分散表現を用いたビジネスメール自動分類器の検討

柴山 翔二郎[1,a)]　中本 千尋[2]　清水 剛[2,b)]　高柳 浩[3]　西田 眞[4]

概要：働き方改革を具現化するため、ビジネスパーソンの生産性を向上させるアプリケーション構築に取り組んでいる。メール種別を自動分類し、その種別に応じたタスク支援機能を提供することを目指している。本研究では、実際のメールを用いて分散表現空間を構築し、一部のメールに教師ラベルを付与した。教師データを学習することで、メールのタスク種類を分類する分類器を構築した。

Study of classifier construction for business mail with distributed representation

Shojiro Shibayama[1,a)]　Chihiro Nakamoto[2]　Takeshi Shimizu[2,b)]　Hiroshi Takayanagi[3]
Makoto Nishida[4]

1. はじめに

少子高齢化に伴う労働人口の減少に加えて労働者の働き方が多様化していることを受け、近年働き方改革が推進されている。ビジネスパーソンは従来よりも効率的に働き、従来と同等の成果を従来よりも短時間で生み出すことが求められている。

近年、企業による AI 活用が進み、社内に蓄積されたデータを活用したビジネス支援検討が急増している。筆者らは、タスク種類ごとにメールを自動分類することで気づきを与えタスク支援を実施することにより、業務の効率と質の向上に資するアプリケーション構築を目指している。

本研究では、ビジネスパーソンの最も重要なタスクのひとつである、日程調整のメールを分類対象とした。研究の実施にあたり実際にやり取りしたメールを収集し、その一部に教師ラベルを付与して教師データとした。教師データを学習し適切に予定調整メールが分類可能な分類器の構築を目指した。

[1] 東京大学大学院 工学系研究科, 東京都 文京区 本郷
[2] 日本ビジネスシステムズ株式会社, 東京都 港区 虎ノ門
[3] はこだて未来大学, 北海道 函館市 亀田中野町
[4] 放送大学秋田学習センター, 秋田県 秋田市 手形学園町
[a)] sshojiro@chemsys.t.u-tokyo.ac.jp
[b)] takeshi.shimizu@jbs.com

2. 分類器の実装

2.1 メールデータおよび教師ラベルの収集

日程調整メールを抽出するために、教師あり学習による分類器構築を実施する。本分類器の構築にあたり、社内営業担当者が保有する約 38 万件のメールデータを収集した。収集したメールデータからランダムサンプリングし教師ラベル (日程調整メールを正例、それ以外を負例とする 2 値) を 3049 件 (正例 266 件) の学習用データと 1814 件 (正例 160 件) の評価用データに手動で付与した。なお、教師ラベル付与の基準 (例えば日程を確定させるメールを日程調整に含むか否か等) は、予備検討を踏まえて決定した。

2.2 分類器構築と評価の枠組み

本研究で収集したメールの数に比較して、教師データ数は非常に少ない。そこで、教師なしのメールデータも活用してビジネスシーンのコーパスを学習し、分類器の精度向上を目指した。

本研究の流れを図 1 に示す。はじめに、メールに対して複数の処理の組み合わせを実施した。次に処理後に得られたメールに対して分散表現 (word2vec[1], fastText[2]) を計算し、ビジネスメールに含まれる単語同士の関係を学習した。各単語について計算された分散表現の平均をメール 1

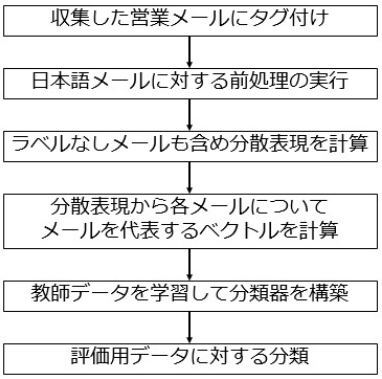

図 1 本研究の流れ

通に対応するベクトルとし [3]、その一部の教師データを学習させ分類器を構築した。なお、分散表現に基づくことで意味の似ている単語を持つメール同士のベクトルの類似度が高くなるように計算されることが期待される。最後に評価用データに対する分類結果に対して計算した、分類精度を示す指標を確認してモデルの評価を行う。

