E-002

LOF を用いたドメイン外発話検出における データ拡張の有効性の検証

山村 崇[1]　真鍋 陽俊[2]　中谷 良平[2]　内田 佳孝[2]
[1] 九州工業大学　[2] ワークスアプリケーションズ

t_yamamura@pluto.ai.kyutech.ac.jp
{manabe_h, nakatani_r, uchida_yo}@worksap.co.jp

1　はじめに

Frequently Asked Questions (FAQ) とは，頻繁に尋ねられるような質問とそれらに対する回答を集めたペアであり，同じような質問に対する対応コストの削減やサービスの品質向上のために用いられている．回答を知りたいユーザがより容易に FAQ を発見できるように，ユーザの問い合わせに対して適切な回答を FAQ 内から検索するような FAQ 検索システム [12] の研究が行われている．また近年では，チャットボットを用いた対話的な質問応答システムの研究 [11] も行われており，ユーザの問い合わせに柔軟に対応可能なタスク指向型対話システムの開発が求められている．

ユーザの多様な問い合わせの特徴や傾向を把握する上で，ユーザとの対話を記録したログデータの分析が重要である．例えば，ログデータから現状のシステムが適切に応答することができないようなユーザの問い合わせを洗い出すことは，システムを改良する上で必要不可欠である．しかし，対話システムにおいて，ユーザは FAQ の内容とは全く関連のない文を入力する可能性が存在する．このようなドメイン外発話は，ユーザの問い合わせの特徴や傾向を分析する際のノイズとなるため，有効な分析の妨げになる．そのため，ユーザの入力文をドメイン外発話であるかどうかを検出することは重要な課題である．

これまでに，ドメイン内・外が明示的にラベル付けされたデータに基づいた統計的なドメイン外発話検出 [3] が取り組まれている．しかしながら，このようなラベル付きデータは十分に用意できない場合が多い．例えば，日々ログデータが収集されるがそれらにアノテーションするコストは高く，逆に FAQ 対話システムを新規に適用した直後などではログデータの数が少ないため，十分なラベル付きデータを確保することが難しい．

そこで，本研究ではドメイン外発話を FAQ データに対する外れ値と捉え，Local Outlier Factor (LOF) [1] によるドメイン外発話検出に取り組む．LOF は，特徴量空間における近傍のデータとの局所密度を計算することで外れ値を検出する手法である．また，ドメイン内・外のラベル付きデータを必要とすることなく，ドメイン内データのみを利用することで，ドメイン外発話の検出が可能である．

しかし，ドメイン内データが少ない場合に LOF ではドメイン内データの局所密度が低くなるため，外れ値であるドメイン外データを検出することが困難である．そのため本研究では，局所密度を高める手法としてデータ拡張に着目する．データ拡張は学習に用いるデータ数を増加させる手法であり，今回の場合，ドメイン内データ数を増やすことで，ドメイン内データ群の密度を高めることができると考えれられる．本研究では，同義語辞書を用いたデータ拡張の手法である Easy Data Augmentation (EDA) [9] を利用したドメイン外発話の検知に取り組む．実験結果より，EDA によるデータ拡張によりドメイン外データ検出の精度が向上することを確認した．

2　関連研究

Lane ら [3] は，ドメイン外データとドメイン内データを用いて入力文に対するトピックを推定することで，ドメイン外データの分類を行なっているが，彼らの手法ではドメイン内・外のラベル付きデータが必要となる．本研究では，ドメイン内･外のラベル付きデータは利用できないケースを想定している．ドメイン内データのみを用いてドメイン外発話を推定する手法として，Autoencoder [4] や Generative Adversarial Network (GAN) [5] を用いた手法がある．これらのニューラルネットによる手法は大量のデータを必要とする一方で，本研究で利用可能なドメイン内データは小規模である．

