舶用ディーゼル機関に関する問い合わせ対応用チャットボットのための類義語辞書の自動生成

美尾樹[1], 小原孝介[1], 多村新平[1], 今上博司[1], 佐藤功一[1], 滝澤一樹[1], 竹内孔一[2]
[1]株式会社三井E&Sマシナリー [2]岡山大学大学院 自然科学研究科
{tatsuru-mio, oharak, tamura_s, hiro4, kochan, kazuki_takizawa}@mes.co.jp,
takeuc-k@okayama-u.ac.jp

概要

技術的な問い合わせへの対応を行うチャットボットが必要とする，特定の分野における類義語を記述した辞書データを，自然言語処理技術を利用して自動生成する．マニュアルや実際の問い合わせ記録からword2vec埋め込みモデルを作成し,同コーパスに含まれる名詞の分散表現を得た．これらの名詞の間で類似度が高いものを2段階で判定することで，チャットボット用の類義語データを作成する．

1 はじめに

舶用ディーゼル機関は，多数のパーツからなる巨大で複雑な機械であり，過酷な条件で用いられることから,適切なメンテナンスは欠かせない.しかし，これらの機関は世界各地で昼夜問わず稼働しており，様々なタイミングで発生するトラブルに対し，技術的な支援を必ずしもタイムリーに提供できていない.さらに，洋上のような孤立した場所でトラブルが発生した場合，この問題はより深刻になる．

一方で，トラブルに関する問い合わせの内容は必ずしも高度な内容ではなく，中にはマニュアルやフローチャートを参照すれば対応できてしまうものも含まれる．こうした問い合わせへの対応を自然言語処理技術の活用により自動化できれば，アフターサービス担当者のスケジュールや時差に囚われず，タイムリーな回答が可能となる．さらに，定型的な対応が可能なトラブルに対応する負担を軽減することで，重点的な対応が必要な案件にリソースを振り向けることが可能となり，全体としてのアフターサービスの質を高めることが可能となると期待できる．

技術的な問い合わせへの対応を自動化するにあたっては，従来通りの自然言語による質問が可能なチャットボットを用いるのがサービスの質を維持する[1]ためにも望ましい．近年，チャットボットの普及は徐々に進んでおり，現在では対話に用いるデータを読み込ませるだけで活用できるパッケージ化された製品も利用可能となっている[1]．しかしながら，これらのソフトウェアは一般的な分野に対するテキストを処理するように調整されており，本研究対象の船用機関のように高度に技術的で，かつ狭い分野のテキストを扱うためには，その分野の類義語データや専門用語データを与えることが望ましい．しかし，専門家にとって専門用語は日常用いる言葉であるため用語であると意識することや，網羅的に書き出すことは簡単ではない．

先行研究において文書に対して word2vec を適用することで文脈が類似している類義語を獲得する手法が提案されている(例えば[3])．専門分野の文書にword2vecを適用し，得られたベクトル間の類似度を利用することで専門分野における類義語を獲得することが期待できる．

そこで，社内に蓄積されてきた専門的な内容の文章データ(コーパス)を形態素解析し，抽出した単語を基にチャットボット用辞書データを作成することを試みた．本稿ではその日本語の類義語辞書データの自動作成について述べる．

2 手法

2.1 コーパス

本手法で用いたコーパスは，(1) ディーゼル機関及び周辺装置の取扱説明書ならびに (2) 技術的な問い合わせとそれに対する返答の電子メール記録に由来する．(1)の取扱説明書はpdf形式やdocx形式のファイルとして存在し，分量は少ないながらも様式・文章ともに整っている．内容の網羅性・正確性も比較的高く，重要な専門用語も多く含むものであることが期待される．学習で利用できるようにテキスト部のみを抽出したところ，これらのデータは約4 MBの大きさとなった(UTF-8，以下同様)．(2)は，実際のトラブルや整備に関する問い合わせと，それ

