日本オペレーションズ・リサーチ学会
2022年 春季研究発表会

# 二部グラフを用いた要約データセットの作成

05000341　株式会社エーアイスクエア　滝本直也　TAKIMOTO Naoya
*******　株式会社エーアイスクエア　呉 涛　Go

## 1. はじめに

ビジネス環境の変化が著しく、環境の変化に対応するために情報収集を行う必要がある。ネットニュースやプレスリリースは他社の状態や取り組みを把握するための重要な情報源である。金融の分野では自然言語処理の技術を利用することでニュースなどの一般に公開されている情報から株価の推移を予測する手法が提案されており、ネットニュースは一定の影響力と信用をもつソースとなると考えられている。そうした中で情報収集という点では文章量や冗長性などから原文のニュース記事は可読性が低い。そのためソースから特に重要と判断された情報をピックアップして要約文を作成し、情報の一覧性を上げた。ここで情報収集にかかるコストを詳細に把握すると、ニュース記事の収集、英語記事の日本語への翻訳、要約記事の作成と整理と複数のステップへと分けられる。これらは自動化によって省力化とスケールが可能である。また浮いた人員をより生産性の高い業務へと割り当てることによって会社としての生産性の向上を図る。

今回のケースでは情報収集のためのニュース記事要約システムを構築した。その内部で数理最適化の手法を利用することができたので以下ではその経緯について述べる。

## 2. 既存手法

文選択による要約の生成は教師なしの手法が基本である。要約の内容の冗長性を目的関数として、その最小化を行うことで必要最小限の文の選択を目指す。McDonald モデル、最小被覆モデル、施設配置モデルなどのアイデアが提案されている。これらのモデルは整数計画問題として記述されており、ソルバーを用いて解くことが一般的である。

また自然言語処理における DNN の飛躍的な進歩に合わせて事前学習済みの言語モデルを用いて文の自然さ (perplexcity) を評価し、それを手がかりとした冗長な単語の削除を行うモデルも提案されている。

教師なし学習は問題に含まれている一般的な性質を利用したものであるため広く適用が可能であるが、そこで用いた性質に含まれない特殊な要求を満たすとは限らない。そのため今回のケースでは教師あり学習をもちいた。

今回のケースでは要約文を作成するためのルールは明文化されたものではないにせよ、すでに共有されたものが組織内に存在していた。そのため、そのルールを踏襲して要約を行うモデルを作成するために、教師ありの二値分類モデルを作成することとした。

## 3. 文章要約システム全体像

今回、文章要約のために作成したシステムの全体像を以下に示す。

1. 教師データの作成
2. 各文のスコアを計算する二値分類モデルの学習
3. ニュース記事から重要文のスコア推定
4. 要約記事の出力

すでに得られたニュース記事と要約記事のペアをもとに二値分類モデルの学習用に教師データを作成する。作成した教師データをもとにニュース記事中の各文の重要度スコアを計算するモデルの学習を行う。学習済みモデルを用いて新規のニュース記事の各文に対して重要度スコアを計算する。スコア順にソートして、単語長や圧縮率などの指定したパラメータを満たすように抜粋を行う。ハイパーパラメータとして圧縮率の指定などが行えるようにした。今回は教師データの作成の段階において数理的手法を用いて適当なデータセットの自動生成を行った。

### 3.1. 分散表現による文書間の類似度評価

単語の特徴を表す方法として分散表現があり、これはベクトルによって文書の特徴を得るものである。word2vec[1] というモデルによって単語の学習した分散表現が得られる。$King - man + woman = Queen$ といった例から単語の意味を捉えた分散表

現が学習されると解釈している。この方法を応用することによって単語だけでなく、文書の分散表現も得られる。ここで分散表現の距離が近いものは近い意味を持っていると解釈する。意味空間上での距離として cos 類似度を用いた。

### 3.2. 抜粋要約モデル

今回の抜粋要約は各入力文に対して、重要か否かというラベルを付与し、重要な文を抜粋することによって要約文を生成する。この要約モデルは各文に対して、二値ラベルの予測をする分類モデルになる。この時、二値ラベルの分類を予測するためにラベルが付与された教師データが必要となる。今回、入手できたデータとしてネットニュースの全文と一部のニュースに対してはすでに作成された要約記事があった。そこで要約記事があるネットニュースについては教師データとして用いることとした。

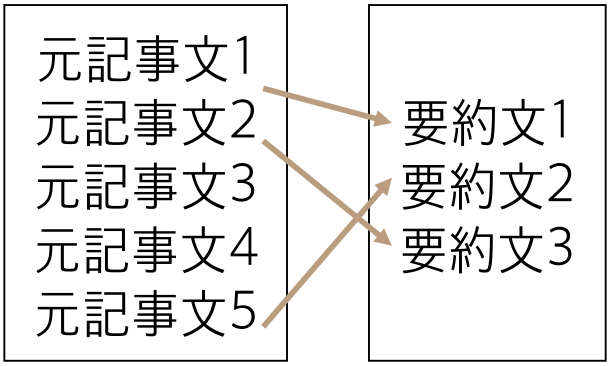


図 1: 元記事と要約記事の対応関係

ここで要約文と元記事の文との対応関係を二部グラフのマッチング問題に定式化することでラベルの付与を行った。

### 3.3. 二部グラフのマッチング問題への帰着

ニュース記事を要約する作業によって得られた要約記事が複数得られていた。そのモデルの訓練のために必要な教師データの作成のためにニュース記事の文と要約記事の文とのペアを作成した。そこで仮定として、

1. 元記事の抜粋が要約文として必要十分な情報を持つ
2. 要約文と対応する文が元記事に含まれる

とした。

文の特徴量として分散表現を用いて類似度を計算し、そのスコアの合計が最も高くなるようなペアの探索を行った。

定式化は以下のものとした。

$$\begin{aligned} \max \quad & f(x) = \textstyle\sum_{i,j}^{n,m} x_{i,j} v_{i,j} \\ \text{s.t.} \quad & \textstyle\sum_i x_{i,j} = n \\ & \textstyle\sum_j x_{i,j} = m \\ & x_{i,j} \in \{0,1\} \end{aligned} \tag{1}$$

ここで $x_{i,j}$ は元記事の文 $i$ と要約文 $j$ のペアが選択されたことを示す。$v_{i,j}$ は mUSE[2] による cos 類似度が既に得られているものとする。

## 4. 実験結果

python の pulp ライブラリを用いて上記の整数計画問題を解いた。またシステム全体としては AWS 上に構築されており、廉価な計算資源を用いて運用されている。詳細については発表にて詳しく述べる。

## 5. おわりに

ニュース記事による業界の情報収集を行っている。情報収集の時間削減のために記事の要約が行われていた。記事の要約を自動化した。自動化のために教師データの作成を二部グラフのマッチングに定式化した。自動化によって一定時間の削減が図れた。

また本システムでは機械翻訳を用いた多言語対応なども行われており、利便性の充実のための改修を随時行っている。

## 参考文献

[1] Efficient Estimation of Word Representations in Vector Space, Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, In Proceedings of Workshop at ICLR, 2013.

[2] Multilingual Universal Sentence Encoder for Semantic Retrieval, Yinfei Yang and Daniel Matthew Cer and Amin Ahmad and Mandy Guo and Jax Law and Noah Constant and Gustavo Hernández Ábrego and Steve Yuan and Chris Tar and Yun-Hsuan Sung and Brian Strope and Ray Kurzweil, ACL, 2020

[3] 文書要約のための数理的手法, 高村大也, 日本オペレーションズリサーチ学会機関誌, 2017,11 月号