テキストデータのためのデータ拡張の手法として，逆翻訳 [10] や言語モデル [2] を用いた手法がある．し

かし，これらの手法はデータ拡張のためのモデルの構築が必要であり，計算コストも高い．

3 提案手法

本研究では，ドメイン内データとして FAQ データ D_{faq} を用いて，ログデータの発話文 $u \in D_{log}$ に対するラベル $l \in \{l_{ind}, l_{ood}\}$ を推定する．ここで，D_{faq} は FAQ データ全体の集合，D_{log} はログデータ全体の発話文の集合，l_{ind} と l_{ood} はそれぞれドメイン内発話 (in-domain) とドメイン外発話 (out-of-domain) を表すラベルである．なお，FAQ データは質問文 q と回答文 a が対のパラレルコーパス $D_{faq} = \{(q_i, a_i) \mid 1 \leq i \leq |D_{faq}|\}$ であり，本手法では質問文のみを用いる．

3.1 入力文の特徴量抽出

ログデータの各発話文 u 及び FAQ データの質問文 q に対して，分散表現を用いて特徴量抽出を行う．各文を形態素解析し，得られた n 単語の単語系列 $W = (w_1, w_2, ..., w_n)$ に対して，各単語の分散表現 $v_{w_i} \in \mathbb{R}^d$ について和を取ることで文の特徴量ベクトル $x \in \mathbb{R}^d$ を作成する [6]．なお，d は構築した分散表現の次元数を表す．

$$x = \sum_{i=1}^{n} v_{w_i} \tag{1}$$

3.2 Local Outlier Factor (LOF)

LOF は，近傍のデータの密度を利用することで，対象のデータ点の外れ値度合いを計算する．データ $x \in D$ の外れ値の度合いを $LOF_k(x)$ とすると，以下のように定義される．近傍点の個数 k はハイパーパラメータである．

$$LOF_k(x) = \frac{\sum_{x' \in N_k(x)} \frac{lrd_k(x')}{lrd_k(x)}}{\|N_k(x)\|} \tag{2}$$

$N_k(x)$ はデータ x の近傍 k 個の集合を表し，以下のように定義される．

$$N_k(x) = \{x' \mid x' \in D, dist(x, x') \leq dist_k(x, x')\} \tag{3}$$

ここで，$dist_k(x)$ は，データ x における k 番目に近いデータとのユークリッド距離を示し，$dist(x, x')$ は二つのデータ x, x' のユークリッド距離を表す．また，式 (2) において，データ x の局所密度 $lrd_k(x)$ は，以下のように表される．

$$lrd_k(x) = \frac{\|N_k(x)\|}{\sum_{x' \in N_k(x)} reachdist_k(x' \leftarrow x)} \tag{4}$$

$$reachdist_k(x \leftarrow x') = max\{dist_k(x), dist(x, x')\} \tag{5}$$

この到達距離 $reachdist_k(x \leftarrow x')$ は，x に対して $x' \in N_k(x)$ であれば k 距離の値をとり，それ以外のデータに対しては二点間の距離の値をとる．

式 (2) で示されるように $LOF_k(x)$ は，x の局所密度 $lrd_k(x)$ と k 近傍内のデータの局所密度の比の平均を外れ値の度合いとしている．つまり，x の外れ値の度合いが大きいときは，近傍の局所密度は高く，自身の局所密度が低い場合である．

3.3 データ拡張

3.2 節で示したように，あるデータ x の外れ値の度合い $LOF_k(x)$ が大きいときは近傍点の密度が高いときである．そのため，ドメイン内データの密度を高くすることで，ドメイン内データ同士がより密度の高いクラスタを形成し，ドメイン外データの外れ値の度合いが相対的に大きくなると考えられる．そこで，本研究ではドメイン内データの密度を高くするために，EDA [9] によるデータ拡張に着目する．

EDA は，小規模なデータセットに対して有効であることが示されており，同義語辞書のみで適用可能なデータ拡張手法である．EDA では，拡張する対象の質問文 q に対して以下の 4 つの操作をそれぞれ行う．

Synonym Replacement (SR)
文 q からストップワード以外の n_q 単語をランダムに選択し，それらを同義語で置き換える．

Random Insertion (RI)
文 q からストップワード以外の単語をランダムに選択し，その同義語を文中のランダムな場所に挿入する．これを n_q 回繰り返す．

