吉野貴晶 のクオンツトピックス：*NO7*
AIによるテキスト情報の解析（経済テキストの指数化）

政府発表の経済テキストを指数化する

- 連載形式でAI（人工知能）と投資手法の関係性を紹介。
- 極性辞書を利用して経済テキストを実際に指数化します。

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。

今回のテーマは経済テキストの指数化です。実例として、月例経済報告を細分化して指数化します。

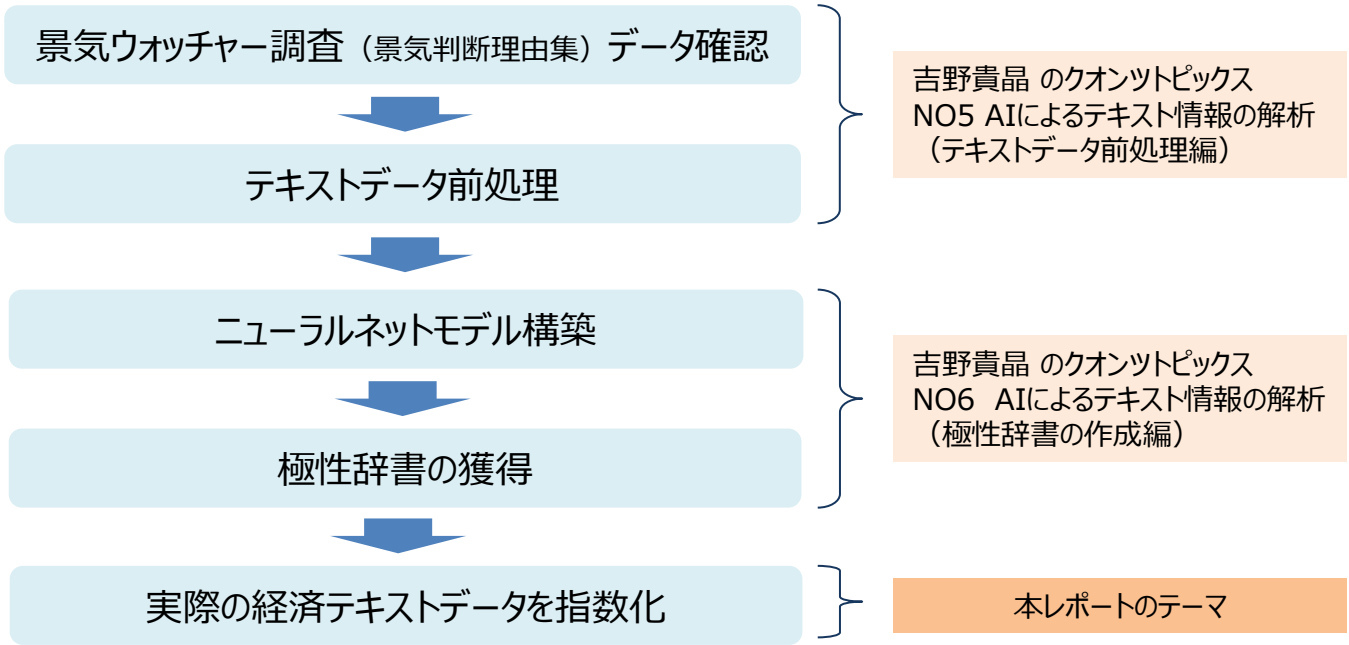
1. 前回レポートまでの復習

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、文章の指数化を実施しするために必要な極性辞書を作成しました。今回は、この極性辞書を利用して、手順④に該当する経済テキストデータの指数化をご紹介します。

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 文章単位で指数化（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

