投資工学開発室

（審査確認番号：他2018-176）

吉野貴晶 のクオンツ トピックス：*NO6*

AIによるテキスト情報の解析（極性辞書の作成編）

単語から「景気に対する見方」を評価する

- 連載形式でAI（人工知能）と投資手法の関係性を紹介。
- 単語の持つ強弱感に着目して文章評価モデルを作成します。

最近、AI（人工知能、以下AI ）に関連するニュースが増えています。投資の分野でも研究開発が盛んに行われており、実際に投資手法として利用可能な段階まで進展しています。本レポートでは、AIと投資手法の関係性をご紹介したいと思います。

今回はAIを利用して、「景気に対する見方」を単語に着目して測定する手法をご紹介します。

1. AIにテキストのポジティブネガティブを判定させる

まずは前回のレポートに続き、テキスト情報の利用についてです。前回は、景気ウォッチャー調査を元データとして、主に手順の①、②に当たるデータ前処理をご紹介しました。今回は、その前処理済みデータを利用して、文章を評価するAIモデルの作成過程をご紹介します。

図1.テキストデータの活用アプローチ

①：テキストデータを取得
②：テキストデータを綺麗な状態に整形
③：AIが読み取れるようにデータを加工（数値情報に変換）
④：単語 または 文章単位でスコア化（AI）
⑤：算出したスコアと投資対象との関連性を確認（スコアとTOPIXとの関係 等）
⑥：実際に投資してリターン獲得

1-1.「文章の評価」を定義する

AIモデルを作成する場合、まず始めにAIに、「どのデータ」で「何をさせるか」を決める必要があります。前回のレポートから引き続き、景気ウォッチャー調査のデータを利用します。このデータから、「文章の評価」をするAIモデルを考えたいと思います。モデル作成にはもう一段掘り下げる必要があります。それは、「文章の評価」の定義です。例えば、TwitterのようなSNSデータを元にする場合は、コメントの内容が「嬉しい」のか「悲しい」のかを判定した方が良いかもしれません。また、企業に届く商品のレビューの場合は、「満足」なのか「不満」なのかを判定すべきかもしれません。このように、評価の尺度はデータや目的により変化します。

1-2. 評価尺度は「景気に対する見方」

対象とする景気ウォッチャー調査のデータは、回答者の景気に対する見方を測る経済統計ですので、「景気に対する見方」を評価尺度にして判定するAIモデルを作成したいと思います。景気に対する見方が良ければプラス（ポジティブ）、悪ければマイナス（ネガティブ）とする判定モデルを目指します。

極性辞書とは？

1-2. 極性辞書

文章のポジティブ、ネガティブを判定させる手法として、文章に含まれる単語に着目する方法があります。これはポジティブ（ネガティブ）な文面にはそれ特有の単語が含まれるはず、という考えに基づいています。そのような特定の単語を集めたリストを極性辞書と呼びます。簡単な例としては、「嬉しい」という単語は感情面での極性値はポジティブ、「悲しい」はネガティブ、といった具合です。また、単にポジティブかネガティブかの情報だけでなく、ポジティブ及びネガティブ度合いを数値化（以下極性値）する場合も有ります。

1-3. 極性辞書の利点

極性辞書は単語レベルでポジティブかネガティブかが分かるので、直感的に分かりやすいというメリットがあります。

1-4. 対象に併せた極性辞書の必要性

極性辞書ですが、既に作成されたものが公開されています。日本語の極性辞書としては、東北大学 乾・岡崎研究室の日本語評価極性辞書（Appendix A-5.参考文献 1及び2）が有名です。但し、モデルの評価尺度と辞書の評価尺度が一致しているか？という点を考慮する必要があります。今回は「景気に対する見方」を判断したいので、既存の感情等をベースにしている極性辞書ではやや不整合かもしれません。実際、米国での研究事例では、金融テキストの評価には金融に特化した極性辞書が利用される事例が多くなっています。