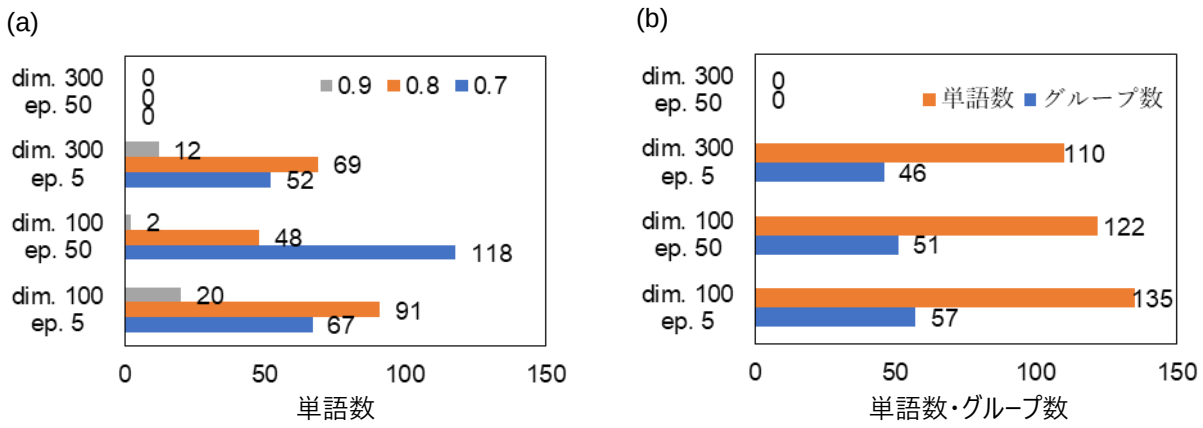

図 1.　(a) 各 word2vec モデルの t_w=0.7, 0.8, 0.9 における抽出類義語数 (b) 各モデルにより抽出された類義語及び類義語グループ数 (dim.: ベクトル次元数, ep.: エポック数)

に対する回答や提案などを記録した，約 11 年に渡る電子メールのやり取りである．こちらについては様式に個人差があり，誤字・脱字もしばしば見られる．しかし，これらはチャットボットが対応すべき業務そのものの生の記録であり，分量も比較的大きい(22 MB)ことから，実践的なデータとしての役割が期待される．和文・英文が混在していることから，仮名・漢字を含む文のみを抽出し，ルールベースの処理でメール特有の無関係な記述(署名や引用されたメールのヘッダを書き出したものなど)を削除した上で用いた．最終的に両テキストは合体させ，約 26 MB の単一のコーパスとした．

2.2 複合名詞のリスト

上記のコーパスから，類義語の候補となるすべての複合名詞を取得した．まず辞書として UniDic (cwj-3.1.0)を用いて MeCab による形態素解析を行い，一続きになっている名詞を収集した．この際，UniDic の詳細な品詞体系を活用し，専門用語によくみられる長い複合語を収集できるようにした．

2.3 単語の埋め込み

類似度算出のための単語埋め込みモデルの構築には，2.1 のコーパスで訓練された word2vec[i]を使用した．以降，単語間の類似度の算出にはコサイン類似度を用い，複数単語からなる語については，構成する単語のベクトルの単純な和を語全体のベクトルとした．

なお，当初は複数単語からなる語を含む未知語の扱いに便利な fastText[ii]を使用することも試みたが，おそらくコーパスのサイズ不足のため語の意味的な類似度を正しく算出できなかった．

2.4 類義語ペアの判定

複合名詞リストの複合語を構成する単語同士の類似度を，上記の埋め込みモデルでのコサイン類似度として 2 段階で算出し，類義語と判定される語のペアを求めた．判定の第 1 段階では，リストに含まれる名詞句のうち，共通の単語を含むもの同士の類似度を算出し，一定の閾値 t_p より高ければ第 2 段階に移る．第 2 段階では複合名詞の間で共通しない単語同士の類似度を算出し，一定の閾値 t_w より高ければこれらの単語を類義語ペアと認める．

例を挙げると，複合名詞のリストに「スティック 防止」と「固着 防止」があったとき，まず「スティック防止」と「固着防止」の類似度が t_p より大きく，かつ両名詞句で共通しない「スティック」と「固着」の類似度も t_w より大きければ，「スティック」と「固着」を類義語と見なすことになる．

2.5 類義語グループの出力

全類義語ペアを辺とするグラフにおいて，各連結成分の頂点の集合を 1 つの類義語グループとする．つまり，ある単語から類義語ペアをたどって結びつくすべての単語を類義語グループとする．最終的にこれを人手で修正することで，本稿の目的であるチャットボット用類義語辞書を得た．

[i] Gensim の word2vec を利用 https://radimrehurek.com/gensim/models/word2vec.html (2022/12/16 access)

