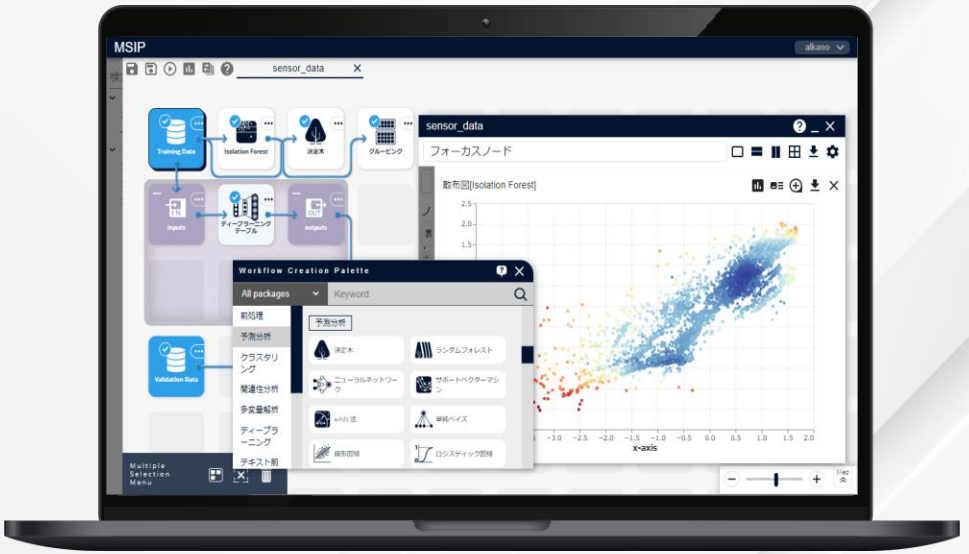
【アルカノ】
Alkano

テクニカルサンプルプロジェクト

日本語学習済みモデルの利用



株式会社 NTTデータ数理システム

NTT DATA Mathematical Systems Inc.

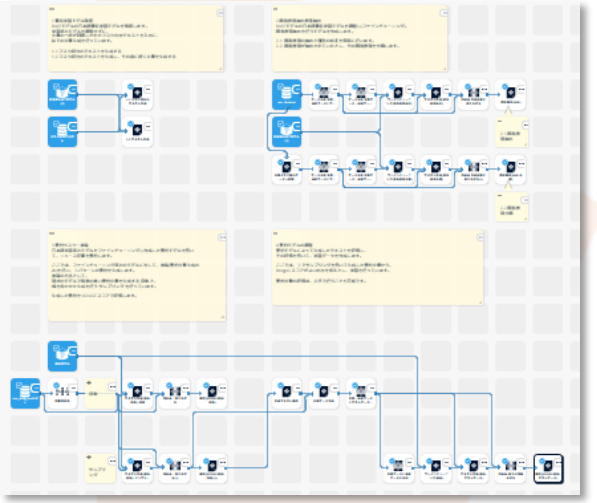
このプロジェクト について

こんな方におすすめします

事前学習済みの自然言語処理モデルを試したい方、お手元のテキストデータで構築したモデルの精度向上をさらに目指したい方

何をするプロジェクト？

事前学習済みの自然言語処理モデルを利用して、追加データでのチューニングを行い文書の要約や固有表現抽出といったタスクに適用する例を紹介します。
近年、自然言語処理の分野において、大量のテキストデータを利用して学習された汎用的な大規模モデルの開発が盛んになっています。さらに、テキストデータから機械学習を用いて解決したいタスクがあるときに、手元のテキストデータのみを利用してモデルを構築するのではなく、タスクに合わせたデータで事前学習済みの大規模モデルをチューニングするといったことが広く行われています。これにより、手元のテキストデータの量が十分でない場合も、既存の優れたモデルを起点にすることでより精度の向上が見込める場合があります。
こういった事前学習済みモデルは、数百億から場合によっては兆単位の内部パラメータを持つ巨大なものであるため、一般にはサービス提供元のサーバ上で運用されているというモデルを利用するという形が採られます。しかし、これより小規模でお手元の環境で動作させることが不可能ではないモデルもあり、特定の言語・特定のタスクであれば、同等の精度以上が見込めることもあります。社内のデータを外部のサービスに投入できないといった事情がある場合にも、お手元の環境にインストールしたAlkanoで事前学習済みモデルを利用することが可能です。
このプロジェクトでは、そういったモデルの一つであるBART(Bidirectional Auto-Regressive Transformer)を利用して、文書の要約と固有表現抽出を行う例を紹介いたします。その他、文書の分類など他のタスクにも適用できる可能性がある技術ですので、ご興味を頂きましたら是非お問い合わせください。



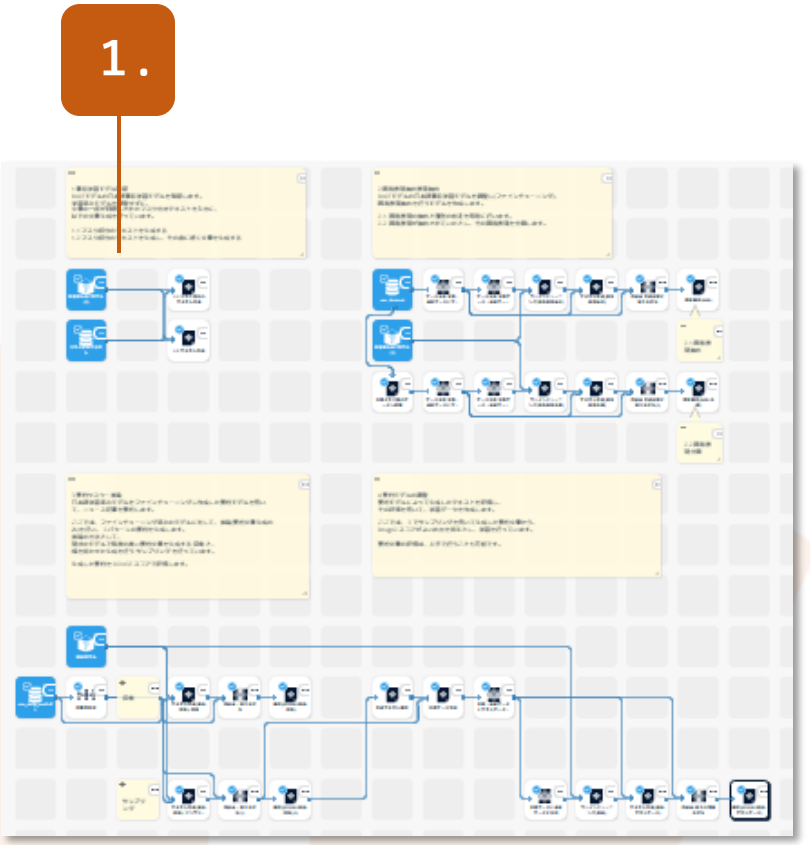
※p.29の「本文書・プロジェクトファイルのご利用にあたって」をお読みのうえ、ご利用ください。

プロジェクト 解説

プロジェクト 解説

1. 本プロジェクトで用いる事前学習済みモデル

本プロジェクトでは、Hugging Face, Inc. によって運営されているAIコミュニティーサイト Hugging Face で公開されている BART* の日本語学習済みモデル**「BART日本語Pretrainedモデル」を使用します。このモデルは、日本語 Wikipedia 全て (約1800万文) を用いて事前学習されています。事前学習では「一部のテキスト(単語列、トークン列)が <mask> に置換された文章に対し、その <mask> 部分に入るテキストを推測するタスク」を行っています。



サンプルプロジェクトの [1] の部分では、以下のマスク付きテキスト

可視化画面:マスク付きテキスト-dataノード

No.	prompt String
1	天気が<mask>から散歩しましょう。
2	私の好きなスポーツは<mask>です。
3	休日はよく<mask>をします。

を用いて

1-1. マスク(<mask>)部分のテキスト生成

1-2. マスク(<mask>)部分のテキスト生成+後に続くテキストの生成

を行い、学習済みモデルでの日本語でのテキスト生成が行えることを確認します。

*Lewis, Mike, et al. “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." *arXiv preprint arXiv:1910.13461* (2019).