図2.日本語評価極性辞書（一部抜粋。Appendix A-5.参考文献 1及び2）

日本語評価極性辞書（用言編）

ポジ（経験）	あこがれる
ポジ（経験）	あじわう
ポジ（経験）	かなう
ポジ（経験）	したう
ポジ（経験）	すがすがしい
ネガ（経験）	あがく
ネガ（経験）	あきらめる
ネガ（経験）	あきる
ネガ（経験）	あきれる
ネガ（経験）	あきれる た
ネガ（経験）	あせる

日本語評価極性辞書（名詞編）
p：ポジティブ、n：ネガティブ、e:ニュートラル

あく抜け	p	～する（出来事）
あこがれ	p	～がある・高まる（存在・性質）
あざやか	p	～である・になる（状態）客観
あたたかさ	p	～がある・高まる（存在・性質）
あいまい	n	～である・になる（評価・感情）主観
あからさま	n	～である・になる（評価・感情）主観
あきらめ	n	～がある・高まる（存在・性質）
あせも	n	～である・になる（状態）客観
あだ	n	～である・になる（評価・感情）主観
ありがた迷惑	n	～である・になる（評価・感情）主観
あいさつ	e	～する（行為）
あいだ	e	～である・になる（状態）客観
あいつら	e	～である・になる（状態）客観
あいまいさ	e	～がある・高まる（存在・性質）
あたしたち	e	～である・になる（状態）客観

1-5. 極性辞書の作成

今回は日本語をベースにした、「景気に対する見方」の極性辞書を作成したいと思います。単にポジティブとネガティブの判定では無く、度合もスコア（極性値）で表現したいと思います。極性辞書の作成にはシンプルなニューラルネットモデル（後述）をベースにします。

不均衡データ

2. データチェックと調整

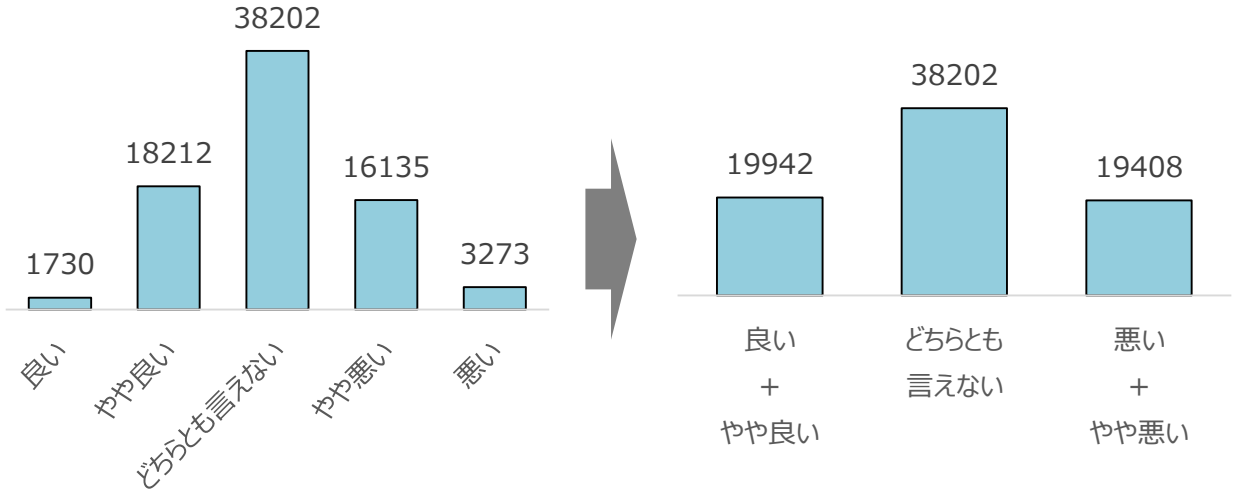
AIによる分析や学習を開始する前の必須事項として、「データの確認」があります。これは、データの特性を確認するステップになります。事前にテキストデータの前処理は実施済みなので、今回はデータの分布にフォーカスを当てたいと思います。