[ii] Gensim の fasttext を利用 https://radimrehurek.com/2enism/models/fasttext.html (2022/12/16 access)

3 実験結果

以上の手順において，t_p=0.6 とし，t_w=0.7, 0.8, 0.9 として結果を比較した．また，類似度の算出に用いる word2vec モデルについても, ベクトル次元数とエポック数が異なるものを 4 種用いて比較した．

結果の評価には，出力された類義語グループのうち，妥当と判断されたグループの数と，それに含まれる単語の総数を用いた．なお，この手法では数十以上の単語が含まれる巨大な類義語グループが出力されることがあるが，これは類義語辞書作成者の負担を軽減するのに役立たないため妥当ではないと判定した．このようなグループは t_w を高くすることによって小さなグループに分かれ，活用できるようになる可能性がある．

まず，各モデル・閾値 t_w について抽出できた類義語数を図 1a に示した．次元数・ベクトル数ともに大きくしたモデル(dim. 300 ep. 50)では，技術的な類義語は抽出されなかった．残りの 3 つのモデルをエポック数の観点から比較すると，エポック数の小さいモデル(5 エポック)では t_w=0.8 の時に最も多くの類義語が抽出できているのに対し，エポック数の大きいモデル(100 次元 50 エポック)では t_w=0.7 の時に最大になっていることが分かる．これは，エポック数を深くするほど，類似度が低く算出される傾向を反映していると考えられる．次元数の観点から比較すると，100 次元のモデルはエポック数を増やしても結果が大きく変化しないのに対して，300 次元のモデルは全く類義語を抽出できなくなってしまっている．これは，今回のコーパスの規模に対して 300 次元が過剰で, 過学習を引き起こしたためと考えられ，コーパス数に合ったモデルを使用することの重要性を示している[iii].

次いで，各 t_w において抽出された類義語を，重複を排してモデルごとに集計したものを図 1b に示した．語数で見てもグループ数で見ても傾向は変わらず，100 次元 5 エポックのものが最も良い結果を示していることが分かった．しかし，これら 3 つのモデルの結果をさらに統合し，全モデルの全条件で得られた類義語の数を求めると，74 グループ 179 語が得られた．これはどのモデル単独よりも大きく，各モデルがそれぞれ異なった類義語の抽出に成功していたことを示している．これによれば，条件を変えながら複数モデルで抽出を行うことで，さらに多くの類義語を得られる可能性があり，より網羅的な類義語辞書を作成する上で役立てられる．

最後に，得られた類義語グループの例を挙げると(運航，運行，航海，航行，運転)，(ブザー，ゴング)，(解放，開放)などがあった. 第 1 の例は船舶の運航に関連する語群で，舶用ディーゼル機関にとっては重要かつ近い意味を持った語であるといえる．第 2 の例は一見奇妙に思えるが，実際に警告音としてゴングが用いられている箇所があることから，狭い分野での特有の類義語の抽出に成功していることを表している．第 3 の例は同義語ではなく誤字に近いものだが，電子メールのやり取りでは「開放」を「解放」と表記する例が非常に頻繁に見られ，その実態を反映した結果になっているといえる．チャットボットに入力する際もこの表記ゆれが発生する可能性は高く，登録しておいた方が利用者にとって便利であると考えられる．

4 まとめ

製品の取扱説明書と，製品に関する技術的な問い合わせの電子メール記録から単語の埋め込みモデルを作成することで，150 語以上の類義語の抽出に成功した．この際，モデルのパラメータを変更することで異なった類義語が得られることも明らかになった．今後は，今回作成した類義語辞書をチャットボットに登録して運用を開始するとともに，チャットボットの運用ログから未登録の類義語を発見することで，継続的に性能を改善できる機構を開発していきたいと考えている．

参考文献

[1] 木下, 薦田, 藤原. チャットボットを用いた社内情報サービス管理方式. FIT2020.

[2] 独立行政法人情報処理推進機構 AI 白書編集委員会「AI 白書 2019」(2019)

[3] 伴, 高橋, 位野木. word2vec を用いた同義語辞書自動作成手法の提案と適用評価，情報処理学会第 81 回全国大会，5N-04（2019）

[iii] サブワード分割のため word2vec より自由度の高い fastText モデルが良い結果を示さなかったこともこれに関連付けられる．