**Formzu/bart-base-japanese・Hugging Face : https://huggingface.co/Formzu/bart-base-japanese

2. 固有表現抽出

サンプルプロジェクトの[2]の部分では、BARTの事前学習済みモデルを用いて、固有表現抽出を行うためのファインチューニング(微調整)を行い、実際に固有表現抽出を行います。

2-1. 固有表現抽出について

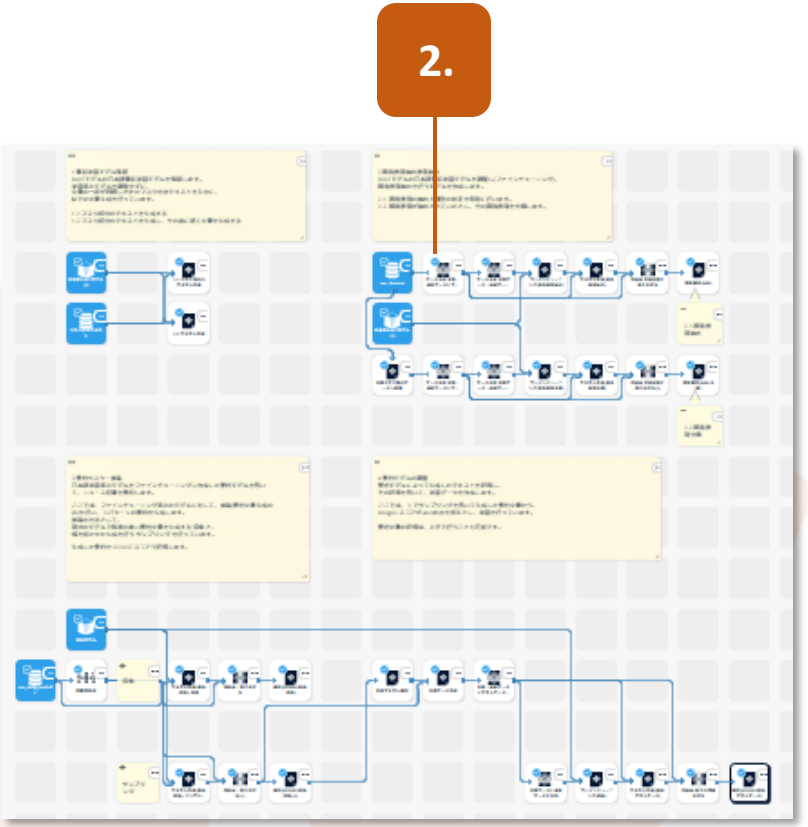
固有表現(named entity)とは、人名や地名などといった固有名詞や、日付、時間などに関する表現を指します。固有表現抽出とは、テキストから固有表現を抜き出し、抜き出した固有表現をあらかじめ定義された固有表現分類に分類するタスクです。
例えば、下記の文章

『2019年現在、日本の航空会社では日本航空、全日本空輸、AIRDOの3社で使用されている。』

で、固有表現の抽出と分類を行うと以下のようになります。

固有表現	固有表現分類
日本	地名
日本航空	法人名

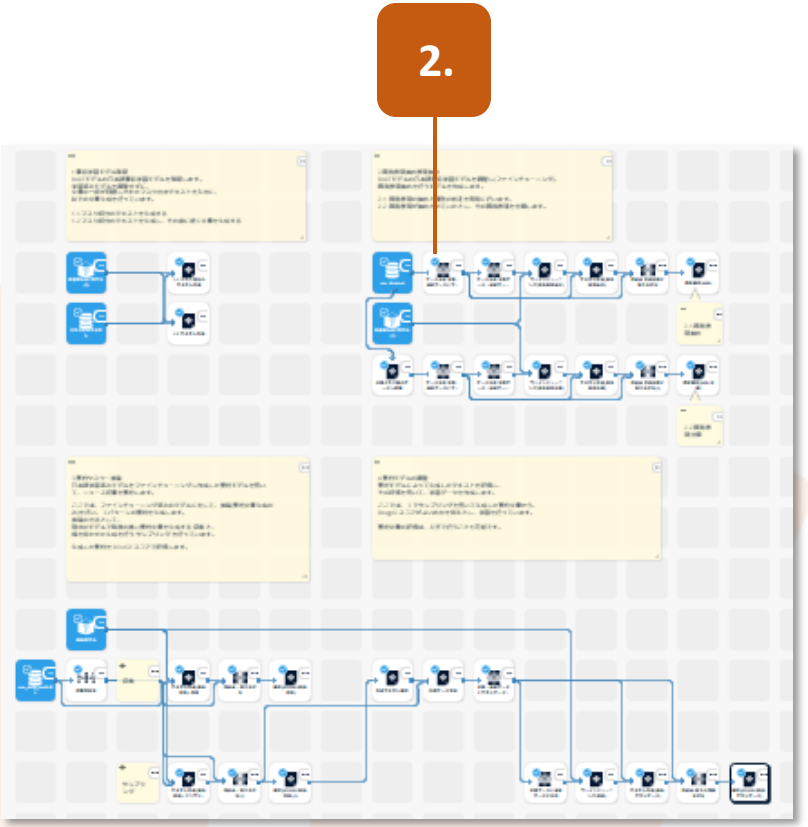
※上記の例は、本プロジェクトの利用データ「Wikipediaを用いた日本語の固有表現抽出データセット」の内容から抜粋したものです。利用データの解説は次頁で行います。



プロジェクト 解説

2. 固有表現抽出
2-2. データセットについて

使用するデータセットとして、株式会社ストックマートが公開している Wikipediaを用いた日本語の固有表現抽出データセット*を用います。これは、対象の文章 (prompt) に対して文中の固有表現の位置と種類(type)がタグ付けされたデータです。
本プロジェクトでは、テキスト生成タスクとしてデータを扱うためデータセットを整形し、下図のように「〇〇は△△です。」(固有表現ラベル)という文章を作成します**。



可視化画面: ner_dataset-data ノード

No.	prompt Category	label Category	type Category
1	SPRiNGSと最も仲の良いライバルグループ。	SPRiNGSはその他の組織名です。	その他の組織名
2	レッドフォックス株式会社は、東京都千代田区に本社を置くITサービス企業である。	レッドフォックス株式会社は法人名です。	法人名
3	レッドフォックス株式会社は、東京都千代田区に本社を置くITサービス企業である。	東京都千代田区は地名です。	地名
4	松友美佐紀は、日本のバドミントン選手。	松友美佐紀は人名です。	人名
5	松友美佐紀は、日本のバドミントン選手。	日本は地名です。	地名
6	ライターの兵庫慎司は普通にアイドルポップスとして出すと売れず、無理にバンドとコラボレー	兵庫慎司は人名です。	人名
7	2019年現在、日本の航空会社では日本航空、全日本空輸、AIRDOの3社で使用されている。	日本は地名です。	地名
8	2019年現在、日本の航空会社では日本航空、全日本空輸、AIRDOの3社で使用されている。	日本航空は法人名です。	法人名
9	2019年現在、日本の航空会社では日本航空、全日本空輸、AIRDOの3社で使用されている。	全日本空輸は法人名です。	法人名
10	2019年現在、日本の航空会社では日本航空、全日本空輸、AIRDOの3社で使用されている。	AIRDOは法人名です。	法人名
11	同月5日には、トヨタファイナンシャルサービス証券株式会社を吸収合併。	トヨタファイナンシャルサービス証券株式会社は法人名です。	法人名

*Wikipediaを用いた日本語の固有表現抽出データセット :
https://github.com/stockmarkteam/ner-wikipedia-dataset
**変換したデータセットは、 ner_dataset.csv として公開します。

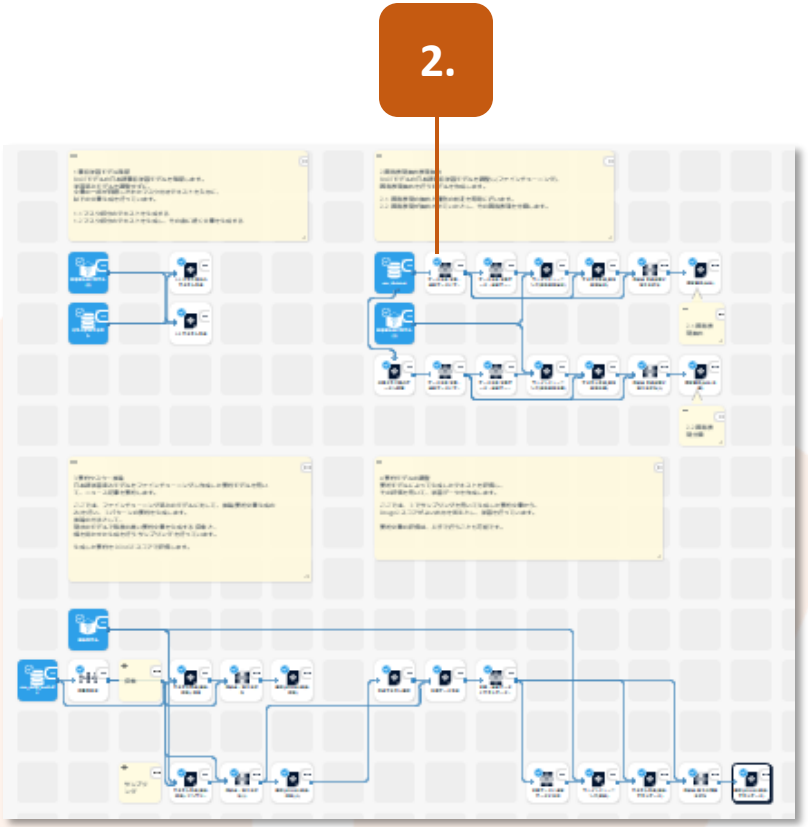
プロジェクト 解説

2. 固有表現抽出
2-3. 固有表現抽出モデル

BART日本語Pretrainedモデルをもとに、対象の文章 (prompt) からラベル (label) を生成するよう学習することで、固有表現抽出タスクを行うモデルを作成します。BART日本語Pretrainedモデルのままですと、日本語の語彙への対応や簡単な日本語のテキスト生成はできるのですが、固有表現抽出には適用できません。 BART日本語Pretrainedモデルをもとに、ファインチューニング (微調整) することで、固有表現抽出を行うモデルを作成します。

