

# ビジネス応用のための自然言語処理入門

ストックマーク株式会社

近江崇宏

Ledge DX Think Week



## 今日の内容：

---

- ✓ 会社紹介・自己紹介
- ✓ 自然言語処理とは？何に使えるのか？
- ✓ この10年での自然言語処理の進化



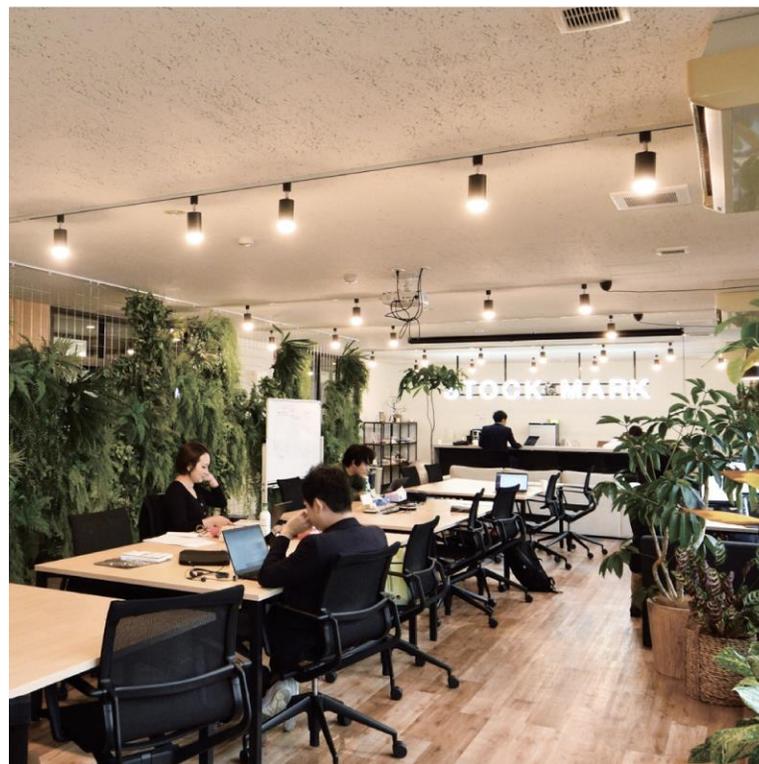
## 今日の内容：

---

- ✓ 会社紹介・自己紹介
- ✓ 自然言語処理とは？何に使えるのか？
- ✓ この10年での自然言語処理の進化

# 会社紹介

会社名	ストックマーク株式会社 Stockmark Inc.
オフィス	東京都港区南青山1丁目12-3 LIFORK MINAMI AOYAMA S209
設 立	2016年11月15日
創業者	代表取締役CEO 林 達 取締役CTO 有馬 幸介
事業内容	自然言語処理技術を活用した AI SaaSプロダクトの提供
従業員数	65名 (2021年11月 現在)
URL	<a href="https://stockmark.co.jp">https://stockmark.co.jp</a>



**Stockmark** 

私たちストックマークは、  
世界中のテキストデータを集め、  
人間が活用できるように整理し、  
人間には発見できない関係性や情報をお届けします。



組織を進化させるインテリジェンス

**Stockmark** 

# 自然言語処理のリーディングカンパニー

日本経済新聞様をはじめ各種テクノロジー関連メディアに登場  
オーム社より待望のBERT本も出版

**Googleの最新AI、読解力も人間超え 驚異の学習法**  
超人間・万能AI (上)

ネット・IT  フォロー済み

2020年1月1日 2:00 [有料会員限定]

保存 



人工知能 (AI) に文章読解は不可能——。米グーグルのAI技術「BERT」がこれまでの常識を覆した

**日経コンピュータ**

画像認識や音声認識などに続き、文章読解の分野でも人工知能 (AI) が人間の平均レベルを超え始めた。米グーグルの新AI技術「BERT (パート)」が壁を突き破った。検査や翻訳といった効率性が飛躍的に高まる可能性が出てきた。

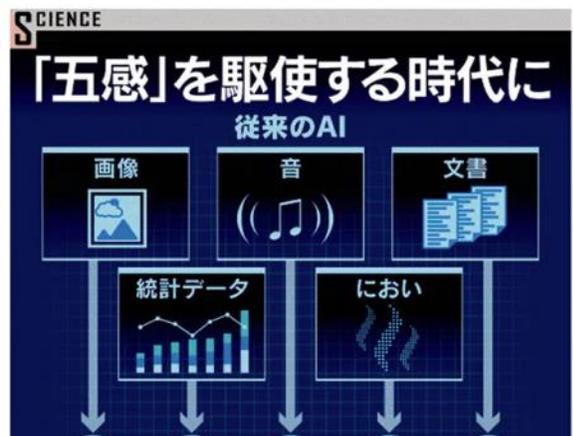
**新世代AI、五感で大進化 「察する力」に人間らしさ**

コラム(テクノロジー)  フォローする

2020年7月11日 2:00 [有料会員限定]

