D-011

対照学習による問い合わせに対するメンテナンス法の推薦

Recommending Maintenance Methods for Queries via Contrastive Learning

西　巧†　　島川 博光†　原田　史子†
Takumi Nishi　Shimakawa Hiromitsu　Harada Fumiko

1. はじめに

機械異常時には製作元の技術者が現地に赴いてメンテナンスすることが多く，手間になっている．ユーザがメンテナンスできれば，現地へ赴くことが本当に必要な時だけになる．また，赴かないといけない事例でも見るべき箇所がわからないとメンテナンスに時間がかかってしまう．

そこで本研究では，問い合わせに対するメンテナンス法を過去の記録から推薦する．現場において，メンテナンス記録の類似度の判別は難しく，ラベルがついているデータも少ない．本研究では，すべてがラベル付けされている場合と，ラベルがない場合での学習結果を比較し，ラベルのないデータへの対応を検討する基準を探索する．

2. 対照学習によるベクトルモデルの訓練

2.1 Sentence-BERT

本研究で扱うメンテナンス記録は自然言語で書かれており，表記揺れがあるので単語に注目するのではなく文の意味的類似性を適切に捉えることが重要である． Sentence-BERT は，文章に注目することができる．事前学習済み日本語向け Sentence-BERT モデルである "sonoisa/sentence-bert-base-ja-mean-tokens-v2" は，各文をベクトル化する．

2.2 対照学習

学習での損失には，CosineSimilarityLoss や TripletLoss，ContrastiveLoss，SoftmaxLoss などがある．TripletLoss 以外の手法では文章の類似度を用いるが，TripletLoss は各標本にラベルが付いていることを仮定している．Triplet Loss を縮小する学習法は，アンカー文，意味的に近い正例文，意味的に異なる負例文の 3 つの文の組を用いて，アンカーと正例の距離を縮めつつ，負例との距離を広げる．

3 テキスト分類によるメンテナンス法の推薦

3.1 トリプレットロスを最小化する過去の検索

過去事例と同様の対処が新しい不具合に当てはまることを期待して，不具合の症状を示す，ユーザからの新しい問合せに類似した過去のメンテナンス記録を検索する．しかし，企業での実メンテナンス記録にはラベルがなく，それを付加する手間は大きい．そこでラベルがないデータセットで学習できる 自己教師学習の精度を，すべてがラベルつけられた教師あり学習を比較する．手法概要を図 1 に示す．

文章レベルで類似性を判定する前に，単語レベルでの前処理を施す．前処理では MeCab で形態素解析を行い，ストップワードを除去する．前処理後の文章を sentence-BERT モデルで 768 次元のベクトルに変換する．

本研究では，全文章にラベルが付いている場合での教師あり学習結果と，ラベルがない文章が多くある場合での自己教師あり学習での学習結果を比較し，自己教師あり学習の能力を調査し，ラベルを手動でどこまでつけるべきかを検討する基準を設ける．

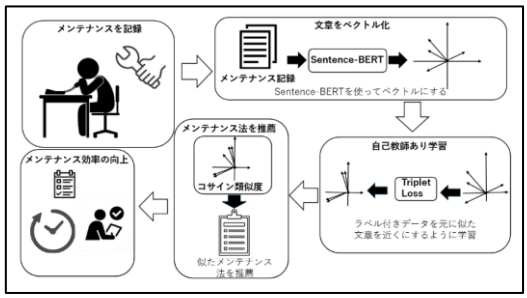

図 1　手法概要図

3.2 学習方法

本研究では教師あり学習と自己教師あり学習の 2 つの手法を使用して sentence-BERT を学習させる．

教師あり学習ではすべてのラベルがついていることを前提とする．アンカーとなる文章をランダムに取得し，同じラベルのものを正例に設定し，違うラベルのものを負例に設定する．この 3 個のデータのペアを作製してこのペアをもとにベクトル化モデルを学習させる．

自己教師あり学習では多くのデータにラベルがついていなくて，小数のデータにしかラベルがついていないデータを前提とする．小数のベクトル化した文章にノイズを追加してそのベクトルを正例に設定し，そのほかのラベルのものを負例に設定する．このようにして作成されたデータを使って TripletLoss を縮小する形でモデルを学習させる．

4 実験

本研究では，データとして日本語ニュース記事データを使用する．日本語ニュース記事データとして RONDHUIT 社が公開している「livedoor ニュースコーパス」[6] のタイトルとラベルの部分を使用する．本コーパスは，9 種類のニュースカテゴリに分類された約 7,000 件の記事で構成されている．データセットのタイトル部分を使用する．各入力文に対して，コーパス内の他の文とのコサイン類似度を計算し，類似度の高い上位 3 つの文章を取り出す．

4.1 教師あり学習

教師あり学習ではランダムに 1 個，文章を選びアンカーに設定する．同じラベルの文章の中からランダムに 1 個を正例として設定する．異なるラベルの文章の中から 1 個をランダムに選び負例に設定する．この方法で 32000 個のサ

†立命館大学情報理工学部データ工学研究室
Ritsumeikan University, Information Science and Engineering, Data Engineering Laboratory

