Sentence-BERT を利用した FAQ 検索におけるデータ拡張手法

加納 渉
岡山大学大学院自然科学研究科
kano-w@s.okayama-u.ac.jp

竹内 孔一
岡山大学学術研究院自然科学学域
takeuc-k@okayama-u.ac.jp

概要

本研究では，質問応答における類似文章検索について，データ拡張を利用した検索精度向上を検討する．質問応答のデータセットは Yahoo!知恵袋データなどが公開されているが，特定の FAQ に対応したシステムを作る場合，データセットが十分に集まらないことも想定される．本研究では，人手でのデータ作成に加え，疑似文章生成モデルを利用してデータ拡張を行い，その検索精度への影響を調べた．実験の結果，拡張データを一定量混ぜた場合，わずかに精度が向上することが確認できた．

1 はじめに

質問応答システムでは，同じ単語でも文脈によって意味が違ったり，ユーザによって表現方法が変わるため，柔軟な質問応答の実現には単語一致のみでは十分ではないことが考えられる．類似文章検索手法としては，分散表現を用いるものがあり，分散表現を取得するモデル例として，Word2Vec[1] や Bidirectional Encoder Representations from Transformers (BERT)[2] がある. 分散表現を使用することで単語間の意味的な距離をベクトルとして測ることができる．本研究では，さまざまなタスクで SoTA を達成し，文章全体の意味を特徴化できるという機能を持つ BERT を使用し，類似文章検索を行う．

岡山大学情報統括センターのホームページ及び岡山大学 Moodle 上には，利用者の質問とその回答が掲載されている場所がある．このウェブサイトで将来的に質問応答システムを実現するため，既に寄せられている質問と回答をデータセットとして機械学習モデルに適用し，類似文章検索精度を向上させたい．しかしそのデータ数はあまり多くなく，質問もひとつずつしか付属していないため，検索モデル構築のためのデータセットとしては不十分なこと，ユーザの入力として想定される表現の幅が少ないことが問題点として挙げられる．本研究では人手による質問データに加えて，疑似文章生成モデルで疑似的な質問文を再現し学習データに追加することで，類似質問検索を行った．以下では，そのデータ拡張手法と，データ追加方法について報告する．

2 関連研究

高橋ら [3] はコミュニティベースの Q&A サイトにおける確率モデルを利用した類似質問検索手法を提案した．これは，質問検索では検索対象文書の長さが短いことを考慮し，適切に回答文情報を利用することで有効性を示した．

中野ら [4] は，対話的情報検索の問い返し文において，質問文の統語構造解析を行い，得られた部分木の一部を欠落させることで曖昧な質問文生成を行なった．生成した曖昧質問を利用し，BERT による質問応答システムを適用することで，質問応答精度を向上することができた．

Mass ら [5] は，FAQ 検索において 3 つのリランカーによる教師なし手法を提案した．ユーザクエリ-質問文，ユーザクエリ-回答間で BERT の Fine-tuning を行うとともに，質問文の言い換えを作成してさらにクエリ-質問文間 BERT を Fine-tuning することで，既存の教師あり手法と同等以上の性能を示した．

3 データセット作成

本研究で使用するデータは，岡山大学情報統括センターのホームページ及び岡山大学 Moodle から集め，整形したものである[1)]．元データの形式は質問と回答のペアとなっており，その中から学習データとして使用可能な 108 件を抽出したが，データセットとして明らかに少ないため，まず研究室の方々に依頼し，ある質問の回答に結びつくような質問を単語の組及び文章形式でそれぞれ 3 つずつ作成した．しかし，人手でのデータ追加はより実用的で自然な文章を考えることができるが，問題点として時間が

