〔経常研究 A〕

企業相談データおよび自然言語処理モデルを活用した技術支援チャットボットの研究開発

福田　純，福井　航，平田　一郎

1　目的

我々工業技術センターへの機器利用や技術相談の問い合わせは、時に曖昧であり、使用したい機器・技術の明確な指定がないことがある。現状では、ハローテクノ担当者がその要求を解釈し、使用すべき機器や担当研究員名を回答している。この業務を AI 技術により自動化できれば、DX（デジタルトランスフォーメーション）の良好なモデルケースを県内企業に展開できる。

2　実験方法

2.1　相談データベースを用いた応答システム

まず、我々は、工業技術センター職員が企業からの問い合わせを受けるたびに記入している、「相談データベース」を学習データとして利用することを考えた。相談データは、企業名・相談内容・相談分野・担当者などの 14 個のカラムを持つ、テーブルデータである。各カラムには文字列形式でデータが保存されている。この相談データベースデータを 2020/08/26 から 2022/05/30 までの期間、合計 239228 件分取得し、AI に学習させる説明変数および目的変数の選定を行った。これらのデータを用いて、相談内容に対して対応可能な担当職員名を返す、検索ベース（すなわち、文章分類）のモデルを訓練した。その際、文章（単語）をベクトル化する必要があるが、この手法は多岐に渡る。本研究では、BoW・TF-IDF・LSI の 3 つの単語出現頻度ベースの手法を用いて比較検討した。モデルの選定に関しては、LinearSVC・NaiveBayes・K-Neighbor の 3 つのアルゴリズムを比較検討し、最も精度の高いものを選定した。

2.2　単語の分散モデルを用いた応答システム

相談データベースを用いたモデルにおいては、相談データ内に存在しないにも関わらず問い合わせが想定される単語（例：「モーションキャプチャ」など）に対処することができない。また、相談データにおいては、問い合わせと機器名が 1:1 で対応しておらず、対応機器を予測することはできない。そこで我々は、単語の分散表現を利用した。単語の分散表現とは、単語を予測するニューラルネットワークの中間層を用いて単語をベクトル化する手法のことである。この手法であれば、Wikipedia など大規模な学習データで学習された学習済みモデルを流用することができ、入力として解釈できない単語が非常に少なくなる。今回は、学習済みモデルとして chiVe[1,2]（国立国語研究所作成データセット NWJC にて学習済み)を使用した。問い合わせ文を chiVe によってベクトル化し、同じく chiVe を用いて作成した研究員の研究ベクトル・機器ベクトルとのコサイン類似度により、問い合わせに対する担当研究員・対応機器を予測した。

3　結果と考察

3.1　相談データを用いた応答システム

相談データは、企業名・相談内容・相談分野・担当者などの 14 個のカラムを持つが、説明変数（AI の入力）には「相談内容」を選定し、目的変数（AI の出力）には、「担当者」を選定した（表 1)。「担当者」は不均衡な分布を持っていたため、データ件数が 1 クラス 2000 件以下となるように、ダウン

サンプリング（間引き）を行った。また、「担当者」は 40 クラスであった。

表 1　相談データ中の「相談内容」「担当者」の例

相談内容	担当者名
淡路産タマネギの成分分析について	吉田和利
無線モジュールの電波評価について	中里一茂

続いて、「相談内容」を複数の方法でベクトル化したところ、「BoW」が最も結果の安定性が高く、複数のモデルにおいても正解率が高かったため、ベクトル化手法は BoW に統一した。この条件下で、複数のモデルについて、担当者予測を行った結果が表 2 である。

表 2　モデルによる担当者分類正解率

	Linear SVC	Naive Bayes	K-neighbor	RNN/LSTM	baseline
train acc	96.10%	69.40%	69.10%	-	11.60%
test acc	75.0%	65.60%	61.00%	65.20%	11.60%

LinearSVC を用いたシンプルな手法が最も正解率が高く、そのテストデータ正解率は 75.0%であった。しかし、学習データにおける正解率との乖離は大きく、過学習が進んでいることが分かる。

