

# BERT による 新しい MoA 抽出と 読解エンジンとしての活用の可能性



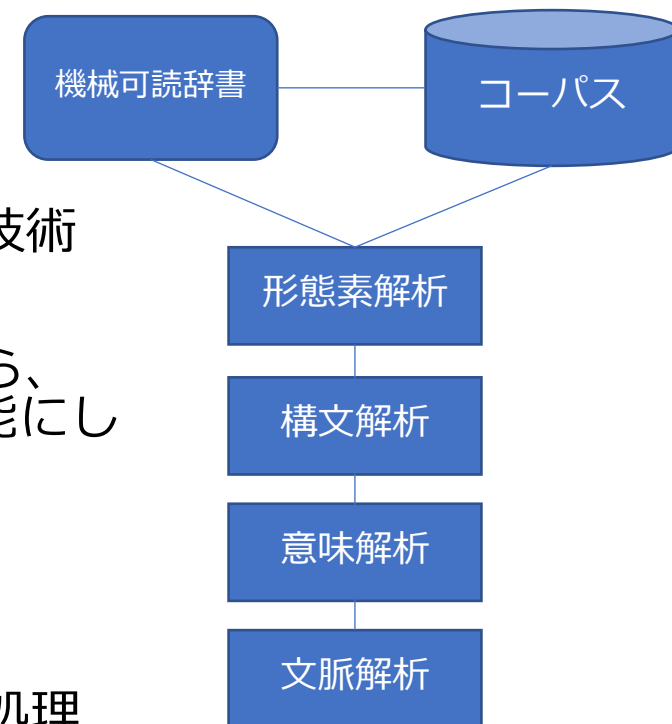
櫻井 祐樹 緑川 淳  
(株)ワールドフュージョン

# BERT による 新しい MoA 抽出と 読解エンジンとしての活用の可能性

- 自然言語処理 と BERT
- BERTを使った新規MoA抽出プロジェクト実行環境
- BERTを使ったGeneRIFの遺伝子vs疾患の関係性の判別

# 自然言語処理 (Natural Language Processing)

- 自然言語処理 (Natural Language Processing) とは
  - 人間の言語 (自然言語) を機械で処理 (言葉が持つ意味を解析)
  - 曖昧性を含め、自然言語で書かれたテキストデータを処理
    - 話し言葉、書き言葉
  - 「かな変換」「Google 検索」「機械翻訳」など身近に利用されている技術
- Word2Vec
  - 言語を機械的に処理するためには、言語を単語や文字などの「記号」から、「数値ベクトル」に変換にする必要があり、Word2Vecはその変換を可能にした。
  - 従来の自然言語処理モデルでは、文章を単一方向からの処理に限定。(Word2Vec)
- BERT
  - BERT<sup>1)</sup>は 2018年10月 Jacob Devlin (Google) らが発表した自然言語処理モデル
    - BERTでは双方向のTransformerによる学習により 精度が向上。
  - スタンフォード大学の文章読解ベンチマークSQuAD 1.1<sup>2)</sup>での評価において、BERTは人工知能で初めて人間の平均の精度を超えた



1) <https://arxiv.org/abs/1810.04805>

2) <https://nlp.stanford.edu/pubs/rajpurkar2016squad.pdf>

# BERT (Bidirectional Encoder Representations from Transformers )

- BERTの特徴

- マスクされた言語モデル (Masked Language Model : MLM)
- 次文予測 (Next Sentence Prediction : NSP)
- 文脈から単語の意味を解釈できる。文章の語順を考慮できる (≠ Word2Vec)
- 「Attention (注意機構)」の導入により離れた単語間の情報も適切に取り入れられる (≠ RNN、LSTMベースモデル)

- BERTのタスク

- 固有表現抽出(Named Entity Recognition: NER)
  - テキストから一般名詞ではない 語句：固有表現(人名、組織名、etc...)を抽出
- MNLI
  - MultiNLIの略称で、自然言語推論のコーパスを意味している。元論文ではこのコーパスは約43万の前提と仮説のペアから構成される。
  - 前提と仮説の内容上の関係に応じて「含意 (entailment)」「矛盾 (contradiction)」「どちらとも言えない (neutral)」といったラベルが付けられる。