2-1.不均衡データ

今回用いるデータに回答例の偏りが無いかをまず確認します。以下の図3(左)が、良いから悪いまでの5パターンにおける回答数です。回答数に大きくバラつきがあることが分かるかと思います。このようなデータを不均衡データ、と呼びます。この不均衡データですが、何も考えずにAIに学習させると、偏ったデータについて学ぶため、予測結果にも偏り（バイアス）が発生する場合があることが知られています。

両端の良いと悪いはデータ数が少な過ぎてAIによる学習に不安があります。ですので、今回は図3(右)のように「良い＋やや良い」と「悪い＋やや悪い」をまとめることにします。このまとめ後のデータを見ると、「良い＋やや良い」が19942件、「悪い＋やや悪い」が19408件であり、差分は538件です。ほぼ同じデータ数なのでこのままでも問題ないとは思いますが、念のためデータ数の調整を実施します。

図3. 回答例の分布と調整

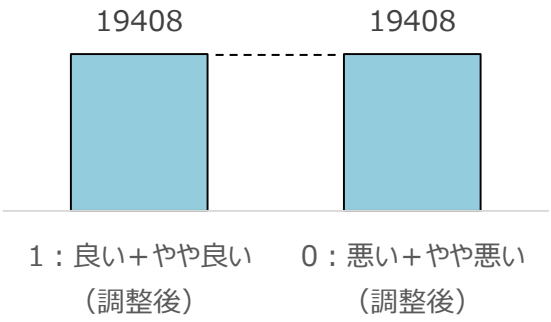


2-2. アンダーサンプリング

不均衡データに対する調整は、過去色々な手法が考案されており、突き詰めると奥が深い領域ですが、今回はシンプルな方法を試します。アンダーサンプリングと呼ばれる手法で、回答例の数が少ない方に、多い方の数を調整して合わせます。図3だと、回答例が少ない「悪い＋やや悪い」のデータ数に、回答例が多い「良い＋やや良い」を合わせます。今回は単純にランダム抽出します。

今回は考慮しませんが、アンダーサンプリングの懸念点としては、元データの特徴を正しく再現できるように抽出できるか、という問題があります。今回はあまり削減されるデータ数が多く無いので問題はないと思われますが、かなりの割合のデータを削減する場合は工夫が必要になります。例としては、元のデータをクラスタリング（分類）することで特徴を持ったグループに分け、そのグループ毎にデータを一定の割合で抽出する、等が考えられます。

図4. アンダーサンプリング後の分布



ニューラルネットモデル

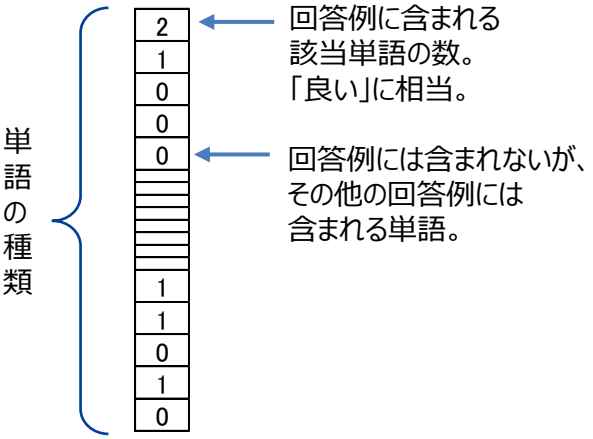
2-3. 文章を頻度ベクトルで表現する

今回は極性辞書を作成するためにニューラルネットモデル を利用します。モデルへの入力データは、文章に出現する単語の頻度ベクトルになります。この頻度ベクトルについて解説します。まず、ある文章中に出てきた単語レベルで考えます。図5(左)にある文章を例にとると、「単価」はこの文章に1回出現します。一方、「良い」は2回になります。このように、ある単語が文章中に何回出てくるか、を表すのが頻度ベクトルになります。具体的には、図5(右)のように、1次元の数の並びで表現されます。この頻度ベクトルを後述するニューラルネットモデルのインプットデータとして利用します。

