ELECTRA 単語分散表現と LightGBM を使った固有表現抽出

徳永 一輝 [1] 石田 希望 [1] 川端 篤 [1,3] 岩山 幸治 [2] 南條 浩輝 [2]

[1] 滋賀大学大学院データサイエンス研究科 [2] 滋賀大学データサイエンス学部
[3] 株式会社帝国データバンク
{s6023129, s6022102, s6023116}@st.shiga-u.ac.jp
{koji-iwayama, hiroaki-nanjo}@biwako.shiga-u.ac.jp

概要

専門分野における固有表現抽出は，ビジネスにおけるニーズが強いが，学習データを作成するアノテーションコストが高いことが課題である．本研究では，対象とするドメインとして自動車部品用語に着目し，少数の自動車部品用語の辞書を元にして，テキストデータから自動車部品用語を抽出し，自動車部品辞書を拡張する方法を提案する．人手でアノテーションした検証データで評価実験を行い，提案手法による固有表現の抽出と辞書の拡張の有効性を確認した．

1 はじめに

自動車業界においては，CASE (Connected / Autonomous / Shared / Electric) による技術変化，および 2050 年カーボンニュートラル宣言に伴う電動化の潮流に伴い，エンジン車から電動車 (BEV/PHV/FCV/HEV) に大きくシフトしていくと予想されている．電動車の中でもバッテリーのみで駆動する BEV (Battery Electric Vehicle) は，走行中に CO2 を全く排出しないことから，電動車シフトの本命と目されている．

BEV においては，バッテリーのほか，モーター・インバーターなど，エンジン車にはない新たな部品が必要となる一方，エンジンやトランスミッション，および関連する吸排気系・燃料系部品や，その制御部品などが不要となると言われている．この不要となる部品の数は，自動車に必要とされる部品約 30,000 個のうち約 11,000 個と想定されている [1]．自動車の電動車シフトに伴い，需要の減少が見込まれるエンジンやトランスミッション等に係るサプライヤーは大きな影響を受け，自動車産業に大きな変化が起きると予想される．

影響を受ける企業に対して早期の支援を行っていくためには，電動化シフトにより需要が増加・減少する部品を把握するとともに，自動車関連企業がどの自動車部品を扱っているかを把握しなければならない．そのために，新しく増える自動車部品に対応した自動車部品用語と，各企業が扱っている自動車部品のリストを収録したデータベースが望まれる．しかし，自動車に必要とされる部品は上記のとおり約 30,000 個にのぼる上に，BEV 等の技術進化により新たな部品が登場するため，全ての自動車部品を把握することは非常に困難である．

そこで，本研究では自動車部品用語の辞書の拡充に取り組む．具体的には，少数の自動車部品用語の辞書を元にして，自動車関連企業の事業内容のテキストデータから自動車部品用語を抽出し，自動車部品辞書を拡張する方法を提案する．

2 関連研究

自動車部品用語などの専門用語の抽出は，自然言語処理の固有表現抽出（NER）のタスクにおいて盛んに研究されている．

固有表現抽出における人名，組織名などの一般的な固有表現については，深層学習モデルを用いて，大量のトレーニングデータを学習したモデルが一般に公開されており，すでに非常に高い精度での抽出が可能となっている．株式会社リクルートの AI 研究機関 Megagon Labs と国立国語研究所との共同研究によりオープンソースとして公開されている ja-ginza-electra [2] では，人名，組織名のほか，地名，施設など 20 種類以上の固有表現の抽出が可能となっている．

化学や金融など特定ドメインにおける固有表現抽出は，ビジネスにおけるニーズが強く，応用の研究が進められている．福田ら [3] によると，特定ドメインの固有表現抽出の研究は，化学ドメインが最も多く，次いで医療・薬事ドメイン，企業情報・金融

ドメインとなっている．

特定ドメインにおける固有表現抽出における課題は，人手によるアノテーションコストが高く，学習データセットの用意が困難ということが挙げられる．そこで，辞書などの外部知識を用いて自動的に学習データセットを作成する Distant Supervision が注目されている [4]．単純な方法では，専門用語辞書を利用してテキストデータへ文字列マッチングを行うことで，人手のアノテーション無しに学習データセットを自動生成することができる．しかし，大量の専門用語を収録した辞書が必要である上に，辞書に含まれない単語は基本的に正解データとならないため，略称や表記ゆれにより学習データセットの質が低くなってしまうという問題がある．本研究の目的である自動車部品用語については，網羅的に収録された辞書は一般に公開されていない．

本研究では，アノテーションコストをかけずに，特定ドメインにおける新たな専門用語を抽出し，辞書を拡張することを目的とする．対象とするドメインとして，自動車部品用語に着目して取り組んでいる．