[https://ainow.ai/2019/05/21/167211/#Next\\_Sentence\\_PredictionNSP](https://ainow.ai/2019/05/21/167211/#Next_Sentence_PredictionNSP)

<https://buildersbox.corp->

[sansan.com/entry/2021/09/21/120000#%E5%9B%BA%E6%9C%89%E8%A1%A8%E7%8F%BE%E6%8A%BD%E5%87%BANER%E3%81%A8%E3%81%AF](https://buildersbox.corp-sansan.com/entry/2021/09/21/120000#%E5%9B%BA%E6%9C%89%E8%A1%A8%E7%8F%BE%E6%8A%BD%E5%87%BANER%E3%81%A8%E3%81%AF)

# BERTを使った新規MoA抽出プロジェクト

- プロジェクト概要
- BERT 用 事前準備
  - 訓練処理用テキストデータ
  - 推論処理用テキストデータ
  - 実行環境
- 結果
- 今後の課題

# プロジェクト概要

- 過去のCBI学会で発表した GO-MoAは、疾患に有意な生物学的機能を提示することで、新たなターゲットや薬剤を提案する有用な方法である。GO-MoAの拡張において新しい Mechanism of Action (MoA) を早期に取得することが課題であった。
- この課題を解決する上でAI技術を調査した結果、深層学習の転移学習モデルである、「BERT」による「固有表現抽出 (NER : Named Entity Recognition) タスク」にて検討することとした。
- 今回 BERTの汎用的な固有表現の学習済みモデルとして、特にライフサイエンス分野の利用目的に特化した「BioBERT」モデルを学習済みモデルとして採用した。

「BERT」の事前訓練データ : Wikipedia (2.5B num of words) ,BookCorpus(0.8B)

「BioBERT」 : 上記 + PubMed Abst (4.5B) ,PMC Full-text(13.5B)

# BERT用事前準備

## 訓練処理用テキストデータ

下記データをSQL処理で組み合わせ、出力した論文テキストデータに対してMoAタームをラベリングすることにより、教師有りの訓練データとした。

1. ChEMBL-DBのDRUG\_MECHANISMテーブル  
( MoA、Target、Chemicalの組み合わせ)
2. LSKB のターゲット（遺伝子）、MoA、化合物オントロジー
3. 「2」のオントロジータームによる直近20年のPubMed論文テキストマイニングデータ

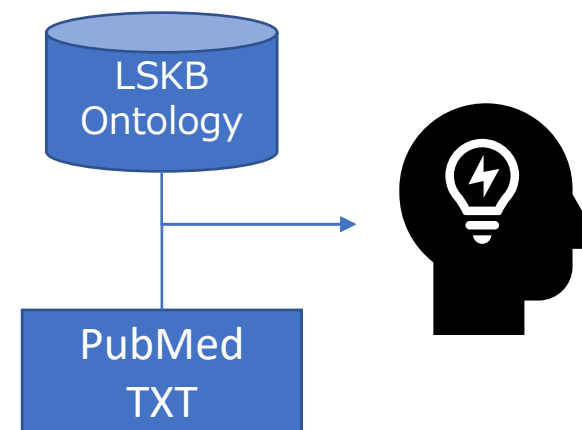
## 推論処理用テキストデータ

下記データをSQL処理で組み合わせ、出力した論文テキストデータを推論処理用データとした。

1. LSKB のターゲット（遺伝子）、MoA、化合物オントロジー
2. 「1」のオントロジータームによる直近20年のPubMed論文テキストマイニングデータ  
(具体的には「2」のマイニングデータの内、化合物タームとターゲットタームは共出現するが、MoAタームは検出されない論文データをテキスト化)