保存 

人工知能 (AI) の行く末を占う鍵となりそうな技術が登場した。「マルチモーダルAI」と呼び、人間が五感を通じて周囲を理解するように、画像や音声、文書など複数のデータをもとに高度な判断ができる。AIがより人間らしくなり、社会の様々な場面にインパクトをもたらす未来が近づいてきた。



SCIENCE

**「五感」を駆使する時代に**

従来のAI

- 画像
- 音
- 文書

統計データ

におい

**BERT**

による

Stockmark株式会社 編  
近江雅宏・金田健太郎・森長隆・江間見聖利 共著

**自然言語処理**

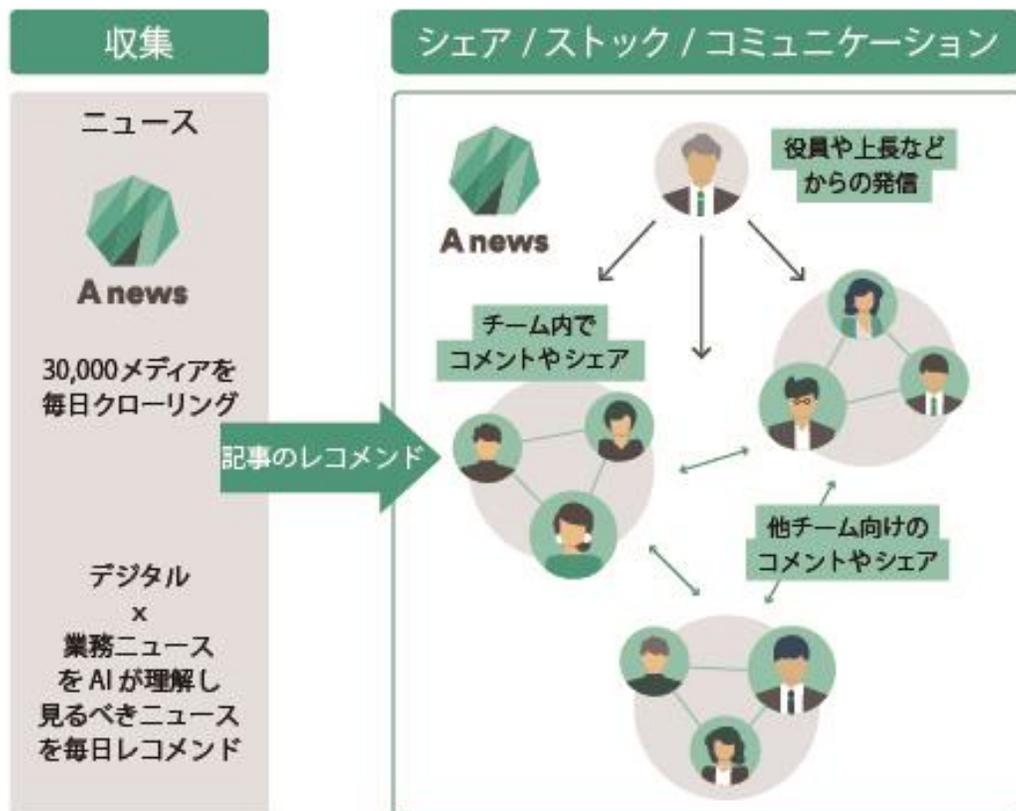
Transformers  
を使った  
実践プログラミング

**入門**

BERTを  
知る・動かす・使いこなす

Stockmark 

さまざまなサービスに応用可能なBERTでの言語タスク解決法を、現役NLP技術者が解説。データセットの処理→ファインチューニング→性能評価までの一連の流れを体験し、BERTを使いこなす力を手に入れよう。



- 世界中の3万メディアからニュースを収集。AIが個人の嗜好に合わせたニュースを毎日配信
- 最先端の自然言語処理で見落としがちなニュースも拾い、教養だけでなく個人や組織のミッションに即したニュースをレコメンド
- **コメント機能**で、簡単にチームにアイデアを共有できる社内の知見者から学び、チームの情報感度を底上げ



