけを見ると「ゲームセンター」は景気が悪いコメントにしか存在しないため、このまま分析すると景気が悪いときに出やすい単語といった結果になる。しかしながら、これらを「娯楽施設」という抽象概念でまとめることで分析対象となるコメントが増え、違和感のある分析結果を避けられる。抽象化の仕方によっては、本来景気の動向と関係する単語が抽象概念に埋もれてしまい見えなくなる可能性もあるが、今回は、全体傾向を見ることを優先し、本分析専用の抽象化辞書を人手で構築している。抽象化辞書は1つの単語が1つの抽象概念に含まれるように構築しており、複数の抽象概念に含まれることはない。また、「食品」という抽象概念の中に「生鮮食品」と「保存食品」という抽象概念を含むような抽象概念を階層化する手法も考えられるが、まず抽象化処理の導入効果を確認することを優先し、今回の分析ではこの階層化は行っていない。

なお、本資料においては、抽象概念は記号「@」を左右に付与することで表現している。本分析で作成した抽象化辞書を資料2に示す。今回は複数人で一から辞書を構築したため、同一の基準で抽象化が行えなかった部分もある。今後は日経シソーラス⁵などを加味することが解決の一つになると考える。



図2.2-6 抽象化の例

⁵ 株式会社 日本経済新聞デジタルメディアが1982年から作成している新聞記事検索のための用語集。 http://telecom21.nikkei.co.jp/help/contract/price/23/help_KIJI_thes.html

2 . 2 - 7 有意表現の抽出

最後に、本ステップにより、ここまでの処理過程を経た単語と特徴量の各々について、それらが「景気が良いコメント」と「景気が悪いコメント」に何件ずつに含まれているかをカウントし、その独立性をカイ二乗値（算出方法は資料 1 参照）として算出する。この値をスコアとして、スコアが高い 400 語を有意表現としてテキストマイニングの出力結果とする。実際に得られた有意表現のうちスコアの高いトップ 40 語を表 2 . 2 - 7 に示す。また、全 400 語については資料 3 に示す。

表 2 . 2 - 7 有意表現の例（トップ 40 語）

有意表現	DF(POS)	DF(NEG)	スコア (カイ二乗値)
@悪い状態@	3996	18254	9540.8
@減少傾向@	3121	16333	9114.0
@良化傾向@	15945	6568	7618.0
@増加傾向@	20072	11114	6711.6
@良い状態@	13753	5776	6103.2
@悪化傾向@	544	7673	5676.4
@強い程度表現@	5664	12975	2340.0
@求人@	3754	1500	1462.6
@弱い程度表現@	8453	5593	1366.4
@少ない@	919	3711	1346.8
@影響@	3227	7606	1311.6
@ない@	1535	4711	1244.2
予想-サ変	5201	3175	971.3
@不安@	470	2300	970.7
@厳しい@	113	1292	828.0
@景気悪化@	101	1145	730.5
@売上増加@	1055	209	725.7
@単価減少@	352	1647	663.6
@円高@	317	1488	599.8
@活発状態@	655	103	503.3

表 2 . 2 - 7 有意表現の例（トップ 40 語）の続き

有意表現	DF(POS)	DF(NEG)	スコア (カイ二乗値)
@ 求人増加@	552	54	499.9
@ 需要@	8158	6906	492.6
@ 需要増加@	2795	1764	455.5
@ 新しい@	1986	1096	438.6
@ 資源@	463	1519	418.2
東日本-名詞	1855	3785	417.0
@ 東日本大震災@	1840	3762	416.9
@ 需要減少@	235	1066	414.7
@ 大震災@	2069	4077	404.9
@ 不透明@	242	1067	404.7
値段-名詞	898	2237	394.0
@ 少しずつ@	1515	777	383.8
競合-サ変	413	1368	380.6
固い-形容詞	46	573	373.1
懸念-サ変	378	1270	359.4
@ 程度表現@	155	818	357.9
見込む-動詞	1642	920	348.7
状態-名詞	627	1693	345.1
@ 店舗（小規模）@	1926	3716	341.5
増税前	1800	1070	334.9