## 実行環境

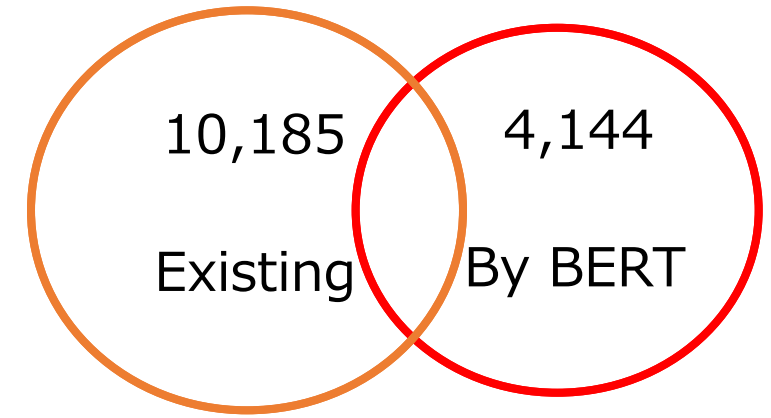
- OS : Oracle Linux 7
- GPU : NVIDIA Tesla P100 16 GB
- Python およびGPU関連モジュール : OS提供Python : 3.6.8
- Tensorflow : 2.5.0



# 結果

- Fine-tuningを実施したBERTモデルのNERタスクによるMoAターム抽出結果、新規MoAターム（候補）4,144件を取得した。
- MoAにアサインできたターゲットは888件であった。
- BERTの推論結果を利用して 抽出された MoAタームとエビデンスとして利用可能な論文情報と記載文章が取得できた。
- モデルの精度は次のとおり。

# of MoA



ラベル	Recall(再現率)	Precision(精度)	F1 Score
MoA	98.3%	95.3%	96.8%

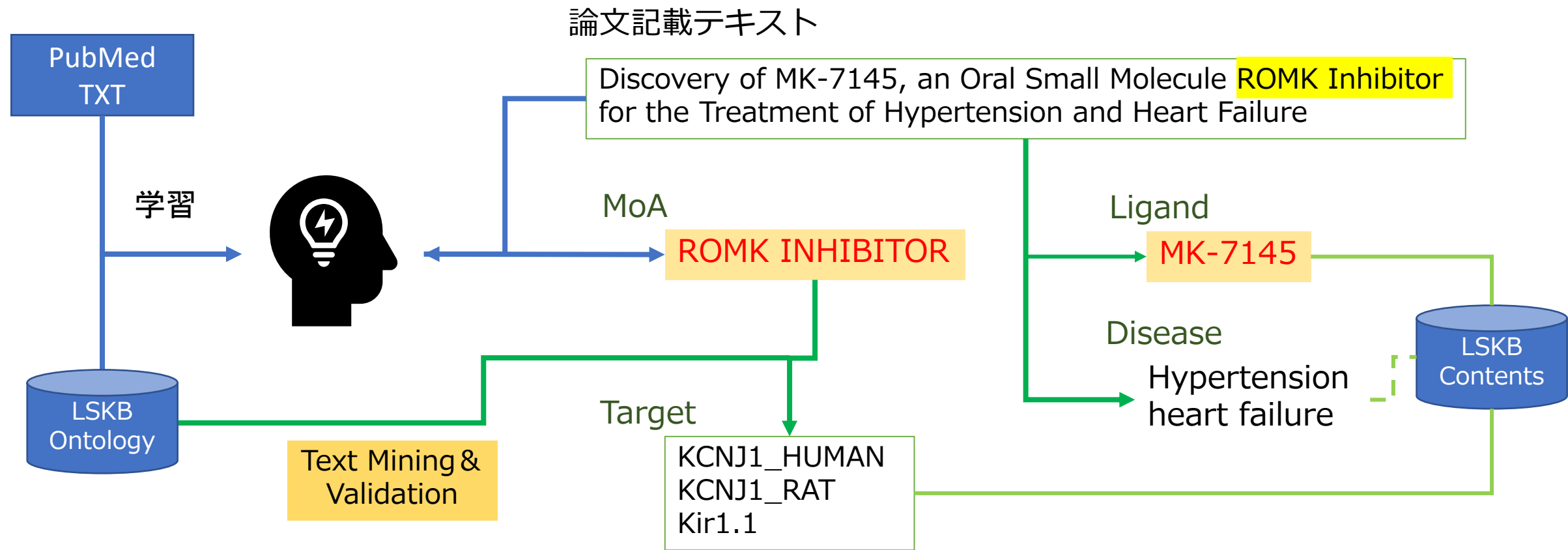
Recall(再現率)：テストデータに存在するとあるラベルのうちモデルが正しく判定した割合