図5. 頻度ベクトルへの変換

＜ 前処理済みデータ 実例＞

単価 良い 来客 数 伸びる いる 新車 納車 後 用品 購入 車検 入庫 台数 増加 要因 オーディオ 価格 下がる いる 傾向 景気 良い



3. モデル構築

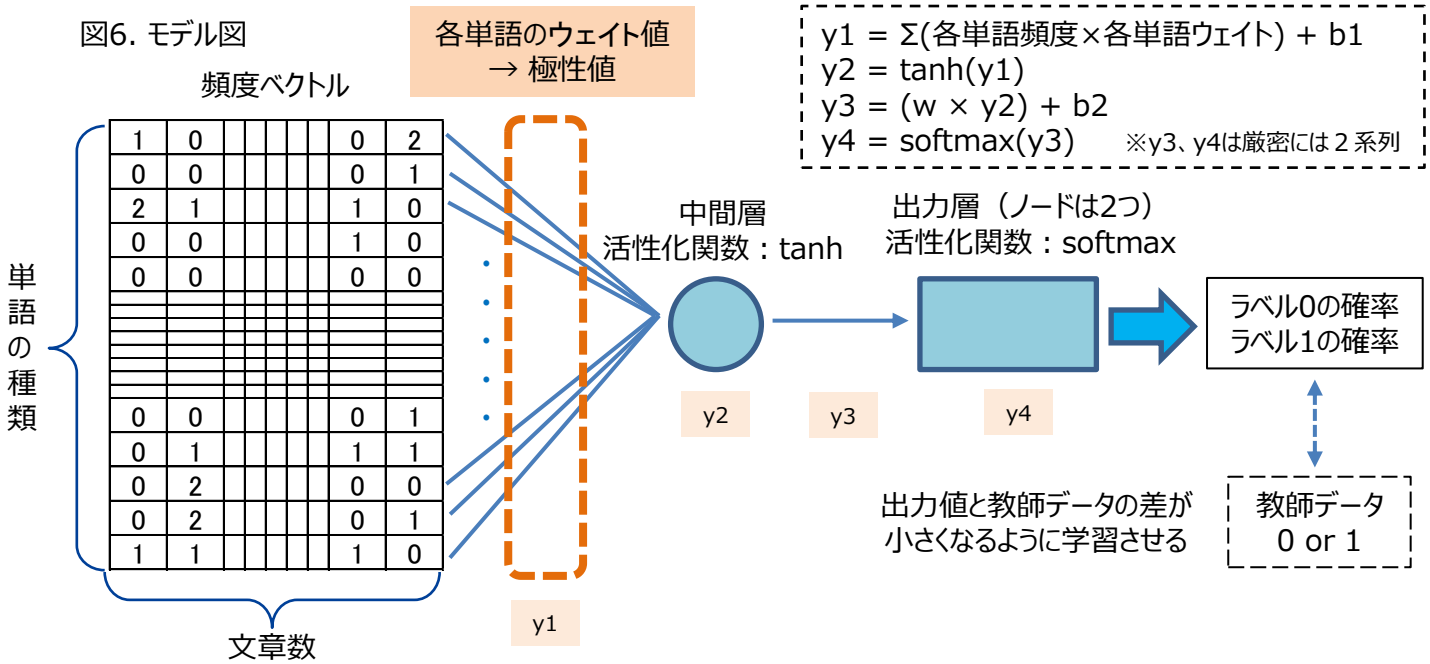
3-1.極性値を得るためのニューラルネットモデル

極性辞書を作成するためにシンプルなニューラルネットモデルを利用します。今回用いたモデルは以下の図6です。入力データに各文章の頻度ベクトル、中間層が1つ、最終的な出力値はポジティブな文章なら1、ネガティブなら0、とした2値分類になります。このモデルは、インプットデータである頻度ベクトルが最終的な文章のセンチメントを説明可能である、と言う前提に立っています。さらに、頻度ベクトルは単語レベルの情報なので、文章に含まれる単語がポジティブ・ネガティブの決定要因である、という考えに基づいています。