1） https://msgs.ccsv.okayama-u.ac.jp/a/faq.php

かかること，外部に依頼した際のコストがある．

本研究では，人手作成データに加えて疑似的な質問を Variational AutoEncoder (VAE)[6] によって生成し，学習データとして利用する．VAE の基本構造は Encoder と Decoder からなり，その間に潜在空間を導入する．潜在空間は確率分布が設定される．文章生成に使用した VAE モデルを図 1 に示す．本モデルは Encoder，Decoder にそれぞれ Transformer[7] の Encoder，Decoder ブロックを使用している Transformer-Based Conditional VAE (T-CVAE)[8] を利用して，入力文章を再現する．潜在変数には乱数を加えており，入力文章と少し異なる文章を出力しやすくなる仕組みとなっている．

このモデルに対し，元データの質問部分を用いて BLEU[9] スコアが約 70 程度になるまで学習させ，それぞれの質問について 10 パターンの疑似質問を生成した．BLEU スコアは語順を考慮しないため，スコアが高くても文章として自然であるか，元文章をそのまま再現しているかはわからない．本実験では生成文章の文法的正しさは問わないものとするため，入力文章とほぼ同じであったり，明らかに文法が破綻しているものも存在する．次に BERT に対して 2 値分類問題で Fine-tuning を行うため，ラベル付き文章ペアのデータセットを作成する必要がある．詳しくは 4 章で述べる．ここまでで，元データのある質問 1 つについて，

- 人手により作成した質問
  - クエリタイプの質問 3 件 (query1, 2, 3)
  - 文タイプの質問 3 件 (sentence1, 2, 3)
- VAE で作成した疑似質問 (10 パターン)

が存在する．元データのタイトル部分を anchor とし，それ以外の人手作成データ，VAE 生成データから類似質問 (positive)，非類似質問 (negative) を選び，組み合わせていく．つまり，ある anchor 一つに対し，(anchor, positive), (anchor, negative) の 2 種類のラベル付きデータが作られる．positive pair は，あるタイトルとそれから作られた人手作成データ，VAE 生成データを組み合わせ，negative pair はあるタイトルと関連のない全ての人手作成データ，VAE 生成データを組み合わせる．人手作成のうち，query3, sentence3 から作られる positive pair はテストデータに使用する．

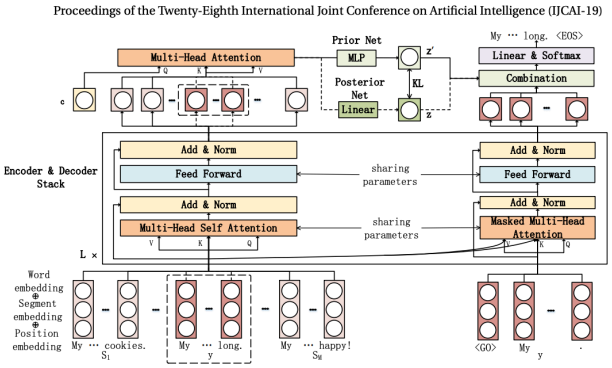

図 1　T-CVAE による文章生成，[8] より引用

4　実験

実験として，ある質問に対して最も類似した各回答の質問文を検索できるか評価する．質問と各回答に属する質問との類似度を計算するために機械学習を利用する．ここでは，2 値分類による BERT の Fine-tuning 及び検索精度評価，改善手法について述べる．

4.1　実験設定

文やクエリ同士の類似度を評価するモデル図は共通であり，図 2 に示す．これは Sentence BERT[10][2] と呼ばれる，文章ペアを利用した学習を行うためのモデルである．類似文章ペアを用いて Fine-tuning を行うとき，単一の BERT で毎回 2 つの文章を入力すると計算量は $O(n^2)$ となるため，学習時間が大幅にかかる．Sentence BERT の特徴は，Siamese Network の構造をとっていること，Pooling 層を導入している点である．2 つの入力の埋め込みベクトルを同時に計算することで，大幅な学習時間短縮を実現している．文献 [10] では，

- Classification Objective Function
- Regressioon Objective Function
- Triplet Objective Function

の 3 つの目的関数が紹介されている．本研究では，入力ペアが positive か negative を判定するため，Classification Objective Function を採用している．以下の実験での BERT はすべて，日本語 Wikipedia 事前学習済み[3] を使用し，Baseline は事前学習済み BERT を用いたものを表す．Sentence BERT の Fine-tuning の際のバッチサイズは 16，epoch は 10 とし，Pooling 手法は文献 [10] で最も高い評価となった MEAN を