対象の文章の例(prompt) :
『かつては東京都三鷹市にオフィスを構えていたが、
大月のキングレコード退職を機に事業を停止。』

ラベルの例(label) :
『大月は人名です。』



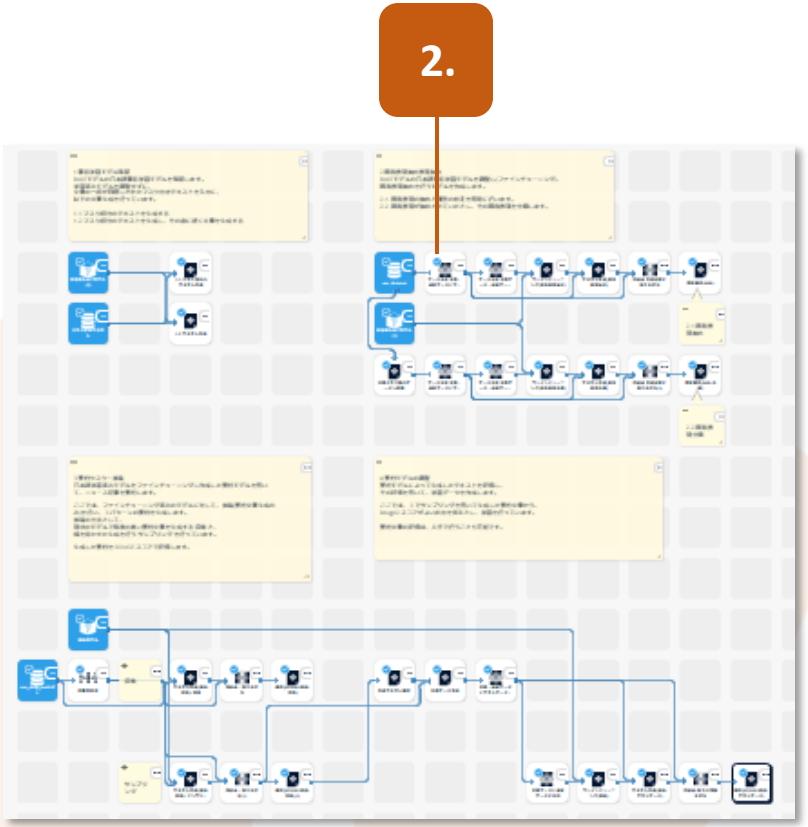
2. 固有表現抽出
2-4. 与えられた固有表現に対して分類を与えるモデル

固有表現抽出では、『〇〇は△△です。』という形で、固有表現自体の抽出(〇〇の部分)と、固有表現の分類(△△の部分)を同時に行っていました。
ここでは加えて、固有表現そのものが既知のものとして与えられた場合に、その固有表現に対する分類(ラベル)を予測する問題を考えてみます。そのために、以下のように説明変数としていた文章の最後に、分類を予測したい既知の固有表現を『〇〇は』として質問として付加してファインチューニングを行います。この学習モデルを利用することで、『〇〇は』の質問に対する回答を、予測結果の文章として得ることができます。

固有表現分類での例:
対象の文章の例(prompt):
『かつては東京都三鷹市にオフィスを構えていたが、大月のキングレコード退職を機に事業を停止。大月は』

ラベルの例(label):
『大月は人名です。』

固有表現分類用データで追加されるテキスト



プロジェクト 解説

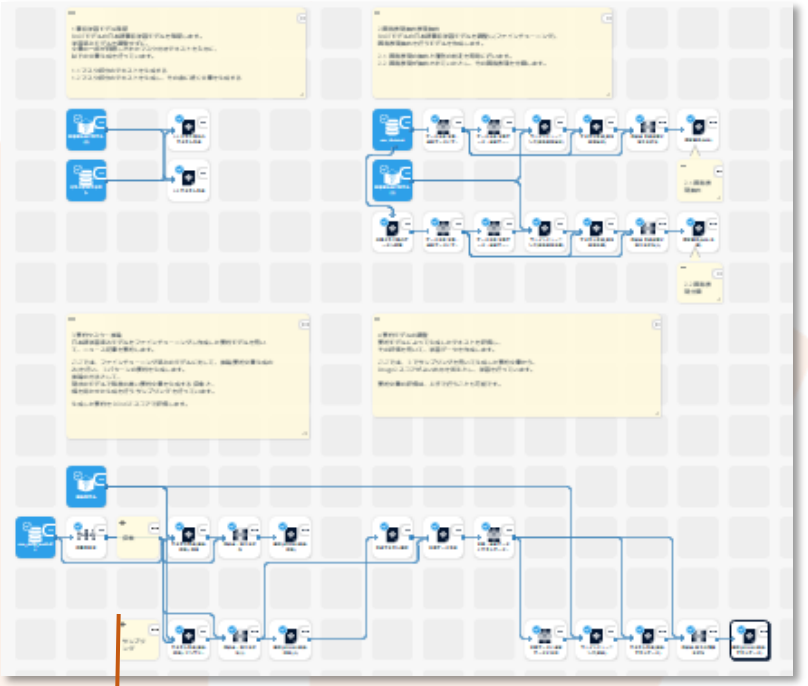
3. 要約モデルでのテキスト生成
3-1. 要約モデルについて

[3.]では、 BART日本語Pretrainedモデルをもとに、あらかじめファインチューニングをして作成済みの要約モデルを用いて、要約したテキストを生成します。要約モデルには、対象となる文章の重要となる部分を抽出する抽出型と、要約したテキストを生成する生成型がありますが、本モデルは後者になります。
要約モデルは、LiveDoorNewsからクロールしたデータに基づく3行要約データセット*を利用し、ニュース記事本文を説明変数、対応する記事の3行要約を目的変数に用いています。

ニュース記事イメージ

記事本文　　3行要約

https://news.livedoor.com/topics/detail/23911842/ より引用



3.

*https://github.com/KodairaTomonori/ThreeLineSummaryDataset

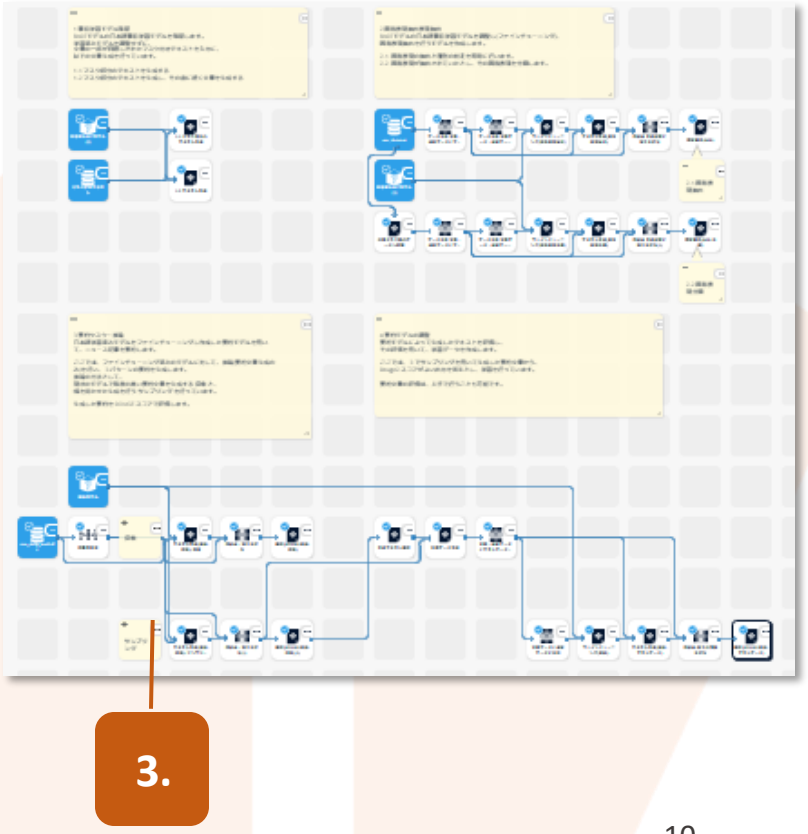
プロジェクト 解説

3. 要約モデルでのテキスト生成
3-2. テキスト生成

要約モデル を用いて、実際に要約したテキストの生成を行います。
入力として、 CNN/Daily Mail データセット*を日本語に翻訳した
データセットを用います**。
下図の記事部分 (article_ja_s) を説明変数 (入力) にし、生成された要
約文章と答えとする要約文章 (highlight_ja) の部分と比較をします。

可視化画面:cnn_daily_mail.dft-data ノード

No.	article_ja String	highlights_ja String	article String	highlights String	article_ja_s String
1	ロンドン（英国）（ロ	ハリー・ポッターのス	LONDON, England (R	Harry Potter star Dani	ハリー・ポッターの主
2	編集部注：「Behind tl	マイアミの精神病患者	Editor's note: In our B	Mentally ill inmates in	マイアミ・デイド州の
3	ミネソタ州ミネアポリ	NEW：「死ぬかと思っ	MINNEAPOLIS, Minn	NEW: "I thought I was	ミネアポリスの橋が崩
4	ワシントン（CNN） -	検査中に5つの小さな	WASHINGTON (CNN)	Five small polyps foun	医師は土曜日、ブッシ
5	(CNN) -- ナショナル	NEW：NFLのチーフと	(CNN) -- The National	NEW: NFL chief, Atlan	ナショナル・フットボ
6	イラクのバグダッド（	両親は誇らしげで、応	BAGHDAD, Iraq (CNN	Parents beam with pri	スーパーマンのシャツ
7	イラクのバグダッド（	援助関係者：暴力と生	BAGHDAD, Iraq (CNN	Aid workers: Violence,	女性たちは恐れと恥ず
8	コロンビア・ボゴタ（	トマス・メディナ・カ	BOGOTA, Colombia (	Tomas Medina Caraca	米国の麻薬取引で起訴