3-2. 極性値はどこに現れるか？

単語の極性値情報はインプットにも出力にもこのままでは現れません。どこから取得するかと言うと、図6においてオレンジ色の太い破線で囲った部分になります。ここの結合重み（ウェイト）が極性値だと定義します。

図6. モデル図



AIモデルを実務で扱う際に気をつけるべき点は？

4. オーバーフィッティング（過学習、Appendix A-1参照）

さて、「モデルの基本設計も決まったし、すぐにAIに学習を」となりがちですが、重要な点を手当てしたいと思います。一般的にオーバーフィッティング（過学習）と呼ばれる事象です。オーバーフィッティングとは、学習用に与えられたデータに対して過剰に適合してしまい、そのデータセットにしか高い性能を発揮できない状態になることを指します。

4-1. テストデータでの性能チェック

作成したモデルがオーバーフィッティングしているかどうかを確認する簡単な方法としては、テストデータでの性能チェックが一般的です。モデルの学習を開始する前に、データセットを学習用とテスト用に分割します。まず学習用データでモデルを学習させます。その学習済みモデルを使い、学習用とテスト用のデータそれぞれに対してモデルによる正答率を算出します。この２つの正答率に大きな乖離がある場合、そのモデルはオーバーフィッティングしている可能性が高い、と言えます。今回のモデルでは学習用とテスト用を70%、30%になるように分割して正答率を比較しました（結果後述）。

5. 初期値依存と再現性（Appendix A-2参照）

一般的に現在AIと言われている手法には、初期値依存の問題があります。簡単にいうと、毎回学習させるたびに結果が変わる、ということです。これは初期値と呼ばれる、最初に与えるパラメーターに依存するためです。毎回結果が変わることを、再現性が低い、と表現することもあります。この問題は、初期値（ランダムシード）を固定することで回避できる場合があります（GPU依存等で一部例外あり）。しかし、初期値を固定すると言うことは、モデルが取りうる多様性を排除しているともいえ、悩ましいところです。