3 データセット

本研究においては，自動車部品用語のデータと，自動車関連企業の事業内容のテキストデータの 2 種類を使用する．ここで，自動車部品用語のデータは，全ての自動車部品を網羅したものではなく，一部のみを収録したデータである．

自動車部品用語のデータとして，マークラインズ社 [5] の公開する自動車部品一覧を使用する．このデータには，エンジンやトランスミッションなどの部品の他，プレス，鍛造などの加工法や，ADAS（先進運転システム）などの車載ソフトウェアの用語が階層構造で収録されている．用語の重複削除などの加工を行い，最終的に 1,138 個の自動車部品用語を用意した．

自動車関連企業の事業内容のテキストデータとして，株式会社 帝国データバンクの信用調査報告書の事業内容欄のテキストデータを使用する．事業内容には，その名の通り企業の事業内容に関する内容が記載されている．また，自動車関連企業の特定として，帝国データバンク産業分類の自動車関連業種（自動車製造，自動車車体製造，自動車内燃機関製造，自動車操縦装置製造，自動車部分品製造）と分類される企業のテキストデータを抽出した．抽出した企業数は 4,069 社，事業内容のテキストデータは平均 456 文字であった．

表 1 データセット

	text 件数	名詞数	うち、自動車部品用語数	
			辞書登録済	辞書未登録
train	4,069	398,300	15,901 (4.0%)	n/a[a]
eval[b]	500	46,278	2,265 (4.9%)	6,016 (13.7%)

[a]: 学習用データにはアノテーションを行っていないため自動車部品用語数は不明
[b]: eval は train に含まれる．抽出対象である辞書未登録用語は，train ではアノテーションされていない

辞書にはない新たな自動車部品用語を取得できているかを評価するために，4,069 社のテキストデータの内 500 社のテキストについて，人手による自動車部品用語のアノテーション作業を行った．アノテーションツールには doccano [6] を使用した．作業は 2 人 1 組で行い，1 人がアノテーションを行い，もう 1 人がチェックするという体制で行った．500 件のテキストの名詞数は 46,278 個となり，うち自動車部品用語は 8,008 個（17.3%）であった．データセットの件数を表 1 に示す．アノテーションを行ったデータセット（eval）においては，全名詞 46,278 個のうち，既に辞書にある自動車部品用語が 2,265 個（4.9%），辞書にはない自動車部品用語が 6,016 個（13.7%）存在しており，辞書にはない自動車部品用語の方が多い．

4 提案手法

4.1 名詞とその埋め込み表現の抽出

自動車部品用語抽出モデルを学習するために，テキストに対して ja-ginza-electra [2] を用いて形態素解析を行い，名詞のみを抽出する．抽出した名詞のうち，自動車部品辞書に含まれているものには，自動車部品ラベル（ラベル 1）を，それ以外のモノには非自動車部品ラベル（ラベル 0）を付与する．用意した自動車部品用語の一覧は 1,138 個と限定的であり，テキストの表現のゆらぎもあるため，全名詞のうち自動車部品（ラベル 1）であるのは 4%のみであった．なお，ラベル 0 が付与された語の中には未知の自動車部品用語が含まれている可能性があるが，ここではそれを考慮せずモデル学習に利用する．

ja-ginza-electra は，GiNZA [7] と呼ばれる形態素解析器と ELECTRA [8] と呼ばれる BERT [9] の派生モデルを組み合わせて実装されている．GiNZA は品詞タグ付けや係り受け解析などのタスクに優れ

た性能を発揮する．BERT の事前学習である MLM では学習時に，マスクされた位置のトークンを予測して学習するのに対し，ELECTRA [8] は一部のトークンを置換してその置換の有無を予測して学習する．これにより ELECTRA は BERT よりも計算効率が高く，同等以上の性能を示している．この ja-ginza-electra で分割した各単語に対する埋め込み表現を取り出すと，ELECTRA が BERT の派生形であることから，その埋め込み表現は文脈を考慮した埋め込み表現となっているといえる．なお，実際には単語が複数の token に分割されていることがあり，その場合は対応する token の分散表現の平均をとって単語の分散表現とできる．

4.2 自動車部品用語抽出モデル

本研究では，大規模なテキストデータに対して学習を行うため，勾配ブースティング決定木モデルである LightGBM [10] を採用した．LightGBM への入力は，ja-ginza-electra によって得られた名詞に対する単語分散表現（768 次元）である．ある自動車部品用語は別の自動車部品用語が出現する文脈にも出現できると仮定すれば，自動車部品用語同士では，ELECTRA の embedding のような文脈を考慮した単語分散表現同士の類似性が高くなると予想され，自動車部品用語とそれ以外の語とでは単語分散表現の類似性は高くないと予想される．したがって，この単語分散表現をそのまま決定木へ入力するための特徴量ベクトルとすることを提案する．