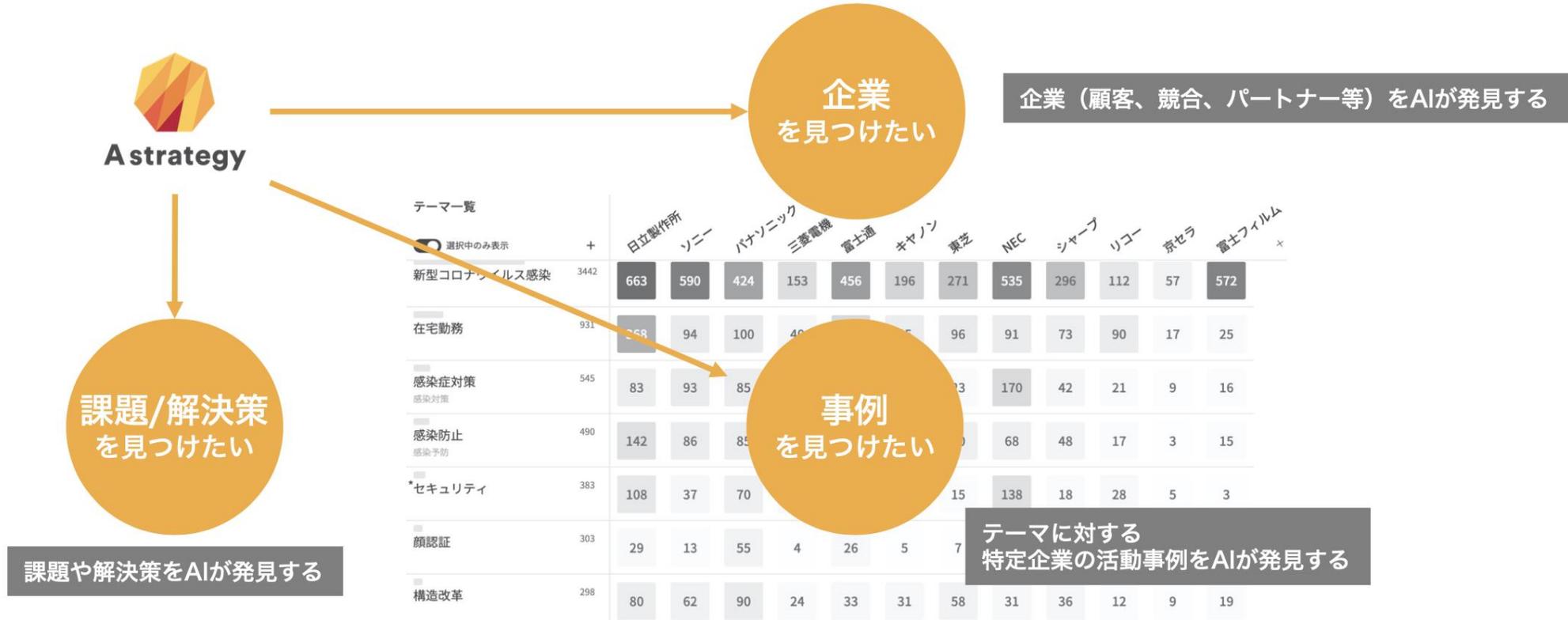


# エンタープライズを中心に累計1,500社以上がご利用



# Astrategy : AIによる市場調査支援

キーワードを入力するだけで数千万のニュース記事から多様な分析を実現



事業企画に欠かせない「課題」「解決策」「顧客」「競合」「自社」の5つの要素をAIが構造的に整理  
 → 従来の情報分析時間を大幅に短縮し、創造的な企画業務に集中できる思考環境を提供



# 自己紹介

---

- ✓ 近江崇宏（おおみたかひろ）
- ✓ 2020年4月にストックマークに入社し、自然言語処理の研究開発に従事
- ✓ アカデミック出身。前職は東京大学（特任准教授）で、時系列解析を中心とした統計学や機械学習の基礎・応用研究に従事。





## 今日の内容：

---

- ✓ 会社紹介・自己紹介
- ✓ 自然言語処理とは？何に使えるのか？
- ✓ この10年での自然言語処理の進化



# 今日話すこと・話さないこと

---

## 今日話すこと：

自然言語処理がどのようなことに使えるか？

特に、文書活用に焦点を絞って解説します。

## 今日話さないこと：

教科書的な解説

技術的詳細



# 文書処理における課題

## 大量の文書の中から欲しい文章を見つけたい

- ✓ お客様の問い合わせ内容に対応する文書を見つけたい。
- ✓ 現在のトラブルと類似の過去のトラブル事例の報告書を見つけたい。

## 大量の文書を分析し、新たな知見を得たい

- ✓ SNSの投稿からある製品に対するユーザーの反応を調べて、製品の改善につなげたい。
- ✓ トラブル報告書から原因を抽出し、今後の施策を決定したい。

しかし、大量の文書を読み、理解し、整理するには多くの時間がかかる。

- 自然言語処理はこれら課題を解決するのに有用な技術
- 近年、自然言語処理の研究が大きく発展し、産業界での活用が進んでいる。





# 自然言語処理（NLP）とは？

人が普段使っている言語（自然言語）をコンピュータで扱うための技術・

学問分野

基礎  
技術

形態素解析

構文解析

固有表現  
抽出

応用  
技術

文書分類

文書検索

情報抽出

機械翻訳

文章生成



# 自然言語はコンピュータには処理が難しい

---

自然言語には様々な曖昧性が存在する。

例) 同一の単語でも状況により意味が異なる場合がある：

- ✓ **ライオン**は来月、消臭・除菌スプレーの新商品を発売する。
- ✓ ○○動物園で**ライオン**の赤ちゃんが生まれた。

人間は文脈からは判断し、前者は**企業**、後者は**動物**であると判断できる。

コンピュータはこのような文脈を踏まえた処理が苦手。 \*最近状況は変わりつつある。



# どのようなデータが使えるか？

---

## 使用可能なデータの収集・電子化

- 日報
- 実験記録
- トラブル・障害・事故報告書
- 仕様書、マニュアル
- 提案書
- 営業メモ
- メール
- チャット
- お客様問い合わせ
  
- 外部オープンデータ（ニュース、論文、特許、、、）



# 自然言語処理をどのように使うか？

ビジネスにおける文書活用の課題は、「分類」・「抽出」・「検索」のタスクに帰着できることが多い。

Ledge.ai 「30年以上研究しても面白い」自然言語処理技術の今と未来をフロントランナーが語る  
<https://ledge.ai/stockmark-nlp-event202107/>