Precision(精度)：とあるラベルと予測したうち実際にそのラベルが当たっていた割合

F1 Score：RecallとPrecisionの調和平均



# 新規MoA 抽出と後処理プロセス概要



ラベル	Recall(再現率)	Precision(精度)	F1 Score
MoA	98.30%	95.30%	96.80%

# 今後の展開

- 抽出した新規MoA に、ターゲット、化合物、関連疾患をアサインし検証プロセスを実行した。
- 結果をDB化するまでの過程において多くの属人的キュレーション作業を要した。今後はLSKB内の2次DBの利用等で、自動化をすすめる。
- 新規MoAを利用して、GO-MoAデータベースを更新する。
- 新規MoA の抽出課程については、新しいデータソースを訓練データに追加し、効率のよいモデル作成を検討する。

# BERTを使ったGeneRIFの遺伝子vs疾患の 関係性の判別

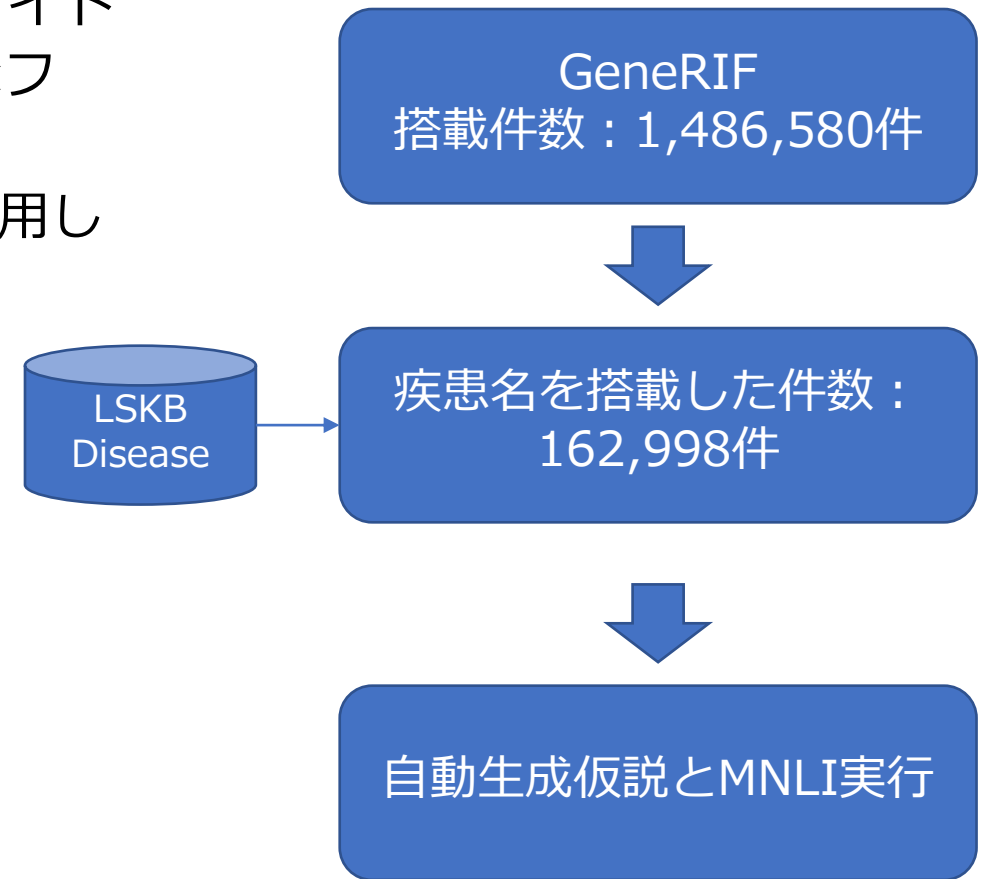
- プロジェクト概要
- GeneRIFとは
- BERT用 事前準備
- 結果
- 今後の課題・展望

