技術報告

自然言語処理技術活用による業務プロセス変革

Natural Language Processing for Business Process Innovation

岩 月 憲 一*　　赤 木 俊 夫　　平 野 弘 二
Kenichi IWATSUKI　　*Toshio AKAGI*　　*Koji HIRANO*

抄　　録

業務知識は文書の形で蓄えられているため，業務プロセス変革のためには自然言語処理技術を活用することが必須である。日本製鉄(株)では，鉄鋼分野に特化した自然言語処理技術を開発すべく，社内文書を用いて Skip-gram モデル及び BERT の学習を行った。本稿では，単語ベクトルモデルにより鉄鋼用語の類義語が獲得できることと，BERT により技術系文書の分類性能が向上することを示す。

Abstract

Since business knowledge is stored in the form of documents, the application of natural language processing to them is mandatory for business process innovation. At Nippon Steel Corporation, the skip-gram model and BERT have been used on in-house documents in order to develop natural language processing technologies for the steel industry. In this report, we show that the word vector model helps to acquire synonyms of technical terms of the steel industry, and classification of technical documents improves using the BERT.

1. 緒　　言

業務に関する知識の多くは，文書の形で蓄えられている。日本製鉄(株)においては，例えば，長年にわたって蓄積された研究開発に関する文書，各製鉄所において記録された設備保全に関するレポート，大量の資機材購入に伴って作成された仕様書等である。

こうした文書をただ蓄えるだけでなく，業務プロセス変革に活用していくためには，自然言語処理技術の適用が必要不可欠である。そのため日本製鉄では，これまで文献データベースの構築[1]，全文検索システムの開発[2]，操業トラブル報告書からの操業知識の抽出[3]等に取り組んできた。

近年，自然言語処理技術の進展はめざましく，word2vec[4]や BERT[5]の登場によって，単語や文の意味を計算機で取り扱うことが容易になった。例えば単語ベクトル（単語の意味を数値によって表現したベクトル）によって，従来莫大なコストをかけて辞書を作成し定期的にメンテナンスを行う必要があった部分をある程度置き換えることができ，あるいは文ベクトル（文の意味を数値によって表現したベクトル）によって類義語や表記ゆれを集約することなく分類が行えるようになった。

これらの手法により学習された一部のモデルは公開され，一般に利用可能である。例えば，Skip-gram モデルでは，“日本語 Wikipedia エンティティモデル ”[6]や chiVe[7]が，BERT モデルでは，京都大学[8]や東北大学[9]が作成したモデルが公開されている。

このように一般公開されているモデルは，Wikipedia を代表に様々な分野の語彙を含む文書を用いて学習されているため汎用性がある一方で，鉄鋼業のような特定の分野における文書に適用するには，不十分な部分がある。そのため，決算短信[10]や医学論文[11]といった特定分野の文書でモデルを事前学習することが行われている。

日本製鉄においても，研究開発分野，設備保全分野，知的財産分野の文書を用いて Skip-gram モデルを作成し，また鉄鋼分野の社内外の文書を用いて BERT モデル[12]を作成している。

Skip-gram モデルは，単語 1 語を入力し，その単語が現れる文におけるその単語の周辺数語を予測するニューラルネットワークによって学習されるモデルである。学習済みのモデルは，単語と単語の数値表現の対として取得できる。

Skip-gram モデルは文書検索に，BERT モデルは文書分類に用いられており，いずれも日本製鉄における業務プロ

* プロセス研究所　インテリジェントアルゴリズム研究センター　特別研究員　博士(情報理工学)　千葉県富津市新富 20-1　〒 293-8511
（現　(株)みらい翻訳　シニアリサーチエンジニア）

セスを大きく効率化することに寄与している。

本稿では，分野に特化したSkip-gram及びBERTモデルの学習と日本製鉄社内における当該モデルを用いた業務プロセス改革の実際について述べる。

2. 類義語の獲得による文書検索業務の効率化

あらゆる部門の業務プロセスにおいて，文書を検索することは欠かせないが，単なるキーワード検索では不十分であることも多い。例えば，研究文書を検索する場合，専門用語の表記ゆれ（略語を含む）があるため，網羅的な検索を行う際には思いつく限りのキーワードを列挙せねばならない。同様のことは特許調査においても生じている。また，資機材調達においても，過去の調達価格を調べるためには，資機材の同定が必要であるが，ここでも同義語・類義語の問題が生じる。