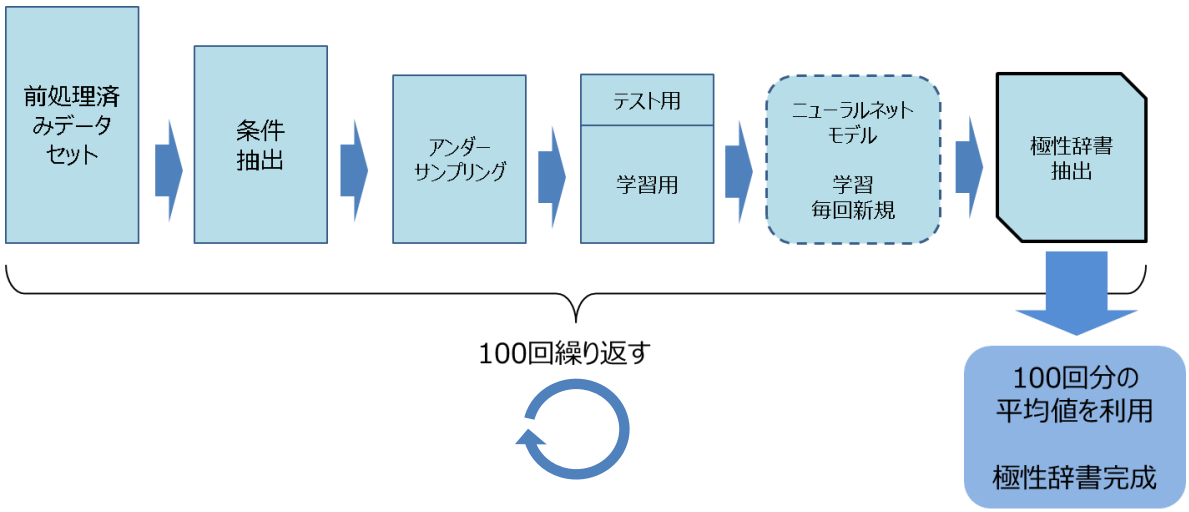
5-1. 試行回数を増やす

今回のモデルでは、初期値依存は受け入れ、解決策としてモデルを複数作成し、結果を平均した値を利用することにします。具体的には、学習モデルを100個作成し、その100個における各単語の極性値の平均を求めます。

6. モデル学習プロセスの全体図

ここまでの内容をまとめると、モデル全体での学習プロセスは以下の図のようになります。アンダーサンプリングも含めて複数回の繰り返し処理になりますので、厳密に言うと初期値依存問題のみではなく、アンダーサンプリング時に抽出されるデータの形状にも依存しています。

図7. モデル全体での学習プロセス



極性辞書の確認

7. 学習結果と極性辞書（Appendix A-3及び4参照）

得られた100個のモデルにおける平均正答率は図8になります。比較的高い正答率、かつ学習用とテスト用データそれぞれに対する結果の差異が小さいことから、オーバーフィッティングしている可能性は低いといえるかと思います。

図8. 正答率の比較

モデル正答率	学習用データ	テスト用データ	差異
	86.4%	83.6%	2.8%

7-1. 獲得した単語とその極性値

目的である極性辞書を確認します。図9が得られた単語と極性値、ならびに100個のモデルでそれぞれ算出された極性値の分散になります。極性値の符号は、ポジティブだと思われる単語なら正に、ネガティブなら負になるように調整しています。極性値が高くポジティブだと思われる単語のリストは、「明るい」、「改善」、「増加」など、景気や経済活動がポジティブな場合に使われそうな単語になっています。対してネガティブだと思われる単語は、「鈍化」、「激減」、「悪化」など素直にマイナスなイメージの単語が選ばれています。また、極性値の分散（極性値の各モデルにおけるブレ幅）を見てもスコアに対して十分に小さく、安定性が確認されます。

図9. 今回得られた極性値（抜粋）

スコア上位	スコア	分散
明るい	0.42	0.03
上向き	0.40	0.03
改善	0.39	0.03
増加	0.38	0.01
活発	0.37	0.03
微増	0.35	0.03
上向く	0.35	0.03
上回る	0.32	0.02
恵まれる	0.30	0.03
順調	0.30	0.03

スコア下位	スコア	分散
激減	-0.57	0.03
鈍化	-0.54	0.03
悪化	-0.46	0.02
下回る	-0.45	0.02
減少	-0.41	0.01
減る	-0.40	0.02
下降	-0.38	0.03
割る	-0.36	0.03
鈍い	-0.36	0.03
悪い	-0.35	0.02

8. 終わりに

今回のレポートでは、景気判断に特化した極性辞書の作成を試みました。また、その際に考慮すべき点についても簡単ではありますが触れてみました。このモデルと極性辞書があれば、文章の景気に対する見方を測定できます。次回レポートでも引き続きAIをテーマに取り扱う予定です。　　（次頁：Appendix）

～執筆者の紹介～

吉野貴晶（写真：右）
「日経ヴェリタス」アナリストランキングのクオンツ部門で16年連続で1位を獲得。ビックデータやAIを使った運用モデルの開発から、身の回りの意外なデータを使った経済や株価予測まで、幅広く計量手法を駆使した分析や予測を行う。