- 社内に蓄えられている文書の利用可能性の向上し、情報の集約や分析が行いやすくなる。
- これらを基盤として、新たな知見につながることを期待される。

文書分類

情報抽出

文書検索



# 自然言語処理をどのように使うか？

---

大量の文書の中から欲しい文書を見つけたい。 文書検索

- ✓ お客様の問い合わせ内容に対応する文書を見つけたい。
- ✓ 現在のトラブルと類似の過去のトラブル事例の報告書を見つけたい。

大量の文書を分析し、新たな知見を得たい。 情報抽出・文書分類

- ✓ SNSの投稿からユーザのニーズを抽出し、新サービスの企画を作りたい。
- ✓ トラブル報告書から原因を抽出し、今後の施策を決定したい。

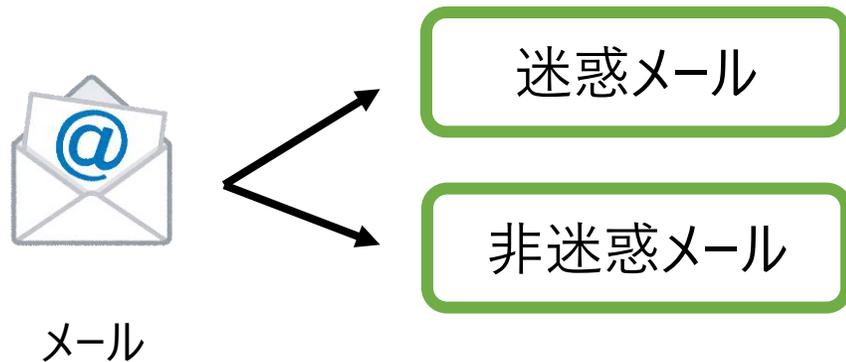


# テキスト分類

文章分類とは、文章をあらかじめ決められたカテゴリーに分類するタスク

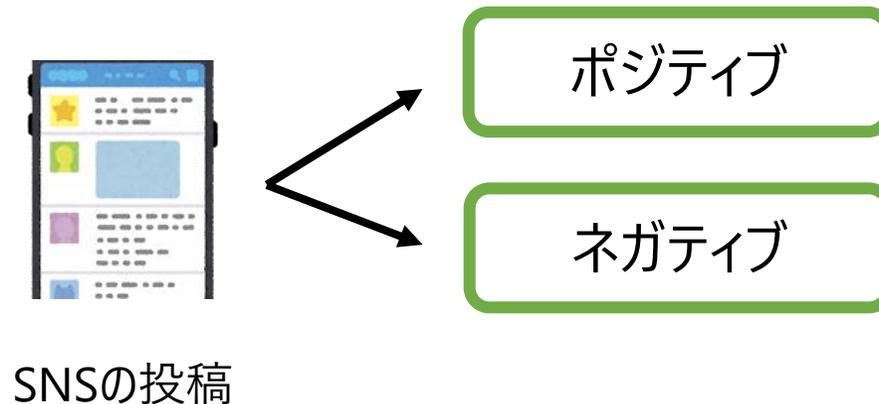
分類しておくことにより、情報の集約や興味のある情報へのアクセスが容易になる。

## 迷惑メールフィルタ



## ネガポジ解析

例) ある製品に対するSNSの投稿が好意的か否定的かを分類



- 製品に対する人々の反応が定量的にわかるようになる。
- ネガティブな投稿だけ、選ぶといったようなことが可能になる。



# 弊社プロダクトでの活用事例

## エンタメフィルタ

- [課題] お客様に配信するニュースからビジネスに関係ないエンタメ記事を除きたい。
- 記事を「エンタメ記事」か「非エンタメ記事」に分類。
- お客様にはエンタメ記事を表示するか、しないかを選べるようにすることで、配信する記事の質を向上させることができる。

## 記事のカテゴリ分類（検証中）

- [課題] ユーザーがどのような記事を見ているのかを分析したい。
- 記事を「経済」、「科学」、「IT」、、、などのカテゴリに分類。
- ユーザーが閲覧した記事のカテゴリを分析することで、ユーザーがどのような分野に興味を持っているかを分析することができる。



# テキスト分類のユースケース

---

- 工場の障害レポート・事故報告書を原因別に分類、深刻度のランク付け。
- お客様の声（問い合わせ、アンケートなど）を要望・感想・クレームなどのカテゴリに分類
- サーバー内の社内文書を「機密文章」か「非機密文書」に分類
- 技術文章を分野ごとに分類
- 日報を重要度に応じて5段階に分類