# プロジェクト概要

- LSKBでは、疾患とターゲット遺伝子の関連情報を元に、疾患からターゲットを探す Target Explorer、ターゲットから関連疾患を探す Disease Explorerを搭載している。
- この情報ソースの一つにGeneRIFがあり、遺伝子に関するAuthorのコメントが記載されているが、その記載は Biomarkerや Variantにおける関連、疾患の原因など 多岐にわたっており、その関係性を判断できれば、更に有用性が向上する。
- この課題を解決する上でAI技術を調査した結果、深層学習の転移学習モデルである、「BERT」による「MNLIタスク」を検討することとした。
  - MNLIでは前提文章と仮説を比較し、「含意 (entailment)」「矛盾 (contradiction)」「どちらとも言えない (neutral)」と判断
  - 比較に用いる仮説はGeneRIFに含まれるキーワードから 仮説を機械的に生成した。

# Gene Reference Into Function (GeneRIF)

- GeneRIFとは
  - NCBIが提供する 遺伝子機能のアノテーション登録サイト
  - PubMedで引用されるPMIDと タイトル以外の簡潔なフレーズ(425文字以下)で 1 つまたは複数の機能を説明
- GeneRIF記載文章から LSKB Disease オントロジーを利用したテキストマイニングで疾患名を取得
- GeneRIFの文章から次の仮説を自動生成
  - 遺伝子と疾患との関連
  - 遺伝子発現の変動と疾患との関連
    - Up/Down regulation
  - SNPが与える疾患への影響
  - 疾患のバイオマーカーとなり得る遺伝子情報



# BERT用事前準備

## 訓練処理用テキストデータ

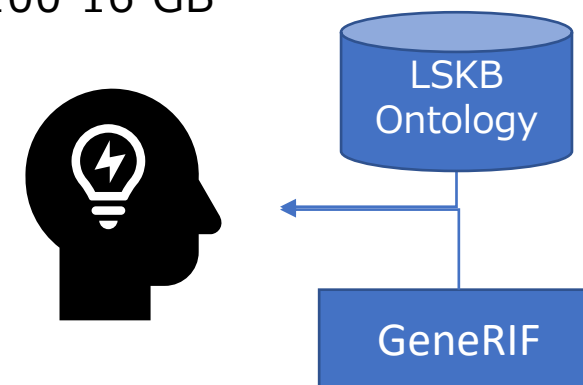
- TensorFlow Datasetsより、entailment, neutral, contradictionのラベル付けをされた、前提文と仮説の組み合わせレコード計391,829のテキストデータを教師有りの訓練データとした。

## 実行環境

- OS : Oracle Linux 7
- Python およびGPU関連モジュール : OS提供Python : 3.6.8
- Tensorflow : 2.5.0