*Dataset: cnn_dailymail
https://huggingface.co/datasets/cnn_dailymail
**日本語に翻訳したデータセットについては、cnn_daily_mail.csv として公開します。

4. 要約モデルのファインチューニング

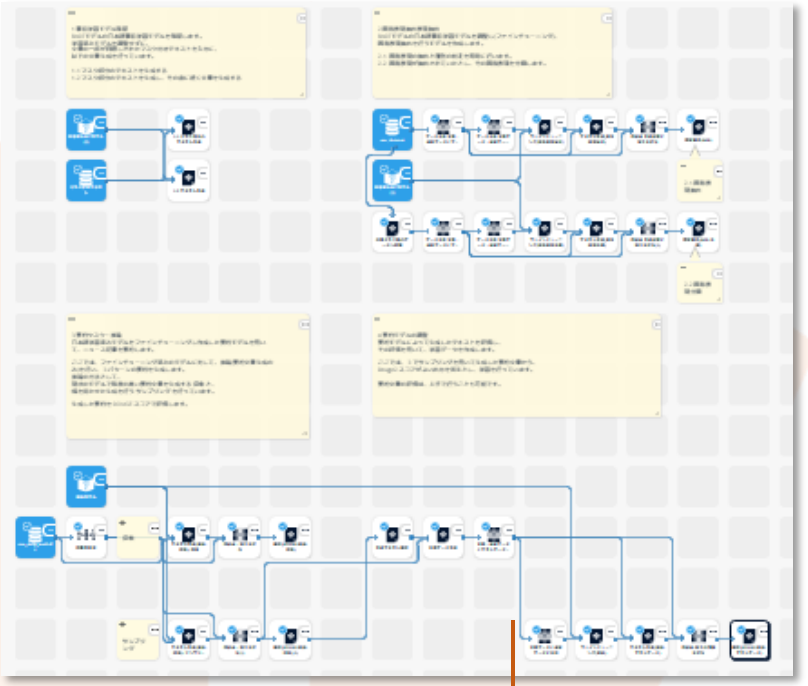
[3.]では、作成済みの要約モデルから要約したテキストを生成しましたが、[4.]では、生成したテキストを評価し、評価に合うようにモデルを再度ファインチューニングします。
複数生成された要約したテキストを評価し、一番よいものを選び、答え (目的変数, ラベル) とする要約テキストとします。

こうすることで、モデルが生成するテキストを評価に合わせたものにすることができます。評価を人手で行った場合、評価者が好みのテキストを生成するモデルに近づけることもできます。

要約したテキスト
(要約したテキストを3種類生成)

可視化画面:テキスト生成(要約-推論)-result ノード

No.	0 String	1 String	2 String
1	１５日に発表された国	１５日に発表された国際サ	１５日に発表された国際サッカー連盟（F
2	ブラジル サッカー 連盟は	ブラジル サッカー 連盟は１６	ブラジル サッカー 連盟は１６日、２０１８
3	ブラジル 代表 の センター	ブラジル 代表 の センター バッ	ブラジル 代表 の センター バック、ルシオが
4	リーネン 監督 が、宮市	リーネン 監督 が、宮市 亮 の	リーネン 監督 が、宮市 亮 の 復調 について言
5	岡崎 慎司 が ２７日、欧	岡崎 慎司 が ２７日、欧州 チ	岡崎 慎司 が ２７日、欧州 チャンピンズリーク
6	２０日 の レスター 戦で	２０日 の レスター 戦で、岡	２０日 の レスター 戦で、岡崎 慎司 が ２ゴー
7	レスター の ヴァーディー	レスター の ヴァーディー が、	レスター の ヴァーディー が、娘 に 対する脅
8	リバプール の シャビ・フ	リバプール の シャビ・アロン	リバプール の シャビ・アロンソ が、チーム
9	ブンデスリーガ ２部 の シ	ブンデスリーガ ２部 の シュツ	ブンデスリーガ ２部 の シュツットガルト は ９
10	２０００年に誕生した	２０００年に誕生した選手	２０００年に誕生した選手が欧州５大リー



4.

プロジェクト 解説

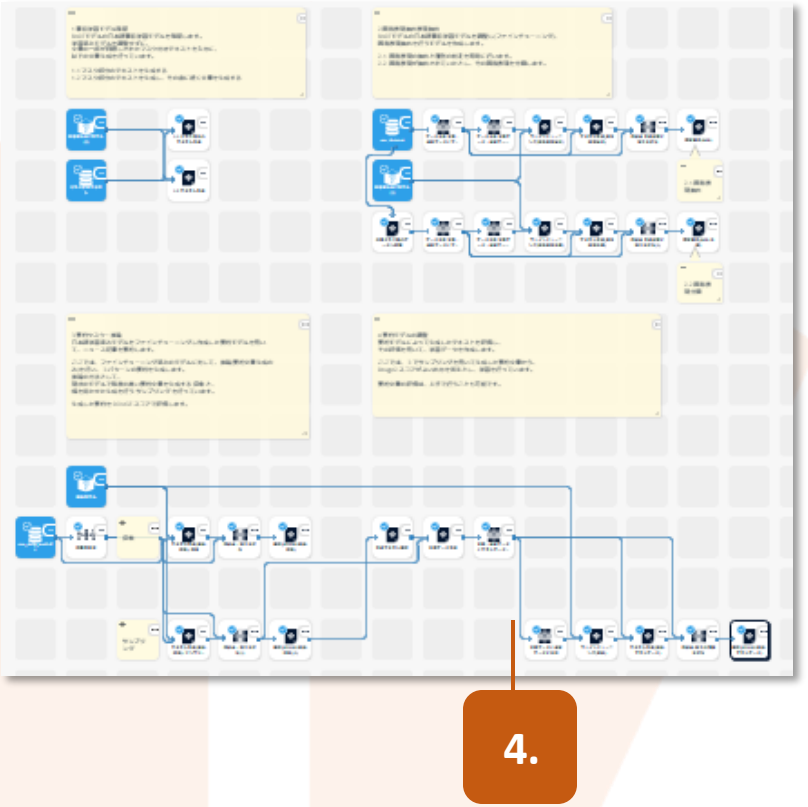
4. 要約モデルのファインチューニング

[4.] では、要約テキストの評価に、答えとなる要約文との一致度を測る Rouge スコアを用います。Rouge スコアにはいくつか種類がありますが、ここでは、 Rouge-2 を用います。Rouge-2 を算出し、Rouge-2 の値が高いテキストを答えの要約テキストとして、再度ファインチューニングを行います。

可視化画面:テキスト生成評価-result ノード

No.	rejected String	choice String
1	2	1
2	2	0
3	0	1
4	skip	skip
5	2	0
6	0	1
7	2	0
8	1	0
9	2	1
10	0	2

Rouge-2 が一番高い要約を”choice”、一番低い要約を”rejected”にする。Rouge-2 のスコアが同じ場合は、評価ができないため skip とし、新たな学習データには含めない。



可視化画面:学習データ生成-result ノード

学習データ生成-result 列数: 5 行数: 118

No.	prompt String	label String	choice String	rejected String	highlights_ja String
1	ロンドン（英国） （ロ	15年公開予定の続	15年公開予定の続	18歳の俳優ダニュ	ハリー・ポッターのス
2	編集部注：「Behind tl	米フロリダ州の刑	米フロリダ州の刑	アメリカの刑務所	マイアミの精神病患者
3	ミネソタ州ミネアポリ	アメリカで29日に	アメリカで29日に	3日、アメリカ中	NEW：「死ぬかと思っ
4	ワシントン（CNN） -	ジョージ・W・ブ	ジョージ・W・ブ	ブッシュ大統領の	検査中に5つの小さな
5	（CNN） -- ナショナル	CNNの報道によ	CNNの報道によ	NFLのスター選手	NEW：NFLのチーフと

ラベル (label) 列に “choice” で選んだ要約した文章が入り、ファインチューニングの目的変数とする

アウトプットの説明

アウトプット(1. 本プロジェクトで用いる事前学習済みモデル)

BART日本語Pretrainedモデルを用いて、マスク付きテキストのマスク部分生成と、その後に続く文章の生成を行います。このモデルは日本語でのテキスト生成を行うための土台となるものですが、間違った文章や日本語として意味の通らいない文章が生成されることもあります。このモデルをもとにして、モデルをファインチューニングし、固有表現抽出・要約を行うモデルを作成していきます。

1-1.マスク部分のテキスト生成

マスク部分のテキストを生成します。日本語として不自然な文章も散見されます。

可視化画面:1-1.マスク部分のテキスト生成-result ノード

No.	prompt String	0 String	1 String	2 String	3 String
1	天気が<mask>から散歩しましょう。	天気 が よく から 散歩 し ましょう 。	天気 が 良く から 散歩 し ましょう 。	天気 が 悪い から 散歩 し ましょう 。	天気 が 良い から 散歩 し ましょう 。
2	私の好きなスポーツは<mask>です。	私 の 好きな スポーツ は 、 です 。	私 の 好きな スポーツ は 私 です 。	私 の 好きな スポーツ は スポーツ です 。	私 の 好きな スポーツ は 1 です 。
3	休日はよく<mask>をします。	休日 は よく 考える を し ます 。	休日 は よく 考え を し ます 。	休日 は よく 知ら を し ます 。	休日 は よく ある を し ます 。