# 情報抽出

---

情報抽出とは、文書から興味ある情報を抽出し、構造化する技術。

- 抽出した情報により、情報の集約や興味のある情報へのアクセスが容易になる。
- 文書分類に比べて、より具体的な情報を得ることができる。

## 固有名詞の抽出

山田さんは大阪府在住だ。

人名

地名

- 抽出した情報に基づく分析
- 個人情報の匿名化

## 関係性の抽出

Appleは新型のiphoneを10万円で発売した。

製品名

価格

- 競合企業が発売する製品とその価格を列挙
- 製品の価格を分析



# 弊社プロダクトでの活用事例

## 企業名の抽出

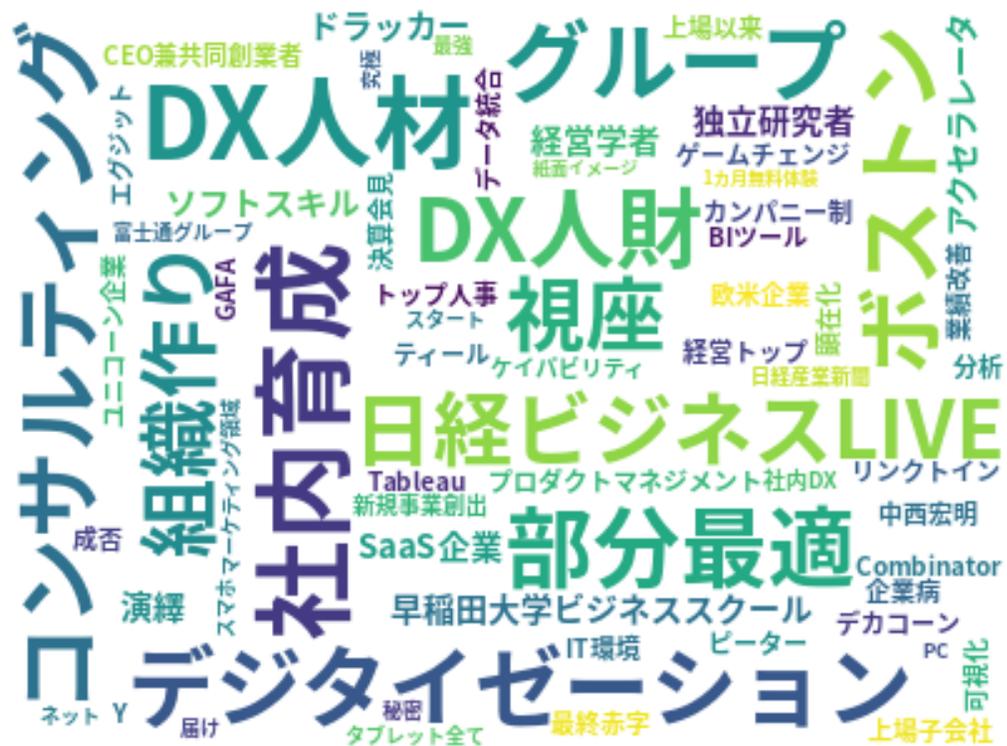
- [課題] 市場調査のため、業界にどのような企業がいるかを知りたい。
- 記事から企業名を抽出。
- 記事を読まなくても、記事に載っている企業がわかるようになる。
- 企業名を軸にしたさらに高度な分析も可能になる。
- 業界に登場した新しい企業を知ることができる。

## ユーザの興味分析（検証中）

- [課題] ユーザーがどのような記事に興味があるかを分析したい。
- 記事からキーワードを抽出
- ユーザーが閲覧した記事から抽出されたキーワードのうち、頻度が高いものを調べることで、ユーザーがどのようなテーマに興味があるのかがわかるようになる。



# 弊社プロダクトでの活用事例（ユーザー興味分析）



弊社CEO



弊社R&D Manager



# 情報抽出のユースケース

---

- 商談メモからお客様のニーズの抽出
- 営業提案書からの商材・企業名の抽出
- 日報からのキーワード抽出
- 化学の学術書から化学物質とその物性を抽出する



# 文書検索

---

文書検索とは、ユーザーの入力に応じて、最適な文章を見つけるための技術  
入力はキーワード、文、文書など様々な形をとりうる。

例)

- 現在のトラブルと類似の過去のトラブル事例の報告書を見つけない。
- お客様の問い合わせ内容に対応する文章を見つけない。
- ある特許文書と関連する特許文書を見つけない。



# 弊社プロダクトでの活用事例

---

## 企業名検索

- [課題] 市場調査のため、ある企業の情報を調べたい。
- 記事から企業名を抽出し、それをもとに文書の検索を行う。
- 興味のある企業についての記事のみを調べることができる。



# 弊社プロダクトでの活用事例

## Google 検索

Google

Q すべて 画像 動画 ニュース 地図 もっと見る ツール

約 191,000 件 (0.24 秒)

**Yahoo!ニュース**  
**珍しい双子のホワイトライオンの赤ちゃん、すくすく育つ ...**  
 赤ちゃんたちの親は、一般的な茶色のアフリカライオンで、飼育員たちは、「まさか白い赤ちゃんが生まれるとは」と驚いたそう。元気いっぱいミルクを飲んだあとは、...

**PR TIMES**  
**コロナ禍における市内の経済活動等を後押しするためライオン株式会社が埼玉県戸田市に除菌・消臭スプレーを約1万4千本寄贈**  
 埼玉県戸田市はこのたび、ライオン株式会社から除菌・消臭スプレー（衣類・布製品用）約14,000本の寄贈を受けることになりました。寄贈品はライオン...