辞書を人手で作成するとなると，相当の労力がかかるだけでなく，メンテナンスの問題も生じる。そこで，単語ベクトルモデルを学習させることで，辞書作成に代える。こうすることで，辞書作成の手間を省き，かつ学習に用いる文書を差し替えるだけでメンテナンスを行うことができるようになる。ただし，類似性の正確さは人手作成の精度ほどではないため，それが許されるタスクで用いることが肝要である。言い換えれば，precisionよりもrecallが重視されるタスクにおいてより効果的である。

単語ベクトルモデルを使用するにあたっては，文書を単語の系列に変換せねばならない。しかし，単語とは何かという問題がある。文章を単語に分割するには，ふつう形態素解析を適用する。この場合，形態素，つまり意味の最小単位に分割される。多くの形態素解析器は辞書を有しており，これが最小単位に影響する。しかし，この分割で目的が満足されるのであれば，既に存在する辞書（あるいはシソーラス）を使えば良いのであって，辞書を作る必要がない。いま問題となっているのは辞書に収録されていない範囲の専門用語なのである。

そこで，辞書を用いない手法として，Byte Pair Encoding[13]（BPE）を適用した。実装にはSentencepiece[14]を用いた。これによって，文書内に頻出する文字列は，ある程度の長さであっても1トークンとして分割されるようになる。例えば，“連続溶融亜鉛めっきライン”という言葉は，辞書にmecab-ipadic-NEologd[15]を用いたMeCab[16]で処理すると，“連続/溶融/亜鉛めっき/ライン”と分割される。後述の方法で学習したBPEによる分割は“連続溶融亜鉛めっき/ライン”である。また，“超音波探傷装置”もそれぞれ“超音波/探/傷/装置”と“超音波探傷装置”である。このように，比較的長めの専門用語がより少ないトークン数によって表現される。そしてそれに対して単語ベクトルを割り当てることができるので，単語ベクトルの計算（例えば2ベクトルの内積をその絶対値の積で割った値であるコサイン類似度）によって容易に専門用語同士の関係性を調べることができる。

こうして分割された単語に対して単語ベクトルを割り当てるべく，Skip-gramモデルを適用した。計算の対象にした文書は3種類である。すなわち，日本製鉄に蓄積されている研究報告書等（以下，研究文書），同じく設備保全に関する報告書等（以下，設備文書），そして鉄鋼各社が出願人である特許文献（以下，特許文書）である。いずれも技術系の文書であるが，その目的と内容は微妙に異なる。

作成したベクトルを用いて，いくつかの単語と最もコサイン類似度（以下，類似度）の大きい単語を調べた。**表1**は“CGL”，**表2**は“ZAM”，**表3**は“モータ”，**表4**は“超音波探傷”，**表5**は“KR”，**表6**は“圧延”について類似度の高い単語を示した。

“CGL”（表1）はcontinuous galvanizing lineの略語である。研究文書では，最も類似度の大きい単語に“CAPL”が来ている。これは，連続焼鈍設備（C.A.P.L.®）のことであるから，工程というジャンルに属する単語同士が近くなるように学習されていると言える。それに対し特許文書ではCGLの同義語と言える単語が近くなっている。設備文書では設備名称が列挙されるが，第2溶融亜鉛めっき鋼板製造設備を意味する“2CGL”が1つの単語として抽出されているのは設備文書ならではと言える。

“ZAM”（表2）は高耐食めっき鋼板の商品名ZAM®である。研究文書においては商品開発に関する情報が載るため単語として抽出されているが，設備文書や特許文書では商品名が記載されにくいため単語として抽出されなかったと考えられる。

“モータ”（表3）はいずれの文書にも登場するが，研究文書では関連語と言うべき語句が列挙されているのに対し，設備文書では表記ゆれを含む同義語が列挙されている。研究文書ではある程度使われる語彙が統一されるのに対し，現場で記入される設備文書では略称や製鉄所によって異なる慣用表現が含まれるため，こうした現象が生じると考えられる。

“超音波探傷”（表4）について着目したいのは特許文書において“漏洩磁束探傷”，“超音波探傷装置”，“渦流探傷”といった極めてニッチな単語が抽出されている点である。特許文献において相当回数出現しているものと考えられる。

“KR”（表5）は溶銑予備処理設備を指す（Kanbara Reactor法によるため）。設備文書においては，設備名称が列挙されているが，研究文書においてはKR法の用途である脱硫を中心に製鋼工程における様々な処理の目的が列挙されている。

“圧延”（表6）について着目したいのは研究文書において“任延”，“庄延”という存在しない用語が抽出されている点である。これは，古い文献をスキャンしOCR処理を行った際の読み取りエラーに起因するものと考えられる。

表１　“CGL” の類義語上位５件
Top-5 synonyms of “CGL”

研究文書	設備文書	特許文書
CAPL	KAP	連続溶融亜鉛めっき
KAP	RCL	焼鈍炉
EGL	ETL	連続溶融めっき
CCL	EGL	直火
APL	2CGL	無酸化炉