髙野幸太（写真：左）
ニッセイアセット入社後、ファンドのリスク管理、マクロリサーチ及びアセットアロケーション業務に従事。17年4月に投資工学開発室に異動後は、主に計量的手法やAIを応用した新たな投資戦略の開発を担当する。

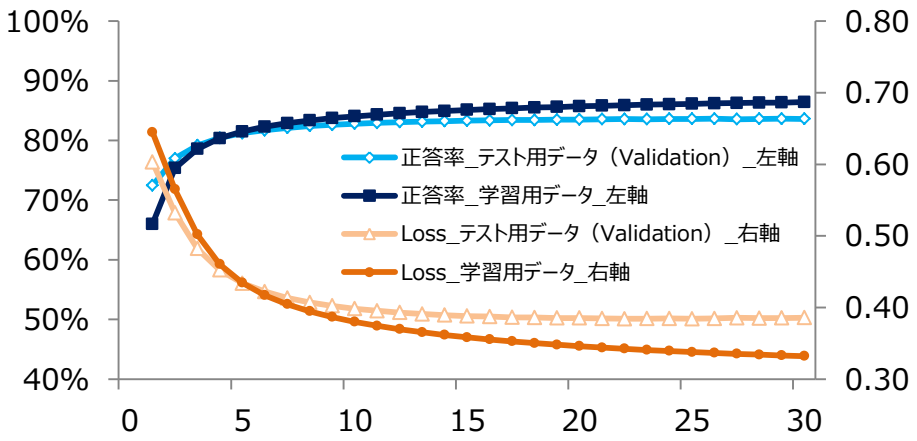
Appendix

Appendix

A-1. 学習時の正答率とloss値の推移

学習過程における、学習用データとテスト用データそれぞれに対する正答率とLossは以下の通り。学習回数が15回を過ぎる辺りから、テスト用データにおける正答率がほぼ上昇しなくなっています。一方、Lossは学習用データで緩やかですが下がり続けており、このまま学習回数を増やすとオーバーフィッティングに陥る懸念があります。

図10. 正答率とloss値の推移



A-2. ループ処理をアンダーサンプリングまで含める意味

今回は初期値依存問題等への対策として、試行回数を100回実施しています。このループ処理にアンダーサンプリングまで含める事で、アンダーサンプリングの過程で極性を持った単語が排除される可能性が少なくなることも狙っています。

A-3. 総単語種類

用意したデータセット全てに含まれる対象単語数は重複無しで11983種類。そこから20事例以上に含まれる単語という条件のもと、2254種類まで絞って極性値を算出しています。

A-4. 極性値の符号調整

学習済みモデルから極性値を取得すると、符号が反転している場合が多発します。これは、分布そのものは似通っているが、極性値に係る係数が中間層から出力層の間にあり、この係数がマイナスの値で学習される場合があるためです。マイナスとマイナスの掛け算がプラスになることから、極性値の符号の傾向が反転した結果も整合性を取ることが可能となります。しかし、符号が反転しているままでは平均値を取れないので、修正のために「鈍化」と言う単語の極性値に着目しました。明らかにネガティブと判定されそうなこの単語に対して、極性値がマイナスで算出されているならそのまま、プラスなら全体にマイナス1をかけて符号を揃えました。

A-5. 参考文献

1.小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一
意見抽出のための評価表現の収集 自然言語処理 Vol.12 No.3 pp.203-222, 2005.

2.東山昌彦, 乾健太郎, 松本裕治
述語の選択選好性に着目した名詞評価極性の獲得 言語処理学会第14回年次大会論文集 pp.584-587, 2008.

3. 伊藤友貴, 坪内孝太, 山下達雄, 和泉潔
経済テキストデータを用いた極性概念辞書構築とその応用 第18回人工知能学会 金融情報研究会資料, 2017.

4. 山本裕樹,松尾豊景
景気ウォッチャー調査の深層学習を用いた金融レポートの指数化
The 30th Annual Conference of the Japanese Society for Artificial Intelligence, 2016.

ニッセイ アセットマネジメント