2）　https://www.sbert.net/
3）　https://huggingface.co/cl-tohoku/bert-base-japanese

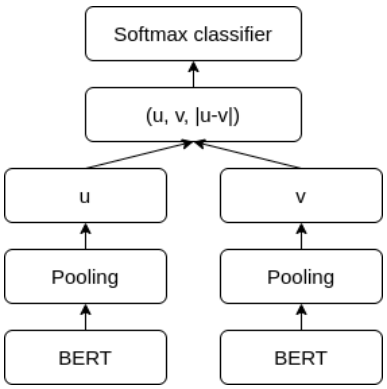

図 2　実験で使用する Sentence BERT モデル

採用する．

4.2　予備実験

予備実験として，元データのタイトル部分と人手作成データ，及び疑似質問を用いて Sentence BERT の Fine-tuning を行った．学習データセットのパターンは以下の通り．

- 人手作成データ
- 人手作成+疑似質問 (1 パターン)
- 人手作成+疑似質問 (3 パターン)
- 人手作成+疑似質問 (5 パターン)
- 人手作成+疑似質問 (10 パターン)

ある positive pair の 1 件に対し，negative pair は 107 件存在する．学習データの positive pair と negative pair の数を合わせるため，107 件から一つ選ぶ必要がある．予備実験での negative pair の選び方はランダムとする．

4.3　改善手法

改善手法として，ある positive pair に対応する negative pair の候補から，ペアの類似度が一番大きいものを選ぶデータセット，小さいものを選ぶデータセットをそれぞれ作成し，再度実験を行なった．事前学習済み日本語 BERT による埋め込みベクトルの cosine 類似度を用いるが，そのために BERT の出力をベクトルに変換する必要がある．本実験では時系列方向に平均をとったベクトルを質問のベクトルとする．

4.4　評価手法

Fine-tuning した BERT による埋め込みベクトルを使用して候補のランキングを出し，精度評価を行なった．本研究の場合は，入力にタイトル部分，回答候補は人手作成データの query3，sentence3 から選ぶ．つまり，一つの入力につき正解が 2 つ存在している．評価指標は 1 位正解率，5 位正解率，Mean Reciprocal Rank (MRR) とする [11].

1 位正解率は，回答候補ランキングを出した時，1 位に正解がある割合である．

5 位正解率は，ランキングを上位 5 位まで出した時，その中に正解がある割合である．今回正解は入力 1 つにつき 2 つあるが，上位 5 位までに 2 つ入っていたとしても 1 つしか入っていない場合と評価は変わらないこととする．

MRR は正解となる回答のランクの逆数をポイントとし，その平均をとったものである．つまり，MRR が高いほどランキングの上位に正解が出ていることになる．今回は上位 5 位までのランキングで評価し，その中に入っていなければポイントは 0 とする．

表 1　予備実験結果

	Manual	VAEpt1	VAEpt3	VAEpt5
1 位正解率	0.606	**0.616**	0.537	0.481
5 位正解率	0.741	0.727	0.644	0.685
MRR	**0.66**	**0.66**	0.577	0.56

	VAEpt10	Baseline
1 位正解率	0.593	0.583
5 位正解率	**0.745**	0.75
MRR	0.651	0.645

4.5　結果

予備実験結果を表 1 に示す．予備実験では，ランダムで選んだ negative pair を追加した．Baseline よりも，人手作成データ及び疑似質問を加えた場合のほうが精度が高くなっている箇所が見られるものの，あまり一貫した精度の向上が見られない．

次に，改善手法を適用した場合の実験結果を表 2 と表 3 に示す．表 2 を見ると negative pair に類似していないものを選んだ場合はどの結果においても精度が下がっており，Baseline よりも全てにおいて低く，特に疑似質問を加えた場合に下がりやすくなっているのが読み取れる．