図2.今までの作業の振り返りと今回のテーマ



月例経済報告とは？

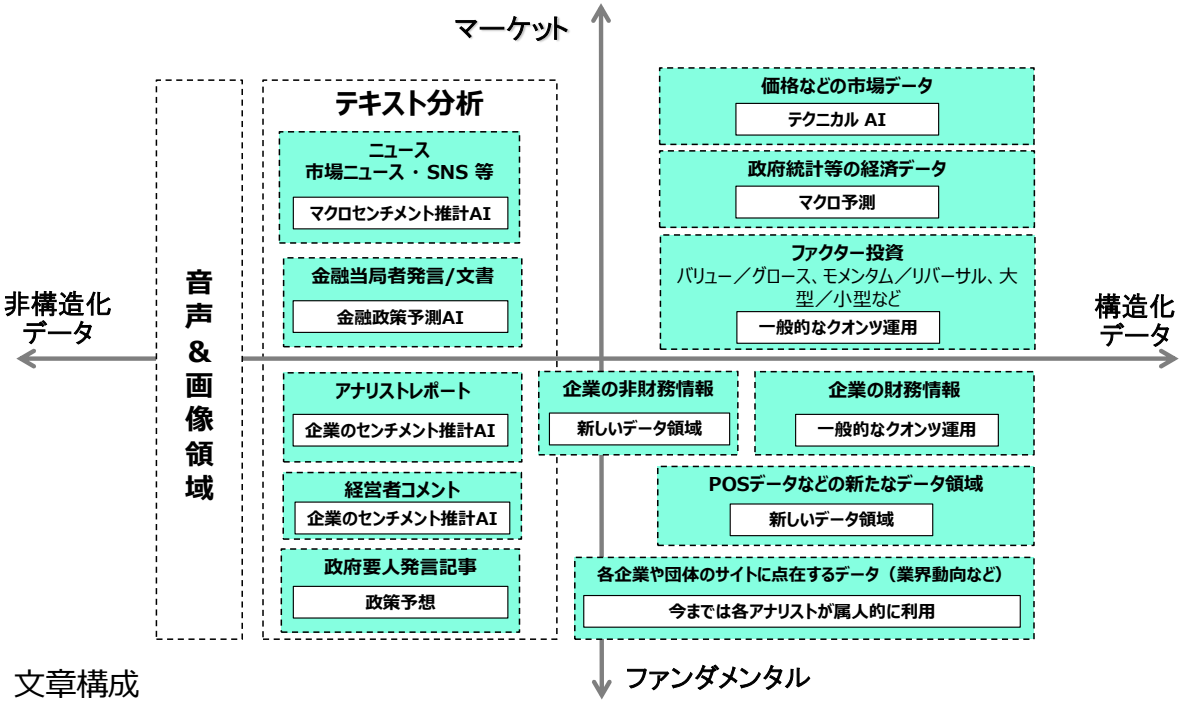
2.月例経済報告

今回は月例経済報告の文章を指数化します。月例経済報告とは、毎月、政府が発表する報告書です。政府による景気への見解が文章で示される資料になります。

2-1. データの分類確認

月例経済報告のデータとしての特徴を確認します。図3は、データの特性を切り口に、データ領域（どのデータを使うか）とAIモデル（どのような結果を目指すか）をマッピングしたものです。今回のデータはテキスト情報なので非構造化データであり、左側に該当します。上下の軸であるマーケットデータとファンダメンタルデータの観点では、切り口にも依りますがマーケットデータ寄りといえるかと思います。結果、図3では左上部分に分類されます。極性辞書を作成する際に利用した景気ウォッチャーの景気判断理由集も同じ領域なので、辞書作成からテキストの評価まで領域の一貫性は保たれているといえます。

図3.各情報の整理学、AIによるモデル化のイメージ



2-2. 文章構成

月例経済報告の文章構成を見てみましょう。まず１ページ目である表紙には、もっとも重要な景気全体の基調判断と先行き（見通し）が記載されます。これが政府の景気全体に対する見解になります。２ページ目には、前回の月例経済報告との比較および変更箇所が記載されます。どこが変更されたかを把握する上で便利です。３ページ目以降では、景気の基調判断において重要な各構成要素についての記載が続きます。この部分を読むと、政府がどの指標を見つつ景気の状況を判断しているかを垣間見ることができます。

図4.月例経済報告の構成

月例経済報告
（毎月1回発表）

表紙：景気の現状判断、先行き

2ページ目：各項目の前月からの変更点まとめ

3ページ目以降：各項目の詳細
個人消費、設備投資、住宅建設、公共投資、
輸出、輸入、貿易･サービス収支、生産、企業収益、
雇用情勢、国内企業物価、消費者物価、世界景気 等