1-2.マスク部分のテキスト生成+後に続くテキストの生成*

マスク部分のテキストを生成し、その後に続く文章を生成します。

可視化画面:1-2.テキスト生成-result ノード

No.	prompt String	0 String
1	天気が<mask>から散歩しましょう。	天気 が よく なって から 散歩 し ましょう 。 天気 の よく 合って いる ところ に いる から 、 そこ から は 、 ここ から とれ ます 。
2	私の好きなスポーツは<mask>です。	私 の 好きな スポーツ は 、 私 が 好きな もの は 私 です 。 私 は 好きな 人 です 」 。 また 、 。 そして 、 それ は 。 （ 中略 ） 。 ）
3	休日はよく<mask>をします。	休日 は よく 考える 日 に し ます 。 休日 の よく 考え ます ように 、 よく 考えて ください 。 」 と お 願い を し ました 。 これ は 、

*テキスト生成アイコンでは、パラメータ設定画面で探索とサンプリングが選択できます。探索は、対象テキストに続く確率が一番高い単語(トークン)を探索することテキストを探索し生成します。サンプリングでは、対象テキストに続く単語の確率にもとづき単語を選びます。探索では、確率の高い単語を探索し生成するため、似たようなテキストが生成されやすいですが、サンプリングは単語の確率で重み付けしテキスト生成を行うので、生成されるテキストにバリエーションが得やすくなります。

アウトプット(2-3. 固有表現抽出モデル)

Wikipedia を用いた日本語の固有表現抽出データセットを用いて、BART日本語Pretrainedモデルをファインチューニングし、テストデータに対して固有表現抽出を行います。1つの文章に対し、 3種類の固有表現ラベル (label) を生成しています。3種類の候補を生成すれば、 6割以上のケースで正しく固有表現を抽出できていることがわかります。

テストデータ :

可視化画面:ner_dataset-data ノード 対象とするテキスト(prompt) 列

No.	prompt String
1	かつては東京都三鷹市にオフィスを構えていたが、大月のキングレコード退職を機に事業を停止。
2	2015年5月7日、第一汽車の新董事長に東風汽車の前董事長だった徐平が任命された。
3	文禄・慶長の役では子・種量らと共に小西軍として従軍したが順天城の戦いで戦死した。
4	文禄・慶長の役では子・種量らと共に小西軍として従軍したが順天城の戦いで戦死した。
5	なお富士製紙釧路工場は、吸収合併により王子製紙釧路工場、戦後の財閥解体により十条製紙釧路工場、さらに合併により日本製紙釧路工場と変遷している。
6	ゲイリーは午後に戻りIBMと協議しようとしたが、彼が秘密保持契約にサインしたかどうか、彼がIBMの代表と会ったかどうかについては、デジタルリサーチ側とIBM側とで話が相反している。
7	国防省管轄下のイスラエル国防軍参謀本部に在り軍事情報を担当。
8	一方、法務省入国管理局によれば、1978年、初めて韓国・朝鮮籍2人が退去強制により送還され、その後1988年までにさらに17人が送還されたとの記録がある。

出力結果(3 つの候補を生成) :

可視化画面:テキスト生成(固有表現抽出)-result ノード

No.	0 String	1 String	2 String
1	東京 都 三鷹 市 は 地名 です 。	キングレコード は 法人 名 です 。	大月 は 人名 です 。
2	第 一 汽車 は 法人 名 です 。	東風 汽車 は 法人 名 です 。	徐平 は 人名 です 。
3	種 量 は 人名 です 。	順天 城 の 戦い は イベント 名 です 。	種量 は 人名 です 。
4	種 量 は 人名 です 。	順天 城 の 戦い は イベント 名 です 。	種量 は 人名 です 。
5	釧路 工場 は 施設 名 です 。	十 条 製紙 は 法人 名 です 。	日本 製紙 は 法人 名 です 。
6	ＩＢＭ は 法人 名 です 。	デジタル リサーチ は 法人 名 です 。	ゲイリー は 人名 です 。
7	国防 省 は 政治 的 組織 名 です 。	イスラエル 国防 軍 参謀 本部 は 政治 的 組織 名 です 。	イスラエル 国防 軍 は 政治 的 組織 名 です 。
8	韓国 は 地名 です 。	韓国 ・ 朝鮮 は 地名 です 。	法務 省 入国 管理 局 は 政治 的 組織 です 。

精度の評価 :

可視化画面:精度評価(NER)-result

No.	1番目の候補に正解がある割合 Float	1から2番目の候補の中に正解がある割合 Float	1から3番目の候補の中に正解がある割合 Float
1	0.206439	0.458333	0.602273

アウトプット(2-4. 与えられた固有表現に対して分類を与えるモデル)

p.8 のように入力データを変更して、BART日本語Pretrainedモデルをファインチューニングし、固有表現の分類を行います。こちらも、1つの固有表現に対し、 3種類の固有表現ラベル (label) を生成しています。

テストデータ :

可視化画面:分類タスク用のデータへ調整-resultノード

No.	prompt String	label String
1	かつては東京都三鷹市にオフィスを構えていたが、大月のキングレコード退職を機に事業を停止。大月は	大月は人名です。
2	2015年5月7日、第一汽車の新董事長に東風汽車の前董事長だった徐平が任命された。徐平は	徐平は人名です。
3	文禄・慶長の役では子・種里らと共に小西軍として従軍したが順天城の戦いで戦死した。文禄・慶長の役は	文禄・慶長の役はイベント名です。
4	文禄・慶長の役では子・種里らと共に小西軍として従軍したが順天城の戦いで戦死した。小西軍は	小西軍は政治的組織名です。
5	なお富士製紙釧路工場は、吸収合併により王子製紙釧路工場、戦後の財閥解体により十条製紙釧路工場、さらに	釧路工場は施設名です。
6	ゲイリーは午後に戻りIBMと協議しようとしたが、彼が秘密保持契約にサインしたかどうか、彼がIBMの代表と	IBMは法人名です。
7	国防省管轄下のイスラエル国防軍参謀本部に在り軍事情報を担当。イスラエル国防軍参謀本部は	イスラエル国防軍参謀本部は政治的組織名です。
8	一方、法務省入国管理局によれば、1978年、初めて韓国・朝鮮籍2人が退去強制により送還され、その後1988年	朝鮮は地名です。
9	不動産だけでなく、巨額の粉飾決算で揺れていた東芝の創業事業である家電部門への海信集団や美的集団による	海信集団は法人名です。
10	不動産だけでなく、巨額の粉飾決算で揺れていた東芝の創業事業である家電部門への海信集団や美的集団による	タカタは法人名です。

出力結果(3 つの候補を生成):

可視化画面:テキスト生成(固有表現分類)-resultノード

No.	0 String	1 String	2 String
1	朝鮮は地名です。	韓国は地名です。	朝鮮は国名です。
2	キー・セーフティ・システムズは法人名	キー・セーフティー・システムズは法人	キーセーフティ・システムズは法人名です。
3	コナミは法人名です。	コナミはその他の組織名です。	コナミは政治的組織名です。
4	佐竹晴雄は人名です。	佐竹晴雄は人名です。	佐竹晴雄は人名です。
5	テレビ新広島は法人名です。	テレビ新広島は地名です。	テレビ新広島はその他の組織名です。
6	船田一雄は人名です。	船田一雄は法人名です。	船田一雄は人名です。
7	スペイン選手団はその他の組織名です。	スペイン選手団は政治的組織名です。	スペイン選手団はイベント名です。
8	菊栄は人名です。	菊栄は法人名です。	菊栄は地名です。
9	東京新聞社は法人名です。	株式会社東京新聞社は法人名です。	東京新聞は法人名です。
10	ジャネット・スーは人名です。	ジャネット・スーは人名です。	ジャネット・スーは人名順です。

精度の評価 :

可視化画面:精度評価(NER-分類)-resultノード

No.	1番目の候補に正解がある割合 Float	1から2番目の候補の中に正解がある割合 Float	1から3番目の候補の中に正解がある割合 Float
1	0.914773	0.945076	0.958333

アウトプット(3. 要約モデルでのテキスト生成)

CNN/Daily Mail データセットに要約モデルを用いて、要約されたテキストの生成を行っています。出力は、 1つの入力文章に対し、3種類の要約テキストを生成しています。また、要約タスクの評価指標として用いられる Rouge スコアを出力しています。

要約文章(3つの候補を生成) :

可視化画面:テキスト生成(要約-推論) -探索 -resultノード

No.	0 String	1 String	2 String
1	18歳になった俳優ダニエル	18歳になった俳優ダニエル	18歳になった俳優ダニエル
2	マイアミ・デイド州の拘置施	マイアミ・デイド州の拘置施	マイアミ・デイド州の拘置施
3	ミネアポリスの橋が崩壊した	ミネアポリスの橋が崩壊した	ミネアポリスの橋が崩壊した
4	ブッシュ大統領の大腸から5	ブッシュ大統領の大腸から5	ブッシュ大統領の大腸から5
5	NFLのスター選手、マイケル	NFLのスター選手、マイケル	NFLのスター選手、マイケル
6	イラク人家族が、重度の火傷	イラク人家族が、重度の火傷	イラク人家族が、重度の火傷
7	3児の母であるカリマ・スハ	3児の母であるカリマ・スハ	3児の母であるカリマ・スハ
8	米国の麻薬取引で起訴された	米国の麻薬取引で起訴された	米国の麻薬取引で起訴された

Rouge スコア :

