

機械学習を用いた日本語 文章の抽象型要約

情報システム学系 4年
知能情報システム研究室 永井ゼミ
J19261 孫宇杰

目次

- 研究の動機と目的
- 使用する言語環境とモデル
- 二種類の要約事例
- 今どこまでやっている
- 問題点と提案
- 検証手法
- 関連研究
- これからの研究計画

研究の動機と目的

■ 動機

- 自然言語処理に興味があるので、機械がどうやって自然言語を取り扱うのか勉強したいので、文章の自動要約を研究することにした。

■ 目的

- 文章をモデルに入力して要約文を生成する。

使用する言語環境とモデル

- Python
- Google T5モデル
 - T5 (Text-to-Text Transfer Transformer) : テキストを入力してテキストを出力するタスクを解く自然言語処理深層学習モデル。

環境

- Google Colab

二種類の要約事例

元の文章

牧場で10日、毎年恒例のヒツジの毛刈りがあり、多くの家族連れで賑わった。東海地方は晴れて5月上旬並みの暖かさに。11日も晴天が続き暖かくなりそうだ。

ヒツジたちの夏バテ防止の為、毎年この時期に毛刈りが行われている。牧場の職員がバリカンで1頭あたり40分をかけ、合計30頭の毛を刈った。抵抗したり暴れたり、ヒツジたちはにぎやかに鳴いていたが、多くは毛刈りを終えると気持ちよさそうに走り回っていた。

刈った毛はその場で販売されたり、牧場内の羊毛工芸体験の原料として使われたりする。毛刈りは11日にも午前と午後の2回行われる。

抽出型

牧場で10日、毎年恒例のヒツジの毛刈りがあり、多くの家族連れで賑わった。ヒツジたちの夏バテ防止のため、毎年この時期に毛刈りが行われている。刈った毛はその場で販売されたり、羊毛工芸体験の原料として使われたりする

抽象型

10日、牧場では、毎年恒例の羊たちの毛刈りがあり、多くの家族連れで賑わった。

毛刈りは夏バテ防止を目的として行われている。

刈った毛はその場で販売、または羊毛工芸体験の原料となる。

毛刈りは11日にも、午前と午後の2回実施を予定している。

WikiHowのデータセットを使ったT5モデルFine-Tuning実行結果

↑ ↓ ↺ 🗨 ⚙ 📄 🗑 ⋮

▶

```
for output, target, input in zip(outputs, targets, inputs):  
    print("generated: " + output)  
    print("actual: " + target)  
    print("body: " + input)  
    print()
```

🔗

generated: ボランティア募集情報を確認する。組織についてよく理解する。組織のウェブサイトをチェックする。採用責任者と連絡先を確認する。
actual: 募集されているボランティアの職を調べる。選んだ会社や団体組織をよく調べる。連絡先を確認する。
body: 関心のある会社や団体組織のウェブサイトを確認しましょう。ボランティア募集情報は通常の求人リストと同じページに、あるいはボランティア専用の求人ページに掲載されているでしょう。複数の仕事内容を調べて自分に最適なものを探

generated: 自分をもっと好きになる時間を作りましょう。自分磨きをしましょう。友達ゾーンから脱出しましょう。
actual: 自己改善の時間を作る。距離を置く。新しい第一印象を与える。簡単に時間を与えない。時間を無駄にしない。彼女の決断を受け入れる。
body: 彼女という時間(もしくは彼女のことを考えている時間)を、自分をもっと好きになる時間に使いましょう。彼女があなたに興味があるかどうかに関わらず、一日の終わりには自信に満ちた新しい自分に成長できるのです。これは彼女のため

generated: 誤字・脱字がないかチェックする。封筒にエントリーシートを入れる。送付状を封筒に入れる。本文を書く。
actual: 間違いがないか見直す。郵送で提出する。メールで提出する。
body: 記入が終わったら、誤字・脱字がないか改めてチェックしましょう。書いていて間違いが発生した場合は、基本は1から書き直します。ただ、エントリーシートが企業独自なもので、1人につき1枚のみと、予備がない場合もあります。そのよ

generated: 髪の毛の長さを測りましょう。
actual: どのような長い髪型を描きたいのか、そしてどちらに毛が流れているのかを想像します。
body: まずはシンプルに、長い髪が斜めに流れてできる曲線を描くようにしましょう。

generated: レモンの皮を剥きます。
actual: レモン果汁を凝乳に絞り入れてよく混ぜます。凝乳を髪に塗ります。何度か繰り返します。
body: レモンの皮を一片、切り取っておきます。レモンの皮を使ってすり込むようにまんべんなく塗ります。やがてフケがなくなります。

