2T-03

情報検索におけるConditional VAEを用いた要約文生成手法の有効性について

阿部　寛之 † 松原　雅文 ‡ Goutam Chakraborty‡ 馬淵　浩司 ‡

岩手県立大学大学院ソフトウェア情報学研究科 † 岩手県立大学ソフトウェア情報学部 ‡

1. はじめに

インターネットにある情報からユーザが目的とする情報を効率的に探すためには，検索エンジンを利用するのが一般的である．しかし，検索エンジンに入力した検索クエリが状況に応じて意味が変化する単語だった場合，Web 検索結果に目的とする情報以外のものを含んでしまう問題がある．

本研究では，Web 検索結果を単語が持つトピックごとにクラスタリングし，対象とする検索クエリについて自動生成した要約文を Web 検索結果とともにユーザに提示する．対象とする検索クエリについて要約文を付与することにより，どのトピックが自分の目的とする情報か曖昧な場合においても，情報発見を容易にすることができると考えられる．

本稿では，Conditional VAE が要約文生成においても有効であることを示す．Conditional VAE を用いることで，要約対象となる文章の文体を適切に保持したまま，各トピックに合わせた要約文の生成が可能になるものと考えられる．

2. 先行研究

2.1. Variational AutoEncoder (VAE)

Diederik らは，ニューラルネットワークを用いて次元削減を行う手法である AutoEncoder [1] を構成する潜在変数に確率分布を導入することで，ランダム性をもたせつつ次元圧縮後の潜在変数を求める手法を提案した [2]．

図 1 に VAE 全体のイメージ図を示す．データ X を入力とし，Encoder でデータがもつ潜在変数 z を学習する．潜在変数をサンプリングするのではなく，平均ベクトル μ と分散ベクトル σ^2 を用いて近似し，基のデータを復元した際の復元誤差と正則化を最適化することによってデータを学習している．

Effectiveness of Summarization Method using Conditional VAE on Information Retrieval

Hiroyuki ABE†, Masafumi MATSUHARA‡, Goutam CHAKRABORTY‡, Hiroshi MABUCHI‡

†Graduate School of Software and Information Science, Iwate Prefectural University Graduate School, ‡Faculty of Software and Information Science, Iwate Prefectural University

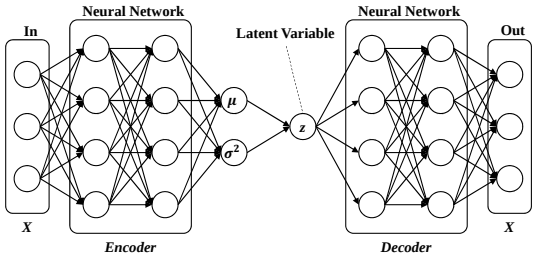

図 1: Variational AutoEncoder

2.2. Conditional VAE (C-VAE)

C-VAE は，VAE の Encoder と Decoder で行うデータの入力に，正解ラベルを付与して学習することによって，半教師あり学習で復元するデータを指定できるようにした手法 [3] である．

本研究では，Encoder には要約文の正解ラベル，Decoder には検索クエリがもつトピックのラベルを入力することにより，要約文がもつ文体を保持しつつ，トピックの要約文を生成する．

3. 提案手法

3.1. 概要

本研究では，Web 検索結果から作成したコーパスを用いて LDA によるクラスタリングを行うことで，検索クエリがもつトピックごとに検索結果を分類する [4]．

日本語版 Wikipedia の記事から学習した潜在変数 z と，Web 検索結果のコーパスから作成したトピックラベルを入力とし要約文を生成する．

3.2. Web 検索結果のクラスタリング

ユーザが入力した検索クエリについて，Google 検索結果上位 1,000 件分の Web ページを収集する．収集した Web ページを構成する HTML を解析し，本文データを抽出する．抽出した本文データを形態素解析し，分かち書きしたものをコーパスとする．

作成したコーパスを基に LDA を用いて，検索クエリがもつトピックごとにクラスタリングを行う．LDA によるクラスタリングを行う場合，トピック数を設定しなければならないが，今回は主観的に設定する．

3.3. 要約文生成

図 2 に要約文生成手法の概要図を示す.