極性辞書を使って経済テキストを指数化する

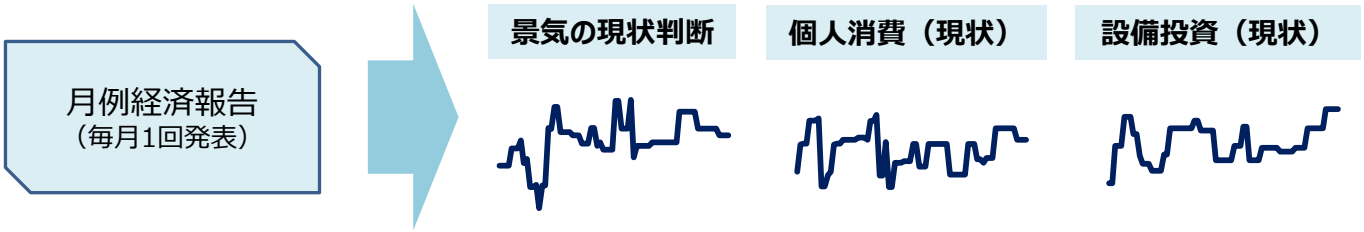
3. 月例経済報告指数を作る

前処理済みのテキストデータを利用して、月例経済報告から指数を作成します。
[過去レポートNo5及びAppendix A_1]

3_1. 先行研究

景気ウォッチャーを利用してモデルを作成し、月例経済報告をスコア化する試みには先行事例[参考文献4]があります。この事例では、LSTMというAI手法を利用し、毎月の月例経済報告に一つのスコアを付与しています。一方、本レポートでは、結果が直感的に分かりやすいという理由から、前回レポートで作成した極性辞書を利用して文章の指数化を試みます。また、表紙に記載される景気全体の基調判断に加え、その構成項目毎についてもスコア化を実施することで、一つの月例経済報告から複数のスコアを作成し、時系列データにすることを目指します。

図5.月例経済報告から極性辞書指数を作成



3_2. 極性辞書を使って指数化「極性辞書指数」

極性辞書に存在する単語が月例経済報告の文章中にある場合、該当する極性値を代入していきます。辞書に無い単語は全て0点とします。最後に文章中の極性値を足し上げれば、その文章の指数値が得られます。2_2 の文章構成にあるように、基調判断に係る項目を全てスコア化しました。（結果は後述）

図6.極性辞書指数値作成までの流れ

元の文章を分かち書き	景気	は	緩やか	に	回復	し	て	いる	
名詞、動詞、形容詞のみ抽出 & 原型化	景気		緩やか		回復	する		いる	
極性値を代入	-0.07		0.19		0.18	-0.00		-0.01	指数値 0.28