可視化画面:評価(ROUGE-要約-推論)-resultノード

No.	0_rouge1 Float	0_rouge2 Float	0_rougel Float	1_rouge1 Float	1_rouge2 Float	1_rougel Float	2_rouge1 Float	2_rouge2 Float	2_rougel Float
1	0.209677	0.016393	0.129032	0.201681	0.017094	0.100840	0.196721	0.016667	0.098361
2	0.213333	0.027027	0.173333	0.206452	0.026144	0.167742	0.207792	0.026316	0.168831
3	0.234375	0.031746	0.156250	0.238095	0.032258	0.158730	0.248062	0.031496	0.155039
4	0.213740	0.015504	0.137405	0.215385	0.015625	0.138462	0.222222	0.015038	0.133333
5	0.244898	0.013793	0.149660	0.255034	0.027211	0.147651	0.246575	0.013889	0.150685
6	0.177778	0.000000	0.133333	0.175182	0.000000	0.131387	0.172662	0.000000	0.129496
7	0.175182	0.000000	0.145985	0.189781	0.000000	0.160584	0.172662	0.000000	0.158273
8	0.217949	0.051948	0.128205	0.216561	0.051613	0.127389	0.222222	0.052980	0.130719

要約文章 0〜2 の3種類に対しての
Rogue-1, Rogue-2, Rogue-L
3種類のスコアを出力しています。
いずれも値が1に近いほど
評価の結果は良好です。

アウトプット (3. 要約モデルでのテキスト生成)

要約文章から一例を抜粋して示します。この例では比較的よく要約できているといえますが、生成された要約テキストの中には、意味が逆になっているものもあります。

入力(prompt) :

『米地質調査所の報告によると、金曜日午前4時42分（日本時間午前7時42分）にマグニチュード4.2の地震がサンフランシスコ周辺を揺らした。この地震により、約2000人の顧客が停電したと、パシフィック・ガス・アンド・ライト社のスポークスマン、デビッド・アイゼンハワー氏は述べた。USGSの分類では、マグニチュード4.2の地震は「軽い」とされ、通常、被害は最小限にとどまるとしています。サンフランシスコ警察のアル・カシアート警部は、「私たちは、通報のかなりの急増がありましたが、ほとんどが問い合わせの通報で、報告されたいかなる怪我やいかなる損害もありませんでした」と述べました。"かなり穏やかだった"。警察が地震直後の心配な電話について説明するのを見る".地震はオークランドから東北東に約2マイル、深さ3.6マイルを震源とすると、USGSは言った。オークランドはサンフランシスコのすぐ東、サンフランシスコ湾を挟んで反対側にある。オークランド警察の派遣隊員はCNNに、この地震で人々の家の警報が鳴った、と語った。CNNの気象学者チャド・マイヤーズ氏は、揺れは約50秒続いたと述べた。USGSによると、マグニチュード4.2の地震は室内で感じられ、食器や窓を割ったり、不安定なものをひっくり返したりすることがある。振り子時計が止まることもある。』

要約テキスト:

『米地質調査所の報告によると、金曜日午前4時42分にマグニチュード4.2の地震がサンフランシスコ周辺を揺らした。USGSの分類では、地震は「軽い」とされ被害は最小限にとどまるとしている。この地震により、約2000人の顧客が停電したという。』

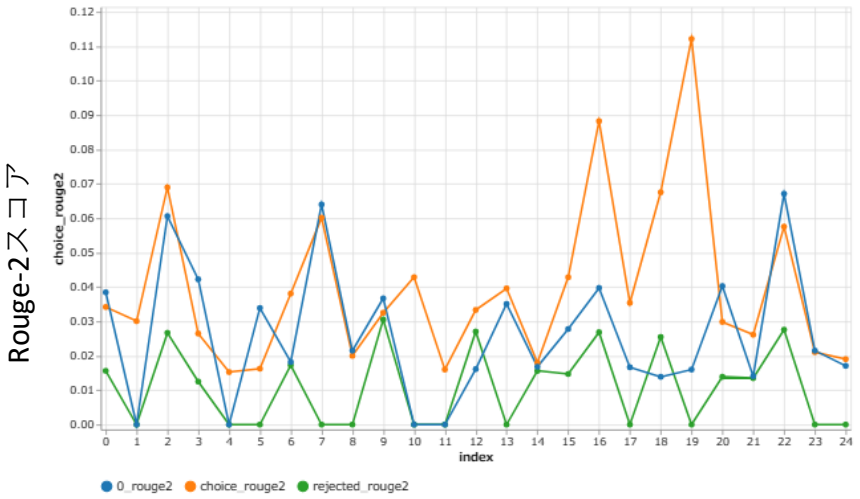
正解要約(label) :

『2,000人が停電、電力会社が発表.マグニチュード4.2の地震で家の警報が鳴った、オークランド警察の派遣者が言う."かなり軽かった"と警察は言う、負傷者や損害の即時報告なし.オークランドの東北東2マイルが中心、深さは約3.6マイルだった.』

アウトプット (4. 要約モデルのファインチューニング)

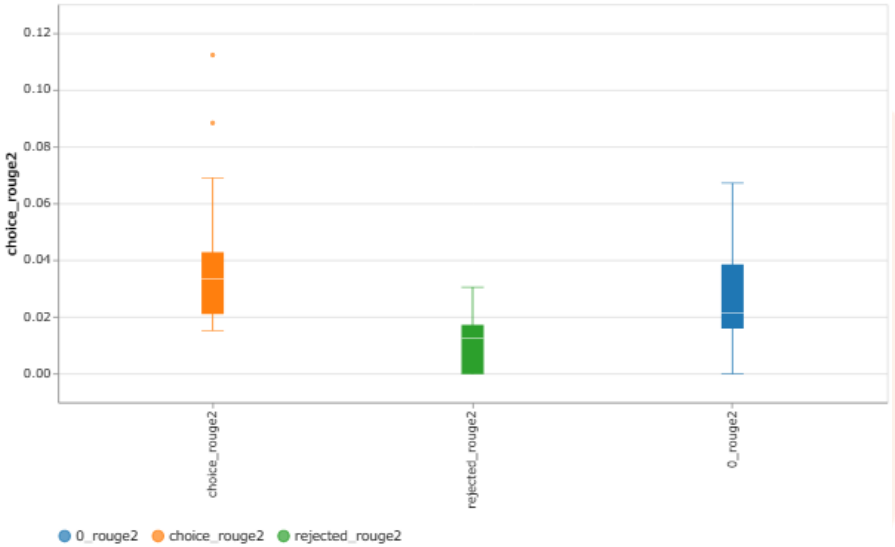
[3.] で生成した要約テキストの中で、Rouge-2 スコアの高かった要約テキストを答えとして作った学習データを用い、要約モデルを調整しました。そのモデルを使い、生成した要約テキストの Rouge-2 スコアを算出しました。
調整前のモデルにおいて3種類の要約文を生成した中で Rogue-2スコア の値が最大であったものの値とその分布を下図橙色の choice_rouge2 として、Rogue-2スコアの値が最小であったものの値とその分布を緑色の rejected_rouge2 として示しています。調整後のモデルで算出した青色の 0_rouge2 では、Rogue-2スコアの値が平均的に上昇していることがわかります。

可視化画面:グラフ一覧
折れ線[評価(ROUGE-要約-テストデータ)-result_y(choice_rouge2,rejected_rouge2,0_rouge2)]



各テストデータ毎の Rouge-2 スコア

可視化画面:グラフ一覧
箱ひげ図[評価(ROUGE-要約-テストデータ)-result_y(choice_rouge2,rejected_rouge2,0_rouge2)]



Rouge-2 スコアの分布

アイコンの設定

アイコンの設定

ファインチューニングアイコンは、説明変数としてテキストを入力し、それに対応したテキストを出力するように学習したモデルを出力します。
ファインチューニングアイコンのパラメータは以下です。

パラメータ名	値の範囲	説明	デフォルト値
目的変数	列を指定	目的変数列を選択する	“label”
説明変数	列を指定	説明変数列を選択する	“prompt”
エポック数	1 以上の整数値	学習データを学習する繰り返し回数を設定する	4
学習率	0 より大きい実数値	学習時のモデルパラメータの更新幅の初期値を設定する	2.0e-5
最終層以外のパラメータの更新を停止する	True/False	True にすると、ファイチューニング実行時にモデルの最終層以外のパラメータ更新を停止する	False
目的変数の最大系列長	1 以上の整数値	目的変数値の長さの上限値を設定する 長い場合は打ち切りを行う	1024
説明変数の最大系列長	1 以上の整数値	説明変数値の長さの上限値を設定する 長い場合は打ち切りを行う	128
学習データのバッチサイズ	1 以上の整数値	一度に計算する学習データの行数を設定する	16
検証データのバッチサイズ	1 以上の整数値	一度に計算する検証データの行数を設定する	16
正則化項の重み	[0,1] の範囲の実数値	ロス関数を汎化させるための重みを設定する	0.1
GPU 使用フラグ	True/False	GPU を使用するか否かを設定する True の場合でも、 GPU が使用できる環境でないと使用できない	False
半精度浮動小数点数の使用フラグ	True/False	GPU を使用するときに、精度を落とした演算(半精度浮動小数点数演算)を使用する	False
乱数シード値	0 以上の実数値	乱数シードを設定する	1
説明変数につける接頭語	選択肢	タスク毎に説明変数の先頭につける接頭語を選択する	要約