2.3 分類器の構築

教師ラベルは入力変数の線形結合では表現できない。そこで、ガウスカーネルを組み込んだ support vector machine（非線形 SVM）[4] によりモデル構築を行った。

非線形 SVM は、マージン最大化の際のペナルティ項の寄与の大きさとガウスカーネルの裾野幅の逆数の 2 つをハイパーパラメータに持つ。これらのハイパーパラメータを決定する際には、汎化性能を維持したまま学習データに対する予測性が高くなるように、k-fold クロスバリデーションによるグリッドサーチを実施する。

2.4 分類器性能評価の実施条件

単語の分散表現の次元数は分類精度が高くなるように 100 次元に設定した。前処理なし、記号の削除、活用語の原形化、記号の削除と活用語の原形化の計 4 通りの処理を施したメールを実験に用いた。処理の後のメールに対して計算された分散表現を使用して実験を行った。SVM のハイパーパラメータは、2-fold クロスバリデーションを実施して F1 スコアが最大となるように決定した。

分類器が日程調整メールを見落とすと、期待するタスク支援が実現しないため、再現率も重要な指標である。再現率とは、本研究では教師ラベルを付与された全ての日程調整メールのうち分類器により日程調整メールであると分類することができたメールの割合を示す指標である。

3. 実験結果と考察

構築した分類器を用いて評価用データを分類した結果を表 1 に示す。単語を原形化させて fastText により分散表現を学習した際に評価用データに対する F1 スコアが最も高くなった。このとき再現率も最も高い結果を示し、2.4 の要請に沿った結果が得られた。

表 1 処理の組み合わせと評価用データに対する分類精度の結果

手法	適合率	再現率	F1 スコア
前処理なし (word2vec)	0.67	0.82	0.74
記号削除 (word2vec)	0.73	0.86	0.79
原形化 (word2vec)	0.77	0.79	0.78
記号削除+原形化 (word2vec)	0.78	0.78	0.78
前処理なし (fastText)	0.76	0.85	0.80
記号削除 (fastText)	0.63	0.86	0.72
原形化 (fastText)	**0.78**	**0.86**	**0.82**
記号削除+原形化 (fastText)	0.75	0.80	0.77

記号を削除して fastText を適用すると、多くのメールが予定調整メールに分類されて、適合率が低く再現率が高くなった。記号を削除した場合を除き、文献 [5] の報告と同様に word2vec に比べて fastText を用いたほうが、高い分類精度が得られた。

fastText は、subwording という機構により接頭辞や接尾辞を考慮した分散表現の計算が可能である。表 1 の結果は日本語について fastText が有効であることを示唆している。この理由として、意味が類似する単語は類似度の高いベクトルとして学習され、単語の意味が適切に学習された分散表現を用いてメール同士の関係がベクトルとして適切に表現できたことが考えられる。

4. おわりに

本研究ではビジネスコーパスを使ったメール分類器の構築について検討した。約 38 万件のメールを集めた上で分散表現を計算した。さらに、サンプリングしたメールに教師ラベルを手動で付与し、教師データを分類器に学習させた。複数の処理および分散表現学習方法を組み合わせることで、予測性の高いモデルを構築する方法を検討した。

今後はメールの処理手法およびモデル構築手法を改善することで、様々なコーパスに対応可能な分類器を構築して高精度な自動分類の実現を目指す予定である。

参考文献

[1] Tomas Mikolov; et al. "Efficient Estimation of Word Representations in Vector Space". *In Proceedings of Workshop at ICLR*, 2013.

[2] Bojanowski, Piotr; et al. "Enriching Word Vectors with Subword Information", *Transactions of the Association for Computational Linguistics*, 5, 135-146, 2017

[3] Dongwen Zhang; et al. "Chinese comments sentiment classification based on word2vec and SVM^{perf}", *EXPERT SYSTEMS WITH APPLICATIONS*, 42, 1857-1863, 2015

[4] B.E. Boser, I.M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers", Proc. COLT, 1992.

[5] 堅山 耀太郎,『Word Embedding モデル再訪』, オペレーションズ・リサーチ, 11, 717-724, 2017