3_3. 比較対象として「人間判断指数」の作成

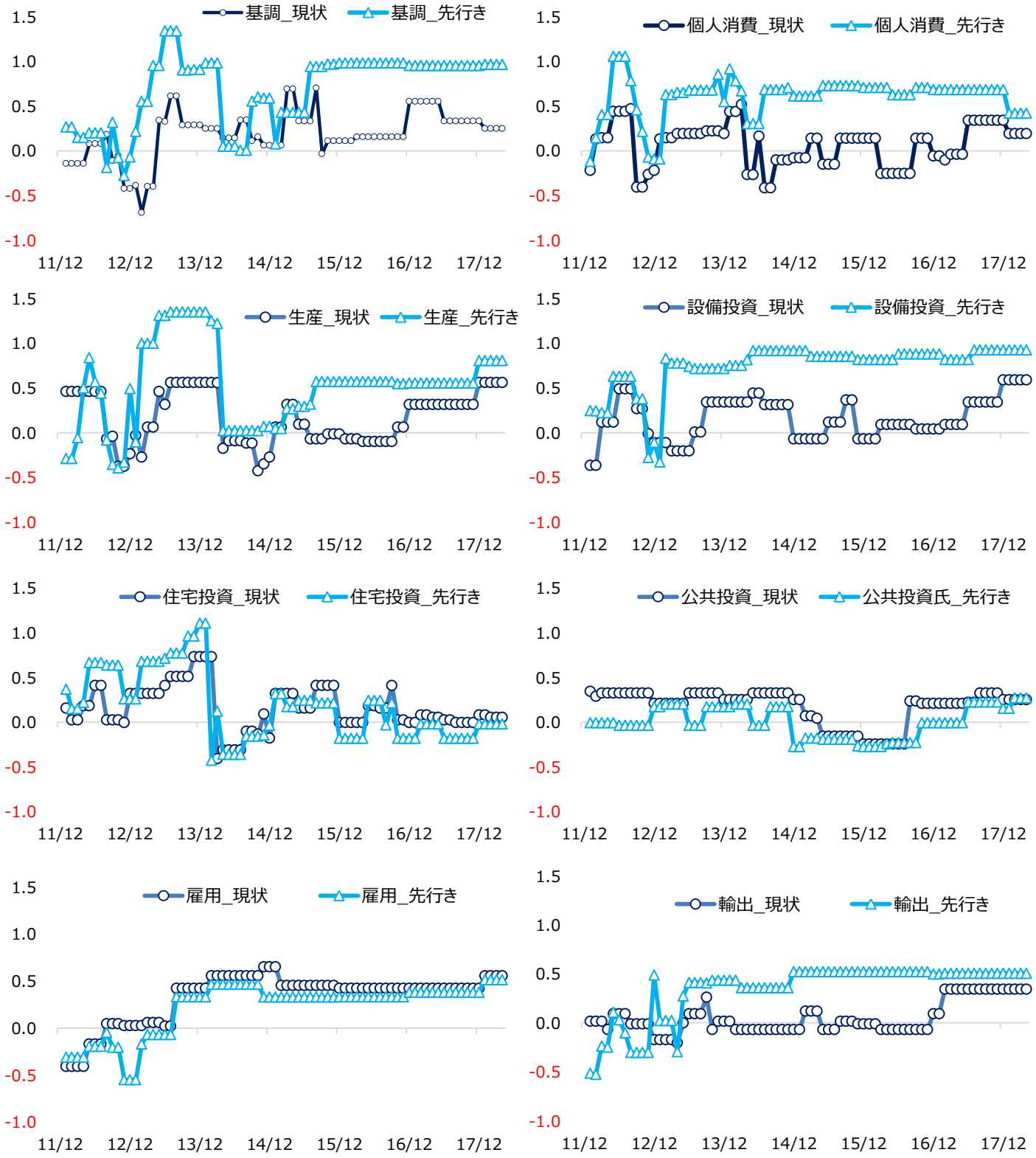
極性辞書指数について、どの程度妥当なのか検証するために比較対象を準備します。比較対象としては、人の感覚を利用します。具体的には、ある月の文章が前月の文章に対してポジティブかネガティブかを人に判断させます。起点の月をゼロとして、前月対比でポジティブならプラス1、ネガティブならマイナス1を前月までの累積値に加えることで指数値を作成します。これを便宜上「人間判断指数」と呼ぶことにしたいと思います。なお、現状判断部分については、月例経済報告が発表されるたびにニュース等で上方修正か下方修正かが報道されています。今回は現状判断は出来る限り報道から政府の基調判断を追って指数化し、明確に発表されなかった月は、筆者の判断で埋めています。なお、月例経済報告における政府の基調判断の変化を利用した累積指数作成の先行研究には[参考文献 5]があります。

極性辞書指数

3_4. 極性辞書指数の可視化

さて、実際に得られた極性辞書指数を図示してみましょう。以下のグラフは、月例経済報告の各項目毎における現状と先行きについて、前回作成した極性辞書を用いて指数化したグラフになります。

図7.極性辞書指数値の推移



AIと人間の判断に差異はあるか？

4.極性辞書指数と人間判断指数の比較と検証

AIで作成した極性辞書指数が、人間判断指数と感覚的に合うのか検証します。

4_1. AIの判断は人間の感覚に合うか？

実際に比較した結果が図8及び9になります。方向感は揃っている月の方が多く、人の感覚値とAIモデルから作成された指数は大きく乖離していない印象です。相関係数も算出していますが、この値からは、両指数間に相関がある可能性が示唆されます。