表２　“ZAM” の類義語上位５件
Top-5 synonyms of “ZAM”

研究文書	設備文書	特許文書
GI	-	-
めっき鋼板	-	-
溶融亜鉛めっき鋼板	-	-
めっき鋼板の	-	-
SGL	-	-

表３　“ モータ ” の類義語上位５件
Top-5 synonyms of “モータ”

研究文書	設備文書	特許文書
モーター	MOT	モータの
電動	モーター	電動機
インバータ	電動機	モーター
トルク	モーターの	モータは
アクチュエータ	Mot	電動機の

表４　“ 超音波探傷 ” の類義語上位５件
Top-5 synonyms of “超音波探傷”

研究文書	設備文書	特許文書
探傷	-	探傷
非破壊検査	-	検査
検査	-	漏洩磁束探傷
欠陥検出	-	超音波探傷装置
探触子	-	渦流探傷

表５　“KR” の類義語上位５件
Top-5 synonyms of “KR”

研究文書	設備文書	特許文書
脱硫	鍋	-
KIP	1REDA	-
脱珪	REDA	-
脱りん	RH	-
脱燐	3RH	-

表６　“ 圧延 ” の類義語上位５件
Top-5 synonyms of “圧延”

研究文書	設備文書	特許文書
圧延の	圧延中	の圧延
任延	次材	圧延機の
庄延	通材	圧延の
圧延で	圧延開始	圧延における
の圧延	パス目	熱間圧延

以上のように，同じ鉄鋼分野における技術系文書であっても，文書の種類によって含まれる単語の分布は大きく異なっており，それぞれのモデルを作成し類義語に考慮した文書検索を行うことには意義があると考えられる。特に，網羅的な検索を必要とする場面，例えば，類似特許の検索，先行研究の調査，設備保全レポートの検索，設備購入仕様書の検索などにおいて，有用である。

3.　鉄鋼BERTによる文書分類業務の効率化

単語ベクトルに留まらず，文ベクトル，文書ベクトルが構築できると，文書活用の幅は一気に広がる。製造業においては，機器障害レポート[17]や設備保全レポートの活用[18]あるいは材料データベースの検索効率化[19]などが提案されている。

こうした例においては，日本語版 Wikipedia 等で事前学習された BERT が用いられていることが多い。しかし前節で示されたように，単語の分布は鉄鋼分野の文献でも大きく異なるのであるから，一般的な文章で事前学習されたモデルよりも鉄鋼分野の文章で学習されたモデルの方が各タスクにおいてより高い性能を発揮できると予想される。実際，金融分野においてそのように報告されている[20]。

これを踏まえ，鉄鋼分野の文書を用いて BERT の事前学習を行った。近年 BERT より規模の大きいモデルが多数提案されているが，鉄鋼分野の文章量は全分野の文章と比べれば少なく，数 GB 程度となることが見込まれたため，BERT を採用した。事前学習に用いた文章は，鉄鋼各社の特許及び技報というオープンなデータと，日本製鉄が所有する研究文書である。オープンなデータのみを用いて作成したモデルと，全てのデータを用いて作成したモデルの 2 つを用意した。また，トークナイザには SentencePiece（BPE）を用い，それぞれの文章を用いて学習した。

モデルの性能を評価するため，東北大 BERT（bert-base-japanese-whole-word-masking, トークナイザも同じ）との比較を次の 2 タスクで行った。1 つは，特許分類[12]である。特許明細書の一部を入力とし，それが属する技術分類を出力するマルチラベル分類問題である。もう 1 つは，社内の研究文書の分類である。これも問題設定は同様である。前者については 2 モデル，後者については 3 モデルを用いて比較した。いずれもファインチューニングを行い，評価を行った。

結果を**表 7**，**表 8** に示す。特許分類においては鉄鋼 BERT（NS-BERT）が汎用の BERT を上回った。また，研究文書分類においては，オープンなデータのみを学習したモデルよりも社内のデータを含めて学習したモデルの方がより良い性能を示した。

以上のように，社内の文書データを用いて学習した BERT モデルにより，より精度の高い文書分類が実現可能である。文書分類は人手で行うには負荷の大きい作業であ

表７　特許分類
Results of classification of patents

Model	Accuracy
NS-BERT (trained on open data)	0.592
General BERT (Tohoku Univ.)	0.563

表８　研究文書分類
Results of classification of in-house research reports

Model	Accuracy
NS-BERT (trained on all data)	0.582
NS-BERT (trained on open data)	0.574
General BERT (Tohoku Univ.)	0.564