**産経ニュース**  
**【マーライオンの目】3度目接種の副反応**

様々な意味の「ライオン」に関する記事が出力される。

## Astrategy

ワークシート タイトルなし ☆ 複製して開く 検索 企業指定  選択解除

絞り込み 記事の言語 日本語 記事の種別 すべて

**記事検索**

3397 件 ① 日付順

**コロナ禍における市内の経済活動等を後押しするためライオン株式会社が埼玉県戸田市に除菌・消臭スプレーを約1万4千本寄贈**

**ライオン、使用済み歯ブラシ16万本回収・再生**

**ライオン、目薬にミネラルオイル配合 新製剤化技術を開発**

**ライオン、コロナ禍におけるオーラルケア実態調査結果を発表**

企業の「ライオン」に関する記事のみが出力される



# 自然言語処理を活用するにおいて

課題に応じて、どのような軸で分析を行うのかを決めるのが重要

Astrategy: 市場調査 -> 「誰」が「何」をしているかが重要：「企業」と「テーマ」が軸



大量の記事を整理し、ビジネスチャンスの発見を支援



## 今日の内容：

---

- ✓ 会社紹介・自己紹介
- ✓ 自然言語処理とは？何に使えるのか？
- ✓ この10年での自然言語処理の進化



# この10年での人工知能の進化

---

深層学習の研究の発展により、これまでコンピュータには苦手とされていた領域において、人工知能の性能が飛躍的に向上した。

- 画像認識
- 自然言語処理
- AlphaGo

当時は、コンピュータが囲碁棋士に勝つまで10年はかかると言われていた。

<https://wired.jp/2016/01/31/huge-breakthrough-google-ai/>



# 自然言語処理（NLP）の発展

---

- ✓ 自然言語処理は人間にとっては容易な課題が多いが、コンピュータにとっては苦手な領域であった。
- ✓ 深層学習によるブレイクスルーにより、多くの自然言語処理のタスクで人間と同等またはそれ以上の性能が得られるようになった。

**BERT**

文章の分析が得意

**GPT-3**

文章の生成が得意



# BERTとは [J. Devlin et al., 2018]

- ✓ 2018年にGoogleにより提案された自然言語処理のニューラルネットワーク。様々な言語タスクで既存のモデルを超えるパフォーマンスを示す。
- ✓ その後の発展に貢献し、最先端のモデルでは人間のスコアを上回るに至る。

## GLUE Benchmark

モデル	スコア
現在のトップのAI	90.8
人間	87.1
BERT	80.5
ELMo	70.0



# BERTが産業界に与えたインパクト

---

- 文章分類・固有表現抽出・文章検索などの自然言語処理の活用事例でよく現れるタスクが得意。
- 日本語に特化したモデルが公開されており、誰でも使える。
- 比較的少数のデータから学習することができる（数百のデータからでもパフォーマンスが出ることもある）。
- End-to-Endで用いることができ自然言語処理の非専門家でも扱いやすい。