## 推論処理用テキストデータ

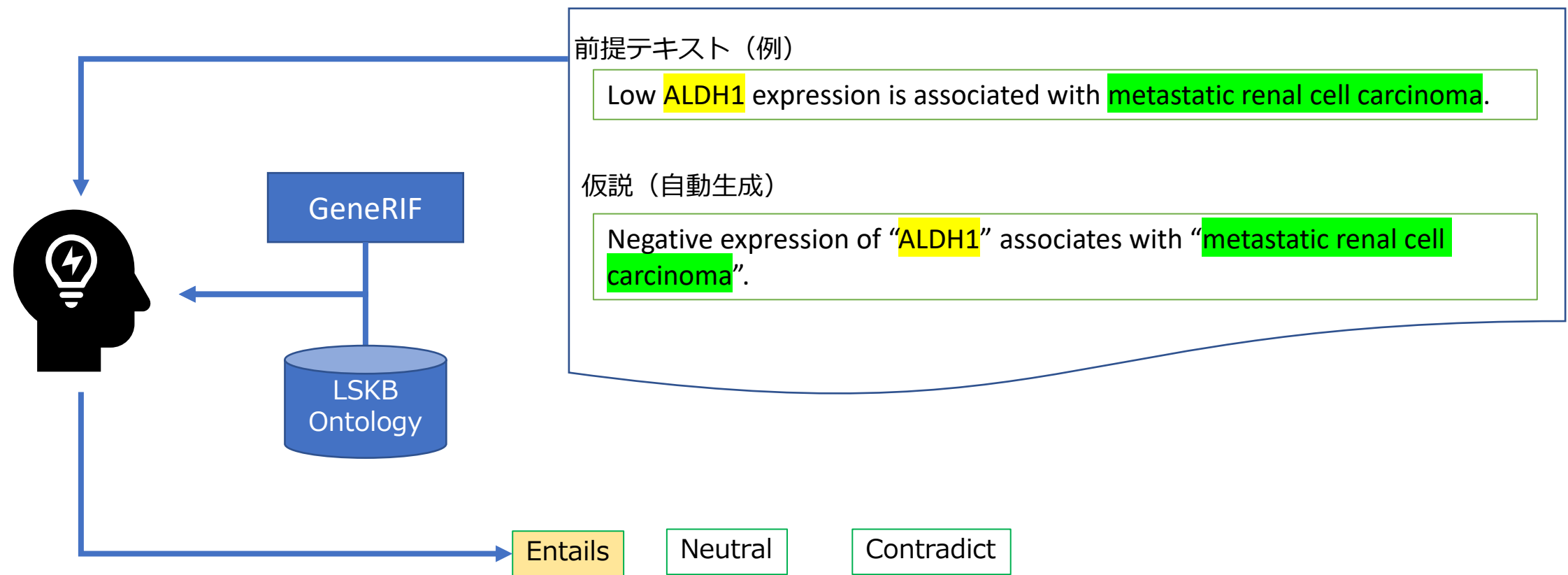
- GeneRIFのテキストデータにLSKBの疾患オントロジーでテキストマイニングを実施し、ヒットしたテキストデータを仮説検証の対象とした。
- 検証用の仮説はGeneRIFのテキストデータに対する特定用語（Activity, Enhance, Inhibitionなど）でのテキストマイニングにヒットした結果を用いて機械的に生成した。



# 結果

- GeneRIFの搭載文章 1,486,580件のうち、162,988件（Gene 19190件）をBERTのMNLI タスクを実行した。
  - 遺伝子の生物種の内訳は Human (10523)、Mouse (5323)、Rat (2671)、Zebrafish (273) であった。
- 9 種類の仮説を自動生成（計 1,499,079 件）した。その内訳は
  - “遺伝子と疾患との関連” 仮説で取得された 疾患は 10502件、遺伝子は18135件であった。
  - “遺伝子発現の負の変動と疾患との関連性” 仮説で取得された 疾患は 3714件、遺伝子は 5673件であった。
- 各仮説の判定結果（Entailment）をランダムに20-25件 前提文章と比較検証したところ、精度は 80-100%であった。
- 実際の利用場面（Entailment）の MNLIのタスクの精度は F1-Score : 0.832であった。

# BERTによるGeneRIFの遺伝子vs疾患の関係性の判別概要



	Recall(再現率)	Precision(精度)	f1-score
contradiction	88.68%	75.03%	81.29%
entailment	83.11%	83.29%	83.20%
neutral	67.60%	82.54%	74.33%