ンプルを作製して学習を行った．損失関数をコサイン距離，マージンを 0.3，学習にはバッチサイズ 16，学習率 2e-5，エポック数を 10 で学習する．

4.2 自己教師あり学習

ニュースコーパスのデータセットではそれぞれのラベルから 5 個ずつサンプルとしてラベルを付けたまま置いておき，それ以外のデータのラベルを消しておく．

1 個のラベル付きデータに対して標準偏差 0.25 の正規分布に従うノイズを加える．これらを各ラベル付きデータ 10 個作成する．アンカーにラベル付きデータを設定し，正例に同じラベルの 4 個とアンカーベクトルにノイズを加えたもの 6 個，負例に別のラベルのデータを設定する．1 個のアンカーに対して正例を 10 個，負例を 10 個作成する．こうして，各ラベル付きデータに対して 100 個のペアを作成する．教師あり学習と同じパラメータで訓練する．

5 結果と考察

本研究でのベースラインとして，ファインチューニングを行っていない状態での sentence-BERT でベクトル化したデータを使用して xgboost でラベル分類を行った．データの 3 割をテストデータ，7 割を訓練データとして xgboost で学習，分類させた結果，正解率 0.74，適合率 0.75，再現率 0.74，F1-score 0.74 が得られた．

5.1 教師あり学習

教師あり学習を行った結果，正解率 0.99，適合率 0.99，再現率 0.99，F1-score 0.99 という結果になった．すべての値が約 0.33 上昇した．このベクトルを主成分分析で 2 次元空間にプロットする．学習前のベクトルが図 2，学習後のベクトルが図 3 である．

教師あり学習でのファインチューニングでは大幅な向上が見られた．本研究では 32000 のデータを作成して学習したが，何件のデータがあれば 99%の正解率を出すことができるのかを検証する必要がある．

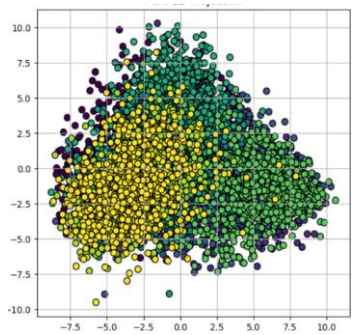

図 2　学習前ベクトル

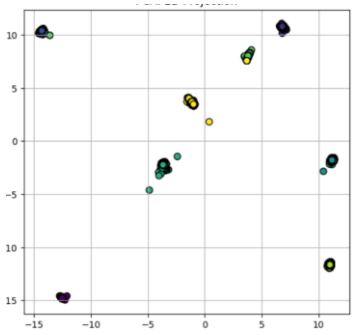

図 3　学習後ベクトル

5.2 自己教師あり学習

自己教師あり学習を実施後，ラベル付きのデータ 45 件で xgboost を学習した．その後，ラベルなしデータのラベルを予測した結果，正解率 0.32%，適合率 0.31，再現率 0.33，F1-score 0.27 になった．

この正解率では実用的ではない．ノイズを加える時の標準偏差を調整することと，ラベル付きとして残したサンプルデータの数を増やすことを検討しなければいけない．正例と負例の数を 10 個ずつで学習したが，合計で 4500 個のデータしかないため，教師あり学習での結果に及ばなかったと考えられる．

5.3 コサイン類似度

教師あり学習でファインチューニングされたベクトル化モデルを使用する．1 つのラベルから 5 文をサンプルとして取り出し，学習済みモデルでベクトル化し，コサイン類似度で似た文章を出力した．結果，同じラベルのものは類似度が 0.999 を超えて出力された．次に高かったラベルは 0.97 のものが出力された．続いて類似度 0.40 のものが出力された．その後の出力の類似度は 0.0 に近かった．

この結果から意味的な類似度を正しく捉えていると考察できる．しかし，異なるラベルの中で 1 番類似しているラベルの類似度が 0.97%と高いので，似ているクラスを近くに配置しすぎている．同じラベルと混同してしまうことが懸念される．

6 おわりに

本研究は，メンテナンスの問い合わせに対し，過去のメンテナンス記録からメンテナンス法を推薦する手法を提案した．実験結果から，ラベル付きデータが少ない場合の精度が低く，今後の改善が必要であることが示唆された．
今後は，自己教師あり学習の精度向上を目指し，ノイズの加え方や，別手法での正例の追加を検討していく．これにより，最適なメンテナンス法を推薦できるシステムの構築を目指す．

参考文献

[1] Chi Sun,Xipeng Qiu,Yige Xu & Xuanjing Huang, “How to Fine-Tune BERT for Text Classification?”,Conference paper, First Online: 13 October 2019 pp 194–206
[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton, A Simple Framework for Contrastive Learning of Visual Representations, Proceedings of Machine Learning Research ,1597-1607
[3] Nils Reimers, Iryna Gurevych, Sentence-BERT: Sentence Enbeddings using Simaese BERT-Networks Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), 3982–3992.
[4] vochicong, 「 Livedoor News Corpus 」 ,Kaggle,2023 https://www.kaggle.com/datasets/vochicong/livedoor-news （ 2025 年 6 月 1 日アクセス）
[5] sonoisa, “sentence-bert-base-ja-mean-tokens-v2,” Hugging Face, 2021. https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2 （2025 年 6 月 1 日アクセス）
[6] livedoor ニュースコーパス https://www.rondhuit.com/download.html