一方，表 3 を見ると，疑似質問を加えたほうが，人手作成データのみよりも精度は良くなる傾向にある．特に，疑似質問を 5 パターン分加えた場合，どの評価指標においても高い値を得ている．さらに，表 1 において，疑似質問を 5 パターン追加した結果

表 2 negative pair(dessimilar) を選択した場合の実験結果

	Manual	VAEpt1	VAEpt3	VAEpt5
1 位正解率	0.426	0.319	0.532	0.111
5 位正解率	0.634	0.468	0.722	0.231
MRR	0.501	0.37	0.601	0.153

	VAEpt10	Baseline
1 位正解率	0.458	**0.583**
5 位正解率	0.620	**0.75**
MRR	0.519	**0.645**

表 3 negative pair(similar) を選択した場合の実験結果

	Manual	VAEpt1	VAEpt3	VAEpt5
1 位正解率	0.491	0.537	0.435	**0.588**
5 位正解率	0.653	0.685	0.634	**0.755**
MRR	0.55	0.590	0.517	**0.650**

	VAEpt10	Baseline
1 位正解率	0.523	0.583
5 位正解率	0.685	0.75
MRR	0.587	0.645

を表 3 の同じ項目で比較すると，negative pair に類似度の一番高いものを選ぶ手法のほうが高い値を示している．

4.6 考察

以上の実験結果より，人手作成データに加え，VAE によって作成した疑似質問から 2 値分類用データセットを作成し，特に negative pair に一番類似度の高いものを選んだ場合，精度向上を確認できた．一方，表 3 の疑似質問を 10 パターン混ぜた場合を見ると，この選び方であっても学習データ数を増やせば精度が上がるわけではないことがわかる．negative pair に一番類似していないペアを選んだ場合，ランダム，または類似しているものを選ぶ場合よりも顕著に精度が下がった．これらの原因として，より異なるペアを学習する回数が増え，最終的な類似度の近いものを選ぶという，元々のタスクの目標に適合しない学習が進んでいるためではないかと考えられる．

5 まとめと今後の課題

本研究では，類似質問検索において学習データが十分に得られない場合のデータ拡張手法を提案した．学習を行なった VAE で生成した文章を用いて 2 値分類学習用データセットを構築する際，negative pair に類似度の一番高いものを選択した場合に，人手作成データのみよりも精度向上が確認できた．しかし，今回は岡山大学情報統括センター，及び Moodle の FAQ のみでの実験のため，その汎用性については検証の余地がある．今後の課題として，VAE による生成文の品質評価，回答部分の利用，Yahoo!知恵袋などの大規模なデータセットでも実験を行い，有効性を検証する必要がある．

謝辞

本研究でのデータ作成にご協力いただいた竹内研究室の諸氏に感謝する．

参考文献

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. **CoRR**, Vol. abs/1301.3781, , 2013.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, June 2019.

[3] 高橋輝, 高須淳宏, 安達淳. コミュニティベース Q&A からの類似質問検索手法. 情報処理学会全国大会講演論文集, 第 72 回, pp. 111–112, 2010.

[4] 中野佑哉, 河野誠也, 吉野幸一郎, 須藤克仁, 中村哲. 問い返し質問文生成によって曖昧性解消を行う質問応答システム. 言語処理学会第 27 回年次大会 発表論文集, 2021.

[5] Yosi Mass, Boaz Carmeli, Haggai Roitman, and David Konopnicki. Unsupervised faq retrieval with question generation and bert. **Association for Computational Linguistics**, 2020.

[6] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In **2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings**, 2014.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[8] Tianming Wang and Xiaojun Wan. T-cvae: Transformer-based conditioned variational autoencoder for story completion. In **Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19**, pp. 5233–5239. International Joint Conferences on Artificial Intelligence Organization, July 2019.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of

machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[10] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3982–3992. Association for Computational Linguistics, November 2019.

[11] 磯崎秀樹, 東中竜一郎, 永田昌明, 加藤恒昭. 質問応答システム. コロナ社, 2009.