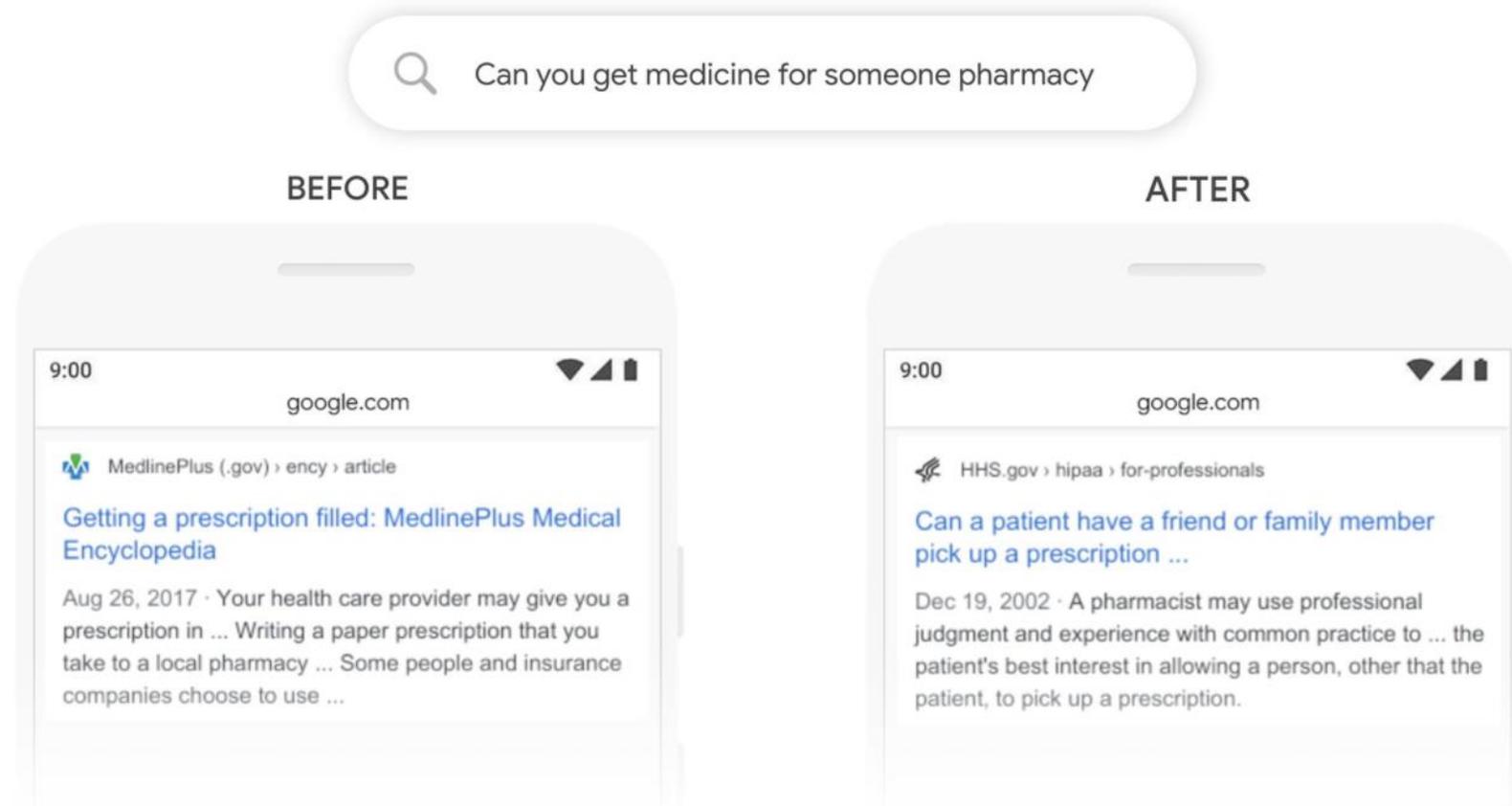
AIの民主化

事業会社などでの利用も進んでいる。



# BERTの実用例

Google検索：自然文に対しても、ユーザーの意図に対応した結果に



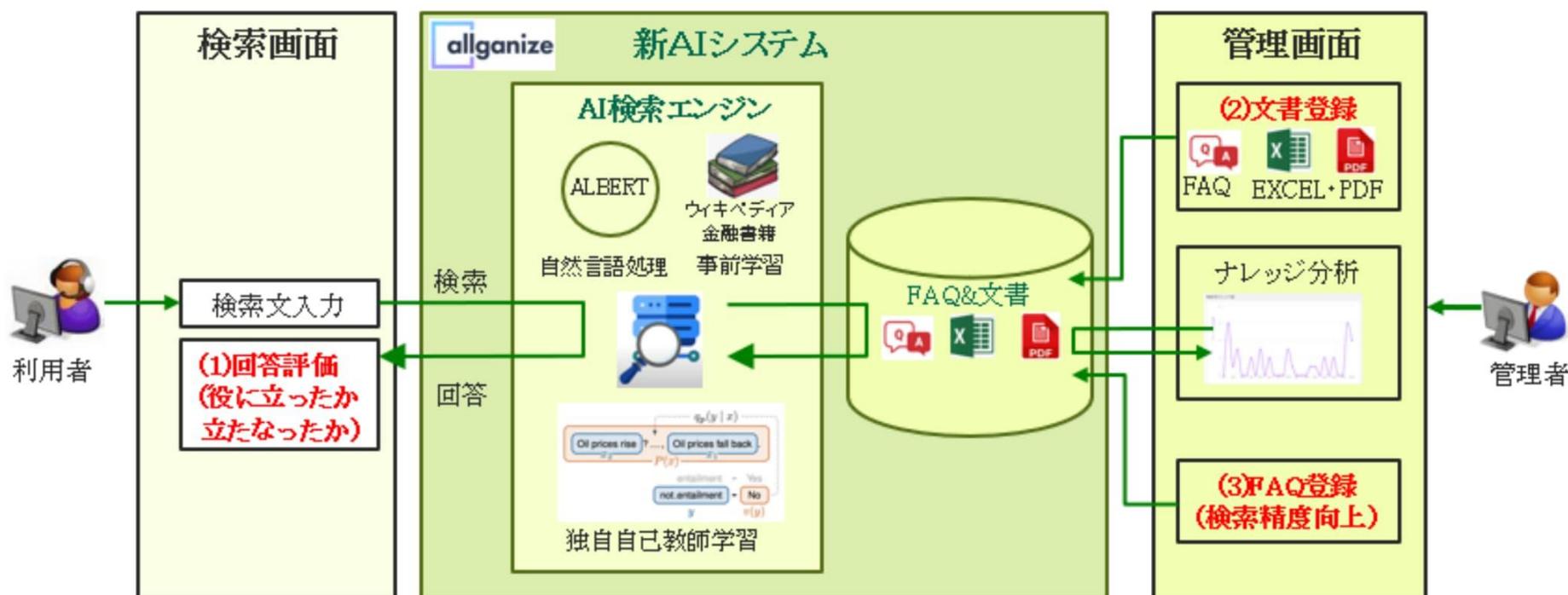
<https://blog.google/products/search/search-language-understanding-bert>