図8.景気の基調判断（現状）における「極性辞書指数」と「人間判断指数」の比較

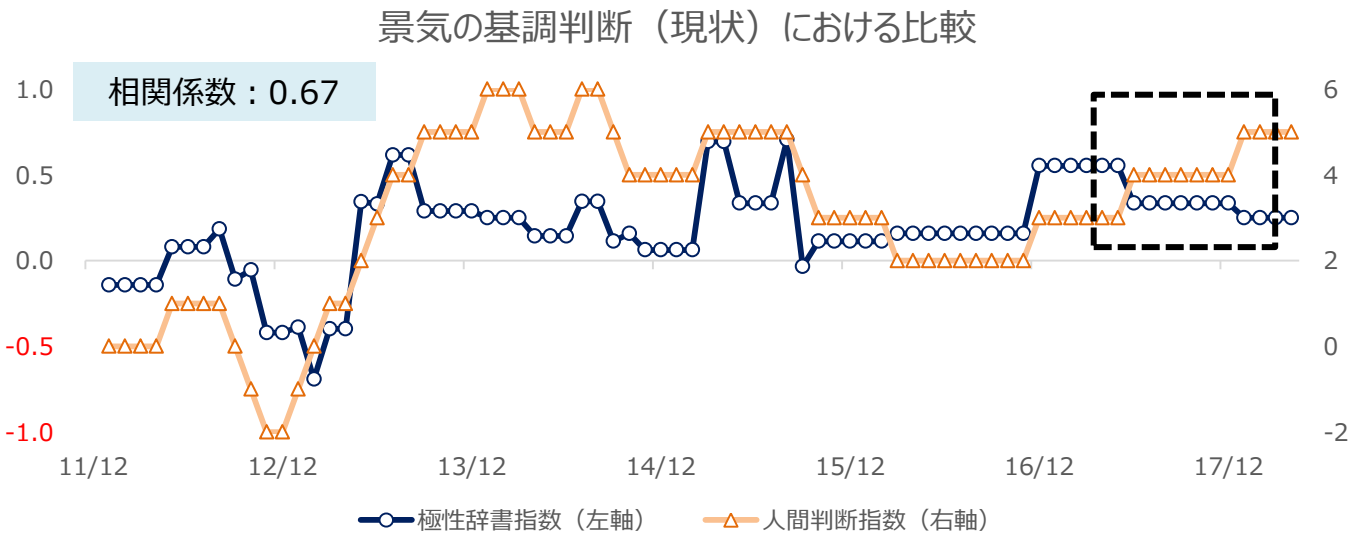
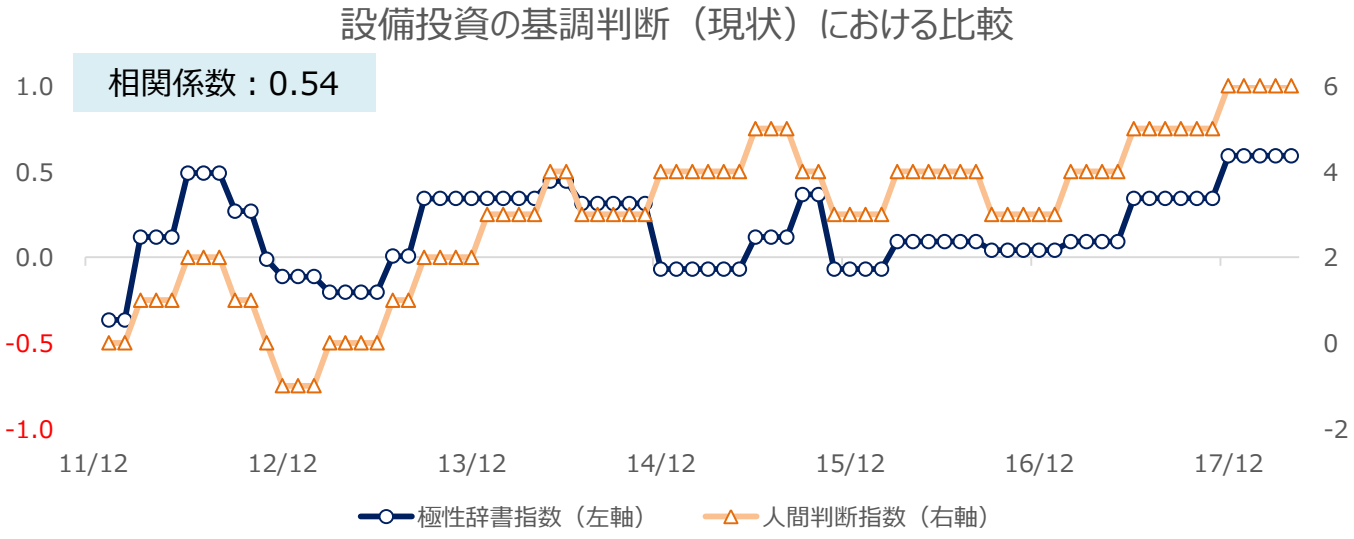