generated: 出生時に両親が日本人であった場合(出生前養子縁組)は、日本国籍を失うことになります。
actual: 両親のどちらかが日本人である。父親が日本人である。日本で生まれた場合。
body: 申請者の出生時に両親のどちらかが日本人であったならば、申請者は自動的に日本の国籍を取得します。日本の国籍法第2条2項は、日本人の父親を持つが、出生前に父親が死亡した場合、その子は出生した段階で日本国民になると定めてい

generated: 材料を準備する。図柄を描く。図柄を切り抜く。図柄を切り抜く。タブを切り抜く。両面テープを貼り付ける。
actual: クラフト用紙または市販のカードを半分に折る。飛び出す飾りを描く。飛び出す飾りの左右に小さなタブを描く。飾りを切り抜いて半分に折る。図柄の中央に直線を引き、半分に折る。図柄が字に立ち上がるようにタブを裏面に折り込む。
body: すでに作ったまたは購入したカードに飛び出す飾りを付けたい場合は、材料を作業台に準備します。シンプルなカードを手作りするには、厚手のクラフト用紙1枚を半分に折りましょう。標準的なカードを作るには短辺どうしを合わせて折り

WikiHowのデータ構造

How to 医療用マスクを着用する方法

方法1: マスクを着用する

1. 手を洗う

清潔なマスクを着用する前に、まずせっけんと水で手を洗いましょう。せっけんを手につけ、20秒ほど手指をこすり合わせましょう。最後に水で洗い流します。必ず清潔なペーパータオルを使って水気を拭き取りましょう。使ったペーパータオルはゴミ箱に捨てます。

2. マスクの状態を確認する

箱や袋から新品のマスクを取り出し、欠陥（穴や破れなど）が見られないか確認しましょう。マスクに何らかの欠陥がある場合は使わずに破棄し、新しいものを再度取り出しましょう。

...

方法2: マスクを取り外す

1. 手を洗う

マスクを取り外す直前に何をしていたかによっては、まず手を洗った方が良いでしょう。あるいは、手袋を取り外してから手を洗い、最後にマスクを取り外しましょう。

...



How to 医療用マスクを着用する方法

Article 1:

清潔なマスクを着用する前に、まずせっけんと水で手を洗いましょう。せっけんを手につけ、20秒ほど手指をこすり合わせましょう。最後に水で洗い流します。必ず清潔なペーパータオルを使って水気を拭き取りましょう。使ったペーパータオルはゴミ箱に捨てます。箱や袋から新品のマスクを取り出し、欠陥（穴や破れなど）が見られないか確認しましょう。マスクに何らかの欠陥がある場合は使わずに破棄し、新しいものを再度取り出しましょう。...

Summary 1:

手を洗う。マスクの状態を確認する。...

Article 2:

マスクを取り外す直前に何をしていたかによっては、まず手を洗った方が良いでしょう。あるいは、手袋を取り外してから手を洗い、最後にマスクを取り外しましょう。...

Summary 2:

手を洗う。...

今どこまでやっている

WikiHowのデータセットを使ったT5モデルFine-Tuning生成例

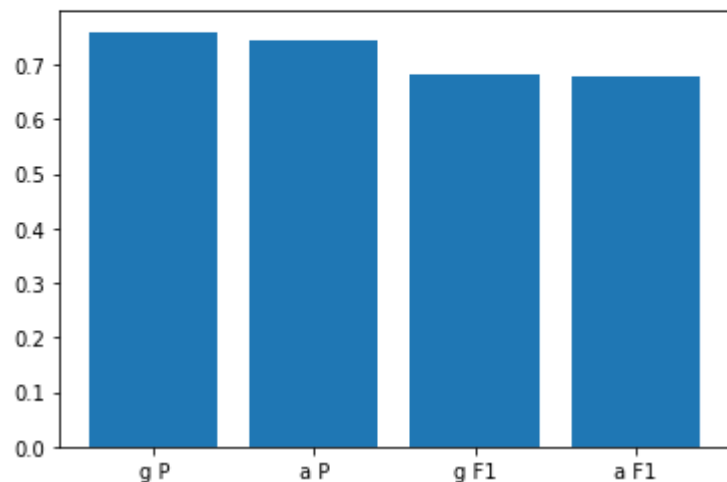
本文	コミュニケーションとは、いろいろな手段(文字、言葉を使わない合図、話し言葉など)を用いて、送り手と受け手の間で信号やメッセージを伝達する一連の作業を指します。また、コミュニケーションは私たちが人間関係の構築や修正に使う方法でもあります。自分の発言が会話を有意義なものにできるのだということを理解し、自信を持つようにしましょう。自分の意見や感情を的確に他人に伝えられるように日頃から自身に問いかける時間を作ります。価値ある発言ができないと思っているがために話すことをためらっているのなら、怖がる必要はありません。ある人にとっては重要性や価値が感じられることでも、それが他人にとっても重要だとは限らず、さらに別の人にとってはもっと大切なことかもしれません。高度なコミュニケーション能力を身に付けるためには、まずは簡単なやりとりから始めます。普段の人づきあいから職場まで、日常のいろいろな場面でコミュニケーション能力を磨くことができます。新しい能力をしっかりと身に付けるためには時間を要しますが、コミュニケーション能力を活用すればするほど、さまざまな可能性が開け、多くの新しい人間関係を築くことができます。
要約例	コミュニケーションとはそもそも何なのかを理解する。思っていることを口に出す勇気を持つ。練習する。
生成例	コミュニケーション能力を身につける。会話を有意義なものにする。日常のコミュニケーションから始める。