本研究では LightGBM を 0 または 1 の値を出力する二値分類モデルとして学習する．出力 1 が自動車部品用語，0 がそれ以外の単語とするモデルである．推論時は，文中の各名詞に対してラベル 1 である確率を計算し，値が高いものから出力する．同じ名詞が複数個所で出てきたときは最も高い確率値を割り当てる．

5 実験

5.1 実験設定

LightGBM を用いた自動車用語の判別モデルの学習をさせるために，学習データの各名詞について，ja-ginza-electra の出力である単語 embedding を求めて，それを LightGBM に学習させる．その際に，単語 embedding に対して不均衡対策を実施する．

表 2 各モデルの評価（precision@k）

	Normal	Over Sampling	Under Sampling	SMOTE-ENN
k=100	0.830	0.810	0.820	**0.840**
k=300	0.713	0.730	**0.737**	0.727
k=500	0.654	0.646	**0.656**	0.638
k=1,000	**0.533**	0.526	0.521	0.505

5.2 不均衡データ対策

先述の通り，学習データにおける自動車部品ラベルがつけられた語は非常に少なく，非自動車部品ラベルがつけられた語との比率が不均衡である．不均衡データで学習した場合，もともとのラベル数が多いデータを予測しやすい識別器を学習してしまうことが多いため，その対策を行うことが多くの場合効果的である．本研究では，不均衡データの対策として，オーバーサンプリング，アンダーサンプリング，そして SMOTEENN [11] の 3 種類を用いて不均衡対策をしない場合と比較する．サンプリングを行なった後の自動車部品用語の割合は，オーバーサンプリングとアンダーサンプリングでは 50.0%，SMOTEENN では 55.9%となった．

SMOTEENN は，オーバーサンプリング手法である SMOTE [12] とアンダーサンプリング手法である ENN [13] の 2 つの手法を組み合わせたサンプリング手法である．SMOTE を用いて少数派クラスのサンプルを合成し，ENN を組み合わせてノイズのあるサンプルを取り除いている．これにより，不均衡なクラス分布への対処と同時にデータ品質の向上が期待される．

6 評価

本研究の目的は辞書にはない新たな自動車部品用語を抽出することであるため，アノテーションを行ったデータセットに含まれる名詞 46,278 語のうち自動車部品辞書に含まれていない 44,013 語を評価対象とした．モデルの出力は，単語（形態素）毎に自動車部品であるかどうかの確率となる．アノテーションした自動車部品用語は複数の単語から構成される場合がある． その際は形態素解析で分割された単語ごとにラベルを割り当て，単語単位で評価を行った．

不均衡対策を行わなかった場合と行った場合の辞書未登録の用語の検出精度を表 2 に示す．辞書未登録の名詞について確率値が上位の k 件 (k=100，300，500，1000) を出力し，自動車部品用語が含まれる割

表 3　各モデルによる確率上位の 20 単語

Normal			Over Sampling			Under Sampling			SMOTEENN		
word	label	probability	word	label	probability	word	label	probability	word	label	prob
ミッション	1	93.90%	ミッション	1	97.30%	ワッシャー	1	98.20%	ワッシャー	1	99.50%
鍛造	1	93.00%	グリル	1	97.00%	ミッション	1	98.10%	ミッション	1	99.50%
ワッシャー	1	92.80%	ワッシャー	1	97.00%	アーム	1	97.70%	アーム	1	99.50%
プレス	1	92.70%	ワイパー	1	97.00%	スプール	1	97.70%	ワイパー	1	99.50%
スロットル	1	92.60%	アーム	1	96.80%	フロント	1	97.60%	トランスミッション	1	99.40%
シェード	1	92.60%	シェード	1	96.80%	インパネ	1	97.60%	スライダー	1	99.40%
焼結	1	92.50%	スプール	1	96.70%	グリル	1	97.50%	プーリー	1	99.40%
鋳物	1	92.40%	スロットル	1	96.50%	ヘッドライト	1	97.50%	ニップル	1	99.40%
ワイパー	1	92.20%	サイドブレーキ	1	96.50%	トランスミッション	1	97.50%	ドアハンドル	1	99.40%
アーム	1	91.90%	ニップル	1	96.50%	ワイパー	1	97.40%	グリル	1	99.40%
ミッションギヤ	1	91.80%	スライダー	1	96.30%	スライダー	1	97.40%	リング	1	99.30%
トランスミッション	1	91.70%	リアサス	1	96.20%	**塗料**	**0**	**97.30%**	サイレンサー	1	99.30%
ハンドル	1	91.20%	トランスミッション	1	96.20%	キャリパー	1	97.30%	スプール	1	99.30%
ドアハンドル	1	91.10%	フランジ	1	96.10%	ショック	1	97.20%	スロットル	1	99.30%
アクセル	1	91.10%	キャリパー	1	96.10%	エアクリーナー	1	97.20%	マニホールド	1	99.30%
カウル	1	90.90%	フロント	1	96.10%	**アウター**	**0**	**97.20%**	アッセンブリー	1	99.20%
スプール	1	90.60%	**ホチキス**	**0**	**96.10%**	アクセル	1	97.00%	レール	1	99.20%
パッキン	1	90.40%	レール	1	96.00%	シリンダー	1	97.00%	シャーシ	1	99.20%
塗料	**0**	**90.40%**	鍛造	1	96.00%	スロットル	1	97.00%	キャリパー	1	99.20%
ホチキス	**0**	**90.40%**	ミッションギヤ	1	96.00%	シャーシ	1	96.90%	**ホチキス**	**0**	**99.20%**