・パラメータ設定画面

アイコンの設定

テキスト生成アイコンは、学習済みモデルと入力のテキストを用いて、文章を出力します。パラメータは以下です。

パラメータ名	値の範囲	説明	デフォルト値
サンプリングを行う	True/False	サンプリングを行うか、探索を行うかを設定する	True
最小の系列長	0 より大きい整数値	生成するテキストの最小の系列長を設定する	10
最大の系列長	0 より大きい整数値	生成するテキストの最大の系列長を設定する	100
生成する系列数	0 より大きい整数値	生成する系列の個数を設定する 探索において、ビーム数より値が大きい場合、ビーム数が優先される	3
説明変数につける接頭語	選択肢	タスク毎に説明変数の先頭につける接頭語を選択する	要約
GPU 使用フラグ	True/False	GPU を使用するか否かを設定する True の場合でも、 GPU が使用できる環境でないと使用できない	False
温度	0 より大きい実数値	サンプリングの際、温度が高いと、確率値が低いトークンが選択されやすくなる	1.0
サンプリング時の候補数	0 以上の整数値(0 で設定が無効)	サンプリングで入力テキストに続く単語(トークン)を、確率の高い候補を保持する個数を設定する	50
サンプリング時の系列に対する累積確率値の下限値	(0,1) の間の実数値	サンプリングで入力テキストに続く単語(トークン)を生成する時、累積確率値が設定値より高いものを選択する	0.975
ビーム数	0 より大きい整数値	探索する場合に保持する個数を設定する	3
同一の n グラムが再出現する上限値	0 以上の整数値	生成される系列中で、同じトークンが生成される個数を設定値より小さくする	2
earlystopping フラグ	True/False	True ：ビーム数分の候補が見つかったら、すぐに探索を終了する False：適切な候補が見つからない可能性が高い場合に探索を終了する	True

・パラメータ設定画面

補足情報

技術的な情報や利用規約について

1. Transformer モデルと言語モデル

大規模言語モデルの構築は、Transformer と呼ばれるモデルを基にした手法が多く用いられています。左下図が Transformer モデルの構造ですが、これは入力テキストを低次元の特徴量に圧縮する Encoder と呼ばれる部分と、低次元の圧縮された情報からテキスト情報への復元を行う Decoder と呼ばれる部分から成り立っています。この Transformer を基にしたモデルは、Encoder 構造を用いたモデル、Decoder 構造を用いたモデル、Encoder-Decoder 構造を用いたモデルの3系統に主に分けられます。

Transformer モデル

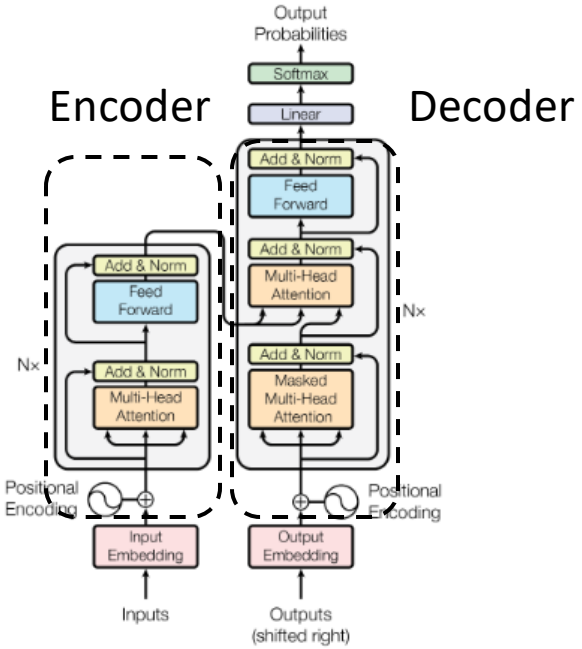
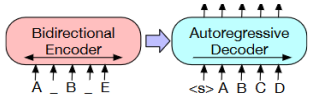


Figure 1: The Transformer - model architecture.

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Encoder 構造　　Decoder 構造



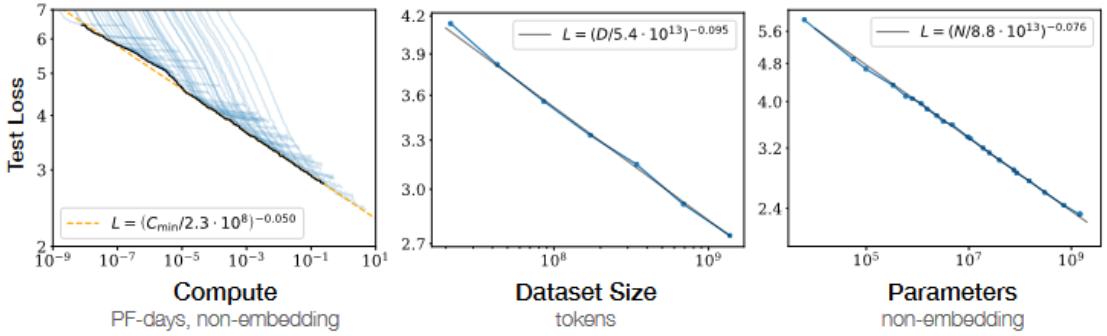
Encoder-Decoder 構造

Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." arXiv preprint arXiv:1910.13461 (2019).

構造	特徴	主な言語モデル
Encoder 構造	事前学習でマスクされた部分のテキストを予測する。 得意なタスク：文書比較や分類	BERT, Distil BERT, XLM, DLECTRA, …
Decoder 構造	ある文章の「次に続く単語」の予測を行う 得意なタスク：テキスト生成	GPT, GPT-2, GPT3, …
Encoder-Decoder 構造	入力と出力の長さを変えられるので、様々な事前学習が行える 得意なタスク：要約や翻訳	T5, BART, BigBird, …

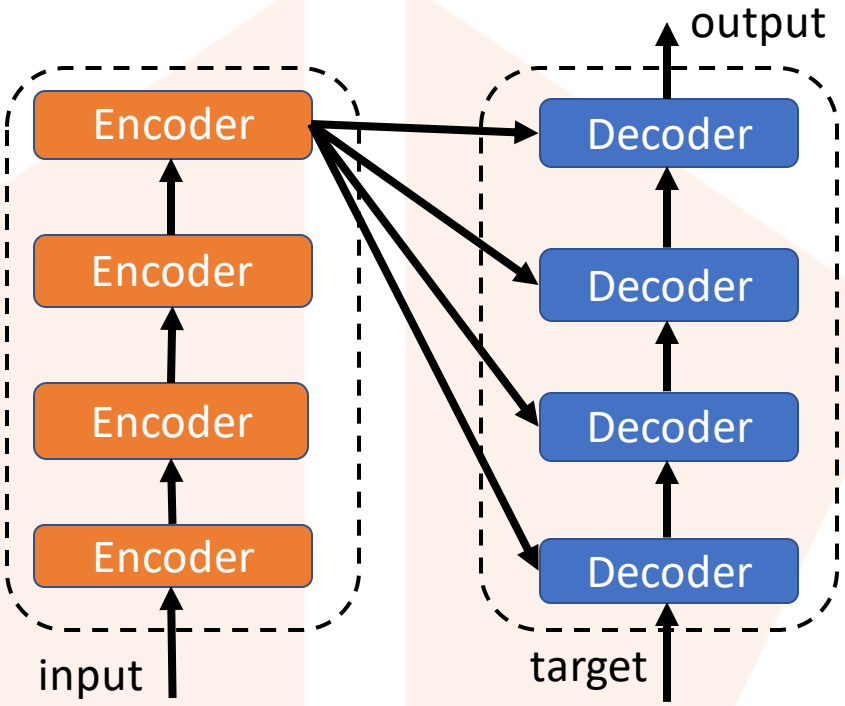
2. 言語モデルの大規模化

Transformer 系のモデルでは、計算量やデータサイズ、パラメータ数、モデルのロス値（誤差値）との間で実験的にベキ乗則が成り立っていることが報告されています。データサイズやパラメータ数を増やせば増やすほど、学習に使われなかったデータでのロス値が下がり、汎化性能が向上していくという状況が続いています。さらに、同程度の計算リソースを消費してモデル構築を行う場合、データサイズを増やすことよりパラメータ数を増やす方がロス値を下げるために効率がよいとされていることもあり、モデルは近年特に大規模化していく傾向にあります。



左からコンピュータの計算量、データサイズ、パラメータを増やした時の、学習時に使われなかったデータでのロス値(クロスエントロピー誤差)