# BERTの実用例

## 三井住友FG：コールセンターのオペレータの質問回答支援

顧客からの問い合わせに対して、オペレーターが端末に検索用の文章を入力すると、対応する回答を掲載した社内FAQのWebページやPDF文書を表示。



# BERTの実用例：ニュース記事からの企業名抽出（ストックマーク）



- Stockmarkは、文脈を考慮できるAIの「BERT」を用いた手法を利用。

詳しく読む：[BERTによるニュース記事の構造化：企業名抽出](#)

- 企業名を正しく発見するには、**文脈から単語の意味を判断する**必要があり、これまでは難しかったが、弊社はそれをいち早く実現。

同じ単語でも文脈により意味が異なる例：

[企業名] スタートアップの**シナモン**は新サービスを発表した。

[誤検出] 店舗で**シナモン**の香りがするスイーツを販売する。

## 企業名抽出パフォーマンス評価

	企業名を 正しく抽出	企業名以外 を誤検出
Stockmark	93%	9%

- ビジネスニュースに特化したモデルを作成したことで、高い精度で企業名を抽出でき、誤検出も少なくなった。
- 企業名抽出が高精度にできるようになったことで、それを基盤とした高次の解析を行えるようになった。



# GPT-3とは [T. B. Brown et al., 2020]

- ✓ 2020年にOpenAIによって開発された1750億のパラメータを持つ超巨大なAI。570GBのテキストから学習。
- ✓ 文章生成に強みを持つ。人間と比べても遜色のない質の文章を生成できる。

## GPT-3の動作例

[入力] 業務の生産性を上げるために何をすれば良いか？

[出力] その解決には、業務システムがあっこそです。歴史的に見て、ビジネスの変化や社会変化に対応できない企業は経営難に陥り、業績を落としていったのです。先進的な企業の分析を見ても、多くの企業は業務システムを導入しています。それは、この業務システムが、各種のビジネスシーンで効果を発揮しているからです。例えば、大手電機メーカーであれば、自動車用・・・

将来的には、レポート作成などもAIが行えるようになるかも、、、

# GPT-3が生成したブログが1位に

Feeling unproductive? Maybe you should stop overthinking.

 Adolos  
Jul 20  27  32 



In order to get something done, maybe we need to think less. Seems counter-intuitive, but I believe sometimes our thoughts can get in the way of the creative process. We can work better at times when we "tune out" the external world and focus on what's in front of us.

I've been thinking about this lately, so I thought it would be good to write an article about it.

So what exactly does this mean? Well, for starters, let's start with some definitions.

Definition #1: Creative Thinking (CT) is the act of using your mind to come up with ideas that are not already present in reality or that have never existed before. It is a form of mental activity that involves the use of logic and reason.

Y

**Hacker News**

new | threads | past | comments | ask |  
show | jobs | submit

wporr (39)  
| logout

1. **Feeling unproductive? Maybe you should stop overthinking** (adolos.substack.com)  
47 points by adolos 1 hour ago | flag | hide |  
26 comments
2. ▲ **'Doomscrolling' Breeds Anxiety. Here's How to Stop the Cycle** (npr.org)  
34 points by mrfusion 1 hour ago | flag | hide |  
24 comments
3. ▲ **Why OKRs might not work at your company** (svpg.com)  
136 points by codesuki 4 hours ago | flag | hide |  
49 comments

<https://adolos.substack.com/p/feeling-unproductive-maybe-you-should>

<https://www.technologyreview.com/2020/08/14/1006780/ai-gpt-3-fake-blog-reached-top-of-hacker-news/>



# GPT-3の実用例

## Github Copilot (<https://copilot.github.com/>)

GPT-3の文章生成をコード生成に応用。  
作りたい機能の説明を書くだけで、それに対応するコードを生成してくれる。

```
sentiment.ts write_sql.go parse_expenses.py addresses.rb

1 #!/usr/bin/env ts-node
2
3 import { fetch } from "fetch-h2";
4
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8   const response = await fetch(`http://text-processing.com/api/sentiment/`, {
9     method: "POST",
10    body: `text=${text}`,
11    headers: {
12      "Content-Type": "application/x-www-form-urlencoded",
13    },
14  });
15  const json = await response.json();
16  return json.label === "pos";
17 }
```

Copilot



## まとめ

---

- 自然言語処理の活用における「文書分類」「情報抽出」「文書検索」の3つの軸について解説しました。
- 自然言語処理の活用する際には、分析する「軸」を決定することが重要
- 今回話したことは、自然言語処理の活用における基盤的な部分であり、これをもとに分析等を行うことにより、各社の課題に応じた有用な知見の発見につながることを期待される。



# さらに勉強されたい方へ

## 参考図書：

- 「自然言語処理の基本と技術」, グラム・ニュービッグ, 萩原正人, 奥野陽, 小町守.  
応用に焦点をおいた自然言語処理の解説
- 「BERTによる自然言語処理入門」, スtockマーク株式会社, 近江崇宏, 金田健太郎, 森長誠, 江間見亜利.  
プログラミング、実践書