Random Swap (RS)
文 q から 2 つの単語をランダムに選択し，それらの単語を位置を入れ替える．これを n_q 回繰り返す．

Random Deletion (RD)
文中の単語をランダムに確率 p で削除する．

SR, RI, RS において，n_q の値は質問文 q の文の長さ l_q に対して以下のように決定される．

$$n_q = \alpha \cdot l_q \tag{6}$$

α は，文中に含まれる単語の内，どのくらいの割合変化させるのかを示すハイパーパラメータである．また，RD において $p = \alpha$ として利用する．これらの操作を適用し，最終的には 1 つの文 q に対して n_{aug} 個の拡張された文の集合が得られる．n_{aug} は一つの質問文に対してデータ拡張によって生成する文の個数を表すハイパーパラメータである．なお，本稿で用いる文の特徴量ベクトルは語順に対して不変であるため，RS によるデータ拡張は行わない．

4 ドメイン外発話検出実験

FAQ 対話システムを構築し，対話システムに入力されたログデータの発話文を対象にドメイン外発話検出実験を行なった．本実験では，FAQ 対話システムに利用される FAQ データを LOF の学習データ，ログデータの発話文を開発・テストデータとして用いて分類を行なった．また，3.3 節で述べたデータ拡張を適用した場合との評価実験の比較を行なった．

4.1 データセット

FAQ 対話システムより得られたログデータの発話文として，ドメイン外発話 119 文とドメイン内発話 4,956 文を使用した．これらを 2:8 の割合で分割し，前者を開発データ，後者をテストデータとして使用した．また，FAQ データとしての 638 文の質問文を使用した．

4.2 実験設定

形態素解析器には Sudachi [7] を利用し，A 単位（短単位相当）で単語の分割を行なった．また，wikipedia データ[1]に対して Sudachi の A 単位で単語分割を行い，Skip-gram with Negative Sampling で学習した 300 次元の分散表現を利用した．LOF の実装には，scikit-learn[2]を利用し，ハイパーパラメータである近傍数 k は開発データに対するグリッドサーチにより決定した ($k \in [2, 4, 8, 16, 32, 64]$) ．なお，contamination パラメータは 0.01 とした．

[1] https://dumps.wikimedia.org/jawiki/
[2] https://scikit-learn.org/stable/index.html

表 1: ドメイン外発話検出の実験結果．

手法	EER
LOF（データ拡張なし）	0.444
LOF（データ拡張あり）	**0.347**

本実験では，データ拡張に用いる同義語辞書として，日本語 WordNet[3]より Synonym として定義されている関係を同義語辞書として用いた．また，データ拡張の際に用いるストップワードとして，SlothLib[4]で定義されている単語リストを利用した．

データ拡張のパラメータは，$\alpha = 0.1, n_{aug} = 9$ とした．つまり，データ拡張を行う設定では，FAQ データの 638 文の質問文に対して 9 倍の 5,742 文が生成され，合計 6,380 文を用いる．

4.3 実験結果

ログデータの発話文に対して，ドメイン外発話検出の評価実験を行なった．ドメイン外発話検出の評価指標として，Equal Error Rate (EER) [3] を用いた．なお，ERR は数値が低いほどドメイン外発話検出の性能が高いことを示す．

表 1 に，ドメイン外発話検出の実験結果を示す．データ拡張なしで LOF を適用した場合とデータ拡張をした場合を比較すると，データ拡張を行った場合に ERR が大きく向上した．これは，データ拡張を行うことで，FAQ データの局所密度が高くなり，式 2 で算出されるドメイン外発話の外れ値の度合いが相対的に高くなったためと考えられる．

5 分析

ドメイン外発話検出におけるデータ拡張の有効性を検証するために，可視化による分析を行なった．図 1 は，データ拡張を行なった場合の事例ごとの分類結果を表しており，本実験で使用した FAQ データとドメイン外発話の文の特徴量ベクトルを t-SNE [8] を用いて 2 次元に圧縮し可視化を行なった．図 1 より，FAQ データ（緑）から外れているドメイン外発話（赤）は正しく検出できている一方で，FAQ データに近いドメイン外発話（青）はドメイン内発話であると誤分類