# 正しくENTIALMENTと判断できた事例

仮説：「遺伝子」の負の発現変動が「疾患」に起因する。

GENE	DISEASE	GENERIF
slc41a1	nephronophthisis	<p><b>Knockdown</b> of slc41a1 expression in zebrafish resulted in ventral body curvature, hydrocephalus, and cystic kidneys, similar to the effects of knocking down other <a href="#">nephronophthisis</a> genes.</p> <p>ゼブラフィッシュにおける<b>slc41a1発現のノックダウン</b>は、他の<b>髄質性嚢胞腎</b> 遺伝子をノックダウンする効果と同様に、腹側体の湾曲、水頭症、および嚢胞性腎をもたらしました。</p>
ALDH1	carcinogenesis	<p><b>loss of</b> ALDH1A1 expression is suggested to promote <a href="#">carcinogenesis</a> especially in the smoking-related lung adenocarcinomas</p> <p>ALDH1A1<b>発現の喪失</b>は、特に喫煙関連の肺腺癌において<b>発癌</b>を促進することが示唆されています。</p>
ALDH1	metastatic renal cell carcinoma	<p><b>Low</b> ALDH1 expression is associated with high stage and <a href="#">metastatic renal cell carcinoma</a>.</p> <p><b>低ALDH1発現</b>は、ハイステージおよび<b>転移性腎細胞癌</b>と関連しています。</p>
BIN1	cutaneous T-cell lymphoma	<p><b>Reduced</b> BIN1 expression is associated with <a href="#">cutaneous T-cell lymphoma</a>.</p> <p>BIN1<b>発現の低下</b>は、<b>皮膚T細胞リンパ腫</b>に関連しています。</p>
Bcl-2	severe congenital neutropenia	<p>a <b>selective decrease</b> in Bcl-2 expression was seen in myeloid progenitor cells of patients with Kostmann syndrome: <a href="#">severe congenital neutropenia</a>.</p> <p>コストマン症候群患者：<b>重症先天性好中球減少症</b>の骨髓前駆細胞でBcl-2<b>発現の選択的減少</b>が見られました。</p>

訳文はGoogle 翻訳をもとに若干修正を加えた。

# 正しく ENTAILMENT と判断できなかった事例

仮説：「遺伝子」の負の発現変動が「疾患」に起因する。

GENE	DISEASE	GENERIF
Gankyrin	hepatocarcinogenesis	Gankyrin expression promotes spontaneous cancer and non-cancerous liver diseases. Gankyrin accelerates liver steatosis, cholangitis, fibrosis, and hepatocarcinogenesis with persistent damage and proliferation in hepatocytes.
		ガンキリンの発現は、自然発生的な癌および非癌性肝疾患を促進します。ガンキリンは、肝細胞の持続的な損傷と増殖を伴い、脂肪肝、胆管炎、線維症、および肝発癌を加速する。
ATP7B	Wilson's disease	Duodenal CTR1 mRNA and protein expression was decreased in Wilson's disease patients, while ATP7A mRNA and protein production was increased. This can be a defense mechanism against systemic copper overload resulting from functional impairment of ATP7B.
		ウィルソン病患者では十二指腸のCTR1mRNAとタンパク質の発現が減少し、ATP7AのmRNAとタンパク質の産生が増加しました。これは、ATP7Bの機能障害に起因する全身的な銅の過負荷に対する防御機構である可能性があります。

訳文はGoogle 翻訳をもとに若干修正を加えた。

# 今後の展開

- 複数の仮説を統合整理し、現在のGeneRIF にアノテーションとして利用する予定である。
  - 遺伝子と疾患との関連性
  - 遺伝子の発現変動と疾患の影響
  - 遺伝子のSNPと疾患の関連性
- 検証プロセスの確立
  - 初期検討では ネイティブのスタッフと検証
- Fine-tuning用の教師あり学習データや 新しい手法など継続して検討する。
- GeneRIF以外のテキストデータにおいてもMNLIを適用し、データの質の向上を試みる

**ご清聴ありがとうございました。**

**アンケートにご協力をお願いいたします。**

**<https://www.lskb.jp/cbi2021>**

**ご質問は**

**[support@lskb.jp](mailto:support@lskb.jp)**

**までお願いいたします。**