検証手法

■ BERTScore評価方法を試してみました。

BertScoreはbertを基づいた生成文の類似度を評価する方法です。

```
generated average P: 0.7596757986715862  
actual average P: 0.742866822414928  
generated average F1: 0.6798034192077698  
actual average F1: 0.6796998897242168  
<BarContainer object of 4 artists>
```



Generated average は生成文の類似度評価の平均値
Actual average はデータセット提供された要約文の類似度評価です

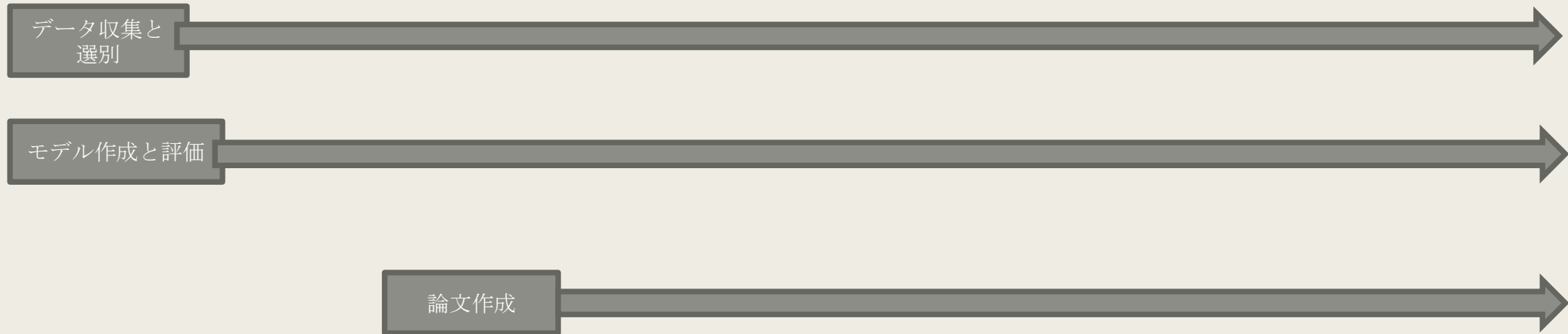
問題点

- 要約データセットが少ない
- データセットの関係で記事文章しか要約できない。

これからの計画

- 訓練データを選別してモデル評価する。
- 評価によってモデル。
- 要約文が不自然の対策を考える。

これからの研究計画



関連研究

- 事前学習モデルを用いた少量データに対する日本語抽象型要約

https://www.anlp.jp/proceedings/annual_meeting/2021/pdf_dir/P1-12.pdf

- Japanese Abstractive Text Summarization using BERT

https://www.astesj.com/publications/ASTESJ_0506199.pdf

- Hie-BART : 階層型 BART による文書要約

https://www.anlp.jp/proceedings/annual_meeting/2021/pdf_dir/B7-1.pdf

- BertSumを用いた日本語ニュース記事の抽象型要約手法の検討

https://www.jstage.jst.go.jp/article/pjsai/JSAI2021/0/JSAI2021_1D40S3c02/pdf/-char/ja

参考文献

- 日本語訓練済みモデル
 - <https://huggingface.co/sonoisa/t5-base-japanese>
 - <https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking>
- Textrank抽出型要約
 - [https://zenn.dev/shikumiya_hata/articles/b18e362e2eae09#%E3%83%91%E3%82%BF%E3%83%BC%E3%83%B32\(textrank\)](https://zenn.dev/shikumiya_hata/articles/b18e362e2eae09#%E3%83%91%E3%82%BF%E3%83%BC%E3%83%B32(textrank))
- Google T5モデル
 - <https://www.ogis-ri.co.jp/otc/hiroba/technical/similar-document-search/part7.html>
- ニュース記事のタイトル生成（一種の文章要約）
 - <https://qiita.com/sonoisa/items/a9af64ff641f0bbfed44>
- データセット
 - <https://www.rondhuit.com/download.html>
- wikiHowから日本語要約データを作成してみた
 - <https://tech.retrieva.jp/entry/2020/07/02/121208>



ご清聴ありがとうございました。