[3] http://compling.hss.ntu.edu.sg/wnja/
[4] http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/Japanese.txt

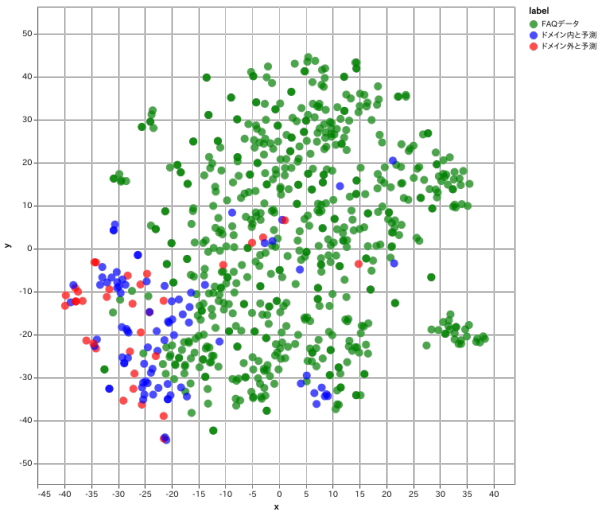

図 1: FAQ データとドメイン外発話の可視化．（緑）：FAQ データ[6]，（青）：ドメイン外発話をドメイン内発話と予測した事例，（赤）ドメイン外発話をドメイン外発話と予測した事例

している事例が存在していることが確認された．この誤分類された事例を分析した結果，ドメイン外発話ではあるが，FAQ データに多く出現する内容語が含まれている事例が多い傾向にあった．そのためこれらの事例は FAQ データと比較的近くなり，誤ってドメイン内発話に分類されたと考えられる．これより，構文情報や単語の系列情報などを考量した文の特徴量ベクトルの作成が今後の課題として挙げられる．

6 おわりに

本研究では，FAQ 対話システムに入力されたログデータの発話文に対するドメイン外発話検出を行なった．実験結果より，データ拡張を適用することでドメイン外データの分類精度が向上することを確認した．今後の課題として，FAQ データに多く頻出する単語を含むドメイン外発話の誤分類を改善するために，文の特徴量抽出の改善に取り組みたいと考えている．

参考文献

[1] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proceedings of the ACM SIGMOD Conference*, pp. 93–104, 2000.

[2] S. Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of NAACL-HLT*, pp. 452–457, 2018.

[3] I. Lane, T. Kawahara, T. Matsui, and S. Nakamura. Out-of-domain utterance detection using classification confidences of multiple topics. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 1, pp. 150–161, 2007.

[4] S. Ryu, S. Kim, J. Choi, H. Yu, and G. G. Lee. Neural sentence embedding using only in-domain sentences for out-of-domain sentence detection in dialog systems. In *Pattern Recognition Letters*, Vol. 88, pp. 26–32, 2017.

[5] S. Ryu, S. K. H. Yu, and G. G. Lee. Out-of-domain detection based on generative adversarial network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 714–718, 2018.

[6] D. Shen, G. Wang, W. Wang, M. R. Min, Q. Su, Y. Zhang, C. Li, R. Henao, and L. Carin. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[7] K. Takaoka, S. Hisamoto, N. Kawahara, M. Sakamoto, Y. Uchida, and Y. Matsumoto. Sudachi: a japanese tokenizer for business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[8] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, pp. 2579–2605.

[9] J. Wei and K. Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2nd Learning from Limited Labeled Data (LLD) Workshop*, 2019.

[10] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le. QANet: Combining local convolution with global self-attention for reading comprehension. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[11] 川端貴幸, 佐藤一誠. 意味と表記の組み合わせによる用例ベースの質問応答モデル, 2017.

[12] 牧野拓哉, 野呂智哉. 自動収集した学習データを用いた文書分類器に基づく faq 検索システム. 言語処理学会第 22 回年次大会 (NLP2016), 2016.

[6]図の簡略化のため拡張を行なった FAQ データは省略している．