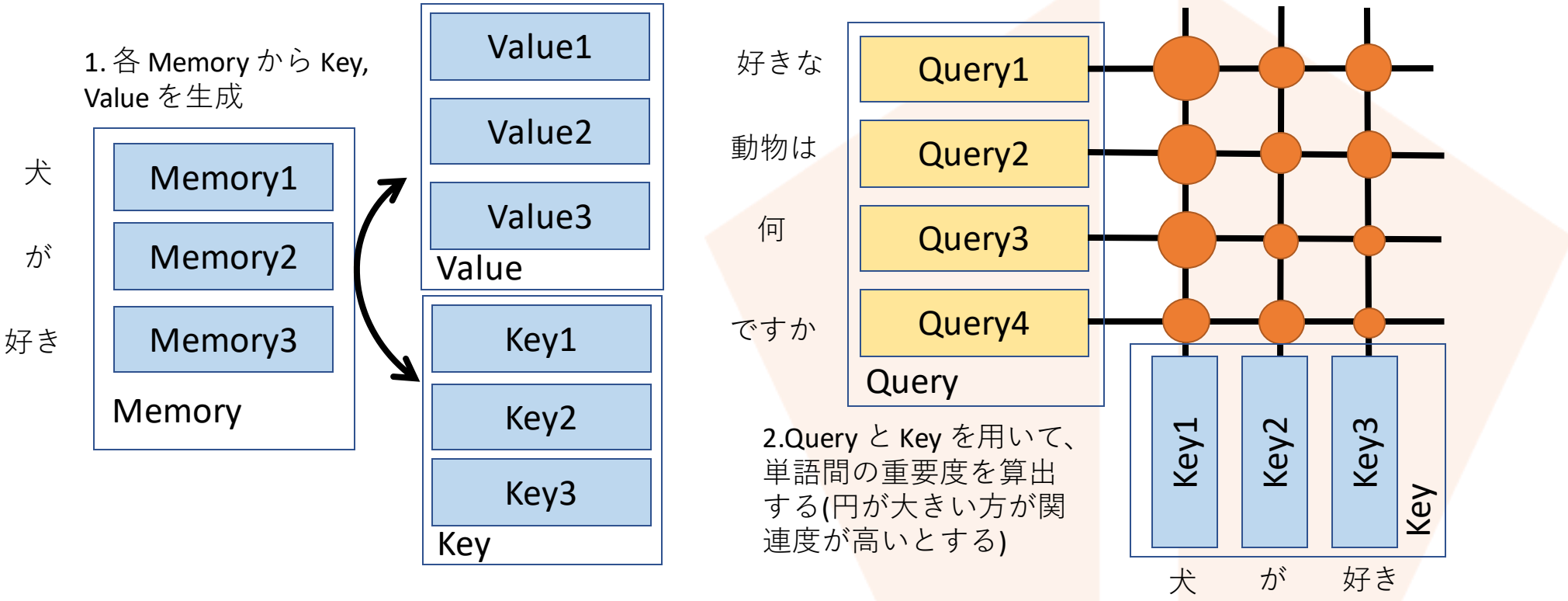
Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.



Encoder-Decoder モデルの構造
Encoder層 や Decoder層 の個数を増やすことで
パラメータ数も増大する

3. Attention 機構

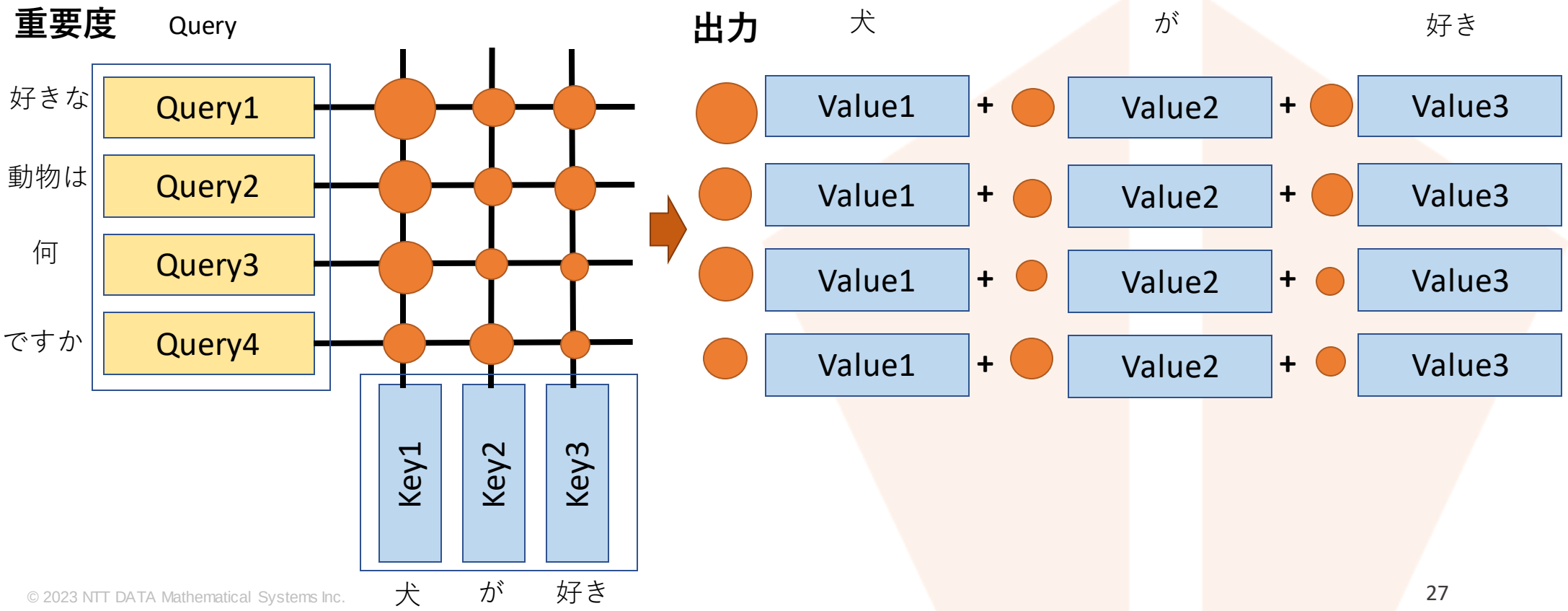
Transformerモデル の中核を成すのが Attention機構 です。Transformerモデルの Decoder 部分で用いられる Source-Target Attention機構 では、探索の対象となる文章を構成する単語から、検索用の Key と情報本体を表す Value を生成します。そして、入力文の単語(下図の Query1～4)に対して、対応する Query の単語と Key の単語との関連度・重要度が高くなるような学習を行います。ここで、Query, Key, Value は全て低次元の特徴量として表現されます。この重要度は、どの単語に注目(= attention)したらよいかという重みととらえることができます。



3. Attention 機構

前頁のステップで単語の重要度を算出したら、情報本体を表す Value にその重みを乗じて出力を生成します。

これらの処理は入力テキストを一度にまとめて処理する機構となっており、この部分がテキストを単語ごとに逐次処理していく手法 (Alkano「ディープラーニング 時系列/テキスト」に搭載されている RNN, GRU, LSTM) とは異なる点です。



4. Rouge スコア

Rouge スコアは、生成要約の精度を正解要約と比較することで求める精度指標です。本プロジェクトでは、Rouge-1, Rouge-2, Rouge-L というスコアを用いています。

- Rouge-1 ：生成要約と正解要約の間の 1-gram（単語、トークン）の共起を評価します。
- Rouge-2 ：生成要約と正解要約の間の 2-gram (隣合う 2 つの単語)の共起を評価します。
- Rouge-L ：生成要約と正解要約間で文の順番に沿って共起している単語の個数の最大値で評価します。

それぞれ、生成要約の n-gram 数/ 単語数での共起部分の割合(precision)と正解要約の n –gram 数 / 単語数での共起部分の割合(recall)の調和平均を取った F1 スコアを用います。

- Rouge-1 計算例 (1 単語毎比較)

生成文:部長/が/沖縄 /に/到着/し/た
正解文: 山田/部長/は/今日/ 沖縄 /に/ 到着/し/ま し/た

共起単語数 : 6 , 生成文の単語数 : 7, 正解文の単語数 :10
precision : (共起単語数)/(生成文単語数) = 6/7
recall: (共起単語数)/(正解文単語数) = 6/10
Rouge-1 : 2/(1/(precision)+1/(recall))≒0.706

- Rouge-L 計算例(共起、一番長い共通部分で比較)

生成文 :部長/が/沖縄/に/到着/し/た
正解文: 山田/部長/は/今日/沖縄/に/到着/し/ま し/た

共起単語数 : 4 , 生成文の単語数 : 7, 正解文の単語数 :10
precision : 4/7, recall: 4/10, Rouge-L: 0.471

- Rouge-2 計算例 (隣合う 2 単語毎比較)

生成文 :
部長/が
が/沖縄
沖縄/に
に/到着
到着/し
し/た

正解文 :
山田/部長
部長/は
は/今日
今日/沖縄
沖縄/に
に/ 到着
到着/し
し/まし
まし/た

共起 2-gram 数 : 3 , 生成文の 2-gram 数 : 6, 正解文の 2-gram 数 :9
precision : 3/6, recall: 3/9, Rouge-2 : 0.4

本文書・プロジェクトファイルのご利用にあたって

本文書ならびにプロジェクトファイルは、（株）NTT データ数理システム （以下「弊社」）が開発・販売する分析プラットフォーム Alkano についての情報提供として弊社が作成を行ったものです。弊社による事前の許可なしに、本文書の再配布や引用の範囲を超える複製といった行為、およびリバースエンジニアリングを禁じます。

本文書ならびにプロジェクトファイルのご利用に際して、ご利用者様および第三者に損害が発生したとしても、弊社は責任を負わないものとします。

プロジェクトファイルは、その中に同梱されているデータを利用し、本文書内で解説している設定可能なパラメータで動作させた場合についてのみ、弊社にて動作の検証を行っております。これを超えるような状況における動作は保証いたしません。

本プロジェクトファイル は、MSIP1.8.0 および Alkano1.2.0 にて動作確認を行っております。

【アルカノ】

Alkano

データ活用の確かなパートナー

お問い合わせ: 株式会社NTTデータ数理システム　営業部
Tel: 03-3358-6681
E-mail: alkano-info@ml.msi.co.jp
WEB: https://www.msi.co.jp/alkano/