3.2　単語の分散モデルを用いた応答システム

研究員一人に対し 3 個のキーワードを選定し、これらキーワードに対応する単語埋め込みベクトルの和を「研究員の研究ベクトル」とした。「研究ベクトル」が妥当なものになっているか確認するため、研究ベクトルの視覚化を行った。研究ベクトルは 100 次元のベクトルであるが、主成分分析を行って 2 次元に次元を落とすことで、平面上に位置を視覚化することができる。この手法により視覚化した結果が図 1 である。

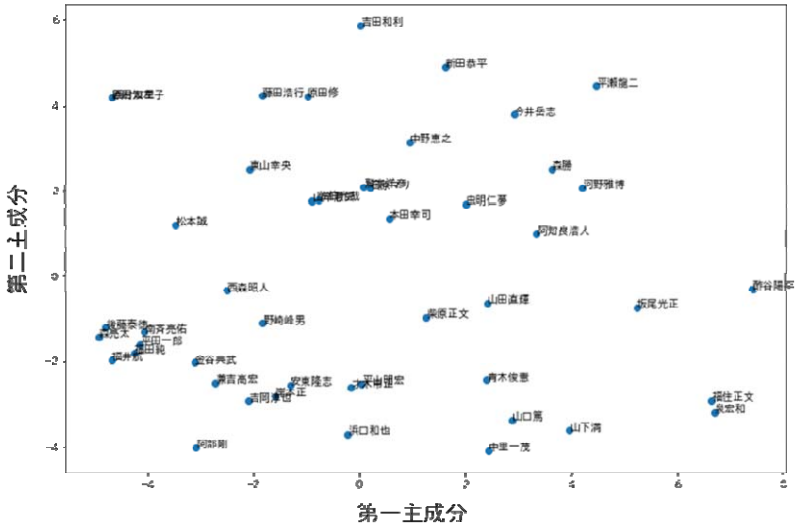

図 1　「研究ベクトル」の視覚化

直観的に研究内容が近いと思われる職員（例：福田・福井・平田など）が凝集しており、もっともらしいベクトル化ができていると考えられる。各種主成分の意味するところの解釈は困難であるが、第一主成分は「材料らしさ」第二主成分は「生物らしさ」のように解釈できると思われる。同様に研究員の担当機器を確認し、担当機器に対応する「研究員の機器ベクトル」を作成した。

3.3　単語の分散モデルを用いた応答システム

以上全てのモデルをまとめ、所内で利用できるウェブアプリケーションを開発した。質問文に対して「相談 DB モデルに基づく回答」「機器ベクトルに基づく回答」「研究ベクトルに基づく回答」の 3 つの観点から、質問に対応する研究員を AI が回答することができる（図 2）。

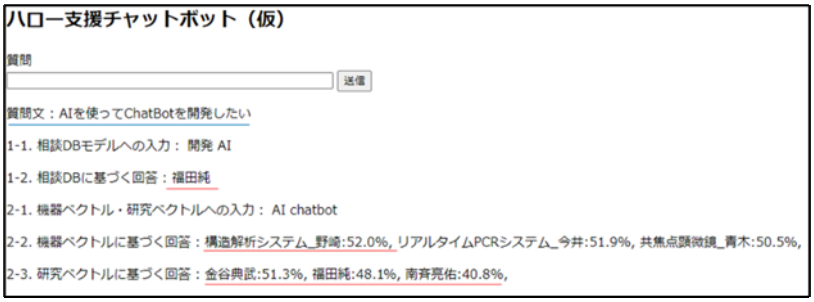

図 2　開発されたシステム

4　結論

本研究では、県内企業における DX 化の推進を念頭に、所内向け技術支援チャットボットを構築した。様々な自然言語を扱う手法を試行し、実用的なアプリケーションを構築することができた。今後、県内企業の自然言語処理に関する問い合わせがあれば、同様の手法で対応することが可能である。

参考文献

1）真鍋陽俊, 岡照晃, 海川祥毅, 髙岡一馬, 内田佳孝, 浅原正幸. 複数粒度の分割結果に基づく日本語単語分散表現. 言語処理学会第 25 回年次大会, 2019.
2）河村宗一郎, 久本空海, 真鍋陽俊, 高岡一馬, 内田佳孝, 岡照晃, 浅原正幸. chiVe 2.0: Sudachi と NWJC を用いた実用的な日本語単語ベクトルの実現へ向けて. 言語処理学会第 26 回年次大会, 2020.

（問合せ先　福田純）