図9.設備投資の基調判断（現状）における「極性辞書指数」と「人間判断指数」の比較



グラフを見て分かる極性辞書指数のメリットとしては、極性辞書指数は毎回の文章の細かな変化を表現できる点かと思います。一方、デメリット、と言いますか、今回の手法の懸念点として、感覚に合わない挙動が時折見られる点が挙げられます。例として、図8の黒破線で囲った部分では、極性辞書指数と人間判断指数の動きが逆になっています。

なぜAIと人で判断が分かれたのか？

4_2. 感覚に合わない文章

図8における破線部分の文章を実際に確認してみます。図10が文章の実例です。極性辞書による指数化では、以下の3事例では上から徐々にスコアが下がっていきます。一方、人の判断ではどうでしょうか？筆者の感覚では、2017年5月の文章よりも2017年6月の文章の方が明らかにポジティブな印象を受けます。2017年6月と2018年1月ではどうでしょうか？どちらも「回復」が主題になっており判断が割れるケースも想定されそうですが、2018年1月の方がやや言い切った雰囲気が感じられます。実際、新聞記事等から知ることが可能な政府の解釈も上方修正でした。

図10.月例経済報告の文章実例

発表日付	文章実例	指数値
20170524	景気は一部に改善の遅れもみられるが緩やかな回復基調が続いている	0.55
20170622	景気は緩やかな回復基調が続いている	0.33
20180119	景気は緩やかに回復している	0.25

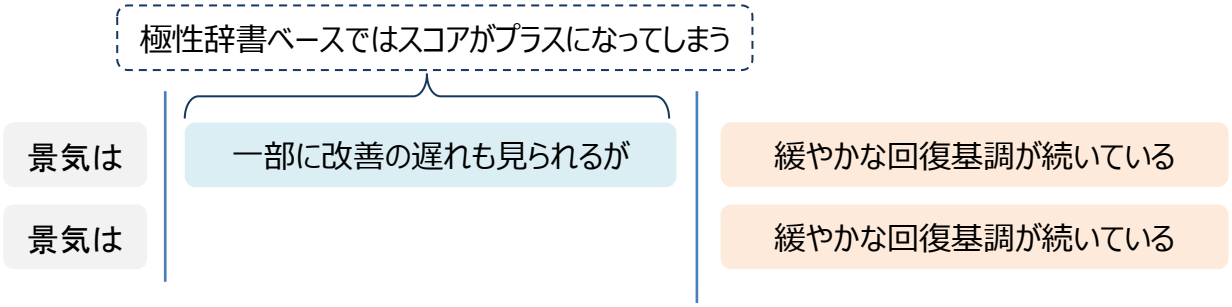
4_3. ケーススタディ

なぜ極性辞書指数では感覚と合わない結果になったのでしょうか？感覚に合わなかった部分の原因を検証してみたいと思います。

ケース1：2017年5月と2017年6月の差異

文章を見てみると、景気については「緩やかな回復基調が続いている」で一致していますが、2017年5月には、状況を説明する文章として「景気は一部に改善の遅れもみられるが」が付随しています。この部分に極性値を割り振った場合、比較的絶対値が大きい極性値を持つ単語は「改善」と「遅れ」になります。極性値を比較すると、「改善」がプラス0.39 、「遅れ」がマイナス0.12となり、足し上げるとプラスが残ります。このため、2017年5月の状況説明部分が極性値プラスの効果を持ち、結果として文章全体でスコアが高くなってしまったようです。