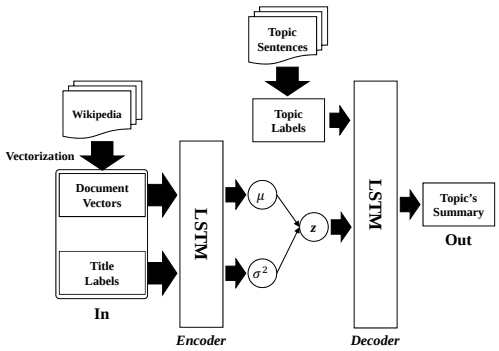

図 2: C-VAE を用いた要約文生成

日本語版 Wikipedia コーパスから生成した文書ベクトルと記事タイトルの one-hot ベクトルを Encoder への入力とする．先行研究とは異なり，Encoder と Decoder のネットワークには LSTM を採用する．Decoder には潜在変数 z と共に，生成したい Web 検索結果のトピック文から生成した one-hot ベクトルを入力する．

Encoder と違い，Decoder ではトピック文から生成したベクトルを正解ラベルとすることで，目的とするトピックの要約文を生成する．

4. 実験

4.1. 実験条件

C-VAE を用いた要約文生成手法が有効であることを示すための実験を行った．要約文を生成する検索クエリには,「指導者」,「読者」，そして「読み取り装置」のトピックをもつ「リーダー」に設定した．生成モデル用の学習データには日本語版 Wikipedia の記事データを使用する．

開発言語には Python 3.6，形態素解析器に MeCab，そして形態素解析用の辞書には mecab-ipadic-Neologed[2]を使用した．C-VAE を用いて学習，及びテストを行うために設定した各パラメータの値を表 1 に示す.

表 1: 学習用パラメータ

エポック数	30
Encoder のユニット数	300
潜在変数と Decoder のユニット数	600
バッチサイズ	45
サンプリング数	10
単語ドロップアウトの割合	0.5

4.2. 実験結果と考察

生成された各トピックごとの要約文を表 2 に示す.

[2]https://github.com/neologd/mecab-ipadic-neologd

表 2: 生成された各トピックの要約文

指導者	の指導者。
読者	本を読む。
読み取り装置	バーコードを読み取る。

「読者」と「読み取り装置」のトピックで生成された要約文はトピックを表す端的な文が生成された．特に「読者」のトピックから生成された要約文は,「読者」と「本」の潜在的意味の関連性が反映されており，良い要約文が生成されたと考えられる．しかし,「指導者」のトピックで生成された要約文は「指導者」という単語を含んではいるものの，助詞の「の」で文が開始されてしまった．これは学習において，構文情報などを特に利用しているわけではないことによるものであると考えられる．

5. おわりに

本稿では，LDA を用いて検索クエリがもつ潜在的トピックを考慮したクラスタリングを行った Web 検索結果に対し，C-VAE を用いた要約文生成手法を提案し，実験結果から C-VAE が要約文生成においても有効であることが示唆された．

実験時に設定した学習用パラメータは暫定的なものであるため，今後はパラメータを調整するとともに，他の検索クエリでの実験を行う予定である．

謝辞

本研究の一部は JSPS 科研費 15K00155 の助成を受けたものである．

参考文献

1) Geoffrey E. Hinton and Ruslan R. Salakhutdinov，”Reducing the Dimensionality of Data with Neural Networks”，Science 28，Vol.313，pp.504-507，2006.

2) Diederik P. Kingma and Max Welling，”Auto-Encoding Variational Bayes”，Proc. International Conference on Learning Representations (ICLR) 2014，Banff，Canada，April 14th-16th 2014.

3) Diederik P. Kingma，Shakir Mohamed，Danilo Jimenez Rezende and Max Welling，”Semi-supervised Learning with Deep Generative Models”，Advances in Neural Information Processing Systems 27 (NIPS 2014).

4) 阿部寛之，松原雅文，Goutam Chakraborty，馬淵浩司，”LDA を利用した Web 検索結果クラスタリング手法の有効性について”，平成 29 年度電気関係学会東北支部連合大会，1F11，August 2017.