合（precision@k）を求めた．なお，Normal は不均衡対策なしの結果である．k=100 では，SMOTEENN が 84%と最も割合が高く，k=1000 では Normal が 53.3%と最も割合が高い結果となった．

予備実験において，辞書登録済みの自動車部品用語の検出を行った際には不均衡対策の効果が大きかったが，辞書未登録の自動車部品用語の検出においては大きな差が見られなかった．いずれのモデルでも上位 100 件の結果は 80%を超えており，出力された候補単語を目視によって確認し辞書へ追加することは実用的と考えられる．追加した辞書を用いて同じ処理を繰り返すことで，また新たに候補単語が得られると考えられ，このような使い方が想定される．繰り返しの回数を減らすために，k=300, k=500, k=1000 におけるさらなる高精度化が課題であることを確認した．

各モデルにおける確率上位 20 単語を抽出した結果を表 3 に示す．ほとんどの単語が自動車部品用語となっており，より上位だけに絞れば人手による確認をほとんど不要とできる可能性も確認できた．

7　おわりに

本論文では，少数の自動車部品用語の辞書を元にして，自動車部品に関連する文書から，辞書にはない新たな自動車部品用語を効率的に抽出する手法を提案した．本提案手法を，1,138 個の自動車部品用語の辞書と 4,069 社の自動車関連企業の事業内容のテキストに適用した結果，高い確率で新たな自動車部品用語を抽出することができた．この新たな自動車部品用語を辞書に追加して再度モデルを実行することで，辞書を拡張しつつ，より多くの新たな自動車部品用語を抽出することができると考えられる．

今後の課題としては，自動車部品名だけではその部品が何であるかを理解することは専門家以外には難しいため，例えばエンジンやブレーキ，トランスミッションなどのように，部品を使用される部位などによってカテゴリ化しておくことが有用であると考えられる．そのために，部品名からカテゴリを予測するモデルを検討する．

謝辞

本研究の遂行にあたり，株式会社帝国データバンクの大里隆也様（滋賀大学データサイエンス・AI イノベーション研究推進センター特任講師）には多大なご助言，ご協力を賜りました．滋賀大学データサイエンス・AI イノベーション研究推進センター准教授の松島裕康先生にはに有益なコメントをいただきました，ここに深く感謝申し上げます．

参考文献

[1] 新素形材産業ビジョン策定委員会. 新素形材産業ビジョン〜我が国のものづくりを支える素形材産業、今後の目指すべき方向性を考える〜, 2013. https://www.meti.go.jp/policy/mono_info_service/mono/sokeizai/sinsokeizaivision.pdf(2024-01　閲覧).

[2] megagonlabs. GiNZA - Japanese NLP Library, 2023. https://megagonlabs.github.io/ginza/（2024-01 閲覧）.

[3] 福田美穂, 関根聡. 国内におけるドメイン依存の固有表現抽出の応用技術の調査. 自然言語処理, Vol. 30, No. 2, pp. 800–815, 2023.

[4] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In **Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP**, pp. 1003–1011, 2009.

[5] マークラインズ株式会社. 『部品辞典』1000部品網羅！クルマの材料・加工法, 2020. https://dictionary.marklines.com/ja/(2024-01 閲覧).

[6] Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. doccano: Text Annotation Tool for Human, 2018. Software available from https://github.com/doccano/doccano.

[7] 松田寛. GiNZA-Universal Dependencies による実用的日本語解析. 自然言語処理, Vol. 27, pp. 695–701, 2020.

[8] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6260–6270, 2020.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019.

[10] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. **In Advances in Neural Information Processing Systems**, pp. 3149–3157, 2017.

[11] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. **SIGKDD Explor. Newsl.**, Vol. 6, No. 1, p. 20–29, jun 2004.

[12] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority oversampling technique. **Journal of artificial intelligence research**, Vol. 16, pp. 321–357, 2002.

[13] Dennis L Wilson. Asymptotic properties of nearest neighbor rules using edited data. **IEEE Transactions on Systems, Man, and Cybernetics**, No. 3, pp. 408–421, 1972.