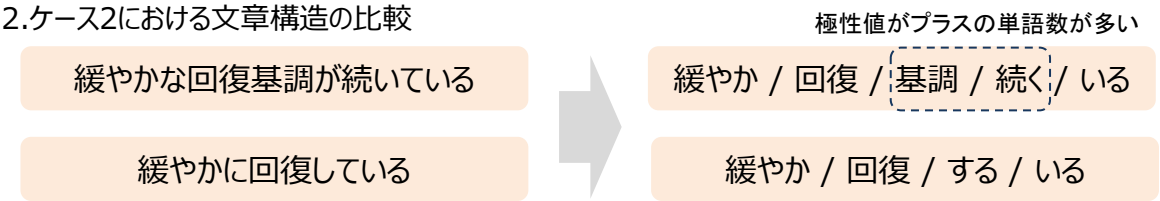
図11.ケース1における文章構造の比較



ケース2：2017年6月と2018年1月の差異

こちらは2017年6月の方が極性値を持つ単語が多いことに起因します。「基調」がプラス0.09、「続く」がマイナス0.01であり、足すとプラスが残ります。この差が2017年6月と2018年1月の差に繋がっています。

図12.ケース2における文章構造の比較



改良可能性の検討

4_4. 改良は可能か？

ケース1及びケース2を対象として改良可能性について簡単に考えてみます。先にケース2を取り上げると、どちらも短い文章であり、かつ「景気は回復」という文章の骨格部分には差異がありません。「～基調が続いている」を言い切りに近い表現である「～している」よりも極性値が下がるように極性辞書型のモデルで表現できるようにするのは簡単にはいかないというのが筆者の感覚です。違うデータセットで辞書を作るか、違うAIモデルを試した方が良いかもしれません。

ケース1については、「遅れ」と「改善」の極性値を足し上げた結果がマイナスになれば条件を満たします。ただ、そのためには極性辞書を作る際に利用した景気ウォッチャー調査の理由集に、「遅れ」が大きくマイナスのインパクトを持つような傾向が見られる必要があります。その傾向が見られない場合は、極性値作成に使うデータを他のデータで試してみる等の工夫が必要です。

一方、ケース1は別の側面で見ると面白い事例ともいえます。これは、接続詞を挟んで複数の文章が存在する場合、どのように処理すべきか、というテーマとして捉えることも可能だからです。対応策としては、係り受け解析を考慮したモデルを組むという選択肢がありますが、日本語の係り受け解析をモデルに落とし込むのは、現状では困難です。他の手段としては、先行事例にもあるようにLSTMという手法が候補に挙がります。ただし、LSTMは極性辞書型とは違い、なぜそうなったのか？といった検証が困難という性格を持ちます。この点を許容できるのであれば、（上手くいくかはケースに依りますが）LSTMは良い選択肢かもしれません。

また、極性辞書のような単語単位、かつ文章評価の理由が分かりやすい手法で対応したい、というニーズの元では、さらに別の手法を探ることも考えられます。例えば、スピンモデルを活用した指数化[参考文献6]が良いかもしれません。[参考文献6]では、隣り合った単語の極性値の向き（符号）が違う場合にペナルティーを導入しています。ケース1の場合だと、前半の文章に「遅れ(極性値マイナス)」、後半に「回復(極性値プラス)」があるので、ペナルティーが発生して従来より指数値が低下します。この方法だと、ケース1は改善の余地があるかもしれません。

5. 終わりに

今回のレポートでは、前回レポートで作成した極性辞書とその極性値を利用して、経済テキストデータである月例経済報告を項目別に指数化しました。また、人間の判断との感覚的な違いが起こる部分に関して、検証と改良案について考えてみました。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. PDFをテキスト化

今回対象とする月例経済報告ですが、内閣府のサイトにはPDF形式でアップロードされており、このPDFからテキスト情報を抽出して利用しています。PDFからテキスト情報を取得する方法は複数ありますが、今回はプログラミング言語pythonのライブラリーを利用しました。しかし、このライブラリーによる作業だけではデータの質は高くはありません。PDFからデータを取得する際に、PDFのページ番号が文章中に混ざり込む、改行がPDFの見え方と同じままになってしまう、うまく取得されない文字列が発生する、等が発生します。これらに対しては地道にテキストデータの前処理を実施しました。

A-2. 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.

5. 山澤 成康
景気指標としての月例経済報告
日本経済研究センター　Discussion　Paper

6.三菱UFJトラスト投資工学研究所
実践　金融データサイエンス