り，その効率化に寄与している。

4. 結　　言

本稿では，鉄鋼分野の文書を用いてトークナイザを学習し，Skip-gram モデルの学習と BERT の事前学習を行った結果を報告した。単語ベクトルは文書の種類によって異なる方向性の類義語を示した。BERT については，汎用のモデルと比べて鉄鋼 BERT が文書分類においてより高い性能を発揮することを示した。

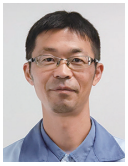
業務プロセス変革という観点においては，網羅的な文書検索において単語ベクトルが適用可能である。また，BERT は文書分類だけでなく，系列ラベリングによる情報抽出にも応用が可能であり，今後の活用課題である。

参照文献

1) 近江晶，明石安央：企業における社内データベースの構築．情報管理．30 (11)，1097-1111 (1988)
2) 中本伸也：特許検索システム PSEARCH/DB における大規模全文検索機能の開発．新日鉄技報．(366)，50-53 (1988)
3) 青山和浩，澤田崇弘，古賀毅，屋地靖人，森純一：操業知識マネジメントにおけるトラブル報告書のテキストマイニングに関する研究：連続鋳造プロセスにおける操業知識マネジメントの研究．日本機械学会年次大会．2012
4) Mikolov, T., Sutskever, I., Chen, K., Corrado, G., S., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26, 2013
5) Devlin, J., Chang, M-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 4171-4186 (2019)
6) Suzuki, M., Matsuda, K., Sekine, S., Okazaki, N., Inui, K.: A Joint Neural Model for Fine-Grained Named Entity Classification of Wikipedia Articles. IEICE Transactions on Information and Systems, Special Section on Semantic Web and Linked Data. E101-D (1), 73-81 (2018)
7) 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 回年次大会発表論文集．1407-1410 (2019)
8) 柴田知秀，河原大輔，黒橋禎夫：BERT による日本語構文解析の精度向上．言語処理学会第 25 回年次大会発表論文集．205-208 (2019)
9) https://github.com/cl-tohoku/bert-japanese
10) 鈴木雅弘，坂地泰紀，平野正徳，和泉潔：事前学習と追加事前学習による金融言語モデルの構築と検証．人工知能学会第二種研究会資料．FIN-028，132-137 (2022)
11) 杉本海人，壹岐太一，知田悠生，金沢輝一，相澤彰子：JMedRoBERTa：日本語の医学論文にもとづいた事前学習済み言語モデルの構築と評価．言語処理学会第 29 回年次大会発表論文集．707-712 (2023)
12) 岩月憲一：ドメインに特化した比較的少量のデータによる事前学習済み BERT の利用可能性：鉄鋼業における事例．言語処理学会第 28 回年次大会発表論文集．741-745 (2022)
13) Sennrich, R., Haddow, B., Birch, A.: Neural Machine Translation of Rare Words with Subword Units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 1715-1725 (2016)
14) Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 66-71 (2018)
15) 佐藤敏紀，橋本泰一，奥村学：単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討．言語処理学会第 23 回年次大会発表論文集．875-878 (2017)
16) Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 230-237 (2004)
17) 本間広樹，小町守，真鍋章，谷本恒野：BERT モデルを用いた障害レポートに対する重要箇所抽出．言語処理学会第 27 回年次大会発表論文集．189-193 (2021)
18) 高橋拓誠，谷口元樹，谷口友紀，大熊智子：設備保全レポート活用に向けたテキスト構造化の検討．言語処理学会第 28 回年次大会発表論文集．230-234 (2022)
19) 蓬田綾香，村瀬文彦，平野徹，三谷陽，坂一忠，飯田哲也，岩堀恵介，竹野貴法：技術ナレッジ活用に向けた Retriever-Reader モデルの検証．言語処理学会第 29 回年次大会発表論文集．2030-2033 (2023)
20) 鈴木雅弘，坂地泰紀，平野正徳，和泉潔：金融文書を用いた事前学習言語モデルの構築と検証．人工知能学会第二種研究会資料．2021

岩月憲一　Kenichi IWATSUKI
プロセス研究所
インテリジェントアルゴリズム研究センター
特別研究員　博士(情報理工学)
千葉県富津市新富20-1　〒293-8511
(現　(株)みらい翻訳　シニアリサーチエンジニア)

赤木俊夫　Toshio AKAGI
プロセス研究所
インテリジェントアルゴリズム研究センター
主席研究員
(現　日鉄テックスエンジ(株)　電計事業本部
開発企画部　部長)



平野弘二　Koji HIRANO
プロセス研究所
インテリジェントアルゴリズム研究センター
生産マネジメント研究室長　Ph.D.